

КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ имени АЛЬ-ФАРАБИ  
ИНСТИТУТ ВЫЧИСЛИТЕЛЬНЫХ ТЕХНОЛОГИЙ  
СИБИРСКОГО ОТДЕЛЕНИЯ РАН

ISSN 1560-7534  
ISSN 1563-0285

## СОВМЕСТНЫЙ ВЫПУСК

по материалам международной научной конференции  
"Вычислительные и информационные технологии в науке, технике и образовании"  
(CITech-2015)  
(24-27 сентября 2015 года)

# ВЫЧИСЛИТЕЛЬНЫЕ ТЕХНОЛОГИИ

Том 20

# ВЕСТНИК КАЗНУ им. АЛЬ-ФАРАБИ

Серия математика, механика и информатика № 3 (86)

## ЧАСТЬ I

АЛМАТЫ – НОВОСИБИРСК, 2015

# Вычислительные Технологии

2015

Том 20

## Редакционная коллегия

### Главный редактор

д.ф.-м.н., академик Ю.И. Шокин

### Ответственный секретарь

к.ф.-м.н. А.В. Юрченко

### Члены редколлегии:

Абдибеков У.С.	д.ф.-м.н., чл.-к. НИИ РК	Казахстан
Баутин С.П.	д.ф.-м.н., профессор	Россия
Бонту П.	профессор	Франция
Бычков И.В.	д.т.н., академик	Россия
Вонг Р.-Х.	профессор	Китай
Голушко С.К.	д.ф.-м.н.	Россия
Данаев Н.Т.	д.ф.-м.н., профессор, академик НИИ РК	Казахстан
Жайнаков А.	д.ф.-м.н., профессор, академик НИИ РК	Киргизия
Жумагулов Б.Т.	д.ф.-м.н., профессор, академик НИИ РК	Казахстан
Ковеня В.М.	д.ф.-м.н., профессор	Россия
Краузе Е.	профессор	Германия
Крейнович В.	профессор	США
Милошевич Х.	профессор	Сербия
Москвичев В.В.	д.т.н., профессор	Россия
Панченко В.Я.	д.ф.-м.н., академик	Россия
Потатуркин О.И.	д.т.н., профессор	Россия
Рознер К.	профессор	Германия
Рябко Б.Я.	д.т.н., профессор	Россия
Рэш М.	профессор	Германия
Смагин С.И.	чл.-к. РАН	Россия
Сойфер В.А.	чл.-к. РАН	Россия
Стемповский А.Л.	д.т.н., академик	Россия
Тайманов И.А.	д.ф.-м.н., академик	Россия
Темирбеков Н.М.	д.ф.-м.н., профессор, академик НИИ РК	Казахстан
Турицын С.К.	д.ф.-м.н., профессор	Великобритания
Федорук М.П.	д.ф.-м.н., профессор	Россия
Федотов А.М.	д.ф.-м.н., чл.-к. РАН	Россия
Хабаша В.Ж.	профессор	Канада
Четверушкин Б.Н.	д.ф.-м.н., академик	Россия
Чубаров Л.Б.	д.ф.-м.н., профессор	Россия
Шайдуров В.В.	д.ф.-м.н., чл.-к. РАН	Россия
Шокина Н.Ю.	к.ф.-м.н.	Германия
Шрёдер В.	профессор	Германия
Юлдашев З.Х.	д.ф.-м.н., профессор	Узбекистан

# Computational Technologies

2015  
Vol 20

## Editorial Board

**Academician Yuri I. Shokin** – Editor-in-Chief  
Institute of Computational Technologies SB RAS  
Academician Lavrentiev Ave. 6, Novosibirsk, 630090, Russia  
shokin@ict.nsc.ru  
Phone: +7(383)330-61-50, Fax: +7(383)330-63-42

**Dr. Andrey V. Yurchenko** – Managing Editor  
Institute of Computational Technologies SB RAS  
Academician Lavrentiev Ave. 6, Novosibirsk, 630090, Russia  
yurchenko@ict.sbras.ru  
Phone: +7(383)334-91-16, Fax: +7(383)330-63-42

Prof. U.S. Abdibekov, Kazakhstan  
Prof. Sergey P. Bautin, Russia  
Prof. Patrick Bontoux, France  
Prof. Igor V. Bychkov, Russia  
Prof. Boris N. Chetverushkin, Russia  
Prof. Leonid B. Chubarov, Russia  
Prof. Nargozy T. Danaev, Kazakhstan  
Prof. Michael P. Fedoruk, Russia  
Prof. Anatolii M. Fedotov, Russia  
Prof. Sergey K. Golushko, Russia  
Prof. W. G. Habashi, Canada  
Prof. V.M. Kovenya, Russia  
Prof. Egon Krause, Germany  
Prof. V.Kreinovich, USA  
Prof. Hranislav Miloshevic, Serbia  
Prof. Vladimir V. Moskvichev, Russia  
Prof. V.Ya. Panchenko, Russia  
Prof. Oleg I. Potaturkin, Russia

Prof. Michael M. Resch, Germany  
Prof. Karl G. Roesner, Germany  
Prof. Boris Ya. Ryabko, Russia  
Prof. Vladimir V. Shaidurov,  
Dr. Nina Yu. Shokina, Germany  
Prof. S.I. Smagin, Russia  
Prof. V.A. Soifer, Russia  
Prof. Wolfgang Shroeder, Germany  
Prof. A.L. Stempkovskii, Russia  
Prof. Iskander A. Taimanov, Russia  
Prof. Nurlan M. Temirbekov, Kazakhstan  
Prof. Sergey K. Turitsyn, UK  
Prof. Ren-Hong Wang, China  
Associate Professor Ziyavidin Kh. Yuldashev,  
Uzbekistan  
Prof. Amanbek Zhainakov, Kirgisia  
Prof. Bakytzshan T. Zhumagulov, Kazakhstan

## Редакционная коллегия

### Главный редактор

д.т.н., профессор, академик НАН РК Г.М. Мутанов

Научный редактор: М.А. Бектемесов – д.ф.-м.н., профессор, КазНУ им. аль-Фараби

Заместитель научного редактора: А.Б. Кыдырбекулы – д.т.н., профессор, КазНУ им. аль-Фараби

Ответственный секретарь: Г.М. Даирбаева – к.ф.-м.н., доцент, КазНУ им. аль-Фараби

### Члены редколлегии:

**Айсагалиев С.А.** – д.т.н., профессор, КазНУ им. аль-Фараби, Казахстан

**Алиев Ф.А.** – д.ф.-м.н., профессор, академик НАН Азербайджана, Институт прикладной математики Бакинского государственного университета, Азербайджан

**Д.Ж. Ахмед-Заки** – д.т.н., КазНУ им. аль-Фараби, Казахстан

**Бадаев С.А.** – д.ф.-м.н., профессор, КазНУ им. аль-Фараби, Казахстан

**Жайнаков А.Ж.** – д.ф.-м.н., профессор, академик НАН Кыргызской Республики, Кыргызский государственный технический университет им. И. Раззакова, Кыргызстан

**Кабанихин С.И.** – д.ф.-м.н., профессор, чл.-корр. РАН, Институт вычислительной математики и математической геофизики СО РАН, Россия

**Калтаев А.Ж.** – д.ф.-м.н., профессор, КазНУ им. аль-Фараби, Казахстан

**Кангуужин Б.Е.** – д.ф.-м.н., профессор, КазНУ им. аль-Фараби, Казахстан

**Майнке М.** – профессор, Департамент Вычислительной гидродинамики Института Аэродинамики, Германия

**Мальшикин В.Э.** – д.т.н., профессор, Новосибирский государственный технический университет, Россия

**Мейрманов А.М.** – д.ф.-м.н., профессор, Белгородский государственный университет, Россия

**Мухамбетжанов С.Т.** – д.ф.-м.н., профессор, КазНУ им. аль-Фараби, Казахстан

**Отелбаев М.О.** – д.ф.-м.н., профессор, академик НАН РК, Евразийский национальный университет им. Л.Н. Гумилева, Казахстан

**Панфилов М.** – д.ф.-м.н., профессор, Национальный политехнический институт Лотарингии, Франция

**Ружанский М.** – д.ф.-м.н., профессор, Имперский колледж Лондона, Великобритания

**Тайманов И.А.** – д.ф.-м.н., профессор, академик РАН, Институт математики им. С.Л. Соболева СО РАН, Россия

**Тукеев У.А.** – д.т.н., профессор, КазНУ им. аль-Фараби, Казахстан

**Шокин Ю.И.** – д.ф.-м.н., профессор, академик РАН, Институт вычислительных технологий СО РАН, Россия

**Юлдашев З.Х.** – д.ф.-м.н., профессор, Национальный университет Узбекистана им. М. Улугбека, Узбекистан

Научное издание

Вестник КазНУ

Серия математика, механика, информатика

№ 3(86) 2015

ИБ № 8514

Подписано 14.04.2015 г. Формат 60x84 1/8. Бумага офсетная.

Печать цифровая. Объем 30,5 п.л. Тираж 500 экз. Заказ № 2617.

Казахского национального университета им. аль-Фараби.

050040, г. Алматы, пр. аль-Фараби, 71, КазНУ.

Отпечатано в типографии издательского дома «Қазақ университеті».

Издательский дом «Қазақ университеті»

©КазНУ им. аль-Фараби, 2015



# The BULLETIN of KAZNU

Mathematics, Mechanics  
and Informatics Issue

2015

№ 3 (86)

## Editorial Board

*Dr. Sci. (Phys.-Math.), Prof., Academician Galymkair M. Mutanov* – Editor-in-Chief

Scientific Editor: *Dr. Sci. (Phys.-Math.), Prof. Maktagali A. Bektemesov, al-Farabi KazNU*

Deputy Scientific Editor: *Dr. Sci. (Tech.), Prof. Almatbek B. Kydyrbekuly, al-Farabi KazNU*

Managing Editor: *Cand.Sci. (Phys.-Math.), Assoc. Prof. Gullazata Dairbayeva, al-Farabi KazNU*

**Serikbai A. Aisagaliev** – *Dr. Sci. (Phys.-Math.), Prof., al-Farabi KazNU, Kazakhstan*

**Fikret A. Aliev** – *Dr. Sci. (Phys.-Math.), Prof., Academician of ANAS, Institute of Applied Mathematics, Baku State University, Azerbaijan*

**Serikzhan Badaev** – *Dr. Sci. (Phys.-Math.), Prof., al-Farabi KazNU, Kazakhstan*

**Darkhan Zh. Akhmed-Zaki** – *Dr. Sci. (Tech.), al-Farabi KazNU, Kazakhstan*

**Amanbek J. Jaynakov** – *Dr. Sci. (Phys.-Math.), Prof., Academician of NAS KR, Kyrgyz State Technical University named after I. Razzakov, Kyrgyzstan*

**Sergey I. Kabanikhin** – *Dr. Sci. (Phys.-Math.), Prof., Cor.-Member of RAS, Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Russia*

**Aidarkhan Kaltayev** – *Dr. Sci. (Phys.-Math.), Prof., al-Farabi KazNU, Kazakhstan*

**Baltabek E. Kanguzhin** – *Dr. Sci. (Phys.-Math.), Prof., al-Farabi KazNU, Kazakhstan*

**Victor E. Malyshkin** – *Dr. Sci. (Tech.), Prof., Novosibirsk State University, Russia*

**Matthias Meinke** – *Prof., CFD-Department of the Institute of Aerodynamics, Germany*

**Anvarbek M. Meirmanov** – *Dr. Sci. (Phys.-Math.), Prof., Belgorod National Research University, Russia*

**Saltanbek T. Muhambetjanov** – *Dr. Sci. (Phys.-Math.), Prof., al-Farabi KazNU, Kazakhstan*

**Mukhtarbay Otelbaev** – *Dr. Sci. (Phys.-Math.), Prof., Academician of NAS RK, Kazakhstan Branch of Lomonosov Moscow State University, Kazakhstan*

**Michael Panfilov** – *Dr. Sci. (Phys.-Math.), Prof., National Polytechnic Institute of Lorraine, France*

**Michael Ruzhansky** – *Dr. Sci. (Phys.-Math.), Prof., Imperial College London, United Kingdom*

**Yurii I. Shokin** – *Dr. Sci. (Phys.-Math.), Prof., Academician of RAS, Institute of Computational Technologies SB RAS, Russia*

**Iskander A. Taimanov** – *Dr. Sci. (Phys.-Math.), Prof., Academician of RAS, Sobolev Institute of Mathematics of SB, Russia*

**Ualsher A. Tukeyev** – *Dr. Sci. (Tech.), Prof., al-Farabi KazNU, Kazakhstan*

**Ziyaviddin Kh. Yuldashev** – *Dr. Sci. (Phys.-Math.), Prof., National University of Uzbekistan named by after Mirza Ulugbek, Uzbekistan*

## Table of Contents

---

### Session I. Technological Process Automation and Control

---

Solving Hard SAT Instances in Volunteer Computing Project SAT@home . . . . .	11
<i>I. Bychkov, S. Kochemazov, M. Manzyuk, I. Otpuschennikov, M. Posypkin, A. Semenov, O. Zaikin</i>	
A Scalable Parallel Algorithm and Software for 3D Seismic Simulation on Clusters with Intel Xeon Phi Coprocessors . . . . .	22
<i>D. Karavaev, B. Glinsky, V. Kovalevsky</i>	
Parallelization of Algorithm of Prediction of miRNA Binding Sites in mRNA on The Cluster Computing Platform . . . . .	28
<i>A. Yu. Pyrkova, A. T. Ivashchenko, O. A. Berillo</i>	
Distributed PIV: the Technology of Processing Intensive Experimental Data-flow on a Remote Supercomputer . . . . .	34
<i>V. Shchapov, A. Masich, G. Masich</i>	
Analysing Modal Behaviour of Hybrid Systems by One-step Parallel Methods . . . . .	43
<i>M. Nasyrova, Y. Shornikov, D. Dostovalov</i>	
Numerical Solution of Three-Dimensional Diffraction Problems Using Mosaic-Skeleton Method . . . . .	51
<i>S. Smagin, A. Kashirin, M. Taltykina</i>	
Seismic Field Simulation on High-Performance Computers in the Problem of Studying the Consequences of Underground Nuclear Tests . . . . .	61
<i>A. Yakimenko, D. Karavaev, A. Belyashov</i>	
The Experience of Implementation of Permutation Tests Using GPU . . . . .	69
<i>A. Yakimenko, M. Grishchenko</i>	

---

### Session II. Information Management, Processing and Security

---

Study of the Problem of Creating Structural Transfer Rules for the Kazakh - English and Kazakh-Russian Machine Translation Systems on Apertium Platform . . . . .	77
<i>B. Abduali, A. Sundetova, N. Zhanbussunov, Zh. Musabekova</i>	
Multicriteria Statistical Analysis of Test Biometric Data . . . . .	83
<i>B. Akhmetov, I. Aleksandr, Y. Funtikova, Z. Alibiyeva</i>	
Solving the Inverse Task of Neural Network Biometrics Without Mutations and Jenkins' "Nightmare" in the Implementation of Genetic Algorithms . . . . .	89
<i>B. Akhmetov, S. Kachalin, A. Ivanov, A. Bezyaev, K. Mukapil</i>	
Module of Lexical and Morphological Analyzer in the Development of Semantic Search Engine for Kazakh Language . . . . .	94
<i>Y.N. Amirgaliyev, A.S. Kalimoldayeva</i>	

Recognition of Isolated Words Using the Bayes' Theorem . . . . .	99
<i>E.N.Amirgaliyev, O.J. Mamyrbayev, T.A.Muratkhanova</i>	
Design of Automated Image Recognition System to Assess the Quality of the Mineral Species Using CASE Technology . . . . .	106
<i>O.E. Baklanova, A.E.Baklanov, O.Ya. Shvets</i>	
Software Implementation of the Cryptographic System Models with the Given Cryptostrength . . . . .	117
<i>R. Biyashev, M. Kalimoldayev, S. Nyssanbayeva, N. Kapalova, R. Khakimov</i>	
The Modified Digital Signature Algorithm Based on Modular Arithmetic . . . . .	122
<i>R. Biyashev, S. Nyssanbayeva, Y. Begimbayeva</i>	
Wireless Sensor Networks and Computational Geometry Problems . . . . .	126
<i>A. Erzin, N. Shabelnikova, L. Osotova, Y. Amirgaliyev</i>	
VNS-Based Heuristics for Communication Tree Optimal Synthesis Problem . . . . .	133
<i>A. Erzin, N. Mladenovic, R. Plotnikov</i>	
Classification of Scientific Documents Based on the Compression Methods . . . . .	140
<i>A. Guskov, B. Ryabko, A. Zubkov</i>	
Paypal E-Commerce and E-Payment - Problems and Solutions . . . . .	145
<i>M. Ilic, Z. Spalevic, P. Spalevic, N. Arsic, M. Veinovic</i>	
One Implementation of the Embedded Database Protection . . . . .	157
<i>S. Ilić, S. Obradović, N. Arsić, V. Petrović</i>	
Choosing the Model for Solving the Problem of Lexical Selection for Kazakh Language on Free/Open-Source Platform Apertium . . . . .	166
<i>A. Karibayeva, D. Amirova, M. Abakan</i>	
Construction of the Database and the Compilation Tools in CANRDB . . . . .	171
<i>V. Kurmangaliyeva, M. Takibayeva, M. Aikawa, N. Takibayev</i>	
Parallel Algorithm of RDF Data Compression and Decompression Based on MapReduce Hadoop Technology . . . . .	175
<i>M. Mansurova, E. Alimzhanov, E. Dadykina</i>	
3D Computer Technologies as a Tool for Contemporary Archaeology . . . . .	181
<i>M. Miłosz, J. Montusiewicz, R. Kayumov</i>	
Using GIM-Technologies for Monitoring of the Ionosphere Over Kazakhstan Region . . . . .	191
<i>S.N. Mukasheva, N.S. Toyshiev, B.K. Kurmanov, G. Sharipova, D.E. Karmenova</i>	
Development of the Kazakh Text-to-Speech Synthesis System on The Basis of Fujisaki Intonation Model . . . . .	196
<i>R. Mussabayev</i>	
Modification of the Encryption Algorithm, Developed on The Basis of Nonpositional Polynomial Notations . . . . .	205
<i>S. Nyssanbayeva, M. Magzom</i>	

Information-Analytical System "ECO Monitoring" .....	209
<i>S. Rakhmetullina, A. Penenko, Ye. Turganbayev, A. Bublikov</i>	
Design of Algorithms for Automated Access Control Based on Business Process Approach .....	218
<i>Z. Rodionova</i>	
User Interfaces for Working with Thesauri and Rubricators in Distributed Heterogeneous Information Systems on the Platform ZooSPACE .....	224
<i>S.A. Santeyeva, O.L. Zhizhimov</i>	
A Case Study of a Knowledge Management System .....	231
<i>A. Savic, E. Kalemi, M. Dëra</i>	
Performance Analysis of Wireless Transmission Channels in the Presence of Eta-Mu Fading and Kappa-Mu Co-Channel Interference .....	237
<i>D. Vučković, S. Panić, H. Milošević, D. Djošić</i>	
Surface Movements in Source Zones by Satellite Data .....	243
<i>Zh. Zhantayev, A. Kim, A. Ivanchukova, V. Junisbekova, A. Turgumbayev</i>	
Системы Распознавания образов в Задачах Автоматизации Распознавания Паспортных Данных .....	250
<i>Е.Н. Амиргалиев, Р. Юнусов</i>	
Полиномиальный Алгоритм для Задачи MSP3 .....	258
<i>М.З. Арсланов</i>	
Методы и Системы Автоматического Реферирования Текста .....	263
<i>А.М. Бакиева, Т.В. Батура, А.М. Федотов</i>	
Логический Подход К Организации Многокритериального Атрибутного Разграничения Доступа .....	275
<i>Р.Г. Бияшев, М.Н. Калимолдаев, О.А. Роз</i>	
Особенности и Требования к Качеству Программных Средств Космического Назначения .....	279
<i>Е. Исмаил</i>	
Особенности Разработки Программно-Технологического Обеспечения для Региональных Геоинформационных Веб-Систем .....	286
<i>А.А. Кадочников</i>	
Вычислительная Технология Обработки Данных Комплексного Мониторинга Природных Геообъектов .....	294
<i>М. Курако, К. Симонов</i>	
Математическое Моделирование Информационных Процессов в Веб-Пространстве ...	300
<i>Ю.И. Шокин, А.Ю. Веснин, А.А. Добрынин, О.А. Клименко, Е.В. Рычкова</i>	
Использование Разнородных Данных при Сегментации Спутниковых Изображений Высокого Разрешения .....	316
<i>Ю.Н. Сняевский, И.А. Пестунов, О.А. Дубровская, П.В. Мельников, С.А. Рылов, Д.В. Лазарев</i>	

Интеграция Географических Метаданных в Современные Системы Организации Цифровых Репозиторийв .....	324
<i>Д.М. Скачков, О.Л. Жижимов</i>	
Системный Подход к Конструированию Интерфейсов Приложений .....	332
<i>И.Н. Скопин</i>	
О Задаче Восстановления и Идентификации Множества Точек Разрыва Геометрических Объектов по Томографическим Данным .....	346
<i>А.П. Полякова, И.Е. Светов, Е.Ю. Деревцов, М.А. Султанов</i>	
Разработка Подсистемы Актуализации Базовых Пространственных Данных по Населенным Пунктам Красноярского Края .....	359
<i>А.В. Токарев</i>	

Session I. Technological Process  
Automation and Control

# Solving Hard SAT Instances in Volunteer Computing Project SAT@home

Igor Bychkov\*, Stepan Kochemazov\*, Maxim Manzyuk\*\*\*, Ilya Otpuschennikov\*, Mikhail Posypkin\*\*, Alexander Semenov\*, and Oleg Zaikin\*

\*Institute for System Dynamics and Control Theory SB RAS, Irkutsk, Russia

\*\*Institute for Information Transmission Problems RAS, Moscow, Russia

\*\*\*Internet portal BOINC.ru, Moscow, Russia

bychkov@icc.ru, veinamond@gmail.com, hoarfrost@rambler.ru, otilya@yandex.ru,  
mposypkin@gmail.com, biclop.rambler@yandex.ru, zaikin.icc@gmail.com

**Abstract.** In this paper we describe a volunteer computing project SAT@home aimed at solving hard combinatorial problems that can be reduced to Boolean satisfiability problem (SAT). Several hard SAT instances (logical cryptanalysis of some stream ciphers and search for sets of mutually orthogonal Latin squares) were successfully solved in SAT@home over recent three years. We also propose the CluBORun tool aimed at utilizing idle computational resources of clusters in volunteer computing projects based on BOINC. The key feature of CluBORun is that it uses only ordinary cluster's user rights. The CluBORun tool makes it possible to temporarily integrate the resources of several computing clusters into SAT@home.

**Keywords:** Volunteer computing, SAT, Latin squares, logical cryptanalysis, A5/1, Bivium.

Volunteer computing [1] is a type of distributed computing which uses computational resources of PCs of private persons called volunteers. Each volunteer computing project is designed to solve one or several hard problems. When PC is connected to the project, all the calculations are performed automatically and do not inconvenience user since only idle resources of PC are used. The first volunteer project was GIMPS<sup>1</sup> project launched in 1996. Nowadays the most popular platform for organizing volunteer computing project is BOINC [2] that was developed in Berkeley in 2002. Today there are about 70 active volunteer projects, the majority of them based on BOINC. Total performance of these projects is more than 10 petaflops. Among the most important results obtained via volunteer computing are the discoveries of the largest prime number in the GIMPS project and of several radiopulsars in the Einstein@home<sup>2</sup> project. Volunteer computing project consists of the following basic parts: server daemons, database, web site and client application. Daemons include work generator (generates tasks to be processed), validator (checks the correctness of the results received from volunteer's PCs) and assimilator (processes correct results). Client application should have versions for the widespread computing platforms. One of attractive features of volunteer computing is its low cost — to maintain a project one needs only a dedicated server working 24/7. Main difficulty here lies in the development of software and in database administration. In addition, it is crucial to provide the feedback to volunteers using the web site of the project and special forums. Another attractive feature of this type of computing is that volunteer project can solve some particular hard problem for months or even years with good average performance. In this paper, we describe a volunteer computing project SAT@home aimed at solving hard instances of Boolean satisfiability problem (SAT). Wide class of problems from modern computer science can be effectively reduced to SAT [3]. SAT problems are usually considered as the problems of search for solutions of Boolean

<sup>1</sup> <http://www.mersenne.org/>

<sup>2</sup> <http://einstein.phys.uwm.edu/>

equations in the form of  $CNF=1$ , where  $CNF$  is a conjunctive normal form. There are many works in which various combinatorial problems are reduced to SAT and solved in this form. For example, such problems can be found in areas of verification, cryptography, combinatorics and bioinformatics. All known SAT solving algorithms are exponential in the worst case since SAT itself is NP-hard. Nevertheless, modern SAT solvers successfully cope with many classes of tests based on the problems from the areas mentioned above. Improvement of the effectiveness of SAT solving algorithms, including the development of algorithms that are able to work in parallel and distributed computing environments, is a very important direction of research. The SAT@home project has been actively functioning since September 2011. On the first stage, the project was used to solve several cryptanalysis problems of the A5/1 keystream generator. On the second stage new pairs of orthogonal diagonal Latin squares of order 10 were found. On the third stage several weakened problems of cryptanalysis of the Bivium cipher were solved. Below we present a brief outline of our paper. In the second section we describe three problems considered and their reduction to SAT. In the third section various decomposition techniques for SAT are discussed. In the fourth section we present the SAT@home project and experimental results obtained. In the fifth section the CluBORun tool aimed at utilizing idle cluster resources in volunteer computing is proposed.

## 1 SAT solved in SAT@home project

### 1.1 Logical cryptanalysis problems

Usually if the cryptanalysis is considered as a SAT problem then it is called a logical cryptanalysis [4]. In this case to find a secret key it is sufficient to find a solution of corresponding satisfiable SAT instance. In the paper we will consider logical cryptanalysis of the A5/1 keystream generator and the Bivium stream cipher. Keystream generator A5/1 is used to encrypt traffic in GSM networks. The algorithm of this generator became publicly available in 1999 after reverse engineering performed by Marc Briceno. The description of the generator A5/1 (see figure 1) was taken from the paper [5]. According to [5] the generator A5/1 contains three linear feedback shift registers (LFSR, see [6]), given by the following connection polynomials: LFSR 1:  $X^{19} + X^{18} + X^{17} + X^{14} + 1$ ; LFSR 2:  $X^{22} + X^{21} + 1$ ; LFSR 3:  $X^{23} + X^{22} + X^{21} + X^8 + 1$ .

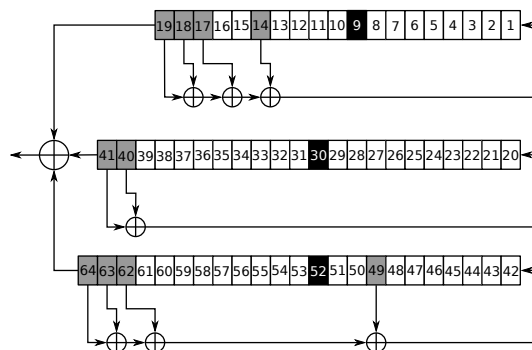


Fig. 1. Scheme of the generator A5/1.

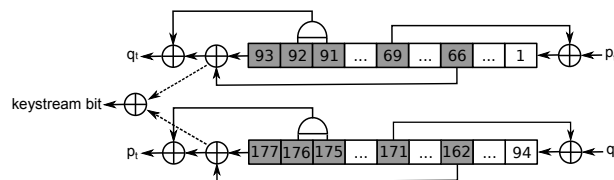
The secret key of the generator A5/1 is the initial contents of LFSRs 1–3 (64 bits). In each unit of time  $\tau \in \{1, 2, \dots\}$  ( $\tau = 0$  is reserved for the initial state) two or three registers are shifted. The register number  $r$ ,  $r \in \{1, 2, 3\}$  is shifted if  $\chi_r^\tau(b_1^r, b_2^r, b_3^r) = 1$  and is not shifted if



$\chi_r^\tau(b_1^\tau, b_2^\tau, b_3^\tau) = 0$ . By  $b_1^\tau, b_2^\tau, b_3^\tau$  we denote the values of clocking bits at the current unit of time. The clocking bits are 9-th, 30-th and 52-nd. The function  $\chi_r^\tau(\cdot)$  is defined as follows

$$\chi_r^\tau(b_1^\tau, b_2^\tau, b_3^\tau) = \begin{cases} 1, & b_r^\tau = \text{majority}(b_1^\tau, b_2^\tau, b_3^\tau) \\ 0, & b_r^\tau \neq \text{majority}(b_1^\tau, b_2^\tau, b_3^\tau) \end{cases}$$

where  $\text{majority}(A, B, C) = A \cdot B \vee A \cdot C \vee B \cdot C$ . In each unit of time the values in the leftmost cells of the registers are added mod 2, and the resulting bit is put to keystream. A lot of attacks on this generator are described, however it is still actively used. The most recent attacks use the technique of rainbow tables [7], however that approach can not guarantee the 100% success if the length of keystream fragment to analyze is small. In the latter cases it is sensible to use logical cryptanalysis. Usually it is believed that to uniquely identify the secret key it is sufficient to consider keystream fragment of length comparable to the total length of shift registers. We used TRANSALG system [8] to make SAT encodings for the problem of cryptanalysis of the A5/1 generator with first 114 bits of known keystream. Obtained template CNF consists of 35454 clauses over a set of 7816 Boolean variables. First 64 variables correspond to the unknown secret key, last 114 variables correspond to the known fragment of keystream. To create a particular SAT instance one needs to add 114 unit clauses (which correspond to the known fragment of keystream) to the template CNF. In more detail the process of creation of SAT encodings for cryptographic problems is described in [8]. The Bivium cipher [9] uses two shift registers of a special kind (see figure 2). The first register contains 93 cells and the second contains 84 cells. To initialize the cipher, a secret key of length 80 bit is put to the first register, and a fixed (known) initialization vector (IV) of length 80 bit is put to the second register. All remaining cells are filled with zeros. An initialization phase consists of 708 rounds during which keystream output is not released.



**Fig. 2.** Scheme of the Bivium cipher.

In accordance with [10,11] we considered cryptanalysis problems for Bivium in the following formulation. Based on the known fragment of keystream we search for the values of all registers cells at the end of an initialization phase. It means that we need to find 177 bits. Therefore, in our experiments we used CNF encodings where initialization phase was omitted. We followed [12,11] and set the keystream fragment length for Bivium cryptanalysis to 200 bits. Using the TRANSALG system we produced SAT encodings for corresponding cryptanalysis problems. Obtained template CNF consists of 12800 clauses over a set of 777 Boolean variables. One can create particular SAT instance for Bivium by adding 200 unit clauses (similar to A5/1). Programs in TA language (special domain specific language) used in TRANSALG and corresponding CNF templates for A5/1 and Bivium can be found in TRANSALG gitrepository<sup>3</sup>.

<sup>3</sup> <https://gitlab.com/transalg/transalg/tree/master>

## 1.2 The search for systems of mutually orthogonal Latin squares

One of the most promising areas of application of SAT approach is the search for combinatorial designs [13]. In particular, combinatorial problems related to Latin squares are very interesting. These problems attract the attention of mathematicians for several centuries. In recent years a number of new computational approaches to solving these problems have appeared. For example in [14] it was shown that there is no finite projective plane of order 10. It was done using special algorithms based on constructions and results from the theory of error correcting codes [15]. Corresponding experiment took several years, and on its final stage a quite powerful (at that moment) computing cluster was used. More recent example is a proof of hypothesis about minimal number of clues in Sudoku [16] where special algorithms were used to enumerate and check all possible Sudoku variants. To solve this problem a modern computing cluster had been working for almost a year. In [17] to search for some sets of Latin squares a special program system based on the algorithms of search for maximal clique in a graph was used. SAT solvers are systematically used to solve similar problems for more than 15 years [18]. One of the most ambitious combinatorial problems consists in answering the question if there exist three Mutually Orthogonal Latin Squares (MOLS) of order 10. Author of [13] notes that he spent more than 10 years on solving this problem using a distributed SAT solver working on 10-40 PCs. However, it was not possible for him to answer the original question. Interesting feature of problems of search for sets of MOLS is that they can be reduced to SAT in many ways because the set of MOLS as a combinatorial design is equivalent to several other designs. Consequently, one can consider the problem of search for such set as a problem of search for any of these equivalent objects. Further we consider one approach to the construction of Boolean encoding for the problem of search for a pair of MOLS of order  $n$ . It was described in many works about the application of SAT approach to the research of Latin squares and Sudoku (see, for example, [19]). In this encoding we consider two matrices  $A = ||a_{ij}||$  and  $B = ||b_{ij}||$ ,  $i, j \in \{1, \dots, n\}$ . We encode each matrix by its own set of  $n^3$  Boolean variables in such a way that its arbitrary cell is associated with  $n$  variables. Further we use the notation  $x(i, j, k)$  and  $y(i, j, k)$  to denote variables which encode cells of matrices  $A$  and  $B$ , respectively. Variable  $x(i, j, k)$  takes the value of true if and only if a cell in the  $i$ -th row and the  $j$ -th column contains  $k$ . Thus, for example, if  $x(1, 5, 3) = 1$  and  $n = 10$ , then  $a_{15} = 3$ . To make sure that matrices  $A$  and  $B$  represent Latin squares we should write down corresponding constraints over their variable sets. Further we show these constraints on the example of matrix  $A$ . It is possible to write these constraints directly in the form of conjunctions of clauses. In every matrix cell there is exactly one number from 1 to  $n$ :

$$\bigwedge_{i=1}^n \bigwedge_{j=1}^n \bigvee_{k=1}^n x(i, j, k) \\ \bigwedge_{i=1}^n \bigwedge_{j=1}^n \bigwedge_{k=1}^{n-1} \bigwedge_{r=k+1}^n (\neg x(i, j, k) \vee \neg x(i, j, r)).$$

Every number from 1 to  $n$  appears in every row exactly once:

$$\bigwedge_{j=1}^n \bigwedge_{k=1}^n \bigvee_{i=1}^n x(i, j, k) \\ \bigwedge_{j=1}^n \bigwedge_{k=1}^n \bigwedge_{i=1}^{n-1} \bigwedge_{r=i+1}^n (\neg x(i, j, k) \vee \neg x(r, j, k)).$$

Every number from 1 to  $n$  appears in every column exactly once:

$$\bigwedge_{i=1}^n \bigwedge_{k=1}^n \bigvee_{j=1}^n x(i, j, k) \\ \bigwedge_{i=1}^n \bigwedge_{k=1}^n \bigwedge_{j=1}^{n-1} \bigwedge_{r=j+1}^n (\neg x(i, j, k) \vee \neg x(i, r, k)).$$

Constraints for variables that correspond to matrix  $B$  are written similarly. After this we need to complete the encoding with a constraint for the orthogonality condition for our two

squares. For example we can write this constraint in the following form:

$$\bigwedge_{i=1}^n \bigwedge_{j=1}^n \bigwedge_{k=1}^n \bigwedge_{p=1}^n \bigwedge_{q=1}^n \bigwedge_{r=1}^n (\neg x(i, j, k) \vee \neg y(i, j, k) \vee \neg x(p, q, r) \vee \neg y(p, q, r)).$$

It is not difficult to see that the number of clauses in a CNF constructed according to this approach grows as  $O(n^6)$  with the increase of  $n$ . We used this encoding to search for pairs of MOLS of order 10 with additional "diagonality" condition. Corresponding CNF consists of 434440 clauses over a set of 2000 Boolean variables. In such pairs in each square both main and secondary diagonals must contain all numbers from 1 to  $n$  where  $n$  is the order of these squares. It is known that pairs of orthogonal diagonal Latin squares of order 10 exist, however all the pairs that we could find are cited in [20]. That is why, in our opinion, it would be interesting to search for new instances of such pairs. It is easy to see that to obtain a Boolean encoding for the problem of search for a pair of orthogonal diagonal Latin squares one should only add to the encoding described above clauses corresponding to the diagonality condition. It does not influence the asymptotics of the total number of clauses in a corresponding CNF.

## 2 SAT partitioning

In [21] various approaches to partitioning SAT were studied. Below we will use the terminology from [21]. Consider a satisfiability problem for an arbitrary CNF  $C$ . Partitioning of  $C$  is a set of formulae

$$C \wedge G_i, i \in \{1, \dots, S\}$$

such that for any  $i, j : i \neq j$  formula  $C \wedge G_i \wedge G_j$  is unsatisfiable and

$$C \equiv C \wedge G_1 \vee \dots \vee C \wedge G_S.$$

When one has a partitioning of an original SAT instance, satisfiability problems for  $C \wedge G_j, j \in \{1, \dots, S\}$  can be solved independently in parallel. There exist various partitioning techniques. For example one can construct  $\{G_j\}_{j=1}^S$  using a scattering procedure [21], a guiding path solver [18] or a lookahead solver [22]. Unfortunately, for these partitioning methods it is hard in general case to estimate the time required to solve an original problem. From the other hand in a number of papers about logical cryptanalysis of several keystream ciphers there was used a partitioning method that makes it possible to construct such estimations in quite a natural way. In particular, in [12,11] for this purpose the information about the time to solve small number of subproblems randomly chosen from the partitioning of an original problem was used. In the paper [23] strict formal description of this idea within the borders of the Monte Carlo method in its classical form [24] was given. Consider a satisfiability problem for an arbitrary CNF  $C$  over a set of Boolean variables  $X = \{x_1, \dots, x_n\}$ . We call an arbitrary set  $\tilde{X} = \{x_{i_1}, \dots, x_{i_d}\}, \tilde{X} \subseteq X$  a decomposition set. Consider a partitioning of  $C$  that consists of a set of  $2^d$  formulae

$$C \wedge G_j, j \in \{1, \dots, 2^d\}$$

where  $G_j, j \in \{1, \dots, 2^d\}$  are all possible minterms over  $\tilde{X}$ . Note that an arbitrary formula  $G_j$  takes a value of true on a single truth assignment  $(\alpha_1^j, \dots, \alpha_d^j) \in \{0, 1\}^d$ . Therefore, an arbitrary formula  $C \wedge G_j$  is satisfiable if and only if  $C \left[ \tilde{X} / (\alpha_1^j, \dots, \alpha_d^j) \right]$  is satisfiable. Here

$C \left[ \tilde{X} / \left( \alpha_1^j, \dots, \alpha_d^j \right) \right]$  is produced by setting values of variables  $x_{i_k}$  to corresponding  $\alpha_k^j$ ,  $k \in \{1, \dots, d\} : x_{i_1} = \alpha_1^j, \dots, x_{i_d} = \alpha_d^j$ . A set of CNFs

$$\Delta(C, \tilde{X}) = \left\{ C \left[ \tilde{X} / \left( \alpha_1^j, \dots, \alpha_d^j \right) \right] \right\}_{(\alpha_1^j, \dots, \alpha_d^j) \in \{0,1\}^d}$$

is called a decomposition family produced by  $\tilde{X}$ . Total number of CNFs in  $\Delta(C, \tilde{X})$  is  $2^d$ . It is possible to estimate total time required for processing this family based on the time of solving of SAT problems for  $k$ ,  $k \ll 2^d$  randomly chosen CNFs from the family. This estimation can be used to plan a computational experiment. In [23] a Monte Carlo algorithm based on tabu search heuristics to search for decomposition sets with good time estimations was suggested. This algorithm was implemented as a parallel MPI program PDSAT. We used PDSAT running on a computing cluster to obtain decomposition sets for A5/1 and Bivium logical cryptanalysis problems. As a result we found the set of 32 variables for A5/1 (see right-hand side of figure 3) and the set of 47 variables for Bivium (see figure 4). On the left-hand side of figure 3 the decomposition set that was found “manually” by analyzing features of the A5/1 generator [25] is located. On these figures cells corresponding to variables from decompositions sets are marked with gray color.

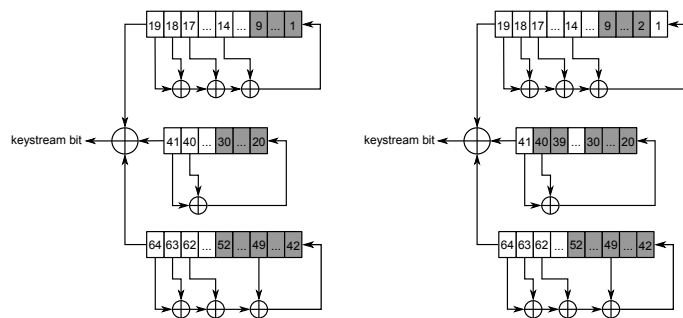


Fig. 3. Decomposition sets for logical cryptanalysis of A5/1.

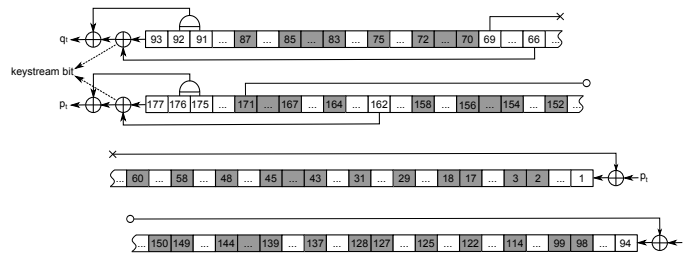


Fig. 4. Decomposition set for logical cryptanalysis of Bivium.

In recent years several desktop grids for solving SAT appeared. For example, in [26] a desktop grid for solving SAT which used conflict clauses exchange via a peer-to-peer protocol was described. Apparently, [27] became the first paper about the use of a desktop grid based on the BOINC platform for solving SAT. Unfortunately, it did not evolve into a publicly available

volunteer computing project. SAT@home<sup>4</sup> [28] is a volunteer computing BOINC-based project aimed at solving hard combinatorial problems that can be effectively reduced to SAT. It was launched on September 29, 2011 by ISDCT SB RAS and IITP RAS. On February 7, 2012 SAT@home was added to the official list of BOINC projects<sup>5</sup> with alpha status. Recently its status was improved to beta. SAT@home server uses a number of standard BOINC daemons responsible for sending and processing tasks (transitioner, feeder, scheduler, etc.). Such daemons as work generator, validator and assimilator were implemented taking into account the specificity of the project. The work generator decomposes the original SAT problem to subproblems according to the decomposition set approach (see section 3). It creates 2 copies of each task in accordance with the concept of redundant calculations used in BOINC. The validator checks the correctness of the results, and the assimilator processes correct results. Client application is based on the SAT solver MINISAT [29]. Characteristics of the SAT@home project as of 5 of June 2015 are (according to BOINCstats<sup>6</sup>):

- 2363 active PCs (active PC in volunteer computing is a PC that sent at least one result in last 30 days) about 80% of them use Microsoft Windows OS;
- 1184 active users (active user is a user that has at least one active PC);
- versions of the client application for CPU: Windows x86, Windows x86-64, Linux x86, Linux x86-64;
- average real performance: 4 teraflops, maximal performance: 10 teraflops (it was achieved during the competition held by BOINCstats).

The dynamics of the real performance of SAT@home can be seen at the SAT@home performance page<sup>7</sup>. An experiment consisting in solving 10 logical cryptanalysis problems of the generator A5/1 was held in SAT@home from December 2011 to May 2012. It should be noted that we considered only instances that could not be solved using the known rainbow tables. Specifically, from randomly generated 1000 cryptanalysis instances for A5/1 (in the formulation described in section 2) 125 could not be solved by rainbow tables. We randomly selected 10 instances and successfully solved them in SAT@home. In that experiment we used the “manually” found decomposition set of 31 variables (see section 3). The client application was based on a modified version of MINISAT-C 1.14.1 (see [25]). On average in order to solve one instance of logical cryptanalysis of A5/1 SAT@home processed about 1 billion of SAT instances (made by decomposing an original SAT instance). During the experiment for three instances there were found two solutions (1 original and 1 so-called collision, see section “Found solutions”<sup>8</sup> on SAT@home site), for other 7 instances there was found exactly one solution. In May 2014 we launched in SAT@home an additional experiment based on the set of 32 variables that was found automatically by PDSAT (see section 3). Until 25th of July first 5 instances (from the same set of 10 instances) have been solved. Current results show that the time required to solve problems considered using these two decomposition sets is quite similar. It should be noted that the estimation for the A5/1 cryptanalysis obtained with the use of PDSAT is close to the average real time spent by SAT@home to solve corresponding SAT instances. It means that with the help of PDSAT one can obtain good decomposition set automatically — i.e. without analyzing a particular original problem. With respect to the estimation obtained by PDSAT the solving of one instance of cryptanalysis of the Bivium cipher in the SAT@home project with its current

<sup>4</sup> <http://sat.isa.ru/pdsat/>

<sup>5</sup> <http://boinc.berkeley.edu/projects.php>

<sup>6</sup> <http://boincstats.com/>

<sup>7</sup> <http://sat.isa.ru/pdsat/performance.php>

<sup>8</sup> <http://sat.isa.ru/pdsat/solutions.php>

performance would take about 4 years (using decomposition set of 47 variables [23]). That is why we decided to solve weakened cryptanalysis problems for this cipher. Below we use the notation *BiviumK* to denote a weakened cryptanalysis problem for Bivium with known values of *K* variables (in corresponding SAT instance) encoding last *K* cells of the second shift register. In particular we considered the *Bivium10* problem. We used PDSAT to find a decomposition set with good time estimation for *Bivium10*. As a result we obtained the decomposition set of 40 variables. From April 2014 to May 2014 with the help of this decomposition set 3 weakened *Bivium10* problems were solved in SAT@home. The client application was based on a modified version of MINISAT 2.2. Time estimation by PDSAT shows good consistency with real solving time of these cryptanalysis instances. As far as we know there are no over publicly available results for such problems. From September 2012 to May 2013 we carried out in the SAT@home project an experiment aimed at the search for orthogonal pairs of diagonal Latin squares of order 10. Client application was based on the MINISAT 2.2 solver. It was slightly modified to reduce the amount of RAM used. The first line of the first Latin square in a pair was fixed to "0 1 2 3 4 5 6 7 8 9". It was done because any pair of orthogonal diagonal Latin squares can be reduced to a pair where one of the squares has such a first line by permutations that do not break diagonality and orthogonality constraints. The decomposition was performed in a following way: we varied values of first 8 cells of second and third lines of the first Latin square. There are about 230 billions possible variants of assignments of corresponding variables that do not break diagonality constraint. We decided to check in SAT@home first 20 millions variants from 230 billions (i.e. about 0.0087% of an original search space). As a result, each variant was formed by assigning values to first 8 cells of second and third lines of the first Latin square (note, that the first line was fixed due to the reasons explained above). It means that values of remaining 74 cells of the first square and of all 100 cells of the second square had to be determined by the SAT solver. Each SAT instance had to be solved within the limit of 2600 restarts of the MINISAT 2.2 solver that is equivalent to approximately 5 minutes of work of one core of modern CPU. Upon reaching the limit the computations were interrupted. The experiment took about 9 months. The computations for the vast majority of subproblems were interrupted due to reaching the time limit, however for 17 subproblems it was possible to find solutions. As a result we found 17 new pairs of orthogonal diagonal Latin squares of order 10 (we compared them with three pairs of orthogonal Latin squares from [20]). All pairs found are available on the project site in the "Found solutions" section.

### 3 CluBORun tool

There are specific volunteer computing projects based only on desktop PCs that have performance greater than one petaflops. Despite this fact there are reasons why resources of computing clusters can be useful in volunteer computing. First, a computing cluster is very reliable, so results obtained on it can be taken as a reference when checking the results from volunteers. Second, a computing cluster can significantly help to increase performance of a new volunteer project with low amount of participants. Some additional argumentation (with recommendations regarding using volunteer computing on clusters) was suggested in the paper [30]. There are several tools that can be used to combine resources of clusters and BOINC-based desktop grids (for example, 3G Bridge [31]). 3G Bridge is aimed at combining resources of several grid system (for example, of a service grid based on cluster and desktop grid based on BOINC). Main feature of such tools is that they require administrators rights of cluster. We have implemented a CluBORun<sup>9</sup> (Cluster

<sup>9</sup> <https://github.com/Nauchnik/CluBORun>

for BOINC Run [32]) tool aimed at utilizing idle resources of computing clusters in volunteer computing projects based on BOINC. CluBORun is a set of several shell scripts and a C++ MPI program. The key features of CluBORun are: it utilizes only idle resources of computing clusters (just as BOINC-manager does it for PCs); it uses only ordinary clusters user rights. BOINC calculations are launched as MPI tasks which are processed by cluster scheduling system (as all other tasks of another cluster users). After being launched on a cluster node the MPI program starts BOINC manager. BOINC manager connects to a project server and performs computations using standard client applications of a project. From the BOINC manager point of view, cluster node is just another PC under Linux OS control. When tasks from another user appear in a cluster queue, CluBORun stops BOINC tasks in queue if new tasks can be launched on freed resources. It should be noted that CluBORun doesn't violate any rules of cluster use since all ordinary user's restrictions apply (limit on total CPU hours or on a number of simultaneously active tasks in queue). The CluBORun tool has been successfully working on the computing cluster MVS-100k (Joint supercomputer center of RAS<sup>10</sup>) from December 2013 until the present moment. As a result of adding resources of this cluster to SAT@home the performance of the project increased by 40 % (i.e. by about 1.5 teraflops) in some periods of time. One of the main difficulties in development of CluBORun is a necessity to make separate version for every new job scheduling system (each system has its own commands to work with cluster queue). At the moment CluBORun can work with the Cleo, SUPPZ and SLURM job scheduler systems. Now we are working on CluBORun version for the PBS scheduling system.

## 4 Conclusions

Since the solving of hard SAT instances can take months or years we decided to create a volunteer computing project SAT@home. Obtained results show that it can be successfully used for solving problems from various areas. In future we plan to solve in SAT@home cryptanalysis problems for other ciphers. Also we will try to find new combinatorial designs, including sets of MOLS. We also plan to continue the development of the CluBORun tool. We hope that this tool will be useful for utilizing cluster resources in other volunteer projects too.

## 5 Acknowledgments

Authors thank Nikolay Khrapov and Vadim Bulavintsev for their help in administering the SAT@home project. We thank Martin Maurer for his valuable feedback that made it possible to improve the effectiveness of the project. We are very grateful to all the SAT@home volunteers. Without your help this work would be impossible. This work was partly supported by Russian Foundation for Basic Research (grants 13-07-00768-a, 13-07-12080-ofi-m, 14-07-00166-a, 14-07-00403-a, 15-07-07891-a and 14-07-31172-mol-a) and by the President of Russian Federation grants (SP-3667.2013.5, SP-1184.2015.5, NSH-5007.2014.9).

## References

1. Nouman Durrani, M., Shamsi, J.A.: Volunteer computing: Requirements, challenges, and solutions. *J. Netw. Comput. Appl.* 39, 369–380 (2014)
2. Anderson, D.P.: BOINC: A System for Public-Resource Computing and Storage. In: *Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing*. pp. 4–10. GRID '04, IEEE Computer Society, Washington, DC, USA (2004)

<sup>10</sup> <http://www.jscc.ru/>

3. Prestwich, S.: CNF Encodings. In: Biere, A., Heule, M., van Maaren, H., Walsh, T. (eds.): Handbook of Satisfiability, Frontiers in Artificial Intelligence and Applications, vol. 185. IOS Press (2009). ch. 2. pp. 5–98 (2009)
4. Massacci, F., Marraro, L.: Logical Cryptanalysis as a SAT Problem. *J. Autom. Reasoning*. 24(1/2), 165–203 (2000)
5. Biryukov, A., Shamir, A., Wagner, D.: Real Time Cryptanalysis of A5/1 on a PC. In: Proceedings of the 7th International Workshop on Fast Software Encryption. pp. 1–18. FSE '00, Springer-Verlag, London, UK, UK (2001)
6. Menezes, A.J., Vanstone, S.A., Van Oorschot, P.C.: Handbook of Applied Cryptography. CRC Press, Inc., Boca Raton, FL, USA, 1st edn. (1996)
7. Güneysu, T., Kasper, T., Novotný, M., Paar, C., Rupp, A.: Cryptanalysis with COPACOBANA. *IEEE Trans. Comput.* 57(11), 1498–1513 (2008)
8. Otpuschennikov I., Semenov, A., Kochemazov S.: Transalg: a Tool for Translating Procedural Descriptions of Discrete Functions to SAT. In: WSCE'2015: Proceedings of the 5th International Workshop on Computer Science and Engineering, pp. 289–294 (2015)
9. De Cannière, C.: Trivium: A Stream Cipher Construction Inspired by Block Cipher Design Principles. In: Katsikas, S.K., Lopez, J., Backes, M., Gritzalis, S., Preneel, B. (eds.) ISC. LNCS, vol. 4176, pp. 171–186. Springer (2006)
10. Maximov, A., Biryukov, A.: Two Trivial Attacks on Trivium. In: Adams, C.M., Miri, A., Wiener, M.J. (eds.) Selected Areas in Cryptography. LNCS, vol. 4876, pp. 36–55. Springer (2007)
11. Soos, M.: Grain of Salt — an Automated Way to Test Stream Ciphers through SAT Solvers. In: Tools'10: Proceedings of the Workshop on Tools for Cryptanalysis. pp. 131–144 (2010)
12. Eibach T., Pilz E., Völkel G.: Attacking Bivium Using SAT Solvers. In: Büning, H.K., Zhao, X. (eds.) SAT. LNCS, vol. 4996, pp. 63–76. Springer (2008)
13. Zhang, H.: Combinatorial Designs by SAT Solvers. In: Biere, A., Heule, M., Van Maaren, H., Walsh, T. (eds.) Handbook of Satisfiability. Frontiers in Artificial Intelligence and Applications, vol. 185. ch. 17. pp. 533–568 (2009)
14. Lam C.W.H., Thiel, L., Swiercz S.: The nonexistence of finite projective planes of order 10. *Canad. J. Math.* 41, 1117–1123 (1989)
15. MacWilliams, F.J., Sloane, N.J.A.: The Theory of Error-Correcting Codes. North Holland Publishing Co. (1988)
16. McGuire, G., Tugemann, B., Civario G.: There Is No 16-Clue Sudoku: Solving the Sudoku Minimum Number of Clues Problem via Hitting Set Enumeration. *Experimental Mathematics*. 23(2), 190–217 (2014)
17. McKay, B.D., Meynert, A., Myrvold, W.: Small Latin squares, quasigroups and loops. *J. Combin. Designs*. 15, 98–119 (2007)
18. Zhang, H., Paola M.B., Hsiang J. PSATO: a Distributed Propositional Prover and its Application to Quasigroup Problems. *J. Symb. Comput.* 21(4), 543–560 (1996)
19. Lynce, I., Ouaknine, J.: Sudoku as a SAT Problem. In: ISAIM'2006: International Symposium on Artificial Intelligence and Mathematics. USA (2006)
20. Brown J., et al.: Completion of the Spectrum of Orthogonal Diagonal Latin Squares. *Lect. Notes Pure Appl. Math.* 139, 43–49 (1993)
21. Hyvärinen, A.E.J.: Grid Based Propositional Satisfiability Solving. Ph.D. thesis, Finland, Aalto University (2011)
22. Heule, M., Kullmann, O., Wieringa, S., Biere, A.: Cube and Conquer: Guiding CDCL SAT Solvers by Lookaheads. In: Eder, K., Lourenço, J., Shehory, O. (eds.) Haifa Verification Conference. LNCS, vol. 7261, pp. 50–65. Springer (2011)
23. Semenov, A., Zaikin, O.: Using Monte Carlo Method for Searching Partitionings of Hard Variants of Boolean Satisfiability Problem. In: Malyshkin, V. (ed.) PaCT. LNCS, vol. 9251, pp. 222–230. Springer (2015)
24. Metropolis, N., Ulam, S.: The Monte Carlo Method. *J. Amer. statistical assoc.* 44(247), 335–341 (1949)
25. Semenov, A., Zaikin, O., Bespalov, D., Posypkin, M.: Parallel Logical Cryptanalysis of the Generator A5/1 in BNB-Grid System. In: Malyshkin, V. (ed.) PaCT. LNCS, vol. 6873, pp. 473–483. Springer (2011)
26. Schulz, S., Blochinger, W.: Parallel SAT Solving on Peer-to-Peer Desktop Grids. *J. Grid Comput.* 8(3), 443–471 (2010)
27. Black, M., Bard, G.: SAT Over BOINC: An Application-Independent Volunteer Grid Project. In: Jha, S., gentschen Felde, N., Buyya, R., Fedak, G. (eds.) GRID. pp. 226–227. IEEE (2011)
28. Posypkin, M., Semenov, A., Zaikin, O.: Using BOINC desktop grid to solve large scale SAT problems. *Computer Science Journal*. 13(1), 25–34 (2012)
29. Eén, N., Sörensson, N.: An Extensible SAT-solver. In: Giunchiglia, E., Tacchella, A. (eds.) SAT. LNCS, vol. 2919, pp. 502–518. Springer (2003)



30. Vyas, D., Subhlok, J.: Volunteer computing on clusters. In: Proceedings of the 12th International Conference on Job Scheduling Strategies for Parallel Processing. pp. 161-175. JSSPP'06, Springer-Verlag, Berlin, Heidelberg (2007)
31. Farkas, Z., Kacsuk, P., Balaton, Z., Gombás, G.: Interoperability of BOINC and EGEE. *Future Gener. Comput. Syst.* 26(8), 1092–1103 (2010)
32. Afanasiev, A.P., Bychkov, I.V., Manzyuk, M.O., Posypkin, M.A., Semenov, A.A., Zaikin, O.S.: Technology for Integrating Idle Computing Cluster Resources into Volunteer Computing Projects. In: WSCE'2015: Proceedings of the 5th International Workshop on Computer Science and Engineering, pp. 109–114 (2015).

# A Scalable Parallel Algorithm and Software for 3D Seismic Simulation on Clusters with Intel Xeon Phi Coprocessors

Dmitry Karavaev, Boris Glinsky, and Valery Kovalevsky

Institute of Computational Mathematics and Mathematical Geophysics SB RAS,  
Prospect Akademika Lavrentieva 6, 630090 Novosibirsk, Russia  
kda@opg.sbcc.ru, gbm@opg.sbcc.ru, kovalevsky@sbcc.ru  
www.sbcc.ru

**Abstract.** In this paper, we present the results of the research in to the development of a scalable parallel algorithm for solving large problems of the forward modeling in geophysics. The problem to be solved is the system of equations of elastic theory representing the wave propagation in elastic 3D media. We have developed a scalable parallel algorithm and a program for the 3D seismic wave simulation on modern multi-core clusters with a hybrid architecture based on Intel Xeon Phi coprocessor. We present this parallel algorithm for solving the above-mentioned problem and the results of the parallel algorithm behavior on the Xeon Phi based cluster for different tests of the parallel program code. In addition, we compare implementation of the proposed parallel algorithm on different computing devices.

**Keywords:** parallel algorithm, seismic simulation, scalability, Xeon Phi, hybrid cluster.

## Introduction

One of the methods for solving inverse geophysical problems is solving the forward problem for a various number of models, which are different in geometry structure and elastic parameters values [1]. Thus, carrying out the simulation, varying elastic parameters and establishing the correspondence with natural geophysical data, one can find a more appropriate geometrical structure and elastic parameters values of a geophysical object under investigation. In addition, one of the useful and well-known methods for solving a 3D seismic simulation problem is a difference method based on 3D grids [2]. The most useful difference methods can be a second or a fourth order of approximation [4,4,6] and have application in modeling elastic or viscoelastic media [7]. The more difficult are geophysical models the more difficult is to calculate them. Because of using the difference method one should handle with large 3D arrays and a great volume of data. Only a 3D grid model of 3D isotropic geophysical media is described with the three parameters: density and two velocities of elastic waves. In this paper, we deal only with isotropic 3D elastic media. The problem of 3D elastic wave propagation is presented in terms of velocity and stress. Therefore, we need to calculate nine 3D values that are 3D arrays. When using the 3D explicit difference scheme with the iterative technique one should use the values at two time steps of iteration. To carry out calculation for 3D models with a detailed representation is a difficult task because of dealing with large 3D arrays that are to be placed in the operation memory of a computer. In such a case, researchers use powerful multi-core calculation systems. Such systems can have different architectures. Modern cluster systems that are in the first places in TOP 500 rating have a hybrid architecture. This means that these systems have special computing devices that are presented by Nvidia GPU or Intel Xeon Phi cluster. Some examples of such clusters are NKS-30T+GPU cluster of the Siberian Supercomputer Center and MVS-10P cluster of the Joint Supercomputer Center. With the use of such computing systems, one can solve large 3D models in parallel. So each of the computing device solved its part of 3D data, all together covering the whole 3D area under study. Therefore, we need to develop a scalable

parallel algorithm and a program code for using such a device in calculations. It is not only a programmable problem but also a researcher's problem. When developing a scalable parallel algorithm we should make special tests of a program code on one computing device to tune it for the 3D elastic wave simulation using difference method to watch the behavior of a program on different number of computing devices and models with different volume of data. There are many program codes realizing difference methods for the seismic wave propagation modeling on clusters with GPUs [8,11]. Using the Intel Xeon Phi coprocessors in simulation is a new and modern approach for parallel computation. Such systems can allow researchers using OpenMP parallel tools make developing programs for large-scale simulation easier. Our main purpose is to describe how the difference method for numerical simulation of seismic wave propagation is implemented on supercomputers with Intel Xeon Phi coprocessors. Section 2 gives a brief description of the problem statement and numerical method. In Section 3 we describe parallel implementation in detail. Section 4 presents computational experiments for different test of a parallel code. We discuss the implementation of the simulation code for the single coprocessor case to tune the script parameters. Second we discuss the multi-device case in order to study parallel algorithm behavior for large-scale problems. Section 5 concludes the main results.

## Problem Statement

We solve the forward geophysical problem of the 3D elastic wave propagation and deal with isotropic and elastic material [9]. Before carrying out the calculation we have to construct the 3D grid model of a geophysical medium under study. Such an object is described by parameters of density, shear wave velocity and longitudinal wave velocity. Thus, we need three material parameters (LamBée coefficients and density) in a difference scheme. All the geometries of elastic media have plane free surfaces. Such a medium can have a difficult geometrical structure and different values of elastic parameters at each point of 3D grid. A problem of the 3D elastic wave propagation is described in terms of components of velocities of displacements  $\mathbf{u} = (U, V, W)^T$  and components of a stress tensor  $\sigma = (\sigma_{xx}, \sigma_{yy}, \sigma_{zz}, \sigma_{xy}, \sigma_{xz}, \sigma_{yz})^T$ . The problem is to be solved with appropriate initial conditions and boundary values. We apply a free-surface condition at the top boundary. We use the Cartesian coordinate system. To numerically solve the simulation problem, we use the difference method [4]. This method is based on using staggered grids. This means that different values are placed at different points of a grid cell. The difference scheme is of second order of approximation with respect to time and space. The government equations of difference scheme will be in the form of (1).

$$\rho \frac{\partial \mathbf{u}}{\partial t} = [A]\sigma + \mathbf{F}(t, x, y, z), \frac{\partial \sigma}{\partial t} = [B]\mathbf{u}; \quad (1)$$

$$A = \begin{bmatrix} \frac{\partial}{\partial x} & 0 & 0 & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} & 0 \\ 0 & \frac{\partial}{\partial y} & 0 & \frac{\partial}{\partial x} & 0 & \frac{\partial}{\partial z} \\ 0 & 0 & \frac{\partial}{\partial z} & 0 & \frac{\partial}{\partial x} & \frac{\partial}{\partial y} \end{bmatrix}, B = \begin{bmatrix} (\lambda + 2\mu) \frac{\partial}{\partial x} & \lambda \frac{\partial}{\partial y} & \lambda \frac{\partial}{\partial z} \\ \lambda \frac{\partial}{\partial x} & (\lambda + 2\mu) \frac{\partial}{\partial y} & \lambda \frac{\partial}{\partial z} \\ \lambda \frac{\partial}{\partial x} & \lambda \frac{\partial}{\partial y} & (\lambda + 2\mu) \frac{\partial}{\partial z} \\ \mu \frac{\partial}{\partial y} & \mu \frac{\partial}{\partial x} & 0 \\ \mu \frac{\partial}{\partial z} & 0 & \mu \frac{\partial}{\partial x} \\ 0 & \mu \frac{\partial}{\partial z} & \mu \frac{\partial}{\partial y} \end{bmatrix}$$

Our modification for the elastic wave propagation simulation is that we use the calculated coefficients in the developed program. This means that coefficients from the problem statement differ from those we use in the difference scheme, including all the summations and multiplications, and are placed in special 3D arrays.

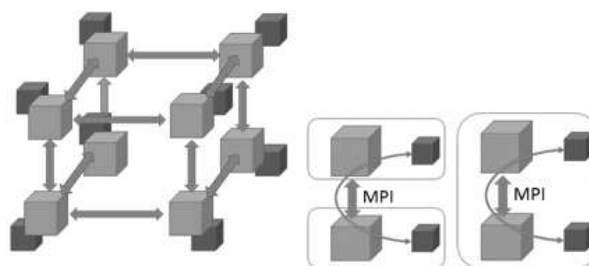
## Parallel Implementation

We consider hybrid parallel implementation, using both CPUs and Intel Xeon Phi coprocessors for computation. We have developed a scalable parallel algorithm based on the difference method with 3D grids and the program code for a cluster with a hybrid architecture and Intel Xeon Phi coprocessors. The developed parallel scheme includes use of technologies for parallel computing such as Message Passing Interface (MPI) and a software for Intel Xeon Phi coprocessor programming. We use MPI and OpenMP, respectively, for parallel computations. Our parallel realization has a data distributed character. We divide a large 3D model into smaller 3D subdomains, Fig. 1. Each of them is calculated independently and in the parallel manner. For the computations, we use the multi-core computing system with Intel Xeon Phi coprocessor. Several CPUs and several Xeon Phi coprocessors (devices) are placed at the computing nodes of such a system. Each Xeon Phi coprocessor can be treated as SMP (Symmetric Multiprocessing) machine. Such a device has 8GB DDR5 memory, 60 computing cores based on x86 architecture, 4HW threads/core, IP addressable, have Linux OS. We can employ up to 240 parallel threads with such a device. The cluster consists of 207 computing nodes with 2 Xeon E5-2650 processors and 2 Intel Xeon Phi 7110X coprocessors, [www.jscc.ru](http://www.jscc.ru). For running the program code on it, one should recompile a program code. We take the 3D model data on the CPUs and initialize necessary 3D arrays on the computing devices. After that, we copy the model data into the computing devices. Then we can carrying out computations using a parallel algorithm. All the calculations for 3D subdomains are conducted only on devices. The CPU is used only for device manipulation and for making exchanges between data placed in the devices. The Xeon Phi coprocessors are used in the offload mode. Thus, direct communication between coprocessors is not available. The data must be sent from coprocessor to the host CPU in order that the data be exchanged with the other coprocessors. In our parallel realization, to make the next time step we should make data exchange among neighbor devices that can be either at one computing node or at different computing nodes of the cluster. We use non-blocking MPI Send/Receive procedures that take place for the data exchange among the computing devices, Fig. 1. Since we use the 3D domain decomposition, we first compute for the points on the sides of subdomains. Second, we start the data copying from Xeon Phi<sup>™</sup>s to CPU cores and run exchange procedures with use of special designed buffers and MPI functions. Then we do the computation for the remaining internal grid points of 3D subdomains and the communication procedures simultaneously. After that, we verify whether all the exchanges have been done and make a data copy from buffers placed at CPUs into buffers at Xeon Phi coprocessors and then into 3D arrays. After Then we proceed to the next time step. Therefore, we overlap the communication and computation by using the non-blocking MPI functions for data transfer. We do the data transfer between the CPUs at nodes concurrent with the computations at Xeon Phi coprocessors.

All the program code was written using C++ language.

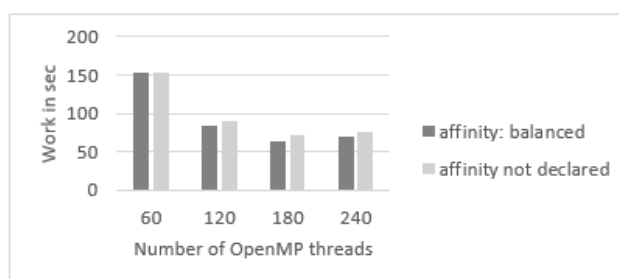
## Studying the Work of Parallel Algorithm on Xeon Phi Cluster

In this section, we present the results of the parallel algorithm behavior for a hybrid cluster architecture with Intel Xeon Phi coprocessor. We have carried out experiments on one computing device to choose appropriate options for large 3D models. We made a comparison of programs running on different computing devices for calculations that is using only CPUs or only Xeon Phi coprocessors. All the program codes were developed by the authors with the use of the proposed parallel algorithm and the designed parallel scheme. In addition, we present the results for different tests for the parallel program code. The first one is a scalability test that reveals



**Fig. 1.** A parallel computing scheme.

that the calculation time should not vary strongly if we do calculations taking into consideration the fact that the number of points in a 3D grid model grows proportional to the number of devices. This means that each computing device will do calculations for the same number of points. Another test is speed-up one. In this case, the number of points in a 3D model is fixed and we show the program behavior running on a different number of computing devices. All the results were carried out using the NKS-30T+GPU cluster of the Siberian Supercomputer Center (SSCC SB RAS), [www2.sccc.ru](http://www2.sccc.ru), and the MVS-10P cluster of the Joint Supercomputer Center of RAS, [www.jscc.ru](http://www.jscc.ru). On one cluster node and on one device, we have carried out experiments with different options of affinity and a different number of threads per core. The 3D model under study has parameters 308x308x308 and 11 iterations. The affinity option has been taken in two versions: B«not declared» or B«balanced». The results of such a research is presented at Fig. 2. The most appropriate is the B«balanced» option and using 60 cores with a 3 threads per core.



**Fig. 2.** A test with affinity option and the number of parallel threads on one device.

The performance of developed multi-coprocessor code is shown in Fig. 3 and Fig. 4. The results of scalability tests presented at figure show the well-done program behavior. When we scale a 3D model as great as 2-fold along each spatial coordinate and the number of devices as great as 8-fold, the program shows a good behavior. In these tests, we use a 3D subdomain size with 308x308x308 grid points and 11 iterations for one device. Maximum eight subdomains were used. In this case we have used 2 x 2 x 2 grid of computing devices. From Fig. 3 we can see that the program was effectively parallelized and the ratio of CPU/Xeon Phi is about x5.7.

The results of the speed-up tests presented in the Fig. 4 show the program behavior with 308x308x308 grid points and 11 iterations for all devices. Figure 4 reveals that the ration of

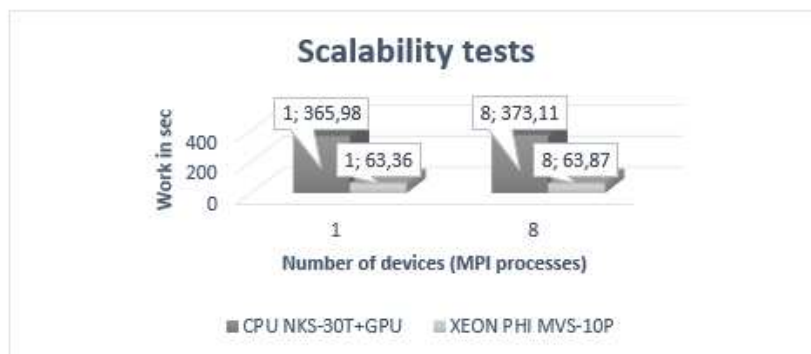


Fig. 3. A scalability test.

CPU/Xeon Phi is about  $\times 5.7$  on one device and  $\times 3.6$  other eight devices. The ratio of 1 to 8 devices for Xeon Phi is about  $\times 7.7$  on eight devices.

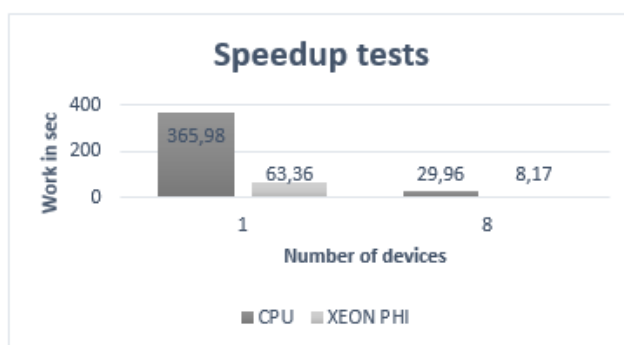


Fig. 4. A speedup test.

Based on the above-mentioned results, we conclude that carrying out computations on Intel Xeon Phi coprocessors for the large-scale seismic field simulation is a promising approach.

## Conclusion

We presented the results of the research into developing a scalable parallel algorithm and program software. We proposed a new software for simulation of the elastic wave propagation in 3D isotropic elastic medium using hybrid supercomputers with Intel Xeon Phi coprocessors. We described the parallel implementation of the difference method based on 3D domain decomposition and using computing devices in offload mode. In the figures presented, the efficiency of a using such computing device for similar difference methods is shown. We have carried out computing experiments and investigated the behavior of the program on one Xeon Phi coprocessor to tune the script parameters to running the program for a greater number of computing devices placed at cluster nodes. It is shown that the 3D difference method with staggered grids can be well parallelized with Intel MIC architecture. We can use the discussed computing devices to simulate big size models. The results of the research done are important and can be of practical use in the field of developing scalable parallel algorithms for exaflops

supercomputers [3] of the future and modeling its behavior on a greater number of computing cores in simulation systems.

**Acknowledgments.** This work was supported in part by the RFBR grants No. 14-07-00832, 14-05-00867, 15-07-06821, 15-31-20150 and MES RK 1760/GF4 and Intel Corporation.

## References

1. Glinsky B.M., Karavaev D.A., Kovalevsky V.V., Martynov V.N: Numerical modeling and experimental research into B«Karabetova MountainB» mud volcano using vibroseismic methods. *Vichislitelnie metody I Programirovanie*. Vol. 11, pp. 95–104 (in Russian)(2010)
2. R. W. Graves: Simulating seismic wave propagation in 3D elastic media using staggered grid finite differences. *Bull. Seism. soc. Am.*, vol. 86, pp. 1091-1106 (1996)
3. A.R. Levander: Fourth-order finite-difference P-SV seismograms. *Geophysics*, vol. 53, issue 11, pp. 1425-1436 (1988)
4. Virieux J.: P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method . *Geophysics*, Volume 51,Number 4, pp. 889–901 (1986)
5. Moczo, P., Kristek, J. and Halada, L.: 3D fourth-order staggered-grid finite difference schemes: stability and grid dispersion. *Bull. Seism. Soc. Am.*, Vol. 90, No. 3, pp. 587–603 (2000)
6. J.O. A, Robertsson, J.O. Blanch, and W.W. Symes: Viscoelastic finite-difference modeling. *Geophysics*, vol. 59, issue 9, pp. 1444-1456 (1994)
7. Dimitri Komatitsch: Fluid-solid coupling on a cluster of GPU graphics cards for seismic wave propagation. *Comptes Rendus MΓ©canique*, Volume 339, Issues 2-3, pp. 125–135 (2011)
8. Dimitri Komatitsch / Dimitri Komatitsch, Gordon Erlebacher, Dominik GΓ¶ddeke, David MichΓ©a /: High-order finite-element seismic wave propagation modeling with MPI on a large GPU cluster. *Journal of Computational Physics*, Volume 229, Issue 20, pp. 7692–7714 (2010)
9. Landau L.D., Lifshitz E.M.: *Theory of Elasticity*. Third Edition, (1986)
10. Chernykh I., Glinskiy B., Kulikov I., Marchenko M., Rodionov A., Podkorytov D., Karavaev D.: Using Simulation System AGNES for Modeling Execution of Parallel Algorithms on Supercomputers. *Computers, Automatic Control, Signal Processing and Systems Science. The 2014 Int. Conf. on Applied Mathematics and Computational Methods in Engineering*, pp. 66–70 (2014)

# Parallelization of Algorithm of Prediction of miRNA Binding Sites in mRNA on The Cluster Computing Platform

Anna Yu. Pyrkova<sup>1</sup>, Anatoli T. Ivashchenko<sup>2</sup>, and Olga A. Berillo<sup>2</sup>

<sup>1</sup> Mechanic-mathematical faculty, Al-Farabi Kazakh National University, Almaty, Kazakhstan

<sup>2</sup> Faculty of biology and biotechnology, Al-Farabi Kazakh National University, Almaty, Kazakhstan  
{Anna.Pyrkova, Anatoli.Ivashenko}@kaznu.kz, %Devolia18@mail.ru

**Abstract.** In presented article the solution of the problem of gene scanning for the purpose of prediction of binding sites of miRNA with matrix RNA (mRNA) is proposed. During the research by authors the following results were received: the mathematical model of optimum process of scanning of genes and miRNA sequences is developed; the algorithm of scanning of genes with miRNA with one gap in miRNA and maximum (in a percentage ratio) free energy is developed and analyzed at coincidence of miRNA and a gene site on the basis of complementarity properties; the constructed algorithm of scanning of genes with miRNA is parallelized on the computational cluster with use of MPJ tools - the MirTarget program; the assessment of overall performance of the parallelized algorithm on the cluster computing platform with consecutive algorithm is performed when processing large volumes of data; the developed program was used for performing researches by search of binding sites of miRNA with matrix RNA (mRNA).

**Keywords:** parallelized algorithm, cluster computing platform, Java MPI, miRNA, mRNA.

## 1 Introduction

After opening of an important role of microRNA (miRNA) in regulation of an expression of genes the problem of a prediction of binding sites of miRNA with matrixRNA (mRNA) has arisen. Some programs which predicted binding sites of miRNA were created. However many of them had unreasonable restrictions for search of binding sites. Earlier it was claimed that binding sites are localized only in 3'UTR. It was established later that binding sites are localized in 5'UTR and CDS. Other programs were based on identification of binding sites with the obligatory requirement to have complementary interactions of a guanine (G) and an adenine (A) in a site of "seed" which corresponds 5'-end of miRNA. Many such programs predicted a large number of false positive sites and did not allow revealing the binding sites located in 5'UTR and CDS. On this and other reasons it is inexact the beginning of binding sites was established and incorrectly schemes of interaction of miRNA with mRNA were formed. Now, in a genome of the human more than 2500 miRNAs are known and it is necessary for each of them to find target genes among 30 thousand genes of the human. Large volume of calculations demands creation of the program, allowing processing these huge data files. We created the MirTarget program which has no shortcomings given above and with big reliability finds binding sites of miRNA with mRNA.

## 2 Scanning gene problem definition

Scanning genes [1], [2] is a process of consecutive comparison of sites of a gene with miRNA with possibility of adding one gap in miRNA in positions with the 3rd on n-2-th, where by n – nucleotide number (length) of miRNA. Thus there is an assessment all of possible comparisons on one site of mRNA with miRNA which is defined according to the value of free energy of compared sequences. It is considered the best that option which is closer (in a percentage ratio) on free energy for coincidence of miRNA and a gene site on the basis of a complementarity.



Scanning of a genome allows to reveal hundreds possible targets for therapy of various diseases. Such scanning is important because knowledge the interacting genes will allow to define that, for what this or that protein and, respectively, what intracellular processes answers are broken at this disease.

The mathematical model of a problem of scanning genes can be formulated in the following view:

Let  $\{u_l\}, l = \overline{1, N}$  is a set of nucleotide or amino-acid sequences *miRNA*,  $N$  is amount of sequences *miRNA*, and  $\{v_g\}, g = \overline{1, M}$  is a set of sequences of genes *mRNA*,  $M$  is amount of sequences *mRNA*, then  $\langle u_l, v_g, Number, Position, Where, Energy, Score, Length \rangle_{l=1, N, g=1, M}$  is scanning genes, where *Number* is order number, *Position* is a position of  $\{u_l\}$  in  $\{v_g\}$ , *Where* is an element from a set  $\{5'UTR, CDS, 3'UTR\}$ , defining site arrangement area  $\{u_l\}$  in  $\{v_g\}$ , *Energy* is value of free energy on the basis of a complementarity, *Score* is value of  $\Delta G/\Delta G_m$ , *Length* is length of  $\{u_l\}$ .

### 3 Algorithm of the program of scanning genes on the cluster

1 step. Verification of attributes in a command line.

```
mpjrun.bat -np
            number\_of\_parallel\_processes
            /path/to/Base
            minimal\_percentage\_of\_convergence
            directory\_with\_genes
            directory\_with\_miRNAs
            file\_name\_of\_result\_mres
            file\_name\_of\_result\_xls
```

For example,

```
mpjrun.bat -np 10 Base 80 gene mir ResultsFull ResultsBrief
```

2 step. The main process (MASTER) invokes the function which forms the geneName array of the files containing genes.

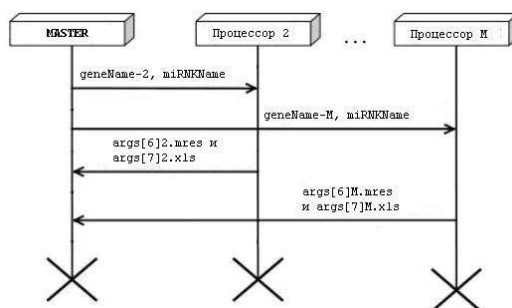


Fig. 1. The protocol of exchanging messages between MPJ processes.

3 step. The main process (MASTER) invokes the function which forms two arrays for miRNAs:

- 1) miRNAName array of the miRNA names;

2) miRNA array of miRNA sequences. Thus there is a check of the format of the file with miRNA.

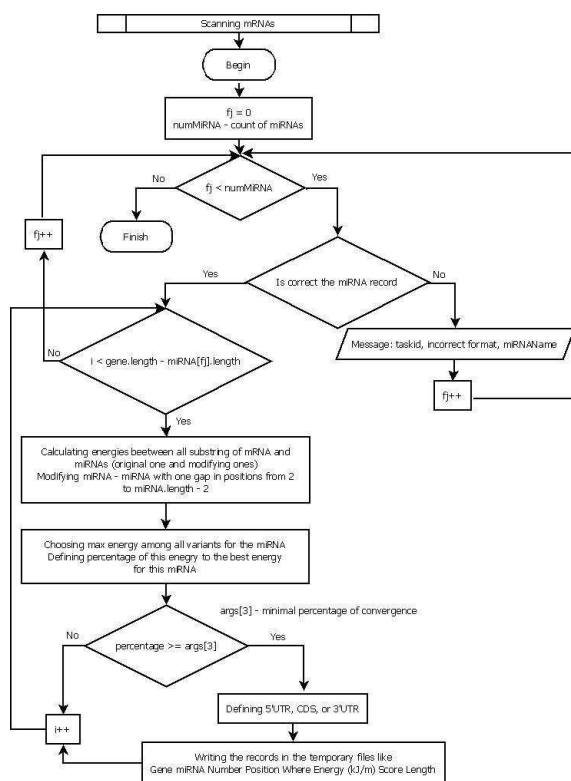
```
>let-7a-2-3p          MIRLET7A2-ex-cod  11
CUGUACAGCCUCCUAGCUUCC
```

**Fig. 2.** File structure with miRNA (.mir).

4 step. The main and parallel processes create files with the names  $args[6] \langle task\_id \rangle .mres$  and  $args[7] \langle task\_id \rangle .xls$  for saving result.

5 step. The main process (MASTER) divides genes between parallel processes and dispatches a certain part of the geneName and the miRNAName and miRNA arrays. The main process also receives part of genes for scanning.

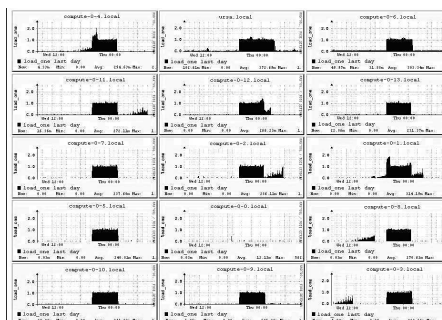
6 step. Processes read out consistently genes from files of the geneName array and check the file format with gene.



**Fig. 3.** Flowchart of procedure of scanning mRNAs.

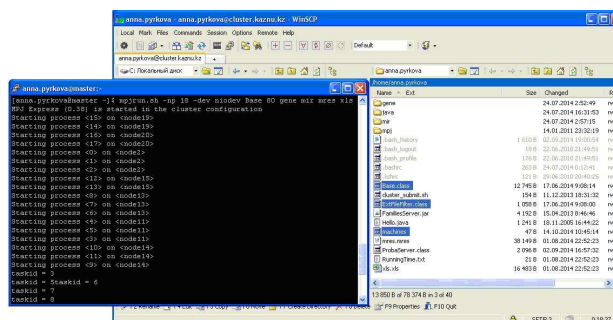
7 step. Processes scan their genes consistently with miRNA, allowing one gap in miRNA in positions from the 3rd to the  $n-2$ -th where by  $n$  – nucleotide number (length) of miRNA, and choosing that option which is closer (in a percentage ratio) on free energy for coincidence of miRNA and a gene site on the basis of a complementarity.

8 step. Processes form the  $args[6] \langle task\_id \rangle .mres$  and  $args[7] \langle task\_id \rangle .xls$  files and after completion of processing of their part of genes send messages about completion to the main process.



**Fig. 4.** Load of the cluster (For the database containing 13016 genes and 100 miRNAs operating time is 2 hours 15 minutes 19 seconds).

9 step. The main process (MASTER) receives messages from parallel processes, copies information from the corresponding *args[6]<task\_id>.mres* and *args[7]<task\_id>.xls* files in the *args[6].mres* and *args[7].xls* files, and deletes the *args[6]<task\_id>.mres* and *args[7]<task\_id>.xls* files.



**Fig. 5.** Running the application on the cluster computing platform cluster.kaznu.kz.

## 4 Conclusion

The MirTarget program has advantages which are not present in known programs of predicting of binding sites of miRNA with mRNA. In literature there are many data about the value of free energy of hydrogen bond between nucleotides in water solution [3]. However there is a wide spacing of value of free energy of this bond and it is difficult to give preference to certain data [4], [5]. It is important to know the relative relations of free energy of hydrogen bond between nucleotides as they are necessary at formation of RNA of secondary and tertiary structures. The analysis of free energy of the hydrogen bond arising between nucleotides at intramolecular interaction of mRNA at formation of its secondary structure showed that between nucleotides of G-C is formed three, between A-U – two and between G-U and A-S – on one hydrogen bond. The relation of free energy of hydrogen bond in G-C and A-U pairs approximately corresponds to the relation of forces of their interaction 3:2 (0.188 nNewton and 0.125 nNewton) [6]. The value of free energy of one hydrogen bond between nucleotides changes in the range from -0.7 to -1.6 kcal/mol [7]. In the MirTarget program free energy of interaction of nucleotides due to hydrogen communications was considered as equal to 6.368 kJ/mol and 4.246 kJ/mol for G-C and A-U pairs, and 2.123 kJ/mol for G-U and A-S pairs, respectively.

Gene	miRNA	Number	Position	Where	Energy (kJ/m)	Score	Length
>SUPT3H-v-3	>miR-11(C20...)	1	2000	3'UTR	-87.0443	81.99	22
5' - AUCCACCACUCGCGCAGCAGCCAA - 3'							
3' - UAUGUA-UGAAGAAAUGUAAGGU - 5'							
>SUPT3H-v-3	>miR-9-3p(...)	2	1904	3'UTR	-87.0451	80.39	22
5' - ACUUCUCGCCAUCUGCCUUUCAU - 3'							
3' - UGAAAAGCCAAUAGAUCCGAAA-UA - 5'							
>SUPT3H-v-3	>miR-24-2-5p(ig)	3	878	CDS	-93.4128	80.00	22
5' - UUAUGGAUUCAGCUCAAUAUGCA - 3'							
3' - GACACA-AAGUCGAGUCAUCCGU - 5'							
>SUPT3H-v-3	>let-7c(...)	4	1215	CDS	-91.2866	81.12	22
5' - GGCCAUUCGACGCCUACAGCCACA - 3'							
3' - UUGGUAUGUUG-GAUGAUGGAGU - 5'							
>SUPT3H-v-3	>let-7c-5p(LIN...)	5	1215	CDS	-91.2866	81.12	22
5' - GGCCAUUCGACGCCUACAGCCACA - 3'							
3' - UUGGUAUGUUG-GAUGAUGGAGU - 5'							
>SUPT3H-v-3	>miR-15b-5p(SMC...)	6	1308	3'UTR	-93.4120	83.01	22
5' - UGACAAACUCUCGAUGUGCCUCCAA - 3'							
3' - ACAUUGGUACUACACG-ACGAU - 5'							
>ARHG8-v-1	>miR-15a-5p(DLE...)	7	4319	3'UTR	-93.4145	83.01	22
5' - CACAAAACACUAHAGUCUUGCUA - 3'							
3' - GUGUUGGUAUA-CAGCAGCGAU - 5'							

Fig. 6. Result of scanning genes on the cluster (the amount of the processed sequences makes 13016 mRNA and 100 miRNA). Speed of work of the algorithm parallelized on 18 nodes is 66.5 times higher than the speed of work of serial algorithm.

Table 1. Scanning genes (time is specified in seconds), total of genes - 13016

	Duration of processing 100 mRNA and 100 miRNA (average length - 21)	Duration of processing 1000 mRNA and 100 miRNA (average length - 21)	Duration of processing 10000 mRNA and 100 miRNA (average length - 21)
Linear algorithm	3	54	532
Parallelized algorithm (dual-core processor)	1	25	69
Parallelized algorithm (a cluster platform of 15 nodes)	1	3	8

The distance between nucleotides G-C and A-U pairs makes 1.03 nm, between nucleotides G-U pair it is equal 1.02 nm and between nucleotides A-S pair equally 1.04 nm [8]. Therefore, formation of hydrogen bonds between these couples of nucleotides allows two-chained structure of mRNA to have a spiral form similar to DNA regular structure. Such structure of mRNA except hydrogen bonds is stabilized by stacking-interactions between the nitrogenous bases [9]. Between nucleotides of couples purine-purine and pyrimidine-pyrimidine distances significantly differ from 1.03 nanometers: distances of A-A, G-A and G-G are equal to 1.23 nm, 1.25 nm and 1.25 nm, respectively. In couples of pyrimidine-pyrimidine distances between nucleotides too significantly differ from 1,03 nanometers: distances of C-C, U-U and U-C are equal to 0.85 nm, 0.81 nm and 1.18 nm, respectively. Therefore, in such couples hydrogen bonds are not formed, and these couples will break regular structure of two-chained miRNA with mRNA reducing stability of

the RISC complex (RNA-induced silencing complex). Therefore, in the program such couples of hydrogen bonds were not considered.

At alignment of the nucleotide sequences of miRNA with mRNA we assume existence of only one admission on miRNA (lack of complementary couple of hydrogen bond) that allows considering binding sites of mRNA longer miRNA on one nucleotide. In this case the regular structure of a spiral is broken and there is its bulg. Free energy of binding miRNA with mRNA of such structure is less, than in alternative case. The program determines free energy of hybridization ( $\Delta G$ , 100 kJ/mole) of miRNA with mRNA and the scheme of their interactions, calculation of the relations  $\Delta G/\Delta G_m$ , levels of reliability (p) and the mRNA areas, where the site (5'UTR, CDS or 3'UTR), since the first nucleotide 5'UTR, is located.  $\Delta G_m$  is equal to free energy of binding miRNA with completely complementary to it a site of nucleotide sequence of miRNA. Level of reliability (p) was defined on the basis of value  $\Delta G$  and its standard deviation. The program outputs the scheme of interaction of miRNA with mRNA, a site position in 5'UTR, CDS or 3'UTR, free energy of interaction of miRNA with mRNA, and its relative value from the maximum energy of binding miRNA. In the program the threshold value of this relation is set, this value allows not considering sites with weak free energy of binding.

## References

1. Lesk Arthur M. Introduction to Bioinformatics. - Oxford: Oxford University Press, 2002. - 255 p.
2. Jones Neil C., Pevzner Pavel A. An Introduction to Bioinformatics Algorithms. - Massachusetts: Massachusetts Institute of Technology Press, 2004. - 435 p.
3. Guckian K.M, Schweitzer B.A, Ren RXF, Sheils C.J, Paris P.L, et al Experimental measurement of aromatic stacking in the context to duplex DNA. // J.Am.Chem.Soc. - 1996. - 118: 8182-83.
4. Turner D.H, Sugimoto N, Kierzek R, Dreiker S.D. Free energy increments for hydrogen-bonds in nucleic acid base pairs. // J. Am. Chem. Soc. - 1987. - 109: 3783-85.
5. Sugimoto N, Kierzek R, Turner D.H. Sequence dependence for the energetics of dangling ends and terminal base pairs in ribonucleic acid. // Biochemistry. - 1987. - 14: 4554-58.
6. Boland T., Ratner B.D. Direct measurement of hydrogen bonding in DNA nucleotide bases by atomic force microscopy. // Proc. Natl. Acad. Sci. USA. - 1995. - V. 92. - Pp. 5297-5301.
7. Kool E.T. Hydrogen bonding, base stacking, and steric effects in DNA replication. // Annu. Rev. Biophys. Biomol. Struct. - 2001. - 30: 1-22.
8. Leontis N.B., Stombaugh J., Westhof E. The non-Watson-Crick base pairs and their associated isostericity matrices. // Nucleic Acids Res. - 2002. - 30(16): 3497-3531.
9. Richard A. Friedman, Honig B.A. Free Energy Analysis of Nucleic Acid Base Stacking in Aqueous Solution. // Biophysical Journal. - 1995. - V. 69. - Pp. 1528-1535.

# Distributed PIV: the Technology of Processing Intensive Experimental Data-flow on a Remote Supercomputer

Vladislav Shchapov<sup>1,2</sup>, Alexey Masich<sup>2</sup>, and Grigoriy Masich<sup>1,2</sup>

<sup>1</sup> Perm National Research Polytechnic University, Perm, Russia

<sup>2</sup> Institute of continuous media mechanics UB RAS, Perm, Russia

{shchapov,mag,masich}@icmm.ru

<http://www.icmm.ru/>

**Abstract.** A model for entering the structured stream of experimental data into a remote supercomputer is developed and the design model parameters of the data transfer path are determined. The infrastructure of the distributed data processing system is shown to elucidate the process of managing the intensive data stream from the source to a remote supercomputer. The results of measurement illustrating the efficiency of the developed architectural solutions are presented. The investigation is supported by the RFBR (grants № 14-07-96001 and № 14-07-96003).

**Keywords:** experimental system, supercomputer, high-speed optic network, data exchange model, throughput measurement.

## Introduction

In the present paper, the process of inputting an intensive data stream into supercomputers is investigated and the developed technology of parallel data transmission is discussed. The most common way of data processing is saving the obtained data on the external data storage devices and their subsequent processing on computer integrated with experimental stand. However, in-situ data processing is not always possible because of the need for high-performance computing and/or large-capacity storage systems. A well-known example of the distributed system is a three-level architecture for processing large amounts of data from LHC (Large Hadron Collider). The classical regime used for processing big data on supercomputers includes three stages: (1) upload of data into a storage system of supercomputer; (2) data processing on supercomputers, (3) download of the processed data from the storage system of supercomputer (shown on the right of Fig. 1). Upload and download of data in/from the storage system (stages 1 and 3) and data processing stage (2) are fulfilled using either file transfer protocols (FTP/Grid FTP and SCP) or a direct access to data storage system using file system protocols (CIFS and NFS/pNFS). We have developed an architectural solution, which allows entering the items of the structured data stream into compute nodes of the remote supercomputer at the rate of their generation and their parallel processing in accordance with the memory-memory model (shown on the left of Fig. 1) bypassing the data storage system.

The structured stream should be understood to mean a sequence of disconnected messages (measurements), which can be processed independently of one another by the applied algorithms and therefore can be handled in parallel. One of the examples of such a data stream is a pair of PIV (Particle Image Velocimetry) images [1] containing tracer particles, the displacements of which allow us to compute vector fields of velocity. However, the performance mismatch between I/O and computing components of current-generation HPC systems has made I/O the critical bottleneck for distributed applications. One of the main reasons of worsening the total computing power of geographically distributed high-speed applications is inadequate end-to-end throughput of widely used TCP protocols. The most effective way to remedy this deficiency is to employ

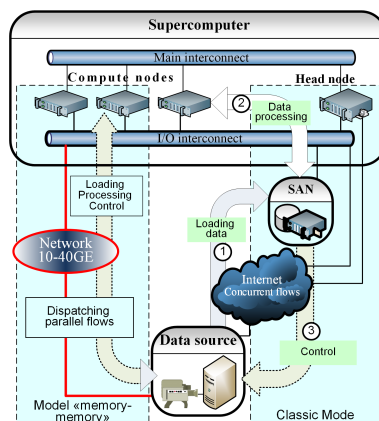


Fig. 1. Classical and developed (memory-memory) scheme of data input.

the mechanism of parallel transmission (Grid FTP, pNFS). In this case, there are two main problems that immediately move to the forefront: implementation of the effective data transfer to the compute nodes of the remote supercomputer and distribution of the data stream items over compute nodes. This paper introduces the diagram of the interaction between the source of intensive data stream and remote supercomputer and presents the estimated parameters of the data path and measurement data illustrating the capabilities of the developed I/O middleware. The paper is organized as follows. The second Section outlines currently available solutions, the third Section describes the data transfer model. In the fourth Section, consideration is given to the infrastructure of the distributed system of data processing and the fifth Section is devoted to the analysis of measurements and estimation of the obtained results.

## Existing solutions

Today there are a few practical ways of organizing data transfer systems in the form of queue structures:

- systems intended for usage as a service accessible via the Internet. The best known example of this technology is Amazon Simple Queue Service [2], which is a constitutive part of the Amazon cloud platform;
- queue management systems, for example, those implementing Message Exchange Protocol AMQP (Advanced Message Queuing Protocol) [3], such as RabbitMQ;
- queue management and message exchange libraries, of which Zero MQ [4] is the most abundant messaging system.

The systems providing services cannot be involved in further comparative estimation due to impossibility of their installation into our infrastructure because of safety and efficiency requirements. The queue management systems execute a total cycle of operations including message dispatching and their storage. Among various implementations of the protocol that employ such systems, the AMQP (Advanced Message Queuing Protocol) has the advantage of being the most open messaging protocol this is an open standard designed to allow dispatching of messages to different components of the distributed system. The logic of the protocol operation declares that the exchange of data between the components is accomplished via specialized AMQP brokers, which perform scheduling and ensure message delivery, distribution of data

streams and subscription for the required types of messages, etc. The best known implementation of this standard is RabbitMQ, which is a cross-platform queue manager written in the Erlang programming language. Among low-level libraries of queue management and message exchange, the ZeroMQ library is in most common use. ZeroMQ is a socket interface superstructure, which offers ample possibilities for creating systems oriented toward the interaction via message transmission. The efficiency of the developed SciMQ queue manager will be estimated by comparing it with the capabilities of RabbitMQ.

## Model of data stream transmission

The necessary condition for parallel processing of the structured stream messages on the super computer is the validity of the relation  $\rho = \lambda/\mu$  where  $\lambda$  is the intensity of arrival of the experimental data stream (messages per second),  $\mu$  is the intensity of processing (servicing) the input data stream by supercomputer and  $\rho$  is traffic intensity of the supercomputer. Clearly, to maintain stable working of such a system, it is necessary that  $\rho < 1$ . This requirement can be fulfilled under the following conditions. First, it is essential that the Net transfer capacity  $v$  should not impede the data stream ( $\lambda \leq v$ ) generated by the experimental stand. Second, the number of compute nodes  $n$  in the supercomputer should be sufficiently large to allow processing of input data at the rate of their arrival. If the time of processing of one measurement by the compute node is  $T_\mu$ , then the condition for vitality of the "Source-Net-Supercomputer" tract can be expressed as

$$\lambda \leq v \leq \frac{n}{T_\mu}. \quad (1)$$

If the accessible end-to-end throughput over a single channel of multi-channel system is  $v_i$ , then the expected result for the system of  $m$ -parallel channels is  $v = m \cdot v_i \cdot \delta_v$ , where  $\delta_v$  is the effectiveness of parallel data transmission. For a perfect parallel system, the parallel efficiency tends to unity ( $\delta_v \rightarrow 1$ ). The investigation of possibilities of organizing the effective parallelism ( $\delta_v$ ) in all components of such measuring system (decrease of  $T_\mu$  and increase of  $v$ ) with the aim to generally improve its throughput is the objective of research directed towards a design of distributed systems. Decrease the time of processing of one measurement by the supercomputer (parameter  $T_\mu$ ) is the task of applied parallel program designers and is not considered in the present paper. Here, our primary concern is to develop the architectural solutions for the problem of input of an intensive data flow in a supercomputer including the evaluation of the required number of compute nodes  $n$  and parallel transport channels  $m$  at the prescribed intensity of measurement stream generation  $\lambda$  and characteristic time of single message processing  $T_\mu$ . The developed model of message transition is based on the idea of arranging the data stream from the Source as a joint queue of messages (Fig. 2), which are distributed by the queue manager over the compute nodes on their request in the FIFO (First In, First Out) order.

The time diagram of single message processing shown in fig. 3 is interpreted in terms of the interaction of the exchange protocol primitives between terminal systems.

**Data source** generates (POST) messages of length  $S_\lambda$  at  $T_\lambda$  interval, which are inserted into the queue of ready-for-dispatch messages from the queue manager.

**Compute nodes** of the supercomputer serve to open TCP connections to the queue manager over the whole period of experiment and send requests (GET) for message reception. The received message (GET REPOSE) is processed by the applied program for time  $t_\mu$  and the obtained computation result of length  $S_\mu$  is returned by the compute node to the queue manager (POST). After receiving the acknowledgement of receipt (POST RESPONSE) the compute node requests (GET) for the next group of data and the cycle is repeated.



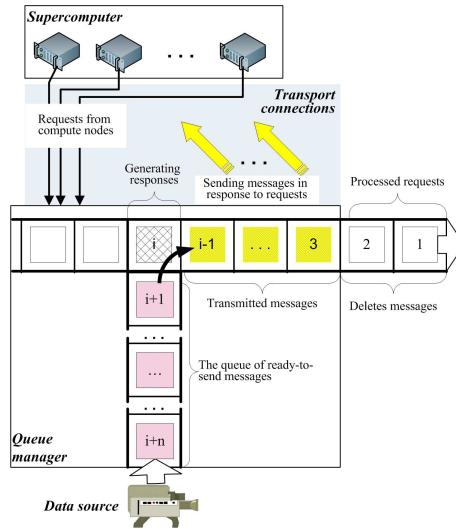


Fig. 2. Model of message distribution over compute nodes.

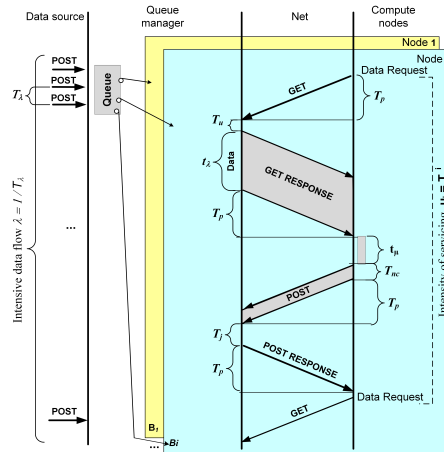


Fig. 3. Time diagram of message processing.

**Queue manager**, on request of the compute node for data (GET), sends the message (GET RESPONSE) to the compute node and translates the delivered message into the "transmission status". The queue manager acknowledges the receipt of computational results from the compute node (POST RESPONSE), assigns to the transmitted message the status "processed" in the log file and deletes the message from the memory. The time it takes for a compute node to process one message is

$$T_\mu = t_\mu + t_\lambda + 4 \cdot T_p + (T_u + T_{nc} + T_j) \tag{2}$$

where  $t_\lambda$  is the transmission time of messages of length  $S_\lambda$ ,  $t_\mu$  is the time of message processing (computation time),  $T_p$  is the time it takes for a signal to travel through the physical medium of data transition. The time of signal propagation through the optical fiber of 1 km is  $0.005 \mu s$ . Then the time of signal propagation through the optical fiber line of length  $L$  km is

$$T_p = L \cdot 0.005, \tag{3}$$

$T_u$  is the time it takes for a manager to remove the next message from the queue,  $T_{nc}$  is the transmission time of computation results of size  $S_\mu$ ,  $T_j$  is the time it takes for the results of computation to be transferred to the applied program of Source to transfer.

Total time spent by the queue manager  $T_m$  to fulfill all these operations is defined by the sum  $T_m = T_u + T_{nc} + T_j$ . Then expression (2) takes the following form:

$$T_\mu = t_\mu + t_\lambda + 4 \cdot T_p + T_m. \quad (4)$$

At  $T_m \rightarrow 0$  the time of processing of a single message is defined by the transmission capacity of the data path of the existing technical systems. Below, we will present the constitutive relations for the design parameters and boundary conditions of the developed model, which have been derived with account for the procedures of primitive interaction shown in the time diagram. At the prescribed intensity  $T_\lambda$  of the input data stream ( $\lambda = 1/T_\lambda$ ) and the known time of processing of a single message by the compute node  $T_\mu$  we obtain from relation (1) the estimated number of compute nodes

$$n = \frac{T_\mu}{T_\lambda}. \quad (5)$$

From the time diagram it also follows that the traffic through one connection is of pulse character with the relative pulse duration defined by  $t_\lambda/T_\mu$ . Then, the number of parallel connections  $m$  providing continuous data stream via the Source-Supercomputer path is found from the relation  $t_\lambda/T_\mu \cdot m = 1$ , which yields

$$m = \frac{T_\mu}{t_\lambda}. \quad (6)$$

According to (5) and (6), the number of required active connections  $m$  for the estimated number of compute nodes  $n$  is equal to  $m/n = T_\lambda/t_\lambda$ . It means that at  $T_\lambda > t_\lambda$  it is necessary to add to the pre-existing  $n$  connections the  $T_\lambda/t_\lambda - n$  number of active connections. Substituting in (6) the values of  $T_\mu$  from relation (4) yields the value of the number of parallel connections  $m$  ensuring a complete utilization of the data transmission path.

$$m = \frac{t_\mu}{t_\lambda} + 1 + 4 \cdot \frac{T_p}{t_\lambda} + \frac{T_m}{t_\lambda}. \quad (7)$$

Since the goal of any design project is to decrease time expenditures of queue manager, in our further considerations we will operate on the supposition that  $T_m = 0$ . Thus we have

$$m_{min} = \frac{t_\mu}{t_\lambda} + 1 + 4 \cdot \frac{T_p}{t_\lambda}. \quad (8)$$

**Let us interpret the obtained results.** From relation (8) it follows that at  $t_\lambda \rightarrow \infty$  it will be sufficient to utilize one active connection  $m = 1$  of  $n$  existing connections at any length of the communication line  $L$ . This is achieved in systems, which have no limits on the data transfer rate  $v = S_\lambda/t_\lambda$ . At  $t_\mu/t_\lambda = 1$  and the length of communication line  $L = 0$  we get  $m = 2$ , which is the case when the supercomputer is located near the source. Finally, for extended communication lines when  $4 \cdot T_p/t_\lambda \gg n$  it will be necessary to organize some connections more in addition to already existing  $n$  connections. The number of newly organized connections is equal to  $4 \cdot T_p/t_\lambda - T_\mu/T_\lambda$ . Thus for the known length of the optical fiber line  $L$  km, the known intensity of the data source generating messages of length  $S_\lambda$  at the interval  $T_\lambda$  and the known characteristic time  $t_\mu$  of one message processing by a compute node, we obtain the following design parameters of the system:

- $v = S_\lambda/T_\lambda$  is the network speed;
- $n = T_\mu/T_\lambda$  is the number of compute nodes;
- $m = (4L \cdot 0.005/t_\lambda) + t_\mu/t_\lambda + 1$  is the number of parallel transmissions;
- $T\mu = t_\mu + t_\lambda + 4L \cdot 0.005$  is the delay in the reception of computation results.

## Infrastructure

Verification of the developed architectural solutions carried out in the context of the problem of processing an intensive data stream obtained on the experimental system PIV (ICMM UB RAS, Perm) by the supercomputer URAN (ICMM UB RAS, Ekaterinburg) in the framework of the Distributed PIV project, which is one of the components of the Cyberinfrastructure developed in the Ural region [5]. Fig. 4 shows the diagram of the distributed system infrastructure.

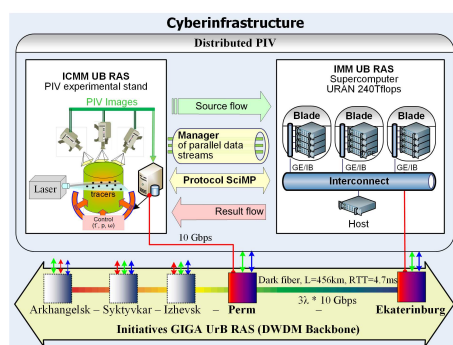


Fig. 4. Infrastructure of distributed system.

The source of data is the experimental PIV system (Particle Image Velocimetry), which is based on the optical method of measuring the velocity fields of liquids or gases in the selected cross section of the flow. The method consists in processing of pairs of images of tracer particles (small particles suspended in the flow) at the instants of time when they are illuminated by a pulse laser generating a thin light beam. The flow velocity is determined by calculating tracer displacements over the time interval between the laser flashes. The intensity of the initiated data stream depends on the number, resolution and operating frequency of the cameras and can reach the value of dozens of gigabytes per second. The PIV system used at the ICMM UB RAS operates in the following basic regimes:

- frame frequency of 4 Hz, image of 1,3 Mpxps;
- frame frequency of 0.5 Hz, image of 11 Mpxps.

Thus, for double-tap shooting the transferred messages will have the following pairs of parameters: ( $T_\lambda = 250ms$ ,  $S_\lambda = 2.6MB$ ) and ( $T_\lambda = 2s$ ,  $S_\lambda = 22MB$ ). In view of the tendency of increasing the frame frequency and resolution of the cameras, the calculations have been made using the following parameter ranges: the Source intensity  $T_\lambda$  is { 50 ms, 100 ms, 250 ms, 500 ms, 1 s, 2 s }; the length of the message  $S_\lambda$  is { 2 MB, 4 MB, 8 MB, 16 MB, 22 MB }; The PIV system is provided with the mechanisms of experiment management (rotation, pressure, temperature), which can be controlled based on the results of current calculations on the supercomputer. **Network.** Connection of the Source to the Supercomputer is accomplished via communication channels of the scientific and educational optical backbone with wavelength

division multiplexing, which is being developed in the framework of the project "Initiative GIGA UB RAS-[4]. The Ethernet ports of the interface boards in the Perm-Ekaterinburg segment of DWDM network provide data transfer rate of 1-10 Gbps. The estimated time for transmission of 2MB and 22 MB of messages at the rate of 1GB/s is 0.26 ms and 2.9 ms, respectively and at the rate of 10 Gbps – 0.026 ms and 0.29 ms. For the Perm-Ekaterinburg fiber optic network of length of 456 km the estimated time of message transmission is  $T_p = 2.28ms$ . The observed time lag is verified experimentally by measuring RTT (Round Trip Time) between terminal switching nodes of Perm – Ekaterinburg highway. With account for delay in the switching cloud (L2 OSI RM) it is equal to

```
rtt min/avg/max/mdev = 5.285/5.300/5.341/0.102 ms (1 GB/s)
rtt min/avg/max/mdev = 5.198/5.210/5.237/0.077 ms (10 GB/s).
```

**Supercomputer URAN** is the MPP system (<http://parallel.uran.ru/node/3>) with the Infiniband and Ethernet Interconnect and peak performance of 240 Tflops. The time of single message/data processing by supercomputer is  $T_\mu$ , which, as mentioned above, is defined by the efficiency of parallel applications. In the following analysis we shall use values multiple of  $T_\lambda$ . Service rate is  $T_\mu = T_\lambda, 5T_\lambda, 10T_\lambda, 50T_\lambda$ .

**Manager/dispatcher** [6,7] is installed on a separate server with 1-10 GE interfaces located close to the PIV system. The SciMQ manager is the main part of the specialized application environment known as the middleware. Applications on the generating and processing levels (end systems) operate together with SciMQ manager. Their interaction is provided by the SciMP protocol. The end system applications interact with the queue server via API provided by the client libraries. The queue server is handled by the developed management software. All components of the software package are written in C++ programming language using Boost library. The window size for transfer protocols with a feedback used in SciMP is determined by the value of  $BDP = bandwidth \cdot RTT$ , which is equal to 0.6625 or 6.5125 MB for the rates of 1 Gbps (RTT=5.3ms and 10Gbps (RTT=5.21ms), respectively. To ensure good performance of the algorithms of TCP overload control, it is essential that the overload window size  $cwnd$  should be not less than BDP.

## Measurements

The scalability of the software applications was verified using two servers HP ProLiant DL360p Gen8 (2x Intel Xeon CPU E5-2660, 2.20 GHz, 16 threads; RAM 128 GB; operating system CentOS 6.5) which are pooled by Data Transfer Network operating at the rate of 10 Gbps.

Fig. 5 shows the dependences of the transmission capacity of the developed queue server SciMQ and the existing queue server RabbitMQ on the message length  $S_\lambda$  and the number  $m$  of parallel queries sent to the server. From the plot it follows that in the case of using the SciMQ software for commonly practiced message length (2-22 MB) the throughput of the system is restricted to available data transfer rate through the communication channel. A comparison of the obtained results with the results of transmission speed tests performed for possible solutions on the basis of RabbitMQ shows that the developed software SciMQ allows dispatching the required volume of messages (more than 1 MB) at the data flow rates, which are 2-4 times higher than the data rates achievable with RabbitMQ.

The stability of the system operation with respect to time was also subjected to testing. To this end, the queue server operating at the rate of 10Gbps was connected to the internal switch of the interconnect of the ICMM supercomputer.

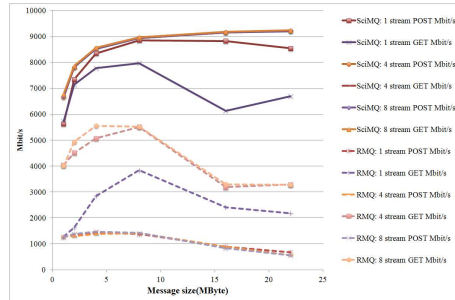


Fig. 5. Graphs of throughputs of SciMQ and RabbitMQ queue servers.

The data stream consists of messages of length  $S_\lambda = 16MB$ , which are transferred every  $T_\lambda = 16ms$  generating traffic at the speed of  $v = S_\lambda/T_\lambda = 8Gbps$ . During the experiment the system transferred 2700000 messages of total size of 41 TB. The number of supercomputer processors employed to receive the intensive data stream is  $n = 376$ . From the graphs (fig. 6) shown in the figure we may draw the following conclusions:

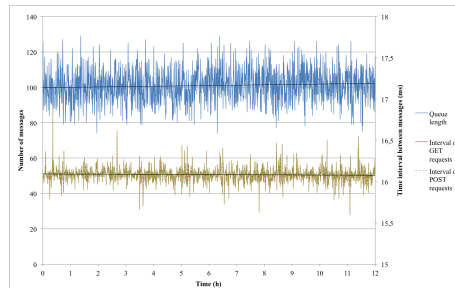


Fig. 6. Queue lengths (including the number of messages in-process) and average interval between the receipt and sending of messages to clients

- the system operates without overload and provides data stream processing without additional delays and message losses;
- the average parameters of the system operation are not variable in time;
- the software is capable of transferring continuous data stream;
- peak consumption of Random access memory for queue server is 2.03 GB ( $m_{max} = 130$  messages of length  $S_\lambda = 16MB$ ), and average consumption is 1.58 GB ( $m = 101$  messages of length  $S_\lambda = 16MB$ ), which coincides with the statistics of resource consumption run by the software.

## Conclusion

The developed technology offers a principally new instrument for conducting unique physical experiments both in the research laboratories and different industries. A distinguishing feature of the developed technology is the asynchronous model of data transfer from the source to remote supercomputer, which is based on the idea of parallel straightforward input of the structured transferred data stream to compute nodes of the supercomputer. High-speed data transmission

channels make it possible to connect the experimental site to supercomputers, which eliminates expenditures for upgrading the in situ computational resources.

This technology can be effectively applied to new-generation measuring systems, including high-tech measuring equipment used for in situ experiments and high performance computational facilities employed at supercomputer centers. At present, considerable progress has been made to incorporate the elaborated technology into the project on the development of technological platform for experimental and computational investigations of high-speed processes of hydroelasticity, which will realize the concept of coupling the measuring system with supercomputer.

## References

1. R. A. Stepanov, A. G. Masich, A. N. Sukhanovsky, V. A. Schapov, A. S. Igumnov, G. F. Masich Processing on the supercalculator of the stream of experimental data // Vestnik USATU. Ufa, Russia, 2012. V. 16, No 3 (48). P. 126133. (In Russian)
2. AWS | Amazon Simple Queue Service (SQS) - Queue Messaging Service. <http://aws.amazon.com/sqs/>.
3. Vinoski S. Advanced Message Queuing Protocol // Internet Computing, IEEE. 2006. Vol. 10, no. 6. P. 8789.
4. The Intelligent Transport Layer - zeromq. <http://www.zeromq.org>.
5. A. G. Masich, G. F. Masich Ot "Iniciativy GIGA UrB RAS" k Kiberinfrastrukture UrO RAN. // Vestnik Permskogo nauchnogo centra (oktjabr-dekabr 4/2009). Perm: Izd-vo PNC UrO RAN, 2009. S. 41-56 ISSN 1998-2097
6. Vladislav Shchapov, Alexey Masich. Protocol of High Speed Data Transfer from Particle Image Velocimetry System to Supercomputer // Proc. of The 7th International Forum on Strategic Technology (IFOST 2012) September 18-21, Tomsk Polytechnic University, Tomsk, 2012, - Vol.2., P. 653-657 DOI: 10.1109/IFOST.2012.6357642.
7. Shchapov V.A., Masich A.G., Masich G.F. A model of stream processing of experimental data in distributed systems // Vychisl. Metody Programm. 2012. Vol.13. P.139-145

# Analysing Modal Behaviour of Hybrid Systems by One-step Parallel Methods

Maria Nasyrova, Yury Shornikov, and Dmitry Dostovalov

Novosibirsk State Technical University,  
Prospekt K. Marksa, 20, 630073, Novosibirsk, Russia.  
Design Technological Institute of Digital Techniques SB RAS,  
Rzhanov st, 6, 630090, Novosibirsk, Russia.  
{maria\_myssak,dostovalov.dmitr}@mail.ru,shornikov@inbox.ru  
<http://www.nstu.ru>,<http://www.kti.nsc.ru>

**Abstract.** This paper discusses the analysis of hybrid models in the context of instrumental environment ISMA\_2015 supported parallel computations. Pj class of hybrid system is considered. The numerical schemes of Runge-Kutta methods with accuracy and stability control are described. The algorithm for choosing the integration step based on the analysis of event-driven function dynamics is presented. The results of performance estimation are given. Simulation results are presented on the example of generated reaction-diffusion problems based on the Lotka-Volterra model.

**Keywords:** hybrid systems, numerical analysis, Runge-Kutta methods, parallel methods, distributed architecture, simulation.

## 1 Introduction

Hybrid systems (HS) theory is a modern and versatile apparatus for mathematical description of the complex dynamic processes in systems with different physical nature (mechanical, electrical, chemical, biological, etc.). HS behavior can be conveniently described as sequential changes of continuous modes [1], [2]. This paper is mainly focus on numerical analysis of such modes. Each mode is given by a set of differential equations with the following constraints:

$$\begin{aligned}y' &= f(y, t), pr : g(y, t) < 0, \\t &\in [t_0, t_k], y(t_0) = y_0, y \in R^{N_y}, t \in R, \\f &: R^{N_y} \times R \rightarrow R^{N_y}, g : R^{N_y} \times R \rightarrow R^S.\end{aligned}\tag{1}$$

The vector-function  $g(y, t)$  is referred to as event function or guard [2]. The predicate  $pr$  determines the conditions of existence in the corresponding mode or state. The inequality  $g(y, t) < 0$  means that the phase trajectory in the current mode should not cross the border. Events occurring in violation of this condition and leading to transition into another mode without crossing the border are referred to as one-sided. Analytical analysis of HS is difficult and often impossible due to gaps in modal behavior. Therefore the research of HS dynamics is performed in special instrumental environments such as Charon (USA), AnyLogic (Russia), Scicos (France), Rand Model Designer (Russia), Hybrid Toolbox and HyVisual (USA), DYMOLA (Sweden), OpenMVLShell (Russia), ISMA (Russia), etc.

Many practical problems are characterized by stiff modes where the surface of boundary  $g(y, t) = 0$  has sharp angles or solution has several roots at the boundary [1]. Numerical analysis of such models by traditional methods is difficult to implement, as it gives incorrect results. To solve moderately stiff problems integration algorithms based on the explicit methods to control accuracy and stability of the numerical scheme can be applied [3], [4].

Furthermore when problem dimension reaches several thousands of equations and more its calculation by sequential methods becomes ineffective and requires the use of multiprocessor computer systems. In this situation parallel computation of local behaviours using cluster technologies can significantly improve the quality and efficiency of calculations. This paper discusses sequential and parallel implementation of algorithms of variable step based on the schemes of Runge-Kutta type. These integration algorithms are well suited for solving hybrid problems including moderately stiff problems.

## 2 Numerical methods

This section is devoted to the integration algorithms of variable step based on two-stage and three-stage explicit methods of Runge-Kutta type that implements schemes of second and third accuracy order respectively.

The algorithms are applied to numerically solve the Cauchy problem for ODE systems of the following form:

$$y' = f(y), y(t_0) = y_0, t_0 \leq t \leq t_k. \quad (2)$$

Consideration of autonomous problem does not reduce the generality because non-autonomous problem always can be cast to autonomous by introducing an additional variable.

Particular attention should be paid to the choice of the integration method. Fully implicit methods cannot be used because they require the calculation of  $f(y)$  at a potentially dangerous area, where the model is not defined. Therefore here we will use explicit methods with solution:  $y_{n+1} = y_n + h_{n+1}\varphi_n, n = 0, 1, 2, \dots$ . As a result we obtain the dependence of the predicted integration step  $h_{n+1}$ .

Considering that explicit methods are known by poor stability this paper examines integration methods with accuracy and stability control. Generally accuracy and stability control are used to limit the size of the integration step. As a result projected step  $h_{n+1}$  is calculated as follows.

The choice of the next integration step size is based on the proved theorem [4] and can be written as follows:

$$h_{n+1} = \max[h_n, \min(h^{ac}, h^{st})], \quad (3)$$

where  $h^{ac}$  and  $h^{st}$  are step sizes obtained as a result of accuracy control and stability control respectively. This formula allows to stabilize the step behaviour in the area of solution establishing where stability plays a decisive role. Because the presence of this area severely limits the use of explicit methods for solving stiff problems.

### 2.1 Two-staged Runge-Kutta method

Suppose that for numerical solution of problem (2) the following implicit two-stage method of Runge-Kutta type is used:

$$\begin{aligned} y_{n+1} &= y_n + 0.5(k_1 + k_2), \\ k_1 &= h_n f(y_n), \\ k_2 &= h_n f(y_n + k_1), \end{aligned} \quad (4)$$

where  $y$  and  $f$  are real  $N$ -dimensional vector-functions,  $t$  is an argument,  $h$  is an integration step,  $k_1$  and  $k_2$  are the method stages and 0.5 is a numerical coefficient defining accuracy and stability properties of (4).



Inequality for accuracy control has the following form [5]:

$$0.5 \| k_2 - k_1 \| \leq \varepsilon. \quad (5)$$

Inequality for stability control in turn looks as follows [6]:

$$v_n = 2 \max_{1 \leq i \leq N} (|k_3^i - k_2^i| / |k_2^i - k_1^i|) \leq 2, \quad (6)$$

where length of stability interval of the scheme is approximately equals to 2;  $k_1^i$ ,  $k_2^i$  and  $k_3^i$  are the components of vectors  $k_1$ ,  $k_2$  and the auxiliary vector  $k_3$ . Stage  $k_3$  coincides with vector  $k_1$  for next step and therefore does not lead to computational costs increasing.

The method is described with more detail in [4].

## 2.2 Three-staged Runge-Kutta method

Consider implicit three-stage method of Runge-Kutta type for solving problem (2), which has the following form:

$$\begin{aligned} y_{n+1} &= y_n + \frac{1}{6}k_1 + \frac{2}{3}k_2 + \frac{1}{6}k_3, \\ k_1 &= hf(y_n), \\ k_2 &= hf(y_n + 0.5k_1), \\ k_3 &= hf(y_n - k_1 + 2k_2). \end{aligned} \quad (7)$$

Inequality for accuracy control has the following form:

$$0.5 \| k_1 - 2k_2 + k_3 \| \leq 6\varepsilon. \quad (8)$$

Inequality for stability control in turn looks as follows:

$$v_{n,3} = 0.5 \max_{1 \leq i \leq N} (|k_1^i - 2k_2^i + k_3^i| / |k_2^i - k_1^i|) \leq 2.5, \quad (9)$$

Other methods implemented with-in ISMA environment are studied in [7].

## 3 Integration algorithms

This section is devoted to the organization of computations. The solver is designed so that sequential and parallel algorithms use the same integration method description. The main difference is in the stage preceded by calculations and in communication. In the stage preceded by calculations, the parallel solver divide the system into data portions representing the part of the equation system. The difference in communication is determined by differences in the types of runtime environments. The sequential algorithm has an access to the common memory therefore the system is solved as a whole in each step. While the parallel solver is executed in the paradigm of the shared memory, where each rank can communicate with each other only by messages. This type of parallel architecture are referred to as MIMD is chosen because it is focused on the cluster systems, which are necessitated by the dimension and the specifics of the solving problems. MPJ Express library implementing the MPI standard for Java platform is chosen for the parallel solver.

### 3.1 Sequential integration strategy

Formal description of the sequential integration algorithm is the following. Let the method (4) is used for numerical solution of problem (2) and let the approximate solution  $y_n$  is known at the moment  $t_n$  with the step  $h_n$ . Then to obtain the approximate solution  $y_{n+1}$  at the moment  $t_{n+1}$  we have the following common algorithm:

1. Calculate the approximate solution  $y_{n+1}$  at the moment  $t_{n+1}$  with the step  $h_n$  according to the performing method.
2. Calculate approximate function value  $f(y_{n+1})$ .
3. Obtain the accuracy characteristics of the integration step.
4. If the solution is accurate then go to 5, else set the integration step  $h_n$  equals to the step  $h^{ac}$  corrected by accuracy according to the performing method and go to 1.
5. Obtain the stability characteristics of the integration step.
6. If the solution is stable then go to 7, else set the integration step  $h_n$  equals to the step  $h^{st}$  corrected by stability according to the performing method and go to 1.
7. Get size of the next integration step using formula (5).
8. Perform the next integration step.

### 3.2 Parallel integration strategy

Developed parallel algorithms are based on the presented above sequential algorithms with the following differences.

For definiteness, we assume that the computer system consists of  $p$  and the size  $N$  of the problem(1) is grater than the number of processors ( $N > p$ ). Let  $k$  is a number of equations per rank. Given these assumptions let us formulate the parallel integration algorithm based on the scheme (2) on each  $i$  node  $1 \leq i \leq p$ :

1. Calculate the approximate solution  $y_{n+1}^i$ ,  $((i-1) \cdot k + 1) \leq j \leq (i \cdot k)$  at the moment  $t_{n+1}$  with the step  $h_n$  according to the performing method.
2. Send the obtained  $y_{n+1}^i$  from each rank to others.
3. Calculate in each rank the approximate function value  $f(y_{n+1}^i)$ .
4. Execute for each rank the sequential algorithm from the 3-d step of the previous section.

## 4 Switching points detection

The correctness of the hybrid model analysis along with the accuracy of calculations is determined by the accuracy of localizing the moments of the local states changes. Therefore, in addition the dynamics of the event function should be taken into account. Let the calculations performed by the numerical scheme (2). [1] contains the proof of the theorem whereby the choice of the integration step according to formula:

$$h_{n+1} = (\gamma - 1) \frac{g_n}{\frac{\partial g_n}{\partial x} \cdot \varphi_n + \frac{\partial g_n}{\partial t}}, \gamma \in (0, 1), \quad (10)$$

provides the behaviour of the event dynamics as the stable linear system which solution is approaching the surface  $g(x, t) = 0$  asymptotically. However the movement direction of the event function is not taken into account in the theorem. Let us formulate the algorithms of the step control considering dynamics and the movement direction of the event function. Let the solution  $y_n$  at the point  $t_n$  is calculated with the step  $h_n$ . In addition the integration algorithms has calculated a new integration step  $h_{n+1}^{int}$  based on the accuracy and stability requirements. Then the choice of the integration step is carried out by the following algorithm:

1. Calculate  $g_n = g(x_n, t_n)$ ,  $\partial g_n / \partial x = \partial(x_n, t_n) / \partial x$ ,  $\partial g_n / \partial t = \partial(x_n, t_n) / \partial t$ .
2. Calculate  $g'_n = (\partial g_n / \partial x) \cdot \varphi_n + \partial g_n / \partial t$ .
3. If  $g'_n < 0$ , take  $h_{n+1} = h_{n+1}^{ac}$  and go to the step 6.
4. Calculate the event step  $h_{n+1}^{ev}$  by the formula  $(\gamma - 1)g_n / g'_n$ .
5. Calculate a new step  $h_{n+1} = \min(h_{n+1}^{ev}, h_{n+1}^{int})$
6. Perform the next integration step.

The direction of the event function is determined on the 3-d step. When approaching the regime border the denominator of (3) is positive and away from the boundary  $g(x, t) = 0$  it becomes negative. Then determining the movement direction of the event function additional constraints on the integration step can be omitted if the event-driven function is removed from the regime border.

## 5 Simulation of Lotka-Volterra problem

Let us consider the use of the proposed approaches on the example of simulation of reaction-diffusion problem in two-dimensional space, which is associated with competition model of Lotka-Volterra [8].

Two kinds of variables  $c^1(x, z, t)$  and  $c^2(x, z, t)$  represent density of competing species in the habitat area  $\Omega = \{(x, z) : 0 \leq x \leq 1, 0 \leq z \leq 1\}$  and in time  $0 \leq t \leq 3$ :

$$\frac{\partial c^i}{\partial t} = d_i \left( \frac{\partial^2 c^i}{\partial x^2} + \frac{\partial^2 c^i}{\partial z^2} \right) + f^i(c^1, c^2), i = 1, 2, \quad (11)$$

where  $d_1 = 0.05, d_2 = 1.0, f^1(c^1, c^2) = c^1(b_1 - a_{12}c^2), b_1 = 1.0, a_{12} = 0.1, f^2(c^1, c^2) = c^2(-b_2 + a_{21}c^1), b_2 = 1000.0, a_{21} = 100.0$ . Boundary conditions are  $\partial c^i / \partial x = 0$  at  $x = 0, x = 1$  and  $\partial c^i / \partial z = 0$  at  $z = 0, z = 1$ . Initial conditions are  $c^1(x, z, 0) = 10 - 5 \cos(\pi x) \cos(10\pi z)$  and  $c^2(x, z, 0) = 17 + 5 \cos(10\pi x) \cos(\pi z)$ .

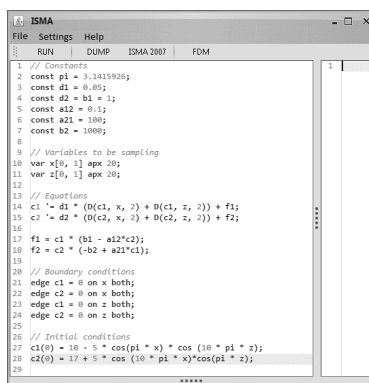
At  $t \rightarrow \infty$  solution becomes spatially homogeneous and tend to periodically solve ODE system of Lotka-Volterra. This ODE system is alternately stiff and non-stiff depending on the solution position in the phase space.

Turning to the grid of size  $J \times K$  by  $x$  and  $z$  respectively we obtain  $\Delta x = 1/(J - 1)$  and  $\Delta z = 1/(K - 1)$  are grid steps by  $x$  and  $z$  coordinates,  $c_{jk}^i$  is approximation of  $c^i(x_j, z_k, t)$  where  $x_j = (j - 1)\Delta x, z_k = (k - 1)\Delta z, 1 \leq j \leq J, 1 \leq k \leq K$ . Thus we obtain differential equations system of  $N = 2JK$  dimension:

$$c_{jk}^i = \frac{d_i}{\Delta x^2} (c_{j+1,k}^i - 2c_{jk}^i + c_{j-1,k}^i) + \frac{d_i}{\Delta z^2} (c_{j,k+1}^i - 2c_{jk}^i + c_{j,k-1}^i) + f_{jk}^i, \quad (12)$$

where  $1 \leq i \leq 2, 1 \leq j \leq J, 1 \leq k \leq K, f_{jk}^i = f^i(c_{jk}^1, c_{jk}^2)$ . Boundary conditions on the grid are the following:  $c_{0,k}^i = c_{2,k}^i, c_{J+1,k}^i = c_{J-1,k}^i$  for  $1 \leq k \leq K$  and  $c_{j,0}^i = c_{j,2}^i, c_{j,K+1}^i = c_{j,K-1}^i$  for  $1 \leq j \leq J$ .

The problem model created in ISMA\_2015 is presented on the Fig. 1. The model description is compact and is maximally close to the original mathematical description.



```

1 // Constants
2 const pi = 3.1415926;
3 const d1 = 0.05;
4 const d2 = b1 = 1;
5 const a12 = 0.1;
6 const a21 = 100;
7 const b2 = 1000;
8
9 // Variables to be sampling
10 var x[0, 1] spx 20;
11 var z[0, 1] spz 20;
12
13 // Equations
14 c1 = d1 * (D(c1, x, 2) + D(c1, z, 2)) + f1;
15 c2 = d2 * (D(c2, x, 2) + D(c2, z, 2)) + f2;
16
17 f1 = c1 * (b1 - a12*c2);
18 f2 = c2 * (-b2 + a21*c1);
19
20 // Boundary conditions
21 edge c1 = 0 on x both;
22 edge c2 = 0 on x both;
23 edge c1 = 0 on z both;
24 edge c2 = 0 on z both;
25
26 // Initial conditions
27 c1(0) = 10 - 5 * cos(pi * x) * cos (10 * pi * z);
28 c2(0) = 17 + 5 * cos (10 * pi * x)*cos(pi * z);
29
*****

```

Fig. 1. The model of the reaction-diffusion problem in ISMA\_2015.

Simulation settings are shown on the Fig. 2 where initial conditions such as simulation interval and initial step size are specified.

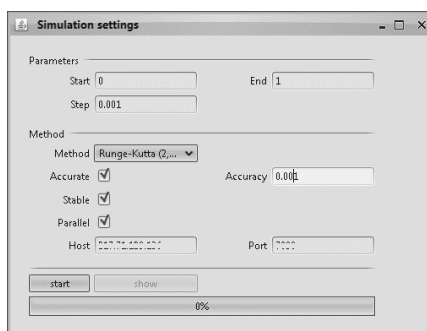


Fig. 2. Simulation settings window in ISMA\_2015.

Two staged Runge-Kutta method with accuracy and stability control enabled is chosen for the model solving. The host and the port of the remote simulation server are specified.

Simulation results are presented on the Fig. 3.

The comparative analysis results of implemented algorithms in sequential and parallel version is shown on the Fig. 4.

Such a significant increase of the computational costs especially for sequential algorithms is related to the costs of construction and inversion of Jacoby matrix of increasing dimension. Also the higher system dimension the advantage of the parallel algorithms becomes even more clearly.

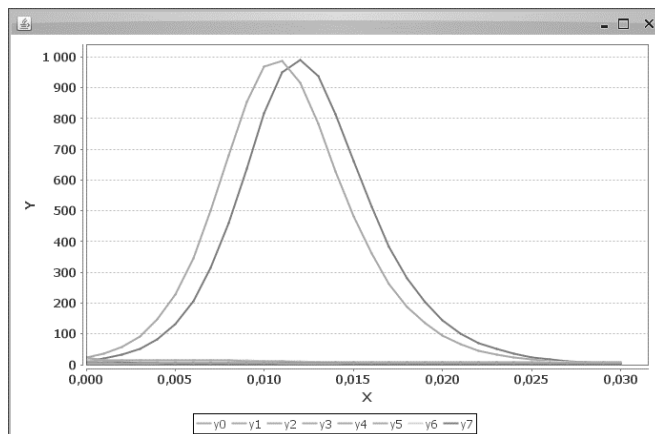


Fig. 3. Simulation results of the reaction-diffusion problem obtained in ISMA\_2015.

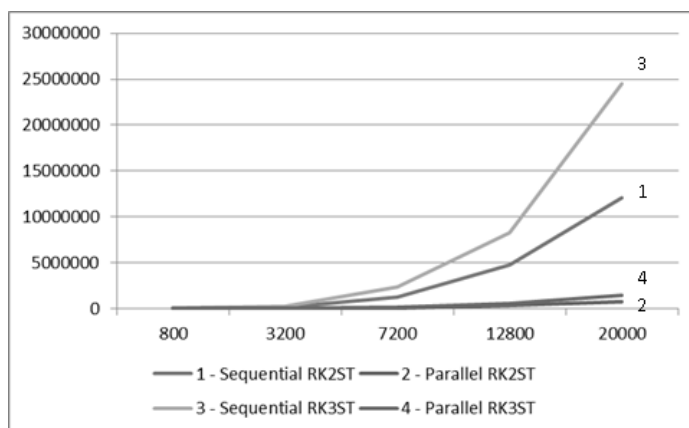


Fig. 4. Comparative analysis of RK22 and RK3.

## 6 Conclusion

This paper presents a mathematical and program software for analysis of the modal behaviour of hybrid systems. The behaviour can be described by a different classes of differential equations such as commonly used ordinary differential equations, differential-algebraic equations of more specific for hybrid systems partial differential equations. The methods suitable for the hybrid model analysis are considered. The algorithms of the control of the integration step are studied. Along with the accuracy and the stability control of the numerical scheme the control of the integration step based on the event detection mechanism is proposed. The evaluations of the calculation costs confirmed the viability of the proposed approaches. Simulation results obtained in ISMA\_2015 simulation environment are presented on the example of generated reaction-diffusion problems based on Lotka-Volterra model.

**Acknowledgments.** This work was supported by grant 14-01-00047-P° from the Russian Foundation for Basic Research. Yu.V. Shornikov is with the Design Technological Institute of Digital Techniques Siberian Branch of Russian Academy of Science, Novosibirsk, Russia (e-mail: shornikov@inbox.ru). M.S. Myssak(Nasyrova), D.N. Dostovalov are with the Department of Automated Control Systems, Novosibirsk State Technical University, Novosibirsk, Russia (e-mail: maria\_myssak@mail.ru, dostovalov.dmitr@mail.ru).

## References

1. Novikov, E.A, Shornikov, Yu.V.: Computer Simulation of Stiff Hybrid Systems: Monograph. Publishing House of NSTU, Novosibirsk (2012).
2. Esposito, J., Kumar, V., Pappas, G.J.: Accurate Event Detection for Simulating Hybrid Systems. In: Hybrid Systems: Computation and Control (HSCC), LNCS, vol. 2034, pp. 204–217. Springer, Verlag (1998).
3. Novikov, E.A., Shornikov, Yu.V.: Numerical Simulation of Hybrid Systems by Runge-Kutta Method of Second Accuracy Order. In: Computing Technologies, vol. 13, no. 2, pp. 98–104. (2008).
4. Novikov, E.A.: Explicit Methods for Stiff Systems. Nauka, Novosibirsk (1997).
5. Novikov, E.A.: The Global Error of One-Step Solution Methods for Stiff Problems. In: Russian Mathematics. In: Vuz, vol. 55, no. 6, pp. 68–75 (2011).
6. Novikov, E.A., Vashchenko, G. V.: Parallel Explicit Runge-Kutta Method 2nd Order: Accuracy and Stability Control. In: Int. J. of Applied and Fundamental Research. Physical and Mathematical Science, no. 1, pp. 101БТ<sup>к</sup>-102 (2011).
7. Shornikov, Yu.V., Myssak(Nasyrova), M.S., Dostovalov, D.N.: Computer Simulation of Hybrid Systems by ISMA Instrumental Facilities. In: Proc. of the 2014 International Conference on Mathematical Models and Methods in Applied Sciences (MMMAS'14), pp. 257–262, Saint Petersburg, (2014).
8. Brown, P.N., Hindmarsh A.C., Matrix Free Methods in the Solution of Stiff systems of ODEs. Lawrence Livermore National Laboratory, p. 38 (1983).

# Numerical Solution of Three-Dimensional Diffraction Problems Using Mosaic-Skeleton Method

Sergei Smagin, Aleksei Kashirin, and Mariia Taltykina

Computing Center FEB RAS. Kim Yu Chen St., 65, 680000 Khabarovsk, Russia

smagin@as.khb.ru, elomer@mail.ru, %taltykina@yandex.ru

<http://www.ccfbras.ru/en>

**Abstract.** Numerical solution of three-dimensional diffraction problems of acoustic waves is considered. The problems are formulated as weakly singular integral equations of the 1st kind with one unknown function. Considered integral equations are approximated by systems of linear algebraic equations with dense matrices. The systems are numerically solved by an iterative method. The most difficult part of the iterative method is a numerical matrix-vector multiplication. The mosaic-skeleton method is used for decreasing computational complexity of the matrix-vector multiplication.

**Keywords:** diffraction problem, Helmholtz equation, boundary integral equation, mosaic-skeleton method.

## 1 Introduction

The study of processes of propagation of stationary acoustic waves in media with three-dimensional inclusions is applied in various areas of science and engineering. It leads to statement of problems of mathematical physics which are customarily called diffraction or scattering problems. Exact analytical solutions of such problems can be constructed only in exceptional cases; therefore, the basic direction of their research is direct computer modeling.

Using a computer assumes preliminary construction of a discrete analogue of the considered problem which can be carried out in various ways. The discrete analogues based on differential statements, leading to finite difference or projective-grid schemes, are not effective. This is explained by the fact that the solutions of the problems are searched in unlimited domains. They depend on three space variables, slowly decrease at infinity, can be fast oscillating functions and must meet radiation conditions at infinity.

From the computing point of view the approach in which such problems are formulated in the form of one weakly singular Fredholm integral equation of the 1st or 2nd kind with one unknown function (density) is more advantageous [1]–[3]. In the given work integral equations of the 1st kind were used for numerical solution of the scalar problems of diffraction. It's important to mention that each problem allows various equivalent statements.

A special method of averaging integral operators with weak singularities in kernels is applied to numerical solution of the received equations. This method has been used earlier for solution of integral equations of boundary value problems of acoustics. The discretization of boundary integral equations leads to systems of linear algebraic equations (SLAEs) with dense matrices. The computational complexity of their solution by direct methods is  $O(M^3)$ , where  $M$  is the order of the SLAE. However, the properties of the resulting matrices are such that approximate solutions can be sought using the generalized minimal residual method (GMRES) [4], which reduces this complexity to  $O(M^2)$ . After approximate solutions of the integral equations have been found, the solutions of the original problems can easily be reconstructed at any spatial point using the integral representations.

The most difficult part of GMRES is a matrix-vector multiplication, whose computational complexity can be reduced using a mosaic-skeleton method [5]–[7]. The complexity of the method is  $o(M^2)$ ,  $M \rightarrow \infty$ . The main idea of the method states that rather large blocks in very large matrices coming from integral formulations can be approximated accurately by a sum of just few rank-one matrices. These sums of matrices can be built and multiplied on a vector by a linear number of arithmetic operations. On the whole, the method allows to decrease the memory demands.

## 2 Integral equations of scalar problem of diffraction

*Problem 1.* In bounded domain  $\Omega_i$  of three-dimensional Euclidean space  $\mathbb{R}^3$  and in unlimited domain  $\Omega_e = \mathbb{R}^3 \setminus \bar{\Omega}_i$ , divided by closed surface  $\Gamma \in C^{r+\beta}$ ,  $r + \beta > 1$ , find complex-valued functions  $u_{i(e)} \in H^1(\Omega_{i(e)}, \Delta)$ , satisfying integral identities

$$\int_{\Omega_{i(e)}} \nabla u_{i(e)} \nabla v_{i(e)}^* dx - k_{i(e)}^2 \int_{\Omega_{i(e)}} u_{i(e)} v_{i(e)}^* dx = 0 \quad \forall v_{i(e)} \in H_0^1(\Omega_{i(e)}), \quad (1)$$

interface conditions on the boundary of division of media from  $\Omega_i$  and  $\Omega_e$

$$\langle u_i^- - u_e^+, \mu \rangle_\Gamma = \langle f_0, \mu \rangle_\Gamma \quad \forall \mu \in H^{-1/2}(\Gamma), \quad (2)$$

$$\langle \eta, p_i N^- u_i - p_e N^+ u_e \rangle_\Gamma = \langle \eta, p_e f_1 \rangle_\Gamma \quad \forall \eta \in H^{1/2}(\Gamma),$$

as well as radiation condition at infinity for  $u_e$

$$\partial u_e / \partial |x| - ik_e u_e = o(|x|^{-1}), \quad |x| \rightarrow \infty, \quad (3)$$

if  $f_0 \in H^{1/2}(\Gamma)$  and  $f_1 \in H^{-1/2}(\Gamma)$  are given functions on  $\Gamma$ .

Here and after  $\langle \cdot, \cdot \rangle_\Gamma$  is a duality relation on  $H^{1/2}(\Gamma) \times H^{-1/2}(\Gamma)$ , generalizing inner product in  $H^0(\Gamma)$ ,  $v^*$  is a complex conjugate function to  $v$ ,  $N^- : H^1(\Omega_i, \Delta) \rightarrow H^{-1/2}(\Gamma)$ ,  $N^+ : H^1(\Omega_e, \Delta) \rightarrow H^{-1/2}(\Gamma)$  are normal derivative operators [8],  $\gamma^- : H^1(\Omega_i) \rightarrow H^{1/2}(\Gamma)$ ,  $\gamma^+ : H^1(\Omega_e) \rightarrow H^{1/2}(\Gamma)$  are trace operators,  $\omega$  is the circular oscillation frequency and

$$k_{i(e)}^2 = \omega (\omega + i\gamma_{i(e)}) / c_{i(e)}^2, \quad \text{Im}(k_{i(e)}) \geq 0, \quad p_{i(e)} = c_{i(e)}^2 k_{i(e)}^{-2} \rho_{i(e)}^{-1}.$$

*Remark 1.* If  $\text{Im}(k_e) = 0$ , then  $u_e \in H_{loc}^1(\Omega_e, \Delta)$ .

**Theorem 1.** *Problem 1 has at most one solution.*

You can find the proof of the theorem 1 in [1],[2].

We introduce the following notation:

$$(A_{i(e)} q)(x) \equiv \langle G_{i(e)}(x, \cdot), q \rangle_\Gamma, \quad (B_{i(e)} q)(x) \equiv \langle N_x G_{i(e)}(x, \cdot), q \rangle_\Gamma, \quad (4)$$

$$(B_{i(e)}^* q)(x) \equiv \langle N_{(\cdot)} G_{i(e)}(x, \cdot), q \rangle_\Gamma, \quad G_{i(e)}(x, y) = \exp(ik_{i(e)} |x - y|) / (4\pi |x - y|).$$

A solution of problem 1 is sought in the form of potentials

$$u_e(x) = (A_e q)(x), \quad x \in \Omega_e, \quad (5)$$



$$u_i(x) = (p_{ei}A_i(N^+u_e + f_1) - B_i^*(u_e^+ + f_0))(x), \quad x \in \Omega_i,$$

where  $q \in H^{-1/2}(\Gamma)$  is the unknown density,  $f_0 \in H^{1/2}(\Gamma)$ ,  $f_1 \in H^{-1/2}(\Gamma)$ ,  $p_{ei} = p_e/p_i$ . Here, the kernels of the integral operators are fundamental solutions of Helmholtz equations and their normal derivatives. Therefore,  $u_{i(e)}$  satisfy identities (1) and radiation condition (3) for  $u_e$ . If they satisfy the first transmission condition in (2), they automatically satisfy the second transmission condition. Substituting potentials (5) into transmission conditions (2), we obtain a weakly singular Fredholm integral equation of the first kind for determining the unknown density  $q$ :

$$\langle Cq, \mu \rangle_\Gamma = \langle f_2, \mu \rangle_\Gamma \quad \forall \mu \in H^{-1/2}(\Gamma), \quad (6)$$

where

$$C = (0.5 + B_i^*)A_e + p_{ei}A_i(0.5 - B_e), \quad f_2 = -(0.5 + B_i^*)f_0 + p_{ei}A_if_1.$$

Problem 1 admits another equivalent formulation in the form of a Fredholm integral equation of the first kind with a weak singularity in the kernel. Its solution is sought in the form

$$u_i(x) = (A_iq)(x), \quad x \in \Omega_i, \quad (7)$$

$$u_e(x) = (A_e(f_1 - p_{ie}N^-u_i) - B_e^*(f_0 - u_i^-))(x), \quad x \in \Omega_e,$$

where  $q \in H^{-1/2}(\Gamma)$  is the unknown density,  $f_0 \in H^{1/2}(\Gamma)$ ,  $f_1 \in H^{-1/2}(\Gamma)$ ,  $p_{ie} = p_i/p_e$ . In this case, Problem 1 is reduced to the integral equation

$$\langle Dq, \mu \rangle_\Gamma = \langle f_0, \mu \rangle_\Gamma \quad \forall \mu \in H^{-1/2}(\Gamma), \quad (8)$$

$$D = (0.5 - B_e^*)A_i + p_{ie}A_e(0.5 + B_i).$$

**Theorem 2.** Suppose that  $f_0 \in H^{1/2}(\Gamma)$ ,  $f_1 \in H^{-1/2}(\Gamma)$ ,  $\gamma_e > 0$  or  $\omega$  is not an eigenfrequency of the problem

$$\Delta u + k_e^2 u = 0, \quad x \in \Omega_i, \quad u^- = 0. \quad (9)$$

Then Eqs. (6) and (8) are correctly solvable in the class of densities  $q \in H^{-1/2}(\Gamma)$  and the solution of Problem 1 is given by formulas (5) and (7).

The proof of the theorem 2 is in [1].

*Remark 2.* If we are more interested in the wave field in the domain  $\Omega_e$ , it is preferable to use Eq. (6), in which case the reflected field can be computed using a simpler formula. For a similar reason, if we are interested in the transmitted wave field in  $\Omega_i$ , it is preferable to use Eq. (8).

### 3 Numerical method

The main idea of the numerical method is that an unknown density is in the form of a linear combination of smooth finite functions generating partition of a unit on the boundary of inclusion. Such approach does not demand preliminary triangulation of a surface and is equally simply realized both on regular, and on irregular grids. During discretization of equations superficial integrals are approximately substituted by expressions containing integrals on  $\mathbb{R}^3$ , which are calculated then analytically, which allows us to find factors of systems of linear algebraic equations, approximating integral equations simply enough. The scheme for its implementation can be briefly described as follows.

The surface  $\Gamma$  is covered with a system  $\{\Gamma_m\}_{m=1}^M$  of neighborhoods of nodes  $x'_m \in \Gamma$ , lying inside the spheres of radii  $h_m$  centered at  $x'_m$ . The corresponding partition of unity is denoted by  $\varphi_m$ . Then

$$\varphi_m(x) = \varphi'_m(x) \left( \sum_{k=1}^M \varphi'_k(x) \right)^{-1}, \quad \varphi'_m(x) = \begin{cases} (1 - r_m^2/h_m^2)^3, & r_m < h_m, \\ 0, & r_m \geq h_m, \end{cases}$$

where  $r_m = |x - x'_m|$ .

Approximate solutions of the integral equations (6) and (8) are sought on the grid  $\{x_m\}$ :

$$x_m = \frac{1}{\bar{\varphi}_m} \int_{\Gamma} x \varphi_m d\Gamma, \quad \bar{\varphi}_m = \int_{\Gamma} \varphi_m d\Gamma,$$

where the nodes are the centers of gravity of the functions  $\varphi_m$ . Assume that, for all  $m = 1, 2, \dots, M$

$$0 < h' \leq |x_m - x_n|, \quad m \neq n, \quad n = 1, 2, \dots, M, \\ h' \leq \sigma_m \leq h_m \leq h, \quad h/h' \leq q_0 < \infty.$$

Here,  $r_{mn} = |x_m - x_n|$ ,  $h, h'$  are positive numbers depending on  $M$ , and  $q_0$  is independent of  $M$ .  $\sigma_m^2 = 0.5\bar{\varphi}_m$ . Instead of the unknown function  $q$  on  $\Gamma$ , we consider a generalized function  $q\delta_{\Gamma}$  defined according to the rule

$$(q\delta_{\Gamma}, \eta)_{\mathbb{R}^3} = \langle q, \eta \rangle_{\Gamma} \quad \forall \eta \in H^1(\mathbb{R}^3).$$

This function is approximated by the expression

$$q(x) \delta_{\Gamma}(x) \approx \sum_{n=1}^M q_n \bar{\varphi}_n \psi_n(x), \quad \psi_n(x) = (\pi\sigma_n^2)^{-3/2} \exp(-(x - x_n)^2/\sigma_n^2), \quad x \in \mathbb{R}^3,$$

where  $q_n$  are the unknown coefficients.

The following is true for  $\eta$  and  $q \in H^1(\Gamma)$

$$\langle q, \eta \rangle_{\Gamma} = \left( \sum_{n=1}^M q_n \bar{\varphi}_n \psi_n, \eta \right)_{\mathbb{R}^3} + O(h^2).$$

Approximating the single layer potential density by a volume density, we derive simple formulas for the approximation of the integral operator  $A_{i(e)}$  in (4). The integral operators from (4) on  $\Gamma$  are approximated by the expressions (see [1])

$$\langle A_{i(e)} q, \varphi_m \rangle_{\Gamma} \approx \sum_{n=1}^M A_{i(e)}^{mn} q_n, \quad m = 1, 2, \dots, M, \tag{10}$$

$$A_{i(e)}^{mn} \equiv A_{mn}(k_{i(e)}),$$

$$A_{mn}(k) = \frac{\bar{\varphi}_m \bar{\varphi}_n}{8\pi r_{mn}} \exp(\mu_{mn}^2 - \gamma_{mn}^2) (w(z_{mn}^-) - w(z_{mn}^+)), \quad n \neq m,$$

$$A_{mm}(k) = \frac{\bar{\varphi}_m^2}{4\pi} \exp(\mu_{mm}^2) \left( ikw(\mu_{mm}) + \frac{\sqrt{2\pi}}{\bar{\varphi}_m} \left( \frac{\bar{\varphi}_m}{\pi\sigma_m} + 2\sigma_m - \frac{k^2\sigma_m^3}{3} \right) \right),$$

$$\sigma_{mn}^2 = \sigma_m^2 + \sigma_n^2, \quad \mu_{mn} = 0.5k\sigma_{mn}, \quad z_{mn}^{\pm} = \mu_{mn} \pm i\gamma_{mn},$$

$$\gamma_{mn} = r_{mn}/\sigma_{mn}, \quad i^2 = -1,$$

$$w(z) = -\frac{2i}{\sqrt{\pi}} \exp(-z^2) \int_z^{\infty} \exp(t^2) dt,$$

$$\langle aq + B_{i(e)}q, \varphi_m \rangle_{\Gamma} \approx \sum_{n=1}^M B_{i(e)}^{mn} q_n, \quad m = 1, 2, \dots, M, \quad a = \pm 0.5, \quad (11)$$

$$\langle aq + B_{i(e)}^*q, \varphi_m \rangle_{\Gamma} \approx \sum_{n=1}^M B_{i(e)}^{nm} q_n, \quad m = 1, 2, \dots, M, \quad (12)$$

$$B_{i(e)}^{mn} = \frac{\eta_{mn}}{4\pi r_{mn}^2} \exp(ik_{i(e)}r_{mn}) (ik_{i(e)}r_{mn} - 1) \bar{\varphi}_m \bar{\varphi}_n, \quad n \neq m,$$

$$B_{i(e)}^{mm} = (-|a| + a + \text{Gs}_m) \bar{\varphi}_m, \quad \eta_{mn} = \sum_{l=1}^3 n_{ml} \frac{x_{ml} - x_{nl}}{r_{mn}}, \quad \text{Gs}_m = -\sum_{n \neq m}^M \frac{\eta_{nm} \bar{\varphi}_n}{4\pi r_{mn}^2},$$

where  $n_{ml}$  are the components of the outward unit normal vector to the surface  $\Gamma$  at the point  $x_m$ . The operators on the left hand sides of Eqs. (6) and (8) are approximated by compositions of operators (10)–(12):

$$\langle Cq, \varphi_m \rangle_{\Gamma} \approx \sum_{n=1}^M C_{ie}^{nm} q_n, \quad \langle Dq, \varphi_m \rangle_{\Gamma} \approx -\sum_{n=1}^M C_{ei}^{nm} q_n, \quad m = 1, 2, \dots, M, \quad (13)$$

$$C_{ie}^{nm} = B_i^{nm} A_e^{mn} - p_{ei} A_i^{mn} B_e^{mn},$$

and the right hand sides of Eqs. (6) and (8) are approximated by the formulas

$$\langle f_2, \varphi_m \rangle_{\Gamma} \approx \sum_{n=1}^M (p_{ei} A_i^{mn} f_{1n} - B_i^{nm} f_{0n}), \quad \langle f_0, \varphi_m \rangle_{\Gamma} = \bar{\varphi}_m f_{0m}, \quad (14)$$

$$f_{lm} = \langle f_l, \varphi_m / \bar{\varphi}_m \rangle_{\Gamma}, \quad l = 0, 1, \quad m = 1, 2, \dots, M.$$

Solving the corresponding SLAEs, we find approximate densities of the integral equations at discretization points. Next, the desired solution of the diffraction problem can be easily and accurately calculated at any point of the space by applying the integral representations.

#### 4 Mosaic-skeleton method

The mosaic-skeleton method consists of three stages. The first stage is building a tree of clusters [5]–[7]. The root of this tree is the cluster containing all the nodes of the grid. On each step the cluster is split into several subclusters. It takes place until the tree of clusters reaches the maximal level.

On the second stage the list of blocks is prepared. Any of the blocks of the matrix is associated with the pair of clusters in the tree of clusters. If the nodes of the pair are apart from each other, the block belongs to the "far" area. Otherwise, it determines the "near" area. The constructed blocks without crossings cover all matrix.

On the last stage the blocks of the "far" area can be approximated by one-rank matrices. These one rank matrices are named skeletons. The easiest way to find the skeletons is applying

of an incomplete cross approximation. The incomplete cross approximation is described in detail in works [6],[7]. The approximation is held before the GMRES and the constructed blocks are saved in the random access memory.

On the GMRES stage multiplication depends on the area of the block. If the block belongs to the "far" area, it is multiplied on the vector by a special function, if the block is from the "near" area, the matrix-vector multiplication from Intel MKL is used.

## 5 Numerical results

The numerical method is verified and validated using test problems of diffraction of a plane wave in domain bounded by a unit sphere and a triaxial ellipsoid. The program of numerical solution of the initial problems is tested on the computer cluster at the Computing Center of the Far Eastern Branch of the Russian Academy of Sciences. The host has 24 CPUs, each CPU speed is 2400 MHz, total memory 94 GB. The parallel versions have been created for the GMRES and the mosaic-skeleton method using OpenMP.

*Example 1.* Consider diffraction problems of a plane acoustic wave on the unit ball centered at the origin and the parameters of the media which are determined as I)  $k_i = 8, \rho_i = 3, k_e = 5.5, \rho_e = 1$ ; II)  $k_i = 15.5, \rho_i = 5, k_e = 9, \rho_e = 4$ ; III)  $k_i = 21, \rho_i = 7, k_e = 30.5, \rho_e = 9.5$ . The complex range of the initial wave field is  $u_0(x) = \exp(ik_e x_3)$ , and  $f_0 = u_0^+, f_1 = N^+ u_0$ .

Number  $M$  of discretization points varied from 1000 to 64000. The resulting SLAEs were solved numerically by applying the GMRES up to  $10^{-7}$ . The error of the incomplete cross approximation is  $\varepsilon = 10^{-5}$ .

The exact densities for Ex. 1 can be found using the formulas

$$q = - \sum_{m=0}^{\infty} \frac{(2m+1) i^{m+1} a_m P_m(\cos \theta)}{k_e b_m j_m(k_e)},$$

$$q = - \sum_{m=0}^{\infty} \frac{(2m+1) i^{m+2} p_e P_m(\cos \theta)}{k_i k_e b_m h_m(k_i)}$$

for the integral equations (6) and (8) respectively.

Figure 1 shows the errors of the solutions of the integral equations (6) (dotted lines) and (8) (full lines) using the mosaic-skeleton method. Here and below, the squares are denoted by the first order of the parameters of the media, the circles are denoted by the second order and the triangles are denoted by the third order. It can be seen that, for sufficiently large  $M$ , the order of the error is at most  $h^2 \sim M^{-1}$ .

The exact solutions of the initial problems for Ex. 1 have the form [9]

$$u_e(r, \theta) = \sum_{m=0}^{\infty} \frac{(2m+1) i^m a_m}{b_m} h_m(k_e r) P_m(\cos \theta), \quad r \geq 1,$$

$$u_i(r, \theta) = \sum_{m=0}^{\infty} \frac{(2m+1) i^{m+1} p_e}{k_e b_m} j_m(k_i r) P_m(\cos \theta), \quad r \leq 1.$$

Where  $r = |x|$ ,  $\cos \theta = x_3/r$ ,

$$a_m = (p_i j_m(k_e r) j'_m(k_i r) - p_e j_m(k_i r) j'_m(k_e r)) \Big|_{r=1},$$

$$b_m = (p_e j_m(k_i r) h'_m(k_e r) - p_i h_m(k_e r) j'_m(k_i r)) \Big|_{r=1},$$

$$j_m(kr) = \sqrt{\frac{\pi}{2kr}} J_{m+1/2}(kr), \quad h_m(kr) = \sqrt{\frac{\pi}{2kr}} H_{m+1/2}^{(1)}(kr),$$

$J_{m+1/2}, H_{m+1/2}^{(1)}$  are the Bessel and Hankel functions of the first kind of the  $(m + 1/2)$ th order, respectively;  $P_m$  are the Legendre polynomials of order  $m$ .

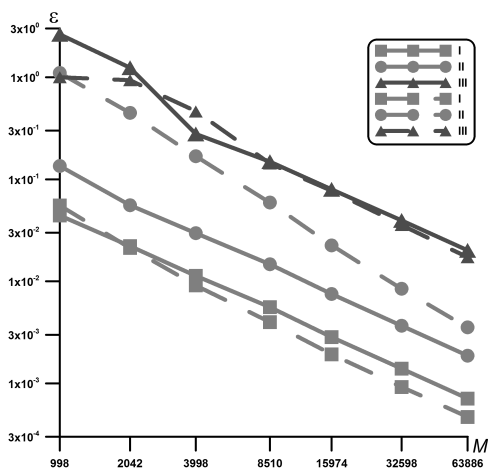


Fig. 1. Errors of the solutions of the integral equations (6) (dotted lines) and (8) (full lines) for Example 1

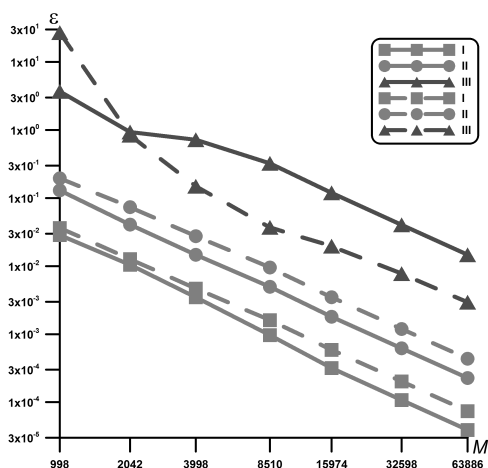


Fig. 2. Errors of solution  $u_i$  (full lines) and  $u_e$  (dotted lines) using the integral equation (6) for Example 1

Figures 2 and 3 depict the relative errors of the solutions of problem 1 founded using the integral equations (6) and (8) respectively. The mosaic-skeleton method was used on the stage of the GMRES. The full lines show the errors of  $u_i$  and the dotted lines show the errors of  $u_e$ . Again, it can be seen that, for sufficiently large  $M$ , the order of the error is at most  $h^2 \sim M^{-1}$ .

On figure 4 you can see the time of solution of the SLAE with the mosaic-skeleton method (dotted lines) and without it (full lines) for the equation (8). The numerical experiments have

shown that the mosaic-skeleton and the GMRES allow to calculate required solutions in a short time. For example, the GMRES with the mosaic method works up to 55 times faster than without it. The results for the equation (6) are similar to ones for the equation (8).

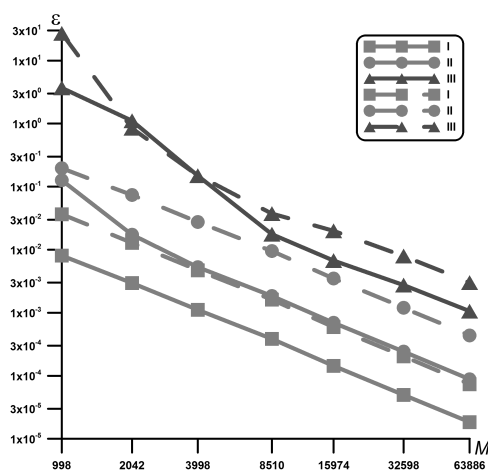
Figure 5 displays how the time of solution of the SLAE depends on the number of the cores of the computer cluster. The time is reduced 1.5 times when the number of cores doubles. The problem can be solved faster using 16 cores.

*Example 2.* Consider diffraction problem of a plane acoustic wave on the inclusion which boundary is the triaxial ellipsoid with semiaxes (0.75, 1, 0.5) centered at the origin and the parameters of the media as in Example 1. The complex range of the initial wave field is the same as in Example 1.

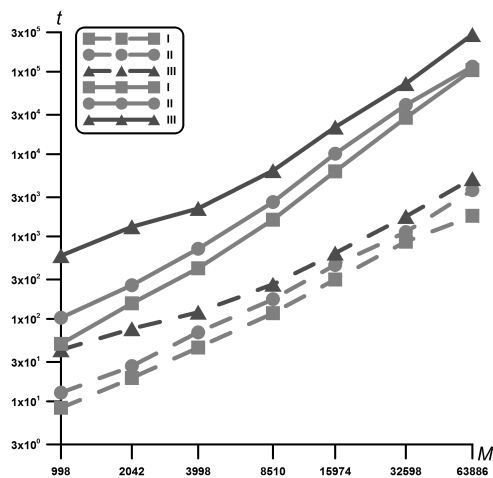
Figure 6 shows the relative errors of the solutions of problem 1 for the integral equations (6) (dotted lines) and (8) (full lines) using the mosaic-skeleton method. The solutions on different grids are compared with the solutions on the number 64139 of discretization points because the exact analytical solutions are unknown. You can see for sufficiently large  $M$ , the order of the error is at most  $h^2 \sim M^{-1}$ .

The numerical results have proved that using the mosaic-skeleton method decreased memory and time expenses in case of Example 2. The speedup of using the mosaic-skeleton method is more than 50 times.

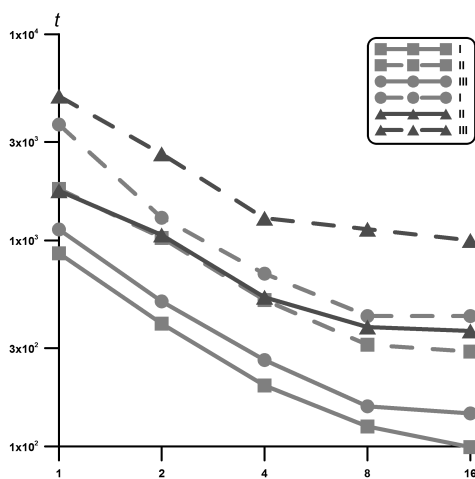
Thus, we recommend using this method for numerical solution of other problems which can be reduced to boundary integral equations.



**Fig. 3.** Errors of solution  $u_i$  (full lines) and  $u_e$  (dotted lines) using the integral equation (8) for Example 1



**Fig. 4.** Time of solution of SLAE using the mosaic-skeleton method (dotted lines) and without it (full lines) for equation (8) for Example 1



**Fig. 5.** Time of solution of SLAE. The orders of SLAEs are 32598 (full lines) and 63886 (dotted lines) for Example 1

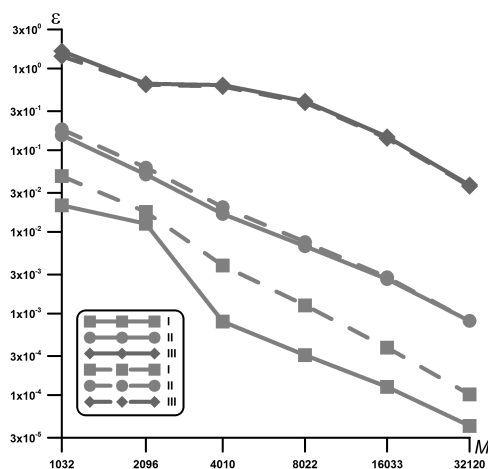


Fig. 6. Errors of solution  $u_i$  (full lines) and  $u_e$  (dotted lines) using the integral equation (6) for Example 2

## References

1. Kashirin, A.A.: Research and Numerical Solution of Integral Equations of Three-dimensional Stationary Problems of Diffraction of Acoustic Waves: Thesis ... of the candidate of physical and mathematical sciences. (in Russian) Khabarovsk (2006).
2. Kashirin, A.A., Smagin, S.I.: Generalized Solutions of the Integral Equations of a Scalar Diffraction Problem. *Differential Equations*, vol. 42, No. 1, pp. 88–100. (2006)
3. Kashirin, A.A., Smagin, S.I.: Numerical solution of integral equations of a scalar diffraction problem. *Doklady Akademii Nauk*, vol. 458, No. 2, pp. 141–144. (2014)
4. Saad, Y., Schultz, M.: GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, vol. 7, No. 3, pp. 856–869. (1986)
5. Goreinov, S.A.: Mosaic-skeleton approximation of matrices generated by asymptotically smooth and oscillatory kernels. In: Tyrtshnikov, E.E.(ed.) *Matrix Methods and Algorithms*, INM RAS. pp. 42–76. (1999)
6. Tyrtshnikov, E.E.: Incomplete cross approximations in the mosaic-skeleton method. *Computing*, vol. 64, No. 4, pp. 367–380 (2000)
7. Savostianov, D.V.: Fast multilinear approximation of matrices and integral equations: Thesis ... of the candidate of physical and mathematical sciences. (in Russian) Moscow (2006).
8. McLean, W.: *Strongly elliptic systems and boundary integral equations*. Cambridge: Cambridge Univ. Press. (2000)
9. Tikhonov, A.N., Samarskii, A.A.: *Equations of Mathematical Physics*. Moscow: Nauka. (1999)



# Seismic Field Simulation on High-Performance Computers in the Problem of Studying the Consequences of Underground Nuclear Tests

Alexander Yakimenko<sup>1,2</sup>, Dmitriy Karavaev<sup>2</sup>, and Andrey Belyashov<sup>3</sup>

<sup>1</sup>Novosibirsk State Technical University,

Prospekt K. Marksa, 20, 630073 Novosibirsk, Russia

<sup>2</sup>Institute of Computational Mathematics and Mathematical Geophysics SB RAS,  
prospekt Akademika Lavrentjeva, 6, 630090 Novosibirsk, Russia

<sup>3</sup>Institute of Geophysical Research, Ministry of Energy,  
Meridian site, 071100 Kurchatov, Kazakhstan

`al--le@yandex.ru, kda@opg.sbcc.ru, abelyashov@igr.kz`

`http://www.sbcc.ru, http://www.nstu.ru`

**Abstract.** In this paper we present the results of the development of software for the numerical simulation of a seismic field. We contrasted the work of a parallel algorithm on a high-performance system with different types of computing devices: a CPU, a GPU. The developed software offers a solution to geophysical problems in the process of studying the effects of underground nuclear tests: the development of the structure of geophysical models of cavernous areas and the study of the structure and properties of a wave field in mathematical modeling. The experimental studies of the vibration of the Earth sounding are an important step in solving on-site inspection tasks in the problem of monitoring underground nuclear tests. Carrying out numerical simulations with the use of the software for clusters gives an opportunity to select the characteristic and distinctive properties of objects under study and enables to extract informative wave groups and to determine their arrival times in the border areas of underground nuclear tests. The possibility to use different numbers of recording geophones in calculations may allow the determination of the minimum amount of geophones required for the necessary resolution of seismograms. The obtained results form the basis for recommendations to determine the areas of an underground nuclear explosion. All numerical calculations were made using the developed software on the NKS-30T+GPU cluster of the Siberian Supercomputer Center of the Russian Academy of Sciences. This work was partially supported by grants of RFBR 14-07-00832, 14-05-00867, 15-07-06821, 15-31-20150 and MES RK 1760/GF4.

**Keywords:** difference scheme, simulation, wave field, underground nuclear explosion, CUDA.

## 1 Introduction

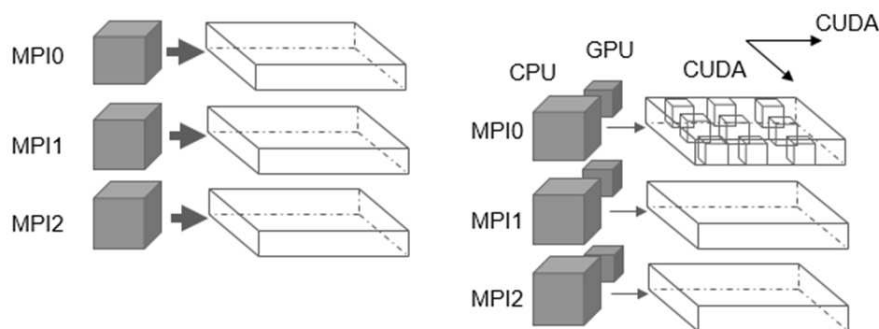
Carrying out field geophysical experiments is quite expensive. At the same time it is necessary to select carefully the location of a monitoring system to do such experiments for getting targeted effects on experimental data. Also, the experimental data themselves are a complex object of study. Calculations in the laboratory will allow to solve some of these problems. Mathematical modeling can help understand the properties of a wave field, create the geophysical model of an object, investigate the structure of a seismic field, specify the location of observation systems [1]. Solving the inverse problem of geophysics and the determination of the structure and parameters of a medium is an issue of high importance. One of the ways to work it out is to conduct computational experiments to resolve a direct problem for the different geometries and values of the elastic parameters [2]. Thus varying these parameters and comparing the simulation results with experimental data it becomes possible to achieve some compliance [3]. One popular way to carry out the simulation is to use a difference method. The use of explicit difference schemes is associated with mesh models. Reporting requirements for models are changing. It is interesting to calculate large-scale, detailed and large-scale models. Such calculations are related

to processing large volume of data. When using the difference schemes high requirements for computing resources are put forward. Therefore multicore computing systems are applied for forward modeling. Typically, such systems are based on multiple-core processors, but they can be constructed using specialized computing devices which are represented by graphics cards and co-processors. It should be noted that the high-end computing clusters are based on Nvidia graphics cards and Intel co-processors. Such clusters are presented in data centers in Russia. Porting a program to another computing architecture is a research task if the user wants to use effectively and efficiently the potential of a computer system. Therefore it is necessary not only to adapt mathematical techniques, but also to develop the structural scheme of parallelization and specialized software to calculate large-scale models on supercomputers. In this work, mathematical modeling of seismic field is carried out for the problem represented in terms of velocity of displacements and stress. The problem is solved in a Cartesian (rectangular) coordinate system with the appropriate zero initial and boundary conditions. A difference method of the second and fourth order of accuracy is used to solve the stated problem [4,5,6]. Computational experiments are carried out on the NKS-30T+GPU cluster of the Center for collective use of the Siberian Supercomputer Center of the Russian Academy of Sciences. This system is presented as CPUs and Nvidia graphics cards.

## 2 Structural Diagram of the Task Parallelization

The application of the difference method to solve the problem of numerical modeling involves the use of a large number of data arrays in calculations. For this reason the calculations of large-scale geophysical models of the objects are realized on multi-core computing systems. It can be both SMP (Symmetric Multiprocessing) servers and clusters with the MPP (Massively multiprocessing) architecture, and systems with specialized computing devices. Typically, a cluster is a set of computing nodes containing several multi-core processors. Some clusters have specialized computing devices. The developed software gives an opportunity to use modern computational tools and technologies for parallel computing. It can be CPUs, Nvidia GPU graphics processors and Intel MIC co-processors. Clusters make it possible to perform calculations for elastic media models using the decomposition of a calculation model for the required number of computing devices and parallel computing processes. The creation of software for such systems occurs due to the development of parallel algorithms, block diagrams and programs with the use of specialized software for the efficient use of computing devices [7]. This paper analyses systems based on a CPU and a GPU. As such a computer system a NKS-30T + GPU cluster was used. There are 40 SL390s G7 servers on this cluster. Each of them has two 6 CPU Xeon X5670 core processors, 96 GB RAM, three Nvidia Tesla M2090 cards on the Fermi architecture. This card has 512 cores, 6 GB of GDDR5 memory. Other compute nodes are presented with blade servers based on Intel Xeon processors of different series. In this paper the authors consider one of the possible and parallel implementations of the numerical algorithm to be used in calculations. This realization implies the application of the parallelization of a computational domain into subdomains along one of the axes (Fig. 1).

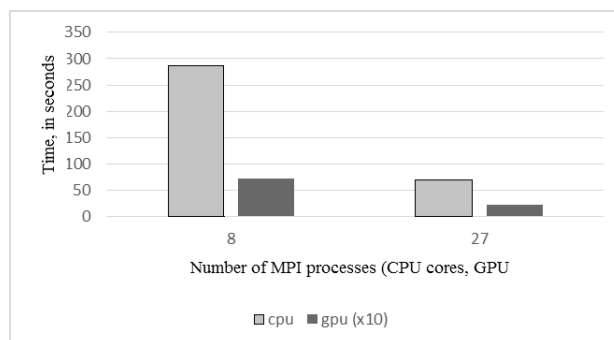
Each of these subareas is calculated in parallel on a dedicated computing device which can be both the CPU and graphics processors. Data exchanges between neighbor computing devices are performed with the use of MPI. To carry out calculations on classical MPP architectures a program is developed using the MPI. For hybrid clusters the MPI and CUDA were applied. Since the CPUs in the NKS-30T + GPU cluster are multicore, the combination of the MPI and OpenMP can be used. In this version of the program 1 MPI process falls on 1 multi-core processor. And the number of OpenMP threads is chosen depending on the number of processor cores. In this



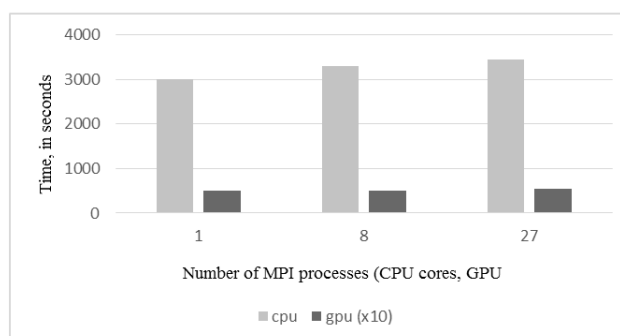
**Fig. 1.** Structural diagram of parallelization.

program the calculations with the difference method of second-order accuracy are implemented. The results of the comparison of the programs for a software implementation on the example of a three-dimensional mathematical modeling are presented in this paper. Graphics cards have a large number of cores and memory in comparison with the CPU. Therefore, an important task was to develop parallel algorithms and software for using graphics cards in calculations. A software implementation was developed for a three-dimensional version on the basis of a one-dimensional decomposition of a computational domain using the scheme of second order accuracy. The use of the graphics card has its own advantages and disadvantages. The calculations can be performed faster but it is necessary to copy the data between the GPU and CPU. The development of the program is complex enough since it is necessary to consider features of the applied algorithms and methods and the specificity of a CUDA technology. When using multiple graphics cards the model region is divided into subregions between calculators. All calculations are carried out only on the GPU. The CPU is used for an interaction between the graphic cards, management, and the implementation of data exchanges between neighbor computing devices. The exchanges of data between the GPUs are implemented through the CPU. To do this, the data are copied from the GPU to the CPU. Between the CPU the MPI exchange functions are executed. Then the data are copied from the CPU to the GPU. Therefore a combination of MPI and CUDA is applied in such a program. The results of the comparison of parallel algorithms for the CPU and GPU were obtained from a scaling test and with the increased amounts of computing resources. While conducting the test the authors applied a three-dimensional homogeneous elastic medium with a  $308 \times 308 \times 308$  mesh and 110 iterations as a model. During the test with the increased amount of resources the model parameters do not vary. During the scaling tests model dimensions change in the proportion to the number of computing devices. The results were obtained on a NKS-30T + GPU cluster for the three-dimensional realization of the parallel algorithm. In figure 2 and 3 the execution time with the use of the GPU is increased 10 times.

Figure 3 shows the behavior of the program when being scaled. So when increasing the size of the computational domain the execution time of the CPU is increasing about 5. When the amount of computing resources in figure 2 increases the ratio of GPU to CPU is  $\times 35$  on average. When compared to the GPU only CPU acceleration on 27 computing devices was 4.2 whereas the GPU figure is 3.4. The developed software is considered to be applicable to solve problems of a two-dimensional numerical simulation.



**Fig. 2.** Test results of varying the amounts of computing devices containing up to 13,824 cores.

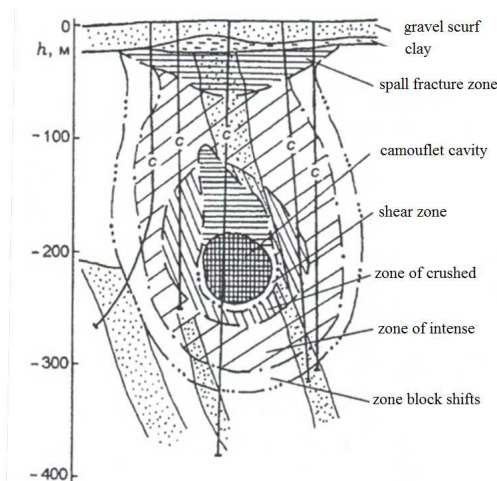


**Fig. 3.** The test results of scaling the parallel algorithm.

### 3 Construction of the Geophysical Models Displaying the Effects of an Underground Nuclear Explosion

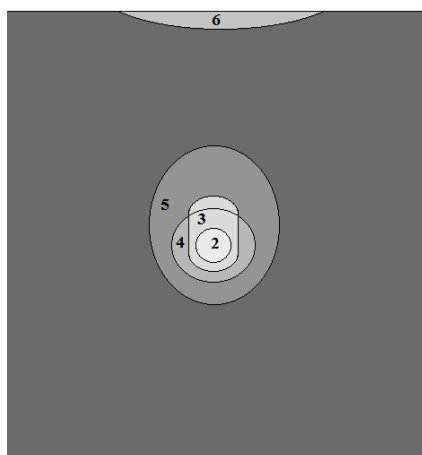
The study of sites where underground nuclear explosions are set off is of direct practical relevance induced by the need of solving problems in the sphere of radiation and geo-ecological safety of territories adjacent to the site of nuclear experiments. Completed bombings had a destructive effect on the surrounding geological structure with the formation of decompressed permeable areas through which radionuclide residues can be carried to the environment (even reaching the daylight surface) [8,9]. The study of the spatial position and configuration of the above-mentioned structures will help monitor and forecast the ecological state of the world's nuclear sites. Also one of the important directions is the development of an on-site inspection technology (of seismic methods in particular) in support of the Comprehensive Nuclear Test Ban Treaty. Numerical modeling of wave fields on the proposed subsurface models can help solve these problems. These snapshots can provide an opportunity to distinguish the group of waves, showing the presence of inclusions in the cavernous environment and assess their geometric dimensions. To study the structure of the wave field which is formed by a seismic radiographic environment different models of cavernous area were simulated. All numerical calculations were carried out using the developed software on the NKS-30T + GPU cluster of the Siberian Supercomputer Center of the Russian Academy of Sciences (ICMMG SB RAS). In this model, the authors studied the effect of geometry on the structure of the wave field of the inhomogeneous medium containing cavity, in order to highlight the distinctive features of the field due to its presence. To create the geophysical model of an underground nuclear explosion the materials [10] were used. Under the scheme of the central core of a nuclear explosion in the wellbore В,<sub>–</sub>102 (Fig. 4) the authors

constructed the geometry of a medium model containing nuclear explosion effects, with basic areas occurring as a result of the test.



**Fig. 4.** Scheme of the central zone of the nuclear explosion in the wellbore B,-102.

Figure 5 shows a part of the developed and explored 2D geophysical model of the inhomogeneous elastic medium with the linear dimensions of 5.0 km along the  $Ox$  axis and 1.0 km along the  $Oz$  axis, containing subareas: a camouflet cavity (2); a shear zone (3); a zone of crushed rocks (4); a zone of intense fracturing (5); a spall fracture zone (6); a host medium (1).



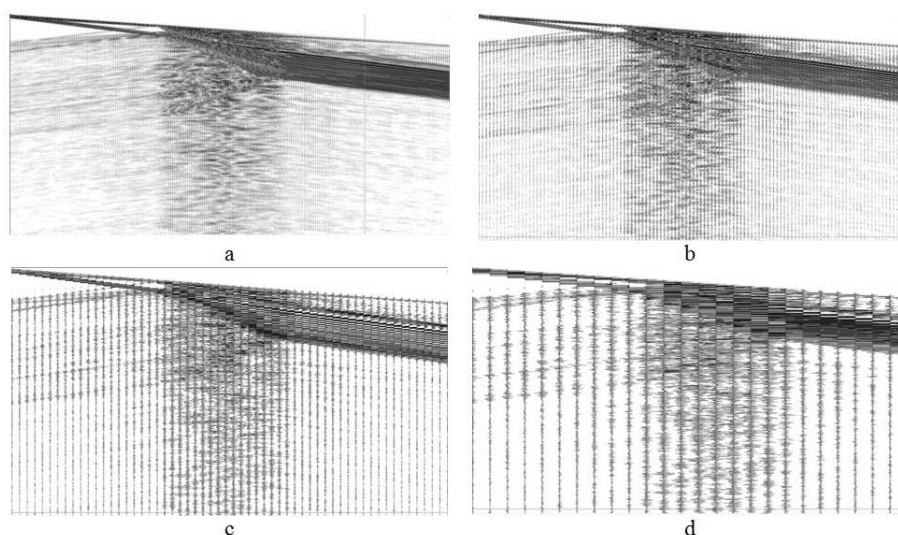
**Fig. 5.** Diagram of the geophysical model of the 2D elastic medium in the  $xOz$  plane. The figure means the number of a model element.

The geometrical dimensions and geophysical characteristics of areas are used by the program to calculate the components of the wave field. The source and the recording geophones are placed on a free surface. The frequency of the source is 30 hertz [11]. The calculations and operation of the developed software for multi-core computing systems result in creating synthetic seismograms and snapshots of the wave field, calculated by the algorithm described in [12]. The paper deals

with the  $U_z$  component of the wave field corresponding to the vertical component of a seismic field [12].

#### 4 The Results of the Numerical Simulation

Numerical experiments on modeling a 2D elastic medium carried out on a cluster NKS-30T + GPU. For the calculations the authors used an MPI program and a CPU. The calculations were performed on 24 cluster cores. Figure 6 shows the synthetic seismograms for the  $U_z$  components for different numbers of geophones in observation system which are presented in a rectangular coordinate system with the x-axis displaying the coordinates of the geophones and the y-axis displaying the arrival times of different waves. For convenience, the same data are presented in another form in figure 7.

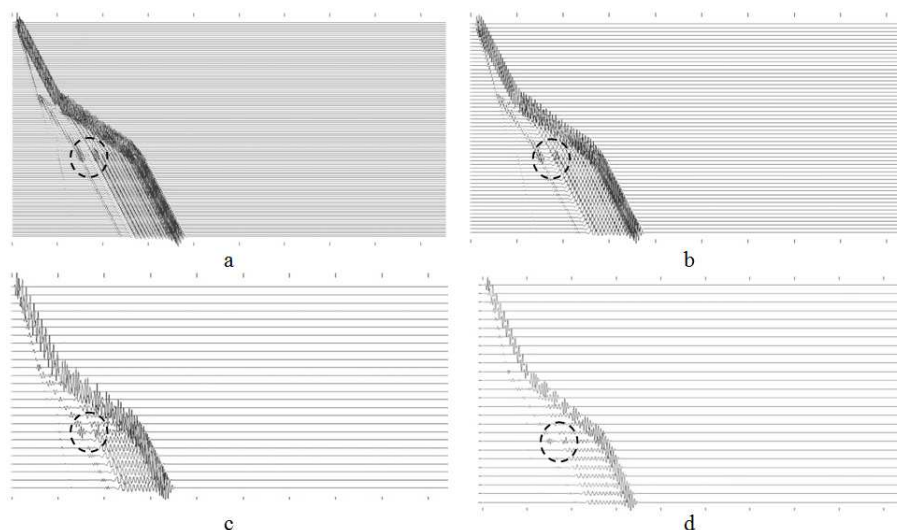


**Fig. 6.** Synthetic seismograms calculated for the  $U_z$  component of the wave field for different numbers of recording receivers (a - 234 receivers, b - 117 receivers, c - 56 receivers, d - 23 receivers).

From figure 6 it is possible to identify a spall fracture zone (the vertical strip of traces with greater amplitude) and determine its extent on the x-axis as the distance between the coordinates of the beginning and end of seismic disturbances. A detailed study of synthetic seismograms can determine the arrival times of seismic waves caused by the aftermaths of underground nuclear tests reflecting from or passing through the border areas. Figure 7 shows how the decrease of number of geophones can affect the allocation of the patches of areas with cavernous inclusions. Thus, if geophones are spaced out it is possible for a small cavity or surrounding area to be skipped. The results are shown for the  $U_z$  components in a rectangular coordinate system where the x-axis presents the arrival times of various waves and the y-axis presents the coordinates of the geophones.

#### 5 Conclusion

A lot of modern problems are related to processing large volumes of data. To solve this problem a lot of multi-core computing systems are being developed and used nowadays. Many of them are



**Fig. 7.** Synthetic seismograms calculated for the Uz component of the wave field for different numbers of recording receivers (a - 234 receivers, b - 117 receivers, c - 56 receivers, d - 23 receivers).

based on graphics cards and co-processors. The software which is being developed by the authors allows numerical modeling of seismic field for environments with complex subsurface structure on different architectures and high-performance computing systems. This paper introduces the technology of mathematical modeling focused on the use of the advanced architectures of exaflop supercomputers. The authors suggest using a block diagram of the parallel algorithm on a variety of computing architectures for the developed software. While testing the behavior of the software it proved possible to show the possibilities of using hybrid systems to significantly speed up the time required for calculations. The attained acceleration allows more knowledge-intensive calculations for large-scale models. A large number of configurable settings available in the process of modeling enables users to build the geometry and define the elastic parameters of models, receive the results of various degrees of accuracy required for each case. The developed software for the clusters of different architectures makes it possible to construct a model of objects and carry out numerical calculations using Nvidia graphics cards. The results of calculations performed on the developed software provide an opportunity to isolate informative wave groups and determine the time they arrive at the border areas of underground nuclear tests. The applicability of different numbers of recording geophones in calculations gives an opportunity to determine their minimum number required for getting the desired resolution characteristics of a seismogram. The obtained results form the basis for developing recommendations to determine the areas of an underground nuclear explosion.

**Acknowledgments.** This work was partially supported by grants of RFBR 14-07-00832, 14-05-00867, 15-07-06821, 15-31-20150 and MES RK 1760/GF4.

## References

1. Rodin G.: Seismology nuclear explosions: Trans. from English. M.: Ed. "The World 190 (1974)
2. Glinski, B.M., Karavaev, D.A., Martynov, V.N., Khairtdinov, M.S.: Numerical simulation of elastic wave propagation in the cavernous environments. In periodic scientific and technical journal National Nuclear Center of the Republic of Kazakhstan, Vol. 3 (43), pp. 96-100. Bulletin NNC RK, Kurchatov, Kazakhstan (2010)

3. Glinsky, B.M., Karavaev, D.A., Kovalevsky, V.V., Martynov, V.N.: Numerical simulation and experimental study of the mud volcano "Mount Karabetova" vibroseismic methods. In: Computational Methods and Programming: new computing technologies, T. 11, в.,- 1, pp. 95-104 (2010)
4. Virieux, J.: P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method. In: Geophysics, Volume 51, April, Number 4, p. 889-901 (1986)
5. Samarskii, A.A.: The theory of difference schemes. M.: Nauka, 656 pages (1977)
6. Bihn, M., Weiland, T. A.: Stable Discretization Scheme for the Simulation of Elastic Waves. In: Proceedings of the 15th IMACS World Congress on Scientific Computation, Modelling and Applied Mathematics (IMACS 1997), Vol. 2, pp. 75-80 (1997)
7. Glinsky, B.M., Rodionov, A.S., Marchenko, M.A., Karavaev, D.A., Podkorytov, D.I., Vince, D.: Using simulation to configure scalable algorithms for high-performance computing. In: Bulletin of the Ufa State Aviation Technical University, T. 17, в.,- 5 (58), pp. 200-209 (2013)
8. Belyashov, A.V., Suvorov, V.D., Miller, E.A.: Seismic study of the upper part of the section on the site of the Semipalatinsk nuclear test site. In: Seismic technology, N3, pp. 64-75 (2013)
9. Belyashov, A.V., Mukusheva, M.K.: Using seismic data to study the effects of underground nuclear explosions on the surrounding geological environment. In: Geophysics, в.,- 6. pp. 36-41 (2011)
10. Adushkin, V.V., Spivak, A.A.: Changing the properties of rocks and arrays in underground nuclear explosions. In: Physics of combustion and explosion, ie. 40, в.,- 6, pp.15 - 26 (2004)
11. Khairetdinov, M.S., Yakimenko, A.A., Karavaev, D.A.: Numerical simulation of wave field in the areas of underground nuclear explosions. In: Periodic Scientific and Technical Journal of the National Nuclear Center of the Republic of Kazakhstan, Issue 2, pp. 76-81. Bulletin NNC RK, Kurchatov, Kazakhstan (2014)
12. Yakimenko, A.A., Karavaev, D.A.: Numerical simulation of elastic wave propagation in media with underground cavities on supercomputers. In: Scientific Bulletin of the NSTU, в.,-2, pp.99-104 (2013)



# The Experience of Implementation of Permutation Tests Using GPU

Alexander Yakimenko<sup>1,2</sup> and Mikhail Grishchenko<sup>1</sup>

<sup>1</sup>Novosibirsk State Technical University,

Prospekt K. Marksa, 20, 630073 Novosibirsk, Russia

<sup>2</sup>Institute of Computational Mathematics and Mathematical Geophysics SB RAS,

prospect Akademika Lavrentjeva, 6, 630090 Novosibirsk, Russia

{alfred.hofmann, ursula.barth,%lncs}@springer.com

<http://www.sccc.ru>, <http://www.nstu.ru>

**Abstract.** This paper proposes an algorithmic approach and program for solving search problems statistically significant over the biological characteristics of genes in a given set. The problem is connected with the implementation of well known in biology permutation (randomization) test. By taking into account the potential parallelism permutation test developed in parallel with the implementation of a program to graphically processors. Provides estimates of the effectiveness of the application of the developed program on empirical material.

**Keywords:** GPU, parallelization, modeling.

## 1 Introduction

This works refers to the solution of the problem of analysis of genetic determination of traits, which an urgent problem for biology. Usually, this research is carried on the basis of analytical criteria (such as t-test, ANOVA, Pearson correlation, etc.) built on a closed logic. The idea about data distribution in the general population may be obtained with the use of samples considered as distributed according to the subjectively given law of data distribution. It might be possible to solve this problem by resampling methods, as they examine sampling data in various combinations as if seeing them from different angles [1,2] rather than require any additional information on the law of data distribution in the general population. Furthermore, another important advantage of resampling methods over analytical methods is that there is no need to continually adjust the levels of statistical significance for the simultaneous testing of many statistical hypotheses reflecting the simultaneous contribution of many factors in the formation of a single trait in the case of the analysis of biological data [3]. Thus, resampling methods are more correct in most biological research in comparison with analytical methods, but they often require huge computing resources for sufficiently accurate estimates of various statistics of the analyzed samples.

Computations can become significantly faster in resampling due to parallelization, and most economical is the parallelization with the use of using graphic processor units (GPUs) [4,5,6]. In this regard, the special software packages allowing us to investigate biological objects (gene groups mainly) by resampling methods were made, among them are the freely available software products RandTestGPU [7] and permGPU [8]. However, the significant disadvantage of those software products is that very simple cases, such as the comparison of two samples of the measured values for genes (expression level, evolution speed, the number of known polymorphic states, etc.) can be tested. This simplified approach to the representation of biological data cannot account for a more detailed and already known experimental information about belonging of genes to different functional and/or structural groups (annotations) and, consequently, determine the fine determination component of the trait under different external or internal conditions.

The aim of this paper is the software implementation and research of effectiveness of parallelization of permutation test aimed at finding the statistically significant overrepresented properties of genes at different external or internal conditions with the use of GPUs. The acceleration obtained in this case substantially increases the performance of the solutions of this problem.

## 2 Problem Definition

The general ideology of organization of the permutation test aimed at finding the statistically significant overrepresented properties of genes is as follows: 1) the value of the statistic is calculated from the empirical material  $G_0(x_1, x_2, x_3, \dots, x_n)$ , where  $x$  is the measured quantitative characterization of the gene (expression level, evolution speed, etc.); 2) the random permutation of quantitative characteristics of genes between samples and measured values is performed; 3) for the same samples with randomly permuted variables  $x_1, x_2, x_3, \dots, x_n$ , we calculate the value of the same statistic  $G_u(x'_1, x'_2, x'_3, \dots, x'_n)$ , which a reflective sample after permutation, and the subscript  $U$  is the number of permutations; 4) permutation procedures 2 and calculations of the statistic 3 are repeated  $U$  times; 5) the error probability is determined in rejecting the zero hypothesis (p-value) as a proportion of the values of  $G_u$  exceeding  $G_0$ .

To implement the permutation test on the graphics card, we should take into account the features of its architecture [5]. It is important to highlight the parts of the code and the parallelizable algorithms. The indivisible operations on data arrays and read/write to a file remain problematic. It is reasonable to distribute the cycles of independent data processing over video card multiprocessors and minimize the number of conditional expressions within such blocks.

The overall structure of the software is shown in Fig. 1. The software has two text files at the input: the one with the permutation test parameters (number of iterations and permutations per iteration) and another one with the input data directly Fig. 2.

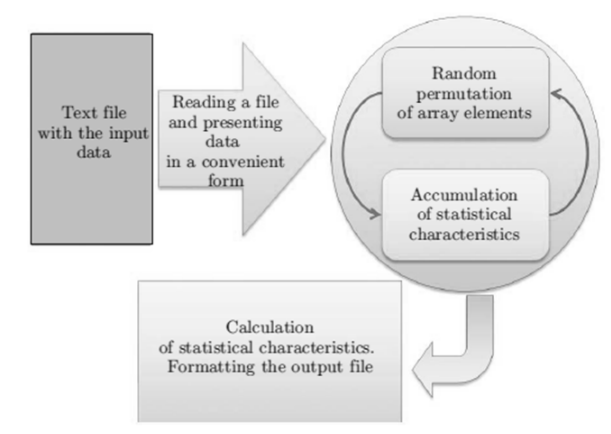


Fig. 1. Generalized permutation test algorithm

## 3 Theory

The whole algorithm works permutation test can be divided into three main stages: 1) reading the input file and forming the data array in a format convenient for further calculations; 2) the

calculation of the sum (or any other quantities, such as average values, variances, etc.) of the measured values of the gene for its different properties (in particular, functional annotations (FAs)), the cycle with mixing of the array elements and collecting the quantities necessary for the statistics; 3) the calculation of p-values and the formation of the file with the results. The most consuming and parallelizable is the second stage. It is presented in more detail in Fig. 3 and Fig. 4.

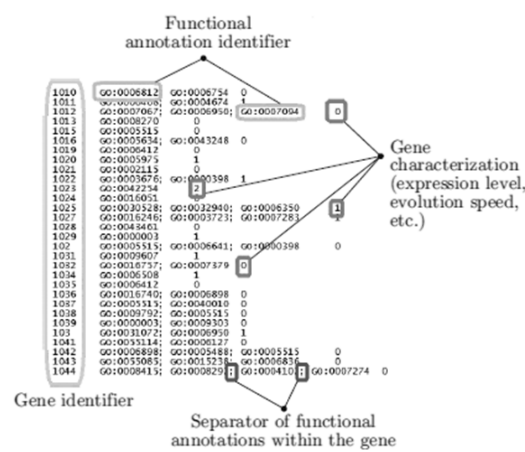


Fig. 2. Data storage structures: gene representation in the input file

We consider an example in more detail, wherein the functional annotations GeneOntology serve as the properties of the genes [9]. Before iterations with permutations, we should calculate the array of sums of real values of the measured characteristics of genes for each available FA. For this purpose, we organized a cycle above the data structure `std::map`, which stores keys `ВТ` gene identifiers and values `ВТ` numerical characteristics of genes, in the sequential version of the developed software algorithm. After this procedure, we start a cycle repeating a random permutation of gene characteristics and further accumulation of statistical values the specified number of times. Upon completion of the procedures described, the output of the p-statistic is performed and the output file with the results is formed, and the significance of certain FAs of genes is concluded on the basis of that file. In addition to the p-statistic, the gene identifiers with the mentioned FAs are given.

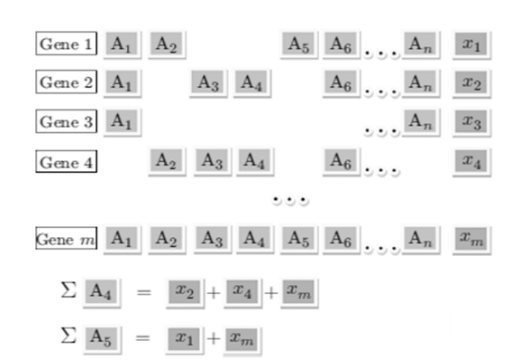
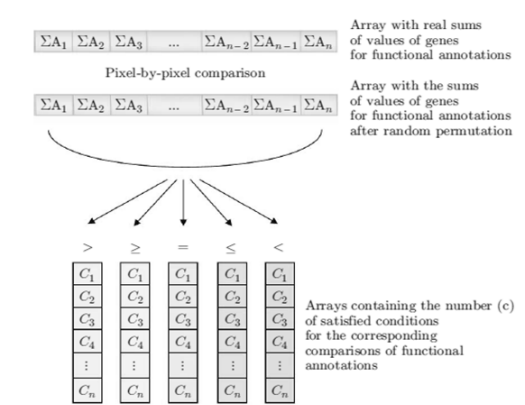


Fig. 3. Data storage structures: gene representation in the input file

It is important to understand that the algorithm features do not allow us to parallelize the larger part of the code. Moreover, the calculation features on graphics cards restrict the use of data storage structures (DSS's) to arrays (the thrust library is not considered) [5]. To solve this problem, we need to review the approaches to implementing the algorithms most profitable in parallelization and move from the more complex DSS's to arrays. This method was used in the algorithm for calculating the sums of numerical characteristics of genes for various FAs.

Fig. 2 shows the structure of a file with input data. It should be counted into memory and presented in DSS's convenient for storage. It is important to note that the figure shows a test sample of input data. In real problems, the identifier of FAs may have a complicated name and a separator between FAs can be any character or a combination of characters. Informative data here are the identifiers of FAs and gene characteristics. It is their representation that plays a big role in enhancing the performance of the entire software.

Fig. 3 shows the gene structure in a schematic form and shows the meaning of the concept of the sum of numeric values of genes for the FAs. For example, this sum consists of the characteristics of the genes  $x_2$ ,  $x_4$  and  $x_m$  for the functional annotation of  $A_4$ .



**Fig. 4.** Algorithm of accumulation of statistical values

In addition, it follows from Fig. 3 that it is reasonable to apply two arrays instead of the `std::map` structure for storing information on the genes: a two-dimensional array reflecting the falling of FAs in the gene and a one-dimensional array containing the measured characteristics of genes. This reduces the amount of data transmitted in the GPU memory and use the standard cuBLAS library to perform matrix-vector multiplication.

The overall structure of the software is shown in Fig. 5. Here the GPU blocks are executed on the GPU. It should be noted that each time we take an original matrix of inclusions of functional annotations for the calculation of real and random sums, and only the array of values of gene characteristics changes. Thus, each iteration requires only one transmission of a one-dimensional array, which reduces the calculation cost. The use of a CPU for the random permutation of the array of values is due to the impossibility of organizing a parallel unordered access to the elements in the array on the GPU. This, in turn, requires the transfer of control from the GPU to the CPU (Central Processor Unit) on each cycle iteration. The algorithm for calculating the real and random sums is in the representation of the set of functional annotations in the form of two-dimensional array with zeros and units (Fig. 6). The available genes are located in the lines, and the FAs included in them are in the columns. Thus, a unity is written in place of the FA included in the gene, otherwise zero is written. Then the matrix-vector multiplication of such transposed

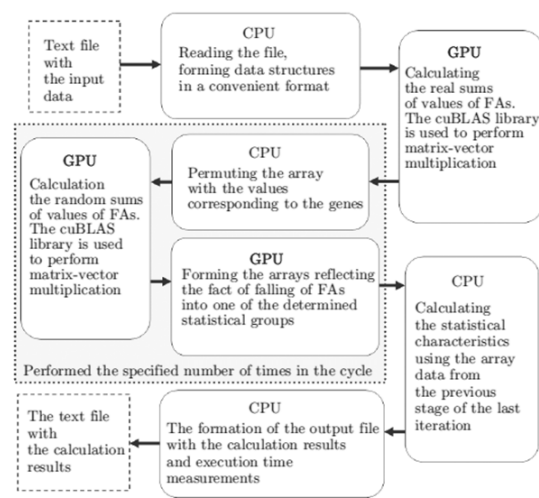


Fig. 5. Software structure

array of dimension of  $M \times N$  ( $M$  is the number of rows and  $N$  is the number of columns) with one-dimensional array of dimension of  $N$ , containing the values of the characteristics of genes results in an array with the sums of FAs for all genes.



Fig. 6. Representation of FA in the GPU memory

Conditional operations are widely used in the algorithm for collecting the statistical values, which reduces the efficiency of execution on the GPU. This is due to the fact that all the GPU threads execute the code completely and cut off the false condition finishing the IF block [5]. Nonetheless, the reached acceleration  $S$  is proportional to  $(S_g \times R_m) / C_p$ , where  $R_m$  is the number of compared real sums of numeric values for the FAs with random sums of these values,  $C_p$  is the degree of parallelization of the algorithm,  $S_g$  in our case equals five and is determined by the number of random sums of the values of FAs for the categories of larger, smaller, equal, larger or equal, smaller or equal real sums of values of FAs.

#### 4 Experimental Results

The software performance was estimated by considering the parallel and consistent implementations with various parameters to run it on two real problems of different dimensions. Fig. 7 shows the time of solution of the first problem with 2256 studied genes containing 782 functional annotations. The graph shows the software execution time for a parallel version on the GPU and a sequential version on the CPU. We performed four runs for each software version: (1) 492 permutations and 10000 iterations, (2) 492 permutations and 100000 iterations, (3) 2256 permutations and 10000 iterations, and (4) 2256 permutations and 100000 iterations.

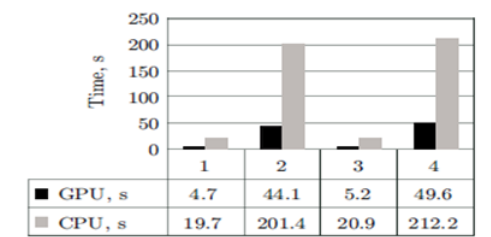


Fig. 7. Software execution time for the first tasks

Fig. 8 shows the execution times of the second problem, which has a significantly larger number of studied genes (19147) with a greater number of functional annotation (898). Unlike the first problem, we give the achieved acceleration of a parallel version of the software with regard to the sequential version rather than the software execution time on the CPU. The number of runs reduced to three: (1) 5000 permutations and 10000 iterations, (2) 19147 permutations and 100000 iterations, and (3) 5000 permutations and 100000 iterations.

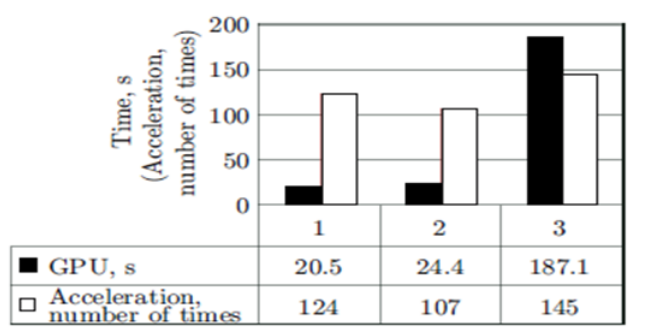


Fig. 8. Software execution time for the second tasks

## 5 Discussion of results

For first task (Fig. 7), an approximately 5-time acceleration is achieved, which is determined by the small size of the problem. The number of iterations increases the execution time linearly, whereas the growth in the number of permutations only slightly slows down the process of counting. For second task (Fig. 8), we achieved an 145-time acceleration. It is expected that acceleration increases in problems with greater dimensionality.

## 6 The system development

Today it is necessary to verify several hypotheses with the same set of functional annotations. In this case values of the gene characteristics are the matrix  $N \times K$  instead of a vector ( $K$  - number of hypotheses), where each column represents a hypothesis. For simultaneous testing of several hypotheses it needs to amend the existing algorithm as follows:

1. To replace the matrix-vector multiplication with matrix-matrix. Then, the resulting matrix will contain a set of arrays each of that fills with sums of all the FA genes for the hypothesis under consideration.

2. To replace elementwise permutation of genes characteristics with arrays permutation of the genes characteristics. This is acceptable because, in general, all the hypotheses are independent to each other and there is no need for independent rearrangements of gene characteristic values.

## 7 Conclusion

The sequential and parallel versions of software for the permutation test aimed at finding statistically significant overrepresented characteristics of genes under different external or internal conditions for computing devices: PC with NVIDIA GPU and hybrid supercomputer NCC-30T+GPU of the Siberian Supercomputer Center, Siberian Branch of the Russian Academy of Sciences. In the course of the work, the problem of parallelizing the most energy-consuming algorithms of permutation test software for the implementation on the GPU is solved. The cuBLAS library of matrix-vector multiplication, which allowed for a transition of this algorithm to the architecture of GPUs, was used. According to the results of performance estimation, the time acceleration of the software was shown on the two considered problems (see Fig. 7, 8) using a GPU, and it amounted to 150 time with regard to the sequential version. It is noted that the software execution time is affected by the input data size (the number of genes and functional annotations) and the number of iterations with permutations. A negligible effect of the number of permutations performed before each iteration of calculation of random sums of values for the functional annotations during the software operation time was demonstrated. Currently, the software is in trial operation at the Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences. This work was supported by the Ministry of Education and Science of Russia (Civil Code No. P-857 "Software Development for HPC in Bioinformatics") and the Russian Foundation for Basic Research (Grant No. 11-04-01771), projects no. 14-07-00518-a, 12-01-00773, grant Novosibirsk State Technical University no. 2.6.1.

## References

1. Efron, B.: *Nontraditional Methods of Multivariate Statistical Analysis*. Finansy i Statistika, Moscow, (1988)
2. Good, P.: *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Springer Verlag, N. Y. (2005)
3. So, H.C., Sham, P.C.: "Multiple Testing and Power Calculations in Genetic Association Studies in Genetics of Complex Human Diseases: A Laboratory Manual Eds. by Al-Chalabi, A., Almasy, L. Cold Spring Harbor Laboratory Press, N. Y., pp. 49–59, (2010)
4. Dematte, L., Prandi, D.: "GPU Computing for Systems Biology," *Brief. Bioinform.* 11 (3), pp. 323–333 (2010)
5. Boreskov, A., Kharlamov, A., Markovskii N.: *Parallel Computations on the GPU. Architecture and Software Model of the CUDA*. Mosk. Univ., Moscow, (2012)
6. Pustovalov, E., Voitenko O., Grudin B., Plotnikov V.: *Graphics Processors in Problems of Electron Tomography*, *Avtometriya* 48 (1), pp. 72–79, (2012)
7. Van Hemert, J. L., Dickerson J. A.: *Monte Carlo Randomization Tests for Large-Scale Abundance Datasets on the GPU*, *Comput. Methods Programs Biomed.* 101 (1), pp. 80–86 (2011). <https://subversion.vrac.iastate.edu/Subversion/RandTestGPU/svn/RandTestGPU/>.
8. Shterev, D., Jung, S. H., George, S. L., Owzar, K.: *PermGPU: Using Graphics Processing Units in RNA Microarray Association Studies*, *BMC Bioinformatics*, 11, p. 329 (2010). <https://code.google.com/p/permgpu/>.
9. Ashburner, M., Ball, C. A., Blake, J. A. et al.: *Gene Ontology: Tool for the Unification of Biology*, The Gene Ontology Consortium. *Nature Genet.* 25 (1), 25–29 (2000).

## Session II. Information Management, Processing and Security



# Study of the Problem of Creating Structural Transfer Rules for the Kazakh - English and Kazakh-Russian Machine Translation Systems on Apertium Platform

Balzhan Abduali, Aida Sundetova, Nurbolat Zhanbussunov, and Zhansaya Musabekova

Al-Farabi Kazakh National University, Information Systems Chair,  
Al-Farabi av., 71, 050040 Almaty, Kazakhstan  
balzhan\_5696@mail.ru, sun27aida@gmail.com, nurbolat\_03.93@mail.ru, zhansaya\_03\_94@mail.ru  
sun27aida@gmail.com, balzhan\_5696@mail.ru, nurbolat\_03.93@mail.ru, zhansaya\_03\_94@mail.ru  
<http://www.kaznu.kz>

**Abstract.** This paper presents the current state of development a shallow-transfer rule-based machine translation (MT) system from Kazakh to English and Kazakh to Russian. The main morphological and syntactic differences between the Kazakh and two languages are presented, and it is described how the MT system was designed to overcome these challenges. We showed current an evaluation of system coverage and outline and future work.

**Keywords:** machine translation, Apertium, syntactic parsing, structural transfer rules, chunk, interchunk, postchunk.

## 1 Introduction

Translating texts between English and Kazakh, Kazakh and Russian faces some challenges. First of all, Kazakh is one of group of Turkic languages, it means that Kazakh language shows clear ordering of morphemes and they are changed by interaction between neighboring morphemes (vowel harmony, sonorization, etc.). On the other hand, syntax of Turkic languages is very different from English or Russian: subject-verb-object order in English as basic order in Russian, but subject-object-verb order in Kazakh, ) [1]. Such kind of syntax reordering and transformation we are going to solve by creating structural transfer rules, which define the transformations needed to convert original sentences and text words into their target language.

Structural rules divides the main task of sentence or phrase on the shares of the phrasal, for example noun phrase (NP), verb phrase(VP) and other(for instance, [NP I] [VP wrote] [NP-acc book]). We obtain the most probable sequences of word to divide sentences into phrases(chunks), which is done by hidden Markov Model [2]. For example, computing probability of sequence "determiner+noun"will be  $P(\text{Det}/N) = C(\text{Det},N)/C(\text{Det})$ , where  $C(\text{Det},N)$  - is number of situation when noun comes after determiner. Secondly, we use the most likely tag sequences to create "chunks"by structural transfer rules on Apertium platform, which is a free/open-source rule-based machine translation (MT) platform, launched in 2005 by the Universitat d'Alacant [1].

## 2 The Apertium platform

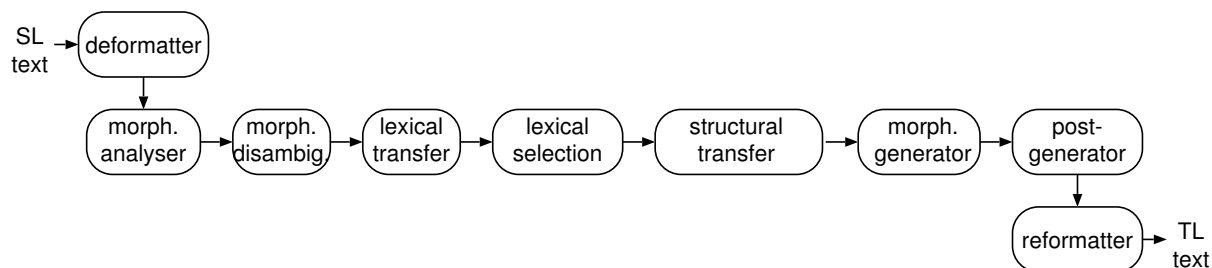
We would like to give a very brief description of Apertium here. The project which we are creating is based on platform Apertium.

The Apertium platform supplies:

- a FOS modular shallow-transfer MT engine with text format management, statistical lexical disambiguation, finite-state lexical processing, and shallow structural transfer based on finite-state pattern matching;

- FOS linguistic data in well-specified XML formats for a wide variety of language pairs;
- FOS tools such as compilers to turn linguistic data into a fast and compact form used by the engine and software to learn disambiguation or structural transfer rules;
- extensive documentation on usage[4];

The Apertium engine is a pipeline or assembly line consisting of the following stages or modules:



**Fig. 1.** The architecture of the Apertium platform.

- A deformatter which encapsulates the format information in the input document as superblanks that will then be seen as blanks between words by the rest of the modules.
- A morphological analyser which segments the text in surface forms (“words”) and delivers, for each surface form, one or more lexical forms consisting of lemma, lexical category and morphological inflection information. It reads a finite-state transducer (FST) generated from a source-language (SL) morphological dictionary (MD) in XML.
- An optional constraint grammar [5] to reduce or remove entirely part-of-speech ambiguity before the statistical PoS tagger, and to provide syntactic and semantic labelling.
- A statistical PoS tagger which chooses, using a first-order hidden Markov model, the most likely lexical form corresponding to an ambiguous surface form, as trained using a corpus and a tagger definition file in XML.
- A Hidden Markov model (HMM) allows us to talk about both observed events (like words that we see in the input) and hidden events (like part-of-speech tags) that we think of as causal factors in our probabilistic model[2].
- A Viterbi Algorithm is perhaps the most common decoding algorithm used for HMMs, whether for part-of-speech tagging or for speech recognition. The term Viterbi is common in speech and language processing, but this is really a standard application of the classic dynamic programming algorithm and looks a lot like the minimum edit distance algorithm[2].
- A lexical transfer module which reads each SL lexical form and delivers the corresponding target-language (TL) lexical form by looking it up in a bilingual dictionary in XML using a FST generated from it.
- A structural transfer, generally consisting of three sub-modules (some language pairs use only the first module and some others call more than three, see below):
  - A chunker which, after invoking lexical transfer, performs local syntactic operations and segments the sequence of lexical units into chunks. A chunk is defined as a fixed-length sequence of lexical categories that corresponds to some syntactic feature such as a noun phrase or a prepositional phrase.
  - An interchunk module which performs more global operations with the chunks and between them. More than one interchunk module can be used in sequence.

- A postchunk module which performs finishing operations on each chunk and removes chunk encapsulations so that a plain sequence of lexical forms is generated.

### 3 Structural transfer rules

The Apertium structural transfer module converts source-language lexical form into target-language lexical form pairs (for example, Russian phrase "v sadu(in garden)" in Kazakh will be "baqshada") and after series of transformations, source language lexical form become target language lexical form. Such transformations include next operations: agreement, reordering, etc.

Structural transfer module contains three stages for Kazakh-Russian, and four stages for English-Kazakh[6]. Currently, by computing the most likely tag sequences, for English-Kazakh were defined 6 main types of chunks (phrases): NP - for noun phrases ("beautiful place"), AdjP - adjective phrases, VP - verb phrases, PP and GenP - adpositional phrases, which include genitive phrases and prepositional phrases. Each chunk is defined by a set of patterns [7], which consists of a category and attributes. Rules create chunks by detecting certain patterns, for example, for "big house" pattern is adjective+noun.

#### 3.1 Linguistic data

To create correct transfer rules, lexical form of each word from context should be clearly and correctly defined in dictionary. Three type of dictionaries are used in Apertium platform for lexical processing: monolingual dictionaries for morphological analysis and generation of English, Russian, Kazakh and bilingual dictionaries for English-Kazakh, Kazakh-Russian lexical transfer.

The Russian (apertium-rus.rus.dix) and English (apertium-eng-kaz.eng.dix) dictionary is used to find each lexical forms of each word from text. Dictionaries are filled with list of all lexical units, source-language alphabet, definition of grammatical categories, such like noun, verb, gender, etc.; each category has paradigms describing groups of correspondences between parts of SFs and LFs:

*Example of "teach" paradigm*

```
<pardef n="t/each_vblex">
  <e i="">
  <p>
<l>each</l>
<r>each<s n="vblex"/><s n="inf"/></r>
  </p></e><e i=""><p>
    <l>each</l>
    <r>each<s n="vblex"/><s n="pres"/></r>
  </p></e><e i=""><p>
    <l>eachs</l>
    <r>each<s n="vblex"/><s n="pres"/><s n="p3"/><s n="sg"/></r>
  </p></e><e i=""><p>
</pardef>
```

(Example from apertium-eng-kaz.eng.dix)

The Kazakh monolingual dictionary uses a morphological transducer, which is called the Helsinki Finite State Toolkit [8]. Two types of formalism are used: first, the `lexc`(`apertium-kaz.kaz.lexc`) defines lexicons by word classes and subclasses, and the `twol`(`apertium-kaz.kaz.twol`) formalism is used for morphophonological rules such like desonorization, vowel harmony, nasalization, etc. For instance, the locative case suffix *-DA* could become *-de*, *-ta*, *-te* by depending on previous harmony: `baqshalar+da`, `mektep+te`, etc.

The English–Kazakh bilingual dictionary provides lexical transfer between English and Kazakh words. There are could be ambiguity: multiple entries will be shortened by lexical-selection rules depending on context [9] And in the Kazakh-Russian dictionary, `apertium-kaz-rus.kaz-rus.dix` is filled with words and their translations:

*Example of inserting paradigms*

```
<dictionary>
  <alphabet></alphabet>
  <sdefs>
    <sdef n="num" c="Имя числительное"/>
    ...
  </sdefs>
<pardefs>
<pardef n="_num_gender">
<e>
  <p>
    <l></l><r><s n="m"/><s n="an"/><s n="sg"/><s n="nom"/>
    </r>
  </p>
</e>
<e>
  <p>
    <l></l><r><s n="m"/><s n="an"/><s n="sg"/><s n="det"/>
    </r>
  </p>
</e>
<e>
  <p>
    <l></l><r><s n="m"/><s n="an"/><s n="sg"/><s n="ord"/>
    </r>
  </p>
</e>
  ...
</pardef>
<e>
  <p>
    <l>бip<s n="num"/></l><r>один<s n="num"/></r>
    </p>
  <par n="_num_gender"/>
</e>
```

(Example from apertium-kaz-rus.kaz-rus.dix)

We create for words new paradigm for numerals. It is for do not write one analyses for all words. In this paradigm we write gender, case, number.

And for Adjectives is created same paradigm with numerals. Adjectives has three degrees of comparison.

*Choosing genitive for "Sorok chelovek"*

```
<pardef n="__adj_sint">
<e><p><l></l><r><s n="m"/><s n="an"/>
      <s n="sg"/><s n="nom"/></r></p></e>
<e><p><l></l><r><s n="m"/>
      <s n="an"/><s n="sg"/><s n="det"/></r></p></e>
<e><p><l></l><r><s n="m"/>
      <s n="an"/><s n="sg"/><s n="ord"/></r></p></e>
<e><p><l><s n="subst"/></l><r>
      <s n="m"/><s n="an"/></r></p></e>
<e><p><l><s n="comp"/></l><r>
      <s n="comp"/></r></p></e>
</pardef>
```

(Example from apertium-kaz.kaz.lexc)

Then in the period of translating some words, which have two meaning, it can be seen that sometimes words in not applied part-of-speech tag right. For example "sorokA chelovek". There is word "sorokA" has two meaning: 1. number - "forty" and 2. view of bird - "magpie". To solve this problem we must write rules for this situation. And in the apertium-rus.rus.rlx we write rule:

*Choosing genitive for "Sorok chelovek"*

```
SELECT Gen IF (0 Num) (1 N + Gen) ;
```

(Example from apertium-kaz.kaz.rlx)

To improve quality of translation it is very important to fill dictionary with words with correct part of speech tags.

## 4 Results

Running English-Kazakh, Kazakh-Russian (and vice versa) systems translate simple phrases and sentences. In English-Kazakh bilingual dictionary total 13685 word entries, in Kazakh-Russian bilingual dictionary contains 9043 word.

## 5 Conclusion

Current structural transfer rules translate some cases of phrases and solve reordering operation. In the future work will be considered: secondary members of the sentence (object, adverbial modifier), ordering case changes on endings, work on the polysemy prepositions in machine

translation from Russian into Kazakh language, the implementation of relations between the parts of speech for the simple sentences for Russian language, the work on the implementation of the structure of interrogative and exclamation sentences on Russian language with the transition to the Kazakh language, some errors on the lexical and syntactic parser generator and machine translation, the work on the elimination of ambiguity of polysemy of words.

In future work is planned to generate chunker rules automatically from parallel corpora and use it with other stages in transfer rules to improve translation quality.

## References

1. Pecherskikh, T.F., Amangeldina, G.A. (2012) "Features of translation of different languages (on example English and Kazakh languages) Young scientist, No.3, 259-261, <http://www.moluch.ru/archive/38/4406/>
2. Daniel Jurafsky and James H. Martin, Speech and language processing, edit.2, Uppel Saddle River, New Jersey 07458
3. Mikel L. Forcada, Francis M. Tyers, Gema Ramirez-Sanchez, The Apertium machine translation platform: Five years on, p: 6, <http://apertium.org/>
4. Documentation on a wide variety of development and usage scenarios can be found on the Apertium Wiki <http://wiki.apertium.org/>
5. Constraint grammar [http://beta.visl.sdu.dk/constraint\\_grammar.html](http://beta.visl.sdu.dk/constraint_grammar.html)
6. Sundetova, A., Forcada, M. L., Shormakova, A., Aitkulova, A.: Structural transfer rules for English-To-Kazakh machine translation in the free/open-source platform Apertium. Proceedings of the International Conference on Computer processing of Turkic Languages, pp. 317–326. L.N. GUMILYOV EURASIAN NATIONAL UNIVERSITY, Astana(2013)
7. Steven Abney, Parsing By Chunks. Principle-Based Parsing, Dordrecht:Kluwer Academic Publishers (1991).
8. Krister Linden and Miikka Silfverberg and Erik Axelson and Sam Hardwick and Tommi Pirinen: HFST–Framework for Compiling and Applying Morphologies. Cerstin Mahlow and Michael Pietrowski. Communications in Computer and Information Science. Vol. 100.978-3-642-23137-7. 67-85 (2011)
9. A.M. Sundetova and A. S. Karibaeva. Creating bilingual dictionary for English-Kazakh machine translation system on Apertium Platform. Proceedings of the international scientific-practical conference "The application of information and communication technologies in education and science dedicated to the 50th anniversary of the Department of Information and Communication Technologies and the 40th anniversary of the Department of "Information Systems" of Al-Farabi Kazakh National University. 22nd november 2013., p: 53–57, 2013

# Multicriteria Statistical Analysis of Test Biometric Data

Berik Akhmetov, Ivanov Aleksandr, Yulia Funtikova, and Zhibek Alibiyeva

<sup>1</sup> A.Yesevi International Kazakh-Turkish University, Turkestan, Kazakhstan  
bahitzhan@rambler.ru

<sup>2</sup> Penza State University, Penza, Russia  
bio.ivan.penza@mail.ru

<sup>3</sup> K.I. Satpayev Kazakh National Technical University, Almaty, Kazakhstan  
bahitzhan@rambler.ru, bio.ivan.penza@mail.ru, alibievajibek@gmail.com

**Abstract.** In the transition to the use of two-criterion of statistical analysis is possible to obtain solutions with a higher reliability. Twocriterial statistical analysis on parallel inspection of two alternative statistical hypotheses about the normal and uniform distribution can reduce the likelihood of errors is proportional to the product of the probabilities of each particular hypothesis testing. This reduces the space requirements of the test sample several times.

**Keywords:** multicriteria statistical analysis, a network of private Pearson parallel validation of two statistical hypotheses, biometric data, the test sample.

## 1 Introduction

One of the most popular in the statistical analysis of the data is Pearson. Chi-squared Pearson devoted entirely to the first part of the recommendations of the State Standard [1], while all other criteria are described in the second part of the recommendations [2]. Detailed description of Pearson in the first part of the recommendations of the State Standard [1], reflects the high demand for this particular industry criteria. Techniques based upon use of the chi-square test, suggest checking out some statistical hypothesis about the distribution of the observed values. The calculations are carried out according to the classical formula:

$$\chi^2 = n \cdot \sum_{i=1}^k \frac{\left(\frac{b_i}{n} - \tilde{p}_i\right)^2}{\tilde{p}_i} \quad (1)$$

where  $b_i$  - number of experiments have fallen to the  $i$ -th interval histogram - the expected theoretical probability of hitting the  $i$ -th interval histogram,  $n$  - the number of tests in the test sample,  $k$  - the number of columns of the histogram.

Unfortunately, standard statistical calculation methods (1) for analyzing biometric data give inaccurate results. In order to achieve the probability of errors at the level of 0.05 we have to use some 100 runs in a test sample.

The main source of error in the analysis of the biometric data is insufficient test data in the test samples [3, 4, 5]. This situation is not characteristic only for testing biometric information security. The same situation occurs in the processing of any biometric data (medical, sports, biological). The problem of improving methods of application chi-square test for statistical processing of fuzzy biometric data received considerable attention magazine B«BiometricsB», which regularly publishes articles on this subject [6, 7, 8] since the 30s of the last century.

In the early 21st century, there is a tendency to solve the problem of bad data values artificially filling the empty spaces in the intervals of the histogram, the so-called "bootstrap method"[9], which destroys the natural correlations in substantially dependent biometric data. Roughly the same effect can be achieved by smoothing the digital histograms real data [10]. For the same

purpose can be used to morph examples crossing-parents, and getting the many examples of offspring [11, 12].

This article focuses on another area of research related to the use of two or more statistical tests. Currently, are known dozens of statistical criteria. The most common statistical tests of hypothesis testing in the analysis of biometric data are shown in Table 1 with the time of their creation.

**Table 1.** The most popular statistical criteria

s.n.	Criterion name and the year of creation	Formula
1	Pearson's chi-squared test ( $\chi^2$ ) 1900 year.	$\int_{-\infty}^{+\infty} \frac{\{p(x) - \tilde{p}(x)\}^2}{\tilde{p}(x)} \cdot dx$
2	Cramer-von Mises test 1928 year.	$\int_{-\infty}^{+\infty} \{P(x) - \tilde{P}(x)\}^2 \cdot dx$
3	Kolmogorov-Smirnov test (K-S test or KS test) 1933 year.	$\sup_{-\infty(x(+\infty))}  P(x) - \tilde{P}(x) $
4	Cramer-Smirnoff - von Mises test 1936 year.	$\int_{-\infty}^{+\infty} \{P(x) - \tilde{P}(x)\}^2 \cdot d\tilde{P}(x)$
5	Gini's test 1941 year.	$\int_{-\infty}^{+\infty}  P(x) - \tilde{P}(x)  \cdot dx$
6	Anderson-Darling test 1952 year.	$\int_{-\infty}^{+\infty} \frac{\{P(x) - \tilde{P}(x)\}^2}{\tilde{P}(x) \cdot \{1 - \tilde{P}(x)\}} \cdot d\tilde{P}(x)$
7	Kuiper's test 1960 year.	$\sup_{-\infty(x(+\infty))} \{P(x) - \tilde{P}(x)\} + \sup_{-\infty(x(+\infty))} \{\tilde{P}(x) - P(x)\}$
8	Watson's test 1961 year.	$\int_{-\infty}^{+\infty} \left\{ \tilde{P}(x) - P(x) - \int_{-\infty}^x [\tilde{P}(x) - P(x)] \cdot d\tilde{P}(x) \right\} \cdot d\tilde{P}(x)$
9	Frotsini's test 1978 year.	$\int_{-\infty}^{+\infty}  P(x) - \tilde{P}(x)  \cdot d\tilde{P}(x)$
10	Gini coefficient 2006 year.	$\int_{-\infty}^{+\infty}  p(x) - \tilde{p}(x)  \cdot dx$

Table 1 shows that the statistical criteria set up gradually from 1900 to the present. The most recent was a differential criterion Gini [13] specifically for the processing of biometric data. This criterion was the most powerful and constructed by replacing the original criteria Gini (line 5 in Table 1 [14], 1941) the probability function on their derivatives (density distribution of probabilities).

Obviously, each of the criteria of Table 1 examines a test sample from its side. They all complement each other. You can try to create some generic criteria that will take into account the data of all 10 criteria in Table 1, built to test the hypothesis, only the first observation of the distribution density - or its equivalent, as a function of probability -. Moreover, we can double the number of statistical tests in Table 1, if each of them to build the right to test two statistical hypotheses, or their integral analogs. The latter assertion is far from obvious. Its numerical proof of the subject of this article.

## 2 Numerical experiment for a description of the distribution of values of the chi-square test in the final test sample.

The popularity of using the chi-square of Pearson in the industry is largely due to the fact that when  $n \rightarrow \infty$  its distribution is described by the gamma function with  $m = k-1$  degrees of freedom:

$$p_{\chi^2}(n = \infty, m = k - 1, x) = \frac{1}{2^{\frac{m}{2}} \cdot \Gamma\left(\frac{m}{2}\right)} \cdot x^{\frac{m}{2} - 1} \cdot e^{-\frac{x}{2}} \quad (2)$$



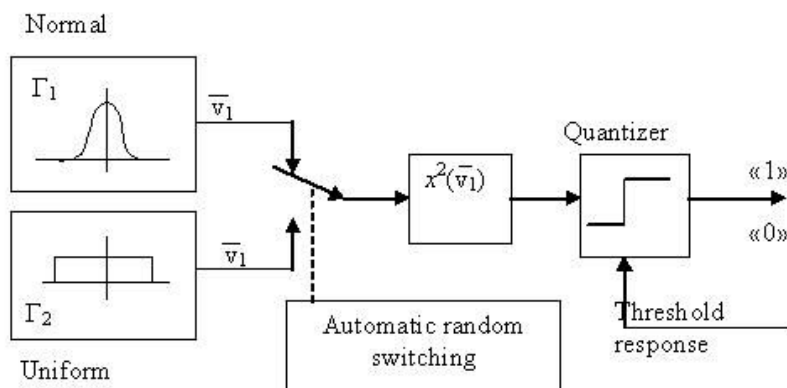
Analytical description (2) was obtained by Pearson in 1904 and played a crucial role in the first half of the 20th century, when the computing capabilities used in the statistical processing of data were very, very limited.

Unfortunately, the traditional use of chi-square test for biometrics gives unsatisfactory results. One of the reasons is an error, resulting from the final test sample. Practice shows that at the final test sample (for example,  $n = 81$ ) in the number of degrees of freedom chi-square distribution is non-integer (fractional) and because of this there is a significant difference:

$$p_{\chi^2}(n = 81, m \neq k - 1, x) \neq p_{\chi^2}(n = \infty, m = k - 1, x) \tag{3}$$

Error due to the finiteness of the test sample can be accounted by a numerical experiment. Today repeating the experiment on a computer 1000000 times quite possible which makes the value of the distribution function values with acceptable accuracy for practical use.

In organizing the numerical experiment, we start from the fact that should be checked two statistical hypotheses. The first hypothesis is that these test samples have normal distribution of values. The second hypothesis is that the data of the same sample may have a normal distribution of values. As a consequence, the organization of a numerical experiment it is necessary to use two pseudo random generator program data, as shown in the block diagram of Figure 1.



**Fig. 1.** Block diagram of the organization of the numerical experiment on power estimation dimensional chi-square test.

Each of the generators of random data G1 (normal data) and G2 (data uniformly distributed) randomly inputted into the calculator values of the chi-square test (1). Then the values chi-square test should be compared with a certain threshold quantizer. If the chi-square value less than the threshold, the decision about the normality of input investigated. If the Chi-square test (1) is above or below the threshold, the decision on the highest validity of one of the hypotheses.

### 3 Checking the first hypothesis of normal distribution of values for the final test sample

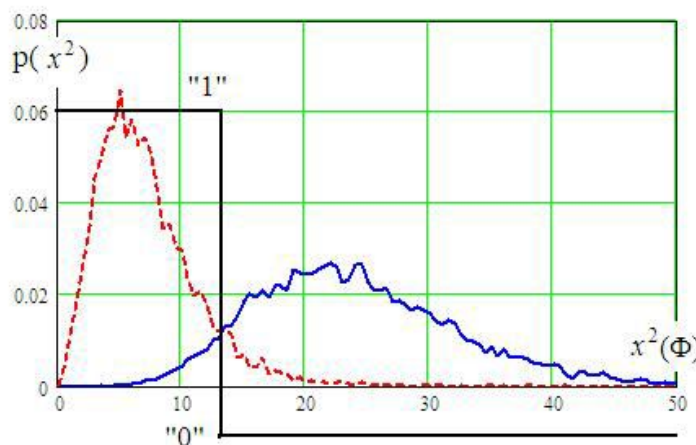
We start from the fact that the chi-square test is required to recognize the emergence of the situation data corresponding to a series of 81 samples obtained from normal generator G1. For this purpose, we calculate the expectation of the test sample -  $E(x)$  and its standard deviation -  $\sigma(x)$ . Next, we will build a histogram consisting of  $k = 9 = \sqrt{81}$  columns evenly covering the

range from the minimum value -  $(E(x) - 3 \cdot \sigma(x))$  to a maximum value -  $(E(x) + 3 \cdot \sigma(x))$ . The values of the chi-square test will be calculated as follows:

$$\chi^2(\Phi) = 81 \cdot \sum_{i=1}^9 \frac{\left( \frac{b_i}{81} - \frac{1}{\sigma(x)\sqrt{2\pi}} \int_{x_i}^{x_{i+1}} \exp \left\{ \frac{-(E(x)-u)^2}{2 \cdot (\sigma(x))^2} \right\} du \right)^2}{\frac{1}{\sigma(x)\sqrt{2\pi}} \int_{x_i}^{x_{i+1}} \exp \left\{ \frac{-(E(x)-u)^2}{2 \cdot (\sigma(x))^2} \right\} du} \quad (4)$$

where the limits of integration  $x_1, x_2, \dots, x_{10}$  - it is boundaries of uniform intervals on which a histogram of occurrences of the data in the test sample.

Figure 2 shows plots of the histogram distribution of values of the chi-square test for the data obtained from the two program generators.



**Fig. 2.** Selecting data from a normal distribution values (dashed line) for the first hypothesis verification

Figure 2 shows that the comparator makes a decision on an input sequence should give a normal state of "1" in the range of 0 to 14. The switching threshold of the comparator in the state "0" 14. In this case, the probability of errors of the first and second kind are identical.  $P_1 = P_2 = P_{EE} = 0.054$ .

#### 4 Checking the second hypothesis of a uniform law of distribution of values for the final test sample

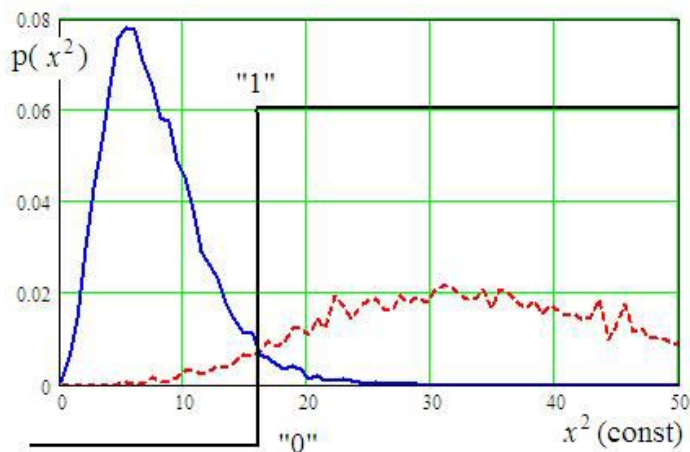
We start from the fact that the chi-square test is required to recognize the emergence of the situation data corresponding to a series of 81 samples obtained from the data generator with a uniform law - T2. For this purpose we find  $\max(x)$  and  $\min(x)$  in the test sample. Next, we will build a histogram consisting of  $k = 9 =$  columns evenly covering the range from  $\min(x)$  to  $\max(x)$ . The values of the chi-square test will be calculated as follows:

$$\chi^2(const) = 81 \cdot \sum_{i=1}^9 \frac{\left( \frac{b_i}{81} - \frac{1}{9} \right)^2}{\frac{1}{9}} \quad (5)$$

where the boundaries of the histogram intervals are as follows:

$$x_i = \min(x) + \frac{(\max(x) - \min(x)) \cdot i}{10} \tag{6}$$

Figure 3 shows plots of the histogram distribution of values of the chi-square test for the data obtained from the two program generators.



**Fig. 3.** Selecting data from a normal distribution values (dashed line) when testing the second hypothesis

Figure 3 shows that the comparator makes a decision on an input sequence should give a normal state of "1" in the range of 17 and above. The switching threshold of the comparator in the state "0" 16. In this case, the probability of errors of the first and second kind are the same.  $P_1 = P_2 = P_{EE} = 0.054$ .

### 5 Generalized chi-square test, taking into account the parallel test of two hypotheses

Chi-square test based verification under the first hypothesis (4) and the chi-square criterion based on test at the second hypothesis (5) - they are two different nonlinear transformation functions having two output comparator configured differently. Since these criteria are complementary, generalize them using logical functions "or":

$$\begin{cases} p(x) = \frac{1}{\sigma(x)\sqrt{2\pi}} \cdot \exp \left\{ \frac{-(E(x)-x)^2}{2 \cdot (\sigma(x))^2} \right\}, & \text{if } (\chi^2(\Phi) \leq 14) \wedge (\chi^2(const) \geq 16); \\ p(x) = const, & \text{if } (\chi^2(\Phi) \geq 14) \vee (\chi^2(const) \leq 16). \end{cases} \tag{7}$$

Practice has shown that the generalized decision rule (7), compared with a simple decision rules is a significant reduction in the probability of errors of the first and second order . That is, private chi-square test (4) and (5) have a high level of agreement making them the right decisions, and their wrong decisions are weakly correlated.

### 6 Conclusion

If we use the chi-square test according to standard procedures [1], then the test sample in 81 count, we obtain the probability of errors at the level of 0.054. However, as soon as we turn to the

accounting errors that occur due to the finiteness of the test sample (3) and apply the generalized chi-square test (7), the probability of error is reduced by about 20 times. This significant reduction in the probability of errors can only be achieved if the size of the test sample is 800 samples, this is equivalent to a 10-fold reduction in the requirements for the size of the test sample.

Should pay the same attention to the fact that this article focuses only chi-square, but all that is contained in the article, is true for other statistical tests in Table 1. By using the generalized criterion that takes into account 10 or 20 private statistical criteria, apparently be able to achieve more than 10-fold reduction in the requirements for the size of the test samples with statistical processing of biometric data.

## References

1. P 50.1.037-2002 Recommendations about standardization. Applied statistics. Rules of check of a consent of skilled distribution with the theoretical. Part I. Criteria like  $\chi^2$ . Gosstandart of Russia. Moscow-2001, 140 p.
2. P 50.1.037-2002 Applied statistics. Rules of check of a consent of skilled distribution with the theoretical. Part II. Nonparametric criteria. Gosstandart of Russia. Moscow-2002, 123 p.
3. Akhmetov B.S., Ivanov A.I., Funtikov V.A., Bezyaev A.V., Malygina E.A. Tekhnologiya of use of big neural networks for transformation of indistinct biometric data to an access key code. // The monograph, Kazakhstan, Almaty, LEM Publishing House LLP, 2014-144 with. (<http://portal.kazntu.kz/files/publicate/2014-06-27-11940.pdf>)
4. Akhmetov B.S., Volchikhin V.I., Ivanov A.I., Malygin A.Yu. Algorithms of testing of biometriko-neural network mechanisms of information security // Kazakhstan, Almaty, KAZNTU of Satpayev, 2013 - 152 p. ISBN 978-101-228-586-4, <http://portal.kazntu.kz/files/publicate/2014-01-04-11940.pdf>
5. Akhmetov B.S., Nadeev D. N., Funtikov V.A., Ivanov A.I., Malygin A.Yu. Otsenka of risks of highly reliable biometrics.//Monograph. - Almaty: Publishing house of KAZNTU of K.I. Satpayev, 2014 - 108 p.
6. Cochran W. G. Some Methods of Strengthening the Common  $\chi^2$  Tests // Biometrics, 1954. - V. 10. - P. 417-419.
7. Gilbert R.J A sample formula for cuterpolating tables of  $\chi^2$  // Biometrics, 1977. - V. 33. - P. 383-385.
8. Pearson E.S. Note on an approximation to the distribution of non-central  $\chi^2$  // Biometrics, 1959. - V. 46. - P. 364-366.
9. Ball R. M., Connell J. H., Pank anti-Sh., Ratkh N. K., Senior E.U. Rukovodstvo on biometrics. Moscow: Technosphere, 2007.-368 p, ISBN 978-594836-109-3
10. Akhmetov B.S., Ivanov A.I., Serikova N.I., Funtikova Yu.V. Algoritm of artificial increase of number of degrees of freedom in the analysis of biometric data on criterion of a consent a chi-square. Bulletin of national academy of Sciences of the Republic of Kazakhstan. No. 5, 2014 P. 28-34.
11. Akhmetov B.S., Ivanov A.I., Kachalin S.V., Seilova N.A., Doszhanova A.A. Addition fuzzy biometric data morphingre production examples of parents in several generations o examples descendants. Wulfenia Jornal vol 21, No. 7; jun 2014. Klagenfurt, Austria, ISSN: 1561-882x: [office@multidisciplinarywulfenia.org](mailto:office@multidisciplinarywulfenia.org)
12. Bakhytzhhan Akhmetov, Alexander Ivanov, Alexander Malyghin, Sergey Kachalin, Nurgul Seilova // Morph-Reproduction Examples of Parents in Several Generations of Examples Descendants//International Conference on Global Trends in Academic Research (ICMRP-December 17-18, 2014) at Kuala Lumpur, Malaysia
13. Malygin A.Yu., Volchikhin V.I., Ivanov A.I., Funtikov V.A. Fast algorithms of testing of neural network mechanisms of biometriko-cryptographic information security / Penza-2006, Publishing house of the Penza state university, 161 p.
14. Kobza player A.I. Applied mathematical statistics. For engineers and scientists. M.: FIZMATLIT, 2006, 816 p.

# Solving the Inverse Task of Neural Network Biometrics Without Mutations and Jenkins' "Nightmare" in the Implementation of Genetic Algorithms

Bakhytzhan Akhmetov<sup>1</sup>, Sergey Kachalin<sup>2</sup>, Alexander Ivanov<sup>2</sup>, Alexander Bezyaev<sup>2</sup>, and Kaiyrkhan Mukapil<sup>1</sup>

<sup>1</sup> K.I. Satpayev Kazakh National Technical University, Almaty, Kazakhstan

<sup>2</sup> Penza Research Electrotechnical Institute, Penza, Russia

bahitzhan@rambler.ru, s.kachalin@gmail.com, ivan@pniei.penza.ru, {bezyaev\underline{ }alex, kaiyrkhan}@mail.ru

**Abstract.** Estimated dimension of the field of possible input states biometric figures. It is shown that the implementation of genetic algorithms can be used two procedures reproduction data after their selection. Using conventional breeding morphing images, parents are less effective in comparison with directional morphing reproduction of figures-parents. Normal morphing creates Jenkins' "a nightmare leading to the degeneration of the correlation matrix biometric figures. Under the directed morph Jenkins' "nightmare" has no correlation matrices biometric figures do not degenerate after a few generations.

**Keywords:** biometrics, neural networks, genetic algorithms, reproduction of biometric figures, large database of test figures, the degeneration of populations.

## 1 Introduction

Currently in Russia, Belarus, Kazakhstan (in the countries of the Customs Union) and NATO countries are actively working to create a biometric authentication. Russia, Belarus and Kazakhstan are on track of using large artificial neural networks [1, 2], the United States and NATO countries follow the path of so-called "fuzzy extractors"[3, 4, 5, 6, 7]. Regardless of the used technology all converters biometrics code are described by equations of the same type and carry identical conversion function.

All converters of biometrics code should minimize entropy code examples belonging to the image of "Its". That is, the initial entropy of continuous biometric data must be reduced to almost zero at the converter output:

$$H(\nu_1, \nu_2, \dots, \nu_N) \gg H("c_1, c_2, \dots, c_n") \approx 0. \quad (1)$$

In contrast, if the input of the converter receives image data "Alien then the gain in entropy of continuous biometric parameters:

$$H(\xi_1, \xi_2, \dots, \xi_N) \ll H("x_1, x_2, \dots, x_n") \approx \frac{n}{10}. \quad (2)$$

Note that the input biometric data is used at least two or three times longer than the length output code, bio-code:

$$N \approx 3 \cdot n. \quad (3)$$

Input redundancy is necessary to compensate for the high value of the relative entropy of the input biometric parameters:

$$E(H(\nu_i/\xi_i)) \approx 0.3. \quad (4)$$

In all cases, when the relative entropy of figures data "Its" acceptable (good biometric data):

$$E(H(\nu_i/\xi_i)) \geq 1.0, \quad (5)$$

no need to the enrichment of neural network. In this case, works well "fuzzy extractors" with self-correcting codes having low redundancy.

Otherwise, when the following condition holds (4), there is no alternative to the use of artificial neural networks, because it is self-correcting codes are not able to detect and rule more than 30% of errors.

## 2 Assessment of multidimensional entropy codes "Aliens"

It should be emphasized that the one-dimensional calculation of the entropy of two-digit codes -  $H("x_1")$ , two digit codes  $H("x_1, x_2")$  or other short codes do not cause any difficulties. For this purpose, you can use the algorithm Shannon. A completely different situation arises when it is necessary to calculate the 256-dimensional entropy long bio-codes [1, 2]. In this case, the sample size of data needed to implement procedures Shannon is huge ( $2^{256+9}$  experiments). Even if we get such a substantial amount of test data on its bust will need tens of years of computer time (using ordinary computers).

Find the way out of the Shannon impasse is possible, if the statistical study of codes in conventional discrete spaces goes to their research in space Hamming distance between code "Its" codes and "Aliens" [8]:

$$"h - \sum_{i=1}^n ("x_i") \oplus ("c_i"). \quad (6)$$

For codes with equal probability states "0" and "1" in each bit " $x_i$ " and long distance  $n \geq 32$  distribution function of the values of the Hamming -  $p(h_j)$  is almost normal. This means that in order to estimate the probability of error of the second kind, it is sufficient to use a test set of about 32 randomly selected figures "Alien" and their response codes  $\bar{x}_j$ . On such a small test sample can be calculated expectation -  $E(h)$  and the standard deviation of the Hamming's distances  $\sigma(h)$ . Both of these values already is a continuum, as calculated by averaging discrete variables. Next, should use the hypothesis of normal distribution of the Hamming distance values and to estimate the probability of rare events (errors of the second kind, when the random "Alien" guesses the code of image of "Its":

$$P_2 \approx \frac{1}{\sigma(h)\sqrt{2\pi}} \int_0^1 \exp \left\{ \frac{-(E(h) - u)^2}{2(\sigma(h))^2} \right\} \cdot du. \quad (7)$$

Integration in the range from 0 to 1 in the expression (7) due to the fact that the discrete distribution of Hamming distances -  $p(h_j)$  is considered as a continuous and entering the event in this interval corresponds to a discrete state " $h - "0"$ " (coincidence of the compared codes in all categories).

Then, to assess the multidimensional entropy use the following conversion:

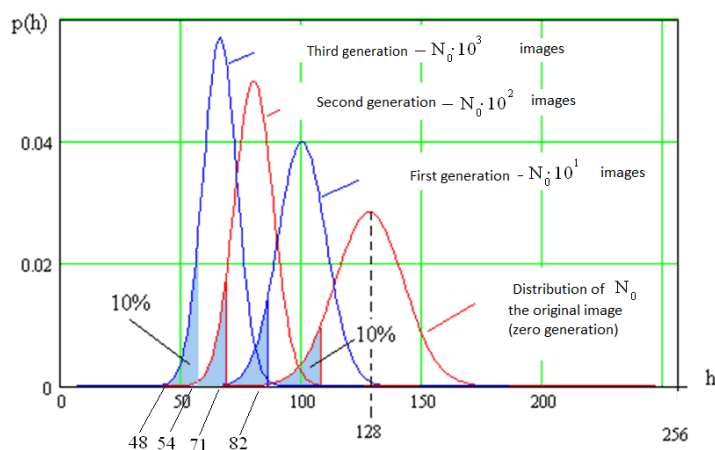
$$H("x_1, x_2, x_3, \dots, x_{256}") \approx -\log_2(P_2). \quad (8)$$

## 3 The matrix inversion of neural network with a known functional code "Its"

Note that calculation chain (6), (7) is an approximation, and has a certain systematic error. In this connection must be confirmed by the estimates obtained. For that reason, [9] in addition to the

transformations (6), (7) recommends the use of a genetic algorithm of neural network functional matrix inversion. In fact, it is about using database consisting of 1024 figures "Alien" formed by the requirements of GOST R 52633.1 [10].

If you consistently apply  $N_0 = 1024$  figure of the "Alien" to test biometrics converter code, then getting the response code 1024. Hamming distance distribution in this case will be  $E(h) = 128$  bits to 256 bits long codes. If then select 10% of the sample (102) biometric figures with the lowest values of the Hamming distance, giving codes with Hamming distance of 82 bits to 112 bits, then getting the original genetic material for the next generation biometric figures. This situation is displayed in the center of Figure 1.



**Fig. 1.** A genetic algorithm for extracting knowledge from trained artificial neural network with 416 inputs and 256 outputs.

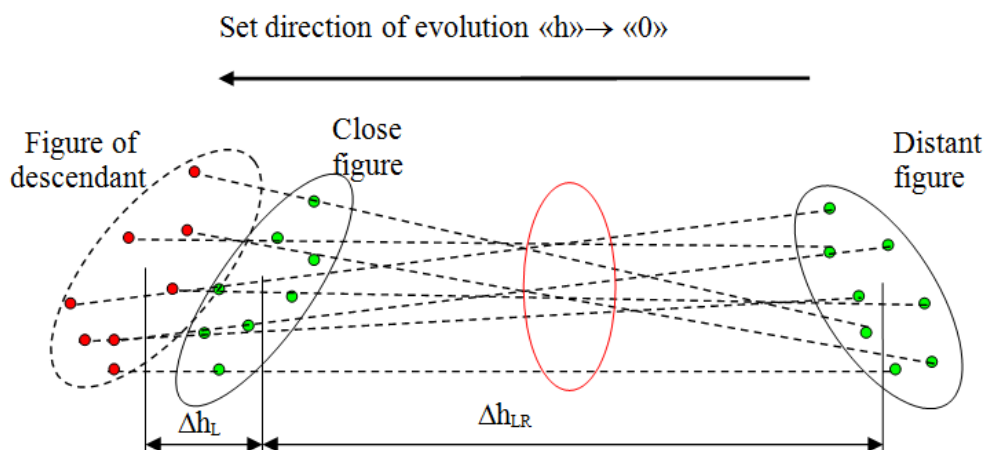
In order to restore the population size from 102 to 1024 must cross between an figures-parents and receive figures-descendant. For example, this can be done by simply averaging the parameters of the two figures parent in accordance with the requirements of GOST R 52633.2 [11]. Experience shows that in such a procedure births a new generation of figures "Aliens" is more similar to the image of "Its" with the expectation of Hamming distances close to the expectation of the Hamming distances previously bleached 102 biometric figures. This situation is displayed in Figure 1 and corresponds to the first generation of synthetic figures "Alien".

Obviously, the selection process described above most similar to the image of "Its" figures "Aliens" can be run multiple times [12]. As a result, we obtain a sequence of iterations increasingly brings us to the figure of "Its". In fact, there is a procedure for extracting knowledge from the training parameters of artificial neural network, which allows to recover the distribution of image data "Its" 96%.

However restored biometric figure of "Its" is defective, it has a degenerate correlation matrix with close to zero elements off the diagonal. This happens due to the fact that when using the reproduction of figures-averaging parent used their data (negative correlation compensate positive correlations) [13, 14].

#### 4 Compensation of degeneration defect data using directional morphing extrapolation to a given direction of motion in advance of evolution

Degeneracy effects of the correlations is illustrated in Figure 2. In this fig., close and distant figure (right and left side of the fig.) have different slopes along the main axis of the ellipse, respectively, averaging their data gives an ellipse without tilting. That is, the correlation coefficients disappeared (compensated for each other).



**Fig. 2.** Getting the figures-descendant, morphing extrapolation to a given side of the directed evolution.

However, if we know the direction of evolution, figure-child can be located just to the left of the closest to the target figure. It is sufficient to carry out the following transformation:

$$\psi_i = \xi_{\nu,i} - \left( \frac{\Delta h_L}{\Delta h_{LR}} \right) \cdot \xi_{\xi,i}, \quad (9)$$

where  $\xi_{\nu,i}$  - the value of the  $i$ -th example of biometric parameter of close figure,  $\xi_{\xi,i}$  is the value of the  $i$ -th example biometric parameter of distant figure,  $\Delta h_L$  - the desired Hamming distance to the left image-descendants,  $\Delta h_{LR}$  Hamming distance between left and right examples of different parents figures.

Figure 2 shows that the most left ellipse of data of figure- descendant has almost the same slope as the closest figure-parent. During the multiplication of directional morphing extrapolation occurs inheritance correlation matrices descendants.

#### 5 Conclusion

In the 19th century Jenkin Darwin noticed the fact that random mutations do not accumulate. The genotype of the white man came to the island to the natives should be dissolved. All the fault is a lot of random mutations large population of Aboriginal people.

In the process considered by us genes of the white man does not dissolve, if you constantly get in the selected sample of evolution. Moreover, for each reproduction of biometric figures in the next generation inherits not only the correlation matrix of the white man, but with each new generation of "white" people getting whiter and whiter, if required by the selected vector of evolution.



The use of directional morphing extrapolation is profitable formerly used of reproduction data of morph-interpolation.

## References

1. Volchikhin, V.I., Ivanov, A.I., Funtikov, V.A., Nazarov, I.G.: Neural Protection of Personal Biometric Data. In: Yazov, J.K. (eds.) M.: Radio, p. 160 (2012)
2. Akhmetov, B.S., Ivanov, A.I., Funtikov, V.A., Bezyaev, A.V., Malygina, E.A.: The Technology of Using Large Neural Networks for Fuzzy Transformation of Biometric Data in the Key Code Access (Monograph). In: Kazakh National Technical University. Almaty, Kazakhstan (2014)
3. Juels, A., Wattenberg, M.: A Fuzzy Commitment Scheme. In: Proceedings of the 6th ACM conference on Computer and communications security, New York, NY, USA, pp. 28–36 (1999)
4. Monrose, F., Reiter, M., Li, Q., Wetzel, S.: Cryptographic Key Generation from Voice. In: Proceedings of the 2001 IEEE Symposium on Security and Privacy, IEEE Computer Society Washington, DC, USA (2001)
5. Dodis Y., Reyzin, L., Smith, A.: Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy. In: cryptology-Eurocrypt, pp. 523–540 (2004)
6. Lee, Y.J., Bae, K., Lee, S.J., Park, K.R., Kim, J.: Biometric Key Binding: Fuzzy Vault Based on Iris Images. In: Proceedings of 2nd International Conference on Biometrics, Seoul, South Korea, pp. 800–808 (2007)
7. Chmorra, A.L.: Masking Key Using Biometrics "Problems of Information Transmission". In: Probl. Peredachi Inf., vol. 47, Issue 2, pp. 128–143 (2011)
8. Akhmetov, B., Ivanov, A., Funtikov, V., Urnev, I.: Evaluation of Multidimensional Entropy on Short Strings of Biometric Codes with Dependent Bits. In: PIERS (Progress in Electromagnetics Research Symposium) Proceedings, Moscow, Russia, pp. 66–69 (2012)
9. GOST R 52633.3-2011 "Information Security. Security Technique. Testing Resistance Means Highly Reliable Biometric Security to Attacks Matching"(2011)
10. GOST R 52633.1-2009 "Information Security. Security Technique. Requirements for the Formation of the Natural Bases of Biometric Images Designed to Test Highly Reliable Means of Biometric Authentication"(2009)
11. GOST R 52633.2-2010 "Information Security. Security Technique. Requirements for the Formation of Synthetic Biometric Images Designed to Test Highly Reliable Means of Biometric Authentication"(2010)
12. Akhmetov, B.S., Ivanov, A.I., Bezyaev, A.V., Kachalin, S.V.: Evaluation of the Computational Complexity of Matrix Inversion Neural Network Functional. In: Reports of the National Academy of Sciences of the Republic of Kazakhstan, No. 5, pp. 49–61 (2014)
13. Akhmetov, B., Ivanov, A., Malyghina, E., Kachalin, S., Serikova, N.: Assessing the Level of Uncertainty of small samples of Multidimensional Biological and Biometric Data. In: International Journal of Engineering Sciences and Research Technology, vol. 7, Issue 3, pp. 284–288 (2014)
14. Akhmetov, B., Ivanov, A., Malyghin, A., Kachalin, S., Seilova, N.: Morph-Reproduction examples of Parents in Several Generations of examples Descendants. In: ICMRP-2014, Kuala Lumpur, Malaysia (2014).

# Module of Lexical and Morphological Analyzer in the Development of Semantic Search Engine for Kazakh Language

Y.N. Amirgaliyev<sup>1</sup>, A.S. Kalimoldayeva<sup>1</sup>

<sup>1</sup> Institute of Information and Computing Technologies MES RK, Almaty, Kazakhstan  
kalimainura@gmail.com, amir\\_ed@mail.ru

**Abstract.** Search engine technology has had to scale sharply to keep up with the growth of the World Wide Web. Search engines link huge amount of web pages involving a proportionate number of distinct terms. They answer to thousands of queries every minute. Despite the importance of massive search engines on the web, very little academic research has been done on them. Moreover, developing a web search engine today is very different from several years ago, because the latest technologies appear every single day [1]. This paper provides description of the module of lexical and morphological analyzer in the development of semantic search engine for kazakh language that is implemented in the Institute of Information and Computer Technologies (IICT) and our main goal is to improve the quality of web search engines by using module for text processing in the Kazakh language and receiving the module of the text with morphological markings in the output.

**Keywords:** semantic web, textual data analysis, search engines, lexical and morphological analyzer Kazakh language, information retrieval.

## 1 Introduction

Nowadays the main driver of innovation progress is a World Wide Web and one of the fastest growing industries is Internet commerce. With the wide range of capabilities the web cost efficient for businesses to make transactions with other businesses. One factor that allows businesses to find each other is search engines. Search engines are part of the reason the web is growing so rapidly. Search engines have many capabilities from using key words or phrases to find what the user is looking for to using general words to browse the web [2]. However what exactly is a search engine? Search engine is an enormous database of web page files that have been assembled automatically by computer.

However, the concept of "information retrieval" came before the Web. It has developed from the numerous issues related to the provision of search and access to information sources. As a matter of fact, basically information retrieval (IR) were applied to scientific publications and to variety of library catalogs, but later it spread to other areas where role of information were vital as well [3]. Much of the scientific research on information retrieval has occurred in these contexts, and much of the continued practice of information retrieval deals with providing access to unstructured information in various corporate and governmental domains.

Internet opened the opportunity to publish tens of millions of information. This explosion of published information would be moot if the information could not be found, annotated and analyzed so that each user can quickly find information that is both relevant and comprehensive for their needs and it would help to save user's time [4]. By the time, developers felt that continuing to index the whole Web would rapidly become inaccessible, owing to the Web's exponential growth in size. However a lot of scientific innovations, the rapidly declining price of computer hardware, superb engineering, and the growth of a commercial underpinning for web search have all conspired to power today's main search engines, which are able to provide excellent results within incessant response times for hundreds of millions of searches a day over billions of web pages [6].

## 2 Module of lexical and morphological analyzer in the development of semantic search engine for kazakh language

The Institute of Information and Computer Technologies (IICT) MES RK, has started to study semantic search issues in 2012 for the first time. Studies have been conducted in the field of semantic web, there were researched search engines, information retrieval, and the specifics of the Kazakh language [6],[7],[8] for the future use and the following results were obtained:

- lexical dictionary;
- dictionary of affixes;
- lexical and morphological analyzer;
- subject classifier;
- search engine;
- and lastly, there were decided to implement them all as a web service.

In the future we plan to add multi-language function (Russian and Turkic languages group). Technology architecture of Web services allows maintaining hardware and software scalability. The hardware part will maintain the vertical and horizontal scalability. Software technology suggests the possibility of using it as both a library of algorithms; as scalable cloud service.

Thus, the web service can serve as a basis for creating different applications that require text mining [10], including:

- monitoring of news and analysis of the information extracted from the traditional periodicals and the Internet;
- analysis of the message tonality
- analysis of the message meaning
- analysis of opinions in social networks and reputation monitoring.

Unlike existing similar projects for processing and analysis of texts, there were developed an original lexical and morphological algorithms based on the Kazakh language, moreover, the inventor's certificate was obtained [11]. This technology may be expanded to other Turkic languages (Kyrgyz, Tatar, Turkish, Azerbaijani, Turkish, Bashkir, etc.) [12].

According to Wikipedia, the population of the Turkic languages groups is approximately 150 million people [13]. The system is implemented as a Web site based on ASPNET technology, in C# programming language, MSSQLServer 2008 was used as the database.

The Web site provides the following services: morphological dictionary, lexical and morphological parser of Kazakh language, an intelligent search of texts from the knowledge base Wikipedia, and the spell checker system.

### 2.1 Search organization

Information retrieval is implemented by using the same methods as search engines. Mostly they are vector model of search or model of a weighted Boolean. A search query consists of a set of keywords. This set can be obtained by using a method of classification graph model. On the basis of the proposed model there were created retrieval query to a search engine. Then, by using the method of calculating of semantic similarity like from text to request, we can obtain a set of documents necessary ranked in descending order.

Thus, by using the graph model of text representation we can perform the model search of semantic text.

## 2.2 Preparation of the Wikipedia knowledge base

As a source of knowledge base system uses encyclopedia of Wikipedia: every Wikipedia article is considered as a concept; each article title and text of a hyperlink to the article is considered to be a term, and the words of the text describing the article, to be used for the topical search system. File Wikipedia in Kazakh language we downloaded from the Internet. The file itself is to Wikipedia is a text file that contains various tags. To build the knowledge base by extracting the titles of articles, texts of articles subject to the rules markup software module was developed WikiParser [14].

Currently, knowledge base stored in the database MSSQLServer, which contains 275 000 articles, more than 17 thousand. Partitions (thematic classes) and more than 20 million words.

## 2.3 Module of lexical and morphological analysis

This module is created for text processing in the Kazakh language and receiving the module of the text with morphological markings in the output. The module contains of two main subsystems, they are lexical analysis and morphological analysis system.

Lexical analysis parses text into separate paragraphs, sentences, words, and assigning each word to lexical descriptors (word with a capital letter, all capital letters, there is a point at the end of a word or it is a separate letter, number, etc.). Morphological analysis is used to find the normal form words and recognize the parts of speech of each word of the text (the grammatical form of the word, the number of nouns, tense of the verb etc.). To perform this task, we used our development that we created last year, they are lexical and morphological analyzer for Kazakh language. Today we continue our project to improve lexical and morphological analyzer. Since the texts of the Internet are "information dump" and recurring task to improve the algorithm for extracting words from the text and updating of the dictionary with new words.

We improved this module by recognition system where we included surnames. The system uses a database with over 65 thousands of Russian and Kazakh surnames and names. But sometimes in the texts of the internet database you can notice new names that are not found in our database [15]. For automatic recognition of unfamiliar names, we have developed an algorithm for recognizing names. This algorithm consists of two rules: 1) unfamiliar word begins with a capital letter, 2) it contains the word syllables included in the list of syllables that are usually included in the end of the names.

The module was developed in the programming language C#, in a visual programming environment VisualStudio 2010.

The module consists of the following classes for working with text data: - DocumentConverter - class for loading documents of various formats and unloading [16]; - IWord - class for the storage and presentation of information elements (words); - Tokenizer - a class that converts text into a list of tokens from the lexical descriptors (the lexical analyzer); - HtmlParser - class for parsinghtml document and the work of this meta-tags; - MorfoAnalyze - class responsible for morphological analysis. - RemoveOmonim - class responsible for disambiguation. - LoadBaze - class responsible for loading dictionaries and affixes from XMLfaylov in structure. All dictionaries Dictionary of 85 thousand [17]. Words dictionary of names of 65 th., 5500 gazetteer loaded into memory. The storage format of dictionaries is given below.

```
<? Xml version = "1.0"encoding = "utf-16"?>
<! - Slovarobschihslov ->
<Slovar>
<Word Value = "and"Part = "interjection"Semantic = />
```

```

<Word Value = "aba"Part = "Nouns"Semantic = />
<Word Value = "aba"Part = "Nouns"Semantic = />
<Word Value = "Abadan"Part = "Appendix"Semantic = />
<Word Value = "Abadan"Part = "Nouns"Semantic = />
.....
<Word Value = "jasper"Part = "Nouns"Semantic = />
</ Slovar>

```

### 3 Conclusion

Overwhelmingly, most of the search engines that are on the web today have many differences. However, all of the search engines vary in speed, size, and content. There are not two search engines uses the exact same ranking schemes. Each has its own method of ranking its search results. Another difference between search engines is not every search engine offers you exactly the same search options. Rankings of search engines all use different methods. Their goal is to return the most relevant pages at the top of their lists. To do this, they look for the location and frequency of keywords and phrases in the web page document and occasionally in the html Meta tags.

This paper provided description of the module of lexical and morphological analyzer in the development of semantic search engine for kazakh language that is implemented in the Institute of Information and Computer Technologies (ИИСТ) and our main goal is to improve the quality of web search engines by using module for text processing in the Kazakh language and receiving the module of the text with morphological markings in the output.

### References

1. Кан Д. А. Применение теории компьютерной семантики русского языка и статистических методов к построению системы машинного перевода Диссертация // Санкт-Петербург 2011, с.131.
2. Дрейзин Ф.А Об алгоритмизации составления алгоритма анализа языка //Математика Наука вып.189 с176
3. Oflazer, K. Two-level description of Turkish morphology // Literary and Linguistic Computing. 1994. – Vol.9, Is.2. – P.137-148.
4. Washington J. N., Salimzyanov I., Tyers F.M. Finite-state morphological transducers for three Kypchak languages // Proceedings of the 9th Conference on Language Resources and Evaluation, LREC 2014. – Granada, 2014.
5. Salimzyanov I., Washington J.N., Tyers F.M. A free/open-source Kazakh- Tatar machine translation system // Proceedings of the XIV Machine Translation Summit. – Nice, 2013. – P.175-182.
6. Бектаев К. Статистико-информационная типология тюркского текста 1978.-87с. 109
7. Бектаев К. Теория вероятностей и математическая статистика на казахском языке 1990.-124с.
8. Бектаев К. и др. Математические методы в языкознании 1974.- 170с.
9. Бектаев К. и др. Математическая лингвистика в соавторстве 1977.- 85с.
10. Бектаев К. Большой казахско-русский русско-казахский словарь. – Алматы: «Алтын Казына», 1999. – 704с.
11. Jonathan North Washington "A Novel Approach to Delineating Kazakh's Five Present Tenses: Lexical Aspect".
12. Altenbek and Wang Xiao-long "Kazakh Segmentation System of Inflectional Affixes", 2010r.
13. Sharipbaev A.A., Bekmanova G.T., Ergesh B.J., Buribaeva A.K., Karabalaeva M.H."The intellectual morphological analyzer based on semantic network".-2012.
14. Амиргалиев Б.Е., Амиргалиев Е.Н. Методы анализа речевого сигнала в системах синтез речи и распознавания // Материалы международной конференции «Современные проблемы математики, информатики и управления», посвященной 60-летию академика МАИ, д.ф.-м.н., профессора М.Б. Айдарханова. – Алматы, 2008. – С. 50-53.
15. Мусабаев Р.Р., Амиргалиев Б.Е. Вопросы разработки информационной системы синтеза и распознавания речи //Труды межд. научно- практической конференции «Информационно-инновационные технологии: интеграция науки, образования и бизнеса». – Алматы: КазНТУ, 2008.– С.151- 158.

16. Шарипбаев А.А. Распознавание и синтез казахской речи 25-28
17. Амиргалиев Б.Е., Мусабаев Р.Р. и др. Унифицированный язык фонетического представления для систем и распознавания речевого сигнала // Материалы международной научно-практической конференции «Актуальные проблемы математики, информатики, механики и теории управления». – Алматы: ИПИУ-КБТУ, 2009. – Ч. 1. – С. 55-57.

# Recognition of Isolated Words Using the Bayes' Theorem

E.N.Amirgaliyev, O.J. Mamyrbayev, T.A.Muratkhanova

Institute of Information and Computational Technologies of MES RK, Pushkin str.125, Almaty, Kazakhstan  
tolganay125@gmail.com

**Abstract.** This paper outlines mathematical solution of building up the speech recognition systems by creating probability conditions and defining percent of made mistakes during the operation, based on-Bayer's theorem. Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format, by digitizing the sound and matching its pattern against the stored patterns. Now-a-day, in the sphere of education there are great deal of useful software programs, one of them is speech recognition system, the benefits of speech recognition applications are numerous. Creating documents, memos, and reports can easily be spoken leaving hands free to complete other tasks. Phone calls can be made via voice over IP software applications, which rely on the ability of the computer to convert voice into data. But unfortunately, many students in our country, do not have feasibility to use such kind of applications, because existing tools do not recognize and process Kazakh, Russian languages, even if some of them recognizes Russian language, they are too expensive to download it.

In the 1990's, a number of innovations took place in the field of speech recognition. The problem of speech recognition, which traditionally followed the framework of Bayes theorem and required estimation of distributions for the data, was transformed into an optimization problem involving minimization of the empirical recognition error. This fundamental change of paradigm was caused by the recognition of the fact that the distribution functions for the speech signal could not be accurately chosen or defined, and that Bayes' decision theory would become inapplicable under these circumstances. After all, the objective of a recognizer design should be to achieve the least recognition error rather than the best fitting of a distribution function to the given (known) data set as advocated by the Bayes' criterion. The method consists of three stages:

1. Estimation of the joint conditional probability  $P(o_1, \dots, o_j; w_i)$
2. Description of processing of incoming data
3. Facilitating the construction of composite models.

For methods of handling voice and video signals and also storing them, performing a relatively simple task, when testing it is defined preliminary parameters of voice and video signals and degree of stability in the working process towards to variations of input signals. For recognition system, working in real conditions and on real data, when testing it is defined percent of made mistakes, i.e. is assessed not the behavior under different conditions, it assessed how effectively system solves its task set before it.

**Keywords:** Speech recognition, voice active detection, segmentation.

## 1 Introduction

Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format, by digitizing the sound and matching its pattern against the stored patterns. Currently available devices are largely speaker-dependent (recognize speech of only one or two persons) and can recognize discrete speech (speech with pauses between words) better than the normal (continuous) speech. Their major applications help people in working around their disabilities. It should not be confused with voice recognition which is used mainly in security devices.

Speech recognition software is fundamental in many fields. For example, in educational process, we see that speech recognition is helping to enhance the educational process for students and teachers alike. It has been shown to improve core reading and writing skills for students of all abilities, including those with physical or language-based learning disabilities, as well as English

Language Learners. Software lets students dictate papers and assignments three times faster than typing — with up to 99% accuracy. It is also widely used in secondary business education and computer applications courses to familiarize students with emerging interfaces. Plus, many teachers rely on the software as a productivity tool to help them manage their overwhelming administrative workload. With all these advantages, it's no wonder that Software is being used in more and more elementary, secondary, and post-secondary school today — with great results. Rudimentary speech recognition software has a limited vocabulary of words and phrases and may only identify these if they are spoken very clearly. More sophisticated software has the ability to accept natural speech.

Designing a machine that mimics human behavior, particularly the capability of speaking naturally and responding properly to spoken language, has intrigued engineers and scientists for centuries. Since the 1930s, when Homer Dudley of Bell Laboratories proposed a system model for speech analysis and synthesis [1,2], the problem of speech recognition has been approached progressively, from a simple machine that responds to a small set of sounds to a sophisticated system that responds to fluently spoken natural language and takes into account the varying statistics of the language in which the speech is produced. Early recognition systems of the 1950's, Olson and Belar of RCA Laboratories built a system to recognize 10 syllables of a single talker [3] and at MIT Lincoln Lab, Forgie built a speaker-independent 10-vowel recognizer [4]. In the 1960's, several Japanese laboratories demonstrated their capability of building special purpose hardware to perform a speech recognition task. Most notable were the vowel recognizer of Suzuki and Nakata at the Radio Research Lab in Tokyo [5], the phoneme recognizer of Sakai and Doshita at Kyoto University [6]. The work of Sakai and Doshita involved the first use of a speech segmented for analysis and recognition of speech in different portions of the input utterance. In contrast, an isolated digit recognizer implicitly assumed that the unknown utterance contained a complete digit (and no other speech sounds or words) and thus did not need an explicit "segmenter." Kyoto University's work could be considered a precursor to a continuous speech recognition system.

In the XXI century one of the most important task for improving speech recognition is to develop and work on it. Now-a-days, we know that England, Japan, USA etc. are doing their best to create the finest recognizer. In Kazakhstan a great deal of researchers are working on SR., group of researchers from the—"Institute of Information and Computing Technologies" Kalimoldaev M.N., Amirgaliev E.N., Musabaev R.R. and Koybagarov K.C. share working on speech recognition since 1991th, and it might be well pointed out that Institute reached good results in the speech recognition sphere as developing speech synthesis (with a focus on the Kazakh language), creating: searching system on the basis of the Kazakh Wikipedia; The system of thematic analysis of textual information; morphological analyzer Kazakh word forms; assembled Dictionary of Kazakh words; As for another researchers for example Sharipbayev A.A. with his disciples developed a mathematical theory of the Kazakh language, where justified and formalized phonetic regularities, built algorithms for analysis and synthesis of words and sentences, and speech recognition algorithms. Developed a mathematical model of grammar rules of the Kazakh language for automation of analysis and synthesis of written and spoken language technology.

Now-a-day, in the sphere of education there are great deal of useful software programs for students, teachers etc. for everyone which can facilitate the training process. One of useful software programs (tools) is speech recognition system, the benefits of speech recognition applications are numerous. Creating documents, memos, and reports can easily be spoken leaving hands free to complete other tasks. Phone calls can be made via voice over IP software applications, which rely on the ability of the computer to convert voice into data. But, students of Kazakhstan cannot use speech recognition tools, because firstly, they are too expensive to download it; secondly, the existing tools do not recognize and process Kazakh, Russian languages.



If we had our own speech recognition tools developed using Bayes theorem we could solve these problems easily.

In many cases, it breaks down to a simple case of mathematics. The average office worker types between 50-70 words per minute. However, speech recognition programs can handle 120 words per minute at 98% accuracy with proper training, even editing can be accomplished through the use of voice activated programming. This makes the use of these types of software applications a way to increase productivity as well as giving individuals with disabilities new ways to interact and work.

### 1.1 Mathematical calculations using the Bayes' Theorem for speech recognition

In the 1990's, a number of innovations took place in the field of pattern recognition. The problem of pattern recognition, which traditionally followed the framework of Bayes theorem and required estimation of distributions for the data, was transformed into an optimization problem involving minimization of the empirical recognition error [7]. This fundamental change of paradigm was caused by the recognition of the fact that the distribution functions for the speech signal could not be accurately chosen or defined, and that Bayes' decision theory would become inapplicable under these circumstances. After all, the objective of a recognizer design should be to achieve the least recognition error rather than the best fitting of a distribution function to the given (known) data set as advocated by the Bayes criterion. The method consists of three stages:

1. Estimation of the joint conditional probability  $P(o_1, \dots, o_j; w_i)$
2. Description of processing of incoming data
3. Facilitating the construction of composite models

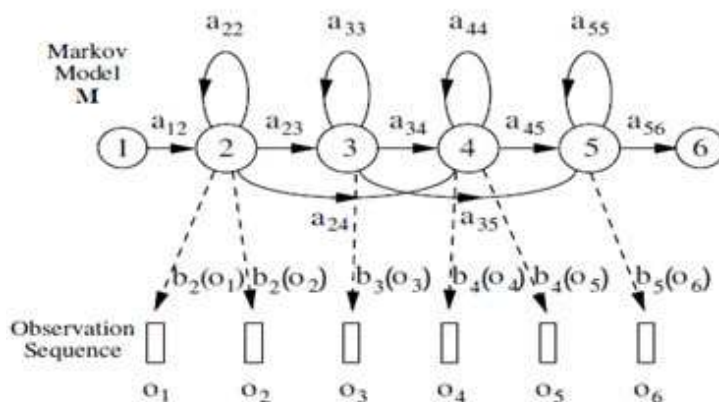


Fig. 1. The hidden Markov model.

A hidden Markov model is a tool for representing probability distributions over sequences of observations. Hidden Markov model falls in subclass of Bayesian networks, for modeling time series data. In time series modeling, the assumption that an event can cause another event in the future, but not vice-versa, simplifies the design of the Bayesian network: directed arcs should flow forward in time. Let each spoken word be represented by a sequence of speech vectors or *observations*  $O$ , defined as

$$O = o_1, o_2, \dots, o_t \quad (1)$$

where  $o_t$  is the speech vector observed at time  $t$ . The isolated word recognition problem can then be regarded as that of computing using Bayes' Rule gives

$$P(w_i|O) = \frac{P(O|w_i)P(w_i)}{P(O)} \quad (2)$$

Thus, for a given set of prior probabilities  $P(w_i)$ , ( $w_i$  is the  $i$ 'th vocabulary word) the most probable spoken word depends only on the probability  $P(O_j|w_i)$ . Given the dimensionality of the observation sequence  $O$ , the direct estimation of the joint conditional probability  $P(o_1; o_2, \dots, o_j|w_i)$  from examples of spoken words [7].

## 1.2 Description of processing of incoming data

In HMM based speech recognition, it is assumed that the sequence of observed speech vectors corresponding to each word is generated by a Markov model as shown in Fig. 3 A Markov model is a finite state machine which changes state once every time unit and each time  $t$  that a state  $j$  is entered, a speech vector  $o_t$  is generated from the probability density  $b_j(o_t)$ . Furthermore, the transition from state  $i$  to state  $j$  is also probabilistic and is governed by the discrete probability  $a_{ij}$ . Fig. 1 shows an example of this process where the six state model moves through the state sequence  $X = 1; 2; 2; 4; 5; 6$  in order to generate the sequence  $o_1$  to  $o_6$ . Notice that in HTK, the entry and exit states of a HMM are non-emitting.

## 1.3 Facilitating the construction of composite models

The joint probability that  $O$  is generated by the model  $M$  moving through the state sequence  $X$  is calculated simply as the product of the transition probabilities and the output probabilities.

The Bayes' theory is a graphical model for representing conditional independencies between a set of random variables. Consider four states  $a_{11}, a_{12}b_2(o_1), a_{22}b_2(o_2), a_{23}b_3(o_3)$ . From basic probability theory we can factor the joint probability as a product of conditional probabilities:

So for the state sequence  $X$  in Fig. 1

$$P(O; X_j M) = P[a_{11}, a_{12}b_2(o_1), a_{22}b_2(o_2), a_{23}b_3(o_3) \dots] \quad (3)$$

$$P[a_{11}, a_{12}b_2(o_1), a_{22}b_2(o_2), a_{23}b_3(o_3) \dots] = P[a_{11}] * P[a_{12}b_2(o_1)|a_{11}] * \\ * P[a_{22}b_2(o_2)|a_{11}, a_{12}b_2(o_1)] * P[a_{23}b_3(o_3)|a_{11}, a_{12}b_2(o_1), a_{22}b_2(o_2)] \quad (4)$$

This factorization does not tell us anything useful about the joint probability distribution: each variable can potentially depend on every other variable. However, consider the following factorization:

$$P(O; X_j M) = P[a_{11}] * P[a_{12}b_2(o_1)] * P[a_{22}b_2(o_2)|a_{11}] * P[a_{23}b_3(o_3)|a_{11}, a_{12}b_2(o_1)] \quad (5)$$

The above factorization implies a set of conditional independence relations. For example variable (or set of variables)  $A$  is conditionally independent from  $B$  given if  $P[A, B, C] = P[A, C]P[B, C]$  for all  $A, B$  and  $C$ . From above factorization we can show that given the values of  $a_{11}$  and  $a_{12}b_2(o_1), a_{22}b_2(o_2)$  and  $a_{23}b_3(o_3)$  are independent:

$$P[a_{11}, a_{12}b_2(o_1)|a_{22}b_2(o_2), a_{23}b_3(o_3)] =$$

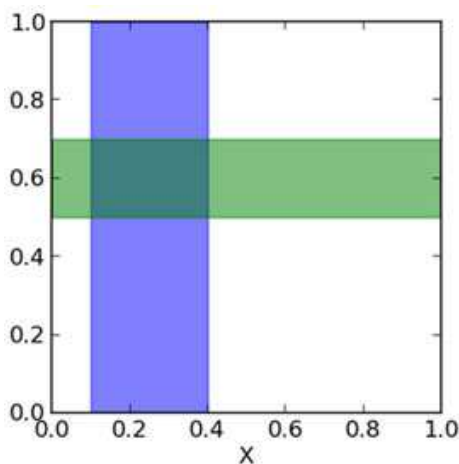
$$\begin{aligned}
 &= \frac{P[a_{11}] * P[a_{12}b_2(o_1)] * P[a_{22}b_2(o_2)|a_{11}] * P[a_{23}b_3(o_3)|a_{11}, a_{12}b_2(o_1)]}{P(a_{22}b_2(o_2), a_{23}b_3(o_3))} = \\
 &= \frac{P[a_{11}] * P[a_{22}b_2(o_2)|a_{11}] * P[a_{23}b_3(o_3)|a_{11}, a_{12}b_2(o_1)] * P[a_{22}b_2(o_2), a_{23}b_3(o_3)]}{P[a_{23}b_3(o_3)]} = \\
 &= P[a_{22}b_2(o_2)|a_{11}] * P[a_{23}b_3(o_3)|a_{11}, a_{12}b_2(o_1), a_{22}b_2(o_2)] \tag{6}
 \end{aligned}$$

In practice, only the observation sequence  $O$  is known and the underlying state sequence  $X$  is hidden. This is why it is called a *Hidden Markov Model*.

By creating of probability conditions based on-Bayer’s theorem, we offer a mathematical solution for building up the speech recognition.

**1.4 The results of recognition of isolated words.**

Suppose sentence is a kind of interval, in this interval numbers i.e. words scattered out of order, as we know that interval-is a set of numbers consisting of all the numbers between a pair of given numbers along with either, both, or none of the endpoints. Based on this it turns out that if words-numbers are not arranged in order it will not be considered as interval as well as the sentence.



**Fig. 2.** Probability of a sequence.

Now assume that any word for example “sunny” is- $y$  is in interval  $[0.5,0.7]$  (see fig.2) [today, rejoiced] in the set there such words as {today, us, tomorrow, late, warm, weather, room, study, spring, rejoiced etc} as we know initial and end boundaries are “today” and “rejoiced”, now we need to determine what is the probability that the word “sunny” will be the second word in sentence, and word “weather” third.

Suppose we want to know what is the probability that  $y$ -i.e. “weather” lies in the interval  $[0.5, 0.7]$ , if  $x$  is already in the interval  $[0.1, 0.4]$  (see fig.2). That is in actual, we have the filter and when we call the pair  $(x, y)$ , we immediately discards those pairs which do not satisfy the condition of finding  $x$  in a given interval, and then filtered pairs we consider those for which  $y$  satisfies our condition and consider the probability as the ratio of the number of pairs for

which  $y$  lies above the interval to the total number of filtered pairs (i.e. for which  $x$  lies in the interval  $([0.1, 0.4])$  i.e. [today,rejoiced]. We can write this probability as  $p(Y|X)$ . Obviously, this probability is equal to the ratio of the area of the dark area (the intersection of the green and blue areas) to the area of the blue region. The area of the dark area is equal to  $(0.4 - 0.1) * (0.7 - 0.5) = 0.06$ , and the area of the blue  $(0.4 - 0.1) * (1 - 0) = 0.3$ , then their ratio is  $0.06 / 0.3 = 0.2$ . In other words, the probability of finding  $y$  in the interval  $[0.5, 0.7]$  despite the fact that  $x$  belongs to the segment  $[0.1, 0.4]$  is equal to  $p(Y|X) = 0.2$ . You can see that taking into account all of the above and all above notation, we can write the following expression

$$p(Y, X) = \frac{p(X, Y)}{p(X)} \quad (7)$$

Briefly play all the previous logic now with respect to  $p(X, Y)$ : we call a pair  $(x, y)$  and the filtered those for which  $y$  lies between 0.5 and 0.7, then the probability that  $x$  lies in the interval  $[0.1, 0.4]$ , provided that  $y$  belongs to the segment  $[0.5, 0.7]$  is equal to the ratio of the area of the dark area to green space:

$$p(X|Y) = \frac{p(X, Y)}{p(Y)} \quad (8)$$

In the two above formulas, we see that the member of  $p(X, Y)$  are the same, and we can exclude:

$$p(X, Y) = p(X|Y) * p(Y) = p(Y|X) * p(X) \quad (9)$$

$$p(X | Y) = \frac{p(Y|X)*p(X)}{p(Y)} \quad (10)$$

For methods of handling voice and video signals and also storing them, performing a relatively simple task, when testing it is defined preliminary parameters of voice and video signals and degree of stability in the working process towards to variations of input signals. For recognition system, working in real conditions and on real data, when testing it is defined percent of made mistakes, i.e. is assessed not the behavior under different conditions, it assessed how effectively system solves its task set before it. For speech recognition systems are used biometric methods for the assessment of percent error in the process of work.

For analyzing system's results we consider 10 speakers in table 1, among them 5 speakers are men and 5 women.

Table 1. Analysis of recognition system

Speakers	System of speech recognition	Multimodal system of speech recognition
Man-1	90,1%	93,4%
Man-2	91%	93,6%
Man-3	90,4%	94%
Man-4	89,8%	91,2%
Man-5	90,7%	93,6%
Woman-1	91,2%	93,6%
Woman-2	91,3%	93,6%
Woman-3	91%	93,1%
Woman-4	92,1%	93,8%
Woman-5	90%	92,7%

From analysis of recognition system (table 1) we get the Comparative analysis of speech recognition system (schedule 1):

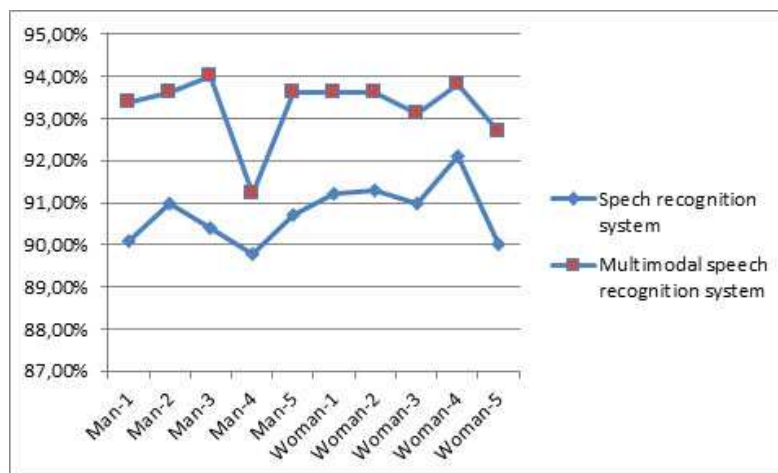


Fig. 3. Shedule 1. Comparative analysis of speech recognition system

For car navigation system it is possible propose created system. Hitherto methods of reading lips have not been effectively used in real time yet, due to difficulty in tracking lips. In this work we propose: the system of tracking lips which can effectively applying as support for an existing speech recognition systems; multimodal system of speech recognition as an effective tool for recognition of Kazakh speech.

## References

1. Mamyrbayev O.J. Mamyrbayev O.J. Types of multimodal speech recognition models. // Proceedings of the business conference "20 years of informatization in Kazakhstan: status, innovations, management development". - Almaty, 101–106. (2011)
2. Shannon C.E. A Mathematical theory of communication // Bell System Tech. J. – Vol. 27, No. 10. - P. 623-656. (1968).
3. Benoit C., Martin J.C., Pelachaud C., Schomaker L., Suhm B. Audiovisual and multimodal speech systems / in D. Gibbon (Ed.), Handbook of Standards and Resources for Spoken Language Systems, SuPlement. - Dordrecht: Kluwer Academic Publishers, 2000. – P. 29-35.
4. H. Dudley, *The Vocoder*, Bell Labs Record, 1939, Vol. 17, pp. 122-126.
5. H. Dudley, R. R. Riesz, and S. A. Watkins, *A Synthetic Speaker*, J. Franklin Institute, 1939, Vol. 227, pp. 739-764.
6. H. F. Olson and H. Belar, *Phonetic Typewriter*, J. Acoust. Soc. Am., 1956, Vol. 28, pp. 1072-1081.
7. J. W. Forgie and C. D. Forgie, *Results Obtained from a Vowel Recognition Computer Program*, J. Acoust. Soc. Am., Vol. 31, No. 11, pp. 1480-1489, 1959.
8. J. Suzuki and K. Nakata, *Recognition of Japanese Vowels—Preliminary to the Recognition of Speech*, J. Radio Res. Lab, Vol. 37, No. 8, pp. 193-212, 1961.
9. J. Suzuki and K. Nakata, *Recognition of Japanese Vowels—Preliminary to the Recognition of Speech*, J. Radio Res. Lab, Vol. 37, No. 8, pp. 193-212, 1961.
10. J. Sakai and S. Doshita, *The Phonetic Typewriter*, Information Processing 1962, Proc. IFIP Congress, Munich, 1962.
11. B.H. Juang, C.H. Lee and Wu Chou, *Minimum classification error rate methods for speech recognition*, Trans. Speech & Audio Processing, T-SA, May 1997, vol. 5, pp. 257-265.

# Design of Automated Image Recognition System to Assess the Quality of the Mineral Species Using CASE Technology

Olga E. Baklanova, Alexander E. Baklanov, and Olga Ya. Shvets

D.Serikbayev East-Kazakhstan State Technical University  
Serikbaeva,19, 070018 Ust-Kamenogorsk, The Republic of Kazakhstan  
{OEBaklanova@mail.ru, ABaklanov@ektu.kz, OShvets@ektu.kz}  
<http://www.ektu.kz/>

**Abstract.** This paper contains design and implement of the automated system for image recognition of mineral species in the mining industry. It is used CASE (Computer-aided software engineering) technology as AllFusion Process Modeler (BPWin), including a function modeling methodology of IDEF0 (Icam DEFinition for Function Modeling) on the functional modeling language of Structured Analysis and Design Technique (SADT) and a graphical representation of Data Flow Diagram (DFD). Now the developed automated system for images recognition for assessment of qualitative structure mineral breeds consists of six subsystems: for the research and receiving micrograph of rock; for input and identification of a rock micrograph; for preliminary processing: quality improvement; definition a threshold of the image reduction; for a choice of a vector of signs for the cluster analysis and the cluster analysis for definition of mineralogical composition of rocks. In the automated system the following Data storages are organized: Gallery of source images in which samples of rocks are stored; Gallery of the connected images where intermediate processing results of source images are stored, Template library where minerals samples are stored. Database with characteristics of rocks is autocompleted at input of images in Gallery of images. It is defined specifies requirements for the subsystem micrographs analysis.

**Keywords:** Petrographic analysis, Digital microscopy, Image recognition, Computer vision system, Case-technology, Mineral rock, Rock sample.

## 1 Introduction

Petrography is the science that studies the material composition of the rocks. Unlike minerals, rocks are aggregates composed of different minerals [1]. Minerals are homogeneous in composition and structure of the rocks and ores. They are natural chemical compounds resulting from various geological processes. Historically minerals initially determined by color and shape. Mineralogical and petrographic characteristics of rocks are determined by the macro- and microscopic examination of samples and thin sections [2]. According to the results of macro- and microscopic studies of rock up a summary petrographic characteristics and define the scope of the possible use of the rock [3]. Macroscopic examination of the rock is carried out visually with a magnifying glass or microscope with the following description of ores and rock cores [4]. In this case, are determined using ISO 25706: - The main rock-forming minerals; - The presence of mineral inclusions adversely affect the durability and decorative; - Availability of secondary minerals, weathering unstable and loose rocks and minerals, rocks crumble during processing; - The presence of inclusions of minerals, rocks impedes treatment, the nature of their distribution among other rock-forming minerals and quantity; - The texture and structure formation; - Fracture; - Translucency; - Color. Microscopic study of rock from thin sections includes a definition using ISO 25706: - Mineralogical composition; - Quantitative composition of rock-forming minerals; - Morphology of minerals and their relationships; - Structure; - Diagnostic petrographic constants of minerals; - The presence of harmful impurities; - Availability of secondary structures (newly-formed minerals, veins, and others.) with their quantification; - Petrographic name of the rock.

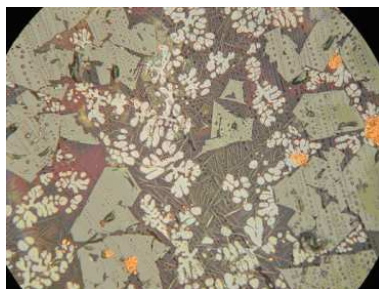
In offered work development of computer vision system of mineral rocks, in particular problems of development of a technique and image recognition technology to assess the qualitative composition of mineral rocks are considered.

## 2 Materials and Methods

### 2.1 Methods of Identification of Mineral Rocks Images

Minerals are homogeneous in composition and structure of the rocks and ores. They are natural chemical compounds resulting from various geological processes. Historically minerals initially determined by color and shape citeFarndon:Minerals.

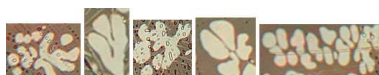
Consider a sample of slag copper anode as an example (Figure 1. Micrographs of this sample were kindly provided Eastern Research Institute of Mining and Metallurgy of Non-ferrous Metals (Kazakhstan, Ust-Kamenogorsk).



**Fig. 1.** Micrograph of a sample of slag copper anode, increasing in 500 times.

According to experts on microscopy of minerals from Eastern Research Institute of Mining and Metallurgy of Non-ferrous Metals at this picture there is no minerals having dependent on the direction of the plane of polarization of light. In this picture you can detect metallic copper and the following minerals: cuprite  $Cu_2O$ , magnetite  $Fe_3O_4$ , Delafosse  $CuFeO_2$ , silicate glass.

Cuprite  $Cu_2O$  can be identified as follows: it is characterized by the shape of a round shape, color - it is light gray (sometimes with a slight bluish tint). Figure 2 shows the graphical representation of cuprite.

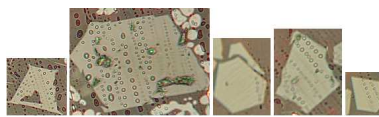


**Fig. 2.** Cuprite on micrographs.

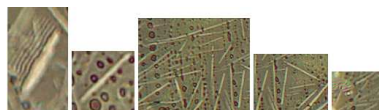
$Fe_3O_4$  magnetite on micrographs may also be detected by color and shape. Color of magnetite on micrographs is dark gray. Shape is angular, as expressed by technologists, "octahedral". Figure 3 shows magnetite apart from other minerals picture.

Delafossite  $CuFeO_2$  micrographs can allocate to the needle shape and gray (with a brownish tint) color. On Figure 4 it can be seen delafossite on the micrographs.

Metallic copper on the micrographs can be found on the following criteria: color - yellow, shape - round, without flat faces. Figure 5 represents a micrograph metallic copper.



**Fig. 3.** Magnetite on micrographs.



**Fig. 4.** Delafossite on micrographs.

Silicate glass - is a dark gray mass fills the rest of the space that is left of the other minerals. These data indicate that for real micrographs slag samples (and some other minerals) it is possible to use automated qualitative assessment of the mineral composition. After receiving the full image it is often needed to treat it, mainly to simplify further analysis.

## 2.2 Modelling Computer Vision System for Mineral Rocks' Images using Case - Technology

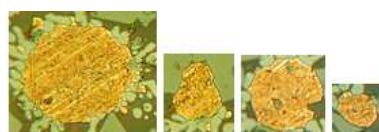
CASE-technology [7] for automated systems development was used to create computer vision system for assessing the qualitative composition of mineral rocks. For this purpose the project was divided into stages of analysis with a description of the main business processes [8]. CASE-means top-level AllFusion Process Modeler (BPWin) [9] and methodology IDEF0 (structural-functional model) and DFD (DataFlow Diagram) were used for the analysis and reorganization of business processes [10]. Description of develop automated image processing system for assessing the qualitative composition of mineral rocks in general, and its interaction with the environment is shown in Figure 6.

Nowadays developed automated image recognition system for assessing the qualitative composition of mineral rocks consists of 6 main subsystems [11]:

1. Research and getting micrograph rock.
2. Input and identification micrograph rock.
3. Pre-processing: improving the quality.
4. Definition of image reduction threshold.
5. Select the feature vector for cluster analysis.
6. Cluster analysis to determine the mineralogical composition of rocks.

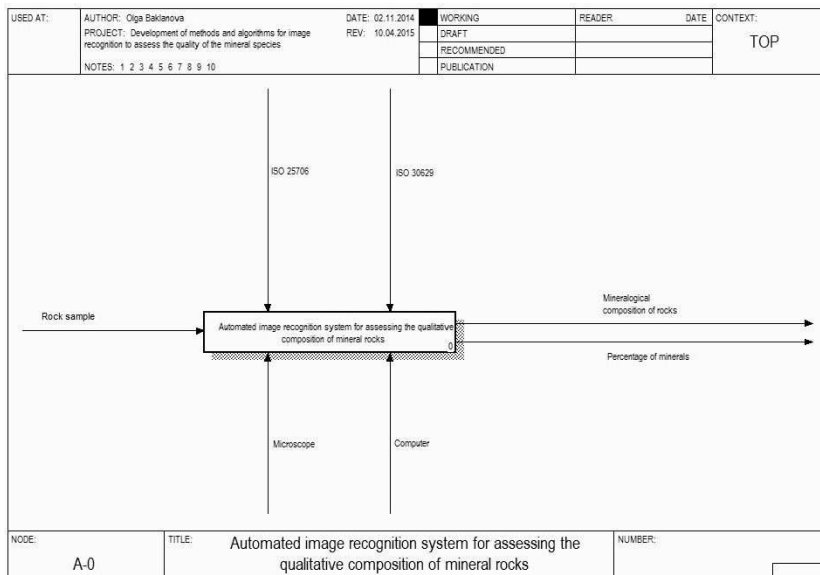
Structural-functional model of the major subsystems of the automated image recognition system for assessing the qualitative composition of mineral rocks in IDEF0 notation is shown in Figure 7.

On the first stage is supposed to implement research and obtain the micrograph of rock. It is necessary to follow these steps:

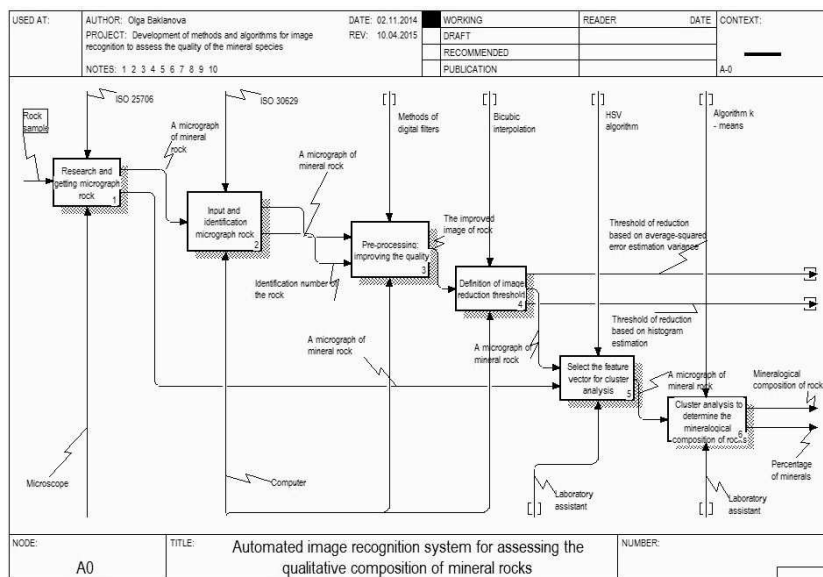


**Fig. 5.** Metallic copper on micrographs.





**Fig. 6.** Structural - functional model of an automated image recognition system for assessing the qualitative composition of mineral rocks.



**Fig. 7.** Structural - functional model of the major subsystems of the automated image recognition system for assessing the qualitative composition of mineral rocks.

1. Implement a preparation of rock sample (cut and polished).
2. Get a scanning electron micrograph of the rock on the microscope.
3. Visually examine the rock sample, determine the types of minerals that contain rocks. In the future, it will be used in the cluster analysis to select initial values of the centroids.
4. Send micrograph of rock on a computer using software supplied with the microscope.

DFD - diagram decomposition subsystem "Research and getting the micrograph rock" is shown in Figure 8.

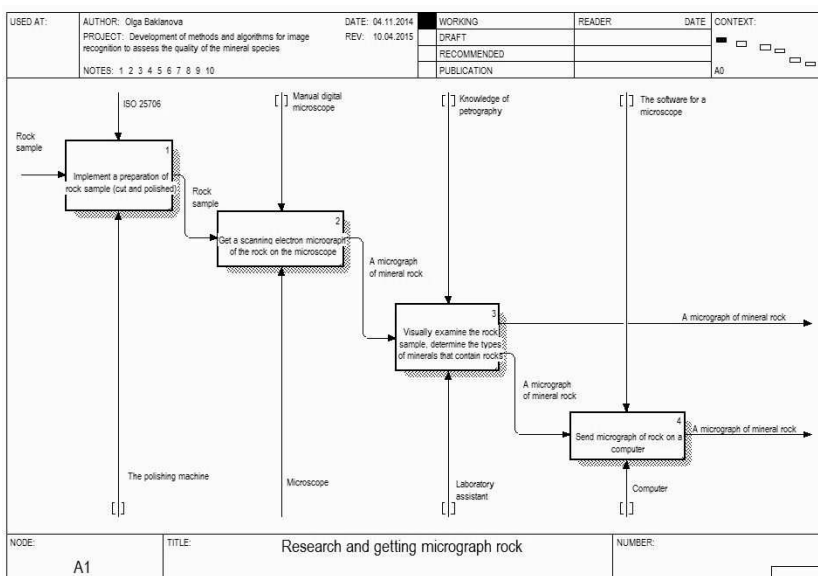


Fig. 8. DFD - diagram decomposition subsystem "Research and getting the micrograph rock".

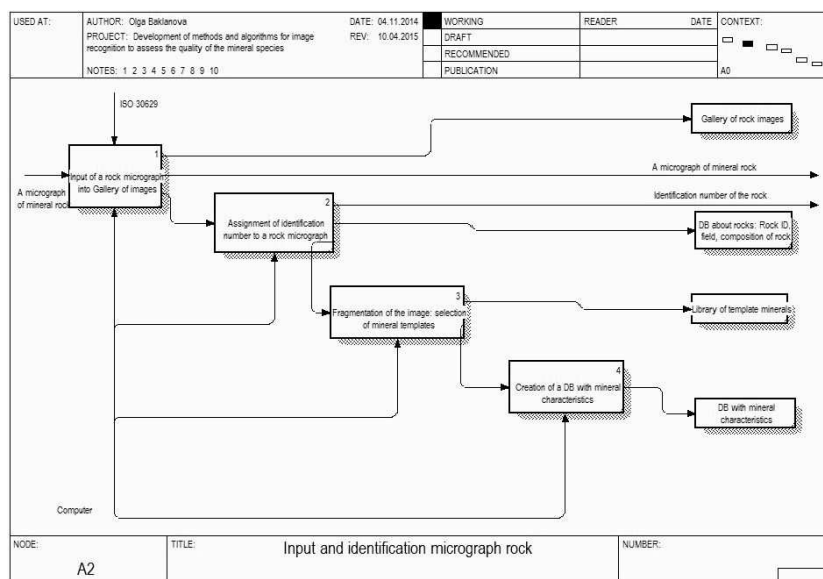
The next stage involves the identification and carry input micrograph rocks. It is necessary to follow these steps:

1. Enter the image in the image gallery. The image is automatically saved in the gallery of images of rocks.
2. When entering the images in the image gallery is automatically populated database with the characteristics of rocks: identification number of the rock; field, where the rock is found; composition of minerals that contain rocks (by visual inspection under a microscope).
3. Process fragmentation image to highlight patterns of minerals. It is formed minerals templates library.
4. When you save the template in the template library database with the characteristics of minerals mineral identification number is automatically fulfilled: the name of the mineral, color, shape, mine.

DFD - diagram decomposition subsystem "Input and identification micrograph of rocks" is shown in Figure 9.

At the next stage it is supposed to implement pre-treatment of images by digital filtering to improve the image quality for the subsequent segmentation and classification [12]. Pre-treatment micrograph includes the following features:

1. Filters for image smoothing to eliminate noise and RF interference. The system included: Box - filters, Gaussian filter, median filter.



**Fig. 9.** DFD - diagram decomposition subsystem "Input and identification micrograph of rocks".

2. Filters for improving image sharpness. Implemented the Laplace filter.
3. Reconstruction filter images distorted famous instrumental function. It is proposed to use the construction of the inverse filter algorithm based combination assembly. This algorithm proposed by the authors.
4. Isolation boundaries detection circuits. The system included a detector Sobel, Prewitt detector, the detector Laplacian Gaussian, Canny detector.

The program is recommended to keep the pre-processed images in the repository related images. DFD - diagram decomposition subsystem "Pre-treatment micrograph: improving quality" is shown in Figure 10.

At the next step it is determined the reduction of the image threshold. Purpose of stage is to reduce the dimensionality of data and as a consequence, increase the speed of algorithms on the following stages. Reducing the size of the image in  $N$  times leads to an increase in processing speed in  $N^2$  times, which greatly increases the efficiency by allowing the use of more "expensive" in terms of time, but better (adequate statement of the problem) algorithms. The main objective of this approach is definition a threshold reduction of digital images. It is proposed two approaches [13]:

1. Assessment of the mean square error dispersion within a sliding window of the original and the reduced images. The criterion is based on splitting the image to the same number of areas (windows disjoint), according to the selected partitioning step, and calculating a mean square error between the inside of the window on the luminance variance for the original and the reduced images.
2. Histogram evaluation. The method is based on a comparison of "forms" of the luminance histogram and the scaled image source. The standard error deviation of the histogram of the original image from the modified histogram serves as a criterion.

The Pyramid of reduced images is realized in the program. Image of Pyramid recommended keeping in storage related images. DFD - diagram decomposition subsystem "Definition of threshold reduction image" is shown in Figure 11.

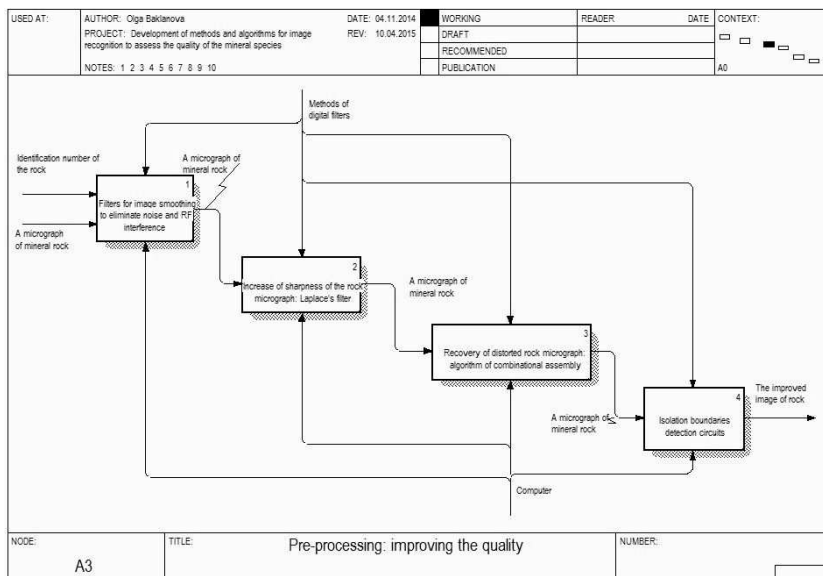


Fig. 10. DFD - diagram decomposition subsystem "Pre-treatment micrograph: improving quality".

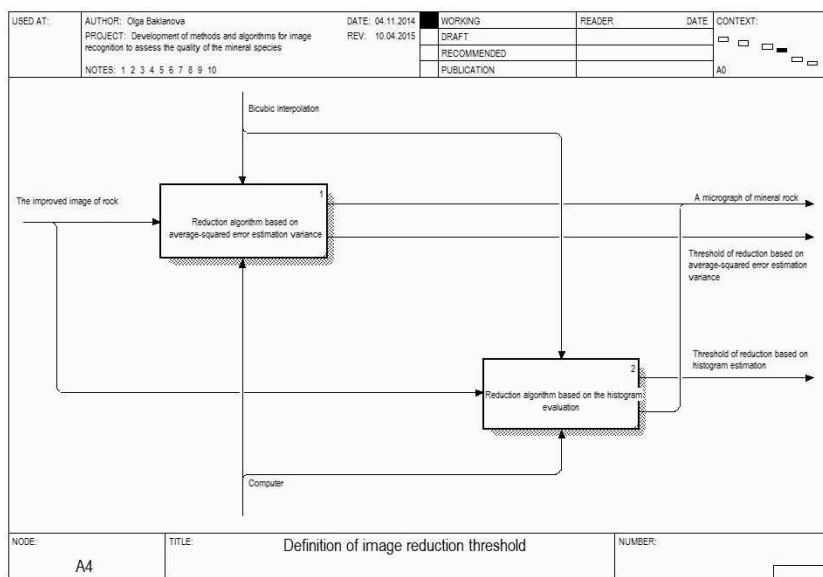


Fig. 11. DFD - diagram decomposition subsystem "Definition of threshold reduction image".

The next step is includes analysis assumes different color spaces for object segmentation using the algorithm proposed in the paper. The purpose of the analysis is to choose the structure of the feature vector for solving the problem of segmentation by cluster analysis method. It is based on their logical and mathematical representations the study of color models, as well as their use as feature space when the cluster analysis led to the conclusion about the correctness of pre-modernized color space HSV (color-saturation-brightness) choice. DFD - diagram decomposition subsystem "Select the future vector for clustering analysis" is shown in Figure 12.

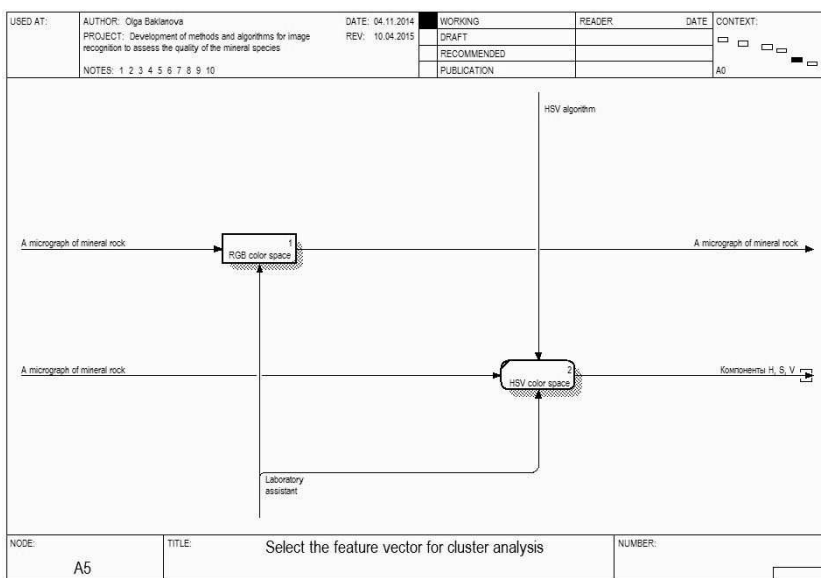


Fig. 12. DFD - diagram decomposition subsystem "Select the future vector for clustering analysis".

The next step is a process of cluster analysis expected for color image segmentation. This algorithm is implemented in two versions [19]:

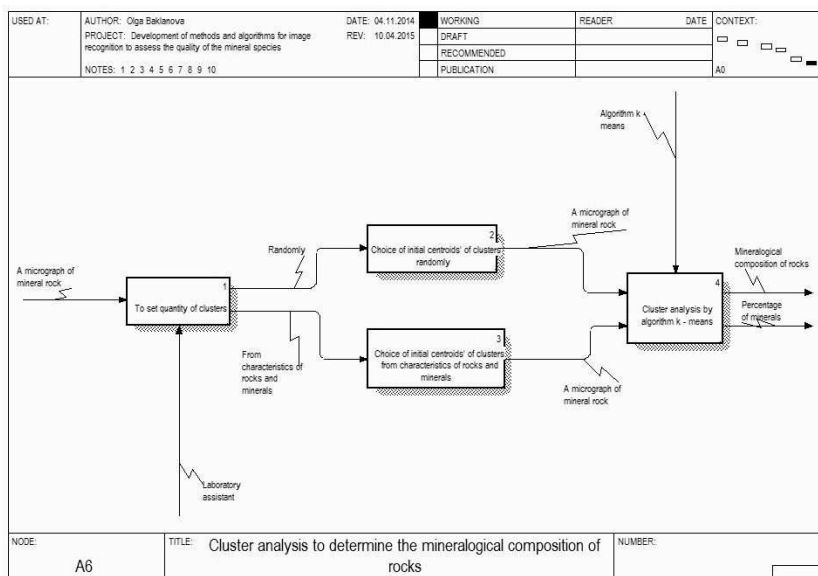
1. Select the number of clusters and initial centroids randomly.
2. Select the number of clusters and initial centroid of the characteristics of rocks and minerals.

DFD - diagram decomposition subsystem "Cluster analysis of rocks" is shown in Figure 13.

On the basis of the offered model the program complex which can be used for quality standard of structure of mineral breeds was developed [14].

### 3 Results and Discussion

It is considered the problem of cluster analysis to segment micro-images in mineralogy. In this case, the cluster is uniform in color-luminance characteristics region (segment) if digital image. And according to the specifics of digital images mineral rocks might be in the same cluster multiple segments, and research method determines homogeneity of individual clusters [16]. Due to two factors it was reasonable to use cluster analysis for the problem of segmentation: there is only one tuning parameter k - sa number of clusters that you want to highlight, and the sets of color-brightness characteristics associated with different types of segments analyzed image are compact [17]. The classic version of cluster analysis focused on a random selection of centroids is unacceptable for an adequate solution to the problem due to variations in the resulting picture



**Fig. 13.** DFD - diagram decomposition subsystem "Cluster analysis of rocks".

segmentation, which, in turn, depends strongly on the order of submission of observations to the input of the algorithm. As follows from the results of the test image processing, segmentation of each picture is different from the obtained segments. It is proposed to develop methods for obtaining initial values of the centroids to solve the problems of inadequate segmentation, and the choice of a set of parameters that form the vector of observations, the most satisfying description of the characteristics shared segments. Initially drawing from the scanner enters in the format RGB. In the future, it can be possible to convert it into color space HSV, XYZ and Lab. It can be concluded that as a feature vector that best satisfy the specifications describe the shared segment must be used with a HSV color space of the cylindrical passage in the Cartesian coordinate system. Accordingly, the image segment can be represented by a set of points in three-dimensional Cartesian coordinate system. To enhance the stability of the algorithm k - means necessary to set the initial values of centroid clustering [18]. In the course of the project have been developed methods for obtaining initial values of the centroids of the clusters. The input parameters of a clustering algorithm are:

- The number of clusters;
- The initial values of the centroids of the clusters.

Once the area is selected it must be recorded in the cluster table. To do this, you must specify the following parameters:

- The name of the cluster;
- Fill color;
- Mineral.

The result of the cluster analysis is shown in Figure 14.

Each cluster includes a certain number of points. Given the ratio of the number of points allocated in each cluster with a number of common points can be displayed relative rates of minerals in rock samples [19]. Various minerals marked in different colors. In this case, the metallic copper is red, magnetite - blue cuprite - orange. Considered sample has the following content of useful elements:

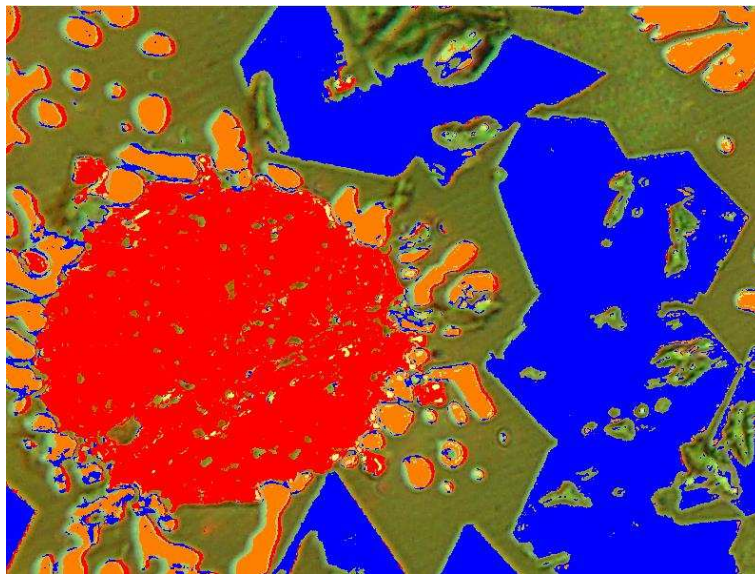


Fig. 14. DFD - diagram decomposition subsystem "The result of cluster analysis".

- Magnetite - 28.45%;
- Metallic copper - 18.45%;
- Cuprite - 7.92%.

#### 4 Conclusion

This work contains development of methods and algorithms of computer vision for image recognition of mineral species. It is described CASE-technology for automated systems development was used to create computer vision system for assessing the qualitative composition of mineral rocks. Methods and algorithms of computer vision of mineral rocks, in particular problems of the algorithm for automatic segmentation of colour images of ores, using the methods of cluster analysis are considered. Results of studies are demonstrated different colour spaces by k-means clustering. It was supposed the technique of pre-computing the values of the centroids. There is formulas translation metrics colour space HSV. The effectiveness of the proposed method lies in the automatic identification of interest objects on the total image, tuning parameters of the algorithm is a number that indicates the amount allocated to the segments. The program complex in the language C Visual Studio 2013 was developed for check of results of research.

**Acknowledgments.** The researches stated in this work were executed within the scientific project on the subject "Development of Methods and Algorithms of Image Understanding for an Assessment of Qualitative Structure of Mineral Breeds in the Mining Industry" of the program 101 "Grant financing of scientific researches" of Committee of science of the Ministry of Education and Science of the Republic of Kazakhstan.

#### References

1. Harvey, B., Tracy, R.J.: Petrology: Igneous, Sedimentary, and Metamorphic. 2nd ed., New York: W.H. Freeman (1995)
2. Chris, P.: Rocks and Minerals. Smithsonian Handbooks. New York: Dorling Kindersley (2002)

3. Shaffer, P. R., Herbert S. Z., Raymond P.: *Rocks, Gems and Minerals*. Rev. ed. New York: St. Martin's Press (2001)
4. Clarke, A. R., Eberhardt, C. N.: *Microscopy Techniques for Materials*. Science Woodhead Publishing, CRC Press, 459 p. (2002)
5. Panteleev, C., Egorova, O., Klykova, E.: *Computer microscopy*. Technosphere, 304 p. (2005)
6. Farndon, J.: *The practical encyclopedia of rocks and minerals. How to Find, Identify, Collect and Maintain the World's best Specimens, with over 1000 Photographs and Artworks*. London: Lorenz Books (2006)
7. Case, A.: *Information Systems Design: Principles of Computer-Aided Software Engineering*. N.J.: Prentice Hall (1988)
8. McClure, C.: *CASE in Software Automation*. N.J.. Prentice Hall (1989)
9. Maklakov, S.V.: *Creation of information systems with AllFusion Modeling Suite*. - Moscow: Dialog, 432 p. (2003)
10. Marko, D.A., McGovan, K.L.: *SADT: Structured Analysis and Design Technique*. N.Y.: McGraw Hill (1988)
11. Baklanova, O. E., Uzdenbaev, Z.S.: Development of methodology for analysis of mineral rocks in the mining industry. Joint issue of the Bulletin of the East Kazakhstan state technical University and Computer technology of Institute of computational technologies, Siberian branch of the Russian Academy of Sciences, Part 1, September, 2013. - pp.60–66. (2013)
12. Gonzalez, R. C., Woods, R. E.: *Digital image processing*. 3rd edition, Pearson Education, 976 p. (2011)
13. Baklanova, O.E., Shvets, O. Ya.: Development of Methods and Algorithms of Reduction for Image Recognition to Assess the Quality of the Mineral Species in the Mining Industry. LNCS, vol. 8671, pp. 75–83. Springer, Switzerland (2014). DOI 10.1007/978-3-319-11331-9
14. Baklanova, O.E., Shvets, O.Ya., Uzdenbaev, Zh.Sh.: Automation System Development For Micrograph Recognition For Mineral Ore Composition Evaluation In Mining Industry. IFIP, vol.436, pp. 604–613. Springer, Heidelberg (2014). DOI 10.1007/978-3-662-44654-6
15. Baklanova, O.E., Kornev, V.A., Shvets, O.Ya.: Quantitative Evaluation of Accuracy of Digital Microscopy System for Automated Petrographic Analysis. In: Proceedings of the 4th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH - 2014), pp.560–566, SCITERPRESS Science and Technology Publications, Lda (2014). DOI: 10.5220/0005025705600566
16. Mandel, J.: *Cluster analysis*. Moscow: Finance and statistics, 176 p. (1988)
17. Odell, P. L., Duran, B. S.: *Cluster Analysis: A Survey*, Springer-Verlag (1974)
18. Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2, pp. 283-304 (1998)
19. Baklanova, O.E., Shvets, O. Ya.: Methods and Algorithms of Cluster Analysis in the Mining Industry. Solution of Tasks for Mineral Rocks Recognition. In: Proceedings of the 11th International Conference on Signal Processing and Multimedia Applications (SIGMAP - 2014), pp. 165–171, SCITERPRESS Science and Technology Publications, Lda (2014) DOI: 10.5220/0005022901650171.



# Software Implementation of the Cryptographic System Models with the Given Cryptostrength

Rustem Biyashev<sup>1</sup>, Maksat Kalimoldayev<sup>2</sup>, Saule Nyssanbayeva<sup>3</sup>, Nursulu Kapalova<sup>4</sup>, and Rinat Khakimov<sup>5</sup>

<sup>12345</sup>Institute of Information and Computational Technologies Ministry of Education and Science, 125 Pushkin str., Almaty, 050010, Republic of Kazakhstan

brg@ipic.kz, mnk@ipic.kz, snyssanbayeva@gmail.com, Kapalova@ipic.kz, relessar@mail.rulncs

<http://www.ipic.kz>

**Abstract.** The models of software implementation of the system of cryptographic protection of information (SCPI) with the specified characteristics are described. This system is designed for using in systems and networks of information transmission and storage. In the developed system of cryptographic protection of information, the nonconventional algorithms of encryption and digital signature developed on the basis of non-positional polynomial notations (NPNs) are implemented. Definable characteristics are the length of the message or the digital signature, as well as cryptostrength of algorithms.

**Keywords:** cryptography, encryption, digital signature, non-positional polynomial notations, cryptostrength, residue, software implementation.

## 1 Introduction

This paper presents the results obtained by the development of the following models:

- software implementation of the digital signature scheme;
- systems of cryptographic information protection based on non-positional polynomial notations (NPNs) with the given cryptostrength.

The implemented system of cryptographic information protection consists of three parts (blocks): the formation of full secret keys for implemented non-traditional cryptographic algorithms, the encryption system of electronic messages with the given cryptostrength and the digital signature scheme with the given cryptostrength [1-4].

## 2 Creation of the Model and the Algorithm of the Block of the Digital Signature Formed by Modules of Several Redundant Polynomial Bases

The algorithm of formation of a digital signature of length  $N_1$  bits for the electronic message of the given length  $N$  bits by modules of several redundant bases is software-implemented and includes three stages.

**Stage 1.** Creation of NPNs.

**Stage 2.** Hashing (compression) of a message from length  $N$  to length  $N_1$  ( $N_1 \ll N$ ) by extrapolation on the redundant (extension) base numbers.

**Stage 3.** Encryption of the obtained hash-value.

Encrypting the hash value is provided by non-traditional method [4].

When checking the digital signature, after receiving a signed message, an addressee computes two hash-values. The first hash-value is determined from the obtained message. The second hash-value is determined as a result of decrypting the obtained digital signature. If the values of both hash-values coincide, then the signature is authentic.

The full key in this algorithm of forming the digital signature is the system of working bases and the system of redundant bases, taking into account the order of their location, and the full key which is used to encrypt the hash value.

To implement the digital signature scheme two models were considered. In the first model, the cryptostrength of the digital signature formation algorithm is determined directly in the block of the electronic digital signature (EDS) system. In the second model, this cryptostrength is calculated in the key formation block and is stored in the database (DB) of full keys.

The creation of different models of implementation for non-traditional cryptography systems allows to construct such system of cryptographic information security which would be simple to be transformed in case of changing the model of implemented cryptographic algorithms.

The structural diagram of the first model of implementation of non-traditional algorithm of the digital signature formation is shown in Figure 1. This diagram shows the structure of the main stages and the EDS formation algorithm and its software implementation.

The choice of systems of working and redundant bases and bases for hash - value encryption, i.e. at each stage of EDS formation algorithm, is made from the database of irreducible polynomials with binary coefficients. The cryptostrength which is compared to the given  $p_{preset}$  is calculated by the chosen base number systems. If the calculated  $p_{rated}$  value turns out to be less than the given cryptostrength  $p_{preset}$ , other base number systems will be chosen.

When the necessary set of base number systems is found, the digital signature by means of encrypting the received hash - value is calculated.

The advantage of this model lies in the fact that the full key is formed and chosen at the time of digital signature formation. This of course reduces the rate of formation of EDS. The increase of the speed of obtaining the signature on this model can be achieved by parallelization of arithmetic operations embodied in the signature formation algorithm, that is by the selected base numbers of each of these three systems.

In the second digital signature scheme model, the full key is formed in the "Formation of full keys" block (respectively, by the length of messages and digital signatures) and is stored in the full keys database. The components of the full key are the system of working bases, the system of redundant bases, and also the pseudorandom sequence (the traditional secret key), and the inverse for pseudorandom sequence key for hash-value encryption.

Then the value of cryptostrength which is registered in the corresponding database is calculated for the EDS system of full keys. Besides these components, other required information can be stored in the database, too.

### 3 Development of the Model and the Algorithm of Implementation of the System of Cryptographic Protection of Information

The assignment of the system of cryptographic protection of information (SCPI) is the cryptography information conversion on the basis of nonconventional algorithms of encryption and electronic digital signature developed on the basis of nonpositional polynomial notations (NPNs) for using in infocommunication information transmission and storing systems [1-5]. The feature of the model is the fact that the created SCPI implements algorithms of encryption and formation of EDS with the given cryptostrength.

The cryptostrength of these encryption algorithms is determined by the total number of possible and distinct from each other variants of choice of full keys. Formulas of cryptostrength of non-traditional algorithms of encryption and formation of EDS were received. Reliability analysis of these cryptoalgorithms showed that using polynomial notations in residual classes with binary coefficients allows to considerably increase their effectiveness by parallelization in

calculation on each base number of the used NPNs. It also allows to decrease the length of digital signature without losing cryptostrength.

For the above-mentioned cryptoalgorithms, the created SCPI includes three interconnected blocks (or subsystems) which realize: the formation of full secret keys, the system of encryption and the digital signature scheme [6]. For software implementation two models of cryptographic protection of information are proposed.

In the first SCPI model the choice of full keys of the realized cryptoalgorithms is implemented from the DB of irreducible polynomials with binary coefficients directly in the blocks of program modules of encryption and digital signature formation. The cryptostrength which is compared to the given one is calculated by the chosen base number systems  $p_{preset}$ . If the value of the calculated cryptostrength  $p_{rated}$  turns out to be more than the given cryptostrength  $p_{preset}$ , other base number systems will be chosen.

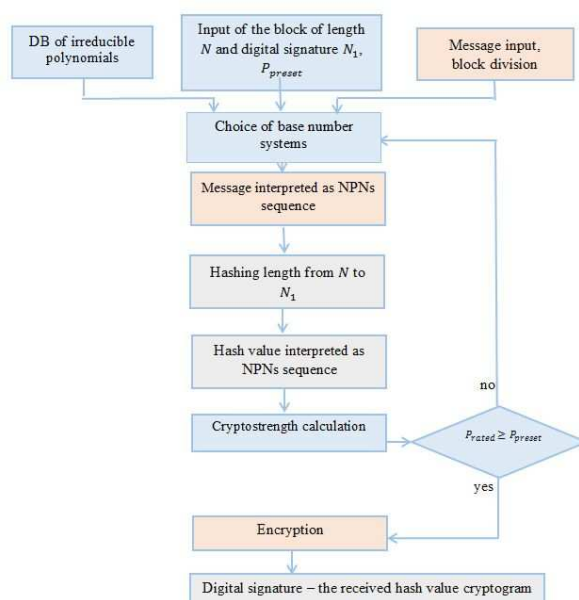


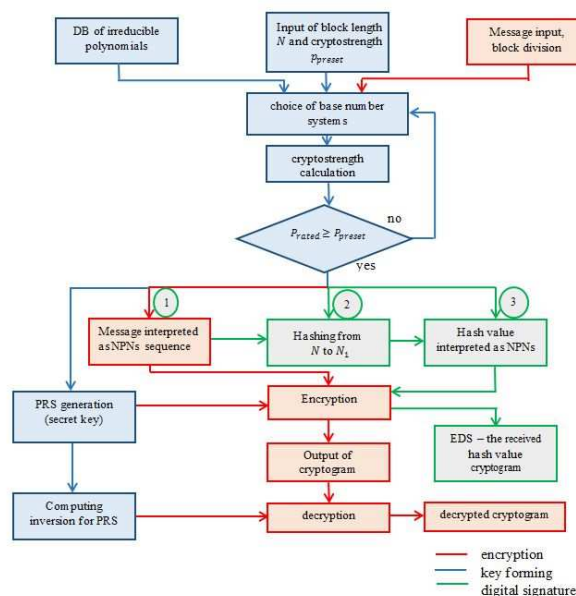
Fig. 1. The arrangement of realization of non-traditional algorithm of digital signature formation

When the necessary full key is found, the digital signature encrypting or calculation is realized.

The advantage of this model lies in the fact that the full key is formed and chosen at the time of cryptoalgorithm performance. This SCPI model allows to provide considerable privacy of the used full keys, but under these conditions the speed reduction of procedures of encryption or formation of the digital signature is possible. It is possible to increase the speed at the expense of parallelization of computing operations in all base numbers used in the realized cryptoalgorithm.

The arrangement of implementation of the first model of the cryptographic information protection system is given in figure 2. It completely shows the structure of SCPI and correlation of SCPI blocks. In this model the database of irreducible polynomials with binary coefficients should be protected.

In the second SCPI, the full key for encryption and EDS computing systems is formed in the "Formation of Full Keys" block with the use of the irreducible polynomials with binary coefficients and is stored in the database of full keys. The components of the full key are the system of working bases, the system of redundant bases, the system of base numbers, the pseudorandom



**Fig. 2.** Structural diagram of cryptographic information protection system based on non-positional polynomial notations

sequence (the traditional secret key), and the inverse for pseudorandom sequence key for hash-value encryption [7]. The database of full keys can be formed in share mode or separately for each block.

The advantage of the second model is the possibility of creating large inventory of full cryptographic keys of different lengths in the database, which, of course, needs to be protected. Therefore, it is assumed to protect this database, to store it in the encrypted form and also to have its reserve copies.

## 4 Conclusions

To ensure the security of the full key database it is supposed to store it in the computer in encrypted form. One of the planned works is to implement the SCPI model on the basis of the time pad, i.e. the full key database will be stored on a removable memory (USB) and identified only by encryption software. The task to prevent reusing of full secret key from the DB is also solved.

The creation of various models of implementation for non-traditional cryptographic systems allows to create such system of cryptographic information protection which would be easily transformed under model changes of the implemented cryptographic algorithms.

The development of the system of cryptographic information protection is carried out in compliance with requirements of legal documents of the Republic of Kazakhstan in the field of informatization.

## References

1. R.G. Biyashev, "Development and investigation of methods of the overall increase of reliability in data exchange systems of distributed ACSs," Doctoral Dissertation in Technical Sciences: Moscow (1985), p.328.

2. B.M. Amerbayev, R. G. Biyashev and S.E. Nyssanbayeva Application of nonpositional polynomial notations at cryptographic protection // Math. Nat. Acad. of Sciences of the Republic of Kazakhstan.- Phys.-Math. - Almaty: Gylym 2005. - No. 3. - pp. 84-89.
3. R. G. Biyashev and S .E. Nyssanbayeva Modular models of ensuring information security // Problems of optimization of complex systems. - Novosibirsk: Inst Calc. Math. Math. Geophysics SB RAS, 2010. - pp. 117-126.
4. R. G. Biyashev and S .E. Nyssanbayeva, "Algorithm for Creation of a Digital Signature with Error Detection and Correction Cybernetics and Systems Analysis. Vol. 48, No 4, pp. 14-23 (2012)
5. R. G. Biyashev and S .E. Nyssanbayeva, N.A. Kapalova Secret keys for nonpositional cryptosystems. - Germany: LAP LAMBERT, 2014. - 136 p.
6. R. Biyashev, M. Kalimoldayev, S. Nyssanbayeva, N. Kapalova, R. Khakimov. Program Modeling of the Cryptography Algorithms on Basis of Polynomial Modular Arithmetic / The 5th International Conference on Society and Information Technologies (ICSIT 2014, march 4-7, 2014- Orlando, Florida, USA) - IIS. pp. 49-54
7. N.A. Kapalova, S.E. Nyssanbayeva, "Analysis of statistical properties of algorithm of generation of pseudorandom sequences Materials of the X-th International scientific-practical conference "Information security P. 2. TIT SFU Publishing house, Taganrog (2008), pp. 169-172.

# The Modified Digital Signature Algorithm Based on Modular Arithmetic

Rustem Biyashev, Saule Nyssanbayeva and Yenlik Begimbayeva

Institute of Information and Computational Technologies of MES RK,  
125 Pushkin str., Almaty, 050010, Republic of Kazakhstan {sultasha1,enlik89}@mail.ru  
<http://ipic.kz>

**Abstract.** In this paper the model of unconventional asymmetric system of digital signature is described. Cryptosystems, developed on the basis of nonpositional polynomial notations (NPNs), are called nonconventional, nonpositional or modular. The model of signature is created on the basis of digital signature scheme the Digital Signature Algorithm (DSA) and NPNs. Application of NPNs allows increasing the cryptostrength of the cryptosystem and reducing the key length

**Keywords:** digital signature, asymmetric scheme, nonpositional polynomial notations, cryptostrength.

## 1 Introduction

Unconventional systems are basis for creation of the proposed model of asymmetric system of the digital signature (DS) on the basis of DSA algorithm [1-3]. In the classical notations in residue number system the bases are prime numbers, and in NPNs bases are irreducible polynomials over  $GF(2)$  [3]. Usage of NPNs allows reducing the key length, increasing the strength and efficiency of the nonpositional cryptographic algorithms [4]. Increased efficiency is ensured by the NPNs rules in which all arithmetic operations can be performed in parallel on NPNs base module.

The developed unconventional cryptographic algorithms of the DS formation are performed for a predetermined length of an electronic message. In these cryptosystems as the cryptostrength criterion used cryptographic strength of DS formation algorithms themselves, which is characterized by the full private key [3-5].

In [3] developed the NPNs arithmetic with polynomial bases and its application to problems of increasing reliability. It is shown that the algebra of polynomials over a field be the irreducible polynomial modulo over this field is the field and polynomial presentation in nonpositional is unique. The rules of arithmetic operations in NPNs and the polynomial recovery by its residues are defined. According to the Chinese remainder theorem all working bases should be different.

## 2 Formation of NPNs

The process of NPNs formation for signing an electronic message  $M$  of length  $N$  bits is as follows. Systems working bases with binary coefficients are selected

$$p_1(x), p_2(x), \dots, p_S(x), \quad (1)$$

where  $p_i(x)$ -irreducible polynomials over the field  $GF(2)$  of degree  $m_i$  respectively,  $i = \overline{1, S}$ . The main working range of NPNs represented by polynomial  $P_S(x) = \prod_{i=1}^S p_i(x)$  of degree  $m = \sum_{i=1}^S m_i$ . All the selected working base should be different from each other (according to the Chinese remainder theorem), even if they are irreducible polynomials of one degree.

In NPNs any polynomial  $F(x)$ , the degree of which is less than  $m$ , has nonpositional representation as a sequence of residues from its division into base  $p_1(x), p_2(x), \dots, p_S(x)$  respectively, and it is unique:

$$F(x) = \alpha_1(x), \alpha_2(x), \dots, \alpha_S(x), \quad (2)$$

where  $F(x) \equiv (\alpha_i(x) \pmod{p_i(x)})$ ,  $i = \overline{1, S}$ . By the form (2) recovering positional representation of the polynomial  $F(x)$  [3,4]:

$$F(x) = \sum_{i=1}^S \alpha_i(x) B_i(x), B_i(x) = \frac{(P_S(x))}{(p_i(x))} M_i(x) \equiv 1 \pmod{(p_i(x))}, i = \overline{1, S}. \quad (3)$$

Polynomials  $M_i(x)$  are selected such as to satisfy the comparison in (3).

In NPNs the electronic message of length  $N$  bits is interpreted as a sequence of remainders of division of some polynomial (denote it also as  $F(x)$ ) according to the working base  $p_1(x), p_2(x), \dots, p_S(x)$  degree not higher than  $N$ , ie in the form (2). Bases are selected from the number of all irreducible polynomials of degree from  $m_1$  to  $m_S$  from the execution condition of the equation [6]:

$$k_1 m_1 + k_2 m_2 + \dots + k_S m_S = N. \quad (4)$$

In equation (4)  $0 \leq k_i \leq n_i$ ,  $i = \overline{1, S}$  - the unknown coefficients and the number of selected irreducible polynomials of degree  $m_i$ . One concrete set of these coefficients is one of the solutions (4) and defines one system of working bases,  $n_i$  - the number of irreducible polynomials of degree  $m_i$ ,  $1 \leq m_i \leq N$ ,  $S = \sum_{i=1}^S k_i$  - the number of selected working bases. Equation (4) defines the number of  $S$  working bases, residues that cover the length  $N$  of the given messages. The full system of residues by polynomials modulo of degree  $m_i$  include all polynomials of degree not higher than  $m_i - 1$ , to record requiring  $m_i$  bit [4-5].

In the NPNs to obtain DS the hash value is used from the signed message.

### 3 Hashing an electronic message in NPNs

For hashing the message  $M$  of length  $N$  bits to  $N_k$  bits the redundant bases

$$p_{S+1}(x), p_{S+2}(x), \dots, p_{S+U}(x). \quad (5)$$

are entered.

These bases are selected randomly from all irreducible polynomials, degree not higher than  $N_k$ . System of redundant bases is formed independently from working bases selecting. Among  $U$  redundant bases may coincide with some of the working bases. Let denote  $a_1, a_2, \dots, a_U$  and  $d_1, d_2, \dots, d_U$  degree and the number of irreducible polynomials, respectively, used in their selection. The number of selected redundant bases in this case determined from the equation:

$$t_1 a_1 + t_2 a_2 + \dots + t_U a_U = N_k, \quad (6)$$

where  $0 \leq t_j \leq d_j$ ,  $0 \leq a_j \leq N_k$ ,  $j = \overline{1, U}$ ,  $t_j$  - the number of selected redundant bases of degree  $a_j$ ,  $U = t_1 + t_2 + \dots + t_U$  - number of selected redundant bases, recording of the residues which covers the hash value of length  $N_k$ . The solution of equation (6) defines one system of redundant bases.

The next stage of calculations of the hash value is to calculate the redundant residues

$$\alpha_{S+1}(x), \alpha_{S+2}(x), \dots, \alpha_{S+U}(x) \quad (7)$$

by dividing the reduced polynomial  $F(x)$  to redundant bases (5). Then, the hash value is interpreted as a sequence of residues:

$$h(F(x)) = (\alpha_{S+1}(x), \alpha_{S+2}(x), \dots, \alpha_{S+U}(x)), \quad (8)$$

where  $h(F(x)) \equiv \alpha_{S+j}(x) \bmod p_{S+j}(x)$ ,  $j = \overline{1, U}$ . The sum of length of the redundant residues (7) is the length of the hash value and the DS.

The nonpositional asymmetric DS system is constructed to obtained hash value.

#### 4 Asymmetric digital signature based on NPNs

Digital Signature Algorithm (DSA) - this is the digital signature scheme [7], which was adopted in 1994, as the US standard, and acting until 2001. This scheme is the variation of a digital signature of the ElGamal scheme and K. Schnorr. DSA reliability is based on the practically insoluble of the particular case of the problem of calculating the discrete logarithm.

The essence of DSA electronic signature scheme is the following. The sender and recipient of the electronic document in computation use large prime integers  $p$  and  $q$ , in the range  $2^{L-1} < p < 2^L$ ,  $512 \leq L \leq 1024$ ,  $L$  multiple of 64,  $2^{159} < q < 2^{160}$ ,  $q$ - prime divisor of  $(p-1)$  and  $g = h^{\frac{p-1}{q}} \bmod p$ , where  $h$  arbitrary integer,  $1 < h < p-1$  such that  $h^{\frac{p-1}{q}} \bmod p > 1$ .

The private key  $b$  is kept in secret and randomly selected from the range  $1 \leq b \leq q$ . Calculated value  $\beta = g^b \bmod p$ . The parameters  $(p, q, g)$  - are the public keys, which published for all users of the information exchange system with DS.

The formation process of the DS for the message  $M$  consists of the following steps:

1. determine hash value  $h$  from the signed message  $M : h = h(M)$ ;
2. choose random integer  $r$ , which is keeping in secret, in range  $1 \leq r \leq q$  and its varying from one sign to another;
3. calculate value:  $\gamma = (g^r \bmod p) \bmod q$ ;
4. calculate  $\delta = (r'(h + b\gamma)) \bmod q$ , where  $r'$  satisfies the condition  $(r'r) \bmod q = 1$ ;
5. DS for the message  $M$  is the pair of numbers  $(\gamma, \delta)$ . They are passed along with the message by open communication channels.

The process of verification of DS  $M$  is consists of the following steps: (Let denote  $M', \delta', \gamma'$  obtained by the addressee version of  $M, \delta, \gamma$ ).

1. checking the conditions  $0 < \delta, \gamma < q$ . Reject the signature if any one of the conditions of the DS is not satisfied these conditions.
2. calculate hash value  $h_1 = h(M')$  from the received message  $M'$ .
3. calculate value  $\nu = (\delta')^{-1} \bmod q$ .
4. calculate value:  $z_1 = (h_1 \nu) \bmod q$  and  $z_2 = (\gamma' \nu) \bmod q$ .
5. calculate value:  $u = ((g^{z_1} \beta^{z_2}) \bmod p) \bmod q$ .
6. the DS is valid if  $\gamma' = u$ . It means that in the transfer process the integrity of the message was not compromised. Otherwise, the signature is invalid.

In the construction of nonpositional asymmetric system of DS initially made modification of the DSA algorithm: excluded the second module  $q$ , because for the DS calculation will use the obtained above hash value in NPNs. Then, for the constructed DSA algorithm by one module  $p$  the nonpositional system of digital signature of length  $N_k$  will be developed.



The obtained electronic document (hash value) of length  $N_k$  (8) is considered in nonpositional polynomial notation, which nonpositional asymmetric system of DS be developed. For this NPNs polynomial bases

$$\eta_1(x), \eta_2(x), \dots, \eta_W(x) \quad (9)$$

are selected similar as the choice of working bases in section 2. Let denote  $q_1, q_2, \dots, q_W$  and  $l_1, l_2, \dots, l_W$  degree and the number of irreducible polynomials, respectively, used in their selection. The number of selected redundant bases in this case is determined from the analog equations (4) and (6).

Then, for bases (9) respective generating elements (polynomials)  $g_1(x), g_2(x), \dots, g_W(x)$  are found, which are analogous to primitive elements in the algorithm DSA.

The private key of the sender  $b$  is selected also in range  $[1, 2^q]$ , where  $q$  is the sum of the degrees  $q_1, q_2, \dots, q_W$ .

Calculate value of the public key  $\beta(x)$ :  $\beta(x) = (\beta_1(x), \beta_2(x), \dots, \beta_W(x))$  by bases modulo (7).

Next, choose a random integer  $r$  in range  $[1, 2^q]$ .

Polynomials  $\gamma(x)$  and  $\delta(x)$  in the modified DSA algorithm by one module  $p$  presented in nonpositional form as a sequence of residues from their division on the bases (9).

The digital signature for the message  $M$  is the pair of polynomials  $(\gamma(x), \delta(x))$ .

## 5 Conclusion

The feature of formation of nonpositional asymmetric system of DS using the DSA algorithm and NPNs is that not all parameters can be used as indicators of the degree. This applies to polynomial bases of NPNs. Computer modeling of the modified cryptosystems based on NPNs will allow developing recommendations for their reliable use and generation of full secret keys.

## References

1. Akushskii, I.Ya., Juditskii, D.I., *Machine Arithmetic in Residue Classes [in Russian]*, Sov. Radio, Moscow (1968).
2. Stallings W., *Cryptography and Network Security*, (4th Edition), Prentice Hall, (2005).
3. Biyashev, R.G., *Development and investigation of methods of the overall increase in reliability in data exchange systems of distributed ACSs*, Doctoral Dissertation in Technical Sciences, Moscow (1985).
4. Biyashev, R.G., Nyssanbayeva, S.E., *Algorithm for Creation a Digital Signature with Error Detection and Correction*, Cybernetics and Systems Analysis, 4, 489-497 (2012).
5. Biyashev, R., Nyssanbayeva, S., Kapalova, N.: *The Key Exchange Algorithm on Basis of Modular Arithmetic. International Conference on Electrical, Control and Automation Engineering (ECAE2013)*, Hong Kong - Monami,S. - P.501-505 (2014).
6. Moisil, Gr.C, *Algebraic Theory of Discrete Automatic Devices [Russian translation]*, Inostr. Lit., Moscow (1963).
7. FIPS PUB 186. *Digital Signature Standard (DSS)*.

# Wireless Sensor Networks and Computational Geometry Problems

Adil Erzin<sup>1,2</sup>, Natalia Shabelnikova<sup>2</sup>, Lydia Osotova<sup>2</sup>, and Yedilkhan Amirgaliyev<sup>3</sup>

<sup>1</sup> Sobolev Institute of Mathematics, Novosibirsk, Russia

<sup>2</sup> Novosibirsk State University, Novosibirsk, Russia

<sup>3</sup> Institute of Information and Computational Technologies, Almaty, Kazakhstan  
{alfred.hofmann, ursula.barth, lnsc}@springer.com

**Abstract.** This paper contains the previously known results, as well as new our results by objective of constructing the least dense covers of the plane regions with disks, ellipses and sectors. Such problems are considered in the context of design an energy efficient wireless sensor networks, which are an example of a distributed network of data collection and transmission. Arising in this connection computational geometry problems are difficult to solve, so basically approximate solutions are searched. We proposed several new coverage models with the least known density in their class.

**Keywords:** Wireless Sensor Networks, Coverage Density.

## 1 Introduction and Problem Formulation

Wireless sensor network (WSN) consists of devices that collect information and transmit it to the base station via radio. Each sensor is equipped with a battery replacement or charging of which is either impossible or impractical. WSN's *lifetime* is a time period during which the network collects and transfers data from a certain region. Energy consumption is associated with sensor's monitoring area, which is called the *coverage area* of the sensor. Energy consumption is proportional to the area covered by sensor. Therefore, the multiple coverage entails excessive energy loss, which leads to a reduction of network's lifetime. Therefore, the problem of energy-efficient monitoring can be reduced to the problem of finding the least dense cover, where the density of coverage an area of  $S$  is the ratio of the sum of squares of elements in the cover to  $S$ . The lower density, the better cover.

Since in the applications a coverage area of a sensor can has different shapes (disk, ellipse, sector) and different size (radius, semi-axes, angle and radius), then the following general computational geometry problem can be formulated.

**General problem.** *For a given plane region, every point of which must be covered by at least one figure, the list of types of figures and admitted regions of their parameters, it is required to define the set of figures in the cover, values of the parameters and to determine the placement of each figure and its orientation in order to minimize the density of the cover.*

Since the set of covers is continual, in practice, typically the *regular* covers are considered. In the regular cover the region is split into the equal polygons (*tiles*), and all the tiles are covered identically. As a result, for the calculation of the density of a cover of the whole region it is sufficient to estimate the coverage density of one tile.

In the cover one can use different types of figures. Thus, the two figures have the different types if they differ not only by shape, but even if they have different sizes. For example, a disk and a sector it is different types of figures. But two disks are different too if their radii differ. On the other hand, the two ellipses have the same type if their half-axis coincide regardless of their orientation.

We introduce the class  $COV(p, q)$  as a set of regular covers, in each of which the tile is covered with  $p$  figures of  $q$  different types. Obviously, the more types of figures used in the cover, the lower density of cover may be. In this paper we restrict ourselves to the covers in the classes  $COV(p, q)$ ,  $q \leq 2$  (with figures of at most two types).

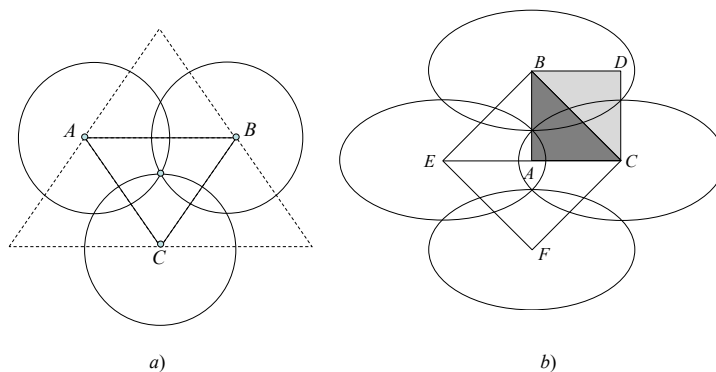
Consider a cover of one tile in the class  $COV(p, q)$ . Renumber the figures covering a tile and denote by  $e_i^t$  the  $i$ -th figure of type  $t$ ,  $t = 1, \dots, q$ ,  $i = 1, \dots, n_t$ , where  $n_t$  is a quantity of figures of type  $t$  covering one tile. Obviously,  $\sum_{t=1}^q n_t = p$ . Denote by  $A(e_i^t)$  tile area, where it can be located centre (for disk or ellipse) or vertex (for sector) of the figure  $e_i^t$ . If one set the number of each type of figures  $n_t$  and the placing domain of each figure  $A(e_i^t)$ , then define the *coverage model*. The choice of the sizes, the centre (from  $A(e_i^t)$ ) and a tilt angle of each figure  $e_i^t$ ,  $i = 1, \dots, n_t$ ,  $t = 1, \dots, q$ , defines a *concrete cover*.

**Problem formulation.** *In this paper, we consider the problem of the construction of minimum density regular covers in the classes  $COV(p, q)$  when one tile is covered by  $p$  figures of  $q$  ( $q = 1, 2$ ) different types.*

## 2 One Type of Figures

If equal disks are used in the cover, then in [13] is proved that regular cover  $D1$  in class  $COV(3, 1)$  is optimal (of minimum density) if the centres of three pairwise intersecting disks form the equilateral triangle – a tile, and only one point belongs to all three disks (Fig. 1a). The density of this cover equals  $2\pi/\sqrt{27} \approx 1.2091$ .

Some covers using ellipses can be constructed from the covers that use disks by applying the *affine transformation* (AT). An AT is a transformation that preserves straight lines and ratios of distances between points lying on a straight line while keeping the coverage density the same. Examples of ATs include translation, expansion, reflection, and rotation. An AT is equivalent to a linear transformation followed by a translation. In [6] we noticed that after applying AT to the cover  $D1$ , one can get the cover  $E1$  with equal ellipses in different classes depending on the tile (Fig. 1b). If tile is a triangle  $EBC$ , then the cover is in the class  $COV(3, 1)$ . If tile is a triangle  $ABC$ , or a square  $ABDC$ , then the cover belongs to  $COV(2, 1)$ . If tile is a square  $EBCF$ , then the cover is in the class  $COV(4, 1)$ .



**Fig. 1.** a) Optimal cover with equal disks. b) Optimal cover with equal ellipses.

Minimum density of the plane coverage with equal disks or ellipses does not depend on their size and is equal to  $2\pi/\sqrt{27} \approx 1.2091$ . A completely different situation with sectors. Denote the sector as a pair  $(R, \alpha)$ , where  $R$  is the radius, and  $\alpha$  is the angle of a sector. Then the density of the cover  $S1 \in COV(1, 1)$  (Fig. 2a) depends on  $\alpha$  and equals  $D(\alpha) = \alpha/\sin \alpha$ , and  $D(\alpha) \rightarrow 1$  when  $\alpha \rightarrow 0$ .

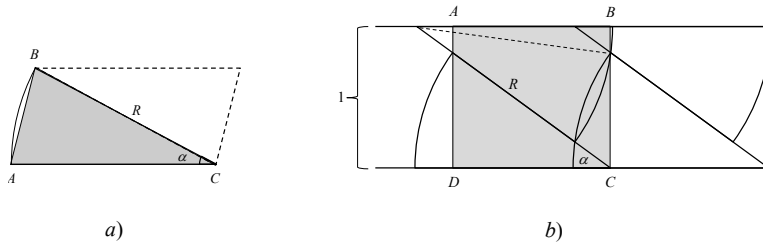


Fig. 2. a) Plane coverage model S1; b) Stripe coverage model M1.

But, starting from a certain angle, the values of the function  $D(\alpha) = \alpha/\sin \alpha$  become more than  $2\pi/\sqrt{27} \approx 1.2091$ , and the optimal coverage with equal disks or ellipses is preferred. So, if the parameters  $(R, \alpha)$  are fixed, it is necessary to find the best coverage model.

Let consider now the stripe coverage with equal disks, ellipses and sectors.

As we showed in [3,5] the coverage density with equal disks or ellipses tends to  $2\pi/\sqrt{27} \approx 1.2091$  when the number of disks covering one tile tends to infinity.

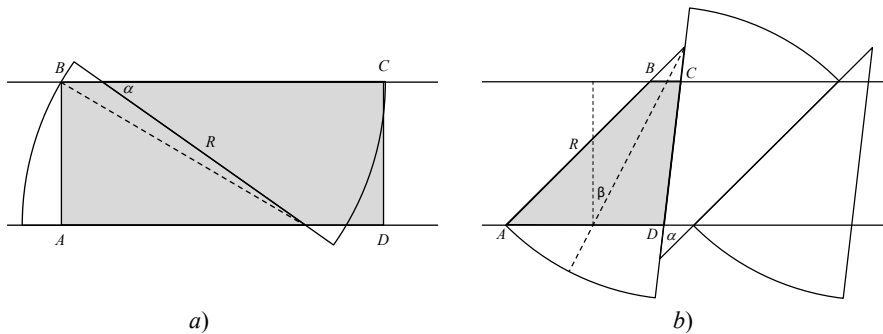


Fig. 3. Stripe covering with equal sectors: a) Model M2; b) Model M3.

If equal sectors are used to cover a stripe, then we propose three efficient models  $M1$ ,  $M2$  and  $M3$  (Fig. 2b and Fig. 3). It was found that (but not published yet), depending on the parameters of a sector, the best (having a minimum density) may be any of these covers. Figure 4 shows the areas of preference of the coverage models, depending on the parameters of the sector. In the blue area model  $M1$  is the best, in the green area model  $M2$  has the lowest density, and in the red area model  $M3$  is preferable. Set, for example,  $\alpha = 36^\circ$ . If  $R = 0.96$ , then model  $M1$  has the lowest density among the models  $M1, M2, M3$ ; if  $R = 1.2$ , then the best model is  $M3$ ; if  $R = 1.6$ , again the best model is  $M1$ ; and if  $R > 1.9$ , then model  $M2$  has the lowest density among the models  $M1, M2, M3$ .

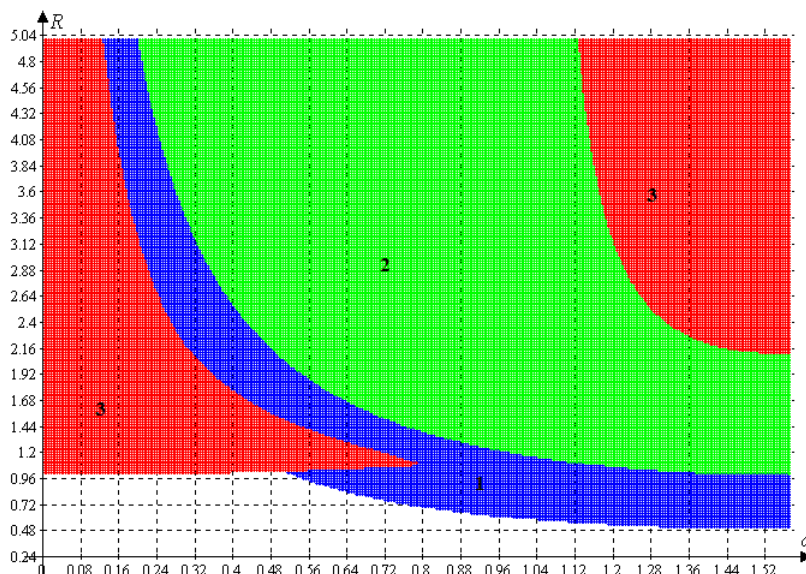


Fig. 4. The areas of preference.

Of course, if there are no any bounds on the  $\alpha$  and  $R$ , then one can set  $R \sin \alpha = 1$ , and the density  $\alpha / \sin \alpha$  of the cover (like  $S1$ ) tends to 1 when  $\alpha$  tends to 0 (in turn,  $R$  tends to infinity).

However, in practice, the sector angle may not be less than some positive number. Recently we proved the

**Theorem 1.** *If  $\alpha \in [\pi/180, \pi/2]$ , then the minimum density of stripe covering with equal sector does not exceed 1.000051.*

### 3 Two Types of Figures

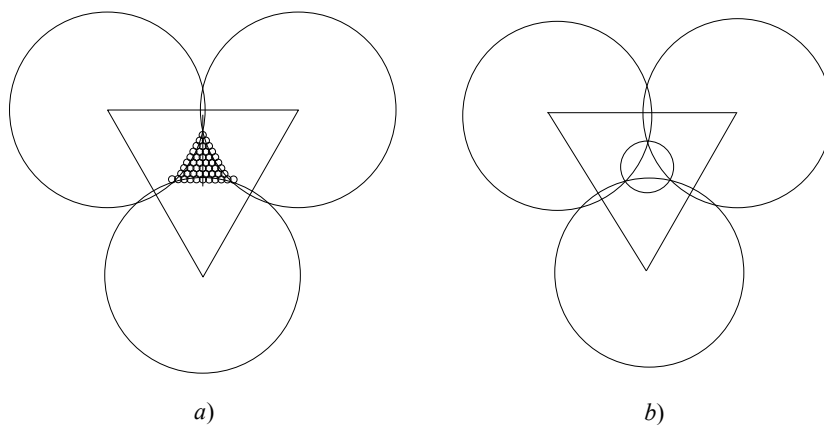
In [9] proposed a cover with two types of disks (Fig. 5a) which density tends to 1.0189 when the number of small disks tends to the infinity. It is a strong result, but not of practical importance because of the unlimited number of circles involved in the covering of one tile.

In [14] we proposed a plane coverage model with two types of disks (Fig. 5b). Its density, equals  $11\pi/\sqrt{972} \approx 1.1084$ , is minimal in the class  $COV(4, 2)$  when a tile is an equilateral triangle. One can apply AT to get the covers with ellipses in the different classes, but having the same density [6].

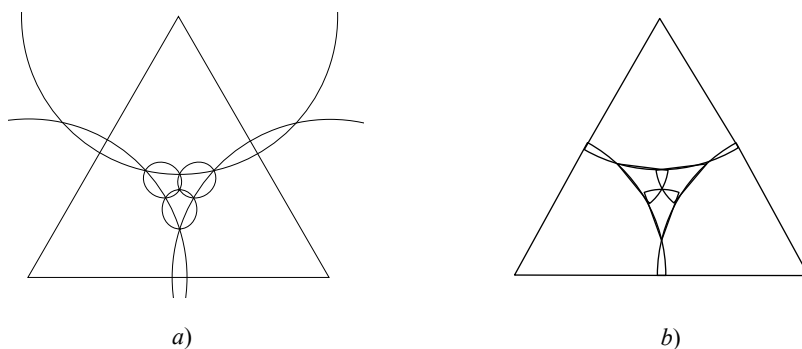
In [6] we proposed a new cover in the class  $COV(6, 2)$  with two types of ellipses having density  $D \approx 1.0786$  (Fig. 6a). If instead of ellipses use sectors, then we get another cover in the same class which density is  $D \approx 1.0321$ .

In the regular cover, a tile typically has the shape of a regular polygon (triangle, square or hexagon). Finally, we consider a regular cover using equilateral triangle with two types of sectors and estimate its density depending on the number of small sectors.

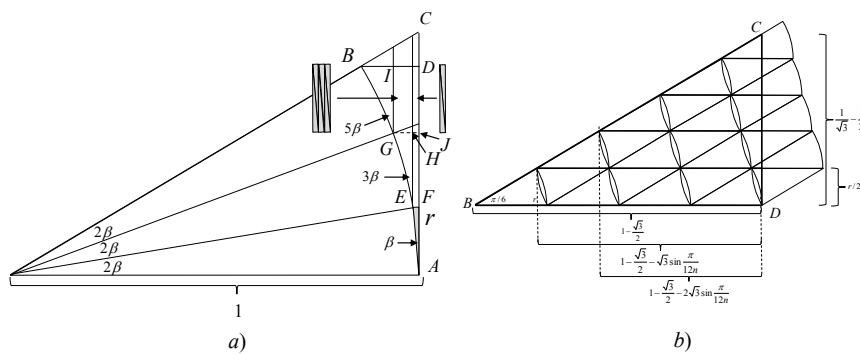
Without loss of generality, set the length of the side of the triangular tile equal 2. The proposed cover  $ST_2$  consists of three equal sectors of radius 1 with vertices at the nodes of the triangle. A uncovered curvilinear triangle in the center of the tile can be split into six equal curvilinear triangles. Let us consider one of them – curvilinear triangle  $ABC$  which is covered with equal sectors  $(r, \beta)$ ,  $r = 2 \sin \beta$ ,  $\beta = \pi/12n$  (Fig. 7a), as follows.



**Fig. 5.** a) Cover in the class  $COV(\infty, 2)$  which density tends to 1.0189 when the number of small disks tends to  $\infty$ ; b) Cover in the class  $COV(4, 2)$ , which density is  $11\pi/\sqrt{972} \approx 1.1084$ .



**Fig. 6.** Cover in the class  $COV(6, 2)$ : a) with ellipses, density is  $D \approx 1.0786$ ; b) with sectors, density is  $D \approx 1.0321$ .



**Fig. 7.** a) Example of the cover  $ST_2$  when  $n = 3$ ; b) Coverage of triangle  $BCD$ .

Triangle  $AEF$  is covered by 1 sector  $(r, \beta)$ , triangle  $EGH$  – by 3 non-overlapping sectors, triangle  $GBI$  – by 5 non-overlapping sectors, and so on (Fig. 7a). Just to cover all curvilinear triangles enough

$$Q_t = \sum_{k=1}^n (2k - 1) = n^2 \tag{1}$$

sectors  $(r, \beta)$ .

Then if we take 2 sectors directed in opposite directions with a common side, they cover the rectangle  $EHJF$ . The rectangle  $GIDJ$  can be covered by 8 sectors. And so on. The total number of sectors covering all rectangles is equal to

$$Q_s = 2 \sum_{k=2}^n (k - 1)^2 = 2 \sum_{k=1}^{n-1} k^2 = \frac{(n - 1)n(2n - 1)}{3}. \tag{2}$$

Note that the form of the triangle  $BCD$  does not depend on sectors  $(r, \beta)$ . The  $2n$  sectors with the common vertex form a sector  $(r, n\beta) = (r, \pi/6)$ . We cover the triangle  $BCD$  with rectangles whose height is  $r/2$  and length is reduced from  $1 - \sqrt{3}/2$  to  $r$ , and each rectangle, in turn, cover with the sectors  $(r, \pi/6)$  (Fig. 7b). As a result, the number of rectangles is

$$L = \left\lceil \frac{1/\sqrt{3} - 1/2}{r/2} \right\rceil = \left\lceil \frac{2 - \sqrt{3}}{2\sqrt{3} \sin \frac{\pi}{12n}} \right\rceil \leq \frac{2 - \sqrt{3}}{2\sqrt{3} \sin \frac{\pi}{12n}} + 1,$$

and the number of sectors  $(r, \beta)$  in  $l$ -th rectangle (counting from the bottom) is

$$q_l = 2n \left( \left\lceil \frac{1 - \sqrt{3}/2 - l\sqrt{3} \sin \frac{\pi}{12n}}{r} \right\rceil + 1 \right) \leq 2n \left( \frac{1 - \sqrt{3}/2 - l\sqrt{3} \sin \frac{\pi}{12n}}{\sin \frac{\pi}{12n}} + 3 \right).$$

Then the total number of sectors  $(r, \beta)$  covering the triangle  $BCD$  is

$$q = \sum_{l=1}^{L-1} q_l + 2n \leq 2n + 2n(L - 1) \left( 3 + \frac{2 - \sqrt{3}}{2 \sin \frac{\pi}{12n}} - L\sqrt{3}/2 \right). \tag{3}$$

**Theorem 2.** *The minimum coverage density of the equilateral triangle with two types of sectors tends to 1.00857 when the number  $n$  of sectors  $(2 \sin \frac{\pi}{12n}, \frac{\pi}{12n})$  tends to infinity.*

*Proof.* The total number of sectors  $(2 \sin \frac{\pi}{12n}, \frac{\pi}{12n})$  sufficient to cover the curvilinear triangle  $ABC$  is  $Q = Q_t + Q_s + q$ . Using (1)-(3), we get

$$Q \leq n \left( n + 2 + (n - 1)(2n - 1)/3 + \frac{((12 - 2\sqrt{3}) \sin \frac{\pi}{12n} + 2 - \sqrt{3})(2 - \sqrt{3})}{4\sqrt{3} \sin^2 \frac{\pi}{12n}} \right).$$

The area of one sector equals

$$s = r^2 \beta / 2 = \frac{\pi}{24n} \left( 2 \sin \frac{\pi}{12n} \right)^2.$$

The density of the cover is  $D(n) = (\pi/12 + Qs)/S$ , where  $S = 1/\sqrt{12}$ . Then

$$D(n) \leq \frac{\pi\sqrt{3}}{6} + \frac{\pi\sqrt{3}}{3} \sin^2 \frac{\pi}{12n} \left\{ n + 2 + \frac{(n - 1)(2n - 1)}{3} + \right.$$

$$\left. \frac{((12 - 2\sqrt{3}) \sin \frac{\pi}{12n} + 2 - \sqrt{3})(2 - \sqrt{3})}{4\sqrt{3} \sin^2 \frac{\pi}{12n}} \right\}. \quad (4)$$

So, it is easy to show that

$$\lim_{n \rightarrow +\infty} D(n) \leq 1.00857.$$

The proof is over.

We found the upper bound for the density depending on the number of sectors in the cover  $n$  (4). In particular,  $\lim_{n \rightarrow +\infty} D(n) \leq 1.00857$  which is much less than the coverage density with disks of two radii (which tends to 1.0189 when the number of disks tends to infinity) [9]. Moreover,  $D(n) \leq 1.00859 < 1.0189$  when the number of sectors is finite, for example,  $n \geq 10000$ .

**Acknowledgments.** This research was supported jointly by the Russian Foundation for Basic Research (grant 13-07-00139) and the Ministry of Education and Science of the Republic of Kazakhstan (grant 0115PK00550).

## References

1. Ai, J., Abouzeid ,A.A.: Coverage by Directional Sensors in Randomly Deployed Wireless Sensor Networks. *J. of Combinatorial Optimization*, 11, 21–41 (2006)
2. Astrakov, S.N., Erzin, A.I.: Efficient band monitoring with sensors outer positioning. *Optimization: A J. of Mat. Programming and OR*, 62(10), 1367–1378 (2013)
3. Astrakov, S.N., Erzin, A.I.: Sensor Networks and Stripe Covering with Ellipses. *Vychislitel'nie Tekhnologii*, 18(2), 3–11 (2013) (in Russian)
4. Deshpande, N., Shaligram, A.: Energy Saving in WNS with Directed Connectivity. *Wireless Sensor Network*, 5, 121–126 (2013)
5. Erzin, A.I., Astrakov, S.N.: Min-Density Stripe Covering and Applications in Sensor Networks. In: Goos, G., Hartmanis, J., Leeuwen, J. (eds.) ICCSA 2011. LNCS, vol. 6784, pp. 152–162. Springer, Heidelberg (2011)
6. Erzin, A.I., Astrakov, S.N.: Covering a Plane with Ellipses. *Optimization: A J. of Mat. Programming and OR*, 62(10), 1357–1366 (2013)
7. Erzin, A., Shabelnikova, N.: Optimal Regular Covering of the Plane with Equal Sectors. In: 20th conf. of the Int. federation of operational research societies (IFORS 2014), pp. 69–69, Barcelona (2014)
8. Fan, G., Jin, S.: Coverage Problem in Wireless Sensor Network: a Survey. *Journal of Networks*, 5(9), 1033–1040 (2010)
9. Fejes Tóth, G.: Covering the Plane with Two Kinds of Circles. *Discrete & Computational Geometry*. 13(3), 445–457 (1995)
10. Fejes Tóth, L.: Lagerungen in der Ebene auf der Kugel und im Raum. Springer-Verlag, Berlin (1953)
11. Guvensan, M.A., Yavuz, A.G.: On Coverage Issues in Directional Sensor Networks. A survey. *Ad Hoc Networks*, 9, 1238–1255 (2011)
12. Ismailescu, D., Kim, B.: Packing and Covering with Centrally Symmetric Convex Disks. *Discrete and Computational Geometry*, 51, 495–508 (2014)
13. Kershner, R.: The Number of Circles Covering a Set. *American Journal of Mathematics*. 61(3), 665–671 (1939)
14. Zalyubovskiy, V., Erzin, A., Astrakov, S., Choo, H.: Energy-Efficient Area Coverage by Sensors with Adjustable Ranges. *Sensors*, 9, 2446–2469 (2009)



# VNS-Based Heuristics for Communication Tree Optimal Synthesis Problem

Adil Erzin, Nenad Mladenovic, and Roman Plotnikov

Sobolev Institute of Mathematics, Novosibirsk, Russia

<http://www.math.nsc.ru/>,

University of Valenciennes and Hainaut-Cambresis, Famars, France

<http://www.univ-valenciennes.fr/>,

Novosibirsk State University, Novosibirsk Russia

<http://www.nsu.ru/>

**Abstract.** In this paper we consider a problem of optimal communication tree construction in edge-weighted graph which occurs in wireless sensor networks while minimizing the power consumption of data transmission. The considered problem is strongly NP-hard, therefore construction of efficient approximation algorithms is one of the most important objective. In this paper we propose several metaheuristics based on variable neighborhood search for the approximation solution of the problem. We have performed extensive comparative analysis among proposed methods and known approaches. The executed numerical experiments demonstrate high efficiency of the proposed heuristics.

**Keywords:** wireless sensor networks, energy consumption, variable neighborhood search.

## 1 Introduction

Elements of different communication networks use wireless communication for data exchange. Herewith energy consumption of a network's element is proportional to  $d^s$ , where  $s \geq 2$ , and  $d$  is a transmission range [1]. In some networks, e.g., in the wireless sensor networks, each element (sensor) has a limited energy storage, and its efficient use results in the lifetime extension of a whole network [10,11,12]. For the rational energy usage modern sensor can adjust its transmission range. Then the problem is to find a transmission range (the transmitter power) for each element that supports a connected subgraph in order to minimize the total energy consumed. If one suppose equal signal propagation in all directions, then all elements inside the disk, which radius is equal to the transmission range, receive the data. In this case we can suppose that the communication network (a spanning subgraph whose edges are used for the data translation) is a complete graph [1,3,9,10]. However, the signal is not always spread equally in all directions and at any distance. Thus, in general case, it is necessary to consider arbitrary communication graph  $G = (V, E)$ . A communication energy consumption over each edge could be arbitrary too. If  $c_{ij} \geq 0$  is a transmission-related energy consumption needed for sending data from  $i \in V$  to  $j \in V$ , then in the connected subgraph  $T = (V, E')$ ,  $E' \subseteq E$  the energy consumption of the node  $i \in V$  equals to  $E_i(T) = \max_{j:(i,j) \in E'} c_{ij}$ . The goal of this paper is development of algorithms for construction of spanning subgraph  $T$ , that minimizes  $\sum_{i \in V} E_i(T)$ . Without loss of generality, we assume that the subgraph  $T$  is a spanning tree. Then the problem can be formulated in the following way.

Given the simple undirected weighted graph  $G = (V, E)$  with a vertex set  $V$ ,  $|V| = n$ , and an edge set  $E$ , find a spanning tree  $T^*$  in  $G$  which is the solution to the problem:

$$(\min)W(T) = \sum_{i \in V} \max_{j \in V_i(T)} c_{ij}, \quad (1)$$

where  $V_i(T)$  is the set of vertices adjacent to the vertex  $i$  in the tree  $T$ , and  $c_{ij} \geq 0$  be the weight of the edge  $(i, j) \in E$ .

Any feasible solution of (1), i.e. a spanning tree in  $G$ , will be called a *communication tree* (subgraph). It is known that (1) is strongly NP-hard [1,5,6,9] and if  $N \neq NP$ , then the problem is inapproximable within a ratio  $1 + \frac{1}{260}$  [6]. Therefore, construction and analysis of efficient approximation algorithms is one of the most important issues regarding to the research of this problem.

It is shown in [1] that a minimal spanning tree (a spanning tree with minimal total edge weights) is a 2-approximation solution to the problem (1). In [5] a more precise ratio estimate for the minimal spanning tree algorithm for this problem is reported. In the same paper several heuristic algorithms are proposed and their a posteriori analysis is performed. Since we were not completely satisfied with the quality of the results obtained, in [4] we propose a hybrid heuristics that combines genetic algorithm and a variable neighborhood search [8]. There two new local search heuristics for the problem (called LI and VND) are proposed. Then they are used as a mutation operator within the genetic algorithm (GA). For the sake of completeness, we will recall both local searches in this paper. The computational results show the high efficiency of the proposed hybrid heuristic.

In this paper we propose different heuristics based on the variable neighborhood search (VNS) metaheuristic [8,2,7]. Contribution of this paper may be summarized as follows.

- New local search that is based on elementary tree transformation (ETT) is proposed. In terms of solution quality it significantly outperforms the previous one (named as LI), but uses more computation time;
- Several Basic VNS and General VNS based heuristics are proposed and tested. Some of these new heuristics give results of better quality than recent state-of-the-art (hybrid heuristic [4]), especially for solving more realistic large size problems.

In the next section the rules of the new VNS based heuristics are given. In section 3 extensive computational analysis is performed, while section 4 concludes the paper.

## 2 Variable neighborhood search based heuristics

As mentioned earlier, we use VNS metaheuristic to get approximate solution of (1). Descriptions of VNS can be found in [7,8]. Let us briefly describe the main ideas of VNS. The *Basic VNS* consists of two phases: *shaking* — obtaining a random solution from the appropriate neighborhood of the given solution and local search. On each step of the algorithm the neighborhood for the shaking is chosen within a predefined neighborhood structure and the local search is performed after that, then an obtained solution replaces a current one iff it is better. *Variable neighborhood descent* (VND) is deterministic algorithm where local search is iteratively performed within elements of appropriate neighborhood structure. *General VNS* is a kind of Basic VNS where VND is used as local search. Further we will call Basic VNS all VNS-based heuristics where the only local search method is used, otherwise (i.e. when the number of used local searches exceeds one) a term General VNS will be used.

The first local search procedure we use is called Local Improvement (LI) [4]. All arcs are sorted by decreasing order of objective function, if the corresponding arc (edge) is added to the tree. Then the best possible deletion is performed. This procedure is repeated while solution has been improved.

The second local search algorithm is based on Elementary Tree Transformation (ETT). Within each iteration of the loop, all non-adjacent vertex pairs are considered for possible addition

to the current tree  $T$ . Then the arc with maximum *drop* from the obtained cycle is removed (we call drop of an edge the increment of objective function obtained by adding the edge into a graph). This procedure is also repeated while solution is improved.

The shaking procedure within a neighborhood  $\mathcal{N}_k(T)$  of the tree  $T$  consists of  $k$ -times repetition of the following arc replacements. First, two different vertices from the original graph that are not in the current tree (in the arc subset  $A$  that defines a tree) are selected. Then this arc is added into a tree and a random arc is excluded from the obtained cycle. As a result one gets a random point from  $\mathcal{N}_k(T)$ . Let us introduce the parameter  $k_{\max}$  — it is the maximum number of arc replacements in the Shaking procedure. The best value of this parameter is estimated experimentally.

We propose two Basic VNS heuristics that use LI and ETT local searches:

- BVNS\_LI: LI is used as a local search within Basic VNS;
- BVNS\_ETT: ETT is used as a local search.

Both algorithms LI and ETT can be combined in General VNS algorithm, using their different order. Let us introduce modifications of LI and ETT — LI\_1 and ETT\_1 where the best neighbor is searched instead of local optimum. The following General VNS algorithms are proposed:

- GVNS\_11: VND local search uses LI first, and then ETT;
- GVNS\_12: VND local search uses ETT first, and then LI;
- GVNS\_21: VND local search uses LI\_1 and then ETT\_1 sequentially;
- GVNS\_22: VND local search uses ETT\_1 and then LI\_1;
- GVNS\_N: VND local search uses  $N_1$ ,  $N_2$  and  $N_3$  proposed in [4].

Thus, we propose and compare five variants of GVNS schema.

### 3 Computational results

All proposed algorithms have been implemented on C++ using the Visual Studio 2010 IDE. A simulation has been executed for  $n = 10, 15, 20, 25, 30, 50, 150, 200$ . For the same dimension  $n$ , 100 different instances have been generated. For each instance  $n$  vertices have been generated in the square and weights of the edges are calculated as squared Euclidean distances. Then a minimal spanning tree as an initial approximate solution in all algorithms has been constructed. As a data structure for storing a feasible solution of the problem we use a tree, where each element stores the pointer to its parent vertex (or null-pointer in a case of root) and the list of its direct successors. The experiment has been executed at Intel Core i5-3470 (3.2GHz) 8Gb machine.

An integer linear programming (ILP) formulation is proposed in [5]. It was used to get optimal solutions using IBM ILOG CPLEX package for small-sized problems. For  $n \leq 30$  we compute the exact value of *ratio*, which is expressed as  $W_A(T)/W(T^*)$ , where  $W_A(T)$  is a value of objective function of the solution constructed by algorithm  $A$ , and  $W(T^*)$  is an optimal value of the objective function. In the case of larger  $n$ , the upper bound of ratio is calculated as  $W_A(T)/LB(W(T^*))$ , where  $LB(W(T^*))$  is a lower bound of  $W(T^*)$ . Value  $LB(W(T^*))$  represents a sum of weights of minimal spanning tree edges (see [5] for details).

It is necessary to define the value of parameter  $k_{\max}$  for the VNS-based heuristics. For this purpose each algorithm executed using different values of  $k_{\max}$ . It appeared that beginning from  $k_{\max} = 30$ , in average, ratio of obtained solution does not decrease significantly, whereas runtime of some algorithms increased up to twice while increasing of  $k_{\max}$  by 10. Wherein, in average,

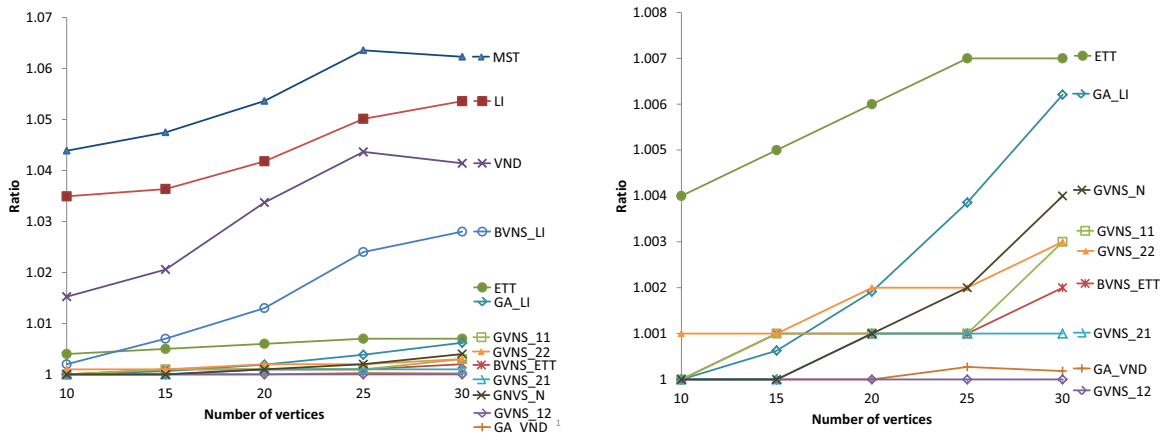


Fig. 1. Ratio

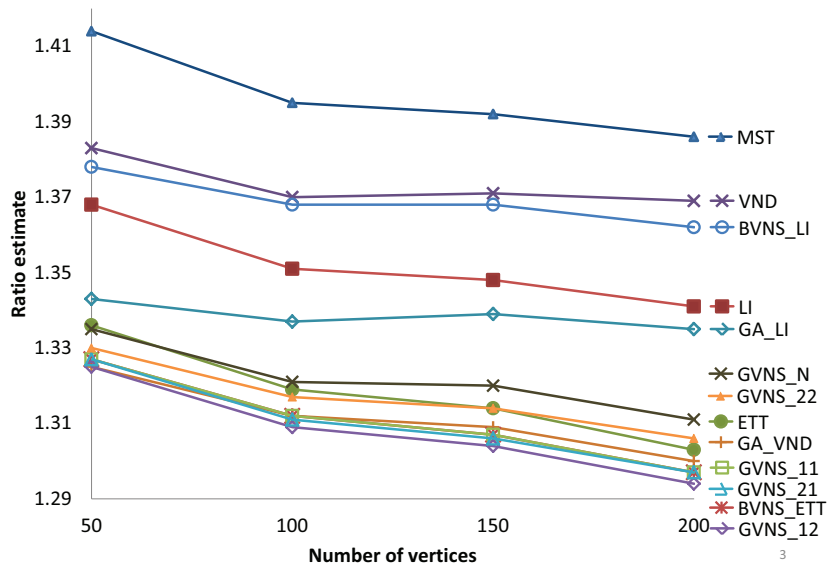


Fig. 2. Ratio upper estimate

runtime of all algorithms remains accessible for  $k_{\max} = 30$ . Therefore, in all VNS-based algorithms we take  $k_{\max} = 30$ .

The following algorithms have been described in section 2: LI, ETT, VND, BVNS\_LI, BVNS\_ETT, GVNS\_11, GVNS\_12, GVNS\_21, GVNS\_22 and GVNS\_N. Old heuristics included in comparison are: hybrid genetic algorithms GA\_LI and GA\_VND were described in [4]. A minimal spanning tree is denoted as MST.

In figure 1 the ratios of solutions yielded by the algorithms for  $n \leq 30$  are presented. In average, in cases when exact values of ratio could be computed (i.e. when  $n \leq 30$ ), algorithms GA\_VND and GVNS\_\* yield solutions which differ from the optimal one not more than 0.6%. At the same time, the most close to optimal are solutions constructed by GVNS\_12 and GA\_VND: they both yield the ratios which do not exceed 1.0003. It appears that a hybrid genetic algorithm GA\_VND, which uses VND as mutation operator, yields better solution on

average than GVNS\_N (that is based on the same VND procedure). However, the second one is faster (see fig. 4). It should be noted that the algorithm ETT in average yields a solution which differs from the optimal one at most 0.7% (For comparison, in average, solution yielded by another local search LI is about 3% from the optimal solution when  $25 \leq n \leq 30$ ).

The experimental results for a cases  $50 \leq n \leq 200$  (see figure 2) show that the main trends found in a case of small values of  $n$  regarding to the majority of algorithms, persist for the larger dimensions. But one can distinguish an algorithm GA\_LI, which solution quality becomes significantly worse in relation to algorithms GVNS\_\*, ETT and BVNS\_ETT while  $n$  grows. Ratio estimate graphics of the last ones are almost parallel to each other, and this means that they have the similar dynamics of ratio estimate changing with  $n$  growth. The most accurate solution is again constructed by GVNS\_12. Further, in decreasing order there depicted the graphics of ratio estimates of the BVNS\_ETT, GVNS\_21, GVNS\_11 (these three algorithms have almost confluent graphics of ratio estimate), GA\_VND, ETT, GVNS\_22, GVNS\_N. Herein the difference between the maximum value of the estimates of the mentioned algorithms and the minimum one does not exceed 2%: e.g., when  $n = 200$  ratio estimate of GVNS\_12 is 29%, and ratio estimate of GVNS\_N is 31%.

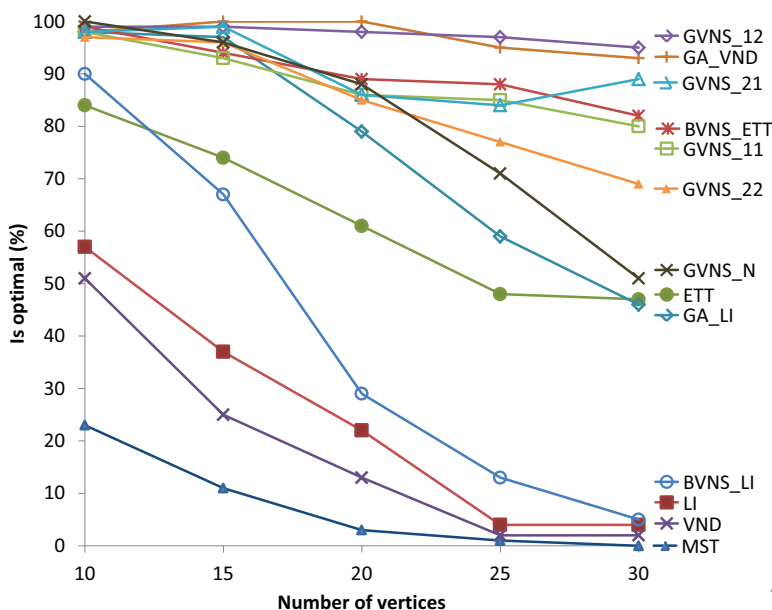


Fig. 3. Percentage of cases when the optimal solution was obtained by algorithm

In figure 3 a percentage of cases when optimal solutions were constructed is presented. One can see that in a case of small dimension both algorithms GVNS\_12 and GA\_VND almost always construct an optimal solution: in a worst case (when  $n = 30$ ) the algorithm GVNS\_12 constructs an optimal solution in 95% of cases, and GA\_VND — in 93% of cases (again, when  $n = 30$ ). It is seen, that the percentage of optimality of other algorithms falls down while  $n$  grows. These algorithms can be conventionally divided into 2 groups: a) GVNS\_21, BVNS\_ETT, GVNS\_11, GVNS\_22, GVNS\_N, ETT and GA\_LI, which optimality percentage is not less than 45%, and b) MST, VND, LI, BVNS\_LI, which optimality percentage in the worst case (i.e. when  $n = 30$ ) does not exceed 5%. In general, it matches results presented on fig. 1 and 2.

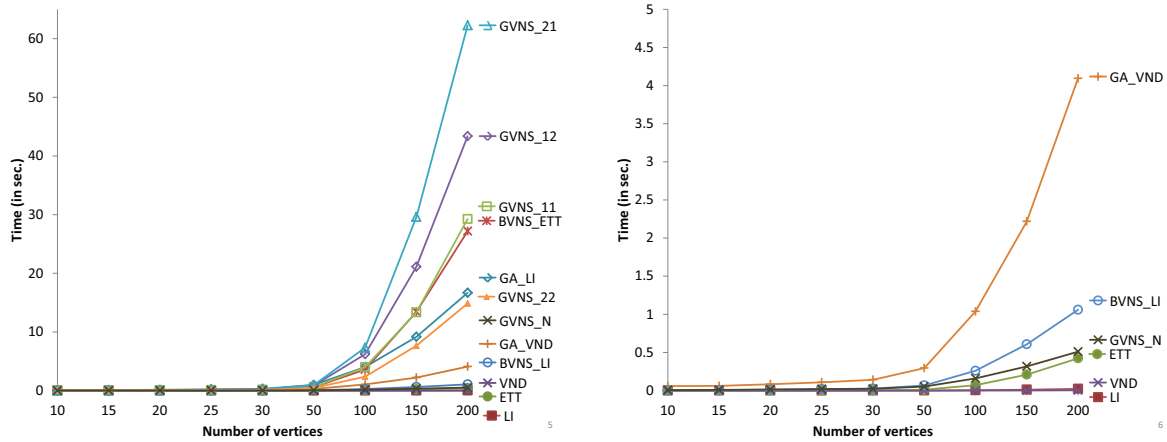


Fig. 4. Runtime

Graphics in the figure 4 represent runtime as a function of dimension  $n$ . Notice that the algorithms GVNS\_N and ETT work rather fast. When  $n = 200$  in average they solve the problem in less than 0.5 seconds; at the same time, they yield rather accurate solution on average. The most accurate algorithm GVNS\_12 spends around 43 seconds when  $n = 200$ , but the algorithm GA\_VND, which ratio close to the ratio of GVNS\_12, solves the problem more than 10 times faster when  $n = 200$ . Here we should note that all genetic algorithms are well-parallelized, and therefore they used 4 parallel threads. Algorithm GVNS\_21 turned out to be the slowest one: on average, it spends about 60 seconds when  $n = 200$ . All other VNS-based heuristics turned out to be faster than GVNS\_12 and slower than GA\_VND. Notice the algorithm BVNS\_ETT which uses on around 29 seconds when  $n = 200$  on average. In a case of larger dimensions, it also provides more exact solution than GA\_VND.

## 4 Conclusion

In this paper we propose new heuristics based on the Variable Neighborhood Search (VNS) approach for finding approximate solutions to the optimal communication tree synthesis problem. Proposed algorithms were compared with each other and also with those recently proposed in [4]. The majority of proposed VNS based variants appeared to be rather effective and efficient. However, the most effective on average appeared to be the General VNS algorithm, named GVNS\_12, where elementary tree transformation (ETT) and so-called Local improvement (LI) are sequentially used as a local search routines within VNS. In a case of large dimension algorithms GVNS\_21 and BVNS\_ETT on average yield the better solution than hybrid genetic algorithm GA\_VND from [4], but spent more computational time. Also a ETT local search proposed in this paper takes more time than the known LI, but it constructs more accurate solution. In general, in a case of successive choice of the neighborhood structures for the local search, the VNS approach is justified for the efficient solution of (1).

Future work may include the use of different intensified shaking procedures to improve the efficiency and the effectiveness of our new heuristics. Parallelization is also an obvious option in that respect.

**Acknowledgement.** This research was supported jointly by the Russian Foundation for Basic Research (grant 13-07-00139) and the Ministry of Education and Science of the Republic of Kazakhstan (grant 0115PK00550).

## References

1. Althaus, E. et al.: Power efficient range assignment for symmetric connectivity in static ad hoc wireless networks, *Wireless Networks* 12 (3) 287–299 (2006)
2. Brimberg, J., Urosevic, D., Mladenovic, N.: Variable neighborhood search for the vertex weighted k-cardinality tree, *European J. of Operational Research* 171 74–84 (2006)
3. Carmi, P., Katz, M.: Power assignment in radio networks with two power levels, *Algorithmica* (47) 183–201 (2007)
4. Erzin, A., Plotnikov, R.: Using VNS for the optimal synthesis of the communication tree in wireless sensor networks, *Electronic Notes in Discrete Mathematics* 47 21–28 (2015)
5. Erzin, A., Plotnikov, R., Shamardin, Y.: On some polynomially solvable cases and approximate algorithms in the optimal communication tree construction problem, *Journal of Applied and Industrial Mathematics* 7 142–152 (2013)
6. B. Fuchs, On the hardness of range assignment problems, Tech. Rep. TR05-113, *Electronic Colloquium on Computational Complexity* (2005).
7. Hanafi, S., Lazic, J. Mladenovic, N. Wilbaut, C., Crevits, I.: New variable neighbourhood search based 0-1 MIP heuristics, *Yugoslav Journal of Operations Research*, DOI: 10.2298/YJOR140219014H (2015)
8. Hansen, P., Mladenovic, N.: Variable neighborhood search: Principles and applications, *European Journal of Operational Research* 130 449–467 (2001)
9. Kirousis, L., Kranakis, E., Krizanc, D., Pelc, A.: Power consumption in packet radio networks, *Theoretical Computer Science* (243) 289–305 (2000)
10. Pottie, G., Kaiser, W.: Wireless integrated network sensors, *Communications ACM* 43 (5) 51–58 (2000)
11. Wu, J., Yang, S.: Energy-efficient node scheduling models in sensor networks with adjustable ranges, *Int. J. of Foundations of Computer Science* 16 (1) 3–17 (2005)
12. Zhang, H., Hou, J.: Maintaining sensing coverage and connectivity in large sensor networks, *Ad Hoc & Sensor Wireless Networks* 1 (1–2) 89–124 (2005)

# Classification of Scientific Documents Based on the Compression Methods

A. Guskov<sup>1,2</sup>, B. Ryabko<sup>2,1</sup>, A. Zubkov<sup>3</sup>

<sup>1</sup>The State Public Scientific Technological Library of Siberian Branch of  
the Russian Academy of Science,

<sup>2</sup>Institute of Computational Technology of Siberian Branch of  
the Russian Academy of Science,

<sup>3</sup>Novosibirsk State University.

## 1 Introduction

World flow of scientific documents is constantly growing: scientometric databases are indexing increasing number of research results, published in a variety of forms: books, articles, conference proceedings and other patents. The annual volume of publications in many disciplines has become so large that it is not possible for an individual researcher or an entire laboratory to be fully aware of the relevant research. In such circumstances, there is a certain crisis of the concept of "expert as a specialist who understands everything that happens in a particular subject area. Under these conditions, the process of information support of scientific researches plays a key role.

One of the most difficult tasks of information support is the process of automation the thematic classification of documents, the result of which is assigning a document to one or more classes (e.g. mathematics, physics, chemistry, etc.). Special attention should be paid to the issues of formation of such classifiers. For existing classifications (e.g. GRNTI, UDC, BBK) are even more difficult questions of comparison and updating [1-3]. The object of this study is the very process of the document referring to one or more classes of a priori given. Currently, this process is mainly carried out in manual mode: expert looks at the title, abstract and full text, and then makes a decision on assigning a suitable category. Another common approach is the automatic assignment to the articles all the categories to which the scientific journals refers itself. In the national citation database Russian Science Citation Index (elibrary.ru) a hybrid approach is used when editor assigns one of the categories to which the journal is related while uploading the articles (amount of such categories is usually in the range of from 3 to 7). Obviously, each of these approaches has significant drawbacks. The work of the expert has a high unit cost, and it is not applicable to big publication sets. Classification of articles "by journal" has a low reliability and could not be used for compartmentalized classifications. The same applies to the combined version, although this method appears suboptimal among others.

How is the process of classification performed by an expert? First, the titles and abstracts of publication are analyzed to identify specific terms and semantic structures. The expert collates them to his own term vocabulary and knowledge. If the expert can "see" similar to the one or more subjects of publications known to him, he assigns the appropriate class to publication. If the information is insufficient, the expert looks through the full text of the publication, repeating the same analytical process. Thus, one of the main criteria for classifying documents is similar terminology between classified paper and the articles corresponding to this class. This study uses this empirical observation to build a mathematical model for this process for the automation purposes.

There was a number of study related problem for clustering documents. In general, clustering algorithms can be divided into the following types: heuristic graph algorithms, statistical



algorithms, hierarchical clustering algorithms. In [4-5] the analysis of document clustering methods for automation tasks is completed, and an approach for clustering based on the extraction of the key terms is proposed. In [6] a modification of the traditional model is proposed, where each document is represented as a vector of the frequency of single words and N-grams. In [7] a new clustering algorithm is developed that uses a hierarchical approach, in which the documents are regarded not just as a set of words, but vectors of concepts based on ontology WordNet. In [8], two algorithms are considered for the descending and hierarchical clustering using the frequency of occurrence of word vectors. In [9-11] an overview of methods and algorithms provided for different types of clustering: graph, hierarchical and statistical.

In this paper, we propose an approach to automatic classification which is based on estimations of the Kolmogorov complexity of texts. The estimations will be obtained by so-called archivers, or programs for lossless compression. This approach was proposed by P. Vitanyi (see [12]) and was successfully applied to the classification of the degree of similarity of natural languages, musical pieces of different genres [12-13], the works of writers, biological "texts"[14], computer viruses and many other objects (see the review in [12]). Independently, the data compression methods applied to problems of mathematical statistics, time series prediction and many other problems; see the review in [15].

The main idea of this approach is based on the concept of Kolmogorov complexity of texts, which, informally, is equal to the minimum length of a computer program that generates the text. It is important that this value is almost independent from the computer (a universal Turing machine). More precisely, if different universal Turing machines are used, the difference between two complexities will be limited by a constant for any text; the formal definition can be found in [16]. It turns out that the length of the "compressed" text is quite effective upper bound of Kolmogorov complexity and it is useful for many applications; see [12,17].

In this paper we propose to use data compression methods in order to automatically determine a thematic affiliation of scientific texts. The obtained results are preliminary, but show that this approach is of practical interest.

## 2 The method

The main idea of the suggested method is quite natural: scientific texts (articles, books, etc.) use similar terminology if they belong to the same area. On the other hand, the archiver uses frequencies of occurrence of words in the text and "compresses" the data the better, the more repeated words. Based on this observation, we suggest the following classification scheme: for any scientific area we form a set of papers, which represents the area. Then the text, whose thematic belonging must be determined, is compressed together with each set of texts representing the thematic areas. Then the text refers to that area for which it is compressed to a minimum size (i.e., "better" is more compressed).

Let us give a more formal description of the method, considering first the situation where the subject areas are not "nested" within each other, i.e. one is not part of another (e.g. "algebra" is a part of the field of "mathematics" but not a part of the field "geometry"). Now, let the given subject area  $A, B, C, D$ . For each of them an expert takes a variety of typical texts, which we denote by  $a_1, a_2, \dots, a_n$ ;  $b_1, b_2, \dots, b_m$ ;  $c_1, c_2, \dots, c_k$ ;  $d_1, \dots, d_l$ , and let  $z$  be a text, which should be attributed to one of these areas. Let an archiver  $U$  be to compress texts. Denote the length of the "compressed" texts  $v_1, v_2, \dots, v_t$  through  $U(v_1 v_2 \dots v_t)$ . Define

$$U(z/v_1 v_2 \dots v_t) = U(v_1 v_2 \dots v_t z) - U(v_1 v_2 \dots v_t).$$

In the suggested method, we first calculate the values of  $U(z/a_1a_2\dots a_n)$ ,  $U(z/b_1b_2 \dots b_t)$ , ...,  $U(z/d_1d_2 \dots d_l)$ , and then decide that the text  $z$  belongs to that area for which this value is minimal. Also we used a modified formula

$$u(z/v_1v_2\dots v_t) = U(z/v_1v_2\dots v_t)/U(z),$$

in order to estimate quantitatively the impact of area knowledge on the degree of compression of the text. Both estimates give the same result when assigning text to the subject area, but the second formula gives additional information about closeness of the text  $z$  to different areas that are of independent interest.

### 3 The results of the pilot experiment

#### 3.1 Preparation of initial data

For a preliminary assessment of the possible practical applications of this method, an experiment was conducted. We used data provided on the website arxiv.org to select subject domains and the formation of the texts describing them. Arxiv.org contains more than a million articles pertaining to various areas of science. When placing the article on the site, an author refers to the work of one of the scientific sections. For our experiment, we have chosen three research fields: Nuclear Theory, Calculation Complexity and Formal Languages (we will call them "classes"). Notable, we have chosen the classes so that the first is thematically "away" from the rest, whereas the second and third are "close".

Each class has been formed its "core"—a set of full texts, which are typical for this class. Formation was carried out by random selection of the paper relating to the corresponding area on the site arxiv.org. Full texts of articles were downloaded in PDF, then text layers were extracted. The experiment showed that for the original PDF-files, which contain a lot of visual information, the method still works, but with much worse results.

#### 3.2 Analysis and interpretation of results

Afterwards, we randomly extracted three full texts (as tests) for each class. The values of  $u(z/v_1v_2\dots v_t)$  were calculated for each combination of the nine tests and three class sets. Its percentage values are given in Table 1. The values, which are supposed to have the best performance ratio (according to our assumptions) are bolded.

As the table shows, the results were as follows:

- 6 test fits the hypothesis well ( $\text{diff} > 1\%$ ): *NTS1*, *NTS2*, *NTS3*, *CCS2*, *CCS3*, *FLS1*;
- 2 test fits the hypothesis moderately ( $0\% < \text{diff} < 1\%$ ): *CCS1*, *FLS2*;
- 1 test does not fit the hypothesis ( $\text{diff} < 0$ ): *FLS3*.

Analysis of the data showed that for the three texts of the Nuclear Theory, the value of  $u(z/v_1v_2\dots v_t)$  is minimal when the typical texts are taken from the same area. Thus, the pilot experiment for the proposed method of papers classification conducted quite successfully. There was only one test (*FLS3*), which was not assigned properly. However, as we noted above, the fields of Calculation Complexity and Formal Languages are thematically quite close, and the "bad classified" text stands at an intermediate position. Thereby, it could equally be assigned to both areas, which is consistent with the process of classification. Thus, a pilot experiment has shown that the proposed approach is efficient and requires more detailed study.

**Table 1.** Experimental results: Compression ratio of tests in combination with different cores (tests in rows, classes in columns).

	Test ID	Nuclear Theory	Calculation y Complexity	Formal Languages
Nuclear theory tests	<i>NTS1</i>	<i>30,92</i>	35,34	35,55
	<i>NTS2</i>	<i>34,05</i>	40,22	40,5
	<i>NTS3</i>	<i>35,61</i>	38,99	38,78
Calculation Complexity tests	<i>CCS1</i>	39,24	<i>36,69</i>	36,97
	<i>CCS2</i>	38,58	<i>33,46</i>	36,26
	<i>CCS3</i>	36,02	<i>33,56</i>	34,91
Formal Languages tests	<i>FLS1</i>	34,63	32,37	<i>27,57</i>
	<i>FLS2</i>	31,38	30,46	<i>30,2</i>
	<i>FLS3</i>	34,08	32,84	<i>33,67</i>

### 3.3 Planning of experiment

The next stage of the study is to conduct a full-scale series of experiments, which would be the base of the technology of automation the process of classification. As can be seen from the description, the proposed method has several parameters affecting the result: the size (number of documents) of the core classes, thematic closeness of classes, "purity" of the text (the presence of the formulas, tables and references), and compression algorithm. A series of experiments is conducted to answer the following questions:

1. What parameters values make this method work most efficiently?
2. What criteria should be used when assigning document to a class?

A series of experiments will be carried out under the following conditions:

1. The full text of the documents will be taken from the site *arxiv.org*.
2. The number of classes is  $N(N > 10)$ , and they will be selected from both adjacent domains and poorly connected to each other areas of science.
3. Each class contains the core of M documents (M=500); whether the document belongs to the core or not will be determined according to the classification from the database *arxiv.org*.
4. K tests will be randomly selected for each class, which are not contained by the core. Thus, the total number of tests will be  $K*N$ .
5. It is planned to conduct separate experiments and compare the results for the following lossless compression algorithms: LZMA, LZMA2, PPMd, Bzip2, Deflate, Deflate64.
6. For the most appropriate compression algorithm, several experiments for the "purification" of the full texts (without formulas, tables and references) and for various core size M (M=100, 300, 500) will be conducted.

Each experiment is a series of compressions of all possible combinations of cores and tests under certain specified test conditions specified in claims 1-6. Thus, the number of compressions for each experiment is  $K N^2$ . The result of each experiment is the matrix of values  $u(z/v_1v_2...v_M)$ .

The statistical analysis of this matrix determines the number of errors of the first and second kinds, and shows how effective the method for the given experimental parameters is. As a result of the series of experiments, the most effective set of parameters is determined, as well as a formal criterion for assigning a class to the document. Probably, this criterion will be formulated in the terms of reaching a certain compression threshold.

## 4 Conclusion

Thus, the pilot experiment demonstrated that the described method is of interest to explore it further. We plan to conduct the bulk of the series of experiments described in September 2015 to present the results at the International Conference "Computational and Informational Technologies in Science, Engineering and Education"(CITech-2015).

## Acknowledgments

B. Ryabko was partially supported by the Russian Foundation for Basic Research (grant no. 15-07-01851).

## References

1. Antopolskii A.B., Belozerov V.N., Markarova T.S., Dmitrieva E.Y. Establishing appropriate GRNTI rubrics for different classification systems of scientific and technical information // Nauchno-tehnicheskaja informacija. Serija 1: Organizacija i metodika informacionnoj raboty. 2015. No 3. PP. 3-18. (in Russian)
2. Shaburova N.N., Belozerov V.N. UDC classification system for the indexing of documents on physics of semiconductor // Nauchno-tehnicheskaja informacija. Serija 1: Organizacija i metodika informacionnoj raboty. 2010. No 9. PP. 34-44. (in Russian)
3. Zaytseva E.M., Anisimova V.P. Digital versions of classification systems: history, current status and technological features // Nauchno-tehnicheskaja informacija. Serija 1: Organizacija i metodika informacionnoj raboty. 2015. No 1. PP. 29-34. (in Russian)
4. Barakhnin V.B., Nekhaeva V.A., Fedotov A.M. Prescription similarity measure for clustering text documents // Vestnik NGU. Serija: Informacionnye tehnologii. 2008. Vol. 6, No 1. PP. 3-9. (in Russian)
5. Barakhnin V.B., Tkachev D. A. Clustering of text documents based on the composite key terms // Vestnik NGU. Serija: Informacionnye tehnologii. 2010. Vol. 8, No 2. PP. 5-14. (in Russian)
6. Miao Y., Keselj V, Milios E. Document Clustering using Character N-grams: A Comparative Evaluation with Term-based and Word-based Clustering <https://web.cs.dal.ca/~eem/cvWeb/pubs/Miao-CIKM-2005.pdf>
7. Baghel R., Dhir R. A Frequent Concepts Based Document Clustering Algorithm // International Journal of Computer Applications 2010 Vol. 4 No.5 C. 6- 12
8. Beil F., Ester M., Xu X. Frequent Term-Based Text Clustering //Proc. 8th Int. Conf. on Knowledge Discovery and Data Mining (KDD '2002), Edmonton, Alberta, Canada, 2002.
9. Schaeffer, S.E. Graph clustering// Computer Science Review 2007 Vol.1 No.1, C. 27-64
10. Voronotsov K.V. Clustering algorithms and multidimensional scaling. Course of lectures. MSU, 2007 (in Russian)
11. Barsegyan A. A., Kupriyanov M. S., Stepanenko V. V., Kholod I.I. Data analysis technologies. Data Mining, Visual Mining, Text Mining, OLAP, BHV-Petersburg, 2007. (in Russian)
12. Rudi Cilibrasi and Paul M. B. Vitanyi. Clustering by Compression. IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 51, No. 4, 2005, pp. 1523 - 1545.
13. R. Cilibrasi, P. M. B. Vitanyi, and R. de Wolf, "Algorithmic clustering of music based on string compression Comp. Music J., vol. 28, no. 4, pp. 49-67, 2004.
14. Ryabko, B., Reznikova, Z., Druzyaka, A., Panteleeva, S. , Using Ideas of Kolmogorov Complexity for Studying Biological Texts. Theory of Computing Systems, Volume 52, Issue 1 (2013), Page 133-147.
15. Ryabko B.Y., Astola J., Malyutov M. Compression-Based Methods of Prediction and Statistical Analysis of Time Series: Theory and Applications. TICSP v.56. Tampere: Tampere International Center for Signal Processing. 2010. 110 p. <http://ticsp.cs.tut.fi/images/8/89/Report-56.pdf>
16. M. Li, P. M. B. Vitanyi. An Introduction to Kolmogorov Complexity and Its Applications, 2nd ed. New York: Springer-Verlag, 1997.
17. B. Ryabko, J. Astola, A. Gammerman. Application of Kolmogorov complexity and universal codes to identity testing and nonparametric testing of serial independence for time series. Theoretical Computer Science, v.359, pp.440-448, 2006.

# Paypal E-Commerce and E-Payment - Problems and Solutions

Milos Ilic, Zaklina Spalevic, Petar Spalevic, Nebojsa Arsic, and Mladen Veinovic

Faculty of Technical Science Kosovska Mitrovica, University of Pristina,  
Kneza Milosa 7, 38220 Kosovska Mitrovica, Serbia

{milos.ilic,petar.spalevic,nebojsa.arsic}@pr.ac.rs

<http://ftnkm.rs/>

Singidunum University,

Danijelova 32, 11000 Belgrade, Serbia

{zspalevic,mveinovic}@singidunum.ac.rs

<http://singidunum.ac.rs/>

**Abstract.** The development of computers and telecommunications has created the conditions for the business globalization. E-commerce sites use electronic payment. E-commerce payment solutions make it easier than ever before for personal business to sell its goods and services on the Internet. Today the development of e-commerce and electronic payment allows Person-to-Person transactions. In such transactions there must always be an intermediary company that will provide a safe and secure handover of goods. In this paper authors describe electronic commerce and electronic payment, like key part of today's electronic business. Authors present advantages and disadvantages of electronic payment through the example of PayPal. Authors describe PayPal structure, security protocols that are in use in this system, and users interaction between which a transaction is performed. Here we present some examples of electronic payment hacking, and solutions for security improvement. Improvement is based on customer authentication, and implementation of patterns for user assessment.

**Keywords:** Electronic commerce, E-payment, PayPal, E-payment problems, E-payment solutions, Data mining patterns.

## 1 Introduction

Development of computer technologies, telecommunication, and rapid development of the Internet caused global market expansion. With technology enhanced flow of information, quick and efficient connection of business participants around the world became main goal in modern society. Different sources are linking emergence of electronic commerce to different time periods: some believe e-commerce emerged in seventies of the twentieth century, while others are of the opinion that it has a much longer history. To find answer on this question, electronic commerce must be defined first. The definition of electronic commerce by [1] is: E-business or e-commerce is exchange of standardized electronic messages in carrying out various tasks in companies, banks, administration, activities of citizens and in all other business transactions. Electronic commerce involves all business processes that use some kind of information and communication technologies, which are collectively marked as electronic technology.

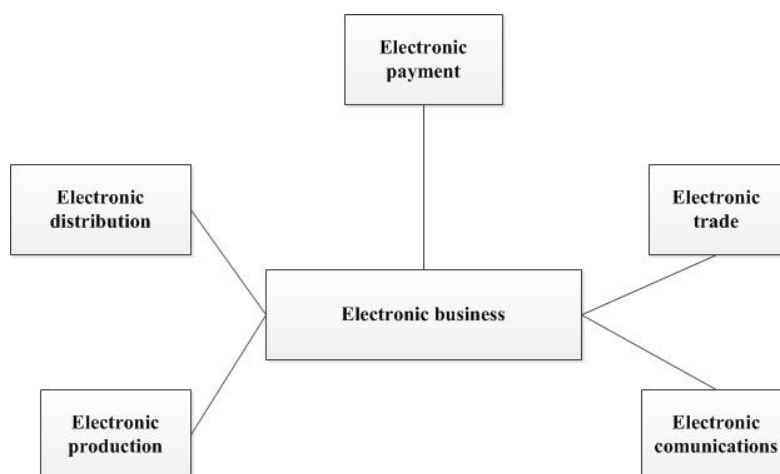
In literature we can find that first electronic money transfer was way back in 1860. That transfer was provided by WesternUnion, and for that job telegraph was used. This is the main proof for people who believe that e-commerce has a long history. However, the modern conception of electronic business implies the application of not only communicational, but also informational technologies, ie digital instead of analogue technologies. That becomes argument for those that reduce history of electronic commerce to the previous forty years. Authors of all reviewed papers agree that electronic commerce got the biggest expansion with wide spread use of Internet, in the last fifteen years. Electronic technology allows users to send large amounts of information over long distances in a short period of time. With it, companies can achieve significant savings

in operating cost, perform their tasks more efficiently, and be more competitive in the market. Electronic based commerce is a powerful tool for achievement of significant benefits, and for getting different possibilities, but only if it is used correctly.

Historically electronic commerce has developed in two directions. The first was supposed to enable companies to quickly and easily transfer money and information among themselves. It was intended only for business to business communication. The second, more recent, is focused on the end users of products or services (customers / clients) [2]. Electronic messaging, including business transactions and information flows, is widely used in triad containing public administration sector, companies and clients.

Electronic commerce has led to the reorganization of almost all business activities, as evidenced by the emergence of many specific forms of e-business, such as: electronic trade, electronic banking, electronic government, electronic marketing, electronic education, etc. Electronic commerce consists of several areas. All of areas represent modern business activities, and their mutual characteristic is that they use electronic technologies. Fig. 1 represents different areas of electronic commerce.

From Fig. 1 we can distinguish five electronic commerce types. Most important electronic commerce types will be further analyzed in continuation of the paper. First type of electronic commerce is electronic trade. Electronic trade represents the purchase and sale of goods or services via the Internet, as well as revenue from advertising, electronic exchange of documents that accompanying the goods, money and services, business through electronic tools. The term electronic trade can be defined as the process of managing financial transactions by individuals or companies online. This process includes both the retail and wholesale transactions. The focus of e-commerce is in the systems and procedures by which there is an exchange of various financial documents and information. Electronic commerce is implemented using one or more telecommunication technologies in order to provide contacts or direct trade between partners [3].



**Fig. 1.** Forms of electronic business

In a narrower sense under the e-commerce we mean buying and selling over the Internet. By this we mean not only money and products exchange, but also electronic production management, logistics organization and customers support. Electronic payments include money transfers and payments using the electronic trade. For electronic transfers of large amounts of money, private information network must be used. Banks and other financial institutions use this type of

networks for their transactions. Low money amount payments can be performed with electronic payment card. Electronic payment system provides electronic payment for all types of goods and services.

Authors in the paper describe types of electronic commerce, operating mode of system for electronic payment, and security mechanisms. Way of work of electronic payment system is described on the example of PayPal. Authors present how PayPal works, describe problems with security in the PayPal transactions, and recommend mechanisms for user protection. Paper is organized as follows. The second section presents the key facts behind electronic commerce and electronic payment. The third section describes payment models over the Internet, their advantages and disadvantages. Fourth sections presents PayPal payment system, describes how this system works, problems with client security and examples of solutions for most common problems. The fifth section represents conclusions and in this section the most important facts of payment systems are summarized.

## 2 Electronic Commerce and Electronic Payment

E-commerce can be defined as electronic performing of business transactions. E-commerce includes performing of business transactions via communication networks, especially the Internet. E-commerce covers all forms of business transactions carried out by companies or individuals. These transactions are based on processing and transfer of multimedia, including video, audio and text. E-commerce can be established between some company and costumers, or between multiple companies. It can also be established between multiple government institutions, and between government institutions and different companies and clients. Also, companies can establish e-commerce with own employees. This means that individual users can electronically order products or services from online sellers. Sellers use information and communication technologies to connect with their suppliers and distributors. Two most common electronic payment models are business to business model and business to customer model. Today's development of electronic commerce and electronic payment provides customer to customer model of transaction too. In such transactions individuals who are end-users can participate in mutual trade [4], but in such transactions there must always be an intermediary company, that will enables safe and secure handover of products.

Like we said above, the development of e-commerce is influenced, among other things, with the rapid growth of the Internet, the emergence of new information and communication technologies, low cost of their implementation, the ability to connect with hundreds of millions of people, the interactive nature of such communication and so on. Today, e-commerce has many benefits over traditional commercial transactions. Some benefits of electronic commerce are:

- e-commerce allows acquaintanceship between seller and a large number of customers from all over the world, with very low capital outlays and operational costs;
- one-to-one transactions and interactivity provides facilitated services and connections with customers;
- e-commerce can reduce the time between payment and reception of goods and services,
- e-commerce enables new business models that increase the competitiveness and profitability,
- e-commerce often provides cheaper products and services for customers, allows them to buy from many places, and to compare the prices online;
- e-commerce provides more choice for customers. They can choose between many products, due to the large number of different sellers;
- e-commerce enables customers to communicate with other buyers and sellers in the electronic community, and to exchange ideas and share experiences;

- e-commerce allows that certain products could be sold at lower prices, so that poorer people could buy them.

Electronic commerce development is strongly binded to the development of Internet technologies, because of their strong impact on it. In recent years electronic commerce architecture become very stable. Part of e-commerce related to multimedia and technology contents is relatively new, and continues to develop. Multimedia and technology content means that customers chooses payment method, and depends of that result of transaction take place.

As in traditional trade for purchased goods it is necessary to pay. In e-commerce the key difference between electronic and traditional trade is that in electronic trade customers do not use paper money. Technology behind electronic trade provides mechanisms for payment without use of paper money. Here is necessary to mention the concept of electronic money used in electronic commerce. Authorized services provide electronic money transfer between end-users. Electronic money can be defined as information of the monetary value which can be transferred through computer networks, or outside the usual channels of payment which are traditionally supported by banks [5]. The concept of electronic money and digital money is very often identified, although there is a difference between them. The difference between those two concepts best we can observe in the case of a telegraph money transfer in which the signal is electronic, but transaction itself is analog. Also, the phone payment order that is forwarded to the bank by phone is not digital transactions, but basically this transaction is an electronic money transfer. Thus, the digital money is associated with the electronic transfer of funds, which is based on informational technologies. Electronic money has been used before the appearance of the Internet. The best example for this fact are POS terminals that enables the realization of cashless payment transactions at point of sale or in a store. Before the appearance of the Internet, users generally were not aware of electronic commerce and electronic money existence. The system itself was much easier to control by banks and other financial institutions, that were involved in a system of electronic money. With appearance of the Internet, there is a massive use of electronic money, which causes a numbers of issues related to its implementation.

### **3 Payment on the Internet**

The massive use of computers, and information technology development, combined with the appearance of electronic money have caused that traditional payment systems have become inadequate in modern conditions. These modern conditions are characterized by accelerating the pace of life, and the need to carry out banking transactions faster, easier, and ultimately cheaper. The emergence of the Internet has opened new questions about electronic money and payment systems. The development of electronic commerce on the Internet prompted the need to adapt the traditional payment system, and the formation of new business concepts that will enable the realize payment transactions in the virtual environment. As a result of these initiatives there has been the emergence of different payment systems on the Internet, which depended of that who initiates the transaction, and how the transaction is processed. Based on these parameters models can be divided into: cash-oriented payment model, model of payment by checks, model of payment by payment card, and model of payment warrant [6].

Cash oriented payment model on the Internet has the same processing cycle like traditional money cycle. Transaction starts when client send request to the bank for the digital banknote. Bank issues digital banknote, and with that banknote client pays for goods or services bought over the Internet. Trader forwards obtained banknote to his bank, where banknote is annulled after necessary checks.



Check oriented payment transaction model starts when a user on his computer forms the check which will be used for payment to the merchant. The user digitally signs the check, adds the banks certificate, and packs it into a digital envelope. The payee receives an electronic check and adds payment instructions so bank can transfer money to his account at the end of transaction. Payee sends packed envelope to the bank. After receiving it, bank must confirm validity of check in order to transfer requested amount of money to payee's account. Payment check verification is performed by checking digital signatures and certificates forwarded by payer.

Model of payment with payment cards begins when the payer on the website of the merchant chooses certain products or services. After that he will be redirected on Web page for payment. On that page payer enters his payment information (type of payment card, number of payment card, and expiration date of payment card). The transaction is processed using standard banking infrastructure through the payment network of some system (Visa, MasterCard), between payer bank and the payee bank. At the end of transactions by payment card, bank which is included in the model inform participants (buyer and seller) of the transaction outcome.

Model of payment warrant is most frequently applied in electronic banking (e-banking). Here we can distinguish two versions of such system. The first older version of these systems was required from the users to have installed appropriate software on computer, and a smart card which containing certificates and digital signatures, in order to freely use e-banking system. New version of the systems require from the user to log on secure website for e-banking with username and password. After logging, the user initiates a transaction by which orders the bank to transfer from his account a certain amount of funds in favor of the payee. By order of the customer, the bank use standard channels for transaction, which means that in the process of clearing closes the mutual positions with the payee bank.

In recent years there has been the emergence of a new form of payment system on the Internet. That is the P2P (Person-to-Person) payment system. These payment systems combine the functions from different payment models and make maximum use of their advantages, which resulting in a large acceptance of the system by the Internet population. The term P2P is also used to indicate the "customer the customer" type of trading. In this type of trade end-user have direct contact with other end-user. This means that individual trade directly with other end-users. The companies that supports these transactions must find some not traditional way to pay their services.

This cost is usually a small percentage of the transaction, membership fees, advertising, or other business combination. Various personal services are offered via the Internet, from teaching to astrology. An increasing number of individuals use the electronic interchange to exchange goods and services via the Internet. It is very important for the customers to be careful while purchasing, because it may be a fraud or computer crime.

The most important advantages of these electronic commerce models are:

- Expanding the potential market,
- Eliminating intermediaries,
- The ability to easily update content,
- Consumers can use it at any time, from any place.

The biggest disadvantages of these electronic commerce models are:

- The absence of quality control,
- There is no guarantee that payment will be made,
- Difficult to pay by checks.

The examples of P2P model use are students who use electronic commerce system to reduce the cost of study. Students one to another sell textbooks and other teaching materials using system based on P2P model. Many web sites offer a service of this kind of trade. Representative example of this payment category is PayPal system. P2P payment systems vary in their complexity, but what they have in common, is to take advantage of standard banking infrastructure, to provide a reliable and inexpensive way for the realization of payment transactions to users on the Internet.

#### 4 PayPal Internet payment system

PayPal payment system was founded in 1999 under the name Confinity. The main purpose is to enable individuals or companies who own e-mail address to securely, easily and quickly send and receive money online. PayPal payment system can be used for payments at online auctions, to purchase goods and services, to send and receive donations, and can even be used to send cash in the form of P2P transfers. The key moment in the PayPal development is when the company was purchased by the giants in the field of online auction e-Bay. E-Bay bought PayPal in 2002 year for a whopping 1.5 billion dollars. This business move has brought a benefit to both companies. From this symbiosis e-Bay got reliable mechanism for processing payment transactions between participants in online auctions, and in other side PayPal got users that originate from e-Bay. Over the years, PayPal has secured a dominant position and suppress other competitors, although they had the support of other Internet giants. With his appearance PayPal took the role of mediator between the various financial institutions involved in the realization of payment transactions. To be able to send the payment through PayPal, the payer first must open an account on this system. To make deposited funds into account on PayPal system, the user has several options: credit card, debit card, bank account, and MoneyPak which represent a type of pre-paid card. The money which user paid into PayPal account, PayPal transfers in the bank accounts that PayPal has in some of the banks. When paying for goods or services on the Internet, payer selects the PayPal payment on the seller web site, or enters the email address of the payee and the funds will be deposited in the recipient's account within the system [7]. On this occasion there is no real movement of money, because the money is all the time on the PayPal account in the bank, simply an administrative funds are transferred from the account of the payer to the account of the payee. In fact, PayPal's business model is based on the basic model of P2P payment system like we said above.

In case if the payer does not have an account on PayPal, the money will be forwarded to the recipient's account by PayPal. PayPal will take the role of mediator between the various financial institutions that process that transaction. That is the greatest strength of PayPal. PayPal uses the existing banks infrastructure and other financial institutions infrastructure to offer to users a simpler and cheaper option for processing payment transactions. This become possible thanks to the effects of economies of scale. This means that PayPal pays far lower fees to financial institutions when processing transactions, than it would have paid users. This is because PayPal provide large volume and the amount of transferred funds. For different categories of users, whether individual or business users, depending on the specific needs PayPal offers the possibility of opening three types of accounts.

Personal account is particularly adapted for online shopping. This account provides possibility for the user to receive and send payments through PayPal system. A limitation for this type of account is that users do not have the option to receive payments by credit and debit cards. For this type of account monthly transaction volume is limited to five hundred dollars. The owners

of accounts that break the limit, have at their disposal the ability to raise its own account at the level of premium or a business account, or simply delete transaction which breaks the limit.

Premium and business accounts are similar. The main difference is that the business accounts must be registered on the company, and premier accounts can be registered on the company, but also on the individual person. Premium accounts have all benefits of personal accounts, but also a couple of special benefits of which is certainly the most important that this account enables payment by credit and debit cards as well as the possibility of processing a large volume of transactions.

Business account is for business users. This account gives them the option of receiving payment on different grounds and different payments methods, and what is most interesting is that provides the ability to receive payments from the users who do not have PayPal account. Special subcategories of PayPal account is so-called student account which can be opened within a personal, premier or business account, and this account is intended for young people. This sub-account enables to young people to take care of their finances and to feel the magic of online payments.

The essence behind PayPal account functionality is that system requires from users to enter data about the account number or a credit card only once, when users register account. The account could be registered without this information, but that account will not be confirmed, and it will not be possible to make payments with that account. To confirm account registration user must enter data from credit card or data about bank account [9].

After that PayPal verify ownership of the entered account or credit card. System charge a small sum of money from users bank account or credit card (approximately about 1.5 US dollars). After that system requires from users to copy four digits about this transaction from bank report, which arrive as some kind of code of collection of small sums charged by PayPal. Once the user copies those digits to the PayPal website, the system returns the charged sum, and account is confirmed. In this way, opening a PayPal account is provided for free. The owner of the account has the ability to transfer payments to anyone who has an active email address, regardless of whether they have a PayPal account or not. Payee will be than informed by e-mail of received payments. After that he must open his own PayPal account. Funds transferred through PayPal will be deposited to the account of the payee, where they will remain until account owner decides to spend them for payment on different grounds or decided to raise funds from the account. PayPal offers to users different ways to raise the money from the PayPal account. If the user was joined bank account to his PayPal account, which passed the verification process, there is the possibility that the funds can be transferred directly to bank account. Fig. 2 represent PayPal structure.

During the realization transactions, the buyer at the website of the trader chooses PayPal as an option for payment. After that web browser will be redirected to the webpage for payment where customer must entries payment instructions. In essence, this kind of transaction, as we have already mentioned, comes down to an internal posting of funds from one to another PayPal account. If the trader website as the only way of payment accepted credit card, the user has the ability to, through PayPal, carry out that transaction too.

This is provided with option PayPal Debit Bar, which enables the user to obtain a virtual MasterCard number. Online PayPal communication structure is presented on Fig. 3. Customer on the website of the trader select MasterCard as a payment option and enter the virtual card number that he received from PayPal. Funds transfer to trade account is performed from the PayPal accounts. Before the appearance of PayPal, trades who want to accept payment by credit cards from their customers had to have opened trade account in the bank. This account is specially designed with the purpose to accept payment by credit and debit cards through the

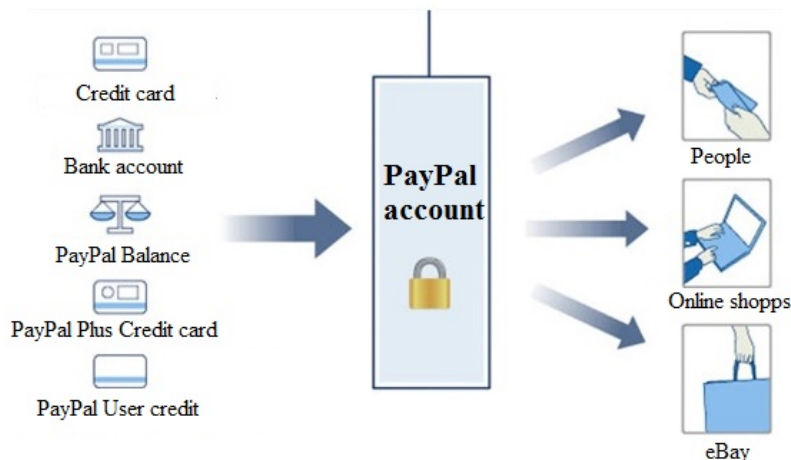


Fig. 2. Structure of the PayPal organization

network which is in use for processing payments with credit cards. This payment method was associated with significant costs in terms of fees (up to five percents of transaction value). PayPal for premium account owners and business accounts owners provides an opportunity to be able to receive payments by credit cards with easy integration of PayPal trade accounts with their websites. This allows significantly lower transaction processing cost (from 1.9 to 2.9 percents of the transaction value), and greater safety of payment cards [8]. PayPal has create a range of safety procedures and protection. For example, although costumer bank card is directly linked to a PayPal account, costumer can ensure against theft by transferring only a small amount of money on card or hold it empty, unless costumer have a need for a purchase. So, when costumer needs money, he can transfer exactly as much as needed, and pay for goods on the Internet from PayPal account. If some transaction is problematic, that transaction can be stopped within seventy-two hours. Product that is bayed and that does not meet specifications can be returned within forty-five days.

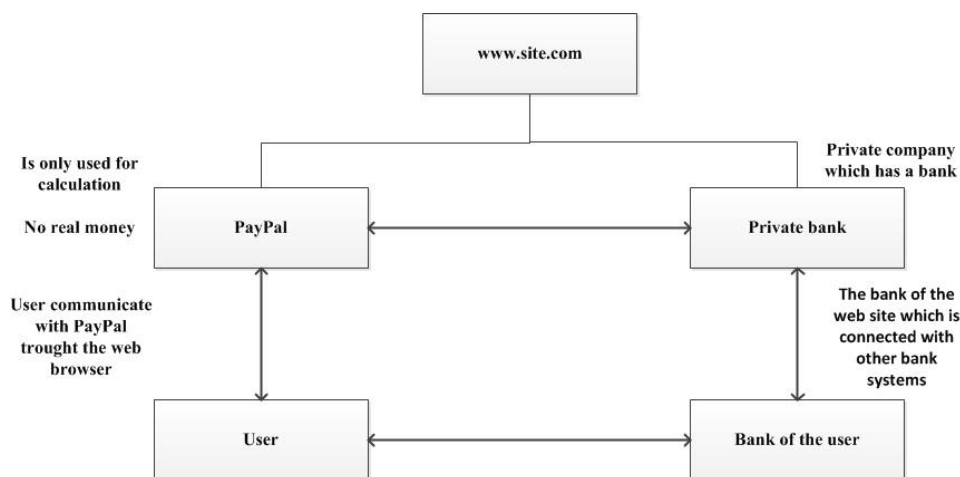


Fig. 3. Communication structure of PayPal

In addition, PayPal have number of teams that preventive explore the Internet trying to detect cheaters before something bad happens. Preventively however despite all security measures, in PayPal system sometimes problems occurs, especially to incautious users.

#### 4.1 Attacks on PayPal system

Like with any other online purchase, user must pay attention to web site through which he purchase, to keep account information confidential and not believe to phishing emails and websites. Phishing is a type of Internet fraud where the goal is to take confidential data that user uses for authentication on some website or some payment system. This includes stealing of passwords, credit card numbers, bank account information and other confidential data [10]. Phishing messages usually take the form of false bank notifications, providers, electronic payment systems, social networks, Internet games and so on. Such notification will attempt to encourage users, from one or another reason, to immediately enter or update their personal data. Common excuses for such requirements are loss of data, crashes of the system, etc. The so-called bank phishing is the most common phishing tactic, which aims to gain access to a user's Internet bank account or e-payment. Then when the person who is behind phishing take the customer user name and password, that person will have access to the customer account. Phishers are very skilled in the development of e-mails that have authentic look, and that are truly copies of official e-mails that came from different organizations.

They use the official logos of organizations and copied the whole style of legitimate correspondence. In e-mail they will suggest to the user to click on the link and to input his personal data. Typically to the users is argued that the reasons for such request are measures that a certain company reportedly take to improve the security of the web site, and because of that the user needs to re-log in to his account. When a user clicks on a link, he will be redirected on a fake website that looks like a legitimate web site in which he must enter his user name and password. The data that the user enters then will be sent to cyber-criminals. Very often, these fake websites contain exploits, which install spyware programs on the computer of the victim. This means that even if user have not entered his user name and password, but only clicked on the link in e-mail message, there is still a possibility that the user was inadvertently downloaded a malicious program on his computer. That program might later steal a range of personal data from user computer. If the e-mail was send from PayPal that e-mail will contains link like this <http://anything.paypal.com/anything>. If in other hand e-mail was not send from PayPal the link within the e-mail will look differently. More precise links in such e-mails instead of [paypal.com](http://paypal.com) contains something different before the slash. For example link may have the following look <http://paypal.confirmation.com/anything>, <http://anything.pay-pal.com/anything>.

Famous are scams of PayPal.com users when criminals take advantage of the similarity of a small letter "l" at the end of the domain names, and replace that letter with a capital "I" (PayPal.com), or as in the latest case, instead on PayPal.com users were redirected to PayPal@az.ru. PayPal as a service for online paying, and transactions over the Internet is constantly the target of attacks by individuals and groups. As each service, PayPal has some omissions in the security measures. Egyptian researcher Yasser H. published some of most critical vulnerabilities of PayPal service. Yasser was exploring the gaps in the PayPal service managed to completely bypass the CSRF (Cross-Site Request Forgery) prevention system that is applied by PayPal.

Cross-site request forgery is a type of attack that occurs when a malicious website, email, blog, instant message, or program causes that a user's web browser perform an unwanted action on a trusted web site for which the user is currently authenticated. The impact of a successful cross-

site request forgery attack is limited to the capabilities exposed by the vulnerable application. For example, this attack could result in a transfer of funds, changing a password, or purchasing an item in the user's context [11]. In effect, CSRF attacks are used by an attacker to make a target system perform a function (funds transfer, form submission etc.) via the target's browser without knowledge of the target user, at least until the unauthorized function has been committed.

Impacts of successful CSRF exploits vary greatly based on the role of the victim. When targeting a normal user, a successful CSRF attack can compromise end-user data and their associated functions. If the targeted end user is an administrator account, a CSRF attack can compromise the entire web application. The sites that are more likely to be attacked are community websites (social networking, email) or sites that have high dollar value accounts associated with them (banks, stock brokerages, bill pay services). This attack can happen even if the user is logged into a web site using strong encryption (HTTPS). Utilizing social engineering, an attacker will embed malicious HTML or JavaScript code into an email or website to request a specific 'task url'. The task then executes with or without the user's knowledge, either directly or by utilizing a Cross-site Scripting flaw. Yasser discovered three critical vulnerabilities in PayPal services, which are described in continuation of the paper.

Reusable CSRF Token: The CSRF token "that authenticate every single request made by the user" which can be also found in the request body of every request with the parameter name "Auth" get changed with every request made by user for security measures, but after a deep investigation was found out that the CSRF Auth is reusable for that specific user email address or user name, this means if an attacker found any of these CSRF tokens, he can then make actions in the behalf of any logged in user [12]. It seems interesting but still not exploitable, as there is no way for an attacker to get the "Auth" value from a victim session.

Bypassing the CSRF Auth System: The CSRF Auth verifies every single request of that user, so what if an attacker "not logged in" tries to make a "send money" request then PayPal will ask the attacker to provide his email and password. The attacker will provide the "Victim Email" and any password. Then he will capture the request, the request will contain a valid CSRF Auth token which is reusable and can authorize this specific user request [12]. Upon further investigation, authors were found out that an attacker can obtain the CSRF Auth which can be valid for all users, by intercepting the post request from a page that provide an Auth token, before the logging-in process. At this point the attacker can CSRF "almost" any request on behalf of this user.

Through examination of the password change process, group of authors have found that an attacker can't change the victim password without answering the security questions set by user. Also the user himself can't change the security questions without entering the password.

Bypassing the Security Questions Change: After further investigation, Yasser have noticed that the request of setting up the security questions which is initiated by the user while signing up is not password-protected, and it can be reused to reset the security questions up without providing the password. Hence, armed with the CSRF Auth, an attacker can CSRF this process too, and change the victim's security questions.

At this point, an attacker can conduct a targeted CSRF attack against PayPal users and take a full control over their accounts. The attacker also can add, remove or confirm the email address, add a premium user under business account, change security questions, change address, method of payment, user settings, and more. These and similar deficiencies testify that there is no perfect service in electronic trade, which can't be misused.

## 4.2 Solutions for better security

Some of the protection measures, or some of the recommendations for the end-user in system like this are given below. User could try to use a newer version of web browser that enables secure transactions, and which encodes the information sent over the Internet [13]. When user sends data he should consider that the proper security authenticity protocols of the transaction participants are used. For the end user this means that a small yellow padlock appears in the status bar, and that the address of the web site starts with https. Online payment services have their own privacy policy, and users must sign that they agree with the terms in the policy. This privacy policy document contains information about how web site owner use all users data, and how customer personal data are protected.. In many cases that document is very boring for reading, and many users do not read any of that, just sign. Each user must save his own personal data. Personal data should not be left to anybody. It is necessary that user first assess whether company or organization are reliable, and think which data should be left and which do not. It is good that user regularly checks bank statement about transaction that are completed, and that he checks how much money he has on the credit card, or current money amount on other accounts. If user noticing suspicious transactions in bank statement, he should immediately contact his bank, and report that. In the case when the money is stolen from the user bank account, money can be refunded if the user proves that behind the suspicious transaction stands anyone else. Good practice is that user should keep all e-mail messages about the transaction that were received or sent.

To improve security for their clients companies that provides online payment systems must think about each detail. One of the common security risks are credit cards. Security risks of credit cards have threatened whole world. The credit card frauds that are very common arise from lost or stolen cards or card numbers, leading the thief to use the card number for criminal activities over the telephone or the Internet [14]. To improve process of credit card risk analysis and customer profiling, data mining methods could be used. One of data mining methods is classification. This is the analytical approach that uses training examples, which are pairs of input and output targets, in order to find a suitable target function also called as a classification model generally. This classification model thus produced is used for descriptive and predictive modelling both. The target function uses techniques that are divided broadly into methods like rule-based, probabilistic, geometric and prototype-based. Classification methods play an important role in risk measurement especially in credit rating problem. In finance, classification approaches are also used in customer profiling by constructing predictive model. The values that were predicted are categorical. Classification predicting and forecasting techniques that are widely used for classification are decision tree, neural networks and Genetic algorithm . Decision tree is the simple, easy to understand and interpret among various classification tools. Decision tree can only solve classification problems but cannot solve regression problem. A neural network is most accurate in bank failure prediction. In comparison with other techniques, neural network models are more accurate, adaptive and robust. Genetic programming is one of the most recent techniques that has been applied in the field of credit card risk assessment.

## 5 Conclusion

Everyday obligations and way of life of today's human population are conditioned by the rapid development of technology. This rapid development causes that people must to resort to use technology to save their time. The development of electronic banking, Internet payments and Internet purchases reduces the time spent on waiting in the line, and filling out extensive

paperwork. System like these provides possibility for many people to perform daily obligations from home. Online shopping and payment systems via the Internet are more and more present in both legal and individual persons. In order to users of such systems feel more comfortable and more secure while using the payment systems, these systems need to constantly develop and improve. Like we said above one of such systems, which is in use for payment and money transactions between private (between individuals), legal (a company that sells services or goods) or both is PayPal. New user verification is very well organized in the process of PayPal account creation. During further transactions safety of the user is conditioned by various factors. As far as the security of transactions it has supported by protocols that are in use in all official payment systems. On the other hand attacks by malicious users on this system are more common. End users must take care about each and every purchase. Also users must take a care about the flow of information which are handled in the process of purchasing and payment for the product. Another problem, which is increasingly common in all purchase and payment systems over the Internet and even in PayPal is the distribution of purchased products. Not so rarely happens that the end user does not receive the purchased product or receive the product which is not in accordance with what was ordered. In such cases, PayPal offers a variety of mechanisms of compensation or money refund, but the fact that the entire procedure takes a certain time. Documented hacker attacks testify that a certain level of security needs to be further improved, and in this field some efforts must be made.

## References

1. Bjelic, P.: Electronic trade. Institute for International Politics and Economy, Belgrade (2000).
2. Koncar, J.: Electronic trade. Faculty of Economics, Subotica (2003).
3. Stojanovic, I.: Electronic trade and shopping through the Internet in Serbia. Master thesis, Singidunum University, Belgrade (2011).
4. Vaskovic, V.: Payment systems in e-business. Faculty of Organizational Sciences. Belgrade (2007).
5. Simovic, V.: Electronic business. College of Information Technology, Belgrade (2013).
6. Zon-Yau, L., Hsiao-Cheng, Y., Pei-Jen K.: An Analysis and Comparison of Different Types of Electronic Payment Systems. International Conference - Management of Engineering and Technology, PICMET '01, Volume 2, pp. 38 - 45. Portland (2001).
7. PayPal Integral Evolution Integration Guide, PayPal, Inc., 22-24 Boulevard Royal, L-2449, Luxembourg (2014).
8. Stanojevic, V.: Electronic banking. Graduate thesis, Higher School of Electrical Engineering. Belgrade (2005).
9. Vuksanovic, E., Tomic, N.: Alternative Payment Mechanisms in Electronic Commerce. The First International Scientific conference Synthesis, DOI: 10.15308/SInteZa-2014-153-15. Singidunum University, Belgrade (2014).
10. Jackson, E.: The PayPal Wars: Battles with eBay, the Media, the Mafia, and the Rest of Planet Earth, <https://books.google.rs/books?id=Jtr3ThHeTEMC&pg=PA2&lpg=PA2&dq=The+PayPal+Wars:+Battles+with+eBay,+the+Media,+the+Mafia,+and+the+Rest+of+Planet+Earth&source=bl&ots=sndbAcbu-w&sig=qumZEKM7wdGT5GxkMSr3lvdNqXI&hl=sr&sa=X&ei=-NOLVbXBD4G6UNCrgIgd&ved=0CFMQ6AEwBg#v=onepage&q&f=false>
11. Cross-Site Request Forgery (CSRF) Prevention Cheat Sheet, [https://www.owasp.org/index.php/Cross-Site\\_Request\\_Forgery\\_\(CSRF\)\\_Prevention\\_Cheat\\_Sheet](https://www.owasp.org/index.php/Cross-Site_Request_Forgery_(CSRF)_Prevention_Cheat_Sheet)
12. Yasser, A.: Hacking PayPal Accounts with one click (Patched), <http://yasserali.com/hacking-paypal-accounts-with-one-click/>
13. Shannon, S., Dave, N., Dave, B.: PayPal Hacks: 100 Industrial-Strength Tips and Tools, [https://books.google.rs/books?id=UJU0cTMj4bsC&pg=PR16&lpg=PR16&dq=paypal+under+the+hood&source=bl&ots=DLA9-muXA8&sig=Rn\\_iJfut8rJ65ew31EY7Q02-e\\_s&hl=sr&sa=X&ei=jtoLVbfyFISoPMX9gNgM&sqi=2&ved=0CE0Q6AEwBw#v=onepage&q&f=false](https://books.google.rs/books?id=UJU0cTMj4bsC&pg=PR16&lpg=PR16&dq=paypal+under+the+hood&source=bl&ots=DLA9-muXA8&sig=Rn_iJfut8rJ65ew31EY7Q02-e_s&hl=sr&sa=X&ei=jtoLVbfyFISoPMX9gNgM&sqi=2&ved=0CE0Q6AEwBw#v=onepage&q&f=false)
14. Srivastava, S., Garg, A.: Data mining for credit card risk analysis: a review. International Journal of Computer Science Engineering and Information Technology Research. Vol. 3, Issue 2, pp. 193-200 (2013)



# One Implementation of the Embedded Database Protection

Siniša Ilić<sup>1</sup>, Slobodan Obradović<sup>2</sup>, Nebojša Arsić<sup>1</sup>, and Vera Petrović<sup>2</sup>

<sup>1</sup> Faculty of Technical Sciences in Kosovska Mitrovica (FTNKM) of  
University of Priština, Kneza Miloša 7, K. Mitrovica, Serbia  
{sinisa.ilic, nebojsa.arsic}@pr.ac.rs  
<http://www.ftn.pr.ac.rs>

<sup>2</sup> The School of Electrical and Computer Engineering,  
Vojvode Stepe 283, Belgrade, Serbia  
slobo.obradovic@gmail.com, vera.petrovic@viser.edu.rs  
<http://www.viser.edu.rs>

**Abstract.** In this paper one implementation of the database configuration and development considering security issues especially when connected to Internet is presented. Sometimes the precautions on security vulnerabilities implemented on other levels of database environment (such as: network, operating system, client application etc) are not enough in order to protect database itself. The attacker can frequently use the public user screen of an application and try to access Database as anonymous user.

We have created the database with one way of self-protection where regular users can't access any of tables, only stored procedures. Permissions on stored procedures are not configured using standard DBMS tools but using embedded security developed by authors that checks the user rights and permissions on different business functionalities and expected parameters' values. The calls to stored procedures with suspicious users and parameters' values are logged for further analysis.

**Keywords:** embedded database security, stored procedures, user permissions.

## 1 Introduction

Very often developers are focused only on delivering functionally reliable software systems of higher quality according to the functional requirements, managers are focused for delivery to be in time frame and budget, testers are usually testing functional, end to end, administration and performance testing.

The security environment is usually predefined and after implementation rarely tested. Software developers are often experienced with security issues when fixing vulnerabilities discovered by clients. Often it is very hard even to foresee them. As the programming of procedures for handling of most frequent security risks is very time consuming compared to the time needed for programming the functional requirements, it becomes clear why lack of security occurs.

The testing of functional requirements is easy - scenario of testing should follow procedures defined by client. Users on the client side are aware on business procedures and they can think on bunch of different alternative scenarios on each use case and test the functionality of the system. However, nobody can predict how many different attacks on security of the system can occur, from outside or from inside, from anonymous or regular users and what kind of threat should be expected. The task of security testing is very hard considering that security threats are rapidly growing and it is difficult to simulate all possible conditions. All parts of infrastructure should be tested: OS, web browsers, web servers, application servers, database servers and communication equipment. Luckily, there are free and commercial tools that can indicate security flaws in these components.

In systems with database, the validation of input data must be performed in both: client application (by using validation scripts or functions) and database (by using constraints, triggers

or stored procedures). By checking data on both levels, the chance that data with wrong values will be inserted in database or some confidential data will be exposed is minimal.

## 2 Related Work

According to [1], security is a process, not a product, service or procedure, but a set that comprises them - with many more elements and measures being implemented continuously. An organisation or institution cannot be considered "safe" at any moment even after the last test performed according to its own security policies. Security is a process of maintaining an acceptable level of risk. So, security is a process, not a final state. That is why the task of managers in charge of security at all levels is very important: they have to analyse potential risks, their effects, impacts - consequences if a risk occurs, probability of risk appearance, handling methods - measures to be taken to ensure that risk will not happen, and who is responsible to take appropriate measures [2].

In the systems with databases, some proposals are created how to react on suspicious querying. As "malicious activities" need not to come from an attacker, there must be ensured enough time for security engineer(s) to investigate if activity is really dangerous or not. During that time, a potential attacker must be convinced that his/her transaction is successful. The method for solving this problem is to build as many "copies" of database as many suspicious users are connected to database [3]. If a "suspicious" user is proven to be an attacker, it is blocked, and if it is not, there is built Merging Algorithm that replaces its trustworthy version (values) with its suspicious version (values), and then removes the suspicious version.

Sometimes a regular user connected to DBMS should have an access only to the data concerning to his/her job description. But, because of the authorization mechanisms in SQL permit access control at the level of complete tables or columns, or on views, users (or applications) are granted database privileges that exceed the requirements of their job function, these privileges may be used to gain access to confidential information. For example: In an academic scenario, institution's database stores information about students, it may be desired to allow students to see only those tuples which store their own marks and/or fees details. On the other hand, a professor should be able to access all grades for a course he/she has taught. That is the reason to introduce permission, policy and condition tables where permissions and conditions are fine grained for users and user groups. Thus, an original query to be sent by application is modified by application [4] according to the conditions defined in mentioned tables for selected user by adding them in WHERE clause of SELECT queries [5].

SQL Injection Attack occurs when an attacker changes the intended effect of an SQL query by inserting new SQL keywords or operators into the query. The extensive review of the different types and injection mechanisms of SQL injection attacks known almost to date, with examples, the detection and prevention techniques against SQL injection attacks are shown in [6]. In order to protect database from SQL Injections it is crucial to validate user inputs on web forms. If malicious SQL query is allowed to be executed on DBMS, the resulting security violations can include identity theft, loss of confidential information, and fraud. Each data item that user sends from client machine to server should be checked and filtered before it comes to database. This process is called sanitising. The input data can be parsed into segments according to the key words of SQL syntax and compared with expected structure [7]. If expected structure is different than actual one (parsed from input field), the input data is not transferred to database.

Another approach is to filter input data by using regular expression search tool [8]. The regular expression search tool enables finding of templates within the text string (by: key words, text in brackets, tags, templates of numbers etc.). The key words of SQL statements, tautologies, URL

encoding and other encoding characters can be put at regular expression search strings. These malicious characters/words can be removed from input fields and not transferred to database. The similar approach can be achieved with Java function HashMap that can be used to convert encoded characters (that attacker can use instead of normal characters in order to skip some standard validation techniques) to standard ones in minimal time [9].

The better security may be applied to database itself. Data in database might be encrypted and even in the case of penetration, an attacker could not use such data [10]. Another solution is using embedding policies into the database itself and enabling these policies to block every attempt to compromise the state of the database, or to alter its configuration in a way that contradicts what has been established and fed into the policy by the system owner [11]. User can access database only by using stored procedures, even for changing parameters of database by database administrators (not the database owner). When a power user or a hacker initiates an attempt to change security configurations (database parameters), the request goes through a process of verification before it can be processed. This step is carried out by database stored procedures that have built-in logic for checking the request against the policies. If the request does not comply with the set policies the request is rejected and the system owner is alerted, the user notified, and an audit trail is recorded.

Similar approach is presented in [12] where user can access database only through stored procedures. The identity of particular user is checked within the procedure although the user belongs to a user role which permission is granted to this stored procedure. Also the parameters of stored procedures are validated within the stored procedures against the width, black list words, black list encoded characters, etc. Calls to stored procedures with suspicious parameters' values are logged to special tables and alerts to DB owner are raised. By using modified MAC (Mandatory Access Control), RBAC (Role-Based Access Control) and DAC (Discretionary Access Control) models it is possible to design a database security system that can individually control user access to data groups of various sizes and is suitable for the situation where the user's access privilege to arbitrary data is changed frequently [13]. In these models user can access any data that has lower or equal security levels, and that is accessible by the roles to which the user is assigned.

### **3 Working Environment**

Every user has a database account. For users who access the database by client application the IP of client machine is assigned as well as the list of eventual IP addresses the user may login. For users of web application the IP address of remote users is sent as parameter of stored procedure. Web application uses SSL that passed appropriate security tests. The assumption of our solution is that an attacker cannot obtain DB owner account credentials, because DB owner cannot log over web application, only from client application (in LAN environment) with already set IP address. An attacker can use web application with compromised credentials to obtain data or to change/insert own data or try to intrude to database from internal network by using client application or some own tools.

### **4 The Implementation of Embedded Database Protection**

The modified solution of the [12] after some deficiencies are noticed is implemented. The following specifications are valid:

1. The standard DBMS users are defined and assigned to user roles that belong to business functionalities; There exists User table with user IDs (no user names and passwords) and binary encrypted permissions on high details of business functionalities;
2. User roles (to which users are assigned) have permissions only to execute stored procedures, only DB owner has permission on tables;
3. Stored procedures have at least one input parameter and its body consists of only static SQL code (there is no `execute_sql(string)` command)
4. The user rights for every stored procedure are controlled within the stored procedure using the check permission function;
5. The parameters' values are validated in stored procedures using the validation function;
6. The values of: user, IP addresses, parameter's values and the number of affected records are inserted in a log table after every execution of stored procedures;
7. For some violations with higher risk, the DB owner is alerted immediately and user suspended temporarily, and for lower risk violations, there are counters that initiate temporary suspension of user when number of violations exceeds limit in defined time-frame;
8. DB owner can change user permissions for running stored procedures through special (admin) stored procedure that can be run only from computer with local IP address; Also, DB owner can set permissions on DB access of users to connect from different IP addresses, to run web or client application and to set ranges of values for each stored procedure's parameter.

User roles in database are defined for business functionalities like: sale, procurement, accounting, hr, etc. and are defined through standard DBMS user roles management. There is no guest account even for the web application. The detailed permissions for particular user within user roles are defined in User table in binary field. A user can be for example the member of "sale" role that consists of the following detailed functionalities like: 1)customers, 2)discounts on products, 3)sale orders and 4)shipping. Then the permission of a user on mentioned functionalities is defined in four bytes (a byte per functionality). The binary value of each byte that defines permissions of functionality customers, lays on sub-functionalities like: add new customer, modify basic details on customers, modify extended customer details, read basic data about customers and read extended data about customers. We can assigned code for each of mentioned sub-functionalities - {0, 1, 2, 3, 4} respectively. Then the byte value can be calculated as  $\sum_i b_i \times 2^i$ , where  $b_i = 1$  if user has permission for particular sub-functionality of weight  $i$  or  $b_i = 0$  if user has not. For example: the byte value for user who can only read and modify basic data is  $10 = 2^1$  (modify basic details)  $+ 2^3$  (read basic data). Some other functionalities may have more (up to 8) sub-functionalities. The functionalities and sub-functionalities are defined in separate tables altogether with their byte positions in User table and codes used. Also, the table with relations between user procedures and sub-functionalities implemented in user procedures is defined.

All parameters of stored procedures have default values, and usually, default values are dummy values. The reason is not to allow attacker to figure out what is the number of parameters. If an attacker in normal case calls stored procedure with some mandatory parameters missing (Figure 1), DBMS raises exception informing him/her that next parameter or type is missing. By adding new parameters in next stored procedure calls until DBMS executes stored procedure without error raised, an attacker will probably figure out the number of parameters.

Also, if an attacker put the value of parameter, but of wrong type (i.e. integer instead of text), DBMS again informs user what type of parameter is expected. When all parameters have default values within the body of stored procedure (and all default values are dummy values), stored procedures will not report errors on number and type of parameters and an attacker will probably not figure out what is missing.

```

[SQL Server]
> execute sp_InsertOrders

Msg 201, Level 16, State 4, Procedure sp_InsertOrders, Line 0
Procedure or function 'sp_InsertOrders' expects parameter
'@datDate', which was not supplied

[Oracle]
> execute sp_InsertOrders
Error starting at line 1 in command:
execute sp_InsertOrders
Error report:
ORA-06550: line 1, column 7:
PLS-00306: wrong number or types of arguments in call to 'SP_INSERTORDERS'
ORA-06550: line 1, column 7:
PL/SQL: Statement ignored

```

**Fig. 1.** The example of error report on calling stored procedure without parameters in SQL Server and Oracle

The design of stored procedure's body is presented in Figure 2. The first command in each stored procedure is call to permission checking function. The parameters of this function are: a stored procedure's name, user id, client IP address, application name and stored procedure's optional parameter - web IP address - sent by web application. Considering that, in our case, we have two application types: web application (where client IP address detected by DBMS is those of Application Server) and client application (where IP address is from known IP range), the optional web IP address is obtained from web application server. Obviously when client IP address is that of Application server, the Web IP address must not be dummy address.

```

Create or alter procedure up_UpdateEmpAge(
    par1 int = -1,
    par2 int = -1,
    par3 varchar(15) = '000.000.000.000'
)
BEGIN
    if not fnValidateIP(par3)
        return
    if not fnCheckPermission(user_id, application_name, client_ip,
        'up_UpdateEmpAge', par3)
        return

    if not fnValidateParam(par1, 'int', 'UpdateEmpAge', user_id)
        return

    if not fnValidateParam(par2, 'int', 'UpdateEmpAge', user_id)
        return

    UPDATE employees
    SET age = par2
    WHERE employee_id = par1

    INSERT into LOG(pname, p1, p2, p3, uid, appname, cliIP, webIP,
        affRows)
    VALUES (up_UpdateEmpAge, par1, par2, par3, user_id, application_name,
        client_ip, par3, @@rowcount)
END

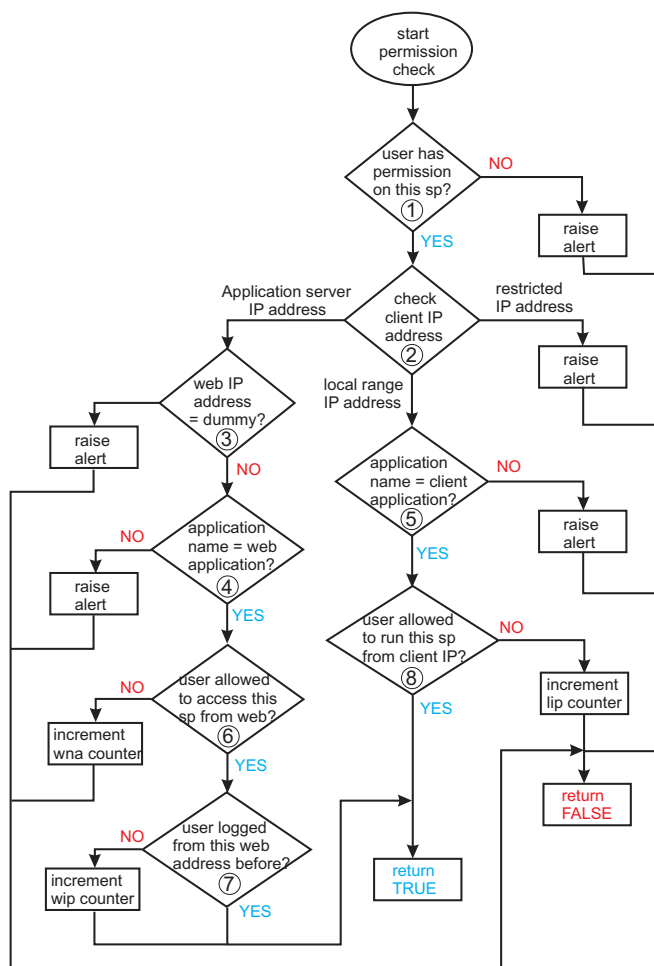
```

**Fig. 2.** The design of Database stored procedure in SQL Server

#### 4.1 Permission Checking Function

As can be seen from Figure 3, in the step 1 check permission function checks if a user has permission to run this stored procedure, then (if answer is Yes) it checks in the step 2 the client IP address. If it is the one of the Application server, the Web IP address is checked against the dummy code in the step 3. If it is, the alert is raised and function returns false. But, if it is not the dummy IP address, the next check (in the step 4) is if procedure is called from the web

application. The alert is raised when the web application name is not recognised. In the step 6, the user is checked if he/she is allowed to access this stored procedure from web.



**Fig. 3.** The flow diagram of checking procedures in check permission function

If user is not allowed to run this procedure from web, the `web_not_allowed` counter for that user is increased and function returns false. But, if user is allowed to run this procedure from the web, the check in step 7 is performed in order to find out if the user logged from this web address before. If not, the `web_IP_not_recognised` counter is incremented, but function returns true until the counter reach predefined limit. If IP address of the client IP address in the output of the step is recognised from the range of allowed IP addresses, then the client application name is checked in the step 5 and check in the step 8 is performed if the user is allowed to run the procedure from that IP address. If not, the counter `local_IP_not_allowed` is incremented for that user and function returns false. Otherwise function returns true. If client IP address does not belong to the predefined range of IP addresses in the step 2, then the alert is raised and function returns false.

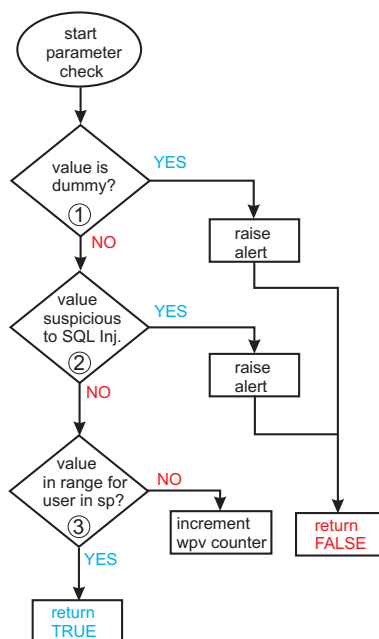


Fig. 4. The flow of checking procedures in check permission function

## 4.2 Parameter Checking Function

When check permission function returns true, the check parameters function starts. The flow diagram of this function is presented in Figure 4. Every parameter's value is checked against its dummy code (when recognised application is used the parameters do not have dummy values). Then the value of parameter is checked against SQL injection key words. If check fails in any of mentioned steps, the function returns false and alert is raised. In the final step it is checked if the parameter's value is within the predefined range of allowed values for that parameter and for that user. There are also some optional parameters the user has to select through the application (optional filters for querying, optional key values for inserting/modifying data) and these values can be checked from appropriate tables. If the values are not within predefined ranges (i.e. some users want to see data about customers they are not in charge to handle, or write new order to some of other user's vendor) the function returns false and the wrong\_par\_value counter for that user is incremented. We must admit that the maintenance of detailed permissions, relations of users and IP addresses; and users and parameters' ranges is not so easy where the number of users connecting to database is high. The proposed solution is implemented to database with less than 100 users.

For all of mentioned counters, the DB Owner can set maximum values and the time-frame for reaching maximum values. Depending on the risk assessment some maximum values are lower and some higher. When any of counters reach own maximum value within the time frame set, the user is blocked temporarily until the problem (the reason why user tried to access the database in inappropriate way) is resolved. When any of two checking functions returns false, the remaining part of stored procedure is not executed and procedure returns nothing that can help eventual attacker to continue with attacks. The record in Issue table is inserted for any path from algorithms presented in Figures 3 and 4 that leads to raising alert or incrementing counters. The internal log table (the log of all successful calls to stored procedures) is transferred to analysis DB at the end of the business day in order not to use too much space in the transactional DB. From analysis DB, the updated aggregated table (AggregatedLog) is generated for validation purposes.

For example when the step 7 from permission checking function is performed, the actual data are compared against aggregated data, because more resources would be spared if comparison would be done against data in historical log tables.

## 5 Results

Before implementation of the solution presented, internal testing was performed in testing environment. The database responded as expected.

In order to simulate initially expected attacks in the production environment, one of IT employee figured out the password of one of colleagues (password consisted of his spouse name and two digits - the year of his marriage) and tried to retrieve data with his credentials using the client application. IT employee initially tried attack from own computer, but the client application didn't start any stored procedure. IP address check was working. But, after the end of working time the IT employee connected to DB from the computer of employee which credentials were compromised and did some retrieval of data, because IT employee knew some of the key parameters the compromised colleague used (a short list of customers). Fortunately the IT employee couldn't modify the data because he didn't know some other parameters related to retrieved customers and the limit for `wrong_par_value` counter was set to be low. The alert was raised after couple of failures. After described case occurred, the stronger password policy is implemented, the additional check on working time is implemented in stored procedures and only database accounts (not the domain accounts) are implemented in database. The following cases were reported since then:

- it happened several times that DB owner didn't update all permissions of some users that were required in order to work with additional business functionalities (in some cases the IP address of user's computer is updated, but not the ranges of parameters' values or vice versa);
- it was recorded that some users had relatively "large number" of typo errors and they reached the limit of `wrong_par_value` counter and were temporarily forbidden to log in;
- some users who hadn't enough permissions for some functions were instructed by other employees (that actually had permissions for these functions) how to proceed to modify data, not figuring out why system didn't react.

The embedded protection had also an impact on data correctness, although it was not the intention. The users who were blocked because of type errors used to take more attention on typing.

## 6 Conclusion

In order to protect software with database from an attacker decently, it is necessary to continuously monitor and investigate all published cases of software vulnerabilities. One approach in improving security of database is to enable its self-protection by controlling the access to database by using only stored procedures with built-in validations on user's permissions and input data. In this way a shield on detailed user permissions and input data that might cause the potential security flaws is embedded. Calls to stored procedures that potentially jeopardise the database security (when any checking function returns false) are logged in special table. By continuous analysis of data in this table some conclusions can be derived and new protection measures can be implemented in order to improve database security.



**Acknowledgements.** This work has been done within the project 'Optimal Software Quality Management Framework', supported in part by the Ministry of Science and Technological Development of the Republic of Serbia under Project No.TR-35026 and Project No. III 44006.

## References

1. Obradović, S., Ilić S.S., and Marković, V.: Responsibility of management related to data security, Proceedings of international conference UNITECH, Gabrovo, Bulgaria (2009).
2. Krause, M, Tipton, H.F.: Handbook of Information Security Management, CRC Press LLC, <http://www.ccert.edu.cn/education/cissp/hism/ewtoc.html>
3. Peng Liu, DAIS: A Real-time Data Attack Isolation System for Commercial Database Applications, Proceedings of 17th Annual Computer Security Applications Conference - ACSAC pp. 219-229 (2001)
4. Stonebraker, M., Wong, E.: Access Control in a relational Database Management System by Query Modification. Proc. ACM Annual Conf., 1974. p.180-186. [doi:10. 1145/800182.810400]
5. Sehta,N, Jain, S.: A Fine Grained Access Control Model for Relational Databases, International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 3 (1), 2012, pp. 3183 - 3186
6. Halfond, W.G.J., Viegas J., Orso, A.: A Classification of SQL-Injection Attacks and Countermeasures, Proceedings of the International Symposium on Secure Software Engineering. Washington D.C. March 2006.
7. Buehrer, G.T., Weide, B.W. and Sivilotti, P. A. G.: Using Parse Tree Validation to Prevent SQL Injection Attacks, Fifth International Workshop on Software Engineering and Middleware - SEM 2005, Lisbon, Portugal (2005).
8. Sunitha, K.V.N. and Sridevi, M.: Automated Detection System for SQL Injection Attack, International Journal of Computer Science and Security (IJCSS), Volume 4: Issue 4, pp. 426-435 (2009).
9. Adi, E., Salomo, I.: Detect and Sanitise Encoded Cross-Site Scripting and SQL Injection Attack Strings Using a Hash Map, Australian Information Security Management Conference, (2010).
10. Yang, Z., Sesay, S., Chen, J. and Du Xu: A Secure Database Encryption Scheme, American Journal of Applied Sciences 1 (4): 327-331, (2004).
11. Jabbour, G. and Menasce, D.A.: Policy-Based Enforcement of Database Security Configuration through Autonomic Capabilities, Proceedings of the Fourth International Conference on Autonomic and Autonomous Systems ICAS'08 (2008).
12. Ilić, S.S., Lazić, Lj., and Spalević, P.: *One approach to the testing of security of proposed database application software*, WSEAS Book "Recent Researches in Computer Science", pp. 475-480, (2011).
13. Jeong, Min A., Kim, Jung-Ja and Won, Y.: A Flexible Database Security System Using multiple Access Control Policies, International Conference on Database and Expert Systems Applications - DEXA 2003, LNCS 2736, pp. 876-885, (2003).

# Choosing The Model for Solving the Problem of Lexical Selection for Kazakh Language on Free/Open-Source Platform Apertium

Aidana Karibayeva, Dina Amirova, and Malika Abakan

Al-Farabi Kazakh National University, Information Systems Chair,  
Al-Farabi av., 71, 050040 Almaty, Kazakhstan

a.s.karibayeva@gmail.com, amirovatdina@gmail.com, mayerabak@gmail.com  
a.s.karibayeva@gmail.com, amirovatdina@gmail.com, mayerabak@gmail.com  
<http://www.kaznu.kz>

**Abstract.** This paper describes process of choosing the model for lexical selection for Kazakh language as a source and target language. We will consider existing models and methods for solving problem of lexical ambiguity. In this paper we will show models which can be applied to Kazakh language. We will consider to rule-based lexical selection and first works which be done to this time for solving this type of problem.

**Keywords:** machine translation, Apertium, ambiguity, lexical selection, HMM, maximum entropy model, MT.

## 1 Introduction

Today it's important to create machine translation for Kazakh language. When we create machine translation we faced with problem of ambiguity. The ambiguity is an open problem of natural language processing and each machine translation system faces it. Solving the task of ambiguity is a difficult task. Today, there are many algorithms and models of resolving it. Linguists distinguish some kind of ambiguity. There are: lexical, morphological, syntactic. We will consider lexical ambiguity. Lexical selection is a choosing one translation of the word in target language by context of source language. Lexical selection is a main tasks of processing language. Kazakh language has a great number of ambiguity, when translating from English. For example, word 'bet' can translated as 'face' and 'page'. We solved this ambiguity by writing rules, which we will consider at the next section [1].

## 2 The Apertium platform

Apertium is machine translation platform. The Apertium platform module's work seems like pipeline. The Apertium machine translation system consists following module [2]:

- Deformatter. It separates the text to be translated from the formatting tags. Formatting tags are encapsulated as 'superblanks' that are placed between words in such a way that the remaining modules see them as regular blanks.
- Morphological analyser. For each surface form (that is, for each lexical unit as it appears in the text), the morphological analyser generates one or more lexical forms composed of: lemma (dictionary or citation form), lexical category (or part-of-speech), and inflection information. The morphological analyser executes a finite-state transducer generated by compiling a morphological dictionary for the source language. Lexical units containing more than one word (multiword lexical units) are analyzed as a single lexical unit. Morphological analyser

uses a finite state transducer based on two-level rules (in the case of Kazakh, `apertium-kaz.kaz.lexc`, `apertium-kaz.kaz.twol`). This module therefore separates lexemes and processes morphological analysis, and then returns possible lexical forms.

- Part-of-speech (POS) tagger. Apertium's POS tagger is based on a statistical model based on hidden Markov models which processes the result of the application of on constraint-grammar rules (Karlsson 2005), which are used to discard some analyses using simple rules (written in `apertium-kaz.kaz.rlx`) based on context.
- Lexical transfer. This module uses a bilingual dictionary (`apertium-eng-kaz.eng-kaz.dix`) which has very simple structure. The module reads each source-language lexical form and finds one or more corresponding target-language lexical forms. Multiword units are translated as a single word.
- Lexical selection. It uses rules that select for those lexical words having many translations, one of the translations in the target language according to context. All rules are written in file `apertium-eng-kaz.kaz-eng.lrx`. Lexical selection is the focus of this paper, and will be described in section 3;
- Structural transfer. This module identifies sequences of lexical forms (phrases or segments), which need syntactical processing (handling of number, prepositions, etc.) to be translated. It uses files with rules, which specify the syntactic transformation as a cascaded process. Transfer rules, which transform lexical-form sequences into a new sequences for the target language, perform the work in this module.
- Morphological generator. From the sequence of target-language lexical forms produced by the structural transfer, it generates a corresponding sequence of target language surface forms. The morphological generator executes a finite-state transducer generated by compiling a morphological dictionary for the target language.
- Post-generator. It takes care of some minor orthographical operations in the target language (for instance, it generates the English form `cannot` from `can` and `not`). This module is generated from file with rules which are very similar in format to dictionary files.

### 3 The lexical selection

The lexical selection is an open problem of each translation system. One of the main tasks of word processing is the problem of lexical choice, which is associated with the task of word-sense disambiguation. It is the correct choice of the word or term in accordance with the context in which they are used. Word-sense disambiguation is used in different areas: to improve the quality of machine translation, improve the accuracy of methods of classification and clustering texts, information retrieval and other applications.

#### 3.1 Rule-based lexical selection

In rule-based free/open source platform Apertium [1] this problem is solved by module of lexical selection (F.M. Tyers, M.L. Forcada 2013), where rules are written by hand. As you know personal pronoun 'ol' from Kazakh is translated as 'he', 'she' and 'it' into English. We wrote the rule of lexical selection in which translation is taken by depending located near words. Generally, hand-written rules do not cover the entire context. So, we want to use statistics methods and models to solve this problem, which connected with training corpora to generate rules automatically.

Rule-based lexical selection is written in file `apertium-eng-kaz.eng-kaz.lrx` for language pairs from English into Kazakh, meanwhile in file `apertium-eng-kaz.kaz-eng.lrx` rules are written to Kazakh into English language pairs. This lexical module in the translation ambiguous word

input language to the target language selected one lexical form of all possible with the help of rules depending on the context. All the rules are written in the XML-format.

*The content of the lexical rules:*

```
<rule> - start of rule;
  <match lemma="the word in english/kazakh" - defining word;
tags="part of speech" - tag of the word's part of speech,
for example, noun - "n", adjective - "adj", and etc.;
  <select lemma="selected word" - selection of a particular ambiguous word translation;
tags="part of speech" - tag of the word's part of speech;
</match>,
</rule> - closing of the relevant tags.
```

*Example of lexical selection rule for 'zhas'*

```
<rule>
  <match lemma="year" tags="n.pl">
  <select lemma="zhyl" tags="n.*"/>
  </match>
</rule>
<rule>
  <match lemma="year" tags="n.pl">
  <select lemma="zhas" tags="n.*"/>
  </match>
  <match lemma="old" tags="adj.*"/>
</rule>
```

(Example from apertium-eng-kaz.eng-kaz.lrx)

### 3.2 Statistical-based lexical selection

Statistical-based lexical selection connected with corpora by counting frequency of collocation or words. When we use statistical lexical selection it means that we choose the most likely translation with their probability.

One of the important part of statistical machine translation system is to make corpora of large volumes. One of the difficult task is a collection of parallel corpora, ie, in our case, to gather the corpus of Kazakh and the corpus of the English. Presently, we have been developing a bilingual corpus, which already contains 4255 sentences. We collect this corpora from fairytales, books. Today we are training this corpora, because Apertium works with trained corpora. To receiving corpora-based lexical selection we need aligned corpora, which is not easy to do. As we mention above, all Turkic language have a complex morphology. So, some words can be aligned to several words.

We want to use both of type of lexical selection, which is meant above. Because, rule-based did not cover all cases for ambiguity. At the first step of creating statistical-based lexical selection we collect and develop bilingual corpus. Today, corpus-based techniques for lexical selection is widely used. Corpus is main part of any corpus-based lexical selection. Preparing corpora depends on complexity of language. As you know Kazakh has complex morphology. We collect

corpus from books, fairytales. At the second step of our work we are training system by adding words to monolingual dictionary of Kazakh (apertium-kaz.kaz.lexc) and English (apertium-eng-kaz.eng.dix) language and adding to bilingual dictionary (apertium-eng-kaz.kaz-eng.dix).

## 4 Models

### 4.1 Hidden Markov model

First suitable model is Hidden Markov model (HMM), which is the main model of statistical modelling in language processing. In HMM model disambiguation is solved by assigning the probability of word. Knowing the most probable tags in context, translation system can decide which translation of word or collocation is adequate. This model requires a big corpus of Kazakh language to receive accurate translation. If a given possible translation appears aligned to a word in a given context more frequently than other possible translations, then we generate a rule which selects the aligned translation in that same context over other translations in that context [3]. We know that the main source of knowledge are dictionaries and encyclopedias. Linguists created thesauri, semantic networks and other specialized structures to establish the relationships between the values. One of the most popular model based on knowledge is a hidden Markov model. Methods based on different variations of hidden Markov model works much better, since it takes into account the context.

The problem of solving the lexical ambiguity can be reformulated as a maximization problem using the formalism of Hidden Markov Models. [16].

Hidden Markov model (Hidden Markov Model) - a statistical model that can be used to solve the classification problem of hidden variables based on observable. A Markov model is a finite state machine, the state transitions performed with a given probability. The process starts from a special starting state, then through discrete points in time it can go into new States.

In a HMM model each state with a given probability corresponds to the observed state. In accordance with the Markov assumption, the current state of the automaton depends only on a finite number of previous, with the change of the law itself states does not change over time. The number of states that need to remember to go to a new state, called the order of the model. The model may be a first order (present condition depends only on the last), n-th order (the current state is dependent on a predetermined number of preceding), and the alternating order (the order is determined according to a certain law).

Hidden Markov models are used to solve three major problems:

- calculating the probability of a observations sequence;
- calculating the most plausible explanation for the observed sequence;
- training model parameters.

For solving the lexical selection problem the words meanings serve as a hidden states, words from the text serve as observations. To use the model, you need to estimate the parameters - the matrix of transition probabilities between states (probability that after a word with the value of the test will met next) and probability matrix of observations (the probability that a given word has a predetermined value). To estimate these parameters using annotated corpus, dictionaries, network documentation. First Order HMM have complexity  $O(N^2T)$ , easy to understand, but for many situations receive insufficient accuracy.

### 4.2 Maximum entropy model

Second model is Maximum Entropy Markov Model (MEMM) that is used to resolve morphological ambiguity. Morphological ambiguity is the main study object of the problem of

determining the word's parts of speech (part of speech tagging). However, modern systems are able to effectively solve the problem using methods of machine learning, such as the support vector machine method or the maximum entropy method, and show the accuracy more than 97

We decided to consider the HMM model as most suitable to Kazakh language and will use this model in machine translation systems, that have Kazakh language as a source and as a target language, namely in MT system from Kazakh into English (and vice versa) and from Russian into Kazakh language.

Maximum entropy lexical selection model includes a set of binary functions and appropriate weights for each function. The feature is defined as combination of 't' and 'c', where 't' is a translation, and 'c' – is a source language context.

## 5 Results

Current version of system, namely English-Kazakh (and vice versa) can translate simple phrases with ambiguity. In English-Kazakh lexical selection rules and n Kazakh-English lexical selection rules contain about 30 rules.

## 6 Conclusion

Current lexical selection rules translate some cases of phrases and solve problem of ambiguity in some cases. We have presented a lexical module for Kazakh language on free/open-source platform Apertium. A lexical selection problem is solved by writing rules for words. In the future we would like to use statistical model for more effective solving the problem of lexical selection, namely maximum entropy model. This model shows the high accuracy. So, now we are preparing parallel corpora for English and Kazakh language, which are collected from fairy-tales and books. Then we would train it as statistical machine translation system has a property of «self-learning». As a method we would use supervised learning. This method based on word-alignment from corpora.

We hope that the using of this model to the problem of disambiguation for Kazakh language will give us the opportunity to have a more accurate translation of texts.

In future work is planned to generate lexical selection rules automatically from bilingual corpora to improve translation quality.

## References

1. The Apertium machine translation platform: <http://apertium.org/>
2. Sundetova, A., Karibayeva A., Tukeyev Ua.: STRUCTURAL TRANSFER RULES FOR KAZAKH-TO-ENGLISH MACHINE TRANSLATION IN THE FREE/OPEN-SOURCE PLATFORM APERTIUM. Proceedings of the International Conference on Computer processing of Turkic Languages "TURKLANG'14 Istanbul(2014)
3. Francis. M. Tyers, Felipe Sanshez-Martinez, Mikel L. Forcada. Flexible finite-state lexical selection for rule-based machine translation (2012)
4. Daniel Ramage. Hidden Markov Models Fundamentals. CS229 Section Notes. 2007

# Construction of the Database and the Compilation Tools in CANRDB

Venera Kurmangaliyeva, Meruert Takibayeva,  
Masayuki Aikawa <sup>a)</sup>, and Nurgali Takibayev

al-Farabi Kazakh National University, av. al-Farabi 71,  
Almaty 050040, Kazakhstan;

<sup>a)</sup> Hokkaido University, Sapporo, Japan;  
{takibayev@gmail.com, aikawa@sci.hokudai.ac.jp}

**Abstract.** The development of nuclear physics, astrophysics, nuclear engineering, radiation medicine and ecology stipulates interest of scientific community to the usage of nuclear data. The Central Asian Nuclear Reactions Database (CANRDB) started in 2012 first for the local nuclear data users. The CANRDB then joined the International Network of Nuclear Reactions. Types and advantages of most common database management systems and server platforms are considered. The description of the main features of the new nuclear data compilation software developed in collaboration by CANRDB and Japanese Charged Particle Reaction Group (Hokkaido University, Japan) is given.

**Keywords:** Nuclear data, CANRDB, compilations, database management system, EXFOR, HENDL.

## 1 Introduction

Nuclear reaction data is essential for research and development in nuclear physics, astrophysics, nuclear engineering, radiation ecology and radiation medicine. These fields require a variety of nuclear reactions data in the form of database accessible to nuclear data users around the world.

The Central Asian Nuclear Reactions Data Base (CANRDB) at al-Farabi Kazakh National University is a member of the International Network of Nuclear Reaction Data Centres (NRDC) under the auspices of International Atomic Energy Agency (IAEA); CANRDB develops and provides access to the nuclear reactions database, reference and educational materials for scientists, nuclear data specialists and students in various fields [1].

One of the necessary tools to handle the massive amount of nuclear reactions database is a database management system (DBMS). A DBMS is a structured system for collecting, retrieving and displaying information, designed to allow the definition, creation, querying, updating, and administration of databases. Well-known DBMSs include MySQL, PostgreSQL, Microsoft SQL Server, Oracle, Sybase and IBM DB2 [2].

Generally, there are two categories of the database management systems: desktop databases and server databases. The difference between these two is that desktop databases are oriented toward single-user applications and reside on standard personal computers, while server databases are designed for organizations and allow to access, manipulate and process large amounts of data efficiently and simultaneously by many users. There is a number of benefits achieved through using a server-based system:

Flexibility. Server-based databases have large functional enabling them to handle various data management problems. They have programmer-friendly application programmer interfaces (or APIs) assuring quick development of the database and the applications. Most platforms are compatible with various operating systems.

Powerful performance. Modern server-based databases can manage multiple high-speed processors, clustered servers, high bandwidth connectivity and fault tolerant storage technology.

Scalability. If the necessary hardware resources are provided, server databases are capable of handling a rapidly expanding amount of users and/or data.

Depending on the way how the data is structured, there can be two types of DBMS: a Network DBMS and Relational DBMS (RDBMS).

In a Network DBMS the relationships among data are of type many-to-many that appears in the form of a network. Thus, the structure of a network database becomes extremely complicated, and one record can be used as a key of the entire database.

A relational DBMS stores data in separate tables rather than puts all the data into one large repository. Such data configuration adds tremendous speed and flexibility: the tables are linked by defined relations making it possible to combine data from several tables upon request. A number of modern database management systems such as Oracle, MS SQL, IBM DB2 and MySQL are RDBMS.

The Central Asian Nuclear Reactions Data Base uses the MySQL database management system, as it fully covers all the requirements of nuclear data storage and has a number of advantages.

## 2 Construction of CANRDB

It is an open source Relational DBMS, which makes it more flexible and allows configuring the database for the special needs of the nuclear reactions database. One of the main requirements for a nuclear database is security: MySQL includes solid data security layers that protect sensitive data from intruders, allows setting rights with various ranges of privileges to the users and has a password encryption tool.

Another advantage of MySQL is that while most relational databases require a basic knowledge of SQL, MySQL is very easy to use. It allows building and interacting with MySQL with only a few simple SQL statements.

In the interest of speed, MySQL designers decided to offer fewer features than other major database competitors do (such as Sybase\* and Oracle\*). However, despite having fewer features than the other commercial database products, MySQL still offers all of the features required by most database developers [3].

The scalability of MySQL allows handling large amount of data up to 50 million rows, and has a possibility to increase the default file size limit from 4 GB up to 8 TB of data. MySQL itself runs across many operating systems and supports several development interfaces including JDBC, ODBC, and scripting (PHP and Perl), making it possible to create database solutions that run on all major platforms, including Windows\* Linux\*, many varieties of UNIX\* (such as Sun\* Solaris\*, AIX, and DEC\* UNIX), OS/2, FreeBSD\*, and others.

MySQL is made available under the GNU General Public License (GPL) free of charge and under commercial license for commercial use.

Many of the world's largest and fastest-growing organizations including Facebook, Google, Adobe and Alcatel use MySQL to power their high-volume web sites, business-critical systems and packaged software.

MySQL DBMS in CANRDB database is realized as a WAMP, "Windows, Apache, MySQL, and PHP" application server platform. Apache is the most popular open source web server and PHP is a widely used general-purpose server-side scripting language designed to produce dynamic web pages. By combining these components, WAMP server platform allows users to set up a server



locally to create dynamic web applications with Apache, PHP and the MySQL database in the same development conditions as on the production server.

One of the great benefits of WAMP server platform is that it allows to develop, upgrade components, perform any web development task and carefully test everything offline first, which reduces the risks of creating problems on the live server.

The International Network of Nuclear Reaction Data Centres (NRDC) uses a special format of nuclear data EXFOR (EXchangeFORmat) containing extensive data on nuclear reactions with photons, neutrons, charged particles, heavy ions, the properties and structure of atomic nuclei.

Systematic collection of experimental neutron nuclear data started in the 1960s and was conducted by four data centres, each using its own data system: Brookhaven National Laboratory (USA), OECD Nuclear Energy Agency (France), Fiziko-Energeticheskij Institut (Obninsk, Russia) and International Atomic Energy Agency (Austria). It became obvious soon that centres' activities required coordination, and in 1970 software experts and physicists from these four centres formulated a unified nuclear data exchange format "EXFOR" in its initial form. Prompt exchange of compiled nuclear data among centres made it available to the increasing community of data users around the world.

Today EXFOR contains data from more than 20,000 experiments with more than 129,000 data tables [4]. Maintenance of such a massive array of various data requires specialized tools with the possibility to compile, input and digitize numerical and graphic information. All centres of NRDC Network use different tools for internal data storage, input, output and retrieval making data available to users.

CANRDB in collaboration with Japanese Charged Particle Reaction Group (JCPRG) at Hokkaido University is developing a new version of a convenient and user-friendly web-based system (HENDL) for nuclear data input with possibility of output in EXFOR and other specific data formats [5].

### 3 Compilation tools in CANRDB

Nuclear data editor, such as HENDL, is a main working tool of a modern nuclear data specialist or compiler. The first version of HENDL was developed in 2001 by Dr. Naohiko Otsuka at the Hokkaido University's Japanese Charged Particle Reaction Group (JCPRG), and is now widely used by a number of centres of the NRDC Network. Yet due to rapid development of data systems and the EXFOR format itself, necessity to update the main tool of nuclear data input becomes inevitable, and such changes dictate a number of requirements [6].

Accessibility. We believe that in drastically and dynamically changing world, accessibility of a working tool is a key to effective and up-to-date data input and exchange. It is essential for a data editor to be available anytime from any device regardless of operating system and the compiler's location. Thus we opt for a browser-based editor, which should be available from all major and popular browsers (Chrome, Opera, Safari, Firefox, etc.). This option is realized in the first version of HENDL, and will be maintained in the new version as well.

Another important requirement is related to the fact that more and more new experiments, institutions, installations and facilities and, accordingly, new data types appear every year, making it necessary to regularly update the EXFOR format itself. Thus a nuclear data editor should be updatable simultaneously with EXFOR updates.

Third requirement concerns the input system of the data editor. EXFOR is a highly-specific and complicated data format with a number of rules and nuances that require large experience and proficiency in work with nuclear data. However, the scope of new experimental nuclear data grow dramatically and centres experience the lack of human resources since the education of

new compilers takes a lot of time. One way to solve this problem is to create a compiler-friendly data editor that would not require deep knowledge of EXFOR and would improve the quality of compilation. This can be realized through inclusion of extensive dictionaries of EXFOR codes and build-in manuals and step-by-step guides to the editor and development of error checking tool.

And the last requirement relates to the usability of nuclear data search engine for outside users. Usually the search tool also demands some knowledge of EXFOR from users, and this seriously reduces the usage of the network. To solve this problem we aim to use both advanced search and a free text search with advised options in our data retrieval system.

All these requirements will be realized in the new version of the HENDL editor, which is now on the stage of active development. The new HENDL editor will include the function of check and graphical visualization of the entered data and will be accessible for the staff of all Nuclear Data Centres.

The development of the CANRDB nuclear reaction database and the new nuclear data editor will increase the activity and effectiveness of CANRDB in extending the scope of experimental nuclear data from the Central Asian region and contribute to the nuclear data development around the world.

The CANRDB takes part in the international cooperation and technical meetings of IAEA. It is important that special educational and reference parts of the Database are formed for students of technical disciplines and junior scientists who are the main users.

**Acknowledgments.** The authors acknowledge Kiyoshi Kato, Naohiko Otsuka and V.V. Varlamov for their help and participation in the CANRDB development.

## References

1. Takibayev, N., Kenzhebayev, N., Kurmangaliyeva, V.: Introduction to the Central Asian Nuclear Reaction Database, Summary Report of the Technical Meeting on International Network of Nuclear Reaction Data Centres (6-9 May 2014, Smolenice, Slovakia), IAEA Nuclear Data Section, Vienna International Centre, Austria, 55-57 (2014)
2. <http://www.databasejournal.com>
3. [http://www.novell.com/documentation/nw65/web\\_mysql\\_nw/data/aj5bj52.html](http://www.novell.com/documentation/nw65/web_mysql_nw/data/aj5bj52.html)
4. Otsuka, N.: Computation of Experimental Nuclear Reaction Data in Central Asia, Proceedings of the 4th Asian Nuclear Reaction Database development Workshop, Almaty, 23-25 October 2013, INDC (NDS)-0633 Distr. G+NC, IAEA, Austria, February 2014, pp. 56-60.
5. Kenzhebayev, N., Kurmangaliyeva, V., Otsuka, N., Takibayev, N.: Joint activities with IAEA on uploading of scientific papers from Kazakhstan and Uzbekistan into the EXFOR database, Proceedings of the Fifth AASPP Workshop on Asian Nuclear Reaction Database Development, BARC, Mumbai, India, 22-24 September 2014, INDC(IND)-0048 Distr. NC, IAEA, Austria, February 2015, pp. 152-154.
6. Aikawa, M.: Japanese compilation tools, Summary Report of the Technical Meeting on International Network of Nuclear Reaction Data Centres (6-9 May 2014, Smolenice, Slovakia), IAEA Nuclear Data Section, Vienna International Centre, Austria, pp. 51-54 (2014)

# Parallel Algorithm of RDF Data Compression and Decompression Based on MapReduce Hadoop Technology

M. Mansurova, E. Alimzhanov, E. Dadykina

Al-Farabi Kazakh National University, Almaty, Kazakhstan  
mansurova01@mail.ru, aermek81@gmail.com  
<http://www.kaznu.kz>

**Abstract.** At present, search systems quite successfully deal with information search in the World Web but processing of semantic inquiries, intellectual use of resources are still of semantic an open problem. The necessity of storing and processing of large volumes of data also contributes to creation of new approaches for quick searching and useful information retrieval. Investigations in this direction have been carried out since 1994 after introduction of the notion "Semantic Web" by Tim Berners-Lee, inventor of World Wide Web.

As, when working in the Web, we mainly deal with unstructured or semi-structured data (electronic documents, web pages, etc.) which are designed for perception of people and not computers, the development of effective technologies and algorithms of machine processing of the data is a field of active investigations in the whole world. Due to scientific results obtained in this direction, the methodological base and technologies for the work with great massive of unstructured data have been formed. At present, owing to the growing volumes of the information being processed, the problems of organization of high performance computing for the work in Semantic Web have become actually. The technology of distributed computing appeared to be the most effective which is explained by their high scalability, flexibility and high productivity. For processing of semi-structured data, this Project proposes to use the model of high productive distributed computing on MapReduce. On the basis of theoretical and experimental results of investigations in this field one can assert that algorithms of compression and decompression of RDF dictionaries are most successfully realized using the model of distributed computing MapReduce. The technology MapReduce Hadoop allows the programmer to concentrate upon the logic of processing; the problems of realization of distributed computing, fault tolerance, load balancing are solved at the level of the technology.

**Keywords:** RDF dictionaries, MapReduce Hadoop, compression and decompression algorithms.

## 1 Introduction

Different research groups [1-3] deal with the problems of development and creation of information search systems which are able to automatically search and retrieve new information from semi-structured data for scientific community. In such systems, such technologies as JSP, JavaScript, PHP, database server MySQL are used as development tools. In spite of the indisputable advantages of these systems their productivity noticeably decrease with the increase in the volumes of the data being processed.

At present, due to the increase in the volumes of the data being processed the problems of organizing high-performance computing for the work in Semantic web have become actual. And the technologies of distributed computing [4-7] proved to be most effective owing to their high scalability, flexibility and high performance.

In this work, to process semi-structured data, the distributed computing model of MapReduce based on Hadoop platform is proposed to be used. On the basis of theoretical and experimental results of investigations in this field [8-9] one can state that algorithms of compression and decompression of RDF dictionaries, scalable distributed reasoning for retrieval of information from Semantic web are most successfully realized using the MapReduce Hadoop technology. The MapReduce Hadoop technology allows a programmer to concentrate on the logic of processing

and such problems as implementation of distributed computations, failure resistance, and load balancing are solved at the technology level.

In this work we describe the process of construction of RDF dictionaries and propose of parallel algorithm of RDF data compression and decompression with MapReduce based on approach of Jacopo Urbani and others [7]. We have implemented a prototype using the Hadoop framework, and evaluate its performance.

## 2 Dictionary coding algorithm

Semantic web contains billions of statements which are realized using the RDF data model (Resource Description Framework). For a better storing and processing of large volumes of data and high performance of RDF applications, the methods of data compression must be used. Parallel algorithms realized using MapReduce Hadoop technology are effective for compression and decompression of a great amount of RDF data [5-6]. One of the most often used methods for compression of data is based on dictionary coding. In the dictionary coding, every element in a set of data is substituted by a numeric identifier. Using a corresponding dictionary, the data can be compressed into the initial uncompressed form. Due to its simplicity this method of compression is widely used in different fields. In this work, a parallel algorithm of compression and decompression of RDF data using MapReduce Hadoop technology is based on the approach of Jacopo Urbani et al. [5]. The algorithm using the method of creation of dictionaries with supporting the initial structure of the data is realized on Hadoop platform. As a model example, in the framework for construction of knowledge base Protege [10], ontology on the theme of RDF model syntax was constructed resulting in the document in the format .rdf of this ontology which was further supplemented in the framework Sesame [11] for storing and processing of RDF data. Then, a program for a parallel algorithm was written using the dictionary coding in the Java language in the Eclipse integrated framework of development. In the course of work, the Jena library [12], a special Java framework for the work with RDF data, was used. As a result of program execution two files were obtained: one with a dictionary with terms and corresponding codes, the other with compressed RDF data.

## 3 Dictionary coding algorithm based on MapReduce Hadoop technology

The parallel algorithm for creation of dictionary coding on MapReduce Hadoop consists of 3 stages. The first stage is finding of popular terms in the input file and entering them into the dictionary. As RDF data are usually large in volume, at this stage only part of the input file is processed. At the second stage, statements are divided into separate terms, and then coding is performed. In the course of division, identifiers of statements and the corresponding place of the term in the triplet are stored. Then, the terms are coded: if the term is already in the dictionary, it is substituted by a numeric identifier, if this term is not in the dictionary, it is entered into the dictionary and a new numeric identifier is assigned for it. Reconstruction of statements is performed at the third stage. When realizing this mechanism of compression, the problems of data tokenization, prevention of excess duplication of data were solved (see Figure 1).

In order to decompress the data which were compressed with the help of dictionary coding, popular terms are processed separately all over again, and the statements are divided into terms and at the end of the statement are collected into triplets (see Figure 2).

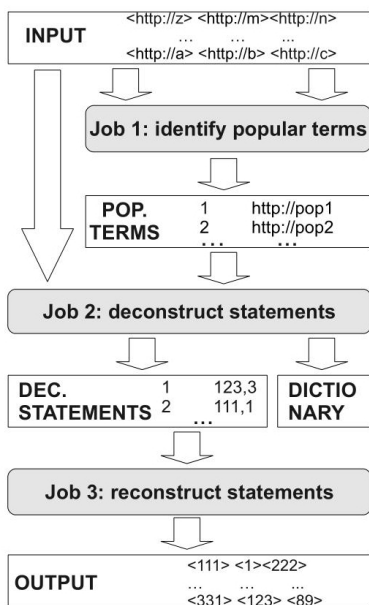


Fig. 1. - Compression algorithm

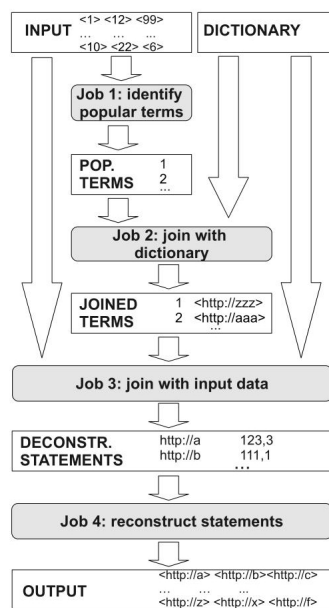


Fig. 2. - Decompression algorithm

## 4 Related work

IDictionary coding is used in widely-known systems of storing RDF data, such as Hexastore, 3Store and Sesame [9-11]. The systems of built-up mechanism of logic output, such as WebPIE [8] and OWLIM [12], also use dictionary coding. A detailed review of compression methods which can be applied to RDF data is presented in [13]. In this article, the authors considered three different methods: standard of compression GZIP, compression on the basis of contiguity of lists and dictionary coding. According to the authors compression considerably depends on the structure of data and the use of dictionary coding is the most effective method if there is a significant amount of identifiers URI in the data sets. In [14], the authors propose a method in which structured coding dictionaries are used for compression of URI in RDF files. This method works at two levels: first, it compresses the URI namespace, and then starts to compress the remaining part of data. Though the work is still at the stage of development, evaluation of the results shows that such method of compression is better than the conventional method GZIP. Another work on compression of URL-addresses is described in [15]. Here, the authors concentrate on the problem of providing an effective web-caching, and they propose a simple algorithm of compression of URL tables. The algorithm is based on the hierarchical decomposition of URL allowing to an aggregate general prefixes and using additional function of hashing to minimize the conflicts between prefixes. Introducing these methods the authors considerably reduced the time and compression of information. Dictionary coding is used not only in Semantic web but also in a number of other fields. For example, in [16], dictionary coding is used for compression of images. In [17], the authors present some parallel methods of compression of data using the existing dictionaries. In some problems, dictionaries are small to be in the main memory; a comparative analysis of the use of different structures of data in the memory is given in [18]. In [19], the authors propose a new structure of data Trie which supports the lines in a sorted comparable to the productivity if tires.

The algorithm of decomposition (Figure 2) requires execution of data connection but the original MapReduce paradigm does not provide ant tool for execution of effective connections. Extension of the programming model MapReduce called MapReduce-Merge [20] is directed to support merging of data. Other structures, such as Pig [21] or Hive [22] built on the basis of Hadoop give SQL-like languages for execution of inquiries on very large data set.

## 5 Experiment design and results

The experimental design and the results of the generated MapReduce program execution on a special deployed Apache Hadoop Mini-Cluster of Laboratory of Computer Science of al-Farabi Kazakh National University are presented below. Apache Hadoop 2.6.0 Mini-Cluster consists of 1 master node and 7 slaves. All slave nodes have Ubuntu 14.04 on board; master node has Ubuntu Server 14.04. Master node hardware characteristics: Hardware: HP ProLiant-BL460c-Gen8, Architecture: x86-64, CPU(s): 4, Model name: Intel(R) Xeon(R) CPU E5-2609 0 @ 2.40GHz. Slaves hardware characteristics: Architecture: x86-64, CPU(s): 4, Model name: Intel(R) Core(TM) i5-2500 CPU @ 3.30GHz. NFS server is configured on master node. Slaves have the same folder mounted with read/write access rights. Nodes are connected using Ethernet devices, this providing up to 1000 Mbps, Intel(R) PRO/1000 Network Connection. Problem sizes are: 12 MB, 46 MB, 93 MB, and 250 MB (12 MB are corresponding to 100 000 statements). From Figure 3 one can see that the data are compressed in the volume up to 80%. The identity of the decoded data is checked by the initial: the loss of data is not detected. Figure 3 presents the dependency of the time of performance of the compression on the volume of data and the number of nodes

on the cluster. It also shows that when we choose a big problem size the computing nodes are more effectively used.

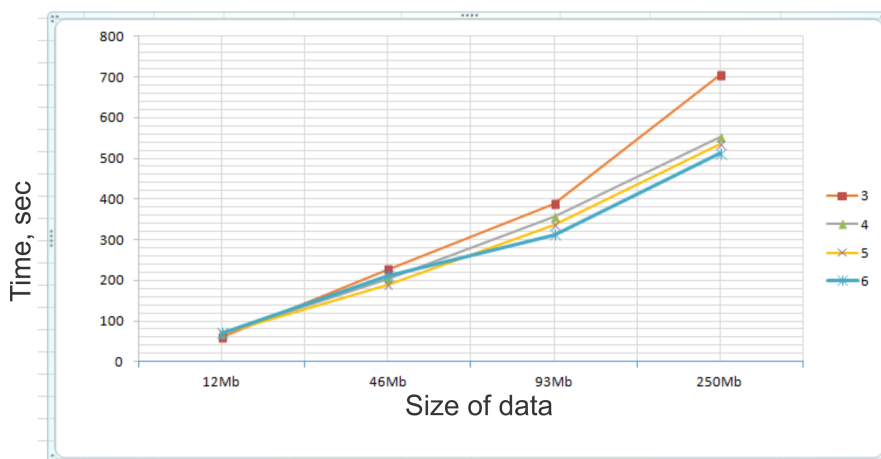


Fig. 3. - Scalability of the compression algorithm on the input size.

## 6 Conclusion

We have implemented a prototype using the Hadoop framework and propose of parallel algorithm of RDF data compression and decompression with MapReduce. The evaluation shows that this approach is able to efficiently compress a large amount of data. This method of data compression can be used further in different applications intended for storing and analysis of RDF data, execution of logic reasoning.

## References

1. Berners-Lee T., Hendler J., Lassila O., *The semantic web*. Scientific American. May 2001; 284(5):34-43.
2. W3C recommendation: Rdf primer. <http://www.w3.org/TR/rdf-primer/>.
3. Linked Life Data. <http://www.linkedlifedata.com>.
4. UK government data website. <http://data.gov.uk>.
5. Official statistics of linked data website. <http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpen-Data/DataSets/Statistics>.
6. Dean J., Ghemawat S., *Mapreduce: simplified data processing on large clusters*. In Proceedings of the USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2004; 137-147.
7. Urbani J., Maassen J., Drost N., Seinstra F., Bal H., *IScalable RDF data compression with MapReduce*. Concurrency and Computation: Practice and Experience. Volume 25, Issue 1, 24-39.
8. Urbani J., Kotoulas S., Maassen J., van Harmelen F., Bal H., *OWL Reasoning with WebPIE: Calculating the Closure of 100 Billion Triples*. In ESWC (1), 2010; 213-227.
9. Abadi D., Marcus A., Madden S., Hollenbach K., *Scalable semantic web data management using vertical partitioning*. In Proceedings VLDB '07 Proceedings of the 33rd international conference on Very large data bases, 2007; 411-422.
10. Weiss C., Karras P., Bernstein A., *Hexastore: sextuple indexing for semantic web data management*. Hexastore: sextuple indexing for semantic web data management.
11. Broekstra J., Kampman A., Van Harmelen F., *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. MIT Press, 2003; 197-222.
12. Kiryakov A., Ognyanov D., Manov D., *OWLIM - a pragmatic semantic repository for OWL*. In Proceedings of the Conference on Web Information Systems Engineering (WISE) Workshops, 2005; 182-192.

13. Fernandez J.D., Gutierrez C., Martinez-Prieto M.A., *RDF compression: basic approaches*. In WWW '10: Proceedings of the 19th International Conference on World wide web. ACM: New York, NY, USA, 2010; 1091-1092.
14. Lee K., Son J.H., Kim G.-W., Kim M.-H., *Web document compaction by compressing URI references in RDF and OWL data*. In ICUIMC, 2008; 163-168.
15. Michel B.S., Nikoloudakis K., Reiher P., Zhang L., *URL forwarding and compression in adaptive web caching*. In Proceedings of INFOCOM 2000. IEEE, 2000; 670-678.
16. Ye Y., Cosman P., *Dictionary design for text image compression with JBIG 2*. IEEE Transactions on Image Processing 2001; 10(6):818-828.
17. Nagumo H., Lu M., Watson K., *Nagumo H., Lu M., Watson K.* In Proceedings of Data Compression Conference, 1995; 162-171.
18. Zobel J., Heinz S., Williams H.E., *In-memory hash tables for accumulating text vocabularies*. Processing Letters 2001; 80(6):271-277. Elsevier Science.
19. Heinz S., Zobel J., Williams H.E., *Burst tries: a fast, efficient data structure for string keys*. ACM Transactions on Information Systems 2002; 20:192-223.
20. Yang H., Dasdan A., Hsiao R., Parker D., *Map-reduce-merge: simplified relational data processing on large clusters*. In Proceedings of the ACM SIGMOD International Conference on Management of Data. ACM 2007; 1040.
21. Olston C., Reed B., Srivastava U., Kumar R., Tomkins A., *Pig Latin: a not-so-foreign language for data processing*. Hexastore: sextuple indexing for semantic web data management. In Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008. ACM 2008; 1099-1110.
22. Thusoo A., Sarma J.S., Jain N., Shao Z., Chakka P., Anthony S., Liu H., Wyckoff P., Murthy R., *Hive: a warehousing solution over a map-reduce framework*. In Proceedings of the VLDB Endowment 2009; 2(2):1626-1629.



# 3D Computer Technologies as a Tool for Contemporary Archaeology

Marek Miłosz<sup>1</sup>, Jerzy Montusiewicz<sup>1</sup>, and Rahim Kayumov<sup>2</sup>

<sup>1</sup> Institute of Computer Science, Lublin University of Technology, Lublin, Poland

<sup>2</sup> Alisher Navoi Samarkand State University, Samarkand, Uzbekistan.  
kayumov@gmail.com, {m.milosz, j.montusiewicz}@pollub.pl

**Abstract.** Modern archeology builds on the achievements of information technology in various fields of its activity. One of them is the creation of virtual three-dimensional (3D) models of sites and artefacts based on the results of excavations. These models are mainly used for visualisation of the past. The authors present the concept of three-dimensional digitisation of fragments of artefacts, recreation of the original shapes of objects and reconstruction of the missing parts in order to create the technological process of material reconstruction of historic artefacts. The study presents the relevant methods, algorithms and modern 3D technology peripheral devices such as scanners and printers, as well as a wide range of research problems whose solution determines the practical applicability of the proposed method.

**Keywords:** Cultural heritage. Archaeological artefacts. 3D models. Technological process of reconstruction.

## 1 Introduction

Information technology supports different areas of human activity, including those associated with a broad concept of Cultural Heritage. It also assists archaeologists by:

- documenting the sites and artefacts (digital photos, scanning, GPS positioning [1], accurate measurements as well as combinations of methods: photogrammetry, 3D characterisation, and geometrical analysis [2], etc.);
- exploration work (non-invasive research methods such as computer and ultrasound tomography [3]);
- reconstruction of objects and sites [4].

There has even arisen the concept of virtual archeology and virtual museum[5] and the protection of Cultural Heritage using 3D technology[6].

Virtual archeology mainly deals with reconstructing archaeological sites and artefacts in virtual reality. Its basic objectives are [4]:

- Dissemination of results of the past play in the virtual world for scientific, educational, tourist or entertainment purposes.
- Reconstruction of sites and artefacts.

Tools producing and disseminating virtual archeology models are identical with those for 3D design used in architecture, technology or computer games. Technology (just as games do) creates 3D models of non-existent objects (or those under construction) with the aim of their production in a technological process. In computer games, on the other hand, artificial worlds are visualised for their exploration by the players. In the case of virtual archeology such artificial worlds are visions of the past, and the generated worlds are virtual museums.

In the case of reconstruction of sites and artefacts, the 3D models created are a tool for further work aimed at conservation and research [7,8] or construction of real-world objects in full or reduced scale [3]. Most of the previous work on 3D models in archaeological research are focused on archiving the appearance and structure [9] of sites and artefacts. Such models can be used, among others, to analyse the progress of degradation of the surface of real objects. An underexplored aspect is the ability to create actual copies of these objects by using 3D replicators.

Physical (not virtual) reconstruction may be associated with the construction of new physical objects (i.e. copies of artefacts), or with complementation of the artefacts by the missing items. 3D modeling in this case is an integral part of the process of reconstruction of historic material artefacts.

## 2 The idea of reconstructing archaeological artefacts

Significant increases in computational power of modern computers and the development of new peripheral devices such as 3D scanners and devices for 3D printing have brought the possibility of their use in archeology. As a result, the idea of reverse engineering and rapid prototyping can be implemented on a large scale also in the area of cultural heritage, and not as heretofore only in manufacturing industries where prototypes of new products were prepared.

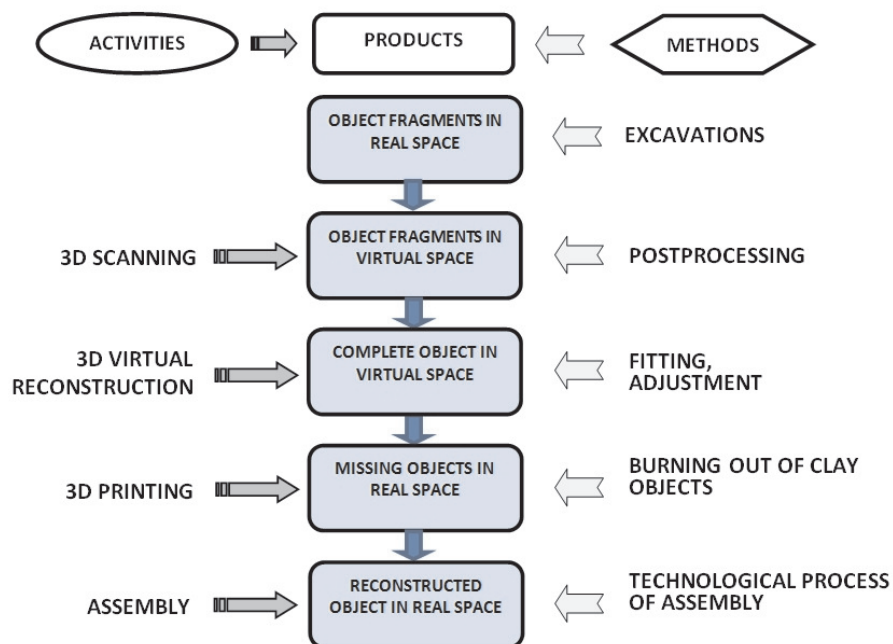
Making copies of museum objects, including archaeological ones, requires a specific approach. On the one hand, these objects are valuable from the point of view of the identity of the nation and the state, on the other hand they often have a very high financial value. These objects may be small, but they must be stored in special climate and safety conditions; others are large enough not to be easily moved. A major problem is the fact that very often archaeological artefacts are damaged and some of their elements are missing. During their reconstruction the missing fragments must be restored.

Figure 1 is a diagram of the automation of the reconstruction process of archaeological objects, together with the assembly process. The elements of the artefact obtained from the excavations are subject to the process of cleaning and segregation. They are then 3D scanned and undergo postprocessing. As a result, solid 3D models are obtained of the fragments of an archaeological artefact and of its surface texture. The adjustment to each other of individual virtual artefact fragments and the recreation of the 3D models of the missing parts results in the construction of a complete virtual object. The missing items are made in a particular material (e.g. clay, which is then burnt out) using 3D printing technology. At the same time the technology is developed of assembling a reconstructed object from fragments (excavated and reconstructed). As a result of the proposed scheme copied or reconstructed (missing) material components of the artefact are obtained and the technological process of their assembly, leading to the emergence of a reconstructed material artefact. The very implementation of the technological process can be manual (as has been happening so far) or automatic, using a precise robot.

## 3 3D scanning and printing

An important element of the proposed idea of reconstructing archaeological artefacts is using specific computer peripheral devices that perform processes of scanning and 3D printing.

The three-dimensional scanning process is a key element of reverse engineering, which in the case of archaeological objects consists in obtaining information about the shape of the elements and storing it in digital form - the so-called point cloud with known coordinates on the XYZ axes. This process, in relation to objects with small dimensions, can be carried out using one of two basic devices, namely laser scanners or scanners using structured light [10,11]. In this paper, laser



**Fig. 1.** General scheme of archaeological object reconstruction in the virtual and real world.

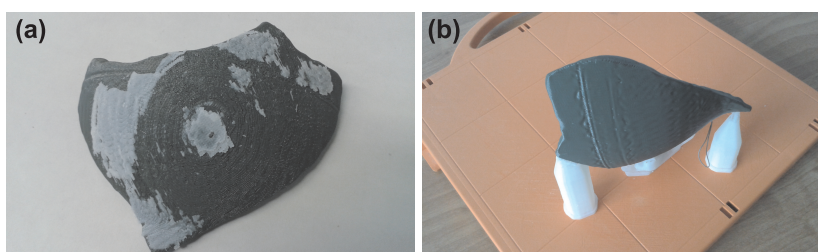
scanners were used: stationary - Roland PICZA LPX-600 and portable - ZScanner ®700, whose suitability for the digitisation of ceramic objects was examined in [12]. The objects of digitisation were elements of dishes made of light scattering material and covered with large and small ornaments, which represented an additional difficulty in the process. An extremely important auxiliary operation in the scanning process is the proper selection of the process parameters in order to accurately reproduce the shapes of the objects, while minimising digitisation time. It should be borne in mind that access to the exhibit can be a one-off, without a second chance of re-scanning. An example of the elements scanned is shown in Figure 2.



**Fig. 2.** Mobile laser scanning two elements of crockery, visible tags allowing to move the device relative to the objects.

Three-dimensional printing is a process of treating excess material, which means that the layers forming the printed item are added to the existing layers and bonded with them in a way depending on the printing technology used. Implementation of 3D printing is possible only after

the post-processing, which allows you to convert the point cloud from the 3D scanning process of an element to a solid body and store it in the appropriate format. The authors applied two replication technologies: powder and FDM (Fused Deposition Modelling). Powder technology requires pre-curing with an adhesive supplied by the head of the machine and final curing after cleaning the item from unnecessary powder. In the FDM printing technology the head supplies molten ABS material, creating directly the replicated object. In this case, there is no curing process, but one should remove the supports formed for retaining the printed element before the hot material stiffens up. A very important step is the positioning of objects in space for printing. In view of the fact that the printing process is carried out in layers of about 0.2 mm (depending on the technology of replication) it may affect the quality and condition of the surface of a printed copy. This is illustrated in Figure 3. The printing process is very long and takes from 4 to 11 hours depending on the number and size of the components placed.



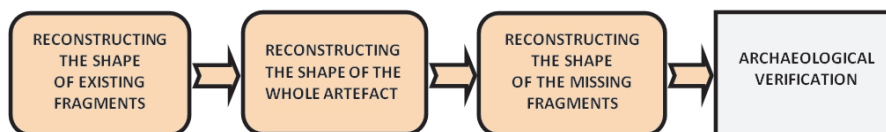
**Fig. 3.** Element printed in FDM technology: a) horizontal arrangement - bad mapping of the ornament (partially unremoved elements of the supports); b) vertical arrangement - good mapping of the ornament details.

#### 4 Research problems of reconstruction

In the practical implementation of the idea of reconstructing archaeological artefacts a whole range of research problems need to be addressed, such as:

- optimisation of the scanning process,
- transition from a 3D point cloud to solid models of objects,
- matching of virtual elements of the artefact,
- restoration of the missing elements and creation of their solid models,
- development of the technological process of the real object's assembly.

The most important processes leading to the reproduction of the shape of the whole artefact and its missing parts are shown in Figure 4.



**Fig. 4.** Steps leading to the reproduction of the missing parts of the artefact.

#### 4.1 Optimising the scanning process

The number of points supplied in the 3D scanning process can be very large. The more points allow to obtain the more accurate the scanning. A large number of them causes specific problems in their storage and postprocessing. Increase in the number of points is matched by greater processing time and memory usage.

The optimisation process allows proper selection of the number of points to the mapped shape of the part. The method applied was that of multicriteria optimisation. The assumed optimisation criteria was the percentage of the number of points used to reconstruct the shape, the volume of the generated file (minimise) and the quality of the resulting image (maximise). The concept of Pareto optimality was employed. A solution is Pareto optimal if the value of any of the criteria  $F_1(x), F_2(x), \dots, F_j(x)$  cannot be improved without simultaneously worsening at least one of them while maintaining the constraints [13]. For the case of minimising all the components of the criteria vector: an element  $x^* \in X$  is called a Pareto optimal solution if and only if the set  $X$  has no element  $x^-$  such that for every  $j \in J$

$$F_j(x^-) \leq F_j(x^*) \text{ and there is } p \in J \text{ such that } F_p(x^-) > F_p(x^*) \quad (1)$$

where  $F_j$  is components of the criteria vector and  $x$  is vector of the design variables.

To reduce the number of alternatives under consideration the concept of optimality in the sense of an indistinguishable interval [14] was used. For the case criteria minimisation, the element  $x^\wedge \in \Omega$  will be optimal in the sense of an indistinguishable interval if and only if the set  $\Omega$  has no element  $x^+$  such that for each  $j$

$$\begin{aligned} \text{when } F_j(x^\wedge) \geq 0 \text{ if } F_j(x^\wedge) < F_j(x^+) \text{ is } (1 + PN_j)F_j(x^\wedge) > F_j(x^+) \\ \text{when } F_j(x^\wedge) < 0 \text{ if } F_j(x^\wedge) < F_j(x^+) \text{ is } (1 - PN_j)F_j(x^\wedge) > F_j(x^+) \end{aligned} \quad (2)$$

where  $\Omega$  is a non-empty set of non-dominated solutions and  $j = 1, 2, \dots, n$  is the index set of criteria.

The search tool in choosing the best solution was the evolutionary strategy of generating representative solutions. Applied for this purpose was the min-max method with weights, using the standard distance function with Chebyshev norm when  $r \rightarrow +\infty$ , which leads to the optimisation problem shown as a dependency (3)

$$p[F(x)] = \left\{ \omega_j \frac{|F_j^0(x) - F_j^k(x)|}{|F_j^0(x)|} \right\} \quad (3)$$

where:

$\omega_j$  - "weight" of the  $j$ -th assessment criterion,

$j = 1, 2, \dots, n$  - the index set of the criteria considered,

$k = 1, 2, \dots, p$  - the index set of the variants considered (non-dominated solutions),

$F_j^0(x)$  -  $j$ -th component of the ideal point [13],

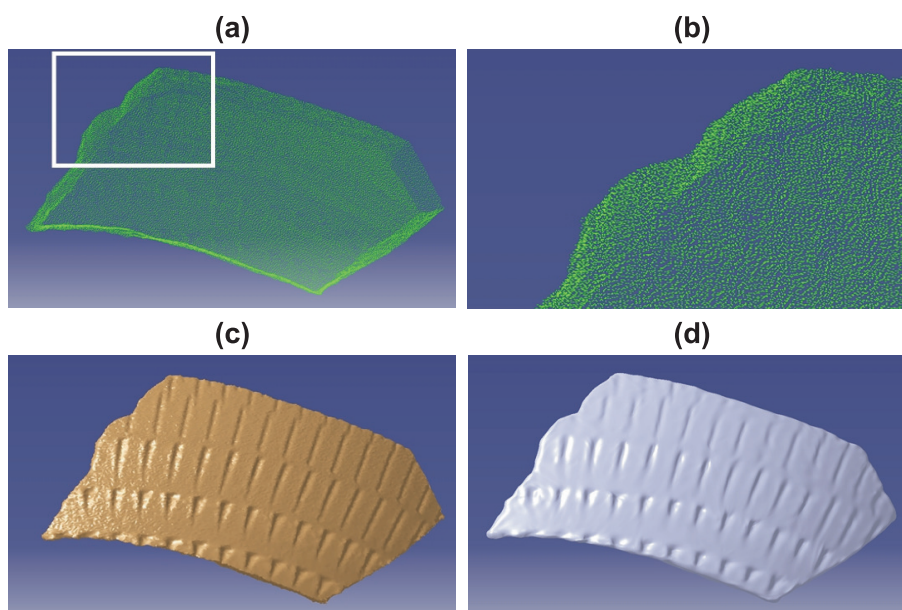
$F_j^k(x)$  -  $j$ -th component of the criteria vector of variant  $k$ .

Final studies made it possible to reduce the number of cloud points to 50%.

#### 4.2 Postprocessing

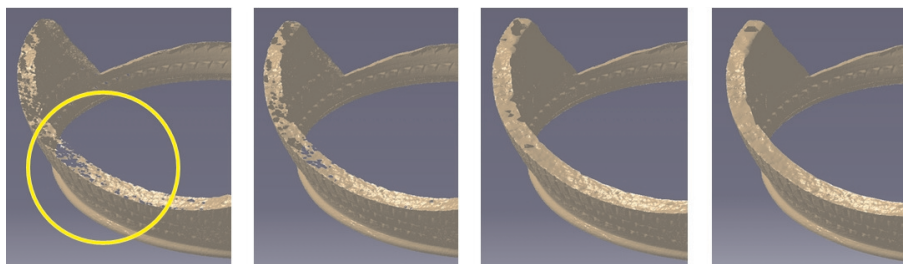
3D scanning provides a point cloud (Figure 5a) - the points on the surface of elements with known XYZ coordinates. On the basis of the point cloud a grid model is created in the process of

postprocessing - Figure 5c, then surface models and lastly a solid model - Figure 5d. A key activity for obtaining the shape of the object is the proper procedure of the process of triangulation. This process involves creating a grid of triangles with adjacent three points from the acquired cloud. The triangulation must take into account that these points have different depth components and reflect three-dimensional shapes. The algorithms used were based on the Delaunay method [15]. The images in Figure 5 show that this process properly reflects the concave elements of the ornaments, which means that it was performed correctly. To carry out the postprocessing, the CATIA and Meshlab programs were used, with one's own solutions introduced as needed.



**Fig. 5.** Steps for creating a 3D model of an object: a) point cloud; b) enlarged detail; c) grid model; d) solid model.

The scanning process does not always properly capture all the points of the digitised surface of the element, which in practice is revealed by the formation of discontinuities in the generated surfaces, popularly called holes. To obtain a solid body (water tight) they must be closed by using appropriate modifiers, selecting the appropriate parameters of their application. Figure 6 shows the progress in obtaining a solid body using MeshLab program.



**Fig. 6.** Steps of refilling surface discontinuities in the surface model.

The final stage of the treatment process of the vessel's components is saving their solid models in a format that allows transferring them to 3D printers.

### 4.3 Adjustment of the artefact's virtual elements

The process of matching elements to each artefact is implemented in the virtual world. It supports joining the obtained solid elements into one artefact using the information contained in the individual elements. The basic distinguishable features of an element are its shape, the curvature of its hyper-surface (Figure 7a), the position of ornament (Figure 7b), fingerprints on the inner surface (Figure 7c), and so on.

**Methodology of fit.** The procedure to be followed when adjusting existing elements can be represented as follows (example for utensils):

- orient the elements relative to the Z axis on the basis of the recognised features; fingerprints and ornaments are situated horizontally;
- determine the values of the horizontal curvature radii of all the elements (for large components three values are to be determined - in the middle and at both ends);
- segregate items into subgroups with similar values of curvature radii;
- move subgroups for elements with the largest diameters to be in the middle;
- look for various subgroups of elements with the largest diameter and the largest and smallest wall thicknesses;
- segregate the items according to thickness ratio; elements with the largest and smallest values can be adjacent;
- search for a subgroup of elements that creates the bottom of the dish;
- retrieve a subgroup of elements forming the spout of the dish (upper bowl);
- adjust items in the various subgroups, starting from the bottom and spout, two disjoint sets of elements are formed;
- match elements from other subgroups to objects already formed;
- match the two groups to each other, rotate about a vertical axis one sub-group of assembled elements, looking for complementary shapes in the second subgroup.

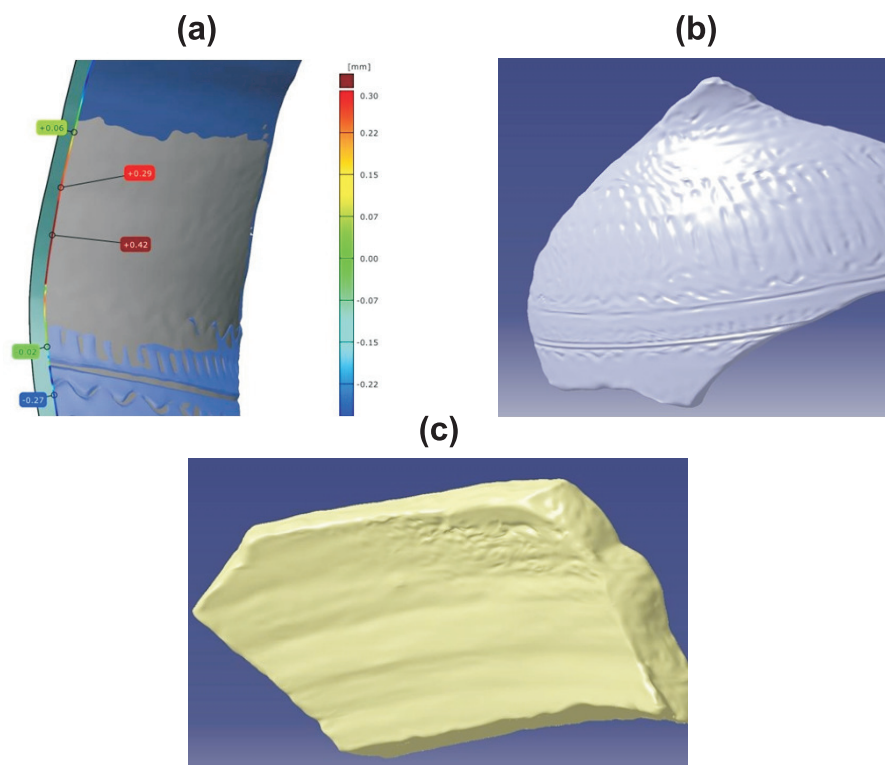
The above procedural methodology can be automated by introducing proprietary software. To this end, for each available element to be determined, one should map out all the edges of their outer surfaces. They are then replaced with broken lines at an arbitrarily assumed constant division value (based on preliminary studies carried out), the same for all elements. With the XYZ coordinates of all the vertices of the broken line thus created, a method of least squares is used, which minimises the sum of squared errors of fitting together two polylines or their fragments, as shown in Figure 8.

The errors calculated concern the distance between the corresponding vertices of two polylines describing two different elements.

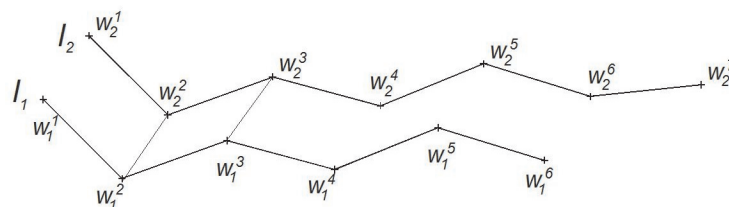
$$\chi^2(l_1, l_2) = \sum_{i=1}^n (w_i^1 - w_i^2)^2 \quad (4)$$

where  $w_i^1$  and  $w_i^2$  - respectively the i-th vertex of the broken line  $l_1$  and  $l_2$ .

Another approach regarding the matching process involves replacing sections of the components of the broken line with vectors.



**Fig. 7.** Examples of selected features on elements of vessels: a) different thicknesses of an element; b) ornament; c) fingerprints.



**Fig. 8.** Broken lines at the edge of two elements.



#### 4.4 Addendum - reconstruction of missing pieces

Reconstruction of the missing elements is based on algorithms that use:

- restored shape of the artefact of the basis of the surrounding structures (usually using Bezier curves).
- existing patterns of the artefact based on the already known shapes of objects from the era,
- manual manipulation of the virtual object by man-archaeologist (manual shaping of the surface, inserting the missing elements, drawings and other surface elements etc.)

Generated elements can be replicated by 3D printers and apply them in the process of assembling a complete real dish, using original elements or previously made copies of them. The artefact thus reconstructed must be evaluated by specialists i.e. archaeologists. With detailed knowledge of the material culture of the era from which the reconstructed object comes, they can bring knowledge that was not available from the analysis and reassessment of existing components. For example, the missing piece of the artefact may contain elements, the existence of which is not indicated by any of the existing ones. Thus, the virtual reconstruction of the vessel should be supplemented by this element, acquiring its shape from another artefact from that era. Its digitisation and postprocessing, sometimes also demanding a rescaling of the object to the size of the reconstructed vessel will also make it possible to obtain the relevant copy.

#### 4.5 The process of the technological installation of a real object

The full technological process of the artefact restoration can be developed using a 3D virtual model of the artefact. This process consists of the following phases:

- manufacture of parts missing (with the selected 3D printing technology and the process of machining of parts after they are printed).
- installation of artefact using the chosen technique (e.g. gluing the elements together).
- the final finish of the reconstituted artefact.

The development of the technological process takes into account the recommendations of restaurateurs and the mounting possibilities (including the order of connecting the parts).

### Conclusions

The development of 3D technology equipment and ever more perfect algorithms for processing virtual three-dimensional objects allowed to develop a support system for archaeological artefacts reconstruction. This system is a complex of hard- and software intended to implement the processes of three-dimensional digitisation of fragments of artefacts, reconstruction of the original shapes of objects and their missing parts in order to create the technological process of material reconstruction of historic artefacts.

The developed system uses the existing hardware components (scanners and printers), software (existing engineering graphics programs) and proprietary software, implementing successive phases of modelling and model transformation.

The system described will be expanded towards the creation of subsystems for the various layers of the artefact (shape and texture, including drawings and paintings), using different models, algorithms and technical measures.

## References

1. Sanders, D.: Why do virtual heritage? In: Clark J.T., Hagemester E.M. (Eds.): Digital Discovery: exploring new frontiers in human heritage. Proceedings from the 34th Computer Applications and Quantitative Methods in Archaeology conference, Fargo, ND, USA, April 2006, Budapest: Archaeolingua, pp. 427-436 (2006)
2. Nadel D., Filin S., Rosenberg D., Miller V.: Prehistoric bedrock features: recent advances in 3D characterization and geometrical analyses. *Journal of Archaeological Science*, 53, pp. 331-344 (2015)
3. Zhao W., Forte E., Tiziana Levi S., Pipan M., Tian G.: Improved high-resolution GPR imaging and characterization of prehistoric archaeological features by means of attribute analysis. *Journal of Archaeological Science*, 54, pp. 77-85 (2015)
4. Rua H., Alvito P.: Living the past: 3D models, virtual reality and game engines as tools for supporting archaeology and the reconstruction of cultural heritage e the case-study of the Roman villa of Casal de Freiria. *Journal of Archaeological Science*, 38, pp. 3296-3308 (2011)
5. Bruno F., Bruno S., De Sensi G., Luchi M-L., Mancusoc S., Muzzupappaa M.: From 3D reconstruction to virtual reality: A complete methodology for digital archaeological exhibition. *Journal of Cultural Heritage*, 11, pp. 42-49 (2010)
6. Zheng R., Zhang D., Yang G.: Protection of Cultural Heritage's Focus on 3D Technologies: A Survey. *Applied Mechanics and Materials*, 513-517, pp. 792-795 (2014)
7. Ruther H., Chazan M., Schroeder R., Neeser R., Held C., Walker S.J., Matmon A., Kolska Horwitz L.: Laser scanning for conservation and research of African cultural heritage sites: the case study of Wonderwerk Cave, South Africa. *Journal of Archaeological Science*, 36, pp. 1847-1856 (2009)
8. Haydar M., Roussel D., Maida M., Otmane S., Mallem M.: Virtual and augmented reality for cultural computing and heritage: a case study of virtual exploration of underwater archaeological sites. *Virtual Reality*, 15, pp. 311-327 (2011)
9. Adriaens A.: Non-destructive analysis and testing of museum objects: An overview of 5 years of research. *Spectrochimica Acta, Part B*, 60, pp. 1503-1516 (2005)
10. Pavlidis G., Koutsoudis A., Arnaoutoglou F., Tsiouka V., Chamzas C.: Methods for 3D digitization of Cultural Heritage. *Journal of Cultural Heritage*, 8, pp. 93-98 (2007)
11. Skarbek K., Kowalski P.: Building the Models of Cultural Heritage Objects Using Multiple 3D Scanners. *Theoretical and Applied Informatics*, 21(2), pp. 115-129 (2009)
12. Montusiewicz J., Czyż Z., Kayumov R.: Selected methods of making three-dimensional virtual models of museum ceramic objects. *Applied Computer Science*, 11(1), pp. 1-16 (2015)
13. Pareto V.: *Cours d'economic politique*. Rouge, Lousanne (1896)
14. Montusiewicz J.: Ranking Pareto optimal solutions in genetic algorithm by using the undifferentiation interval method. [in:] Burczyński T. (ed.) *Evolutionary Methods in Mechanics*. Kluwer Academic Publishers, pp. 265-276 (2004)
15. Miller G.L., Talmor D., Teng S.-H., Walkington N.: A Delaunay based numerical method for three dimensions: Generation, formulation, and partition. *Proceedings of the Twenty-Seventh Annual ACM Symposium on the Theory of Computing, Las Vegas*, pp. 683-692 (1995)

# Using GIM-Technologies for Monitoring of the Ionosphere Over Kazakhstan Region

S.N. Mukasheva<sup>1</sup>, N.S. Toyshiev<sup>1</sup>, B.K. Kurmanov<sup>1</sup>, G. Sharipova<sup>1</sup>, D.E. Karmenova<sup>2</sup>

<sup>1</sup>Institute of Ionosphere, National Center for Space Research and Technology, Almaty, Kazakhstan,

<sup>2</sup> Al-Farabi Kazakh National University, Almaty, Kazakhstan

admion1@mail.ru

**Abstract.** Modern information technologies in science received a new impetus to the development and use of the spacecraft. Radio sounding of the ionosphere by signals of the global GPS navigation system now allows the continuous monitoring of the Earth's ionosphere. The so-called GIM technology (Global Ionospheric Maps), which was developed by some research centers, is a powerful tool for monitoring and investigation of global and local structure of ionosphere (Afraimovich and Perevalova, 2006). This geoinformation technology allows obtaining qualitatively new information about the state of the ionosphere and is one of the most reliable means for monitoring the ionosphere during the disturbances. This paper shows the application of GIM-technologies for monitoring the ionosphere over Kazakhstan region. We use GIM-maps, designed by the Swiss center CODE (Center for Orbit Determination in Europe, University of Berne, Switzerland) using data from more than 150 GPS sites (list of stations is given in IONEX file) that contains the UNIX format online ftp: [//cddis.gsfc.nasa.gov/pub/gps/products/ionex](http://cddis.gsfc.nasa.gov/pub/gps/products/ionex). This paper gives a brief description of the calculation of the absolute values of the total electron content for any regions of Kazakhstan using IONEX files. The results of the analysis of the total electron content variations at different helio-geophysical conditions are shown.

**Keywords:** ionosphere, total electron content, navigation spacecraft, Global Ionospheric Maps.

## 1 Introduction

On Kazakhstan territory, which is the 9th largest country in the world in terms of area (2 724.9 thousand sq. Km), there functions only one station of vertical sounding (Almaty [43.38°N; 77.38°E]), whose data do not provide an ionospheric prediction for the the whole country. Vertical sounding method enables us to study the ionosphere to the heights of 250-300 km, up to altitudes of the main maximum of the ionosphere, hmF2. Radiophysical experiments on receiving signals from geostationary and the orbital Artificial Earth Satellites (AES) provide a unique opportunity not only to expand the range of heights of the observation of the ionosphere up to 2000 km, but also to expand the range of tasks in terms of studying the physics of the ionosphere. The undiminishing interest of the world community to expansion and all-round development of the satellite communication foreground ionospheric researches with use of the method of radio-sounding from satellites to the rank of urgent tasks, by putting forward demands of studying regional characteristics of the ionosphere. The development of the Global Positioning System (GPS) allows, in addition to solving navigation tasks, remote diagnostics of the ionosphere. Radio-sounding of the ionosphere by using signals of the global GPS navigation system now allows continuous monitoring of the Earth's ionosphere (Mannucci et al., 1998; Afraimovich and Perevalova, 2006; Yasukovich et al, 2010; Yasyukevich et al, 2010; Ashkaliev et al, 2012; Afraimovich, et al., 2013; Zolesi and Cander, 2014). The method of determining the values of total electron content, TEC, by measurement of increments of the phase and group path of the transionospheric radio signal is described in detail and justified in the series of publications (eg. Afraimovich and Perevalova, 2006). One of the key parameters of the ionosphere, the total

electronic content of I (t), is calculated from measurements of increments of phase and group of the transionospheric radio path according to the following formula:

$$I = \frac{1}{40.308} \cdot \frac{f_1^2 f_2^2}{f_1^2 - f_2^2} [L_1 \lambda_1 - L_2 \lambda_2 + const + nL,]$$

$L_1 \lambda_1$  and  $L_2 \lambda_2$  - are the phase paths of the radio signal, caused by the signal delay in the ionosphere (m);  $L_1$  and  $L_2$  - stand for the number of the phase rotations of the radio signal obtained during its propagation; and  $\lambda_1$  and  $\lambda_2$  - wavelength (m) for frequencies  $f_1$  and  $f_2$ ; const - an unknown initial phase distance (m);  $nL$  - error in determining the phase path (m). Phase measurement in the GPS system are made with a high degree of accuracy, at which an error in determining the TEC at 30 second averaging intervals is less than 1014 electrons / m<sup>2</sup>, although the initial value of TEC remains unclear, which allows the detection of inhomogeneity of ionization and wave processes in the ionosphere in a wide range of amplitudes (up to 10-4 by diurnal TEC changes) and periods (from days to 5 minutes). TECU (total electron content unit) is the universally accepted measure of the total electron content, eq. 1016 el/m<sup>2</sup>. The GIM (Global Ionospheric Maps) technology, which was developed on the basis of the interpolation TEC data, obtained from the global network of GPS receivers, in several international research centers, is a powerful modern instrument for monitoring and studying global and local structures of the ionosphere (Afraimovich and Perevalova, 2006). With help of GIM-technology daily and seasonal features of Earth's plasma shell - ionosphere condition can be studied, it can also be explored how external factors, such as variations in solar activity, affect the dynamic structure of the Earth, in particular, the near-Earth space (Schaer et al., 1998a; Schaer et al., 1998b; Polyakova, et al., 2009; Zhivetev, 2009; Perevalova et al., 2010; Afraimovich et al., 2013; Huang et al., 2014; Zolesi and Cander, 2014). The purpose of this work is to show the possibility of using GIM-technologies for monitoring the ionosphere over Kazakhstan region.

## 2 Methodical recommendations for working with the global map of the total electron content

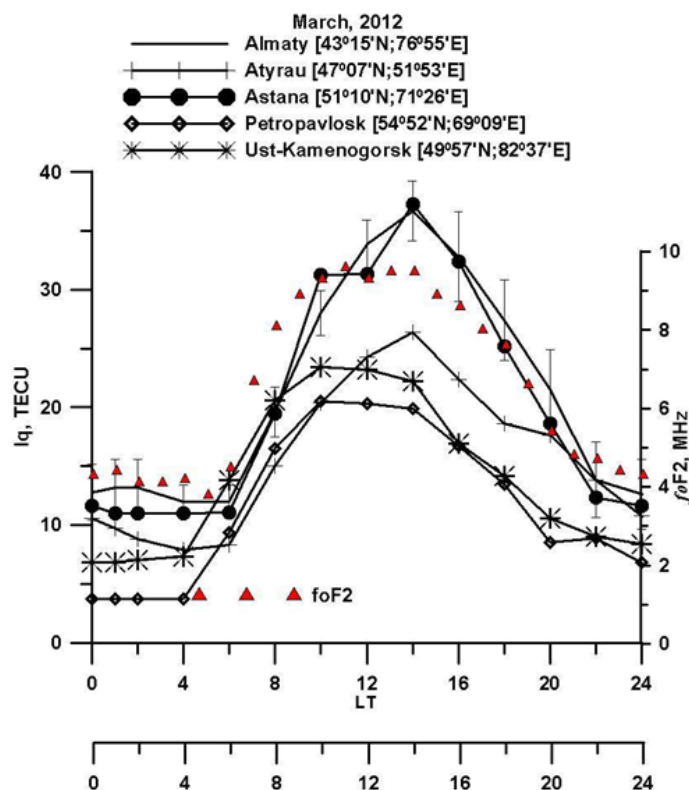
GIM technology provides development of global maps of the absolute vertical values of the total electron content by interpolating the data received by the global network of the global navigation satellite system receivers. For storage and transfer of TEC maps there was digitally developed a special standard format IONEX (standard of the files Technology Global Ionospheric Maps). Each IONEX file contains the world map with global distribution of the absolute vertical TEC and the corresponding map with errors of calculation of TEC per day on a scale of universal time UT with a time resolution of 2 hours, which corresponds to longitude-latitude grid with a resolution of 5 ° in longitude and 2.5 ° by latitude. The size of a GIM-cell by latitude is 279 km and the longitude - 436 km. The vertical total electron content is calculated with taking into account the state of the solar-geomagnetic conditions on the spherical harmonic formulas. For each time point from the IONEX-files there can be read off the values of the vertical total electron content in 5183 cells in TECU units (Afraimovich and Perevalova, 2006). The TEC maps in the IONEX format can be found on the website <ftp://cddisa.gsfc.nasa.gov/pub/gps/products/ionex>. The ftp-server directory is organized as follows: the data stored in the folders chronologically, the information inside folders is sorted by days, where three characters of DDD format mean index number of days in a year, within the folder there is archived data from the international research centers on a two-hour TEC cards. Global GIM-maps are calculated according to the international network of GPS receivers at various scientific centers: 1) Geodetic Survey Division of Natural Resources Canada (EMRG) (<http://www.nrcan-rncan.gc.ca>); 2) Center for Orbit Determination in

Europe, University of Berne, Switzerland (CODG) (<http://www.cx.unibe.ch>); 3) Jet Propulsion Laboratory of California Institute of Technology (JPLG) (<http://www.jpl.nasa.gov>); 4) Grup Universitat Politecnica de Catalunya (UPCG) (<http://www.upc.es/>); 5) European Space Agency Group (ESAG) (<http://www.esa.int/ESA>) and others. In the paper there were used the GIM-maps, which were calculated by Swiss research center CODG (Center for Orbit Determination in Europe, University of Berne, Switzerland) based on data from more than 150 GPS signal reception centers. The GIM-maps are available online in the IONEX format (<ftp://cddis.gsfc.nasa.gov/pub/gps/products/ionex>). From the archived data, the files `codgddd0yyi.Z` was selected, where `codg` is a four-character code for the center, which created the file; number `ddd` means the number of the day; `yy` - the last two digits of the year. For example, the file `codg1720.14i` contains two-hour TEC values which were calculated by the Swiss center CODG for June 21, 2014. The IONEX-file format is described in detail in many papers, but we recommend using the monograph Afraimovich and Perevalova, 2006 for further instructions. The IONEX-file is a text file with fixed record length of 80 symbols. In free access there are a few programs for work with the IONEX-files which give an opportunity to collect, process and use the GPS-monitoring measurements of Earth's upper atmosphere at minimum cost. Guidelines for working with global maps of total electron content are given in Ashkaliev et al, 2012. By GIM the absolute values of the vertical TEC can be restored for a GIM-node, whose coordinates are the closest to the geographic coordinates of our points of interest. So for the following cities of Kazakhstan the absolute values of the vertical TEC were calculated in the most closely spaced GIM-nodes: Almaty [43°15 N; 76°55 E] - [42,5°N; 75°E], South Kazakhstan; Atyrau [47°07 N; 51°53 E] - [47,0°N; 50°E], Western Kazakhstan; Astana [51°10 N; 71°26 E] - [50°N; 70°E], Central Kazakhstan; Petropavlovsk [54°52 N; 69°09 E] - [55,0°N; 70°E], Northern Kazakhstan; Ust-Kamenogorsk [49°57 N; 82°37 E] - [50°N; 82°E], East Kazakhstan. For example, calculations were made for March 2012, when there was an anomalously low solar activity maximum. The period of the vernal equinox - March 2012, was selected because in equinoctial periods, the diurnal TEC variations reach their maximum, which allows to study TEC variations particularities for different regions of Kazakhstan. For most geomagnetic quiet conditions, with  $K_p$  not more than 3, of the spring equinox (March, 2012), we have restored the TEC diurnal variation for Almaty (GIM node [42.5°N; 75°E]) from the IONEX maps. The average diurnal variation of TEC,  $I_q$ , for 10 geomagnetic quiet days, were shown with a solid black line in the Figure 1.

The vertical segments show squared average, calculated by using:

$$\sigma = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n (I_i - I_q)^2}$$

where  $I_q$  - TEC value, averaged for chosen geomagnetic quiet periods,  $I_i$  - current TEC value. Squared average of current  $I_i$  from average  $I_q$  peaks in March, 2012, when their values equals to about (5-3) TECU (Fig. 1). The red triangles show the diurnal variations of the critical frequency of the F2 ionosphere layer, which were averaged for the same periods, as the TEC variations (Fig. 1). In diurnal variations of critical frequencies the same peculiarities as in TEC variations were observed, because main contribution to TEC is made from the part of ionosphere, which is located about maximum of ionization. Figure 1 shows the reduced value of the TEC, 21 March, 2012: Atyrau (GIM node [47,0°N; 50°E], Western Kazakhstan) - a curve with cross; Astana (GIM node [50°N; 70°E], Central Kazakhstan) - a curve with circles; Petropavlovsk (GIM node [55,0°N; 70°E], Northern Kazakhstan) - curve with rhombuses; Ust-Kamenogorsk (GIM node [50°N; 82°E], East Kazakhstan) - curve with asterisks. Diurnal variation of TEC for March 2012 for different regions of Kazakhstan are qualitatively similar: an increase of the electron density begins after sunrise (04-05 LT) and extends continuously to the local noon (12-13 LT), followed by the lowering of



**Fig. 1.** The March, 2012 diurnal variations of the total electron content  $I_q$  in Almaty, Atyrau, Astana, Petropavlovsk and Ust-Kamenogorsk, which were calculated with use of GIM-technology and averaged for magnetic calm days. The vertical lines indicate the mean square deviations. The red triangles show the diurnal variations of the critical frequency of the F2 ionosphere layer, which were averaged for the same periods, as the TEC variations.

the electron density until midnight. After midnight until sunrise electron concentration is kept at the same level - squared average of current  $I_i$  from average  $I_q$  when their values equals to about (3 2) TECU. That does not contradict to the existing ideas about the TEC variations at middle latitudes of the Northern Hemisphere.

### 3 Conclusion

Thus, by using IONEX maps, there can be built the diurnal course of the absolute vertical electron content, as well as distribution maps of the total electron content of the region for the territory of Kazakhstan. The quality of the TEC maps, constructed with using GIM-technology, was tested by comparing with the numerous measurements by using data from other spacecraft and the ionosphere model IRI. The comparison showed good agreement between the results (Mannucci et al., 1998; Afraimovich and Perevalova, 2006; Perevalova et al., 2010; Zolesi and Cander, 2014). While using GIM-technology there should be borne in mind that data of the IONEX-file is interpolated from measurements of a network of GPS receivers. The TEC values, which are shown on the maps, are reliable for regions with a large number of GPS receivers and not very reliable for the regions with a little number of GPS receivers. This paper was written as part of a grant project No. 0079/GF4.

## References

1. Afraimovich E.L., Perevalova N.P., 2006. GPS-monitoring of the Earth's upper atmosphere/- Irkutsk: SC RRS SB RAMS.-480 p. (in Russian).
2. Afraimovich, E., Astafyeva, E., Demyanov, V., Edemsky, I., Gavrilyuk, N. et al., 2013. A review of GPS/GLONASS studies of the ionospheric response to natural and anthropogenic processes and phenomena. *J. Space Weather and Space Climate*, 3, A27, doi:10.1051/swsc2013049.
3. Ashkaliev J.F., Bibosinov A.Z., Breusov N.G., Zhumabaev B.T., Kurmanov B.K., Mukasheva S.N., Nurgaliyeva K.E., Sadykov K.A., 2012. Preparation of Global Navigation Satellite System Gps Data for Identification Of Seismic-Ionosphere Effects (Workbook).- Almaty: Gylym, -2012. -43p. (in Russian).
4. Mannucci A.J., Wilson B.D., Yuan D.N., Ho C.M., Lindqwister U.J., Runge T.F., 1998. A global mapping technique for GPS derived ionosphere TEC measurements // *Radio Sci.*-V. 33, N 3. -P. 565-582.
5. Perevalova, N.P., Polyakova, A.S., Zalizovsky, A.V., 2010. Diurnal variations of the total electron content under quiet helio-geomagnetic conditions. *J. Atm. Sol.-Terr. Phys.* 72(13). 997-1007.
6. Polyakova A.S., 2009. Investigation into diurnal variation of "vertical" TEC under quiet geomagnetic conditions//BSFF-2009. Section A. Physics of near-Earth space.-P.174-177. . (in Russian). Polyakova A.S., Perevalova N.P., 2012. Diurnal variations in the total electron content in the East Siberian region in August 2009 // *Probing Earth's surface synthetic aperture radar*. Institute of Solar-Terrestrial Physics SD RAS, Irkutsk, Russia. -S. 259-268. (in Russian).
7. Schaer S., Beutler G., Rothacher M., 1998a. Mapping and predicting the ionosphere // *Proceedings of the IGS AC Workshop*. Darmstadt, Germany. February 9-11.- P. 307-320.
8. Schaer S., Gurtner W., Feltens J., 1998b. IONEX: The Ionosphere Map Exchange Format Version1// *Proceedings of the IGS AC Workshop*. Darmstadt, Germany. February 9-11- P. 233-247.
9. Yasyukevich Y.V., Zhivetev I.V., Lukhnev A.V., 2010. Regional electron content in the Baikal rift zone// *Electronic collection of reports of the Russian conference «Sounding Earth's surface synthetic aperture radar Section 2 "Radiofizicheskie diagnostic methods environment."*Russian Federation, the Republic of Buryatia, Ulan-Ude, 6-10.09.2010. -P. 195-205 // <http://jre.cplire.ru/alt/library/Ulan-Ude-2010/conf.pdf>. (in Russian).
10. Zhivetev I.V., 2007. Ionospheric disturbances at different phases of the 23rd solar cycle, according to a global network of GPS: Author. ... *Cand. agricultural Sciences: 25.00.29*. - Irkutsk: Nauka,- 22p. (in Russian).
11. Zolesi B, Cander LjR., 2014. *Ionospheric Prediction and Forecasting* Springer-Verlag Berlin Heidelberg. Library of Congress Control Number: 2013942473. ISBN 978-3-642-38429-5 ISBN 978-3-642-38430-1 (eBook): DOI 10.1007/978-3-642-38430-1.

# Development of the Kazakh Text-to-Speech Synthesis System on The Basis of Fujisaki Intonation Model

Rustam Mussabayev

Institute of Informational and Computational Technologies, Laboratory of Analysis and Modeling of Informational Processes,  
Pushkin st. 125, Almaty, Kazakhstan  
rustam@ipic.kz  
<http://www.ipic.kz>

**Abstract.** In the given article the task of the Kazakh speech synthesis with different intonation characteristics is considered. For its solution the methods of semantic analysis of Kazakh language texts have been used. Using the obtained semantic information the optimal values of the Fujisaki model parameter set have been selected. On the basis of this model it is possible to modeling the intonation contour of the synthesized speech signal with good natural characteristics.

**Keywords:** semantic analysis, text-to-speech synthesis, Fujisaki model, intonation contour.

## 1 Introduction

One of the problems of modern text-to-speech synthesis systems is the low level of capacity for the synthesis of emotional (expressive) speech [6]. This fact greatly reduces the scope of existing speech synthesizers. The synthesized speech perceived by human as a natural only if it has some intonation and emotional coloring depending on the semantic content of the pronounced textual information. As intonation and emotional variability is very important to eliminate the effect of fatigue during the listening to a long monotonous speech (for example, at the time of listening to synthesized audio books). This is due to the fact that a person psychologically perceives the intonationally and emotionally colored speech is much better than the monotone pronounced and emotionally neutral. The quality of sound, in turn, depends on the ability of the electronic announcer to give a speech with a different tone depending on the text content. It is very significant improvement in the quality of making emotional and expressive speech in the right color depending on the context moments. Intonationally and emotionally variable speech is apprehended much better then monotone narration and fatigue when listening to such speech is significantly less.

For the qualitative solving of the emotional text-to-speech synthesis task the development of the new algorithms for semantic analysis of the text information is required, which should provide the highest possible degree of extraction of the informative features from the text data for automatic decision making in the process of emotional speech synthesis. With regard to the Kazakh language the development of these types of algorithms is necessary for increasing of the quality of Kazakh speech technologies.

The development process of the automatic Kazakh text-to-speech synthesis system includes selection of the intonational (prosody) synthesis model, which allows implementing Kazakh speech intonational process modeling in a wide dynamic range including various intonation undertones. With regards to the above, the main selection criteria were based on proximity of model intonation contour to natural speech contour. The task of intonational synthesis has been divided into several subtasks: selecting intonational process models, software implementation of the selected model, parametric identification of the model for voice signal, development of



the algorithm for automatic specification of model parameters for arbitrary Kazakh texts to be synthesized.

The previous scientific research in this field is mainly focused on the following issues:

- The formation of speech corpora and databases of emotional speech [7];
- Study of acoustic and prosodic features of emotional speech [8,9,10,11,12];
- Automatic emotional speech synthesis [13,14,15,6];
- Study of intonational properties of Kazakh speech [16,17].

Currently on the Internet in the open source code the following projects are available:

- Emofilt [18] – the software for modeling of emotional expression by speech synthesis. For the implementation of the synthesis process the MBROLA [19] software synthesis module is used. Formally it is not the complete and full-featured synthesizer. This is a tool that allowing to carry out the manipulation of melody and rhythm of the synthesized phrases using the special phonetic descriptions.
- OpenMary [20] - Multilingual emotional text-to-speech synthesis system that is available in the following languages: German, English and Tibetan. This system is capable to synthesize the emotionally expressive speech both on the basis of the compilative diphone synthesis and on the use of unit-selection synthesis method. The conducted analysis of the works in the field of emotional synthesis and classification showed that this direction is progressive, and requires additional efforts of researchers and developers to produce better results, as well as for the development of this sector in other countries for their national languages.

The first stage included investigation of several models of intonational processes, which have been successfully applied for different languages: AM-model (autosegmental-metrical) [1], Fujisaki model [2], Tilt-model [3], INTSINT model [4] and others. After investigation of these models, it was decided to use the Fujisaki model because it has been successfully used in speech synthesis systems for a variety of languages, and specifically for Turkish language. Fujisaki model is often described as that most accurately reflecting biological aspects of speech formation and reproducing articulatory mechanisms. The second stage included programming of this model in Delphi programming environment. The third phase included parametric model identification for voice signals in Kazakh language. A speech signal in Kazakh language was supplied to the system inputs. Intonational contour was extracted from this signal. The challenge was to carry out optimized selection of model parameters in order to ensure that the model contour is in accordance with the real speech contour. Ultimately, in the process of identifying of parameters using the technique proposed in [5] the adequately acceptable results were obtained with the inaccuracy of 3%. In order to properly address the challenge at hand, a unified language of phonetic representations has been designed [3], which allows defining and setting out various phonetic and intonation forms of speech. All the original data pertaining to models are described using the unified language representation that allows flexible inter-system interaction.

## 2 Extracting of semantic parameters of the input text data

In the Laboratory of Analysis and Modeling of Informational Processes the issue of constructing of an intonation model of Kazakh language and subsequent algorithmic realization of full-featured Kazakh speech synthesizer is carried out. The purpose of this project is to develop Kazakh speech synthesis system, which will be presented in a form of complete software product. The algorithm for identifying of model parameters for arbitrary synthesized Kazakh text was developed. In order to address this challenge, it has been broken into several stages:

- Developing of algorithm for text to syntagms dividing.
- Accentual words detection inside syntagms;
- Automatic specifying of phoneme sounding durations;
- Identifying of Fujisaki model parameters according to [4].

The developed algorithm of text into syntagms dividing was broken into the several sub-algorithms:

- Text into sentences dividing
- Sentences into punctuative syntagms dividing;
- Punctuative syntagms into lexical syntagms dividing.
- Detection of accentual words inside lexical syntagms.

The process of dividing syntagmatic text in Kazakh language lacks proper research. Corresponding research activities are carried out to automate this process as part of the current project.

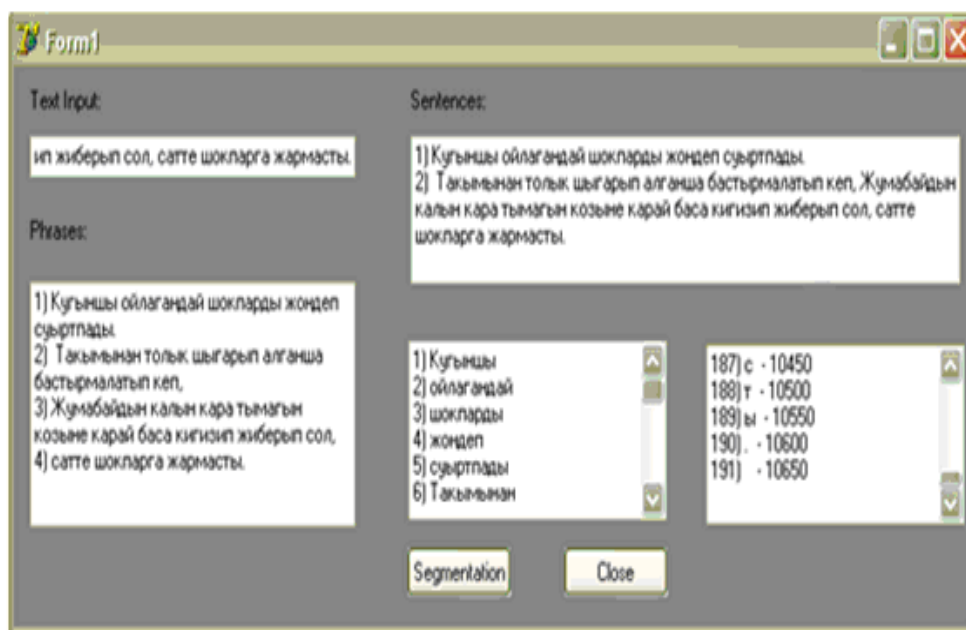


Fig. 1. Preliminary results of length placements and dividing punctuative syntagms.

Fig. 1 shows graphical user interface of a system being developed for the purpose of dividing Kazakh syntagmatic text. Additionally, in order to address challenges associated with high-quality intonation synthesis, construction of respective Kazakh language intonation database is planned, which shall be used as a base for calculating model parameters through statistical techniques.

### 3 The solution of the text-to-speech synthesis task

The most important qualitative indicator of synthesized speech is represented by its natural sounding properties and intelligibility. Speech naturalness is what determines the proximity of synthesized speech to the natural one. Furthermore, speech intelligibility is what determines the

degree of perception of synthesized speech by humans. Synthesizer must have both high degree of naturalness and intelligibility [22]. The majority of state-of-the-art speech synthesis systems are based on usage of the following models:

- Formant synthesis [23].
- Concatenative synthesis [24]:
  - Phoneme synthesis.
  - Diphone synthesis.
  - Dynamic selection of the most appropriate synthesis unit (unit-selection synthesis) [25].
- Synthesis based on vocal apparatus models [26].

Each of these technologies has both advantages and disadvantages. The majority of state-of-the-art high quality speech synthesis systems for English and Japanese languages are constructed on concatenative synthesis process. In addition, Unit Selection process is utilized as one of the varieties of concatenative synthesis. It is these two approaches that have been selected as a base for implementing Kazakh speech synthesis system. Using unit-selection approach prior to synthesis, the database searches for vocal fragments, which is followed by extracting the majority of the most suitable fragments necessary for synthesis purposes. Moreover, fragments derived from the database, may have different dimensions (half-phoneme/semi-phoneme, diphones, phonemes, syllables, words, and even entire sentences). Database comprises of many attributes, which correspond to each fragment, thus being characterized by these attributes:

- Length
- Location.
- Various prosodic characteristics
- Attributes of adjacent phonemes, etc.

In addition, the number of variations of signal being synthesized is limited to the number of and variance of fragment characteristics, which are stored in the database. Basic sections of speech signal are obtained by way of fragmenting lengthy recording of speaker speech reading specified text. Recorded voice signal is consecutively divided into elements in a manner reduce their dimensions: To start with, individual sentences are highlighted, and then phrases, words and syllables, etc. are highlighted. Each of the highlighted fragment is classified, many of its attributes are determined and all this information is therefore written into the database. Such process of constructing a database is very time consuming. For this purposes, specialized software commercially named STS (Speech Transcription System) has been developed as part of this Project. This software allows implementing a multi-level mark-up of speech signal followed by formation of marked-up vocal corpus. This software product has originally been developed as a language-independent system with flexible internal settings. This system fully supports Unicode encoding that allows to be used with the majority of current languages including Kazakh language. The current user interface is designed in English language. Further features to be incorporated will include shifting between interface languages to Kazakh and Russian languages, and this shall have a positive effect on commercialization phase. This software product will be of interest to the following categories of users: Speech synthesis and recognition system developers, language science researchers, law enforcement authorities, audio forensic experts, national security officers and other. The main distinction of this software product from others: Unlimited number of speech signal mark-up levels, setting of any number and types of regions to be marked-up, marked-up regions properties editor, built-in capability to create directory of meanings and other. Main window view of STS software is shown in Fig. 2.

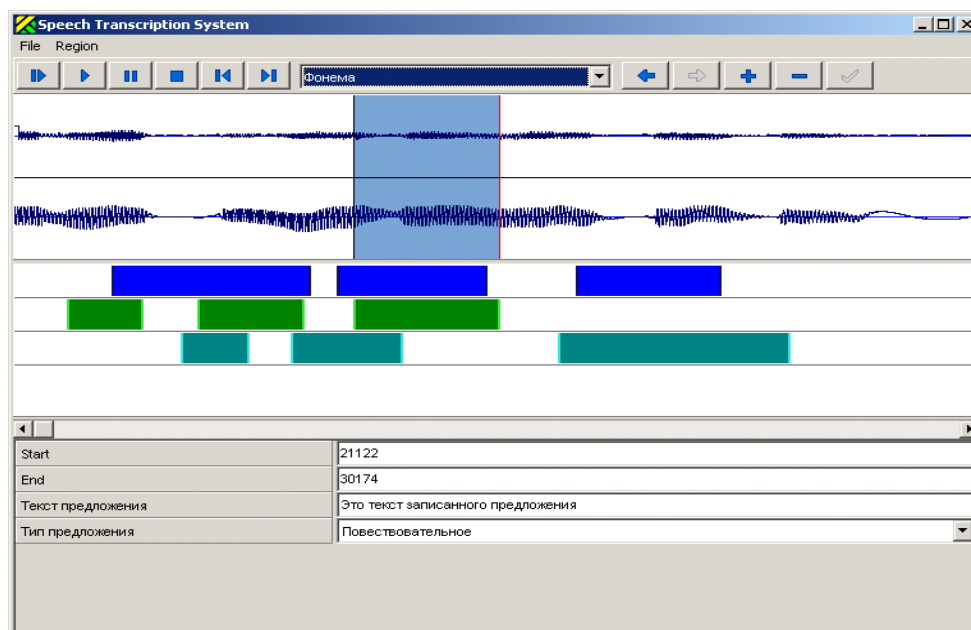


Fig. 2. Main window view of STS software (Speech Transcription System)

Fig. 3 shows the view of editors for directory of region types and properties in STS software. Thus, an intermediary product has been obtained during the phase of Kazakh speech synthesis system development. STS software product will be used in a proactive manner at various stages of works creating Kazakh vocal corpus for the purpose of developing multilingual speech synthesis and recognition software.

#### 4 Model selecting for synthesis of Kazakh speech intonation processes

This phase of project work includes selection of intonation synthesis model, which allows implementing Kazakh speech intonation process modeling in a wide dynamic range including various intonation undertones. With regards to the above, the main selection criterion was based on proximity of intonation contour model to natural voice contour. Objective in respect of intonational synthesis has been divided into multiple subtasks:

- Selecting intonational process models;
- Software implementation of model selected;
- Parametric model identification for voice signal;
- Developing algorithm for setting parameters of model for random Kazakh language text to be synthesized.

The first stage included review of several models of intonation processes, which have been successfully applied for different languages. Among these are AM-model (autosegmental-metrical) [1], Fujisaki model [2], Tilt-model [3], INTSINT model [4] and others. Upon review of these models, it was decided to use Fujisaki model because it has been successfully used in speech synthesis systems for a variety of languages, and specifically for Turkish language. Fujisaki model is often described as that most accurately reflecting biological aspects of speech formation and reproducing articulatory mechanisms. The second stage included programming of this model in Delphi programming environment. Mathematical model description was set by the following equations:

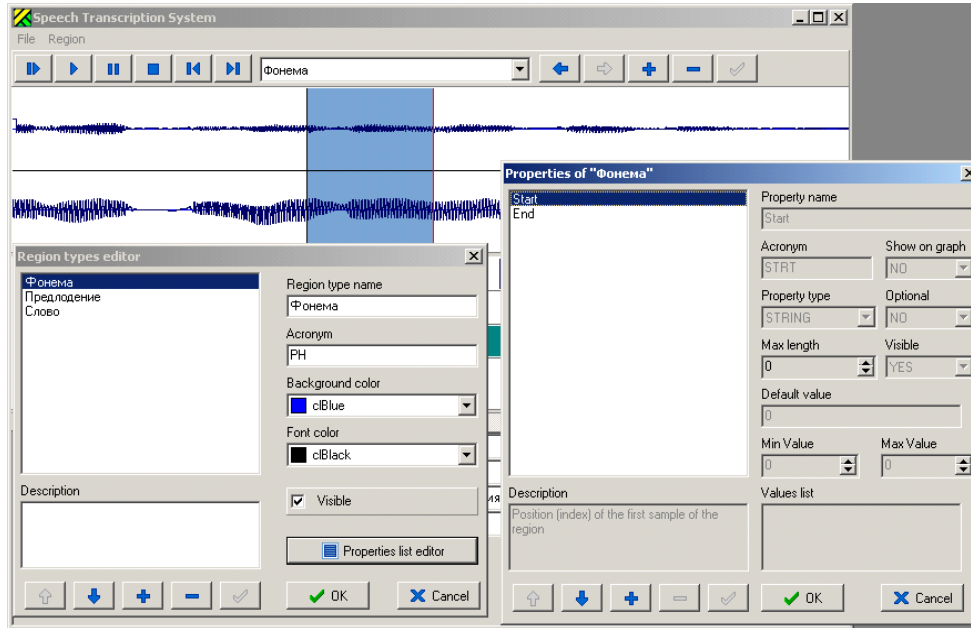


Fig. 3. View of editors for directory of region types and properties in STS software.

$$\ln F_0 = \ln F_b + \sum_{k=1}^{N_p} A_{p,k} G_p(t - T_{0,k}) + \sum_{k=1}^{N_a} A_{a,k} [G_a(t - T_{1,k}) - G_a(t - T_{2,k})],$$

where

$$G_p(t) = \begin{cases} \alpha^2 t e^{-\alpha t}, & t \geq 0 \\ 0, & t < 0 \end{cases}$$

and

$$G_a(t) = \begin{cases} \min((1 - (1 + \beta t)e^{-\beta t}), \theta), & t \geq 0 \\ 0, & t < 0 \end{cases}$$

where

$F_b$  – minimum frequency / under frequency,  $N_p$  – number of phrasal component/entry,  $N_a$  – number of accentual component/entry,  $A_{p,k}$  – magnitude of  $k$  phrasal component/entry,  $A_{a,k}$  – magnitude of  $k$  accentual component/entry,  $T_{0,k}$  – time of  $k$  phrasal component/entry,  $T_{1,k}$  – start of  $k$  accentual component/entry,  $T_{2,k}$  – end of  $k$  accentual component/entry,  $\alpha_i$  – angular frequency of  $i$  component phrasal command control mechanism,  $\beta_j$  – angular frequency of  $j$  component accentual command control mechanism,  $\theta$  – parameter indicating peak level of accentual component.

Fig. 4 shows graphical interface of software implemented Fujisaki model.

The third phase included parametric model identification for voice signal in Kazakh language. A speech signal in Kazakh language was supplied to the system inputs. Intonation contour was extracted from this signal. The challenge was to carry out optimized selection of model parameters in order to ensure that the model contour is in accordance with the real speech contour. Ultimately, while selecting settings using the technique proposed in [5], adequately acceptable product was derived resulting in an inaccuracy level/tolerance limit of 3%. Fig. 5 shows the results of Fujisaki model parametric identification.

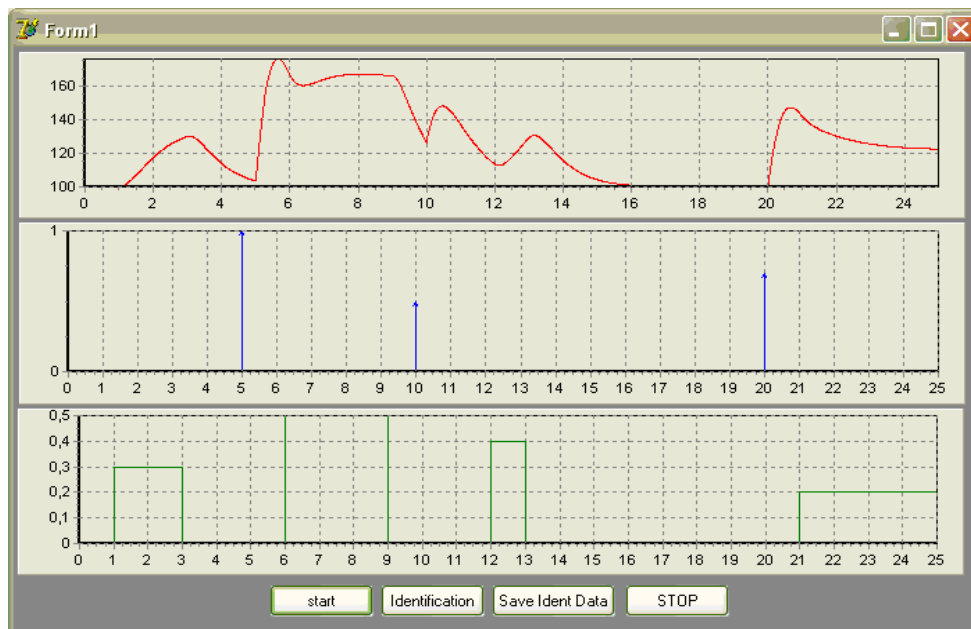


Fig. 4. Graphical interface of software implemented Fujisaki model

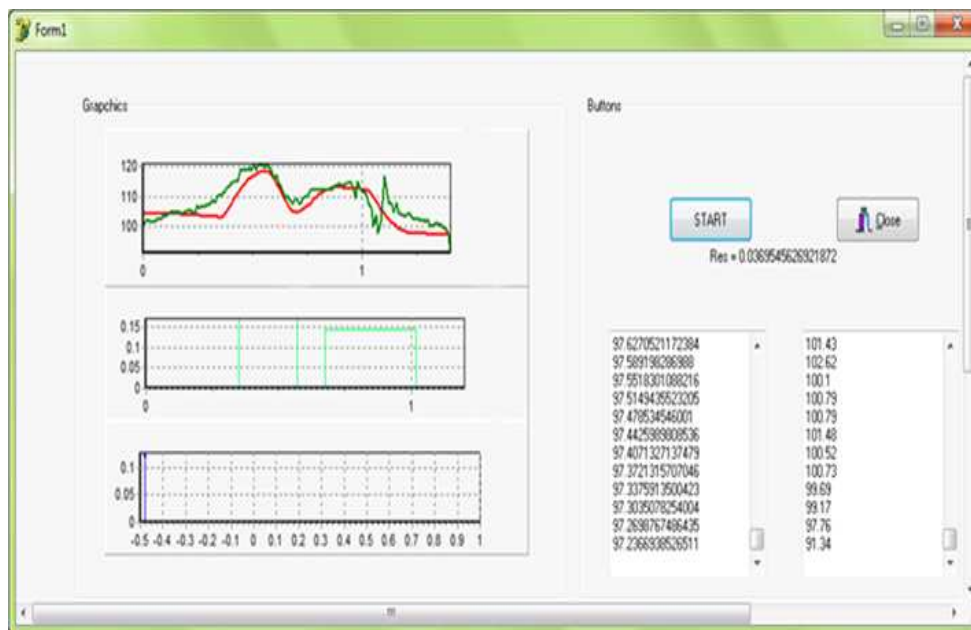


Fig. 5. Fujisaki model parametric identification (Green graph - real intonational contour, red graph - modeled contour)

## 5 Conclusion

Developed mathematical techniques and models allow implementing a wide range of modifications to intonation features of reference voice signal sets based on multiple adjustable parameters. Software implementation of these techniques and models proved adequate efficiency and reliability of these techniques respectively. Additionally, the synthesized product is of high quality performance indicators. All major research and development activities have been conducted with the use of Kazakh language as an example. For the purpose of automating synthesis processes, the corresponding phonetic-acoustic structure has been studied, a complete classification of phonetic structure conducted, statistical information obtained and language databases varying in functions and contents have been constructed and are available to be used for executing individual science studies in the field of Kazakh literature. Variety of software programs developed represents a toolset, which will facilitate and contribute to scientific research works studying intonation structures of various languages. With this system available for use, the process of researching Kazakh intonation may be implemented at a whole new level - utilizing processes of speech synthesis. Additionally, this system may be used both as a toolset required for research and as a platform for developing higher quality synthesis systems, utilizing intonation model of language.

## References

1. M. Liberman, *The Intonational System of English*. PhD thesis, MIT, 1975. Published by Indiana University Linguistics Club.
2. H. Fujisaki, and H. Kawai, Modeling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation. In Working Group on Intonation, 13th International Congress of Linguists (1982).
3. P. A. Taylor, *A Phonetic Model of English Intonation*. PhD thesis, University of Edinburgh, 1992. Published by Indiana University Linguistics Club.
4. D. Hirst, and A. Di Cristo, *Intonation Systems: a survey of twenty languages*. Cambridge University Press, 1998.
5. H. Mixdorf, "A novel approach to the fully automatic extraction of Fujisaki model parameters Proceedings of ICASSP, pp. 1281-1284, 2000.
6. M. Schroder, "Emotional speech synthesis: A review. Proceedings of Eurospeech 2001, Aalborg, Denmark, vol. 1, 2001, pp. 561-564.
7. Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B., "A Database of German Emotional Speech In Proc. Interspeech, Lisbon, 2005, pp. 1517-1520.
8. Douglas-Cowie E., Campbell N., Cowie R., Roach P., "Emotional Speech: Towards a New Generation of Database," *Speech Comm.*, Vol. 40, nos. 1/2, 2003, pp. 33-60.
9. Vroomen J., Collier R., Mozziconacci SJL, "Duration and Intonation in Emotional Speech Eurospeech 93, 1993, vol. 1, pp. 577-580.
10. Murray IR, Arnott JL, Rohwer EA, "Emotional stress in synthetic speech: Progress and future directions *Speech Communication*, vol. 20, iss. 1-2, 1996, pp. 85-91.
11. Banse R., Scherer KR, "Acoustic Profiles in Vocal Emotion Expression *J. Personality Social Psychology*, vol. 70, no. 3, 1996, pp. 614- 636.
12. Mozziconacci S., "Prosody and Emotions Proc. First Int'l Conf. Speech Prosody (Speech Prosody '02), 2002, pp. 1-9.
13. Iida A., Campbell N., Higuchi F., et al., "A corpus-based speech synthesis system with emotion *Speech Communication*, vol. 40, is. 1-2, 2003, pp. 161-187.
14. Murray Ir., Arnott JI., "Implementation and testing of a system for producing emotion-by-rule in synthetic speech *Speech Communication*, vol. 16, iss. 4, 1995, pp. 369-390.
15. Burkhardt F., Sendlmeier WF, "Verification of acoustical correlates of emotional speech using formant-synthesis In *SpeechEmotion-2000*, 2000, pp. 151-156.
16. Z.M. Bazarbaeva, *Kazakh intonation*. - Almaty: Dyke Press, 2008.-284 p.
17. J.A. Aralbayev, *Vowels Kazakh language (Essays in experimental phonetics and phonology)*. - Alma-Ata: Nauka, 1970 - 179 p.

18. Emotional Speech Synthesis project: [electronic resource]. URL: <http://emofilt.syntheticspeech.de/>. (Date: 08.07.2012).
19. The MBROLA Project Homepage: [electronic resource]. URL: <http://tcts.fpms.ac.be/synthesis/mbrola.html>.
20. OpenMary: Open Source Emotional Text-to-Speech Synthesis System: [electronic resource]. URL: <http://mary.dfki.de/Download/openmary-open-source-emotional-text-to-speech-synthesis-system-released>.
21. M. Kalimoldaev, Ye. Amirgaliev, R. Musabayev, The Method of Speech Signal Intonation Synthesis Based on Spline Approximation. Computer Modelling and New Technologies, 2011, Vol. 15, No. 2, 65-68
22. GOST R 50840-95. Speech transmission through communication channels. Techniques for quality, intelligibility and recognizability assessment. - Since 11.21.95, Moscow, Russia State Standards (GOSSTANDARDS), Standards Revised, 1995.
23. T. Styger, E. Keller, "Formant synthesis Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges, pp. 109-128, 1994.
24. H.M. Torres, J.A. Gurlekian, "Acoustic speech unit segmentation for concatenative synthesis Computer Speech and Language, Vol. 22, pp. 196-206, 2008.
25. R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech Speech Communication, Vol. 52, pp. 394-404, 2010.
26. H. Altun, K.M. Curtis, T. Yalc?noz, "Neural learning for articulatory speech synthesis under different statistical characteristics of acoustic input patterns Computers and Electrical Engineering, Vol. 29, pp. 687-702, 2003.



# Modification of the Encryption Algorithm, Developed on The Basis of Nonpositional Polynomial Notations

Saule Nyssanbayeva and Miras Magzom

Institute of Information and Computational Technologies of MES RK,  
125 Pushkin str., Almaty, 050010, Republic of Kazakhstan {snyssanbayeva,magzomxzn}@gmail.com  
<http://ipic.kz>

**Abstract.** A model of the encryption algorithm, developed on the basis of nonpositional polynomial notation, is proposed. The possibility of modifying the model, using a Feistel network and encryption modes, is considered. The proposed model of the cryptographic algorithm will considerably improve statistical characteristics of resulting ciphertexts.

**Keywords:** cryptographic system, encryption algorithm, modular arithmetic, Feistel network, encryption mode.

## 1 Introduction

For symmetric block ciphers one of the criteria of cryptographic strength is the length of the secret key. In the encryption system under consideration as a criterion of cryptographic strength the cryptostrength of algorithm itself is used, which is characterized by a complete secret key. Its structure, apart from the standard secret key, also includes secret parameters of the cryptographic algorithm based on nonpositional polynomial notations (NPNs). Synonyms of NPNs - classical notations in residue number system (RNS), polynomial notations systems in RNS, modular arithmetic.

Classical modular arithmetic, or residue number system (RNS), is based on the Chinese remainder theorem, which states that any number can be represented by their remainders (residues) from its division by the base numbers systems, which are formed by pairwise coprime numbers [1,2]. In contrast to the classical RNS, proposed cryptographic procedures are considered in polynomial number system in residue classes, where bases are not prime numbers but are irreducible polynomials in [3,4]. Cryptographic algorithms and methods, based on NPNs are called nonconventional, modular or nonpositional. Nonconventional cryptographic methods and algorithms, developed on the basis of nonpositional polynomial notations (NPNs), allow increasing the reliability of the encryption algorithm and reduce the length of the key. Cryptostrength in this case is defined by full key, which depends not only on the length of a key sequence, but also on choice of a system of polynomial bases and the number of permutations of these bases in the system. If the length of the full encryption key in NPNs is larger, then there are more choices of systems of working bases. Therefore, the cryptostrength of the proposed encryption algorithm based on NPNs significantly increases with the length of the electronic message [3].

## 2 Nonconventional encryption algorithm

The encryption algorithm based on NPNs includes the following steps. For an electronic message of the length  $N$  bits from the set of all irreducible polynomials of degree not exceeding  $N$  a system of working base numbers is selected

$$p_1(x), p_2(x), \dots, p_S(x). \quad (1)$$

According to the Chinese remainder theorem, all the base numbers must be different even if their degrees are equal. The main working range in this system is defined by the polynomial  $P(x) = p_1(x), p_2(x), \dots, p_S(x)$  of the degree  $m$ :

$$m = \sum_{i=1}^S m_i, \quad (2)$$

where  $S$  – is a number of selected working base numbers. In this system any polynomial, which degree is less than  $m$ , has a unique representation in the form of sequence of residues of its division by the working base numbers (1). Therefore, the message of the length  $N$  bits could be represented in the form of sequence of residues  $\alpha_1(x), \alpha_2(x), \dots, \alpha_S(x)$  from division of some polynomial  $F(x)$  by the working base numbers  $p_1(x), p_2(x), \dots, p_S(x)$ :

$$F(x) = (\alpha_1(x), \alpha_2(x), \dots, \alpha_S(x)), \quad (3)$$

where  $F(x) \equiv \alpha_i(x) \pmod{p_i(x)}, i = \overline{1, S}$ .

In the same way the key of the length  $N$  bit is also interpreted as a system of residues  $\beta_1(x), \beta_2(x), \dots, \beta_S(x)$  from the division of another polynomial  $G(x)$  by the same working base numbers:

$$G(x) = (\beta_1(x), \beta_2(x), \dots, \beta_S(x)), \quad (4)$$

where  $G(x) \equiv \beta_i(x) \pmod{p_i(x)}, i = \overline{1, S}$ .

Then the cryptogram  $\omega_1(x), \omega_2(x), \dots, \omega_S(x)$  is considered as some function  $H(F(x), G(x))$ :

$$H(x) = (\omega_1(x), \omega_2(x), \dots, \omega_S(x)), \quad (5)$$

where  $H(x) \equiv \omega_i(x) \pmod{p_i(x)}, i = \overline{1, S}$ .

According to operations of nonpositional notation system, operations in functions  $F(x), G(x), H(x)$  are executed in parallel on the modules of polynomials  $p_1(x), p_2(x), \dots, p_S(x)$ , which are selected as the base numbers of NPNs.

In software implementation of this nonconventional algorithm the encryption method [4] is used. The ciphertext is obtained from multiplication of the polynomials (3) and (4) in accordance with the properties of comparison to the double modulus:

$$F(x)G(x) \equiv H(x) \pmod{P(x)}, \quad (6)$$

i.e. represented as remainders of division of products  $\alpha_i(x)\beta_i(x)$  to the respective base numbers  $p_i(x)$ .

In the decryption process of the cryptogram  $H(x)$  with known key  $G(x)$  for each  $\beta_i(x)$  an inverse polynomial  $\beta_i^{-1}(x)$  is calculated, completing the following comparisons

$$\beta_i(x)\beta_i^{-1}(x) \equiv 1 \pmod{p_i(x)}, i = \overline{1, S}. \quad (7)$$

The result is a polynomial which inverse to a polynomial  $G(x)$ . Then, the original message is restored over:

$$F(x) \equiv G^{-1}(x)H(x) \pmod{P(x)}. \quad (8)$$

### 3 Application of the modified Feistel network

In the development of symmetric block cipher the cryptosystem called Feistel scheme has gained wide popularity. It was first used by Horst Feistel in 1973 in the development of the cipher Lucifer[5], and then used in many developments of block ciphers, including standards DES and AES [6].

Feistel scheme is a method of blending the sub-blocks of the input text in the cipher through the repeated use of the key-dependent non-linear functions, called F-functions and performance of permutations of the sub-blocks. Round of a block cipher is a transformation that connects the sub-blocks of the input block by the F-function and permutations of sub-blocks. In the standard Feistel network, the plaintext is divided into two sub-blocks of the same length.

In general case, the Feistel network can split an input block into  $n \geq 2$  sub-blocks. Further assumed that all sub-blocks are of the same length, so that each sub-block may be involved in the transposition with any other sub-block. A generalized exchange scheme is a permutation of  $n \geq 2$  sub-blocks in the round.

The developed encryption algorithm based on NPNs is the basis for solving the problems of its practical use. In development of the model of unconventional encryption algorithm, it is planned to use the modified Feistel network to develop its application mode. The aim of this works is to improve the statistical characteristics of nonpositional cryptograms. In this regard, it is planned to consider several models of the Feistel scheme.

Unlike traditional Feistel network where the input data is a plain text message, in the developed model the input is a bit sequence of the ciphertexts obtained in (5).

A necessary condition for the strength of the cipher is achievement of complete diffusion. The diffusion process of the cipher is characterized by the distribution of influence of the one of the input bit on many output bits. Cipher is called complete if every output bit depends on all input bits [7]. In the considered model, all F-functions are implied to be complete.

Most ciphers with Feistel network architecture use function F that each round only depends on one of subkeys generated from the main encryption key. A network with such dependence of the function is called heterogeneous and homogeneous otherwise. The use of heterogeneous networks can significantly improve the properties of the cipher, as uneven changes of the internal properties of the network within the permissible limits make study of the characteristics of the cipher rather difficult task.

For example, consider a model in which the input block of 128 bits is divided into two sub-blocks of equal length  $R$  and  $L$ .

When using a homogeneous network, at each round of encryption a separate key sequence  $K_i$  is used:

$$L_i = R_{i-1} \quad (9)$$

$$R_i = L_{i-1} \oplus F(R_{i-1}, K_i) \quad (10)$$

When using a heterogeneous network at each round the encryption function F depends not only on the round key  $K_i$ , but also on the chosen system of base numbers (1):

$$L_i = R_{i-1} \quad (11)$$

$$R_i = L_{i-1} \oplus F(R_{i-1}, K_i, P(x)) \quad (12)$$

During computer modeling of developed modified algorithms, the statistical characteristics of the resulting ciphertexts will be analyzed. Verification of the model for satisfying the strict avalanche criterion will be conducted by examining the received bit sequence through statistical

tests of uniformity (frequencies) - Frequency (Monobit) Test for cryptographic functions from National Institute of Standards and Technology [8]. "NIST Statistical Test Suite statistical package consisting of 16 tests designed to check the randomness of binary sequences produced both by technical means and by software.

#### 4 Conclusion

The proposed modification of the nonpositional cryptographic algorithm is the basis for its software implementation. On implemented model, the statistical characteristics of obtained ciphertexts will be investigated. The results of computer modeling will allow making recommendations on the use of the described encryption model.

#### References

1. Akushskii, I.Ya., Juditskii, D.I., *Machine Arithmetic in Residue Classes [in Russian]*, Sov. Radio, Moscow (1968).
2. Bijashev, R.G., *Development and investigation of methods of the overall increase in reliability in data exchange systems of distributed ACSs*, Doctoral Dissertation in Technical Sciences, Moscow (1985).
3. Bijashev, R.G., Nyssanbayeva S.E., *Algorithm for Creation a Digital Signature with Error Detection and Correction*, Cybernetics and Systems Analysis, 4, 489-497 (2012).
4. Nyssanbayev, R.K., *Cryptographical method on the basis of polynomial bases*, Herald of the Ministry of Science and Higher Education and National Academy of Science of the Republic of Kazakhstan, 5, 63-65 (1999).
5. Feistel H., *Cryptography and Computer Privacy*, Feistel H., Scientific American, V. 228, N.5, 15-23(1973).
6. Bassham L., Burr W., Dworkin M., Foti J., Roback E., *Report on the Development of the Advanced Encryption Standard (AES)*, Computer Security Division, Information Technology Laboratory; NIST: Technology Administration; U.S. Department of Commerce, 116 p. (2000).
7. Schneier B., Kelsey J., *Unbalanced Feistel Networks and Block-Cipher Design*, Fast Software Encryption, Third International Workshop Proceedings (February 1996), Springer-Verlag, 121-144 (1996).
8. Rukhin A., Soto J., [8] *A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications*, NIST Special Publication 800.-22, 154 p. (2001).

# Information-Analytical System "ECO Monitoring"

Saule Rakhmetullina, Alexey Penenko, Yerken Turganbayev, and Alexey Bublikov

East Kazakhstan State Technical University named after D.Serikbayev,  
69 Protozanova Street, 070004 Ust-Kamenogorsk, Kazakhstan,  
Institute of Computational Mathematics and Mathematical Geophysics SB RAS,  
prospect Akademika Lavrentjeva, 6, Novosibirsk, Russia  
Novosibirsk State University,  
Pirogova Str., 2, Novosibirsk, Russia,

rakhmetullinas@mail.ru, a.penenko@yandex.ru, eturganbaev@ektu.kz, bublikov.alexey@gmail.com

**Abstract.** This paper proposes a new approach to the development of information systems of environmental monitoring based on joint use of actual data of automated observing system, modern and efficient algorithms of data assimilation and sources of pollution localization, WRF computer model of atmospheric dynamics. Information-analytical system "ECO Monitoring" that allows solving of tasks of environmental monitoring such as atmospheric dynamics simulation, modeling of air pollution, sources of pollution localization, visualization of simulation results has been developed. Also a mobile version of the information system has been designed.

**Keywords:** Environmental information system, mathematical models, pollution, source localization.

## 1 Introduction

One of the important directions of the development of environmental information systems is the design of analytical support systems for the monitoring of air pollution in cities([1], [2], [3], [4], [5]). To date, due practice extensive experience in developing information systems that solve certain tasks of environmental monitoring ([6], [7], [8], [9]) is gained. Analysis of existing environmental monitoring systems showed that tasks such as data assimilation, pollution sources localization are rarely solved, and often only for research purposes. In the field of ecology, attempts to create a multipurpose information system are not implemented, the local use of powerful computers for optimizing individual processes does not bring the desired effect, we need a holistic interconnected and interdependent information system that solves the problem of environmental monitoring in complex with the use of actual measurement data.

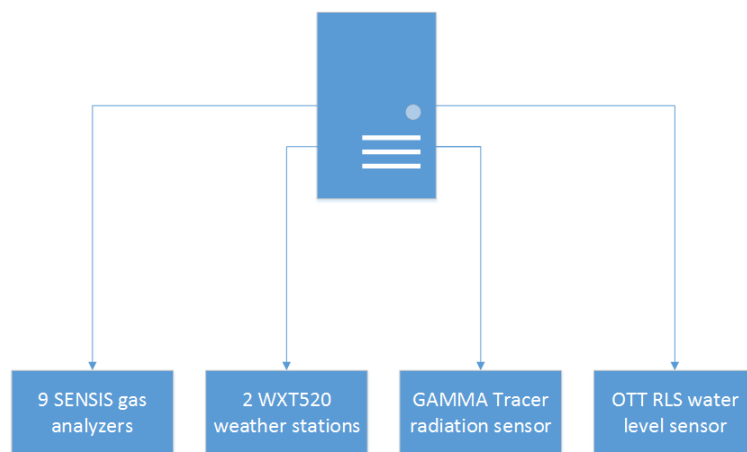
A research team of the D.Serikbayev EKSTU is actively working in this field and performs funded research works ([10], [11], [12]). The choice of the direction of research is determined by the requests of environmental institutions of the region and unfavourable ecological situation. In the framework of the research the team has developed an information system of environmental monitoring "ECO Monitoring". Based on the data of an automated air quality observation system the information system allows to compute meteorological fields, concentrations of pollutants, localization of emission sources and to visualize data. The system is based on a client-server architecture. User interaction with the system is performed by web-interface.

The aim of the paper is to design an information analytical system of the ecological monitoring of the atmosphere of an industrial city. The purpose of the system is to forecast the propagation of pollutants and to determine sources of pollution.

To address these challenges, the system uses data of the automated monitoring system. For the calculation of climate fields the Weather Research and Forecasting (WRF) model is used. The obtained simulation results are compared with those of the automated monitoring system.

## 2 Information analytical system "ECO Monitoring"

**Architecture of the system.** In the framework of environmental monitoring of the Ust-Kamenogorsk city, an automated system of monitoring of air quality, meteorological parameters, background radiation and water level operates (Fig. 1). As a result of introduction of this system it has become necessary to use operational data of air monitoring and retrospective observational database for solving problems related to environmental forecasting. The problems of modeling and forecasting of air pollution; localization and evaluation of capacities of pollution sources; determination of the areas most exposed to sources are very urgent tasks in the system of air quality monitoring of an industrial city. Combining of methods of mathematical modeling, modern information technologies, WEB technology and observational data is an effective tool for solving environmental monitoring problems.



**Fig. 1.** Automated observation system

We developed an information-analytical system of ecological monitoring "ECO Monitoring". The system is based on a client-server architecture, which is shown in Fig. 2. Within the information-analytical system the interaction of main subsystems is carried out: observations, mathematical modeling, data management.

Informational support of the system is formed based on observational database containing data from all observation points and database of modeling that includes data on main sources of pollution, meteorological data fields and results of simulation. The acquisition of data from the automated system is performed in real-time with 20 minutes interval of measurement of concentrations. There are 9 observation points that measure concentrations of the following pollutants: nitrogen dioxide (NO<sub>2</sub>), sulfur dioxide (SO<sub>2</sub>), chlorine (CL<sub>2</sub>), carbon monoxide (CO), hydrogen chloride (HCL), hydrogen fluoride (HF), formaldehyde (HCOH), hydrocarbons (C<sub>x</sub>H<sub>y</sub>).

Users interact with the system by the web interface (Fig. 3-5). The development of the application was directed to the cross-browser with a bias towards smartphones and tablets. Thus, we adapted the "ECO Monitoring" web application for mobile devices using the technology of adaptive web design.

Modeling system consists of several subsystems: prediction of propagation of pollutants, localization of sources of pollution, assimilation of observation data, visualization.

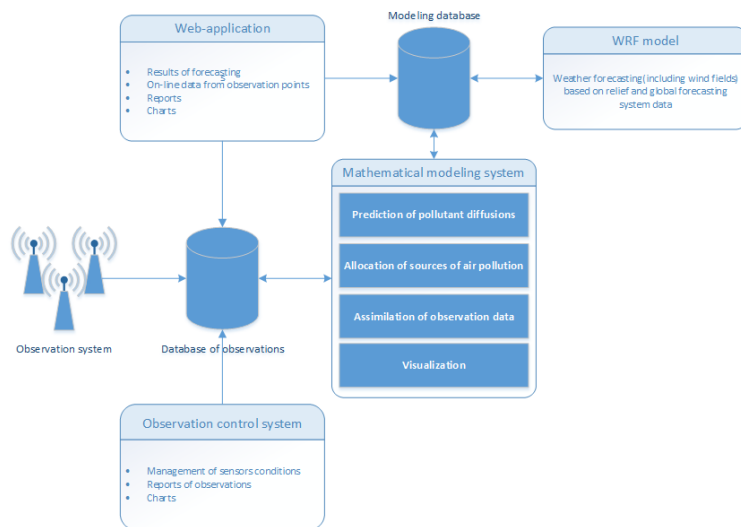


Fig. 2. Architecture of the information analytical system

Physically, the system is placed on two servers, because of both the difference in the platforms on which the system is implemented, and the participation of various organizations in the implementation of the monitoring system. Data is transmitted via TCP / IP protocol.

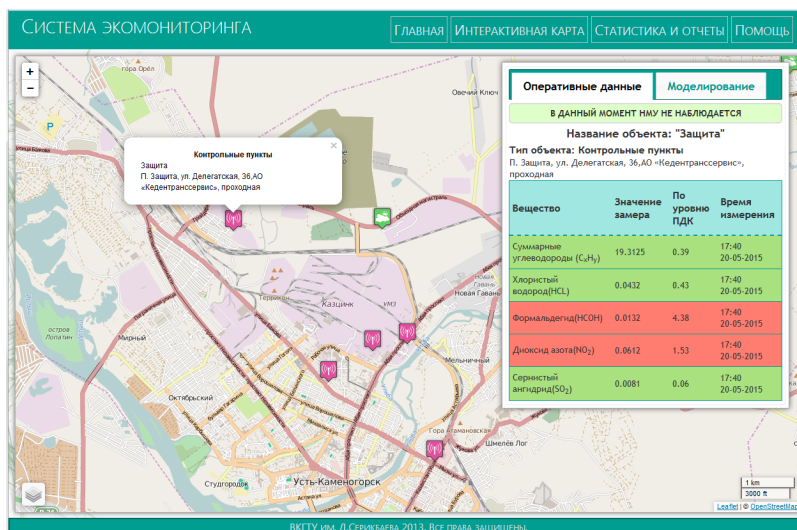


Fig. 3. Web-application (Interactive Map)

On the server with Server Windows Server 2008:

- Management System
- Database of observations
- Management of observation points
- Web-application “ECO Monitoring”

On the server with Server Ubuntu 12.04:

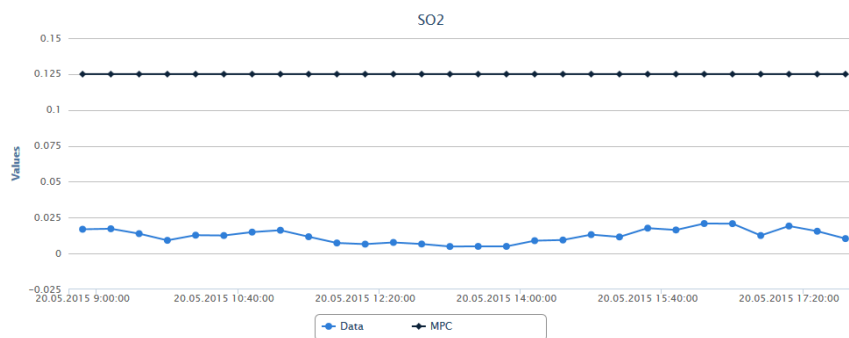


Fig. 4. Web-application (Observation data chart)

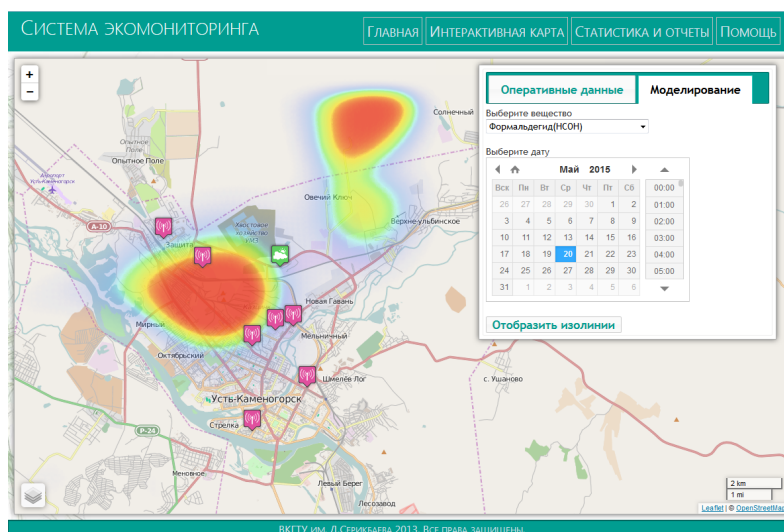


Fig. 5. Web-application (Modeling visualization)



- WRF Model
- System of mathematical modeling
- Database of modeling

The information-analytical system was developed using the following software:

- Microsoft Windows Server 2008 and Microsoft SQL Server 2008 R2 was used to provide functioning of the System of Observations (management of gas sensors, data storage of observations).
- UNIX was used for the operation of WRF model and integration of data of meteorological fields.
- Qt4 was used for the processing of results of forecasting.
- Web application "ECO Monitoring" was developed using: ASP.NET MVC Framework 4, JavaScript Library Leaflet (as an API for working with maps), OpenStreetMaps - as a map service.
- The system of mathematical modeling is implemented on the C++ language, applications run under Ubuntu 12.04, modeling database uses MySQL.

### 3 The mathematical part of the informational-analytical system

The mathematical part of the subsystem of the localization of sources and the subsystem of monitoring data assimilation was developed using approaches to the source estimation based on variational principle with the use of adjoint equations for the convection-diffusion models with source-term uncertainty.

The localization of sources of pollution is carried out based on the following approaches ([13], [10], [11]).

- The algorithm of the estimation of the capacity of sources of pollution is based on finding of fundamental solutions of the direct problem for each source and use of measurement data.
- In the algorithm of localization of sources of pollution based on functions of observability, we define area where supports of sensitivity functions are in a range above the specified level of significance.
- In the algorithm of localization of sources of pollution based on the definition of the generalized normal solution, we construct a system of equations of feedbacks expressing the dynamics of changes of model parameters proportionally to the corresponding functions of the sensitivity of the functional to variations in the parameters.

In data assimilation we use an approach based on the formulation of data assimilation problems, taking into account uncertainties in models and results of observations. A variational principle is constructed with the conditions of minimization of functionals containing total measure of uncertainty. Data assimilation algorithm is applied to the convection-diffusion model.

#### 3.1 Variational data assimilation algorithm for convection-diffusion models

Multidimensional mathematical models, describing processes of transport and transformation of heat, moisture, radiation and substances:

$$L(\mathbf{Y})\phi \equiv \frac{\partial \phi}{\partial t} + \text{div}(\phi \mathbf{u} - \mu \text{grad } \phi) = \mathbf{f}_a + \mathbf{r},$$

$$\phi^0 = \phi_a^0, \quad R_{\text{bound}}(\phi) = \mathbf{g}_a, \quad \mathbf{Y} = \mathbf{Y}_a.$$

Measurement data is connected to the state function with the observation operator  $\mathbf{H}$

$$\Psi = \mathbf{H}(\phi) + \eta,$$

$\phi$  - model state function,  $\{\mathbf{u}, \mu\} = \mathbf{Y}$  - model parameters,  $\mathbf{f}_a, \mathbf{g}_a, \phi_a^0$  - *a priori* source functions,  $\Psi$  - measurement data,  $\mathbf{r}, \eta$  - control (uncertainty) functions in the sets, that are introduced to the rigid model structure to assimilate data. Find  $\phi$  for  $t \geq t^*$  with measurement data for  $t \leq t^*$ .

**Additive-averaged splitting scheme.** On the time interval  $t^j \leq t \leq t^{j+1}$  consider additive-averaged splitting scheme (analogous to [14]), for the partition  $\sum_{\beta=1}^2 \gamma_\beta = 1$ .

– Convection-diffusion processes

$$\gamma_\beta \frac{\partial \phi_\beta}{\partial t} + L_\beta \phi_\beta = \mathbf{f}_\beta + \mathbf{r}_\beta, \quad \phi_\beta(t^j) = \phi(t^j), \quad \beta = 1, 2.$$

– Next step approximation

$$\phi = \sum_{\beta=1}^2 \gamma_\beta \phi_\beta.$$

On the grid lines containing measurement data a standard direct model step is substituted to the minimization of the corresponding extended target functional

$$\begin{aligned} \bar{\Phi}_x(\phi_x^{j+1}, r_x^{j+1}, \phi_x^*) &= \sum_{i=0}^{N_x-1} \left( \frac{(\phi_x^{j+1})_i - \Psi_i^{j+1}}{\sigma_i} \right)^2 M_i^{j+1} \tau + \alpha_l \sum_{i=0}^{N_x-1} (r_{x_i}^{j+1})^2 \tau \\ &+ \langle ((I + 2\tau A_x) \phi_x^{j+1} - \phi^j - \tau r_x^{j+1})_l, \phi_x^* \rangle. \end{aligned}$$

$$\begin{aligned} \bar{\Phi}_y(\phi_y^{j+1}, r_y^{j+1}, \phi_y^*) &= \sum_{l=0}^{N_y-1} \left( \frac{(\phi_y^{j+1})_l - \Psi_l^{j+1}}{\sigma_l} \right)^2 M_l^{j+1} \tau + \alpha_i \sum_{l=0}^{N_y-1} (r_{y_l}^{j+1})^2 \tau \\ &+ \langle ((I + 2\tau A_y) \phi_y^{j+1} - \phi^j - \tau r_y^{j+1})_l, \phi_y^* \rangle. \end{aligned}$$

### 3.2 Implicit data assimilation for one-dimensional convection-diffusion model

Non-stationary one-dimensional convection-diffusion model

$$\gamma_\beta \frac{\partial \phi_\beta}{\partial t} + L_\beta \phi_\beta = \mathbf{f}_\beta + \mathbf{r}_\beta, \quad \phi_\beta(t^j) = \phi(t^j), \quad \beta = 1, 2.$$

approximated on a spatial-temporal grid. To do this one can use an upwind scheme.

This leads to tri-diagonal matrix systems

$$\begin{aligned} -a_i \phi_{i+1}^{j+1} + b_i \phi_i^{j+1} &= \phi_i^j + \tau r_i^{j+1}, \quad i = 0, \\ -a_i \phi_{i+1}^{j+1} + b_i \phi_i^{j+1} - c_i \phi_{i-1}^{j+1} &= \phi_i^j + \tau r_i^{j+1}, \quad i = 1, \dots, N-2, \\ b_i \phi_i^{j+1} - c_i \phi_{i-1}^{j+1} &= \phi_i^j + \tau r_i^{j+1}, \quad i = N-1, \end{aligned}$$

**Data assimilation problem solution.** For contact measurement case we can introduce measurement system mask  $M_i^{j+1}$  that is equal to 1 in measurement point and 0 in others.

Data assimilation problem solution is the minimum of the functional

$$\Phi(\phi^{j+1}, r^{j+1}) = \left( \sum_{i=1}^{N-1} \left( \frac{\phi_i^{j+1} - \Psi_i^{j+1}}{\sigma_i} \right)^2 M_i^{j+1} + \alpha \sum_{i=1}^{N-1} \left( r_i^{j+1} \right)^2 \right) \frac{\tau}{2},$$

wrt direct problem where  $\sigma_i$  are standard deviation for the measurement instrument errors. Introducing Lagrange multipliers (adjoint functions):

$$\begin{aligned} \bar{\Phi}(\phi^{j+1}, r^{j+1}, \phi^{*j+1}) &= \left( \sum_{i=0}^{N-1} \left( \frac{\phi_i^{j+1} - \Psi_i^{j+1}}{\sigma_i} \right)^2 M_i^{j+1} + \alpha \sum_{i=0}^{N-1} \left( r_i^{j+1} \right)^2 \right) \frac{\tau}{2} \\ &+ \sum_{i=0}^{N-1} \left( -a_i \phi_{i+1}^{j+1} + b_i \phi_i^{j+1} - c_i \phi_{i-1}^{j+1} - \phi_i^j - \tau r_i^{j+1} \right) \phi_i^{*j+1}. \end{aligned}$$

**Matrix system.** A system consisting of direct and adjoint problem can be assembled to the matrix equation [15], [16], [17]

$$\begin{aligned} -A_i \Phi_{i+1}^{j+1} + B_i \Phi_i^{j+1} &= F_i^{j+1}, \\ -A_i \Phi_{i+1}^{j+1} + B_i \Phi_i^{j+1} - C_i \Phi_{i-1}^{j+1} &= F_i^{j+1}, \\ B_i \Phi_i^{j+1} - C_i \Phi_{i-1}^{j+1} &= F_i^{j+1}, \end{aligned}$$

where

$$\begin{aligned} A_i &= \begin{pmatrix} a_i & 0 \\ 0 & c_{i+1} \end{pmatrix}, \quad B_i = \begin{pmatrix} b_i & -\tau \\ \frac{M_i \tau}{\alpha \sigma_i^2} & b_i \end{pmatrix}, \quad C_i = \begin{pmatrix} c_i & 0 \\ 0 & a_{i-1} \end{pmatrix}, \\ \Phi_i^{j+1} &= \begin{pmatrix} \phi_i^{j+1} \\ \phi_i^{*j+1} \end{pmatrix}, \quad F^{j+1} = \begin{pmatrix} \phi_i^j \\ \frac{M_i \tau}{\alpha \sigma_i^2} \Psi_i^{j+1} \end{pmatrix}, \end{aligned}$$

which is solved with the matrix sweep method.

### 3.3 External parameters for data assimilation

One of the input parameters for the convection-diffusion model is a function describing dynamics of the atmosphere. We reviewed models describing the atmospheric processes at the mesoscale level and found that of them are implemented in software and can be used for short-term and long-term weather forecasts, atmospheric modeling, etc. ([18],[19],[20]). Based on the analysis of models of atmospheric dynamics we chose the WRF model (dynamic kernel of AWR) because it is a model of common use, does not demand any licensing conditions and is widely used in the Weather Forecast Office of Kazakhstan. To control the mesoscale model and to retrieve the output of meteorological data for further use in the process of modeling of pollution transport and localization of sources we developed a program module. To retrieve data we used a library written in C++ language, which provides access to the data file, all measurement and direct access to the data.

The module allows the correct start of all components of the WRF model, the correct extraction of data from the output file and correct record of data to the database. The use of the module in the information system "ECO Monitoring" allows to simulate meteorological dynamics of atmosphere in the air pollution modeling, assimilation of data of environmental monitoring and localization of sources of pollution.

#### 4 Conclusion

Thus, in this paper we propose a new approach to the development of information systems of environmental monitoring based on joint use of actual data of automated observing system, modern and efficient algorithms of data assimilation and sources of pollution localization, a computer model dynamics of atmospheric WRF. We developed the information-analytical system "ECO Monitoring" that allows to solve such tasks of environmental monitoring as the modeling of atmospheric dynamics, localization of air pollution sources, visualization of simulation results.

Analytical part of the information system is based on modern algorithms of variational data assimilation and localization of pollution sources under conditions of uncertainty in the real-time environment. Combining splitting schemes and data assimilation schemes allows to construct computationally effective algorithms without iterations for data assimilation of in situ measurements to convection-diffusion-reaction models.

The performance of the system has been tested using data from the automated observation system. The results of simulation are consistent with the observation data. The system has passed operational testing at the Center for Environmental Monitoring of the city of Ust-Kamenogorsk.

The use of the developed system allows efficient use of the data of the automated observation system and improvement of environmental decision-making.

**Acknowledgments.** This research was supported by the Research Grant of the Ministry of Education and Science of the Republic of Kazakhstan.

#### References

1. Bakirbayev B, Danayev NT. (2002) Mathematical Modeling of Climate Change Due to Natural and Anthropogenic Factors (in Russian), 314, 212-225.
2. Penenko VV, Andreeva IS, Belan BD, Borodulin AI, Buryak GA, Zhukov VA, et al. (2001) Variability of the Content of Live Microorganisms in the Atmospheric Aerosol in Southern Regions of Western Siberia "Doklady Biological Sciences: Proceedings of the Academy of Sciences of Russia" 381, 530-534.
3. Penenko VV, Tsvetova EA. (2007) Mathematical Models of Environmental Forecasting "Journal of Applied Mechanics and Technical Physics" March, pp.428-436.
4. Penenko VV, Tsvetova EA. (2009) Optimal Forecasting of Natural Processes with Uncertainty Assessment "Journal of Applied Mechanics and Technical Physics" Feb. pp. 300-308.
5. Zambakas JD, Retalis DA, Mavrakis DC. (1985) A Simultaneous Interpretation, by Wind Speed and Direction, of Ambient Air Polar Conductivities in Athens "Archives for Meteorology, Geophysics and Bioclimatology" Greece, Apr. pp.381-388.
6. NPO Firma Garant, consulted 20 March, 2014 <http://garant.hut.ru>
7. Logos Plus, consulted 20 March, 2014 <http://lpp.ru>
8. EkologPS, consulted 20 March, 2014 <http://www.ekolog-ps.ru>
9. NPP Logus, consulted 20 March, 2014 <http://www.logus.ru>
10. Zhumagulov BT, Temirbekov NM, Rakhmetullina SZ, Turganbayev YM, Denisova NF. (2013) Information System of Ecological Monitoring and Application of Variational Algorithms (in Russian). "The Bulletin of National Engineering Academy of the Republic of Kazakhstan" Kazakhstan, Jan. pp. 10-17.
11. Rakhmetullina SZ, Turganbayev YM, Gromaszek K. (2012) Application of Variational Data Assimilation Algorithms in the Ecological Monitoring System "Informatics Control Measurement In Economy and Environment Protection" Apr. pp. 33-35.

12. PenenkoAV, RakhmetullinaSZ (2013) Localization of pollution sources of atmospheric air by using data of environmental monitoring system (in Russian) "Computational Technologies" Sep. pp. 152-164.
13. Penenko VV, Baklanov A, Tsvetova EA. (2012) Direct and Inverse Problems in a Variational Concept of Environmental Modeling "Pure and Applied Geophysics" March, Volume 169, Issue 3, pp. 447-465.
14. A.A. Samarskii and P.N. Vabishchevich. Computational Heat Transfer. Vol.1,2. Wiley, Chichester, 1995.
15. Penenko, A.V. and Penenko V.V. (2014), Direct data assimilation method for convection-diffusion models based on splitting scheme. Computational technologies, 19 Issue 4, 69 (In Russian).
16. V. V. Penenko *Variational methods of data assimilation and inverse problems for studying the atmosphere, ocean, and environment* Num. Anal. and Appl., 2009 V 2 No 4, 341-351.
17. A. Penenko *Some theoretical and applied aspects of sequential variational data assimilation (In Russian)*, Comp. tech. v.11, Part 2, (2006) 35-40.
18. Consortium for Small-scale Modeling, consulted 20 March, 2014 <http://www.cosmo-model.org>
19. The international research programme HIRLAM, consulted 20 March, 2014 <http://www.hirlam.org>
20. The Weather Research and Forecasting Model, consulted 20 March, 2014 <http://www.wrf-model.org/index.php/>

# Design of Algorithms for Automated Access Control Based on Business Process Approach

Zinaida Rodionova

Novosibirsk State University of Economics and Management,  
Computer Science Department,  
Novosibirsk, Russian Federation  
z.v.rodionova@nsuem.ru

**Abstract.** The article describes access control problem solving to information system resources in the context of common change of business. The algorithms are designed for automated generation and actualization of access control model elements (RBAC, MAC, and DAC). Input information for the algorithms is data from business process model. Estimating algorithms complexity was made on account of necessary operations quantity. The main research methods were set theory, algorithm theory, and structure analysis. The findings of the article can be helpful for IT and security professionals.

**Keywords:** algorithm, access control, business process, security, access control model.

The efficiency of the modern computer-based information systems directly depends on the correspondence between the user credentials and their responsibilities. The formalized credentials in the form of access rights are reflected in the settings of the access control system, the secure building of which is determined by the formal model. Despite the high level of theoretical research in the field of formal access models, their practical implementation encounters significant difficulties relating to the interpretation, i.e. ensuring the compliance of abstract entities and model processes with real objects and information system rules and updating access rights due to the constant changes in the business processes.

This problem is supposed to be solved by using algorithms to form and update a set of access rights on the basis of business processes analysis [1].

To realize the possibilities of access rights formalization and updating in the conditions of access delimitation systems, which function on the basis of different formal (role, discretionary, mandatory) models, there is a generalized model of access rights delimitation. Business processes are analyzed in order to get necessary data to build a generalized access model.

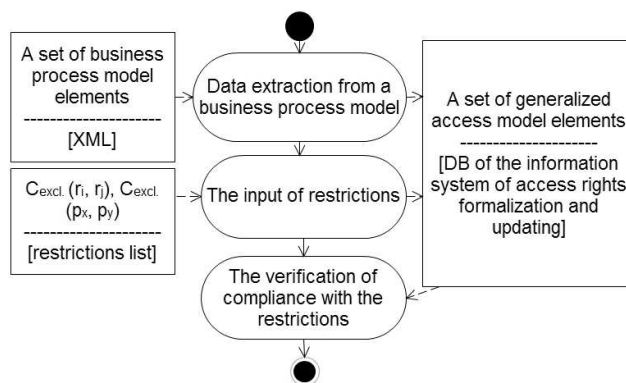
Let  $FF = \{ff_1, ff_2, \dots, ff_a\}$  be a finite nonempty set of functions performed by the members of a business process (performers). For each function  $ff \in FF$  in the business process, there is a set of performers  $PP(ff) = \{pp_1(ff), pp_2(ff), \dots, pp_d(ff)\}$ , the performers are denoted by

$PP = \bigcup_{ff \in FF} PP(ff)$  Let  $PP = \{pp_1, pp_2, \dots, pp_e\}$  denote a set of performers (the position, the functional role) which cannot be empty. We shall introduce the owner-member relationship  $<$  on set  $PP$ . Assume that  $pp_1 < pp_2$  if one of two conditions is fulfilled: a).  $pp_2$  is the direct supervisor of  $pp_1$ ; b).  $\exists$  chain  $pp_1^0 = pp_1, pp_1^1, pp_1^2, \dots, pp_1^n = pp_2, pp_1^i$  is the direct supervisor of  $pp_1^{i-1}$  for all  $i = 1, n$ . Every performer  $pp$  is bounded with set  $FI(pp) = \{fi_1(pp), fi_2(pp), \dots, fi_g(pp)\}$  - the name of the performer, the set of all performers is denoted by  $FI = \bigcup_{pp \in PP} FI(pp)$ .

Each element of "the performer's name"  $fi \in FI$  is labeled by  $sfi(pp)$ , which is contained in its properties and will be associated with the access level. Let  $SFI = (sfi_1, sfi_2, \dots, sfi_y)$  denote the set of properties of the performer's name. The finite set of information systems  $IS = \{is_1, is_2, \dots, is_o\}$  is defined. As a certain number of functions is performed automatically

and it is reflected in the business process model, we can introduce the binary relation  $fis = \{(ff, is), \text{ where } ff \in FF, is \in IS\}$ ,  $ff$  is performed in the information system  $is$ . The operations on a finite set of information objects  $IO = \{io_1, io_2, \dots, io_u\}$  are performed with the help of automated functions. Each data object  $io \in IO$  is labeled by  $sio(io)$  which is contained in its properties and will be associated with the access level. We denote the set of properties  $sio = (sio_1, sio_2, \dots, sio_f)$ . The information system is connected with an information object set  $IS(io) = (is_1(io), is_2(io), \dots, is_b(io))$ .

The kind of operations is determined by the type of the performer's access to the information object, a set of access types is denoted by  $AT$ . The performer's operation with an information object is defined by the three  $Op = (pp, io, at)$ , where  $pp \in PP, io \in IO, at \in AT$ , at  $AT$ ,  $pp$  performs the operation by means of  $io$ . Using the denotations of the generalized access model and formalized representation of a business process model, the algorithm for extracting the necessary data to formalize and update access rights can be represented in a general form (Fig. 1):



**Fig. 1.** Algorithm for automated formation of the elements of the generalized model of access rights delimitation from a business process model

- the data extraction from a business process model, e.g. from a XML file.
- the input of restrictions ( $C_{excl.}(r_x, r_y)$ ) - the mutual exclusion of roles,  $C_{excl.}(p_b, p_c)$  - the mutual exclusion of access rights,  $V(r_d)$  - a quantitative restriction on the role possession.
- the verification of compliance with the restrictions (the consistency check).

The block "Data extraction from a business process model" is presented in details in Fig. 2.

We have made the estimate of the algorithm complexity based on the number of operations required to implement it. The basic operations of the algorithm are connected with the linear search of the elements of several sets and their recording in other sets. The linear search is performed via loops (Fig. 3).

Let  $T_x$  denote the number of operations for loop  $X$ ,  $T(is)$  is the estimate of the complexity of data extraction algorithm for a fixed information system is ( $1 \leq is \leq |IS|$ ).

Then the estimate of the algorithm complexity  $T$  can be obtained by organizing loop  $E$ , in which  $T(is)$  is summed for all information systems:

$$T = \sum_{is=1}^{|IS|} T(is) = T_E = O(|IS|) \cdot (T_c + T_d) \tag{1}$$

We estimate the algorithm complexity  $T(is)$  for a fixed information system with number  $is$ .

The number of operations in loop  $D$  for every information system is estimated by the constant, so  $T_d$  is determined by the number of information objects of system  $T_d|IO(is)|$ , and it can be

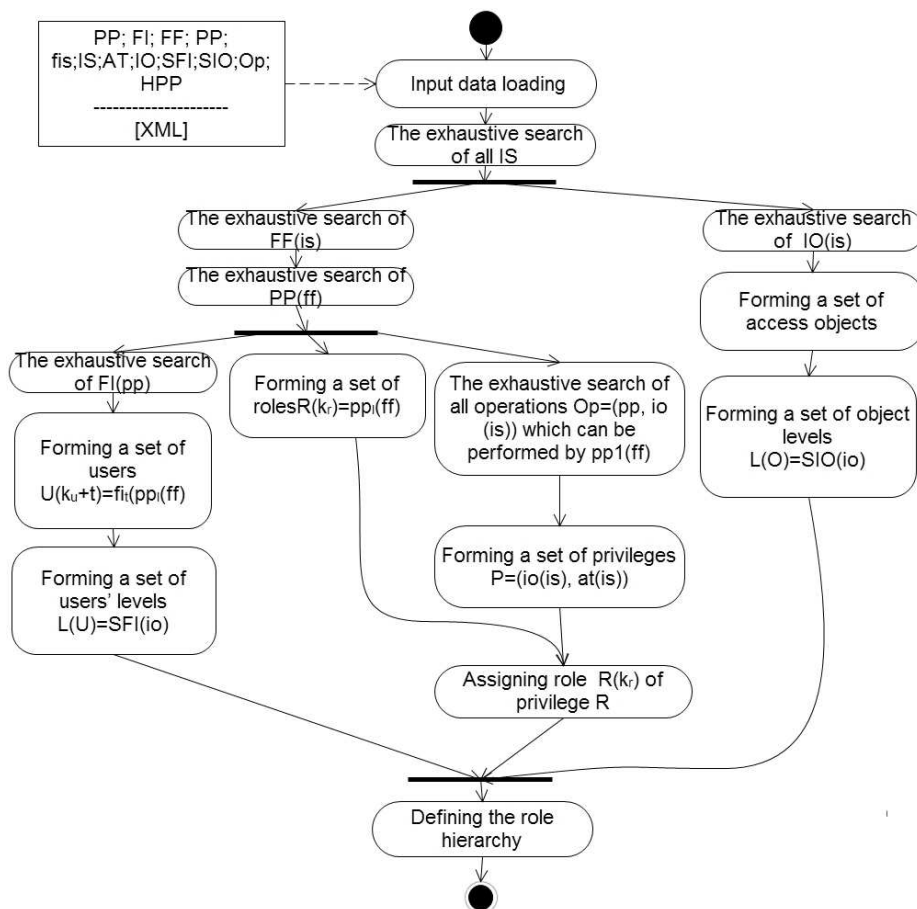


Fig. 2. The block "Data extraction from a business process model"

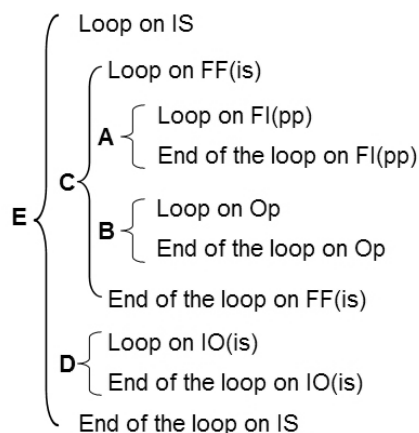


Fig. 3. The schematic representation of algorithm loops



represented as:

$$T_D = O(|IO(is)|) \quad (2)$$

The exhaustive search in loop  $C$  is performed over all the functions of the business process implemented in information system  $is$ , the number of which  $|FF(is)|$  determines the repetitions number in this loop. In addition, the loop body includes two other consecutive loops  $A$  and  $B$ , performed for every fixed function with number  $1 \leq ff \leq |FF(is)|$ . Thus, the complexity of loop  $C$  can be recorded as:

$$T_C = O(|FF(is)|) \cdot (T_A + T_B) \quad (3)$$

We estimate the number of operations performed in loops  $A$  and  $B$ .

The number of operations in loop  $A$  can be expressed as  $T_A = O(|FI(pp)|)$  because this loop organizes the exhaustive search of all names of the performers  $pp$  of function  $ff$ . In fact, the number of the names is equal to the number of system is users  $|FF(is)|$ , which allows us to write:

$$T_A = O(|FI(is)|) \quad (4)$$

Such situation is quite possible, because in many information systems, every user has at least one type of access to most information objects. It should be noted that in any case, this number does not exceed the number of the employees (individuals) who work in the organization, i.e., in the worst case, when the vast majority of the organization's employees (individuals) perform some of the functions in all or almost all information systems, the search in loop  $A$  can be limited by the number of the employees in the organization, and in compliance with (4) can be rewritten as:

$$T_A \leq O(|FI|) \quad (5)$$

The complexity of loop  $B$  is determined by the amount of searching the elements of set  $Op$ :  $T_B = O(|Op|)$ . The elements of set  $Op$  in relation to a given information system  $is$  and function  $ff$  are the three of the form  $op = (pp, io, at)$ , where

-  $pp$  is the performer of a business process (their number is determined by the number of available functional roles  $|PP(is)|$  in the information system, and does not exceed the number of posts in the organization  $|PP(is)| \leq |PP|$ , while the worst (in terms of computational complexity) case corresponds to the situation where each organization official unit is represented by its own role in every information system);

-  $io$  is an information object, wherein  $|IO(is, ff)|$  is the number of information objects involved in the implementation of function  $ff$  in information system  $is$  - it is limited by  $|IO(is)|$ , which in turn does not exceed the total number of information objects of business process  $|IO|$ , i.e.,  $|IO(is, ff)| \leq |IO(is)| \leq |IO|$  (worst case corresponds to the situation where all or nearly all of the functions of any information system are used by the vast majority of business process information objects of all);

-  $at$  is an access type in an information system (the number of different access types depends only on the information system, in practice it is relatively small and may be limited to a constant of the order of 10).

Thus, if we assume that in the worst case, every function requires all information objects of the system and each function is used by all system users, and, besides, if we consider that  $|AT(is)| = O(1)$ , we obtain  $|Op| = O(|PP(is)| \cdot |IO(is)|)$ . In this case, the estimate of the operations number for loop  $B$  is given by the following formula:

$$T_B \leq O(|PP(is)| \cdot |IO(is)|) \quad (6)$$

We substituting the expressions (5) and (6) into (3) and get

$$T_C \leq O(|FF(is)|) \cdot (O(|FI|) + O(|PP(is)| \cdot |IO(is)|)) \quad (7)$$

Using (2) and (7), we obtain an expression for estimating the complexity of the extracting data algorithm for information system  $is$ :

$$\begin{aligned} T(is) &\leq O(|FF(is)|) \cdot (O(|FI| + O(|PP(is)| \cdot |IO(is)|))) + O(|IO(is)|) = \\ &= O(|FF(is)|) \cdot O(|FI|) + O(|FF(is)|) \cdot O(|PP(is)| \cdot |IO(is)|) + O(|IO(is)|) = \\ &= O(|FF(is)| \cdot |FI|) + O(|FF(is)| \cdot |PP(is)| \cdot |IO(is)|) + O(|IO(is)|) = \\ &= O(|FF(is)| \cdot |FI| + O(|FF(is)| \cdot |PP(is)| \cdot |IO(is)|)) \end{aligned} \quad (8)$$

For the overall assessment of the algorithm complexity according to formula (1), we note the following circumstances arising from the practice of business modeling. First, as a rule, the number of information systems  $|IS|$  is significantly less than the number of functions in business process  $|FF|$ . Second, information systems typically automate the different functions of a business process, whereby the intersection of the functions of different information systems can be considered negligible in comparison with the total number of functions in the system. This provides the basis for the assumption  $|FF(is_1) \cap FF(is_2)| \approx 0$  if  $is_1 \neq is_2$ . But with this assumption, the total number of loops repetition after the summation over all information systems will be comparable with the total number of the automated features of the business process, which with the high level of automation, is close to the total number of business process  $|FF|$  functions. Third, the performers number usually does not exceed the number of the performers names (with the delimitation of discretionary access, these values are equal). Therefore, we can write:

$$\sum_{is=1}^{|IS|} O(|FF(is)| \cdot |FI|) \quad (9)$$

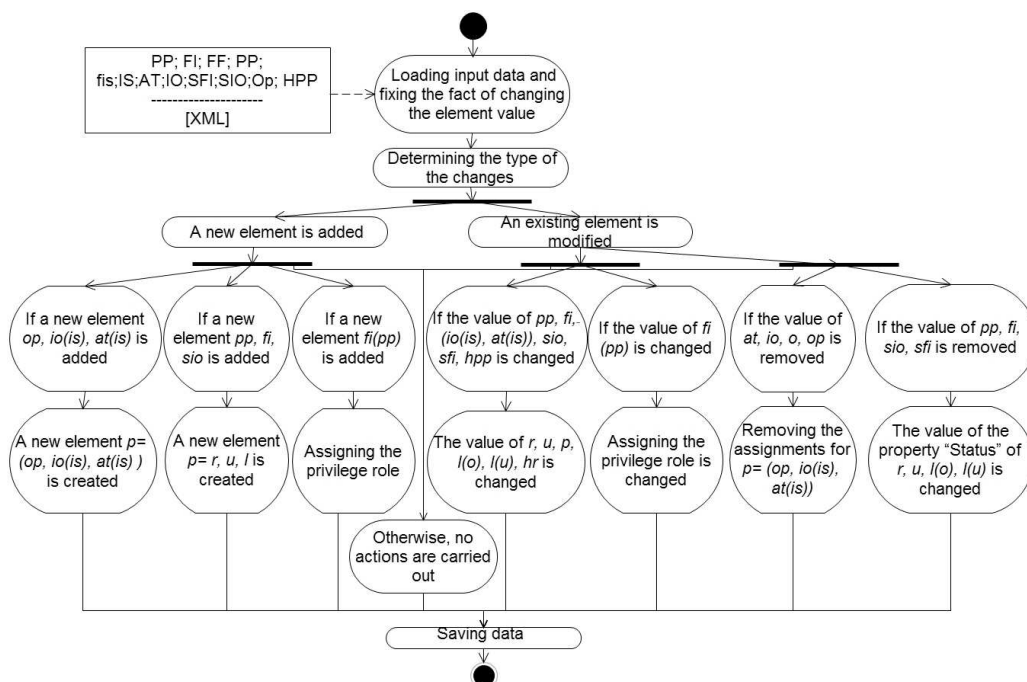
$$\sum_{is=1}^{|IS|} O(|FF(is)| \cdot |FI|) = O(|FF| \cdot |FI|) \sum_{is=1}^{|IS|} O(|FF(is)| \cdot |PP(is)| \cdot |IO(is)|) \quad (10)$$

which by substituting the formula (1) yields the final result:

$$T = \sum_{is=1}^{|IS|} T(is) \leq O(|FF| \cdot |FI|) + O(|FF| \cdot |FI| \cdot |IO|) = O(|FF| \cdot |FI| \cdot |IO|) \quad (11)$$

The main parameters that determine the dimension of this exhaustive search problem, are the functions number of the business process, the number of users (individuals) in the system and the number of access objects, whereas the algorithm for extracting data from the business process model for access control system has only a linear dependence on each of these parameters. We can note that this estimate is valid even in the worst case. Thus, the algorithm for the automated elements of the generalized model of access rights delimitation from the business process model falls into the category of effective algorithms.

After the initial formalization of access rights, there is the stage of their updating, carried out by comparing the state of the business process models before and after making any changes. Access rights updating shall be in accordance with the established classification of the company's activity [2]. For each classification attribute, the response rules have been defined. The algorithm for updating the rights of the access to the resources of automated information systems based on changes in the models of business processes is represented schematically in Fig. 4.



**Fig. 4.** Algorithm for updating the rights of the access to the resources of automated information systems on the basis of changes in business process models

The block of inputting data and fixing the fact of the change of a business process element value is carried out similarly to the block of data extraction from the business process model of the algorithm for the automated formation of the elements of the generalized model of access rights delimitation. The exhaustive search of all sets of access model elements is performed.

The complexity of the algorithm depends on the block of input data loading, the complexity of which has been defined previously.

On the basis of these algorithms, the information system of formalization and updating of access rights, which implements the approach based on the analysis of business processes, has been developed and registered in ROSPATENT. This system has been implemented in a number of businesses, and one of its purposes is to form access matrices within the establishing and operating systems of personal data protection information systems.

## References

1. Rodionova, Z. The technology of access rights change management on the basis of the business-processes analysis, vol. 1. Vestnik NSUEM, 2011.
2. Rodionova, Z. Information system of access control, vol.4-11, In the World of Scientific Discoveries, 2010.

# User Interfaces for Working with Thesauri and Rubricators in Distributed Heterogeneous Information Systems on the Platform ZooSPACE

S.A. Santeyeva<sup>1</sup> and O.L. Zhizhimov<sup>2</sup>

<sup>1</sup> Novosibirsk State University, Novosibirsk, Russia

<sup>2</sup> Institute of Computational Technologies SB RAS, Novosibirsk, Russia  
saya\_santeyeva@mail.ru, zhizhim@mail.ru

**Abstract.** This work describes technologies used to form user interfaces for working with thesauri, rubricators and ontologies in heterogeneous information environment. Discusses various aspects of building user interfaces for: navigation rubricators, thesauri and ontologies (hereinafter - the thesauri); search for and view articles thesauri; see the relationships between different thesauri; view dynamic links between thesauri and other databases with the possibility to manage the list of connected databases; view dynamically related records in other databases. As examples of the user interfaces of the subsystem ZooSPACE-W platform ZooSPACE.

To work with heterogeneous distributed information systems and databases, the user must provide the widest possible range of organizations search for information and view it in the form requested. One of the desirable elements of this spectrum is the possibility of navigating rubricators and thesauri with dynamic binding their articles (terms) to various databases which list must also be dynamic and can be configured by the user. With this in mind the heterogeneity of information sources requires decisions based on standardized components, such as protocols, access to the information resources used by the data schema, queries and formats the retrieve records. The need to create a software component that provides the required functionality and which has a sufficiently high level of generality, was realized with the development of appropriate graphic interfaces user platform ZooSPACE [3] (subsystem ZooSPACE-W). In accordance with the logic of work with information resources it is possible to allocate the following set of necessary graphical interfaces:

1. Interfaces that implement navigation on rubricators, thesauri and ontologies (hereinafter - the thesauri);
2. Interfaces that implement the functions of searching and browsing articles thesauri;
3. Interfaces that implement communication between different thesauri;
4. Interfaces implement dynamic links between thesauri and other databases with the possibility to manage the list of connected databases;
5. Interfaces that implement the view dynamically related records in other databases.

This paper describes the technological approaches and the implementation of the described set of interfaces for platform ZooSPACE.

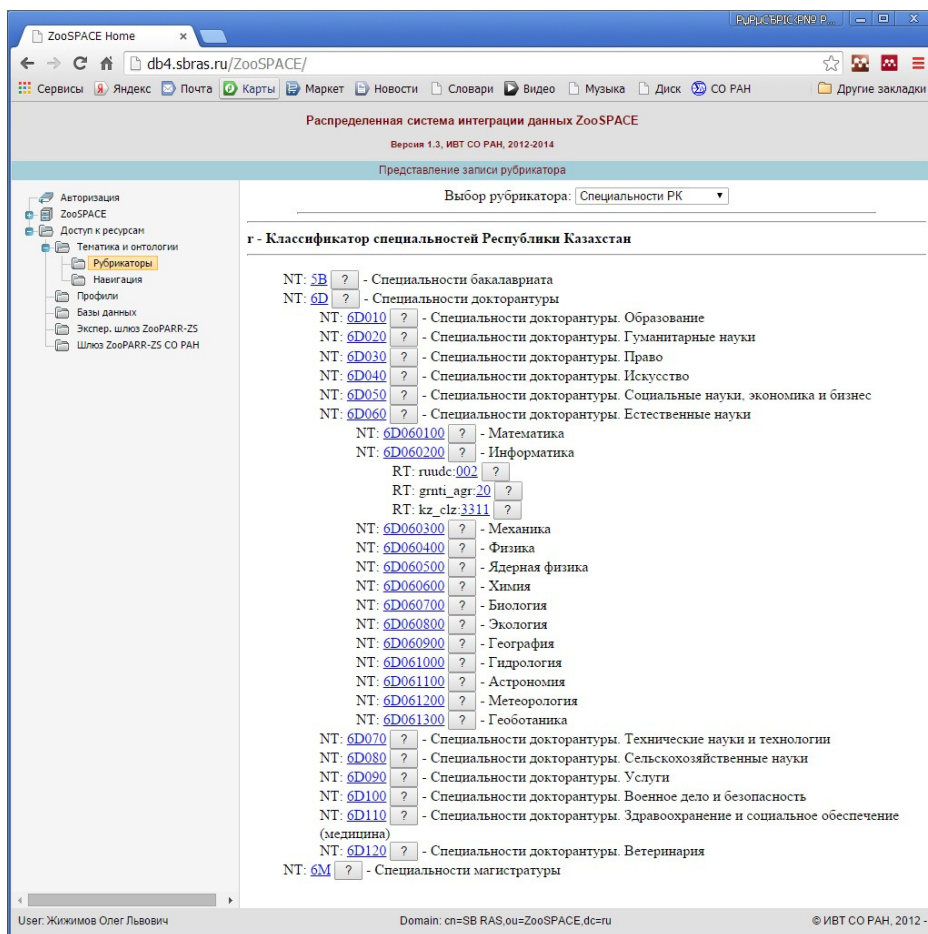


Fig. 1. To navigate by thesauri

As the platform ZooSPACE initially focused on working with information resources on the protocols Z39.50 Z39.50 [2]–[1], SRW and SRU [4] using a standard search query (RPN, PQF, CQL), a standard data schemas and standard of their presentation formats (XML, ISO2709, GRS-1) (see, e.g., [3,1]), as the basis for working with thesauri were selected:

1. Protocols of access to resources - Z39.50 and/or SRU;
2. Data schema thesauri - Zthes [5]–[20];
3. The format of the records thesauri - XML.

As a information resources for debugging algorithms and interfaces used

1. Database "Classification of specialties of higher and postgraduate education of the Republic of Kazakhstan for a three-level structure of training specialists" bachelor-master-doctorate PhD / on profile "
2. State Classifier of the Republic of Kazakhstan
3. Database "Rubricator SRSTI"
4. Database "Rubricator UDC"

All of these databases are structured in accordance with international and Russian standards (ISO 25964 [7], GOST 7.25-2001 [10], GOST 7.24-2007 [12]), are integrated into the platform ZooSPACE, which provides access by the Z39.50 protocol and / or SRU within scheme Zthes.

According to a profile Zthes to work with thesaurus used the following elements scheme:

The screenshot displays the ZooSPACE web application interface. The browser address bar shows the URL `db4.sbras.ru/ZooSPACE/`. The page title is "Распределенная система интеграции данных ZooSPACE" (Distributed data integration system ZooSPACE). The interface includes a navigation menu on the left with options like "Авторизация", "Доступ к ресурсам", and "Тематика и онтологии". The main content area shows a search query: `@attr ZTHES-attset 1=1 @attr 5=1 {6D}`. Below the search bar, a table displays search results:

Идентификатор	Заголовок источника данных	Записей	Найдено	Время, с	Ошибка	Просмотр
kz_spec	Рубрикатор специальностей РК	578	192	0.58		SRU

Below the table, the selected record is displayed in a detailed view. The record is identified as "Запись: 1 из 192" and is in "Обычное" (Normal) format. The thesaurus entry is:

Расширяющие термины: `g` `rus` Классификатор специальностей Республики Казахстан

Термин тезауруса: `6D` `rus` Специальности докторантуры

Уточняющие термины:

- `6D010` `rus` Специальности докторантуры. Образование
- `6D020` `rus` Специальности докторантуры. Гуманитарные науки
- `6D030` `rus` Специальности докторантуры. Право
- `6D040` `rus` Специальности докторантуры. Искусство
- `6D050` `rus` Специальности докторантуры. Социальные науки, экономика и бизнес
- `6D060` `rus` Специальности докторантуры. Естественные науки
- `6D070` `rus` Специальности докторантуры. Технические науки и технологии
- `6D080` `rus` Специальности докторантуры. Сельскохозяйственные науки
- `6D090` `rus` Специальности докторантуры. Услуги
- `6D100` `rus` Специальности докторантуры. Военное дело и безопасность
- `6D110` `rus` Специальности докторантуры. Здравоохранение и социальное обеспечение (медицина)
- `6D120` `rus` Специальности докторантуры. Ветеринария

The interface footer shows the user "User: Жижимов Олег Львович" and the domain "Domain: cn=SB RAS,ou=ZooSPACE,dc=ru".

Fig. 2. Search and view articles thesauri

Table 1.

1	Zthes/term/termId	local identifier term
2	Zthes/term/termName	term name
3	Zthes/term/termNote	term note
4	Zthes/term/termQualifier	identifier term
5	Zthes/term/termType	term type
6	<b>Zthes/term/relation</b>	<b>element relation to other articles thesaurus</b>
7	Zthes/term/relation/termID	local identifier related term
8	Zthes/term/relation/relationType	C,relation type
9	Zthes/term/relation/termName	name of the related term
10	Zthes/term/relation/sourceDB	URI other thesaurus
11	<b>Zthes/term/postings</b>	<b>element related with other databases</b>
12	Zthes/term/postings/sourceDb	URI external databases
13	Zthes/term/postings/fieldName	index name external database
14	Zthes/term/postings/hitCount	amount of records in an external database

1. To navigate by thesauri (see Fig. 1) used scheme elements 1-9(see Tab. 1). In this processing the following relationship between the terms: BT- relationship the related term is more general than the current term, NT- relationship with related term, USE- relationship with term, that is the related term should be used in preference to the current term, UF- the current term should be used in preference to the related term, USE and RT- relationships, which determined related term.
2. To search and view articles thesaurus (see Fig. 2) using standard interfaces ZooSPACE (Z39.50 protocol and SRU, requests CQL, RPN and PQF, formats XML and the GRS-1).
3. To view the relationship between articles of different thesauri processed static elements of the schema 10 (see. Tab. 1). It uses the information present in the databases about relationships between articles. Visually, the hierarchical structure of the thesaurus is dynamically supplemented by the according branch related thesaurus, addressed which initiates an asynchronous request to the database related thesauruses with retrieval and visualization of the according related terms with all of its secondary relations.
4. The interfaces are implemented by dynamic relations between thesauri and other databases with the ability to manage a list of connected databases based on the use of elements 11-14 schema data Zthes (see Tab. 1). These elements are dynamic - they are formed at the time of processing the according request to display articles thesaurus using a user - formed list of information resources (databases) through interfaces ZooSPACE. At the time of formation of the structure records articles of the thesaurus adressed (via Z39.50 or SRU) to external databases and get the amount of records for according articles thesaurus - content elements 14 (hitCount) records Zthes (see Fig. 4). Request to an external database is implemented as a request to search for records with the specified code rubricators(code GRNTI, UDC, etc.), but can be implementation algorithms with more complex search queries.
5. Interfaces implementing view dynamic related records of other databases fully according interfaces view records, implemented at the ZooSPACE-W (see Fig. 5) (see also[3]).

It should be noted that at the time the described algorithms are implemented in the form of a working prototype of an individual module, which is part of the server subsystems ZooSPACE-W.

In conclusion it should be noted that the described user interfaces significantly extend the functionality of the client software for access to distributed heterogeneous information resources.

The screenshot shows the ZooSPACE web application interface. The browser address bar displays `db4.sbras.ru/ZooSPACE/`. The page title is "Распределенная система интеграции данных ZooSPACE" with version "1.3, ИБТ СО РАН, 2012-2014". The main content area is titled "Представление записи рубрикатора" and shows a dropdown menu for "Выбор рубрикатора" set to "Специальности РК". Below this is the section "г - Классификатор специальностей Республики Казахстан".

The list of specialties includes:

- NT: 5B ? - Специальности бакалавриата
- NT: 6D ? - Специальности докторантуры
- NT: 6D010 ? - Специальности докторантуры. Образование
- NT: 6D020 ? - Специальности докторантуры. Гуманитарные науки
- NT: 6D030 ? - Специальности докторантуры. Право
- NT: 6D040 ? - Специальности докторантуры. Искусство
- NT: 6D050 ? - Специальности докторантуры. Социальные науки, экономика и бизнес
- NT: 6D060 ? - Специальности докторантуры. Естественные науки
- NT: 6D070 ? - Специальности докторантуры. Технические науки и технологии
- NT: 6D070100 ? - Биотехнология (по отраслям и областям применения)
- NT: 6D070200 ? - Автоматизация и управление
- RT: ruide:004 ?
- NT: 004.2 ? - Архитектура вычислительных машин
- NT: 004.3 ? - Вычислительная техника. Вычислительные машины. Аппаратные средства. Специальные определители
- NT: 004.4 ? - Программные средства
- NT: 004.5 ? - Человеко-машинное взаимодействие. Пользовательский интерфейс
- NT: 004.6 ? - Данные
- NT: 004.7 ? - Общение компьютеров. Сети ЭВМ. Вычислительные сети
- NT: 004.8 ? - Искусственный интеллект
- NT: 004.9 ? - Прикладные информационные (компьютерные) технологии
- NT: 004.01 ? - Документация
- NT: 004.02 ? - Методы решения задач
- NT: 004.03 ? - Типы и характеристики систем
- NT: 004.04 ? - Ориентация вычислительного процесса
- NT: 004.05 ? - Качество систем и программ
- NT: 004.07 ? - Характеристики памяти
- NT: 004.08 ? - Средства ввода, вывода и хранения данных
- RT: kz\_clz:2132 ?
- RT: grnti\_agr:50 ?
- NT: 6D070300 ? - Информационные системы (по отраслям)
- NT: 6D070400 ? - Вычислительная техника и программное обеспечение
- NT: 6D070500 ? - Математическое и компьютерное моделирование
- NT: 6D070600 ? - Геология и разведка месторождений полезных ископаемых
- NT: 6D070700 ? - Горное дело
- NT: 6D070800 ? - Нефтегазовое дело

The footer of the page shows "User:" and "Domain: ou=ZooSPACE,dc=ru" on the left, and "© ИБТ СО РАН, 2012 -" on the right.

Fig. 3. View articles related thesauri вЂ“ conversion from specialty of RK to UDC



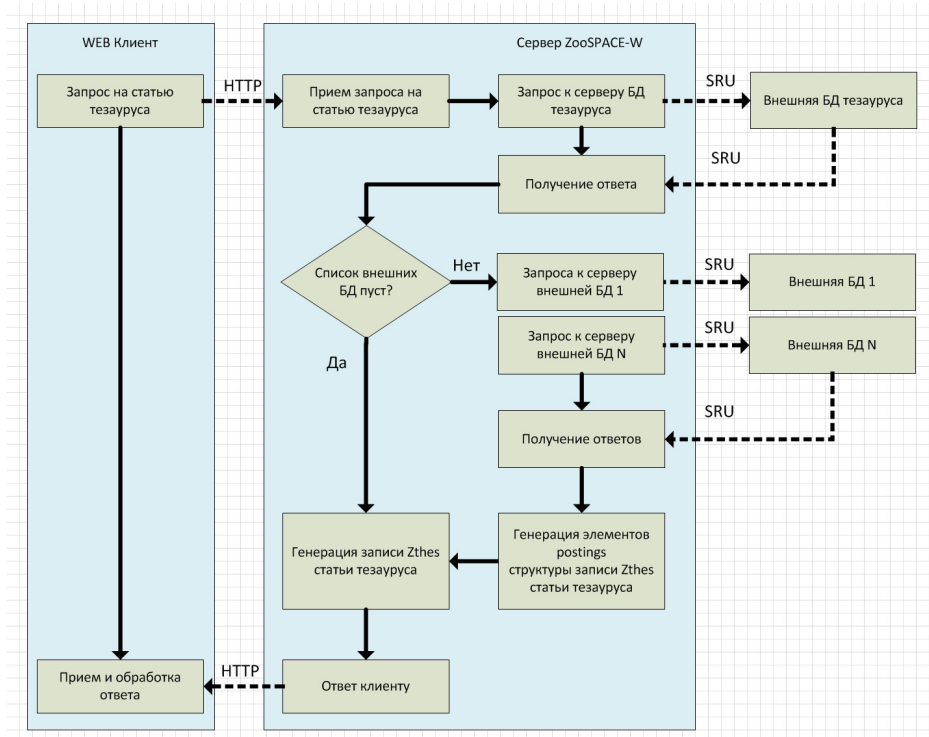


Fig. 4. Formation the structure records articles of the thesauri with dynamic relations to external databases

Идентификатор	Заголовок источника данных	Записей	Найдено	Время, с Ошибка	Просмотр
AB	РЖ Автоматика	1594630	6	0.558	SRU
AC	РЖ Астрономия	445759	0	1.093	

Запись: 2 из 6 Представление: Обычное Формат: XML Схема: F

Коды: ГРНТИ: 50.33.39; УДК: 681.325

Авторы: Бродский И.И.; Козлачков В.А.; Коршевер И.И.; Нестерихин Ю.Е.; Павлов С.А.; Ремель И.Г.

Заглавие: Высокопроизводительный периферийный векторный процессор А-12

Реферат: Представлен высокопроизводительный периферийный векторный конвейерный процессор А-12, разработанный в ИАиЭ СО АН СССР в 1981 г. Подключение такого процессора к малой ЭВМ открывает новые возможности создания автономных ВС, предназначенных для науч. расчетов, и систем реального времени. Описаны архитектура процессора, элементы его программного обеспечения, различные конфигурации, в которых процессор м.б. использован в архитектуре ВС высокой производительности. Ил. 3. Библ. 13.

Ключевые слова: ПРОЦЕССОРЫ; КОНВЕЙЕРНЫЕ УСТРОЙСТВА; ВЕКТОРНЫЕ УСТРОЙСТВА; АРХИТЕКТУРЫ; ЭЛЕМЕНТЫ; ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ;

Включено в: Автотметрия. (ISSN 0320-7102). - 1984. - N 4. - С. 29-35.

Fig. 5. View related records from external sources

## References

1. Zhizhimov O.L., Fedotov A.M., Shokin Yu.I. Technology platform mass integration of heterogeneous data // Bulletin of the Novosibirsk State University. – Series: Information Technology, 2013. Т.11. В.,– 1. pp. 24–41. ISSN 1818-7900.
2. ISO 23950:1998 – Information and documentation – Information retrieval(Z39.50) – Application service definition and protocol specification [electronic resource]. – 1998. – URL: [http://www.iso.org/iso/catalogue\\_detail.htm?csnumber=27446](http://www.iso.org/iso/catalogue_detail.htm?csnumber=27446)
3. Zhizhimov O.L. Introduction Z39.50: 4-st rewrite and addition. – Novosibirsk: NSRSB, 2003. –262 p.
4. SRU – Search/Retrieve via URL.The Library of Congress [electronic resource]. – 2013. – URL: <http://www.loc.gov/standards/sru/>
5. The Zthes specifications for thesaurus representation, access and navigation [electronic resource]. – 2006. – URL: <http://zthes.z3950.org/>
6. Mazov N.A., Zhizhimov O.L. Unification of building and accessing thesauri and classification schemes in distributed information systems by protocol Z39.50 // II All-Russian Scientific Conference В«Digital Libraries: Advanced Methods and Technologies, Digital Collections В» - RCDL'2000 (РЦС'ЪРҢС,ПИР'ПҢРҢ, Russia, 26.09 - 28.09.2000): Collections of works conference. – Protvino: GNC IFVE, 2000. pp. 230-233.
7. 25964–1:2011–Information and documentation – Thesauri and interoperability with other vocabularies – Part 1: Thesauri for information retrieval [electronic resource]. – 2011. – URL: <https://www.iso.org/obp/ui/ru/#iso:std:53657:en>
8. GOST 7.25-2001- system of standards on information, librarianship and publishing. Monolingual thesaurus for information retrieval. Rules for its development, structure, composition and form of presentation, – Minsk: IPK Publisher standards 2002. p 16.
9. GOST 7.74-96 SIBID. Information retrieval languages. Terms and definitions - // Minsk: IPK Publisher standards, 1996. 38 p.

# A Case Study of a Knowledge Management System

Ana Savic<sup>1</sup>, Edlira Kalemi<sup>2</sup>, and Mirjeta Dëra<sup>2</sup>

1 - School of Electrical Engineering and Computer Science, Belgrade , 2 - University of Tirana

**Abstract.** Albanian key Cross is a humanitarian association that works to improve the lives and dignity of people in need, in accordance with basic principles of motion. He is assistant to the authorities for humanitarian affairs nationwide. Albanian Red Cross today is a "wealth" of the very real humanitarian values of the Albanian civil society. This property is present not only as a concept and existence of traditional and prestigious idea into action, but also with its institutional structures, leadership and management. Our study focuses on the problems that appear in the Association of the Red Cross on the functioning of the department and how those can be solved or made easier to manage through a Knowledge Management System (KMS). Though the construction of ontology we managed to solve the problem that existed between departments. Full implementation of Microsoft Dynamic GP reach to improve and give solutions malfunction of the Albanian Red Cross.

**Keywords:** Knowledge Management System, Ontology, Microsoft Dynamic GP.

## 1 Introduction

Albanian Red Cross is the oldest humanitarian organizations in Albania. It was established on 4 October 1921. Since this year is a member of the Federation of Societies of Red Cross and Red Crescent. This organization is part of the International Movement of the Red Cross and Red Crescent today counts 187 organizations worldwide. It is a voluntary humanitarian organization that operates independently on the basis of fundamental principles of international humanitarian Movement of the Red Cross. The organization is a subsidiary of the authorities for humanitarian affairs nationwide. This organization aims to be a humanitarian association developed, which works to improve the lives and dignity of people in need, in accordance with the Fundamental Principles of the International Red Cross and Red Crescent. Albanian Red Cross counts 39 branches spread across the country, which operate humanitarian in their district. He has achieved a good development, becoming the largest humanitarian organizations in the country and the largest donor of blood donation in Albania. Albanian Red Cross helps, through its humanitarian activities, the most vulnerable people by mobilizing the power of humanity to society.

Currently has as main purpose and acting in fulfillment of the Strategic Plan 2010-2015 in the areas of dissemination of humanitarian values, providing enough blood in all state hospitals in Albania, preparing for more effective interventions in disaster relief, health education and strengthening the health of the population, meeting some basic human needs and social care delivery, organizational development and increase the sustainability of the organization, based on its internal resources. System Albanian Red Cross currently uses five programs for the management and maintenance of all activities to achieve the objectives and goals. These systems are:

1. Finance department program that records all movements in relation to the ark, bank, salaries, state of warehouses, costs and revenues by donor for all Red Cross branches across the country.
2. The blood donation program becomes determining where blood will be collected and where will happen organization in order to be communicated to all concerned for voluntary blood donation.

3. First aid program where the Albanian Red Cross is able to organize training courses for giving first aid to the general public, training courses for instructors and volunteers.
4. The program for the registration of members and volunteers of the organization related to cycle to follow the volunteers during their assignment at the NRC, which is related to the recruitment, training, motivation, commitment and removal of volunteers and maintains data about with registered members and volunteers in the organization.
5. Search and Rescue program keeps information about the person sought and the applicant of that person. [10]

Humanitarian organizations Albanian Red Cross has an internal organization divided into several departments responsible for several key areas to achieve its goal, but that often has difficulties the current system. Some of the problems of the system of the Albanian Red Cross are high operational costs, low productivity, high costs of storage and maintenance, poor integration of donor management, providing information not in real time and proper uncoordinated between departments considered Key shortcomings which slow down daily operations, excess inventory and unified system as across the country. In general will talk about knowledge management systems, what are they and how operate to solve a problem, what solution offers this system. Second will also handle general ontologies and ontology built for the Albanian Red Cross. Finally we give our solution to the system of the Albanian Red Cross is implementing Microsoft Dynamics GP which improves and solves the problems mentioned above and conclusions.

## 2 Knowledge Management System and Ontologies

Knowledge management systems refer to any kind of IT system that stores and retrieves knowledge, improves collaboration, locates knowledge sources, mines repositories for hidden knowledge, captures and uses knowledge, or in some other way enhances the KM process. [9] "Knowledge management is a discipline that promotes an integrated approach to identifying, capturing, evaluating, retrieving, and sharing all of an enterprise's information assets. These assets may include databases, documents, policies, procedures, and previously un-captured expertise and experience in individual workers."(Michael E. D. Koenig)

Purpose of Knowledge Management System are:

- Improved performance
- Competitive advantage
- Innovation
- Sharing of knowledge
- Integration

Activities in Knowledge Management are:

- Start with the business problem and the business value to be delivered first.
- Identify what kind of strategy to pursue to deliver this value and address the KM problem.
- Think about the system required from a people and process point of view.
- Finally, think about what kind of technical infrastructure are required to support the people and processes.
- Implement system and processes with appropriate change management and iterative staged release.

In recent years the development of ontologies explicit formal specifications of the terms in the domain and relations among them (Gruber 1993) has been moving from the realm of Artificial-Intelligence laboratories to the desktops of domain experts. Ontologies have become common on the World-Wide Web. An ontology defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them.

Why would someone want to develop an ontology? Some of the reasons are:

- To share common understanding of the structure of information among people or software agents (Musen 1992; Gruber 1993).
- To enable reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge (McGuinness and Wright 1998).
- To analyze domain knowledge (McGuinness et al. 2000).

An ontology is a formal explicit description of concepts in a domain of discourse (classes (sometimes called concepts)), properties of each concept describing various features and attributes of the concept (slots (sometimes called roles or properties)), and restrictions on slots (facets (sometimes called role restrictions)). An ontology together with a set of individual instances of classes constitutes a knowledge base. In reality, there is a fine line where the ontology ends and the knowledge base begins.

Classes are the focus of most ontologies. Classes describe concepts in the domain. A class can have subclasses that represent concepts that are more specific than the superclass.

In practical terms, developing an ontology includes:

- defining classes in the ontology,
- arranging the classes in a taxonomic (subclass–superclass) hierarchy,
- defining slots and describing allowed values for these slots,
- filling in the values for slots for instances.

We can then create a knowledge base by defining individual instances of these classes filling in specific slot value information and additional slot restrictions.

### 3 The implementation of Microsoft Dynamics GP

The Albanian Red Cross provides disaster relief and emergency services for thousands of people every year. In addition to responses to disasters, organizations serving local communities in education, training, and support materials that prepare volunteers to respond to disasters and other emergency life-threatening. The organization also helps the community to prevent, prepare for and respond to emergencies through training programs and coordinated volunteer efforts. To support its mission, in order to reduce financing needs reducing costs through streamlining of operations, including financial and storage processes. Currently working with Microsoft Office package we suggest the implementation of Microsoft Dynamics GP to achieve more efficient processes and to provide more extensive information on the current situation of the organization. With the implementation of Microsoft Dynamics GP is believed to dramatically increase productivity, the annual costs of storage sit and sit down new requirements for additional staff. We suggest that within the next year Albanian Red Cross to replace its current system of information with Microsoft Dynamics GP, which provides greater integration of donor

management and its other programs, to provide information in real time the state of the organization in all departments. Besides the implementation of Microsoft Dynamics GP 10.0 to improve internal operations, the initiative includes the implementation of Microsoft Dynamics CRM to manage and management training volunteers to help recruit volunteers more effectively in the event of national or regional disaster to help in emergency situations. The benefits from the implementation of Microsoft Dynamics GP are cost reduction, increased focus on core mission, updating a central system and provides efficiency in reporting.

#### 4 The solution offered

Because Microsoft Dynamics GP is similar and works with applications in the Microsoft Office system, employees can use more efficiently. They can pass information from applications in the Microsoft Office system directly into Microsoft Dynamics GP. This new system implemented requires service staff, department of finance and fundraising department should automatically start working together as soon as they begin to see the movement of data from their desktop to the desktop of everyone else. In this way the data are available to them at any time they want if they have the appropriate permissions. Microsoft Dynamics GP allows administrators to have more control over business processes. They can create approval processes so donations system routes pass through the appropriate channels with automatic messages that appear in desktop of employees. For example, employees must go through the proper channels before you add data to a donor the database of donors. This increases accountability and makes the organization less charge. By implementing the new system, employees will avoid unnecessary tasks. In the case of negotiation of contracts (for the purchase of relief from the proceeds of donations), staffers use the system reports to determine the benefits and the amount of goods needed to ensure agreement in accordance with the rules of the tender. Employees can also create minimum inventory levels in Microsoft Dynamics GP so they automatically receive notifications in Office Outlook when supplies are below this level. Also Microsoft Dynamics GP allows sending automatic notifications to the person responsible for the management of blood storage, which eliminates the destruction documents and sends blood units has expired.

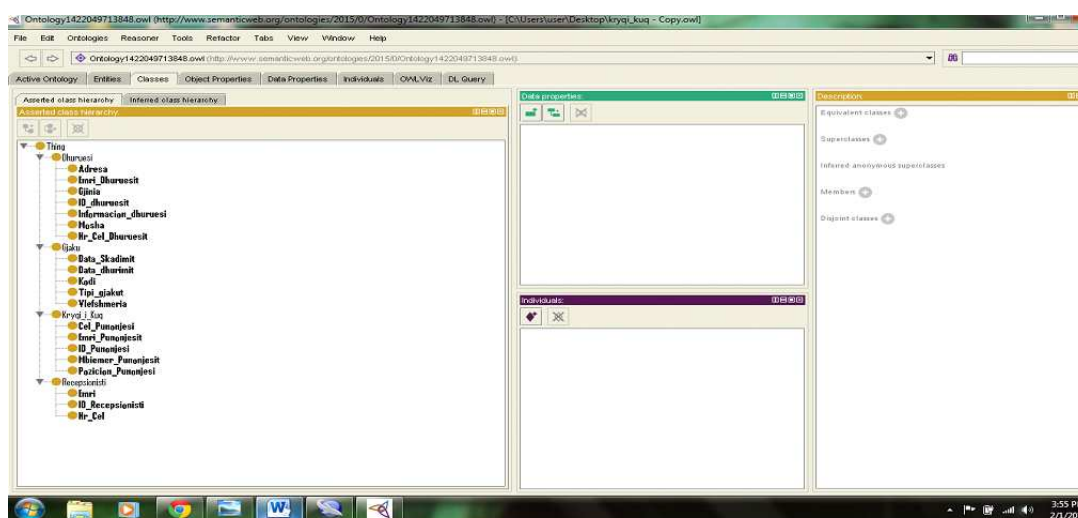


Fig. 1. Ontology of the Albanian Red Cross.

## 5 The system of blood donation

Blood donation system is one of the most important Albanian Red Cross and as such should be given special attention. For this reason the new system upgrade that we thought we implement aims to improve and make more efficient the current system. This system should function as shown below.

Red Cross should enhance interactivity with state hospitals across the country. The amount of blood for every blood type sent to these hospitals should be in accordance with the specific requirements of each hospital. As compared to the current system where whole blood collected is distributed between hospitals without a criterion of need, this kind of new system aims efficient distribution of blood units not wasted any unit.

Blood unit has a limited lifespan of six weeks in which to stay on hold. Therefore, the new system aims at determining the quantity needed by each hospital. Under the system improved after the admittance of blood, the latter is stored in the blood bank of the Red Cross. The unit of blood destruction takes daily a list of those entities who spent their lives waiting, they collapsed and updated with current state store date available.

Blood Center also distributes blood to all state hospitals who require blood. Requests usually come for a specific type of blood. Blood bank refrigerators Blood prepares and distributes the amount of blood when they come to pick up the load. Blood Bank takes a list for each hospital and specific units of blood to every hospital must provide the processing unit.

## 6 Conclusions

Albanian Red Cross is the only humanitarian organization helping people. It helps, through its humanitarian activities, the most vulnerable people by mobilizing the power of humanity to society. Humanitarian organizations Albanian Red Cross has an internal organization divided into several departments responsible for several key areas to achieve its goal. One important factor that enhances the performance and effectiveness of each department is communication and coordination that these departments have with each other. Given that this institution provides assistance in disaster or calamities, the current program has used them often complicates the problem and help you need to give as well as organizations serving local communities in education, training, and support materials that prepare volunteers to respond to disasters and other emergency life-threatening. We suggested and implemented Microsoft Dynamic GP system, which facilitates problems and is very effective. Through ontology system was built for the Albanian Red Cross, which was made possible link between blood needs according claims or state hospitals and other institutions'. Also was established and the possibility of communication between donors and assistants or doctors to facilitate and reduce the excessive duration and high costs. Through Microsoft Dynamics GP organizations managed to reduce inventory levels, which increased the space available and allow employees to spend less time moving donations in stock, issued by placing boxes or on shelves or refrigerators and sat holding larger quantities of blood units currently needed. Microsoft Dynamics GP increased concentration of employees to prepare reports and activities and will reduce the amount of time employees spend on administrative tasks. Also helped to eliminate inefficient processes so that our employees to spend more time to organize volunteer training and in meeting other important tasks. IT team is now able to update the system at central and spend less time on maintenance.

## References

1. Kephart, Jeffrey O., Chess, David M., *The Vision of Autonomic Computing*, IEEE Computer society (2003).

2. Russell B., *Artificial Intelligence: a modern approach*, Wiley, (2002).
3. Kephart, <http://www.kmworld.com/Articles/Editorial/What-Is-.../What-is-KM-Knowledge-Management-Explained-82405.aspx>
4. [http://protege.stanford.edu/publications/ontology\\_development/ontology101.pdf](http://protege.stanford.edu/publications/ontology_development/ontology101.pdf)
5. Subashi.Sh, *Red Cross ninety years humanism* (2011).
6. [http://www.kksh.org.al/editor\\_files/file/content/Revista%2049\\_v3.pdf](http://www.kksh.org.al/editor_files/file/content/Revista%2049_v3.pdf)
7. Roussey.C, Pinet.F, Kang.M and Corcho.O, *An Introduction to Ontologies and Ontology Engineering*
8. <http://www.kksh.org.al/>
9. <http://www.knowledge-management-tools.net/knowledge-management-systems.html>
10. [www.tutorialspoint.com/management\\_information\\_system/knowledge\\_management\\_systems.htm](http://www.tutorialspoint.com/management_information_system/knowledge_management_systems.htm)
11. <https://www.microsoft.com/en-us/dynamics/erp-gp-overview.aspx>



# Performance Analysis of Wireless Transmission Channels in the Presence of Eta-Mu Fading and Kappa-Mu Co-Channel Interference

Darko Vučković<sup>†</sup>, Stefan Panić<sup>\*</sup>, Hranislav Milošević<sup>\*</sup>, and Danijel Djošić<sup>\*</sup>

<sup>†</sup>-Faculty of Technical Science, University of Priština,  
Knjaza Miloša 7, 38320 Kosovska Mitrovica

<sup>\*</sup>-Faculty of Science and Mathematics, University of Priština,  
Lole Ribara 29, 38320 Kosovska Mitrovica

**Abstract.** Mathematical characterization of various types of propagation environments and accompanying transmission phenomena has arisen as a major task in the processes of computing and evaluating the performances of modern generation wireless communication systems services. In this paper we will analyze standard performance criteria of wireless transmission, when communication is carried out over eta-mu faded channels, in the interference limited environment, with total co-channel interference influence modelled with kappa-mu model. Infinite series expressions will be derived for the probability density function (PDF) and cumulative distribution function (CDF) of received signal-to-interference ratio (SIR). Capitalizing on these expressions, outage probability (OP) and average bit error probability (ABEP), will be efficiently evaluated, graphically presented and discussed in the function of transmission parameters. Finally, possible performance improvement will be considered through the possibility of SC diversity reception appliance. Capitalizing on analysis and evaluations presented, system designers could perform trade-off studies among the various modern communication system parameters in order to determine the optimal choice in the presence of their available constraints.

**Keywords:** performance computation, wireless transmission, co-channel interference, selection combining reception, fading models.

## 1 Introduction

Emerging development of modern generation wireless communication applications and services, provides constant necessity of analyzing the possibility for their performance improvement. However, wireless transmission is often accompanied by various side effects such as multipath fading and co-channel interference (CCI), which limit system performances, spectral efficiency and channel capacity. Enhanced necessity for computing and evaluation of the performances of these wireless communication services, provides a need for mathematical characterization of observed transmission phenomena [1]. Any wireless system tends to conserve the existing disposable spectrum by reusing allocated frequency channels in areas that are geographically located as close to each other as possible. However, the amount of CCI that occurs determines constraints in distance for reusing wireless channels [2]. CCI is defined as the interfering signal from another source that has the same carrier frequency as the useful information signal. So, the main goal is then to determine how CCI affects well-accepted criteria of performance of wireless systems services, such as outage probability (OP) and average bit error probability (ABEP), in order to implement practical system realizations with satisfied predetermined minimum performance levels.

Different fading models for describing desired wireless signal variations in the presence of multipath fading are presented in the literature. The  $\eta - \mu$  model, general fading model has been presented in [3]. This model is expressed in the function of parameters:  $\mu$  and  $\eta$ , corresponding to

number of multipath clusters through which observed signal propagates, and the ratio of powers of independent processes, which form the resulting fading process, respectively. When parameter  $\mu$  takes the value of  $\mu = 0.5$ , this model reduces to Nakagami- $q$  model, with the  $q$  parameter corresponding to  $\mu/2$ . Nakagami- $m$  model can be also approximated when  $\eta \rightarrow 0$  and  $\eta \rightarrow \infty$ , with  $m$  parameter corresponding to  $\mu$  [4].

In [5], it has been shown that sum of arbitrary number of interfering signals, can be efficiently modelled with  $\kappa - \mu$  signal model. This model encompasses scenarios when each of interferers could have not only scattered waves with equal powers, but also a dominant component with arbitrary power. In such case parameter  $\kappa$  is related to the total dominant/scattered components powers ratio of all interfering signals, while parameter  $\mu$  is related to the total number of interferers.

## 2 System model

Desired information signal with a  $\eta - \mu$  distributed random amplitude process can be presented by [3]:

$$f_R(R) = \frac{4\sqrt{\pi}h_d\mu_d^{1/2+\mu_d}R^{2\mu_d}}{H_d^{\mu_d-1/2}\Gamma(\mu_d)\Omega_d^{1/2+\mu_d}} \exp\left(-\frac{2\mu_d h_d R^2}{\Omega_d}\right) \times I_{(\mu_d-1/2)}\left(\frac{2\mu_d H_d R^2}{\Omega_d}\right) \tag{1}$$

with  $\Omega_d = E[R^2]$ , denoting the desired signal average power, while  $I_n(x)$  is the  $n$ -th order modified Bessel function and  $\Gamma(a)$  denotes Gamma function [8, Eq.8.310.1].  $H_d$  and  $h_d$  are desired signal parameters, written in the function of parameter  $\eta_d$  as [3]:

$$H_d = \frac{\eta_d^{-1} - \eta_d}{4} \quad h_d = \frac{2 + \eta_d^{-1} + \eta_d}{4}; \tag{2}$$

In a similar manner, resultant interfering signal can be presented as [5]:

$$f_r(r) = \frac{2^{\mu_c} \kappa_c^{\frac{(1-\mu_c)}{2}} (\kappa_c + 1)^{\frac{(\mu_c+1)}{2}} r^{\mu_c}}{\exp(\kappa_c \mu_c) \Omega_c^{\frac{(\mu_c+1)}{2}}} \exp\left(-\frac{(\kappa_c + 1)\mu_c r^2}{\Omega_c}\right) \times I_{(\mu_c-1)}\left(2\mu_c \sqrt{\frac{\kappa_c(\kappa_c + 1)r^2}{\Omega_c}}\right) \tag{3}$$

with  $\Omega_c = E[r^2]$ , denoting the CCI signal average power.

If the instantaneous SIR,  $\lambda$ , is defined as  $\lambda = R^2/r^2$ , while average SIR,  $S$ , defined as  $S = \Omega_d/\Omega_c$ , then by using the relation [6]:

$$f_\lambda(\lambda) = \frac{1}{2\sqrt{\lambda}} \int_0^\infty f_R(r\sqrt{\lambda}) f_r(r) r dr \tag{4}$$

$$f_\lambda(\lambda) = \sum_{p=0}^\infty \sum_{l=0}^\infty \frac{2\sqrt{\pi} H_d^{2p} \mu_d^{2\mu_d+2p} \mu_c^{\mu_c+2l} \kappa_c^l (1 + \kappa_c)^{\mu_c+l+1} \Gamma(2\mu_d + \mu_c + 2p + l)}{2 \exp(\mu_c \kappa_c) \Gamma(\mu_d) \Gamma(\mu_d + p + 1/2) \Gamma(\mu_c + l) p! l!} \times \frac{\lambda^{2p+2\mu_d-1} S^{\mu_c+l}}{(\mu_c(1 + \kappa_c)S + 2\mu_d h_d \lambda)^{2\mu_d+\mu_c+2p+l}} \tag{5}$$

The double infinity sum in (5) converge rapidly, since only about 20-30 terms should be summed in order to achieve accuracy at 5th significant digit for various values of corresponding system parameters. PDF of instantaneous SIR for various values of system parameters is presented at Figure 1. Capitalizing on (5), and by using the same mathematical transformations as in [7], closed-form expression for the cumulative distribution function (CDF) of the instantaneous SIR can be presented as:

$$\begin{aligned}
 F_{\lambda}(\lambda) &= \int_0^{\lambda} f_{\lambda}(t) dt \\
 &= \sum_{l=0}^{\infty} \sum_{p=0}^{\infty} \frac{\sqrt{\pi} H_d^{2p} \mu_d^{2\mu_d+2p} \mu_c^{\mu_c+2l} \kappa_c^l (1 + \kappa_c)^{\mu_c+l+1} \Gamma(2\mu_d + \mu_c + 2p + l)}{2 \exp(\mu_c \kappa_c) \Gamma(\mu_d) \Gamma(\mu_d + p + 1/2) \Gamma(\mu_c + l) p! l!} \\
 &\quad \times B\left(2j + 2\mu_d, 2k + 2\mu_c, \frac{2\mu_d h_d \lambda}{(\mu_c(1 + \kappa_c) S + 2\mu_d h_d \lambda)}\right)
 \end{aligned} \tag{6}$$

with  $B(a, b, z)$  denoting the well-known incomplete Beta function [8, eq. (8.391)].

### 2.1 Selection Combining

Space diversity reception is an efficient tool for improving transmission reliability, which is based on providing the receiver with multiple faded replicas of the same desired signal [1]. In that way transmission power and bandwidth increase is avoided. Simplest diversity reception, that process only one of the diversity branches is SC reception. SC diversity chooses and outputs the branch with the largest instantaneous SIR (or SNR or signal level), namely,  $\lambda_{out} = \max(\lambda_1, \lambda_2, \dots, \lambda_N)$ , with  $\lambda_i$  denoting the instantaneous value of SIR at  $i$ -th received branch. This type of SC in which the branch with the highest SIR is selected, can be measured in real time both in base stations and in mobile stations using specific SIR estimators as well as those for both analog and digital wireless systems [9]. Since co-phasing of multiple branches is not required, SC can be used in conjunction with either coherent or differential modulation. Calculating CDF of  $N$ -branch SIR-based SC output can be carried out according to [10][11]:

$$F_{\lambda}(\lambda) = F_{\lambda_1}(\lambda) F_{\lambda_2}(\lambda) \dots F_{\lambda_N}(\lambda) = \prod_{i=1}^N F_{\lambda_i}(\lambda) \tag{7}$$

where corresponding CDF's for the uncorrelated input branches are defined with Eq.(6). Similarly calculating PDF of  $N$ -branch SIR-based SC output can be carried out according to [10]:

$$f_{\lambda}(\lambda) = \frac{dF_{\lambda}(\lambda)}{d\lambda} = \sum_{j=1}^N \left( f_{\lambda_j}(\lambda) \prod_{i=1, i \neq j}^N F_{\lambda_i}(\lambda) \right) \tag{8}$$

where corresponding PDF's for the uncorrelated input branches are defined with Eq.(5).

## 3 Performance computing

### 3.1 Outage probability

Outage probability (OP) is performance measure often used for controlling the CCI level. This measure is crucial in helping the wireless communications system designers to meet the quality of service (QoS) and grade of service (GoS) demands [12]. In interference-limited environment,

OP is defined as the probability, that the output SIR falls under defined protection ratio, which depends on applied modulation technique and expected QoS [1]:

$$P_{out} = F_{\lambda}(\lambda < \gamma_{th}) = \int_0^{\gamma_{th}} f_{\lambda}(t)dt = F_{\lambda}(\gamma_{th}) \tag{9}$$

### 3.2 Average Bit Error Probability

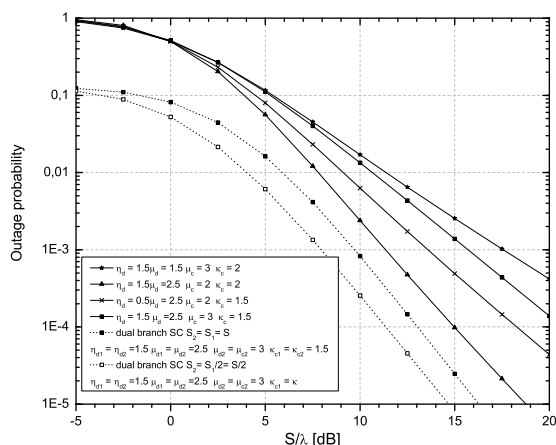
Performance measure, which in best way describes the nature of the wireless communication system behavior is the average symbol error probability (ASEP), or alternatively average symbol error rate (ASER). If number of bits per symbol is equal 2, then this measure is equivalent to the measure known as average bit error probability (ABEP), or alternatively average bit error rate (ABER). Otherwise, if we want to obtain ABER values, signal energy per symbol should be converted into signal energy per bit. ABEP values are obtained capitalizing on conditional SEP relations, which are conditioned over fading statistics which impairs the communication. Conditional SEP are functions of the instantaneous SNR, and functional dependency is determined by the type of modulation scheme performed. Namely, when observing non-coherent binary modulation transmission, conditional SEP is denoted with  $\bar{P}_e = \exp(-g\lambda)$ . Then by averaging over instantaneous SIR values,  $\lambda$ , ABEP can be obtained as [13]:

$$P_e = \int_0^{\infty} f_{\lambda}(t) \frac{1}{2} \exp(-gt)dt \tag{10}$$

where  $g$  denotes modulation constant, i.e.,  $g = 1$  for Binary Differentially Phase Shift Keying (BDPSK) modulation and  $g = 1/2$  for Non-coherent Frequency Shift Keying (NCFSK) modulation.

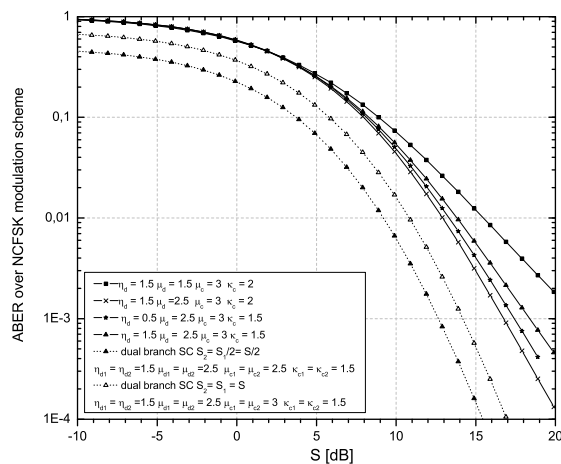
### 3.3 Numerical results

According to Eqs. (9), (7) and (6) OP was efficiently evaluated and presented at for some values of system parameters. It can be seen that lower values of OP (better performances) are obtained



**Fig. 1.** Outage probability improvement obtained by using SC reception for some values of system parameters in the areas of smaller values of  $\mu_c$  and  $\kappa_c$ , and higher values of  $\mu_d$ , since for higher values of  $\mu_c$  and  $\kappa_c$  CCI signal level arises and for smaller values of  $\mu_d$  desired level deteriorates. Also can be seen, that significant improvement of OP values improvement can be obtained by using dual SC reception, especially when balanced reception is applied ( $S_1 = S_2$ ).

According to Eqs. (10), (8), (6) and (5) ABEP for NCFSK transmission was efficiently evaluated and presented at for some values of system parameters. Effects of improvement



**Fig. 2.** ABEP improvement for NCFSK transmission obtained by using SC reception for some values of system parameters

obtained by applying SC reception are also clearly visible here. One can also notice how ABEP performances deteriorate in the areas of system parameter values that correspond to higher levels of superimposed CCI signal.

## 4 Conclusion

In this paper an approach to the performance computation of wireless transmission in general propagation environment subjected to the CCI influence has been presented. Major contribution is obtained in the generality and wide range of applicability of rapidly converging PDF and CDF expressions from Eqs. (5) and (6). Capitalizing on them standard performance measures OP and ABEP are efficiently evaluated, graphically presented and analyzed in the function of system parameters. Possible improvement of calculated performances was also observed through the sight of SC reception technique appliance. Obtained performance analysis could serve as calculation basis for implementing practical wireless system realizations with predetermined minimum performance levels.

**Acknowledgments.** This work has been funded by the Serbian Ministry of Science under the project III 044006.

## References

1. Panic, S., et.al.: Fading and Interference Mitigation in Wireless Communications. CRC Press, USA, (2013)
2. Stavroulakis, P.: Interference analysis and reduction for wireless systems. Artech House, INC, London, (2003)
3. Da Costa, D., Yacoub, M. : The  $\eta - \mu$  joint phase-envelope distribution. IEEE Ant. and Wirel. Propag. Lett., vol. 6, pp. 195-198, (2007)
4. Simon, M., Alouini, M-S.: Digital Communication over Fading Channels. John Wiley Sons, USA, (2000).
5. Filho, J., Yacoub, M.: Highly accurate  $\kappa - \mu$  approximation to sum of M independent non-identical Ricean variates. Electronics Letters, vol. 41, (6), pp. 338-339, (2005)
6. Stefanovic, M., et al.: The CCI Effect on System Performance in kappa-mu fading channels. TTEM, vol. 7, no. 1, pp. 88-92, (2012)
7. Panic, S., et. al.: Performance analysis of selection combining diversity receiver over  $\alpha - \mu$  fading channels in the presence of co-channel interference. IET Communications, vol. 3, no. 11, pp. 1769-1777, (2009).

8. Gradshteyn, I., Ryzhik, I.: Tables of Integrals, Series, and products. Academic Press, New York, (1980)
9. Austin, D., Stuber, L.: In-service signal quality estimation for TDMA cellular systems. in Proc. Sixth IEEE, PIMRC '95 Toronto, ON, Canada, pp. 836-40 (1995)
10. Stuber, G.: Mobile communication. 2nd edition, Kluwer, USA, (2003)
11. Suljovic, S., et. al.: On the wireless transmission in the presence of kappa-mu fading and eta-mu CCI. International Journal of Electronics, accepted for publication. (2015)
12. Moraes, A., Da Costa, D., Yacoub, M. : An outage analysis of multibranch diversity receivers with cochannel interference in  $\alpha - \mu$ ,  $\eta - \mu$ , and  $\kappa - \mu$  fading scenarios. Wireless Pers. Commun., vol. 64, no. 1, pp. 3-19, (2012)
13. Stefanovic, M., et. al.: Performance analysis of system with selection combining over correlated Weibull fading channels in the presence of cochannel interference. AEU - International Journal of Electronics and Communications, vol. 62, issue 9, pp. 695-700, (2008)

# Surface Movements in Source Zones by Satellite Data

Zh. Zhantayev, A. Kim, A. Ivanchukova, V. Junisbekova, A. Turgumbayev

JSC National Center of Space Researches and Technologies,  
LLP Institute of Ionosphere, Almaty, Kazakhstan  
kim.as@mail.ru

**Abstract.** The experimental study of the modern movements of the earth's surface was developed and based on the processing and analysis of satellite GPS data from international center SOPAC and catalogue of primary data for 2000-2013 years was formed. The slow crustal movements of the Northern Tien Shan in tectonic faults was investigated using the methods of satellite radar interferometry and GPS data processing[1][2]. Taking into account the geological conditions of the Almaty city and fault distribution in geological structure was studied the slow movements of the earth surface. Initial data for the Almaty city, located in a high seismic activity, was selected from ENVISAT ASAR satellite. Archive data from 2003 to 2010 totaled 90 images for the two tracks area of 100x100 km, covering the city area. A result was performed in the StaMPS software package. The resulting product of PS method processing refers to the measurement of vertical displacements and gives the output values of height and velocity for individual reflectors (points). Values of the vertical movements of the earth surface calculated using the PS method identified with millimeter precision. The study of slow tectonic motions on Northern Tien Shan region based on GPS data processing and analysis using GAMIT/GLOBK software package. Maps of the velocity distribution in tectonic faults of the Northern Tien Shan for 2003-2013 years in the reference system related to the Eurasian continent.

**Keywords:** GPS, radar interferometry, vertical displacements.

## 1 Introduction

Seismicity of the mountain regions and surrounding flatlands of Kazakhstan is caused by their accessory to the difficult Euroasian continent in the geodynamic relation within which orogenesis is a consequence of interaction of large lithospheric plates. The Northern Tien Shan region characterizes with the greatest activity for the last 125 years where there were strongest earthquakes. The urbanized agglomeration having a dense population and including the city of Almaty located in this seismoactive area[3]. All these listed factors define a relevance of this research.

Data of satellite measurements are actively used in studying of earth's surface movements at the present stage. Intensity of tectonic earth's movements usually are measured by millimeters a year and to distinguish them from other processes happening both on the Earth's surface and in subsoil, it is possible only applying various tool methods and taking high-precision long measurements. The modern movements come to light by complex application of various methods, including: geological, geomorphological, geophysical, geochemical, astronomical, geodetic. For definition slow earths movements of the Northern Tien Shan in tectonic breaks zones were used methods of a spaceborne radar interferometry and processing of GPS data. The method of a satellite interferometry is based on the effect of an interference of electromagnetic waves and mathematical processing of coherent amplitude-phase measurements of several images on the same site on the Earth's surface. For interferometric processing of Almaty city input data were chosen images from the satellite ENVISAT ASAR.

Experimental studies on studying of modern movements of the Earth's surface were conducted on the basis of processing and the analysis of satellite GPS data of the international center

SOPAC and the primary data catalog was created for 2000-2013 years. Also data of local and regional networks of GPS measurements were processed in the GAMIT/GLOBK software. For more detailed analysis temporary time series of daily increments were received according to the processed GPS data from local and regional networks. As a result of processing were constructed maps of the distribution of the velocity of the earth's surface in tectonic fault zones of the Northern Tien Shan for the 2000-2013 years in the reference frame concerning to the Eurasian continent.

## 2 Research on radar interferometry data

The European Space Agency (ESA) saved up a big data archive from various radar satellites. Within the new policy of the agency there was an opportunity to receive archival images free of charge under the offer on scientific researches. The advantage of the archived data is that processing of radar images allows to restore long-term dynamics of a relief. Under the grant of the European Space Agency (ESA) for the territory of Almaty city were received 90 images from the satellite ENVISAT ASAR which were suitable for further interferometric processing. The data archive from 2003 to 2010 consist 90 images for two tracks of 100x100 km covering the territory of the city. Radar imaging was carried out in the Image mode with a vertical combination of polarization and the 30 m spatial resolution. The processing and analysis were carried out in the StaMPS software which realizes a full cycle or separate stages of interferometric processing. As a result of the processing of the Persistent Scatterers method the map of vertical movements of the Earth's surface was received for the period from 2003-2010 according to SAR imaging on the territory of the Almaty city. The resultant product of processing of the PS method belongs to measurement of linear shifts and gives output values of height of separate reflectors (points)[4]. The values of vertical movements of the Earth's surface calculated on the PS method are defined with millimeter accuracy.

The preliminary results received by data processing of SAR imaging on the territory of the Almaty city showed that the most part of the settling reflectors are located in foothill areas and surrounding territories. The map of vertical displacements of the Earth's surface of the Almaty city for 2003-2010 years according to satellite radar imagery is submitted in the Google earth environment Fig. 1.

The map of vertical displacements of the Earth's surface for the Kaskelen settlement is submitted in Fig. 2. On the basis of the received the map of vertical displacements of the Earth's surface according to the satellite radar interferometry it is possible to allocate areas of a terrestrial surface where raisings prevail. Results of processing allow seeing displacement dynamics for each date of image for each point. The graph of displacements on 145 points is provided for the interested area in a radius of 500 meters.

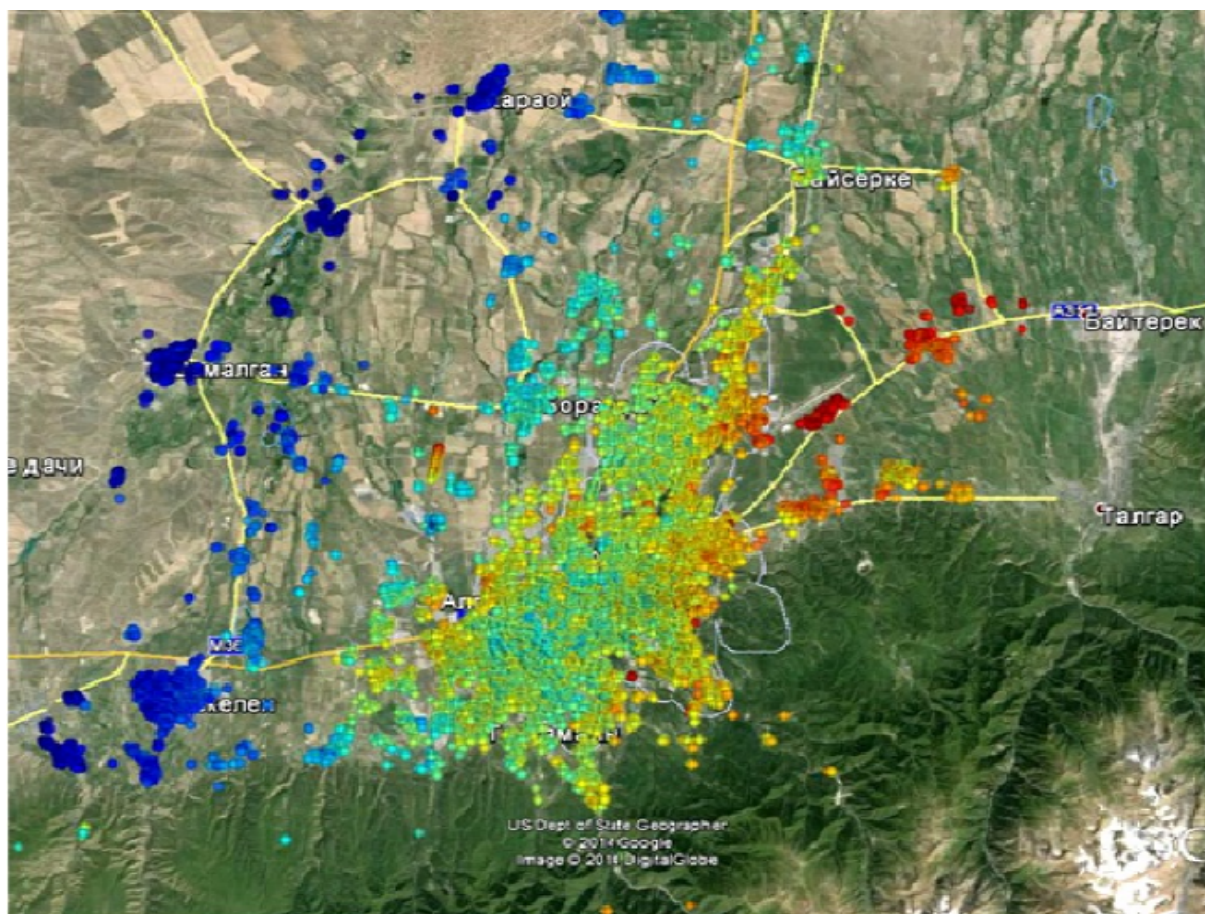
## 3 Research on GPS data measurements

The studied region is shown in a rectangle and includes 30 stations of the International Ground Station (IGS) Network and 10 GPS stations of a local network Fig. 3.

Currently there are multiple types of interpretative processing which are based on data of preprocessing. Results of processing of this kind possess certain distinctions, connected generally with a choice of basic systems of the world network. The maximum errors definitions of movement speeds are given in table 1.

For comparison parameters of modern movements of the Earth's surface of the region were calculated in two reference systems, rather Euroasian continent and the Earth's center. Processing





**Fig. 1.** The map of vertical displacements of the Earth's surface of Almaty city for 2003-2010 years according to satellite radar imagery in Google earth.

of GPS data was carried out by the GAMIT/GLOBK software. For each GPS point of a network graph of annual temporary displacements were analyzed on each of three components in the directions the South-North (SN), the West-East (WE), up-down (UP) with removal from the subsequent calculations of abnormal deviations of the technogenic nature. On abscissa axis values of days of a year are presented. On ordinate – average value of displacement of days in mm. Over graphs values of average shift in a year are shown,  $\pm$  an error, a normal average mean square error and the weighted mean square error. Time series according to the TURG GPS station for 2013 were given in Fig. 4.

Seasonal fluctuations of vertical component are allocated on time series. The modern movements observed on the Earth's surface in the form of movements of separate points of measurements are defined by a set of actions of at the same time happening various processes of the linear and nonlinear nature. One of ways of display of motion is the binding to each point of a surface of the movement velocity vector. In practice velocity in separate points or separate velocity components can be measured in various ways. As a result of GPS monitoring as a result received a full vector of velocity of the measured points on a surface. At the same time the horizontal component of velocity is accepted to consider separate from the field of the vertical component. It is caused considerably by a bigger error of determination of vertical velocity in connection with features of a relative positioning of wave signals of satellites at the determination of coordinates of points. The velocity field of movement of the Earth's surface of the Northern Tien Shan for

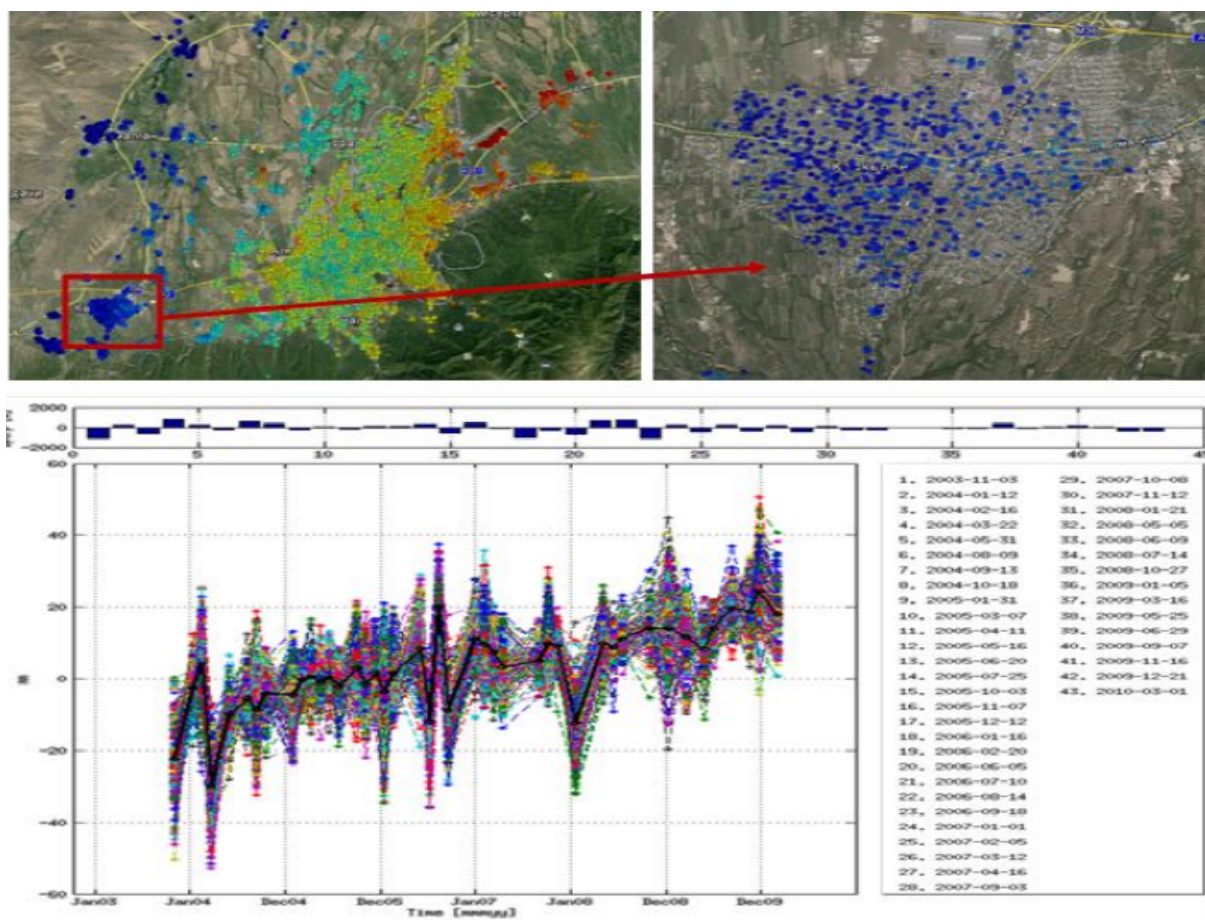


Fig. 2. Kaskelen settlement, where raisings prevail. The graph is constructed on 145 points in a radius of 500 meters.

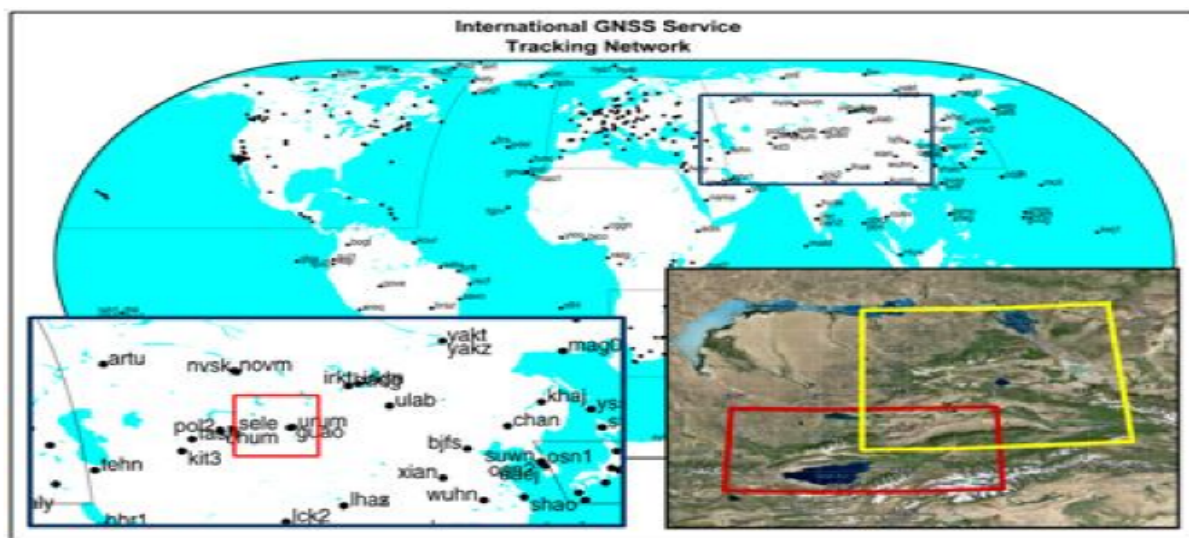


Fig. 3. International Ground Station (IGS) Network



Table 1. Errors of determination of speeds on GPS networks of various researchers

Network name	The East-West component (mm/year)		The North-South component (mm/year)		The Vertical component (mm/year)	
	maximum	average	maximum	average	maximum	average
IGS network	2,59	1,6	2,89	1,9	4,49	3,5
JSC “NC “KGS” network	5,50	4,2	6,34	4,6	9,16	8,6
Institute of Ionosphere LLP network	1,41	1,2	1,42	1,3	3,31	2,2
Temporary measurements of the Almaty city polygon	10,0	8,5	10,0	8,5	10,0	9,5

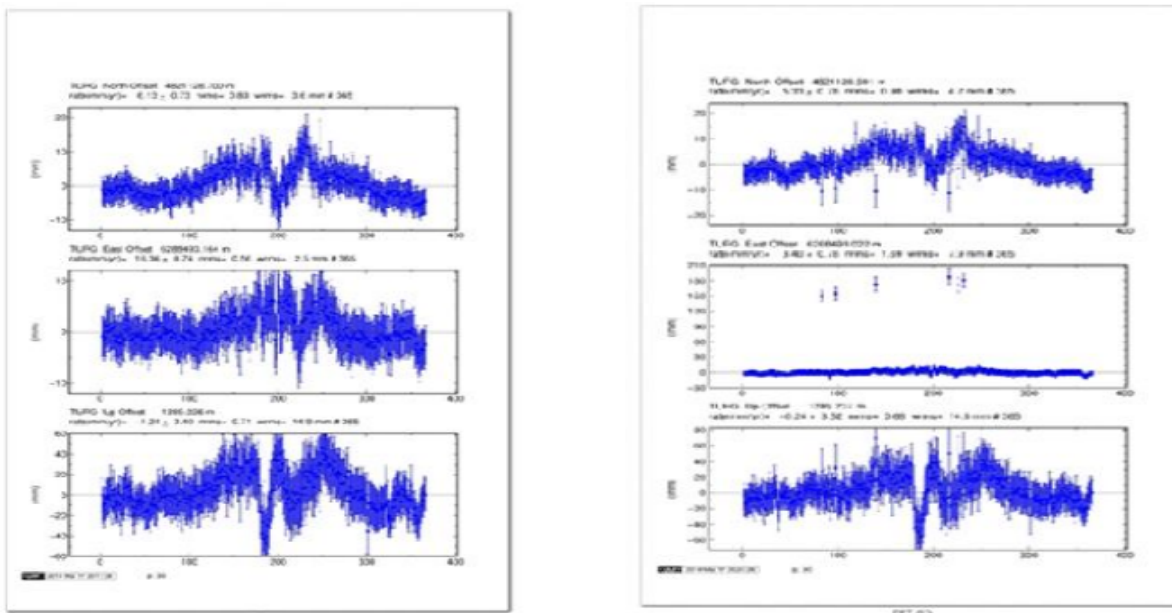
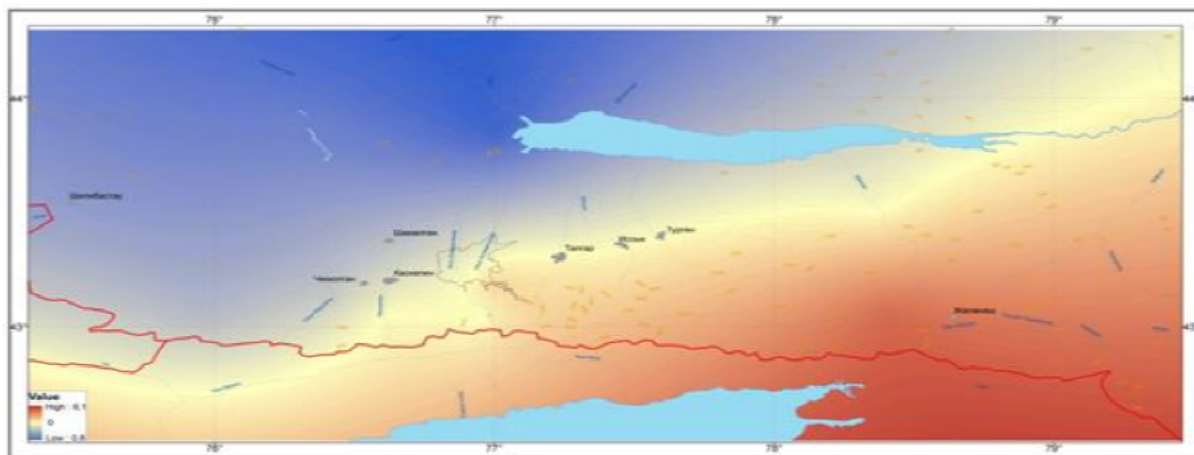
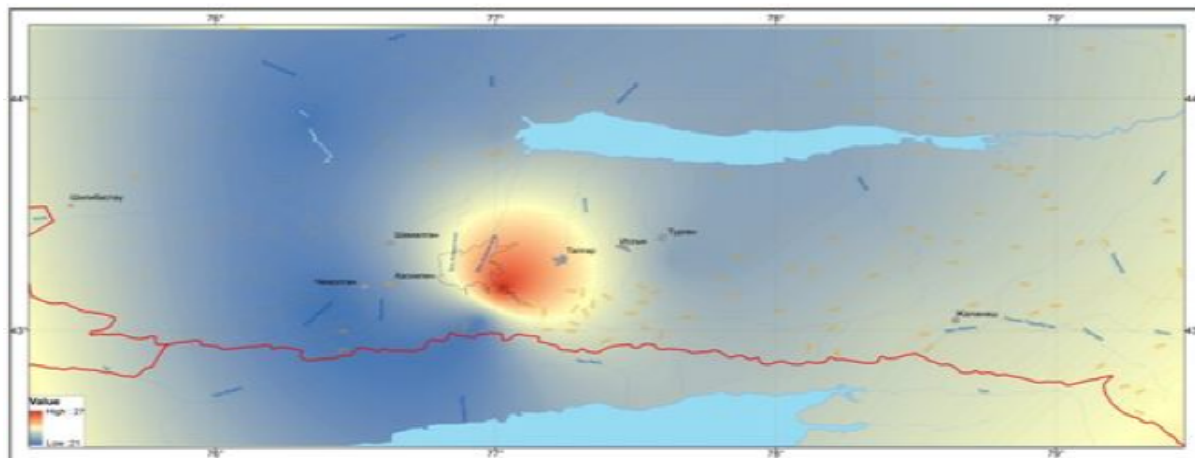


Fig. 4. Time series of components – SN, WE, UP of point displacement of a local GPS network concerning the Earth’s center (at the left) and the Eurasian continent (on the right) for 2013.

2000-2012 years processed in the reference system of the rather Euroasian continent is presented in Fig. 5. Maps of the velocity field of movement of the Earth's surface of the Northern Tien Shan for 2000-2013 processed in a geocentric reference system concerning the Earth's center of gravity are submitted in Fig. 5-7.

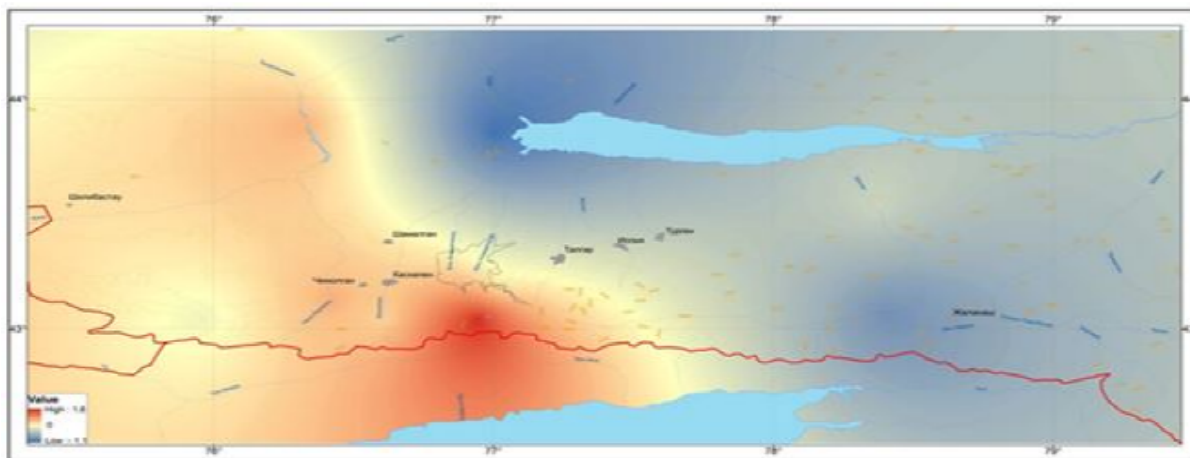


**Fig. 5.** The map of movements velocities of the Earth's surface of the Northern Tien Shan in the South-North SN direction concerning Earth's center of gravity in mm/year



**Fig. 6.** The map of movements velocities of the Earth's surface of the Northern Tien Shan in the West-East WE direction concerning the Earth's center of gravity in mm/year

Concerning the Earth's center the prevailing horizontal movement of the region is the movement in the direction the East North East at values of the velocity of 1-6 mm/year for northern the SN components (Fig. 5) and 21-28 mm/year for east WE (Fig. 6). In the vertical movement was established the steady raising of the western part of the territory with a velocity of 1.5-3.0 mm/year and lowering east with velocity of 1.5-2.0 mm/year (Fig. 5-7). The vertical displacement velocities of the Northern Tien Shan territory by GPS data don't conflict with modern concepts of seismotectonic zoning and confirmed by the latest figures zoning epi-platfornal orogens on motion mode. Sign-variable sites of vertical velocities coincide with areas of uplift and subsidence (foothill and intermountain hollows) of various temporary duration and reflect the modern active geodynamic movements of the territory.



**Fig. 7.** The map of movements velocities of the Earth's surface of the Northern Tien Shan in the vertical UP direction concerning the Earth's center of gravity in mm/year

#### 4 Conclusion

The analysis of spatio-temporal distribution of velocities and displacements of the Earth's surface of the Northern Tien Shan from 2000 to 2013 was carried out by results of GPS monitoring on stationary points. The created maps reflect the general trend of the direction of the movement of the Earth's surface in the region and confirm structurally tectonic constructions according to geologic-geophysical data.

The vertical movements of the earth surface in the region of Almaty were studied based on radar images from satellite ENVISAT ASAR. Based on these maps of vertical displacements of the earth's surface by satellite radar interferometry were noted intensive vertical movements in fault zones of the southern part of the considered Almaty region.

#### References

1. Zubovich, A.V., Trapeznikov, J.A., Bragin, V.D., Mosiyenko, O.I., Shchelochkov, G.G., Rybin, A.K., Batalev, V.Y.: Deformation field, deep crustal structure and the spatial distribution of the seismicity of the Tien Shan. *J. Geology and Geophysics*. vol.42, № 2, pp.1634–1640 (2001)
2. Zubovich, A.V., Wang, X.-q., Scherba, Y.G., Schelochkov, G.G., Reilinger, R., Reigber, C., Mosienko, O.I., Molnar, P., Michajljow, W., Makarov, V.I., Li, J., Kuzikov, S.I., Herring, T.A., Hamburger, M.W., Hager, B.H., Dang, Y.-m., Bragin V.D. and Beisenbaev. R.T.: GPS velocity field for the Tien Shan and surrounding regions. *J. Tectonics*. Vol. 29, doi:10.1029/2010TC002772 (2010)
3. Timush, A.V.: Seismotectonics of a lithosphere of Kazakhstan. Scientific publication. Almaty, pp.5–7 (2011)
4. Hooper, A., Zebker, H., Segall, P. and Kampes B.: A new method for measuring deformation on volcanoes and other non-urban areas using InSAR persistent scatterers. *J. Geophysical Research Letters*. Vol. 31 (2004)

# Системы Распознавания Образов в Задачах Автоматизации Распознавания Паспортных Данных

Е.Н. Амиргалиев<sup>1</sup>, Р. Юнусов<sup>2</sup>

<sup>1</sup> Институт информационных и вычислительных технологий МОН РК, Алматы, Казахстан

<sup>2</sup> Университет им. Сулеймана Димереля, Каскелен, Казахстан  
yunussov@gmail.com, amir\_ed@mail.ru

**Аннотация.** Рассмотрены следующие задачи: 1. Идентификация границ документа. Для идентификации границ были применены алгоритмы идентификации смены градиента интенсивности пикселей на основе алгоритма CannyEdge[2]. 2. Разворот документа до получения минимального расхождения с горизонтальной плоскостью. Для поиска таких линий использовался алгоритм HoughLineTransform[3]. Данная процедура позволяет решить сразу несколько задач: Получение горизонтального положения текстовой информации; Проведение поиска шаблонов типов документа; Извлечение лица человека при наличии. 3. Идентификация типа документа путем сравнения шаблонов. Рассмотрены следующие задачи: Извлечение лица из документа; Извлечение машиночитаемой зоны; Распознавание текстовой информации.

**Ключевые слова:** нейронные сети, распознавание образов, растр, обработка изображений.

## 1 Введение

Системы безопасности, построенные на использовании современных аппаратных и программных разработок получили широкое распространение, как за рубежом, так и в Республике Казахстан. Следует отметить программу "Безопасный город"[1], в рамках которой было реализован ряд мероприятий по установке видеокамер в крупных городах. Системы автоматической регистрации пассажиров на авиалиниях [2], основанные на алгоритмах распознавания лиц. Множественное проникновение таких систем в различные отрасли человеческой деятельности говорит о большом потенциале разработок в этой сфере. А все более острые проблемы безопасности, встающие перед государством и корпоративным сектором, требуют разработок более лучших и эффективных методов решения задачи. Предлагаемая в работе модель позволяет автоматизировать процесс регистрации персоны путем автоматического распознавания его паспортных данных и занесения метаинформации извлеченной методами оптического распознавания символов с использованием нейронных сетей. Применения похожих моделей рассматривали Young-Bin Kwon and Jeong-Hoon Kim [3]

## 2 Обзор исследуемой области

Проблема автоматического распознавания паспортных данных рассматривается в рамках решения различных задач, в которых требуется увеличить производительность труда по идентификации владельца паспорта в устоявшихся бизнес процессах, таких как фиксация миграционных процессов на границе, ведение журналов посещения режимных объектов и пр. На текущий момент большинство стран приняло стандарт ИКАО 9303, регламентирующий вид идентифицирующих документов и машиночитаемой зоны, что существенно упрощает процесс создания систем с использованием алгоритмов автоматически извлекающих информацию из растровых изображений паспорта, получаемых путем сканирования или же фотографирования. В данной работе рассматриваются проблемы идентификации





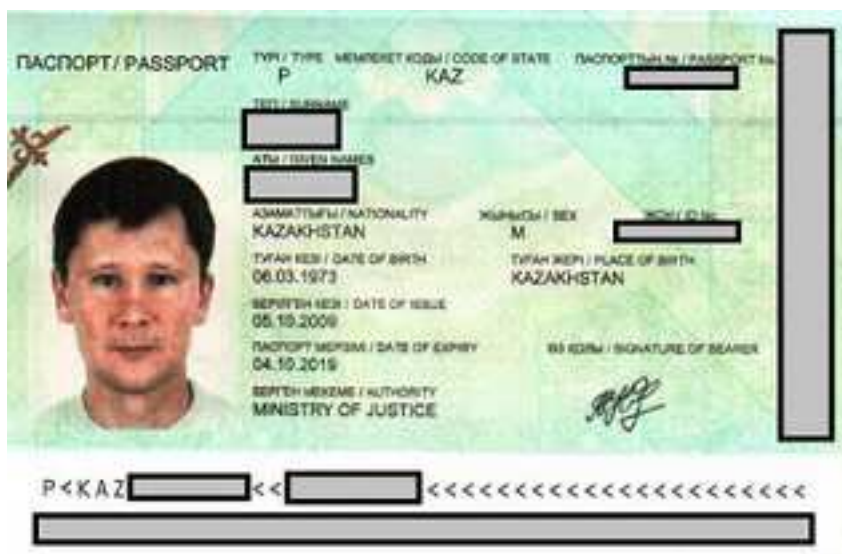


Рис. 4. Паспорт гражданина РК

В работе использовался сканер формата А6 Fujitsu-65fi. При этом сканируемая площадь превышает площадь представленных видов документов, что накладывает условия вариативности окружающей среды поверхности сканирования и ставит задачу поиска границ документа в различных условиях окружающих пикселей.

Для решения задач были использованы алгоритмы:

1. Поиска границ на основе анализа градиента интенсивности пикселей[4];
2. Поиска прямых линий [5];
3. Поиска шаблонов на основе вычисления корреляции [6];
4. Геометрического соответствия шаблонов;
5. Поиска человеческого лица на изображении[7];
6. Распознавания машиночитаемой зоны на основе оптической идентификации символов с использованием нейронных сетей.

### 3 Описание модели

В работе рассматриваются проблемы извлечения информации из растровых изображений и пути оптимизации качества извлекаемой информации путем принятия решения о применении нейронных сетей и обучения их на тестовых наборах. Среди существующих подходов по распознаванию символов существуют два основных направления - это анализ шаблонов растровых представлений или же извлечение характерных инвариантных признаков символа. В работе предпочтение было отдано анализу шаблонов растровых представлений ввиду простоты имплементации нейросетевого алгоритма классификации символов в условиях одного типа шрифта, регламентированного стандартом ICAO 9303.

Анализ растрового изображения паспорта сводится к последовательности действий, отображенной на рисунке 5.

#### 3.1 Идентификация границ документа

Проблема поиска границ документа на сканированном изображении является актуальной, так как размеры видов документов отличаются, а положение его на сканирующей поверхно-



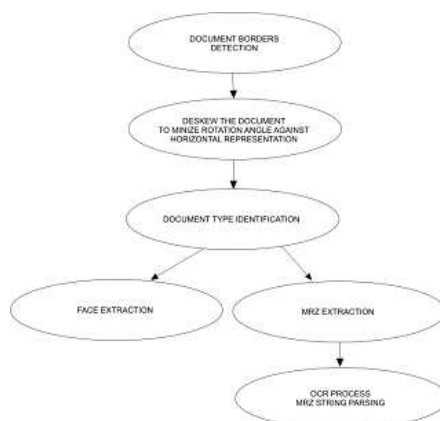


Рис. 5. Последовательность действий модели

сти зависит от человеческого фактора - т.е. смещенное положение документа относительно ожидаемой позиции и его незначительные повороты - ситуация неизбежная. Так же проблема усложняется фактором изменчивости окружающей среды сканируемого документа, так как заранее не известно - будет ли сканирование проводиться при закрытой крышке или открытой, а также какова площадь сканируемого документа. Для идентификации границ были применены алгоритмы идентификации смены градиента интенсивности пикселей на основе алгоритма Canny Edge[4]. Применение алгоритма и его коэффициентов дает различный результат, который зависит от разрешения исходного растра (в рамках работы использовалось разрешение 300 DPI) и типа документа. Для уменьшения неинформативных элементов, получаемых на выходе было проведено уменьшение размера изображения и линейное размытие. Полученные границы документа позволяют приступить к следующему этапу обработки - поиска наклона документа и компенсация разворота относительно горизонтальной плоскости.

### 3.2 Разворот документа до получения минимального расхождения с горизонтальной плоскостью

Полученные результаты обработки изображения на поиск границ документа дает информацию, доступную для поиска связанных пикселей, которые можно описать уравнением прямой линии. Для поиска таких линий использовался алгоритм Hough Line Transform[5]. При этом не исключены случаи нахождения линий, не соответствующих реальным границам документа. Однако такие линии имеют случайный характер, и ими можно пренебречь применив простой статистический анализ появления линий определенного градуса относительно горизонтальной плоскости. В работе был применен алгоритм учитывающий статистику линий не превышающих отклонение от горизонтальной плоскости в 10 градусов.

После получения прямых линий границ документа можно определить наклон документа относительно горизонтальной плоскости и компенсировать этот наклон путем разворота всего сканированного изображения на величину, обратную повороту документа относительно горизонтальной плоскости. Данная процедура позволяет решить сразу несколько задач:

1. получение горизонтального положения текстовой информации;
2. Проведение поиска шаблонов типов документа;
3. Извлечение лица человека при наличии.

### 3.3 Идентификация типа документа путем сравнения шаблонов

На каждом типе документа и для каждой стороны документа выделены уникальные участки, не повторяемые по положению, текстурным особенностям и размерам на других видах документах. Для каждого документа было выбрано не менее 5 характерных признаков. Характерные признаки были сохранены в виде шаблонов для дальнейшего применения алгоритма поиска их на изображении[6]. Пороговым значением при поиске шаблонов было выбрано достаточно низкое значение в 60. При этом возникает ситуация ложноположительных ответов алгоритма. Для снижения вероятности неверной идентификации типа документов были применены алгоритмы топологического соответствия взаимного расположения шаблонов. Т.е. все расстояния и смещения шаблонов относительно друг друга. При не соответствия геометрической топологии найденный фрагмент принимается как ложно положительный и исключается из выборки.

### 3.4 Извлечение лица из документа

При идентификации типа документа (стандарта РК) становится очень просто выделить участок с наличием фотографии человеческого лица. Однако в работе предполагается использование паспортов, вид которых не известен (шаблоны для которых не были собраны, например паспорт гражданина США). Тогда такой паспорт принимается как стандарт ICAO 9303 с двумя линиями строк в зоне MRZ. Однако положение человеческого лица может варьироваться в зависимости от страны выпуска. Для решения этой задачи были применены алгоритмы поиска лица человека[7]. Данный алгоритм инвариантен к размерам человеческого лица, так как использует пирамидальный спуск при поиске соответствия.

### 3.5 Извлечение машиночитаемой зоны

При извлечении машиночитаемой зоны решаются проблемы выбора пороговой величины для бинаризации изображения. Машиночитаемая зона отличается следующими характеристиками - значимые символы имеющие диапазон значений интенсивности пикселей от 0 до 100, и фон - либо монотонная текстурная поверхности, либо незначительный белый шум в виде узоров обычно в диапазоне интенсивности пикселей от 120 до 255. Как видно граница между значимыми и незначимыми пикселями не велика. При этом имеют место наличие шумов различного характера - потертости, измятости, частичное отсутствие символа, наличие пятен. Так же в случае с документом типа А появляется явление просвечивания документа, что увеличивает сложность определения пороговой величины для бинаризации. В рамках работы были применены алгоритмы адаптивной и линейной пороговой бинаризации[8].

### 3.6 Распознавание текстовой информации

На рисунке отображено изображение MRZ, получаемое со сканера. Как видно, имеется наклон. Такой наклон может оставаться даже после второго этапа обработки изображения (идентификация поворота всего документа по границам). Поворот направления текста определяется как относительное положение крайнего левого символа и крайнего правого символа в каждой строке. Повороты компенсируются афинными преобразованиями путем вращения вокруг центральной оси всего документа и повторения всего алгоритма с первого шага.

После проведения афинных преобразований проводится бинаризация изображения, для выделения только букв и стирания заднего фона, каким бы он не был. При этом следует



Рис. 6. Изображение скана зоны MRZ



Рис. 7. Бинаризованное изображение зоны MRZ

принимать такие факты, как шумы, грязный паспорт, мятый паспорт, частично стертые названия.

После чего выделяется контур каждого символа и сравнивается с имеющимся шаблоном алфавита. При этом рассчитывается суммарная степень достоверности алгоритма и степень достоверности распознавания каждого символа.



Рис. 8. Выделение границ каждого символа

И на последнем этапе производятся алгоритмические проверки контрольных сумм, регламентированных стандартом ICAO 9303.

#### 4 Предположения

Коммерческое применение оптического распознавания текста начинает свою историю с 1955 года. Среди работ по автоматическому распознаванию текстовой информации можно выделить исследования в следующих направлениях:

1. Адаптивная система оптического распознавания символов (Adaptive OCR), которая охватывает такие проблемы как распознавание различных шрифтов и языков, распознавание монотонных шрифтов, автоматическая сегментация документа и математические модели распознавания [9].
2. Система распознавания рукописного текста, которая остается в стадии активных изысканий и охватывает такие проблемы как распознавание рукописного текста в заданной позиции, распознавание росписи, системы автоматического распознавания свободного текста и распознавание рукописного текста на специализированных устройствах [10].
3. Предварительная обработка изображений [11], которая охватывает проблемы подбора правильных фильтров для получения презентабельной выборки и последующего применения систем автоматической классификации.
4. Системы интеллектуальной пост-обработки [12], которые охватывают проблемы обработки при условиях высокого шума.

5. Системы распознавания текстов в мультимедиа [13], охватывающая проблемы распознавания текстов на простых фотографиях и имеют дело с выделением контуров, проектными и нелинейными искажениями.

Следует отметить, что абсолютная точность при распознавании может быть достигнута только путем последующего редактирования человеком. Поэтому проблемы распознавания текстов и по сей день являются предметом активных исследований.

## 5 Заключение

В рамках работы были затронуты такие вопросы как:

1. Устойчивые к помехам алгоритмы идентификации человеческого лица
2. Устойчивые к помехам алгоритмы распознавания печатных символов
3. Алгоритмы идентификации паттернов на основе корреляционного анализа
4. Применение геометрических топологических особенностей при поддержке принятия решения

Достигнуты высокие значения достоверности автоматического извлечения паспортных данных в условиях сканирования при разрешении 300 DPI и незначительных поворотах паспорта - до 10 градусов относительно горизонтальной оси сканирования. При этом нейронная сеть по распознаванию символов обучалась только на одном образе каждого символа без учета поворотов и наклонов. Это влияет на качество распознавания и решения неоднозначностей при распознавании похожих друг на друга символов в условиях повышенных шумов. Именно поэтому является критичным максимальное снижение отклонения положения паспорта относительно горизонтальной оси сканирования. При распознавании символов вычисляется общая достоверность алгоритма, после которой можно пересчитать наклон текста, анализируя первую и последнюю букву в строке и повторить процедуру распознавания текста. Данный подход позволяет увеличить точность распознавания паспортных данных без необходимости повторного сканирования документа. Остались вопросы неоднозначности идентификации символов с применением алгоритмов на основе нейронных сетей. Наиболее яркий пример символ "0 "Ноль" и "O "Буква алфавита". В большинстве случаев расстояние между двумя представлениями крайне низкое и процент возникновения ложноположительной идентификации символа становится достаточно высок при низком качестве растрового изображения. Поэтому необходимы дополнительные исследования в части контекстного определения положения исследуемого шаблона и моделирования контекстной нейронной сети.

## Список литературы

1. "Безопасный город" в Казахстане - краткий исторический обзор, Журнал "Рубеж" №1 Апрель 2013, Главные редактор Михаил Динеев, Издатель и учредитель ООО "Компания Р-Медиа" [Электрон. ресурс]. - 2014. - URL: <http://www.aips.kz/ru/otchety/otchety-o-vystavke-ot-partnerov/bezopasnyj-gorod-v-kazahstane> (дата обращения: 01.06.2014)
2. Company bets on airport of the future: passing security with an iris scan. Ministry of innovation/Business of Technology, ArsTechnica, Sept 2012, by Cyrus Farivar. [Электрон. ресурс]. - 2014. - URL: <http://arstechnica.com/business/2012/09/company-bets-on-airport-of-the-future-passing-security-with-an-iris-scan/> (дата обращения: 01.06.2014)
3. Young-Bin Kwon and Jeong-Hoon Kim, Recognition based Verification for the Machine Readable Travel Documents [Электрон. ресурс]. - 2014. - URL: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.89.9373&rep=rep1&type=pdf> (дата обращения: 01.06.2014)

4. J. Canny, "A computational approach to edge detection," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 8, No. 6, pp. 679- 698, November 1986.
5. Bhattacharya, P., Rosenfeld, A., and Weiss, I. 2002. Point-to-line mappings as Hough transforms. Pattern Recognition Letters 23, 14, 1705-1710.
6. Template Matching [Электрон. ресурс]. – 2014. – URL: [http://docs.opencv.org/doc/tutorials/imgproc/histograms/template\\_matching/template\\_matching.html](http://docs.opencv.org/doc/tutorials/imgproc/histograms/template_matching/template_matching.html) (дата обращения: 01.06.2014).
7. Cascade Classifier [Электрон. ресурс]. – 2014. – URL: [http://docs.opencv.org/doc/tutorials/objdetect/cascade\\_classifier/cascade\\_classifier.html](http://docs.opencv.org/doc/tutorials/objdetect/cascade_classifier/cascade_classifier.html) (дата обращения: 01.06.2014).
8. Thresholding Operators [Электрон. ресурс]. – 2014. – URL: <http://docs.opencv.org/doc/tutorials/imgproc/threshold/threshold.html> (дата обращения: 01.06.2014).
9. Heath E. Nielson and William A. Barrett. Consensus-based table form recognition. In Proceedings of the International Conference on Document Analysis and Recognition, volume II, August 2003.
10. Ching Y. Suen, Shunji Mori, Soo H. Kim, and Cheung H. Leung. Analysis and recognition of Asian scripts - the state of the art. In Proceedings of the International Conference on Document Analysis and Recognition, volume II, August 2003.
11. Kristen Summers. Document image improvement for OCR as a classification problem. In Document Recognition and Retrieval X, volume 5010, 2003.
12. Yefeng Zheng, Huiping Li, and David Doermann. Text identification in noisy document images using markov random field. In Proceedings of the International Conference on Document Analysis and Recognition, volume I, August 2003.
13. Paul Clark and Majid Mirmehdi. Finding text regions using localised measures. In Majid Mirmehdi and Barry Thomas, editors, Proceedings of the 11th British Machine Vision Conference, pages 675-684. BMVA Press, September 2000.

# Полиномиальный Алгоритм для Задачи MSP3

М.З. Арсланов<sup>1</sup>

<sup>1</sup> Институт информационных и вычислительных технологий, Алматы, Казахстан

**Аннотация.** Рассматривается следующая открытая проблема теории расписаний. Даны  $m_i$  маленьких отрезков длины  $l_i, i = 1, 2, 3$ , которые необходимо разместить в как можно меньшем числе отрезков длины  $L$ . Эта проблема называется MSP3 (multiprocessor scheduling problem with 3 job length). Библиографию по данной проблеме можно найти в [1]. В этой статье представлен полиномиальный алгоритм для задачи MSP2 (multiprocessor scheduling problem with 2 job length) и отмечены трудности, возникающие при построении полиномиального алгоритма для MSP3. Подробно исследован так называемый делимый случай задачи. Для этого случая построен полиномиальный алгоритм.

**Ключевые слова:** теория расписаний, полиномиальный алгоритм.

## 1 Основные результаты

В данном докладе рассматривается следующая открытая проблема теории расписаний.

Даны  $m_i$  маленьких отрезков длины  $l_i, i = 1, 2, 3$ , которые необходимо разместить в как можно меньшем числе отрезков длины  $L$ . Эта проблема называется MSP3 (multiprocessor scheduling problem with 3 job length). Библиографию по данной проблеме можно найти в [1]. В этой статье представлен полиномиальный алгоритм для задачи MSP2 (multiprocessor scheduling problem with 2 job length) и отмечены трудности, возникающие при построении полиномиального алгоритма для MSP3. Будем рассматривать так называемый делимый случай этой задачи, когда отрезок длины  $L$  полностью покрывается  $M_1$  отрезками длины  $l_1$  или  $M_2$  отрезками длины  $l_2$  или  $M_3$  отрезками длины  $l_3$ . Множество целочисленных точек многогранника задачи о рюкзаке будет представляться в следующем виде:

$$P = \{(x_1, x_2, x_3) \mid \frac{x_1}{M_1} + \frac{x_2}{M_2} + \frac{x_3}{M_3} \leq 1, x_i, M_i \in Z_+ = \{0, 1, 2, \dots\}, i = 1, 2, 3.\}$$

Каждая из этих точек представляет некоторый раскрой отрезка длины  $L$ , поскольку  $(x_1, x_2, x_3) \in P \Leftrightarrow x_1 l_1 + x_2 l_2 + x_3 l_3 \leq L$ .

Тогда задачу MSP3 можно сформулировать как задачу целочисленного линейного программирования, если перенумеровать все точки  $P$  индексом  $j = 1, 2, \dots, D$

$$\sum_j z_j \rightarrow \min$$

при условии

$$\sum_j z_j x_{j,1} \geq m_1$$

$$\sum_j z_j x_{j,2} \geq m_2$$

$$\sum_j z_j x_{j,3} \geq m_3, z_j \in Z_+.$$

Несмотря на то, что задачи целочисленного линейного программирования хорошо изучены, для так поставленной задачи MSP3 вряд ли возможно построение эффективного полиномиального алгоритма. Важную роль в исследовании делимого случая MSP3 играет релаксация к задаче линейного программирования связанной с ней задачи линейного целочисленного программирования. Эта релаксация выглядит следующим образом

$$z = z_1 + z_2 + z_3 \rightarrow \min$$

при условии

$$z_1 M_1 \geq m_1, z_2 M_2 \geq m_2, z_3 M_3 \geq m_3.$$

Легко показать, что решение этой задачи дается формулой

$$z^c = z_1^c + z_2^c + z_3^c$$

$$z_1^c = \frac{m_1}{M_1}, z_2^c = \frac{m_2}{M_2}, z_3^c = \frac{m_3}{M_3}.$$

Здесь индекс  $^c$  означает, что решается непрерывная (continued) задача линейного программирования.

В это связи краткая запись раскроя отрезка  $L$

$$(z_1, z_2, z_3) : z_1 + z_2 + z_3 \leq 1.$$

означает, что  $z_1$  доля отрезка  $L$  кроится на отрезки  $l_1$ ,  $z_2$  доля отрезка  $L$  кроится на отрезки  $l_2$ ,  $z_3$  доля отрезка  $L$  кроится на отрезки  $l_3$ .

Для задачи MSP2 аналогичная релаксация дает фактически решение исходной задачи целочисленного программирования. Дело в том, что для MSP2 справедлива следующая формула (так называемое свойство IRUP (integer round up property))

$$z^* = \lceil z^c \rceil, \quad (1)$$

где  $z^*$  есть решение задачи MSP2.

Это позволяет использовать для оптимального решения MSP2 только 3 раскроя, в первых двух из которых отрезки длины  $L$  раскраиваются на одинаковые отрезки  $l_1, l_2$ . Легко видеть, что если  $\{z_1^c\} + \{z_2^c\} > 1$ , то количество таких раскроев в оптимальном решении равно  $\lceil z_1^c \rceil, \lceil z_2^c \rceil$  и больше раскроев нет. Если же  $\{z_1^c\} + \{z_2^c\} \leq 1$ , то количество таких раскроев равно  $\lfloor z_1^c \rfloor, \lfloor z_2^c \rfloor$  и к ним еще добавится один раскрой отрезка  $L$  на  $M_1 \{z_1^c\}$  отрезков длины  $l_1$  и  $M_2 \{z_2^c\}$  отрезков длины  $l_2$ . Приведенное формульное решение задачи MSP2 намного проще, чем в [1]. Легко показать, что алгоритмическая сложность этого формульного решения имеет линейную трудоемкость, что отличает его в положительную сторону по сравнению с алгоритмом квадратичной сложности в [1].

Таким образом, структура оптимального раскроя для MSP2 очень проста. На первый взгляд кажется, что нечто подобное должно выполняться и для MSP3. Однако, исследование затрудняется, в силу того, что для MSP3 не выполняется уравнение (1). Для MSP3 справедливо неравенство

$$\lceil z^c \rceil \leq z^* = \lceil z^c \rceil + 1, \quad (2)$$

Это свойство называется MIRUP (modified integer round up property).

Приведем один из известных примеров того, что для MSP3 не выполняется свойство IRUP.

В этом примере  $L = 132, l_1 = 44, l_2 = 33, l_3 = 12, m_1 = 2, m_2 = 3, m_3 = 6$ . Можно показать что  $z^c = \frac{259}{132} < 2$ , однако 2 отрезка длины 44, 3 отрезка длины 33 и 6 отрезков длины 12 никак не могут уместиться в двух отрезках длины 132.

Известно, что любая задача из MSP3 обладает свойством MIRUP. Будем обозначать через MIRUP множество задач, обладающих свойством MIRUP, не рискуя впасть в смешение терминов. Таким образом  $MIRUP = MSP3$  в теоретико-множественном смысле. Аналогично через IRUP будем обозначать множество задач из MSP3, для которых справедливо свойство IRUP. Интерес при этом представляют задачи собственно MIRUP, то есть задачи, не принадлежащие IRUP. Множество этих задач будем обозначать IRUP', имея в виду справедливость теоретико-множественного равенства

$$IRUP' = MIRUP \setminus IRUP.$$

Каждой задаче MSP3 соответствует вектор

$$(z_1, z_2, z_3) = \left( \frac{m_1}{M_1}, \frac{m_2}{M_2}, \frac{m_3}{M_3} \right).$$

Пусть

$$1 < z = z_1 + z_2 + z_3 \leq 2.$$

Без ограничения общности можно считать, что  $z_1 \geq z_2 \geq z_3$ . Если задача принадлежит IRUP', то очевидно, то  $z_1 < 1, z_2 < 1, z_3 < 1$ , ибо иначе существовал бы раскрой одного отрезка  $L$  на  $M_1$  отрезков  $l_1$  и второго отрезка  $L$  на  $(m_1 - M_1, m_2, m_3)$  отрезков  $(l_1, l_2, l_3)$ . Кратко этот раскрой можно записать в виде

$$(1, 0, 0), (z_1 - 1, z_2, z_3).$$

Будем называть вектор

$$(z_1, z_2, z_3) : 1 > z_1 \geq z_2 \geq z_3, 1 < z = z_1 + z_2 + z_3 \leq 2$$

плохим. Если вектор  $(z_1, z_2, z_3)$  плохой, то соответствующая задача может принадлежать IRUP' (но может и не принадлежать). Но если  $z_1 \geq 1$ , то вектор  $(z_1, z_2, z_3)$  будет соответствовать задаче из IRUP. Поэтому будем его называть хорошим. Аналогично можно определить понятие плохого и хорошего вектора  $(z_1, z_2, z_3)$  для случая  $2 < z_1 \leq z_2 \leq z_3 \leq 3$  и т.д.

Пусть ищем вектор  $(z_1, z_2, z_3) : (z_1 + z_2 + z_3 \leq 1, z_1 \geq z_2 \geq z_3)$  такой, что векторы  $(2z_1, 2z_2, 2z_3), (3z_1, 3z_2, 3z_3), (4z_1, 4z_2, 4z_3), (5z_1, 5z_2, 5z_3), (6z_1, 6z_2, 6z_3)$  были плохие. Тогда оказывается, что вектор  $(7z_1, 7z_2, 7z_3)$  будет хорошим.

В самом деле, вектор  $(2z_1, 2z_2, 2z_3)$ , это значит что  $2z_1 < 1, 2z_2 < 1, 2z_3 < 1$ . Пусть вектор  $(3z_1, 3z_2, 3z_3)$  плохой. Ясно, что  $3z_1 \geq 1$ . Чтобы вектор  $(3z_1, 3z_2, 3z_3)$  был плохим необходимо, чтобы  $3z_2 < 1, 3z_3 < 1$ . Ибо, если бы  $3z_2 \geq 1$ , то имели бы раскрой трех отрезков длины  $L$  который записывается в виде векторов  $(1, 0, 0), (0, 1, 0), (3z_1 - 1, 3z_2 - 1, 3z_3)$ . Пусть вектор  $(4z_1, 4z_2, 4z_3)$  плохой. Ясно, что  $1 < 4z_1 < 2, 4z_2 \geq 1$ . Поэтому  $4z_3 < 1$ , ибо иначе имелся бы следующий раскрой 4 отрезков  $L$

$$(1, 0, 0), (0, 1, 0), (0, 0, 1), (4z_1 - 1, 4z_2 - 1, 4z_3 - 1).$$



Пусть вектор  $(5z_1, 5z_2, 5z_3)$  плохой. Так как  $z_3 < \frac{1}{4}$ ,  $z_2 < \frac{1}{3}$ , то легко показать, что  $z_1 > \frac{2}{5}$  и стало быть  $5z_1 > 2$ . Поскольку же  $5z_2 > 1$ , то с необходимостью получаем, что  $5z_2 < 1$ , ибо иначе имелся бы следующий раскрой 5 отрезков  $L$

$$(1, 0, 0), (1, 0, 0), (0, 1, 0), (0, 0, 1), (5z_1 - 2, 5z_2 - 1, 5z_3 - 1).$$

Вектор  $(6z_1, 6z_2, 6z_3)$  будет автоматически плохим, поскольку  $6z_1 < 3$ ,  $6z_2 < 2$ ,  $6z_3 < 2$ .

Рассмотрим, наконец вектор  $(7z_1, 7z_2, 7z_3)$ . Хотя  $z_3 < 1/5$ , но  $z_3 > 1/7$ , поскольку  $z_1 < 1/2$ ,  $z_2 < 1/3$ . Так как  $z_2 < 1/3$ , то  $x > 6/13$ . И так как  $1 - e < z_1 + z_2 + z_3 \leq 1$  для достаточно маленького  $e$ , то  $z_2 > 2/7$ . Поэтому имеем неравенства  $7z_1 \geq 3$ ,  $7z_2 \geq 2$ ,  $7z_3 \geq 1$ . Таким образом имеем следующий раскрой 7 отрезков  $L$

$$(1, 0, 0), (1, 0, 0), (1, 0, 0), (0, 1, 0), (0, 1, 0), (0, 0, 1), (7z_1 - 3, 7z_2 - 2, 7z_3 - 1).$$

То есть вектор  $(7z_1, 7z_2, 7z_3)$  хороший.

Другими словами, справедлива следующая лемма.

**Лемма 1** Для любого вектора  $(z_1, z_2, z_3) : (z_1 + z_2 + z_3 \leq 1, z_1 \geq z_2 \geq z_3)$  среди векторов  $i(z_1, z_2, z_3), i = 2, 3, 4, 5, 6, 7$  есть один хороший.

Пусть теперь дана делимая задача MSP3, которая представлена в виде вектора  $K(z_1, z_2, z_3), z_1 + z_2 + z_3 \leq 1$ . Если  $K > 7$ , то с помощью хорошего вектора  $i(z_1, z_2, z_3), 2 \leq i \leq 7$  раскраиваем  $i$  отрезков  $L$  и у нас остается задача  $((K - i)(z_1, z_2, z_3), z_1 + z_2 + z_3 \leq 1$ .

Таким образом, следующий алгоритм решает делимый случай задачи MSP3.

Шаг 1. Представить задачу MSP3 в виде вектора

$$K(z_1, z_2, z_3), z_1 + z_2 + z_3 \leq 1.$$

Шаг 2. Найти хороший вектор  $i(z_1, z_2, z_3), 2 \leq i \leq 7$ .

Шаг 3. Вычислить  $K_1 = \lceil K/i \rceil$ ,  $K_2 = K - K_1 i$ .

Шаг 4.  $K_1 i$  отрезков длины  $L$  раскраиваем с помощью хорошего вектора  $i(z_1, z_2, z_3)$ .

Шаг 5. Оставшуюся задачу

$$K_2(z_1, z_2, z_3), z_1 + z_2 + z_3 \leq 1$$

решаем с помощью известного алгоритма Ленстры для задачи линейного целочисленного программирования с фиксированным числом переменных, поскольку формулировка оставшейся задачи с помощью целочисленного программирования приводит к задаче линейного целочисленного программирования с числом переменных меньше  $21=7*3$ .

Поскольку как легко видеть все шаги предложенного алгоритма имеют полиномиальную сложность, то справедлива

**Теорема 1** Делимый случай задачи MSP3 принадлежит классу  $P$  (классу задач, решаемых полиномиальным алгоритмом).

## 2 Заключение

Задачи теории расписаний являются сложными  $NP$ -трудными задачами. В данном докладе для делимого случая задачи MSP3 (multiprocessor scheduling problem with 3 job length) впервые построен полиномиальный алгоритм.

## Список литературы

1. McCormick S.T. A polynomial algorithm for multiprocessor scheduling with two job lengths //Mathematics of Operations Research. – 2001. – Vol. 26(1). –P. 31-49.

# Методы и Системы Автоматического Реферирования Текста

А.М. Бакиева<sup>1</sup>, Т.В. Батура<sup>2</sup> и А.М. Федотов<sup>3</sup>

<sup>1</sup> Новосибирский государственный университет, Новосибирск, Россия

<sup>2</sup> Новосибирский государственный университет, Новосибирск, Россия

<sup>3</sup> Новосибирский государственный университет, Новосибирск, Россия  
m\_aigerim0707@mail.ru, Tatiana.v.batura@gmail.com, fedotov@nsu.ru

**Аннотация.** Статья представляет собой обзор различных методов автоматического реферирования текстовых документов. Существует три основных подхода к решению задачи создания реферата: частотно-лингвистический, семантический и гибридный. В работе рассмотрены некоторые наиболее распространенные алгоритмы автоматического реферирования, а также приведено сравнение функциональных возможностей систем, реализующих эти алгоритмы. Перспективным направлением в данной области представляется создание универсальных систем, не накладывающих ограничений на тематику документов и позволяющих обрабатывать большие объемы разнородной информации.

**Ключевые слова:** краткое изложение содержания документов, статистические методы, позиционные методы, индикаторные методы, семантические методы, обработки большого объема данных и коммуникации, алгоритм информационно-аналитической деятельности.

## 1 Введение

Автоматическое реферирование (Automatic Text Summarization) — это составление коротких изложений материалов, аннотаций или дайджестов, т.е. извлечение наиболее важных сведений из одного или нескольких документов и генерация на их основе лаконичных отчетов [1]. В современном мире возрастает актуальность применения методов автоматического реферирования и аннотирования. В настоящее время существует проблема информационной перегрузки. Рефераты и аннотации дают возможность установить основное содержание документа и определить необходимость обращения к первоисточнику. Автоматическое реферирование и аннотирование помогает человеку эффективно обрабатывать большие объемы информации.

Существует много путей решения этой задачи, которые довольно четко подразделяются на два направления — квазиреферирование и краткое изложение содержания первичных документов. Квазиреферирование основано на экстрагировании фрагментов документов — выделении наиболее информативных фраз и формировании из них квазирефератов [2].

В рамках квазиреферирования выделяют три основных направления, которые в современных системах применяются совместно:

- статистические методы, основанные на оценке информативности разных элементов текста по частоте появления, которая служит основным критерием информативности слов, предложений или фраз;

- позиционные методы, которые опираются на предположение о том, что информативность элемента текста зависит от его позиции в документе;

- индикаторные методы, основанные на оценке элементов текста, исходя из наличия в них специальных слов и словосочетаний — маркеров важности, которые характеризуют их содержательную значимость. После выявления определенного (задаваемого, как правило, коэффициентом необходимого сжатия) количества текстовых блоков с наивысшими весовыми коэффициентами, они объединяются для построения квазиреферата [2].

Преимущество методов квазиреферирования заключается в простоте их реализации. Однако выделение текстовых блоков, не учитывающее взаимоотношений между ними, часто приводит к формированию бессвязных рефератов. Некоторые предложения могут оказаться пропущены, либо в них могут встречаться слова или фразы, которые невозможно понять без предшествующего пропущенного текста. Попытки решить эту проблему, в основном сводятся к исключению таких предложений из рефератов. Реже делаются попытки разрешения ссылок с помощью методов лингвистического анализа [1].

Краткое изложение содержания первичных документов основывается на выделении из текстов наиболее важной информации и порождении новых текстов, содержательно обобщающие первичные документы. В отличие от частотно-лингвистических методов, обеспечивающих квазиреферирование, подход, основанный на базах знаний, опирается на автоматизированный качественный контент-анализ, состоящий, как правило, из трех основных стадий. Первая — сведение исходной текстовой информации к заданному числу фрагментов — единиц значения, которыми являются категории, последовательности и темы. На второй стадии производится поиск регулярных связей между единицами значения, после чего начинается третья стадия — формирование выводов и обобщений. На этой стадии создается структурная аннотация, представляющая содержание текста в виде совокупности концептуально связанных смысловых единиц.

Семантические методы формирования рефератов-изложений предполагают два основных подхода: метод синтаксического разбора предложений и методы, опирающиеся на понимание естественного языка. В первом случае используются деревья разбора текста. Процедуры автоматического реферирования манипулируют непосредственно деревьями, выполняя перегруппировку и сокращение ветвей на основании соответствующих критериев. Такое упрощение обеспечивает построение реферата — структурную «выжимку» исходного текста.

Второй подход основывается на системах искусственного интеллекта, в которых также на этапе анализа выполняется синтаксический разбор текста, но синтаксические деревья не порождаются. В этом случае формируются семантические структуры, которые накапливаются в виде концептуальных подграфов в базе знаний. В частности, известны модели, позволяющие производить реферирование текстов на основе психологических ассоциаций сходства и контраста. В базах знаний избыточная и не имеющая прямого отношения к тексту информация устраняется путем отсечения некоторых подграфов. Затем информация подвергается агрегированию методом слияния оставшихся графов или их обобщения. Для выполнения этих преобразований выполняются манипуляции логическими предположениями, выделяются определяющие шаблоны в текстовой базе знаний. В результате преобразования формируется концептуальная структура текста — аннотация, т.е. концептуальные «выжимки» из текста [3].

Многоуровневое структурирование текста с использованием семантических методов позволяет подходить к решению задачи реферирования путем:

- удаления малозначащих смысловых единиц. Преимуществом метода является гарантированное сохранение значащей информации, недостатком — низкая степень сжатия, т.е. сокращения объема реферата по сравнению с первичными документами;
- сокращения смысловых единиц — замена их основной лексической единицей, выражающей основной смысл;
- гибридного способа, заключающегося в уточнении реферата с помощью статистических методов, с использованием семантических классов, особенностей контекста и синонимических связей.

Существуют общедоступные программы квазиреферирования, например, в состав сервисных возможностей системы Microsoft Word 97 входит сервис «Автореферат».

Автоматическое реферирование получило значительную актуальность в связи с развитием Internet и каталогов информационных ресурсов. Для экономии времени поиска пользователям предлагаются каталоги аннотаций и рефератов источников. Формирование рефератов и аннотаций вручную требует колоссальных человеческих ресурсов, поэтому и возникла задача создания методов автоматического реферирования и аннотирования. Автоматическое реферирование и аннотирование — одно из направлений компьютерной обработки естественно-языковых текстов. И в этом качестве оно относится к фундаментальным технологиям ИИ.

Основные тенденции для данной области:

- аннотированные каталоги перерастают в гипертекстовые (с их минусами и плюсами);
- на всех крупных сайтах Internet предусматривают оглавления (карта сайта — sitemap) и функции поиска по сайту;
- использование онтологических словарей-тезаурусов общего и специализированного назначения, а также методов ИИ.

Потребности в средствах автоматического реферирования и аннотирования испытывают: корпоративные системы документооборота; поисковые машины и каталоги ресурсов Internet; автоматизированные информационно-библиотечные системы; каналы вещания; службы рассылки новостей и др.

## 2 Подходы к автоматическому реферированию

В теории автоматического реферирования различают [5] три основных подхода. Первый из них не предполагает опоры на знания, связанные с текстом на естественном языке. В системах такого типа применяется универсальная база правил, не зависящая от языка текста. Второй подход предусматривает выделение различных уровней понимания текста, что требует использования наряду с универсальными правилами базы знаний и базы лингвистических правил, зависящих от языка. Третий подход является гибридным. Он сочетает лучшие стороны первых двух.

В системах первого типа [5] (т.е. воплощающих первый подход) применяется метод составления выдержек. Он реализуется в два этапа. На первом проводится сопоставление текста и фразовых шаблонов, в результате чего выделяются блоки наибольшей лексической и статистической релевантности. Это автоматическое определение частот использования отдельных слов и сочетаний в исходном документе. На втором — путем соединения выделенных фрагментов формируется итоговый документ.

Для реализации первого этапа используют модель линейных весовых коэффициентов. В соответствии с ней каждому блоку  $U$  текста оригинала автоматически (на основании определенных правил) приписываются весовые коэффициенты:

- $k_1$ , зависящий от расположения блока  $U$  в оригинале;
- $k_2$ , зависящий от частоты появления блока в оригинале;
- $k_3$ , зависящий от частоты использования блока в ключевых предложениях;
- $k_4$ , отражающий показатели статистической значимости блока.

Затем по значениям  $k_1$ ,  $k_2, k_3$  и  $k_4$  и коэффициентам настройки программы реферирования  $\alpha_1$ ,  $\alpha_2$ ,  $\alpha_3$  и  $\alpha_4$  вычисляется коэффициент важности блока  $B(U) = \alpha_1 k_1 + \alpha_2 k_2 + \alpha_3 k_3 + \alpha_4 k_4$ . По коэффициентам важности выполняется отбор блоков в реферат.

Для вычисления каждого весового коэффициента используется своя группа правил. Для  $k_1$  они учитывают расположение блока:

- во всем тексте или некотором разделе;
- в начале, середине или конце текста;
- во вводной части, заключении и т.д.

Для  $k_2$  правила учитывают результаты автоматической индексации документа (например, соотношение между частотой появления термина в документе и в наборе документов).

Для  $k_3$  учитывается наличие в блоке таких ключевых фраз и выражений, как «в заключение...», «в данной статье...», «согласно результатам анализа...», «отличный от...», «мало-значущий...» и т.п.

Для  $k_4$  правила учитывают вхождение термина в заголовки, колонтитулы, первый параграф текста, пользовательский профиль запроса и т. п. Настройка с помощью коэффициентов  $\alpha_1$ ,  $\alpha_2$ ,  $\alpha_3$  и  $\alpha_4$  позволяет управлять степенью сжатия.

На рис. 1 изображена обобщенная архитектура системы автоматического реферирования первого типа.



Рис. 1. Рис. 1. Обобщенная архитектура системы автоматического реферирования.

Рис. 1. Обобщенная архитектура системы автоматического реферирования

Главное достоинство описанной модели линейных весовых коэффициентов заключается в простоте ее реализации, а главный недостаток связан с возможностью формирования бес-связных рефератов, не учитывающих контекст. Для его устранения вводится этап ручного редактирования результатов. Схема автоматического определения критериев адекватного выбора фрагментов оригинала для реферата используется в системе Inxight Summarizer (рис. 2).

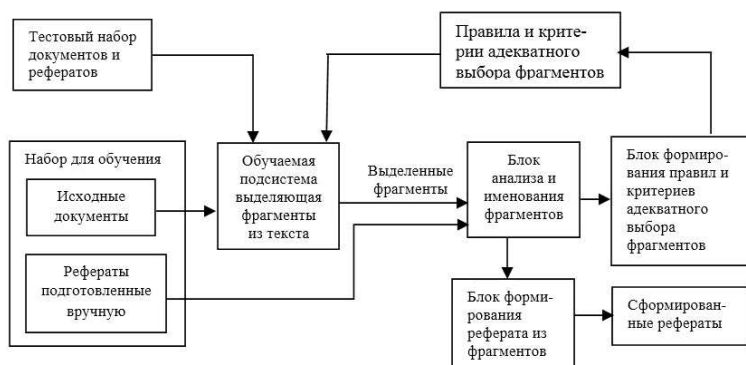


Рис. 2. Рис. 1. Обобщенная архитектура системы автоматического реферирования.

Рис. 2. Схема автоматического определения критериев адекватного выбора фрагментов

Обучение (настройка) системы осуществляется на наборах текстов и рефератов, составленных для них вручную при различных критериях сжатия.

Человеку, уловившему общий смысл информации, легче выделить главное и кратко изложить содержание. Это и обуславливает создание реферирующих систем второго типа [5]. Для таких систем требуются:

- мощные вычислительные ресурсы;
- развитые грамматики и словари;
- развитые средства синтаксического разбора;
- средства генерации естественно-языковых конструкций;
- онтологические справочники.

В этих системах реализуются три подхода:

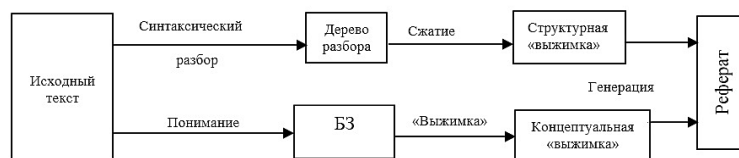
- 1) традиционный метод синтаксического разбора;
- 2) подход с опорой на понимание естественного языка;
- 3) комбинированный подход.

В первом случае для построения деревьев разбора используется синтаксическая информация. Процедуры сжатия манипулируют деревьями с целью сокращения скобок, подчиненных предложений и т.д. При этом дерево разбора упрощается до «структурной выжимки».

При втором подходе в результате разбора строится не дерево, а семантическая сеть текста. Другими словами, в ходе разбора выделяются концептуальные репрезентативные структуры исходного текста. Из них удаляется избыточная информация: поверхностные суждения, концептуальные подграфы. Далее выполняется агрегирование и обобщение информации: слияние некоторых концептуальных графов на базе правил. В результате получается «концептуальная выжимка».

Обобщенная схема для этих двух методов представлена на рис. 3.

Стадии синтеза реферата в обоих подходах почти совпадают (используется генератор текста).



**Рис. 3.** Рис. 1. Обобщенная архитектура системы автоматического реферирования.

Рис. 3. Два основных подхода к формированию реферата в системах с опорой на знания. Для функционирования подобных систем необходимы:

- исчерпывающие словари (тезаурусы) типа WordNet;
- онтологические справочники типа Cus и Penman Upper Model;
- большие объемы тестовых файлов с текстами (например, The Wall Street Journal или Perm Treebank от Linguistic Data Consortium).

### 3 Виды алгоритмов автоматического реферирования

Во всем мире были разработаны различные алгоритмы, из всех более используемых были выбраны различные виды алгоритма различных методов. По способу построения текста методы автоматического реферирования и аннотирования делятся на две группы: извлекающие и генерирующие. При использовании извлекающих методов из исходного текста выделяются наиболее важные фрагменты (предложения, абзацы). При этом данные фрагменты не обрабатывают, а извлекают в таком порядке и виде в каком они приведены в

тексте. Оптимальным решением данной задачи мы выбрали несколько алгоритмов, анализ которых приведен ниже. Это алгоритмы Freq, LRU-K, семантический анализ, алгоритм частотного анализа.

Алгоритм Freq [6], учитывает частоту слов в окне длиной в 1000 слов вокруг анализируемого фрагмента (то есть фрагмент находился в середине окна). Отбирается 10 слов, которые встречались в данном фрагменте наиболее часто. Для расчетов используется следующая формула (1) для вычисления веса:

$$W_{freq} = Wb + \text{Sum}x(\log_2 Fk), \text{ где (1)}$$

$Wb$ — вес, вычисленный по базовому алгоритму;

$Fk$ — сколько раз слово встретилась в окне в 1000 слов, включающем фрагмент.

Таким образом, наибольший вес получают те фрагменты, которые кроме того, что содержат наибольшее число слов запроса, но и большее количество слов часто встречающихся в документе. Нам неизвестно упоминаний в литературе использования подобных алгоритмов для анализа текста.

В последнее время определенную популярность приобрели более сложные статистические модели, обычно основанные на Марковских цепях. Однако эти модели достаточно сложны и имеют еще более высокую вычислительную сложность. Алгоритм LRU-K [7], является вариантом алгоритма «последний недавно использованный». Мы исходили из известного в психологии предположения, что человек в быстрой памяти сохраняет только относительно малое количество объектов, поэтому алгоритмы класса «последний используемый» должны показывать хорошие результаты.

Алгоритм делает следующее:

1) при инициализации создается 3 структуры данных: массивы слов и 2 массива с указателями на слова ( $array1$  и  $array2$ ) и длинами  $k$ ;

2) для каждого слова при обработке документа производится следующие действия:

а) поиск в массиве слов.

1. Если слово не найдено, то ссылка на него помещается в массив  $array1$  в первую позицию, остальные позиции в массиве сдвигаются, самое последнее слово удаляется из  $array1$  и из массива слов.

2. Если слово найдено и встречается в  $array1$ , то оно из него удаляется и переносится на первую позицию в  $array2$ . При этом, если  $array2$  полностью заполнен, то последнее слово из него так же удаляется, как и в первом случае.

3. Если слово найдено и уже есть в  $array2$ , то оно просто перемещается на первую позицию.

Легко можно показать, что если бы слова в тексте имели равную вероятность появления, то после обработки фрагмента текста, содержащего слов намного больше  $k$ , содержимое массива  $array2$  совпадало бы с  $k$  наиболее часто встретившихся слов. То есть данный алгоритм можно рассматривать как один из вариантов оценки локальной частоты терминов, при предположении равномерного распределения слов.

Однако, предлагаемый алгоритм, кроме этого, должен выделять слова, которые имеют не только высокую частоту, но и равномерно распределенные вблизи выбираемого фрагмента.

Для автореферирования необходимо определять семантическую близость между предложениями текста [11]. Алгоритм семантического анализа может решить эту задачу. Назовем этот метод семантической матрицей. Каждому предложению текста соответствует определенная семантическая мера, это частота вхождения слова в иерархию классификатора умноженная на глубину иерархии. Потом необходимо определить семантическую связность пар предложений текста по формуле (2):  $Scoup = Mcom - Msent$ , где (2)



$M_{com}$  — совокупная мера двух предложений;

$M_{sent}$  — мера каждого предложения.

Получаем наборы семантической связности предложений, из которых потом формируем матрицу. В связи с особенностью алгоритма, матрица симметрична, на главной диагонали получаем нули, так как связность предложения с самим собой в данный момент не нужна при анализе. Элементы матрицы выстраиваются по принципу того, что пара предложений с наибольшей связностью — это первая часть элементов аннотации. Следующая пара предложений проверяется на семантическую связность с группами, которые имеют большую связность, и добавляются к той или иной группе, связность с которой наибольшая, либо образуют независимую группу, если связность отсутствует. Как результат получается несколько фрагментов с семантической связностью. В целом алгоритм реферирования таков:

- 1) морфологический анализ;
- 3) устранение омонимии слов в предложении;
- 4) построение семантической матрицы;
- 5) выделение семантически связных групп предложений.

Этот алгоритм, при реализации поможет пользователю выполнять такие функции как задание степени сжатия, порог семантической связности, количество групп рефератов одного текста.

Алгоритм частотного анализа подробно описан в [14]. Вычисление базовых весов словосочетаний. Существуют шаблоны (согласованное прилагательное + существительное, например: «первая леди», «последний звонок») и (существительное + существительное в родительном падеже, например: «замок зажигания», «карта мира»). Базовый вес словосочетания вычисляется как удвоенный  $TF \cdot IDF$  менее редкого слова, входящего в словосочетание.  $TF \cdot IDF$  [14] — увеличивает значимость слов, которые часто встречаются в документе, но редко в обучающем корпусе, где  $TF$  — частота термина в текущем документе,  $IDF$  — логарифм отношения количества всех документов к количеству документов содержащих данный термин;

Далее словосочетания рассматриваются наряду с другими ключевыми словами. Алгоритм частотного анализа.

Вычисление весов ключевых слов. Для вычисления окончательного веса слова базовый (частотный) вес умножается на коэффициент

$$K = 1 + KB + KU + KI + KT + KH + KQ, \text{ где}$$

$KB, KU, KI$  — равны двум, если слово выделено соответственно жирным шрифтом, подчеркиванием или курсивом;

$$KT = 10, \text{ если слово присутствует в заголовке};$$

$$KH = 5, \text{ если слово встречается в подзаголовках Н1..Н4};$$

$$KQ = 500, \text{ если слово присутствует в запросе.}$$

Все коэффициенты подобраны эмпирически, могут настраиваться в программе.

Вычисление весов предложений. Граница предложения определяется на основе шаблонов:

- [.|?!|!|... ] [пробел] [A..Z|A..Я|0..9|];
- [<P>|</TITLE>].

Итоговый вес предложения вычисляется по формуле:

$$P = L \cdot I \cdot e^{-\left(\frac{SL-OL}{K}\right)^2} \cdot \left(1 + \frac{2q^2}{QL}\right) \cdot \sum_{i=1}^{SL} W_i, \text{ где}$$

$L$  — повышающий коэффициент для первых и последних четырех предложений документа;

$I$  — понижающий коэффициент для вопросительных предложений;

$SL$  — длина предложения в словах;

$QL$  — длина запроса в словах;

$Wi$  — вес  $i$ -го слова в предложении;

$q$  — количество слов запроса в предложении;

$OL$  — «оптимальная» длина предложения для реферата;

$K$  — коэффициент уменьшения веса предложения при отклонении от «оптимальной» длины.

Формирование реферата с заданным количеством предложений. Предложения сортируются в соответствии с вычисленным весом по убыванию. Первое предложение помещается в реферат. Каждое следующее предложение берется из списка и сравнивается с предложениями реферата. Предложение отбрасывается, если оно имеет 80 процентов или более общих слов с предложениями реферата. Процесс повторяется, пока не будет отобрано заданное количество предложений. Отобранные предложения выдаются в том порядке, в каком они находились в тексте.

Таким образом, мы можем сделать вывод, что современные подходы к автоматическому реферированию и аннотированию отличаются разнообразием используемых методов. Материалами для реферата и аннотации могут выступать не только тексты, но и числовые данные.

#### 4 Существующие системы автоматического реферирования

На международном рынке представлено множество программных продуктов, которые позволяют создавать авторефераты для текстовых файлов. Ориентированы они преимущественно для файлов, содержащих текст на английском языке. Существуют три наиболее популярные программы автореферирования: МЛ Аннотатор, Золотой ключик, TextAnalyst. Методы автоматического реферирования и аннотирования подразделяются на поверхностные и глубинные. «Поверхностные методы» базируются на «экстрагировании» текста, т.е. извлечении из него фрагментов, оцениваемых системой как важнейшие, и объединении их в реферат или аннотацию. Важность фрагментов определяется:

- по маркерам важности (оборотам типа «идея ... состоит в...», «главным результатом ... является...», «в заключении нужно сказать, что...» и т.д.);
- по количеству заданных в запросе ключевых слов, входящих во фрагмент, и др.

При объединении выделенных предложений в реферат или аннотацию учитываются их зависимости друг от друга (удаленность выделяемых мыслей). «Стыки» между предложениями (фрагментами) «сглаживаются». «Глубинные методы» развиваемые в настоящее время, базируются на применении тезаурусов и развитых механизмов синтаксического разбора текста [11].

Традиционные отечественные системы автоматического реферирования и аннотирования, реализующие поверхностные методы приведены в таблице 1.

Таблица 1. Список отечественных систем автоматического реферирования и аннотирования, реализующие поверхностные методы

Таблица 2. Список зарубежных систем автоматического реферирования и аннотирования, реализующие следующие функции

Перечисленные средства обеспечивают выбор оригинальных фрагментов из исходных документов и соединение их в короткий текст. Сделаем два замечания. Во-первых, источниками информации для рефератов и аннотаций могут служить не только тексты, но и видеозаписи, разнообразные табличные документы и т.д. Во-вторых, краткое изложение

Наименования отечественных систем	Основные функции
Microsoft Word 97	функция автоматического реферирования
ОРФО 5.0	(разработчик — компания Информатик), включающую функцию автоматического аннотирования русских текстов
Либретто	(разработчик — компания "МедиаЛингва"), обеспечивающую автоматическое реферирование и аннотирование русских и английских текстов (система встраивается в Word)
"МедиаЛингва Аннотатор" SDK 1.0 Следопыт	служащий инструментарием для реализации функций автоматического реферирования и аннотирования в прикладных ИАС [18]. поисковую система включающую в средства автоматического реферирования и аннотирования документов
Поисковая машина «Золотой Ключик»	Это программная библиотека, работающая по принципу фильтрации на базе тезауруса. Как входные данные программе подается произвольный текст на русском языке, на стандартном выходе программа формирует аннотацию данного текста и список рубрик, к которым относится данный текст. В качестве аннотации используются предложения из входного текста, наиболее полно отражающие тематику текста. При рубрикации текста используется фиксированный список заранее определенных рубрик [18]
Inxight Summarizer	выделяет наиболее весомые предложения из текста используя статистические, алгоритмы, либо слова-подсказки [16]
eXtragon	набор исходных данных, созданный на основе оценивавшихся запросов дорожек поиска по Веб-коллекции и по коллекции нормативно-правовых документов
Galaktika-ZOOM	интеллектуальный поиск по ключевым словам с учетом морфологии русского и английского языков, а также и формирование информационных массивов по конкретным аспектам
InfoStream	Технология позволяет создавать полнотекстовые базы данных и осуществлять поиск информации, формировать тематические информационные каналы, автоматически рубрицировать информацию, формировать дайджесты, таблицы взаимосвязей понятий (относительно встречаемости их в сетевых публикациях), гистограммы распределения весовых значений отдельных понятий, а также динамики их встречаемости по времени.
TextAnalyst	Программа создана в Московском Научно-производственном Инновационном Центре «МикроСистемы» [20] TextAnalyst работает только с русским языком, выделяя именные группы и строя на их основе семантическую сеть — структуру взаимозависимостей между именными группами.

предполагает передачу основной мысли не обязательно теми же словами. Из рассмотренных программных продуктов, на данный момент можно выделить «Золотой ключик» как наименее гибкий и функциональный инструмент для задач автореферирования [12]. TextAnalyst [12] как программный продукт, основанный на алгоритмах создания семантических сетей, проявляет гибкость при работе с базами знаний и алгоритмами формирования смыслового портрета. Тот факт, что в Аннотаторе [12] применяются алгоритмы определения семантических весовых коэффициентов предложений и специальные вероятностные модели, но при этом нет возможности создания смыслового портрета, позволяет относить Медиа Лингво Аннотатор [12] к промежуточному классу программных продуктов между «Золотой ключик» и TextAnalyst. Рассмотренная программа Extractor [12] в большей степени подготовлена к работе в сети Интернет (например, в составе поисковых машин). Это делает Extractor более популярной и востребованной на международном рынке услуг автореферирования и поиска информации. Наибольшие перспективы в данной области видятся в развитии взаимодействия и совмещения алгоритмов формирования семантических сетей и алгоритмов поисковых машин в глобальной сети Интернет. И создание на базе совмещённых алгоритмов новых, общедоступных сервисов интеллектуального поиска информации, а также систем автореферирования больших объёмов текстовой информации. Использование общедоступных

Наименования зарубежных систем	Основные функции
Extractor	Способы определения наиболее вероятных ключевых фраз, используя контекстную информацию, служат основой для идеи выявления в тексте переформулированных смысловых конструкций [17].
Autonomy Knowledge Server	анализ текстов и идентификации ключевых концепций в пределах документов путем анализа корреляции частот и отношений терминов со смыслом текста.
InterMedia Text, Oracle Text	В ходе обработки текст каждого документа подвергается процедурам лингвистического и статистического анализа, в результате чего определяются его ключевые темы и строятся тематические резюме, а также общее резюме — реферат.
SemioMap	SemioMap поддерживает разбиение материала по «папкам», создание отдельной базы данных для каждой папки. Связи между понятиями, которые выявляет SemioMap, базируются на совместной встречаемости фраз в абзацах исходного текстового массива.
Text Miner	Позволяет определять, насколько правдив тот или иной текстовый документ. Обнаружение лжи в документах производится путем анализа текста и выявления изменений стиля письма, которые могут возникать при попытке исказить или скрыть информацию [15]
WebAnalyst	Представляет собой интеллектуальное масштабируемое клиент/серверное решение для компаний, желающих максимизировать эффект анализа данных в Web-среде. Сервер WebAnalyst функционирует как экспертная система сбора информации и управления контентом Web-сайта. Модули WebAnalyst решают три задачи: сбор максимального количества информации о посетителях сайта и запрашиваемых ими ресурсах; исследование собранных данных и генерация персонализированного, на основе результатов исследований, контента.
Intelligent Text Miner (IBM)	Технология эффективного анализа текстовых данных. Представляет собой набор отдельных утилит, запускаемых из командной строки или скриптов независимо друг от друга. Данная система является одним из лучших инструментов глубокого анализа текстов [13]
Oracle Context	Разнообразие источников, форматов, запросов
RCO FX Ru	Программный продукт предназначен для аналитической обработки текста на русском языке. Основной сферой применения программы являются задачи из области компьютерной разведки, требующие высокоточного поиска информации. Например, к ним можно отнести автоматический подбор материала к досье на целевой объект или же мониторинг определенных сторон его активности, освещаемых в СМИ. [21]

сервисов по поиску и автореферированию позволит значительно облегчить задачу. Одним из возможных решений в этой ситуации может стать, создание систем составления краткого изложения полнотекстовых документов на базе общедоступных сервисов. Представляется возможным проектирование и разработка совмещённых поисковых систем с системами автореферирования.

Составив обзор по всем существующим программам, хочется добавить следующие критерии и требования с точки зрения пользователя. Это пользовательские требования и характеристика интерфейса.

Пользовательские требования:

- Работа с метаданными документов по информационным ресурсам
- Использование ключевых терминов с помощью тезауруса

Характеристики интерфейса:

- исследование целесообразности использования тех или иных технологий,
- разработка архитектуры системы,
- реализация алгоритмов реферирования,
- исследование инструментов тестирования систем автоматического реферирования
- оценка качества работы системы.

- создание систем составления краткого изложения полнотекстовых документов на базе общедоступных сервисов
- Развитие методов с помощью онтологий и семантических методов
- Исследование алгоритмов реферирования текстов
- Генерация ответов на сложные вопросы при помощи краткого содержания нескольких документов.

## 5 Заключение

Можно выделить следующие новые задачи, связанные с компьютерным реферированием.

1. Создание одноязычных рефератов из источников на разных языках из одного источника при наличии тезауруса.

2. Построение рефератов по гибридным источникам, включающим как текстовые, так и числовые данные в разных формах (таблицы, диаграммы, графики и т. д.).

3. Создание рефератов на основе массивов документов. Например, построение единого реферата по сборнику тезисов докладов научной конференции. Одна из областей применения подобных средств — формирование новостных сообщений по газетным источникам.

Задача автоматического составления короткого реферата текстового документа возникает в контексте Веб-поиска. В списке результатов поисковой машины наряду с заголовком и адресом обычно присутствуют сниппеты (англ. snippets) — фрагменты документа, содержащие слова запроса. Назначение сниппетов — помочь пользователю составить представление о документе и решить, имеет ли смысл обращаться к оригиналу [5].

В статье были рассмотрены различные методы для задачи автоматического реферирования текстов, различные алгоритмы для сравнения работ систем и программ по автореферированию и приведен обзор систем, реализующих эти методы. Как правило, выделяют два подхода к автоматическому составлению короткого реферата (или аннотации) текстовых документов. Первый предполагает извлечение наиболее важных фрагментов текста из одного или нескольких документов, второй основывается на знаниях о морфологии, синтаксисе и семантике конкретного языка с целью генерации лаконичных отчетов. Наиболее эффективным методом является семантический метод, который работает следующим образом: из выбранного текста удаляется избыточная информация: поверхностные суждения, концептуальные подграфы. Далее выполняется агрегирование и обобщение информации: слияние некоторых концептуальных графов на базе правил. В результате получается концептуальная выжимка. Важной характеристикой при выборе метода реферирования для Web-приложений является универсальность, т.е. выбранный метод не должен накладывать ограничений на тематику. Для работы с тематическими каталогами существенной является возможность настройки на конкретную предметную область.

Наибольшие перспективы в данной области видятся в развитии взаимодействия и совмещения алгоритмов формирования семантических сетей и алгоритмов поисковых машин в глобальной сети Интернет. И создание на базе совмещённых алгоритмов новых, общедоступных сервисов интеллектуального поиска информации, а также систем автореферирования больших объёмов текстовой информации.

## Список литературы

1. Luhn H. The automatic creation of literature abstracts // In IBM Journal of Research and Development, —New York, 1958. — Vol. 2(2). — P. 159–165.
2. Гинкул А.С. Сравнительный анализ существующих систем автоматического реферирования текста // Політ. сучасні проблеми науки — Киев, 2012. —С. 255.

3. Луканин А.В. Автоматическая обработка естественного языка — Челябинск: Изд. центр ЮУрГУ, 2011. — 70 с.
4. Adwait Ratnaparkhi. Learning to parse natural language with maximum entropy models. // Machine Learning, —New York, 1999. —341(3) — P.151-176
5. Автоматическая обработка текста. [Электрон. ресурс]. — 2006. — URL: <http://aot.ru/> (дата обращения: 12.03.2015)
6. Анализ алгоритмов автоматического реферирования текста Е.А.Гридина // Восточно-Европейский журнал передовых технологий, —Харьков, 2011.3/2 ( 51 ) —с.36-38
7. Алгоритмы кэширования. [Электрон. ресурс]. — 2006. — URL: [http://ru.wikipedia.org/wiki/Алгоритмы\\_кэширования](http://ru.wikipedia.org/wiki/Алгоритмы_кэширования). (дата обращения: 12.03.2015)
8. Хан У., Мани И. Системы автоматического реферирования. [Электрон. ресурс]. — 2000. — URL: [http://www.osp.ru/os/2000/12/067\\_print.htm](http://www.osp.ru/os/2000/12/067_print.htm) (дата обращения: 12.03.2015)
9. Jurafsky D. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition / D. Jurafsky, J.H. Martin. — New Jersey: Prentice Hall, 2000. —р.934
10. Дж. Солтон. Динамические библиотечно-информационные системы // Издательство «Мир», —1979. С.41-50
11. Интеллектуальные информационные технологии [Электрон. ресурс]. — 2000. — URL: <http://www.irkinfo.ru/intellektualnye-informatsionnye-tehnologii-str34.html> (дата обращения: 12.03.2015)
12. SoftLine. МЛ Аннотатор 1. 0 для Windows 95 и Windows NT, [Электрон. ресурс]. — 2011. — URL: [http://www.softline.ru/products/MediaLingua/MlAnnotator/MlAnnotator1Win\\_full.asp](http://www.softline.ru/products/MediaLingua/MlAnnotator/MlAnnotator1Win_full.asp) (дата обращения: 12.03.2015)
13. Кутукова. Е.С. Технология Text mining // SWorld: Перспективные инновации в науке, образовании, производстве и транспорте. — Одесса, 2013. —с.136-138
14. Павел Браславский. eXtragon: экспериментальная система для автоматического реферирования веб-документов, [Электрон. ресурс]. — 2007. — URL: [http://www.romip.ru/romip2005/03\\_extragon.pdf](http://www.romip.ru/romip2005/03_extragon.pdf) (дата обращения: 12.03.2015)
15. Харламов А.А. Автоматический структурный анализ текстов // Открытые системы. —Москва —2002. — № 10. —С. 16–22.
16. Kupiec J., Pederson J. and Chen F. A trainable document summarizer. // In Proceedings of the 18th ACM/SIGIR Annual Conference on Research and Development in Information Retrieval, Seattle, 1995. — P. 68–73.
17. А. Михаилян. Некоторые методы автоматического анализа естественного языка, используемые в промышленных продуктах, — [Электронный ресурс]. —2000. —URL:<http://www.inteltec.ru/publish/articles/textan/natlang.shtml>. (дата обращения: 02.05.2015)
18. Ступин В.С. Система автоматического реферирования методом симметричного реферирования // Компьютерная лингвистика и интеллектуальные технологии. Труды межд. конференции «Диалог 2004». — М.: Наука, 2004. —С. 579–591.
19. Г. Леліков, В. Сороко, О. Григор'єв, Д. Ланде Моніторинг діяльності органів виконавчої влади із застосуванням комп'ютерної системи контент-аналізу електронних ЗМІ // Вісн. держ. служби України. —2002. — № 2. —С. 21–38.
20. Танатар Н.В., Федорчук А.Г. Интеллектуальные поисково-аналитические системы мониторинга СМИ // Научно-практический и теоретический сборник. —Киев, —2008. —477 с.
21. RCO Fact Extractor Desktop , — [Электронный ресурс]. —2012. —URL:<http://axoft.ru/software/RCO/RCO-Fact-Extractor-Desktop/> (дата обращения: 02.05.2015)

# Логический Подход к Организации Многокритериального Атрибутного Разграничения Доступа

Р.Г. Бияшев<sup>1</sup>, М.Н. Калимолдаев<sup>2</sup> и О.А. Рог<sup>3</sup>

<sup>1</sup> Институт информационных и вычислительных технологий КН МОН РК, Алматы, Казахстан

<sup>2</sup> Институт информационных и вычислительных технологий КН МОН РК, Алматы, Казахстан

<sup>3</sup> Институт информационных и вычислительных технологий КН МОН РК, Алматы, Казахстан  
brg@ipic.kz, mnk@ipic.kz, olga@ipic.kz

**Аннотация.** Построена формальная логическая теория для моделирования многокритериального атрибутного разграничения доступа путем построения систем, осуществляющих разграничение доступа субъектов к информационным объектам по ряду параметров на основе одновременного применения ряда политик безопасности, представляемых унифицированным образом в виде задач удовлетворения ограничениям.

Синтаксис формальной системы описывается многосортным языком логики с переменными в виде множеств. Параметрическая интерпретация языка дает набор политик разграничения доступа в виде семантических доменов атрибутов, на основе которых создаются модели для разграничения доступа по отдельным параметрам. Набор этих моделей составляет создаваемую систему.

Предлагаемый путь конструирования систем многокритериального атрибутного разграничения доступа позволяет осуществить их дальнейшую реализацию на языках логического программирования с ограничениями.

**Ключевые слова:** защита информации, атрибутное разграничение доступа, АВАС, множественная категоризация, семантический домен, задача удовлетворения ограничений, формальная теория, язык логики, интерпретация.

## 1 Многокритериальное атрибутное разграничение доступа и логический подход к его реализации

Известные модели безопасности, регулируя доступ субъектов к объектам, рассматривают одно свойство сущности в качестве параметра разграничения доступа. К их числу, в частности, относятся модель безопасности Белла и ЛаПадулы основанная на предоставлении доступа по результатам сравнения уровней конфиденциальности субъектов и объектов; ролевая модель разграничения доступа (RBAC), которая разграничивает доступ по видам операций с информационными объектами, объединенных в группы - роли; тематическая модель, предоставляющая доступ на основе принадлежности документа той или иной тематике, и организационная - принадлежности отделу организации. [1].

Разграничения доступа по отдельному признаку недостаточно для отражения реальных ситуаций, имеющих место в условиях функционирования информационных систем, где данные характеризуются рядом разнотипных свойств, требующих учета в качестве параметров разграничения доступа.

В настоящее время разрабатываются методы атрибутного разграничения доступа (АВАС), призванные преодолеть ограничения, присущие широко распространенным моделям, основанные на представлении характеристик сущностей - объектов и субъектов разграничения доступа - в виде атрибутов.

Атрибут является парой имя-значение, выражающей то или иное свойство сущности. Сущности в моделях атрибутного разграничения доступа заменяются наборами своих атрибутов, доступ субъектов к объектам разрешается на основе оценки их значений.

Применение атрибутно-ориентированного разграничения доступа оправдано в быстро развивающихся открытых распределенных системах, характеризующихся наличием большого числа пользователей и ресурсов, таких, например, как грид- и облачные вычисления, где сущности описываются своими характеристиками, а не предопределенными идентификаторами. Кроме того, подобные вычислительные среды организуются на основе сетей доменов, имеющих собственные наборы политик безопасности, что требует наличия возможностей гибкой настройки правил разграничения доступа.

К числу основных недостатков атрибутного разграничения доступа (АВАС) следует отнести сложность инжиниринга атрибутов, в частности, необходимость исключения возможности пересечения подмножеств атрибутов, подчиненных разным группам [2-4].

Эти и другие недостатки предполагается устранить в предлагаемой модели многокритериального атрибутного разграничения доступа, в которой применяется единообразное представление различных политик безопасности и соответствующих им моделей, дающее возможность их одновременного применения в рамках одной системы, что обеспечивает управление разграничением доступа по нескольким критериям. Для этого модель производит множественную категоризацию сущностей путем их многокритериальной оценки [5].

Множество атрибутов делится на категории, каждая из которых представляет определенную политику разграничения доступа. На множестве всех возможных значений атрибутов категории устанавливается отношение частичного порядка и определяются операции сравнения значений, что позволяет реализовать для каждой политики отдельный контроллер доступа.

Различные политики и соответствующие им модели разграничения доступа определяются подвидами структур категорий.

Так, линейно упорядоченное множество уровней классификации соответствует модели безопасности Белла и ЛаПадулы. Структура в виде дерева используется в моделях, в которых привилегии доступа образуют иерархические группировки. Например, названия тематик и отделов в тематической и организационной моделях разграничения доступа. В ролевой модели разграничения доступа операции доступа, такие как чтение, запись, корректировка, объединяются в группы "автор" "редактор" "рецензент" "администратор". Структура категории в виде скалярного множества может быть использована вариантами моделей безопасности без иерархической группировки привилегий.

В соответствии с особенностями защиты информации данной вычислительной среды, производится отбор моделей, из которых конструируется многокритериальная система разграничения доступа, обеспечивающая доступ субъектов к объектам при выполнении всех условий, выдвигаемых выбранными моделями безопасности.

В данной работе предлагается следующий подход к реализации подобных систем.

Множество значений атрибутов категории с определенными на нем отношениями и операциями рассматривается как алгебраический тип субъектов и объектов разграничения доступа. Типизация сущностей в системе означает, что субъектам и объектам в качестве меток безопасности по категории присваиваются значения этого типа с дальнейшим применением операций сравнения для разрешения/запрета доступа по категории. Многокритериальная система атрибутного разграничения доступа осуществляет полиморфную типизацию сущностей путем присвоения им значений атрибутов разграничения доступа по всем категориям [6].

Определенный таким образом алгебраический тип категории является реализацией механизма разграничения доступа для политики безопасности, соответствующей данной категории. Объединение механизмов разграничения доступа по всем категориям образует механизм разграничения доступа всей системы.



Показана возможность представления алгебраических типов различных категорий в виде задач удовлетворения ограничений, задаваемых на конечных доменах со структурами в виде скалярных множеств, линейных списков или деревьев - подструктур частично упорядоченного множества, являющегося компонентой исходного алгебраического типа [7-8].

Для моделирования систем многокритериального атрибутного разграничения доступа сначала строится концептуальная модель, которая преобразуется в ряд проектных моделей, реализующих различные политики атрибутного разграничения доступа, из которых затем составляются конкретные варианты систем.

В качестве концептуальной модели создается формальная теория, содержащая много-сортный язык логики с константами и переменными в виде множеств, функциональные и предикатные символы которого обозначают операции определения значений меток безопасности сущностей и сравнения их значений на конечном структурированном домене.

Далее формальная теория интерпретируется в зависимости от структуры множеств, на элементах которых формулируются задачи удовлетворения ограничений. В результате первого этапа интерпретации получается ряд проектных моделей в виде программ на языке логического программирования с ограничениями, соответствующих политикам разграничения доступа, отобранным для использования в конструируемой системе. Данный этап соответствует стадии настройки структуры и значений категорий с помощью команд администратора системы многокритериального атрибутного разграничения доступа.

Второй этап интерпретации заключается в подстановке значений в полученные программы и соответствует стадии функционирования, когда в системе регистрируются пользователи и ресурсы, при этом значениями их меток безопасности заполняется матрица доступа. Пользователи выдают запросы на доступ к информации, система, путем оценки значений их меток безопасности выдает разрешение или запрет на доступ.

Подобный подход позволяет моделировать как различные политики разграничения доступа, так и архитектуру и функционирование многокритериальных систем разграничения доступа в которых они применяются.

## 2 Заключение

Построена многокритериальная модель атрибутного разграничения доступа, синтаксис которой описывается многосортным языком логики с переменными в виде множеств. Параметрическая интерпретация языка дает ряд моделей, использующих частично упорядоченные множества с их алгебраическими операциями для реализации различных атрибутных политик безопасности.

## Список литературы

1. Гайдамакин Н.А. Теоретические основы компьютерной безопасности. – Екатеринбург: издательство Уральского университета, 2008. – 212 с.
2. Jin X., Krishnan R., Sandhu R. A unified attribute-based access control model covering DAC, MAC and RBAC // Proceedings of the 26th Annual IFIP WG 11.3 conference on Data and Applications Security and Privacy (DBSec'12), 2012. – P. 41-45.
3. Wang L., Wijesekera D., Jajodia S. A logic-based framework for attribute based access control // Proceedings of the 2004 ACM workshop on Formal methods in security engineering (FMSE '04). ACM, New York, NY, USA, 2004. – P. 45-55.
4. Khan A.R. Access control in cloud computing environment // ARPN Journal of Engineering and Applied Sciences. – 05, 2012. – № 7(5). – P. 613-615.
5. Калимолдаев М.Н., Бияшев Р.Г., Рог О.А. Формальное представление функциональной модели многокритериальной системы разграничения и контроля доступа к информационным ресурсам // Проблемы информатики. – 2007. – № 1(22). – С. 43-55.

6. Бияшев Р.Г., Калимолдаев М.Н., Рог О.А. Полиморфная типизация сущностей и задача конструирования механизма многокритериального разграничения доступа // Известия НАН РК. Серия физико-математическая. – 2014. – № 5. – С. 33-41.
7. Щербина О.А. Удовлетворение ограничений и программирование в ограничениях // Интеллектуальные системы. – 2011. – № 15(1-4). – С. 53-170.
8. Петров Е.С. Опыт интеграции логического программирования и программирования в ограничениях // Программирование. – 1998. – № 3. – С. 40-49.

# Особенности и Требования к Качеству Программных Средств Космического Назначения

Есмагамбет Исмаил

Институт космической техники и технологий, Алматы, Казахстан  
ismaile@mail.ru

**Аннотация.** В настоящей статье рассматриваются вопросы обеспечения качества программных средств космического назначения (ПСКН). С целью установления особенностей и требований к качеству ПСКН проведена их классификация с учетом их назначения, условий эксплуатации, требований к надежности, безопасности и др. Проанализированы особенности ПСКН критического применения. Обоснованы общие требования и принципы обеспечения качества ПСКН. Широкий спектр требований к качеству ПСКН, в зависимости от их назначения, принципиальных особенностей и условий эксплуатации, приводит к необходимости адаптации и детализации рекомендаций существующих базовых стандартов, регламентирующих качество программного обеспечения.

**Ключевые слова:** программное средство, космическая система, качество, безопасность, особенности, категория критичности, требования, принципы обеспечения качества.

## 1 Введение

Как показывает практика, высокий уровень зависимости выполнения космической системой основной целевой функции и безопасности от используемого в ней программного обеспечения порождает необходимость придания применяемым программным средствам заданных свойств качества и безопасности, способности противостоять разрушению, нарушениям функционирования системы, сбоям. Ситуация еще более усложняется, когда речь идет о критичном программном обеспечении, от правильного функционирования которого напрямую зависит успешность выполнения миссии и безопасность космической системы. К качеству и надежности подобного программного обеспечения (программных средств космического назначения - ПСКН), предъявляются особо высокие требования.

Особенности и характеристики качества ПСКН зависят от того, для какой цели, для какого потребителя и для каких условий эксплуатации они предназначены. Один и тот же программный продукт, произведенный для различных целей и при разных условиях применения, может иметь несколько различных представлений и оценок качества. В соответствии с принципиальными особенностями программных средств должны выбираться номенклатура и значения показателей качества, необходимых для его эффективного применения пользователями, а также требования к процессу верификации.

В реальных проектах часто отсутствуют или недостаточно четко формулируется понятие качества ПСКН, характеристики которыми оно описывается, как их следует измерять и сравнивать с требованиями технического задания или спецификации [1].

В связи с этим является актуальной задачей анализ особенностей ПСКН с целью установления требований к характеристикам их качества.

## 2 Анализ особенностей программных средств космического назначения

С целью установления особенностей и требований к характеристикам качества и безопасности представляется целесообразным провести классификацию ПСКН с учетом их назначения, условий эксплуатации, требований к надежности, безопасности и др.

С учетом существующих подходов к классификации программных средств целесообразно провести классификацию ПСКН по следующим признакам:

- принадлежность ПСКН к объектам космической техники;
- функциональное назначение;
- степень апробированности;
- влияние на безопасность.

По принадлежности к объектам космической техники можно выделить ПСКН:

- бортовых комплексов управления пилотируемыми и автоматическими космическими аппаратами;
- бортовых вычислительных комплексов ракет-носителей, разгонных блоков;
- технических и стартовых комплексов, наземных автоматизированных комплексов управления космическими аппаратами, наземного оборудования и сооружений;
- полезных нагрузок;
- экспериментов и моделирования.

Принадлежность ПСКН к объектам космической техники определяет специфические требования, например требования к безопасности ПСКН космических аппаратов, полезных грузов отличается от требований к безопасности ПСКН для экспериментов и моделирования.

К программному обеспечению (ПО), входящему в состав бортовых комплексов, традиционно предъявляются повышенные требования к надежности и безопасности. Помимо этого, к важным свойствам такого программного обеспечения относят высокое качество, поддающееся проверке, непротиворечивость, возможность повторного использования, быстрая интеграция с аппаратными средствами, возможность переноса на другие платформы [1].

По функциональному назначению ПСКН подразделяют на:

- общее (или системное);
- прикладное (или функциональное);
- технологическое (или инструментальное), которое используют при разработке, тестировании и верификации.

Признак "функциональное назначение" определяет специфические требования, которые предъявляют к инструментальным средствам, необходимым для реализации соответствующих функций, прикладного (функционального) ПО и общего (системного) ПО.

По степени апробированности различают следующие типы ПСКН:

- новое, разработанное впервые;
- существующее собственное (разработанное ранее) или существующее приобретенное;\*
- конфигурируемое из типовых модулей.

Признак "степень апробированности" определяет объем требований к разработке и верификации ПСКН в зависимости от принадлежности к объектам космической техники и назначения, категории безопасности.

Разработка нового ПО требует более высокой квалификации разработчиков, больших материальных затрат и усилий. Верификация должна проводиться после каждого этапа жизненного цикла ПО в полном объеме, с уровнем независимости, обусловленным категорией безопасности функций ПО.

Вариант разработки нового программного продукта имеет то преимущество, что управление процессом верификации можно осуществлять в начале.

Повторное использование существующего собственного ПО имеет то преимущество, что весь объем ПО является всегда доступным. Дополнительные затраты требуются только на создание необходимых изменений, приложений для подготовки данных, конфигурации и дополнительной верификации.

Для имеющегося приобретенного ПО исходный код и первичная документация, в большинстве случаев, не доступны для верификации. В этом случае необходимо проводить анализ опыта эксплуатации и функциональное тестирование.

Конфигурируемое программное обеспечение разрабатывают с использованием типовых простых и надежных модулей (базовых процессов, таких как ввод сигнала, проверка сигнала, операции инициализации, логического контроля, управления данными и др.). Интеграция модулей осуществляется с использованием стандартных моделей настройки объектов и процессов и может сопровождаться введением данных для определения и изменения характеристик для определенного объекта. Верификацию проводят в полном объеме и на всех диапазонах входной информации, констант управления и на диапазонах регулирования.

По влиянию на безопасность различают ПСКН, которое:

- влияет на безопасность (критическое ПО);
- не влияет на безопасность.

Признак "влияние на безопасность" определяет требования к ПСКН по реализации критических функций в зависимости от принятых категорий опасности. Например, в зависимости от категории опасности предъявляются разные требования к объему, полноте, документированию отчетности и независимости процесса верификации.

Под понятием "критическое ПО (Safety-Critical Software)" обычно понимают программное обеспечение, выполняющее критические функции, важные для безопасности, отказ в выполнении которых (потеря или деградация) могут привести к катастрофическим или критическим последствиям [7]. Иногда этим же термином называют программы, разработанные в соответствии со специальными стандартами, принятыми для критически важных областей.

### 3 Особенности ПСКН критического применения

Особенности и требования к ПСКН зависят от реализуемых критических функций, требований к безопасности системы (программно-технического комплекса - ПТК), в состав которой она входит.

В настоящее время существует классификация критического программного обеспечения по категориям безопасности принятые для атомных станций, авиационных и космических систем [2-6]. Эти классификации основаны на установлении категорий опасности отказных ситуаций систем или объекта управления, вызванные сбоем или отказом программного обеспечения. Уровень критичности или категория безопасности ПС определяется тяжестью последствий его аномального функционирования с учетом вероятности их наступления.

В европейском стандарте ECSS-Q-ST-80C [2] определены следующие категории критичности программного обеспечения космического назначения (таблица 1).

### 4 Общие требования к качеству ПСКН

Основой для формирования требований к ПСКН является анализ свойств, характеризующих качество его функционирования с учетом его назначения и условий эксплуатации. В

Таблица 1. Категории критичности ПСКН

Категория	Характеристика
А	Программное обеспечение, которое в случае неисполнения или неверного исполнения, или аномального поведения, может вызвать или способствовать отказу системы, приводящему к: катастрофическим последствиям (гибель людей, угроза их жизни, разрушение, потеря техники);
В	Программное обеспечение, которое в случае неисполнения или неверного исполнения, или аномального поведения, может вызвать или способствовать отказу системы, приводящему к: критическим последствиям (ущерб, не угрожающий жизни людей, значительное повреждение техники, вредное влияние на окружающую среду);
С	Программное обеспечение, которое в случае неисполнения или неверного исполнения, или аномального поведения, может вызвать или способствовать отказу системы, приводящему к: существенным последствиям (существенное снижение возможностей объекта управления или способности персонала справиться с неблагоприятными режимами);
Д	Программное обеспечение, которое в случае неисполнения или неверного исполнения, или аномального поведения, может вызвать или способствовать отказу системы, приводящему к: незначительным или ничтожным последствиям (незначительному уменьшению безопасности объекта управления и требует действий персонала, которые осуществимы в пределах их возможностей).

соответствии с принципиальными особенностями ПСКН при проектировании должны выбираться номенклатура и значения показателей качества, необходимых для его эффективного применения пользователями, которые впоследствии отражаются в технической документации и в спецификации требований на конечный продукт.

Номенклатура показателей качества ПСКН должна устанавливаться с учетом:

- назначения и условий эксплуатации;
- результатов анализа требований пользователя (заказчика), поставленных задач управления качеством;
- состава, структуры и специфики характеризующих свойств.

Для каждого вида (группы), а иногда и конкретного ПСКН необходимо устанавливать свою номенклатуру показателей качества, учитывающую специфику назначения и условий применения. Каждый показатель качества может использоваться, если определена его метрика и может быть указан способ ее оценивания и сопоставления с требуемым эталонным значением.

В общем случае требования к качеству ПСКН должны обязательно включать следующие показатели:

- функциональное соответствие;
- безотказность (reliability) [7,8];
- живучесть (survivability) [7,8];
- функциональная безопасность (functional safety) [9].

Другие характеристики качества должны быть заданы при формировании спецификации или технического задания на разработку ПСКН.

Разработчик программного обеспечения после проведения функционального анализа требований к проекту должен определить категории безопасности выполняемых ПСКН функций и установить категорию его критичности (категию безопасности). В соответствии с категорией критичности ПСКН должны устанавливаться требования к характеристикам качества и безопасности, а также требования, относящиеся к верификации, валидации и уровням доказательств.

Для ПСКН критического применения должны быть:

- проанализированы возможные источники отказов;
- определены последствия проявления дефектов ПО;
- предусмотрены необходимые программные средства для исключения отказов по общей причине или уменьшения их последствий до приемлемого уровня;
- проведен анализ влияния такого ПО на безопасность, задокументированы результаты анализа и приняты меры.

Требования к качеству ПСКН критического применения должно базироваться на реализации в полном или сокращенном объеме следующих функций:

а) прогнозирование возможности появления (проявления) дефекта и возникновения отказа вследствие этого дефекта (fault forecasting);

б) предупреждение появления (проявления) дефекта и возникновения отказа (fault prevention);

в) выявление появления (проявления) дефекта, ошибки вычислений, отказа (fault detection);

г) идентификации причины, вида и места дефекта (отказа) (fault diagnosis);

д) парирования последствий дефекта и возникновения отказа (fault tolerance). Эта функция включает:

1) отключение элементов (компонентов, модулей архитектуры), которые отказали, и / или изолирование искаженной информации (fault isolation);

2) реконфигурацию структуры (архитектуры) путем удаления компонента, который отказал из конфигурации и замены его работоспособным (fault removal);

е) восстановление вычислительного процесса путем формирования правильной информации или возврата к предыдущей точке и продолжения функционирования (fault recovery).

С учетом особенностей ПСКН как программного обеспечения критического применения, для обеспечения их надежности и безопасности необходимо придерживаться следующих принципов: единичного отказа, резервирования, независимости, многообразия, защиты от отказов по общей причине.

Приоритетным требованием качества для ПСКН критического применения является гарантия качества или гарантоспособность (dependability), под которой понимается - доказанная уверенность способности ПС надежного и безопасного выполнения необходимых функций в соответствии с назначением [7,8].

Требования к характеристикам качества ПСКН должны устанавливаться с учетом совокупности различных факторов.

Это технические факторы:

- новизна разработки;
- сложность и объем;
- уровень критичности;
- наличие требований к повторному использованию;
- уровень использования готовых коммерческих компонентов или существующего ПО;

- уровень стабильности требований пользователя.

Эксплуатационные факторы, которые необходимо учитывать при установлении к характеристикам качества ПСКН:

- назначение ПСКН в соответствии с типом космических систем или их частей (например, беспилотные или пилотируемые аппараты, пусковые установки, полезные грузы, эксперименты);
- количество потенциальных пользователей;
- предполагаемое время использования;
- количество систем, в которых ПСКН будет использоваться;
- ограничение режимов работы, технической поддержки, использования в других системах и изъятие из использования.

Также необходимо учитывать организационные факторы:

- необходимые для разработки ПСКН объемы работ и времени;
- необходимые для разработки и эксплуатацию ПСКН финансовые, человеческие ресурсы;
- приемлемый для проекта уровень риска;
- тип модели жизненного цикла;
- требования графика разработки ПСКН.

Обеспечение и подтверждение качества ПСКН, как сложных ПС с высокими требованиями к качеству, должно базироваться на проверках и испытаниях:

- качества требований к ПСКН;
- качества выполнения процессов жизненного цикла ПСКН;
- качества готового программного продукта и документации;
- качества проверки выполнения требований.

Главным и достаточным условием обеспечения качества ПС является гарантия (доказанная уверенность) правильного, надежного, достоверного и устойчивого выполнения необходимых функций в течение заданного времени, невзирая на возникшие внутренние и внешние возмущения.

## 5 Заключение

Сложность процесса разработки и сопровождения ПСКН во многом обуславливается особыми требованиями, предъявляемыми к их качеству. Неполнота, неопределенности и разная трактовка в определении и формализации характеристик качества ПСКН и требуемых их значений оставляют широкое поле для произвола при описании и оценивании их качества. Эти факторы обосновывают необходимость разработки и применения для каждого проекта ПСКН специальных планов и программ, методологии и инструментальных средств, формализованных методов описания и оценки качества, обеспечивающих требуемое качество, надежность и безопасность функционирования. Методы оценки качества ПСКН должны базироваться на следующих основных компонентах:

- модели качества ПСКН, содержащей механизмы для формального определения характеристик качества и их отношений;
- модели метрик ПСКН, формирующей механизмы для измерения показателей качества;
- методики оценки качества, определяющей процессы оценки качества ПСКН.



В качестве исходной информации для оценки соответствия ПСКН требованиям качества должны использоваться требования к системе, требования к ПСКН, описание его архитектуры, данных, а также программная документация.

Широкий спектр требований к качеству ПСКН, в зависимости от их назначения, принципиальных особенностей и условий эксплуатации, приводит к необходимости адаптации и детализации рекомендаций существующих базовых стандартов, регламентирующих качество программного обеспечения. Прежде всего, это относится к ПСКН критического применения.

## Список литературы

1. Тюгашев А. А., Ильин И. А., Ермаков И. Е. Пути повышения надежности и качества программного обеспечения в космической отрасли // Управление большими системами. Выпуск 39. – М.: ИПУ РАН, 2012. – С. 288-299.
2. ECSS-Q-ST-80C-2009 Space product assurance: Software product assurance.
3. ГОСТ Р МЭК 61226-2011 Атомные станции. Системы контроля и управления, важные для безопасности. Классификация функций контроля и управления.
4. RTCA DO-178B Software Considerations in Airborne Systems and Equipment Certification.
5. ГОСТ Р 51904-2002 Программное обеспечение встроенных систем. Общие требования к разработке и документированию.
6. СОУ-Н ДКАУ 078-2014 Верифікація програмного забезпечення програмно-технічних комплексів критичного призначення
7. ECSS-Q-80-03-2006 Space Product Assurance - Methods and Techniques to Support the Assessment of Software Dependability and Safety.
8. ECSS-Q-80B-2003 Space Product Assurance: Software Product Assurance
9. ГОСТ Р МЭК 61508-1-2007 Функциональная безопасность систем электрических, электронных, программируемых электронных, связанных с безопасностью. Часть 1. Общие требования.
10. Basic Concepts and Taxonomy of Dependable and Secure Computing / A. Avizienis, J.C. Laprie, B. Randell, C. Landwehr // IEEE Trans. on Dependable and Secure Computing. - 2004. - Vol. 1. - P. 11 - 33

# Особенности Разработки Программно-Технологического Обеспечения для Региональных Геоинформационных Веб-Систем

А.А. Кадочников

Институт вычислительного моделирования СО РАН, Красноярск, Россия  
scorant@icm.krasn.ru

**Аннотация.** На примере системы «Банк пространственных данных Красноярского края» рассматриваются особенности разработки региональных геоинформационных систем и сервисов в Интернет. Основное назначение таких систем – мониторинг и публикация данных состояния окружающей природной среды, мониторинг различных экономических или социальных процессов для систем поддержки принятия решений на уровне Красноярского края. В работе востребованы методики и программные средства, которые позволят формировать оценки состояния территорий на базе основных показателей в наглядном виде. Важную роль играет использование современных средств визуализации данных с использованием ГИС-технологий. Значительное внимание уделяется веб-сервисам и программным интерфейсам. Решается ряд задач, связанных с обменом данными и метаданными о пространственной информации, возникающих при разработке совместных проектов различных научных институтов, университетов и подразделений органов власти.

**Ключевые слова:** ГИС, каталог ресурсов, веб-сервисы, геоинформационный Интернет-сервер, веб-картография, геопространственные данные.

## 1 Введение

Банк пространственных данных – государственная информационная система, предназначенная для межведомственного взаимодействия и интеграционных проектов Красноярского края по линии каталогизации, хранения, аналитической обработки и публикации геопространственных данных (<http://24bpd.ru/>). Разрабатывалась по заказу Министерства информатизации и связи Красноярского края.

Функционально назначение банка пространственных данных – создание распределенной системы идентификации, адресации и позиционирования объектов управления на территории края с использованием средств цифровой картографии и геоинформатики в виде банка пространственных данных, состоящего из тематических электронных карт и космических снимков высокого разрешения. Он призван обеспечить оперативное решение следующих задач:

- навигация по информационным картографическим ресурсам, визуализация и анализ пространственно-ориентированных данных на унифицированных цифровых картах;
- ведение, хранение цифровых картографических материалов, растровых снимков территории;
- предоставление картографических веб-сервисов и ресурсов для сторонних прикладных информационных систем.

На примере системы «Банк пространственных данных Красноярского края» рассматривается задача, связанная с формированием картографических программных интерфейсов. Основное назначение программных интерфейсов к каталогу пространственных данных –

обеспечение доступа к этим данным, различным службам и веб-сервисам. Цель его создания – информационное обеспечения задач мониторинга состояния социальной и природной среды и ресурсов в региональной ГИС.

В качестве основы использовались разработанные программные средства для анализа пространственных данных в среде геопортала Института вычислительного моделирования СО РАН [1] с использованием технологий, предлагаемых международным консорциумом OGC (Open Geospatial Consortium) и программного обеспечения MapServer и GeoWebCache. Программные инструменты содержат средства для хранения цифровых картографических материалов, растровых снимков территории, сервисы для навигации по распределенному каталогу пространственных данных, сервисы для пространственного анализа и математического моделирования на унифицированных цифровых картах. Основным элементом геопортала является каталог метаданных о пространственных данных.

Каталог метаданных содержит информацию по доступным слоям и картам. Основной особенностью каталога пространственных данных является возможность использования различных форматов пространственных данных и организация доступа для пользователя к этим данным с помощью современных стандартов и технологий. Для оформления карт и картографических слоев применяется SLD (Styled Layer Descriptor) – язык описания стилей, используемый для отображения объектов на карте в WMS (Web Map Service) и WFS (Web Feature Service) и WCS (Web Coverage Service) серверах, а также собственный формат описания стилей, разработанный для геопортала [2]. Пользовательский интерфейс для каталога метаданных, для систем мониторинга тематических показателей и для информационно-аналитических систем в региональном управлении выполнен в виде геоинформационного веб-приложения.

## 2 Архитектура системы

Программные модули каталога пространственных данных в созданных региональных геоинформационных веб-системах состоят из следующих основных элементов:

1. Хранилище пространственных данных. Хранилище состоит из файл-сервера с геоданными в популярных форматах ГИС, а так же сервера PostgreSQL/PostGIS с набором баз геопро пространственных данных. Поддерживаются и сторонние источники данных, размещенных отдельно от каталога на внешних серверах.
2. Каталог ресурсов. База данных метаописаний всех информационных ресурсов портала, а также набор программных библиотек (API) для различных операций по их обработке [3]. Содержание каталога информационных ресурсов составляют объекты различных типов: информационные ресурсы (картографический слой, карта, атрибутивные данные, аналитический сервис с веб-доступом, публикация, и др.); элементы множественной классификации информационных ресурсов; информационно-навигационные элементы (HTML документы) и др.
3. Веб-клиент для доступа к ресурсам каталога:
  - пользовательский веб-интерфейс каталога пространственных метаданных – веб-приложение, предназначенное для навигации по зарегистрированным в системе ресурсам и поиску среди них;
  - подсистема картографической веб-визуализации – отображение карт и отдельных слоев геоданных портала через веб-интерфейс с развитыми интерактивными возможностями.
4. Пользовательские интерфейсы управления:

- административный веб-интерфейс – подсистема управления каталогом ресурсов;
  - редактор стилового оформления – специализированная программа-редактор тематической раскраски слоев и карт.
5. Картографические и служебные веб-сервисы – программные и пользовательские интерфейсы для получения и предоставления геоданных на основе стандартных протоколов OGC (WMS, WMTS, WFS и т.д.); библиотеки функций и программных интерфейсов для интеграции разных элементов разработки в единое целое.
  6. Прикладные веб-сервисы – ресурсоемкие вычислительные задачи, выполняемые на стороне сервера: адресный поиск, геокодирование, прокладка маршрутов транспорта по графу дорожной сети, построение водотоков по графу речной сети и т.д.

### 3 Технологии и программное обеспечение

Разработка программных средств выполнялась на языке сценариев PHP. Для хранения данных использована СУБД PostgreSQL – свободно распространяемая объектно-реляционная система управления базами данных, наиболее развитая из открытых СУБД в мире и являющаяся реальной альтернативой коммерческим базам данных. Немаловажным преимуществом является наличие дополнительного модуля, который облегчают работу с пространственными данными PostGIS (свободная ГИС библиотека, которая позволяет работать с географическими объектами и функциями).

Для построения клиентской части веб-приложения, использующего карту региона, подходят несколько технологий – DHTML, Adobe Flash, SVG (Scalable Vector Graphics – масштабируемая векторная графика), WebGL (Web-based Graphics Library). Их возможностей достаточно для реализации клиентской логики картографического веб-интерфейса. Одним из интересных решений и популярных на сегодняшний день является применение технологии динамического HTML с методами асинхронного обмена данными без перезагрузки страницы (Remote Scripting, AJAX) [4]. Практически все современные веб-браузеры поддерживают эти технологии без использования дополнительных модулей. Эта технология позволяет уменьшить объем передаваемой информации по сети и улучшить «качество» пользовательского интерфейса. В результате, можно говорить, что пользовательская часть системы является клиентским приложением, а не набором динамических страниц, генерируемых сервером. Использование такого подхода дает возможность частично разделить логику клиентской и серверной частей, что приводит к более высокой гибкости всей системы.

При разработке картографического компонента веб-интерфейса были проанализированы два способа представления картографической информации для пользователя. Первый способ – карта отображается с использованием растровых фрагментов (тайлов). Основным преимуществом такого способа является скорость получения визуальной информации пользователем и малая нагрузка на сервер при отображении статической информации. Процесс формирования карты на клиентском компьютере состоит из нескольких этапов, с использованием дополнительных программных потоков, механизма кэширования, очереди загрузки фрагментов и др. При таком способе отображения карты пользователю процесс построения композиции карты позволяет оптимизировать процесс загрузки, снизить нагрузку на веб-браузер и более равномерно ее распределить по времени. Однако при отображении меняющихся тематических данных, необходимых для информационно-аналитических систем, такой способ снижает скорость доступа пользователя к пространственным данным и увеличивает нагрузку на сервер. Для решения этой проблемы используется второй способ отображения информации – по запросу пользователя генерируется одно растровое изображение,

либо формируется слой с векторными объектами. В зависимости от типа представляемой информации пользователю в программном интерфейсе системы используется комбинация этих двух способов.

Сегодня существует большое число библиотек с открытым исходным кодом для создания готового пользовательского интерфейса с картографическим интерфейсом. Однако функционала существующих библиотек было недостаточно для решения поставленной задачи и было разработано веб-приложение с использованием библиотеки OpenLayers. OpenLayers – это JavaScript библиотека с открытым исходным кодом, предназначенная для создания карт на основе программного интерфейса, поддерживает технологию AJAX и анимацию. При разработке серверной части веб-приложения для работы с картой Красноярского края используется программное обеспечение MapServer, предназначенное для обеспечения доступа через Интернет к интерактивным картам. MapServer представляет собой открытую и свободно распространяемую среду разработки Интернет-приложений для работы с электронными картами широко распространенных среди множества геоинформационных систем векторных и растровых форматов, обладающую большим числом функциональных возможностей.

Для создания карты из фрагментов использовалось программное обеспечение GeoWebCache. GeoWebCache использует спецификацию WMS Tile Caching (WMS-C), которая явилась результатом конференции FOSS4G в 2006 г [5]. Сервисы WMS разрабатывались с учетом большой гибкости и богатого функционала, но это оборачивается высокими требованиями к вычислительной мощности сервера [6]. Серверы WMS-C по протоколам совместимы с WMS, поэтому их можно встроить между клиентом и сервером WMS, что позволяет существенно увеличить скорость реакции и разгрузить сервер. Рассмотрены альтернативные решения для создания каталога фрагментов, такие как ka-map Cache (<http://ka-map.ominiverdi.org>), TileCache (<http://tilecache.org>), MapCache (<http://mapserver.org>) и др. Источником пространственных данных для сервера с программным обеспечением GeoWebCache послужил WMS сервер с картой Красноярского края на основе программного обеспечения MapServer. Реализована система сервисов, которые поддерживают кэш растровых изображений на сервере с GeoWebCache в актуальном состоянии при обновлении исходных данных на WMS сервере.

В результате объединения различных технологий представления карты пользователю на стороне клиента реализован вариант, в котором карта состоит из двух слоев: подложка и тематический слой (рис. 1).

#### 4 Сервисы и интерфейсы

Разработанный банк пространственных данных включает в себя программные средства, предоставляющие пользователям различный набор инструментов для работы с системой посредством различных клиентов (веб-клиент на основе веб-браузера, windows-клиенты), а также ряд средств администрирования самой системы и базы данных. Хотя в общем случае в Интранет-системе могут использоваться все возможные службы Интернет. При этом доступ к различным компонентам системы можно осуществлять дополнительно через прямое подключение к базе данных, что используется в основном для ввода и редактирования данных.

В рассматриваемой системе реализован ряд вспомогательных сервисов, которые могут использоваться в связанных информационных системах:

1. Сервисы OGC. Доступ к ресурсам каталога для удобства пользователей возможен через сервисы WMS и WFS. Эти сервисы позволяют работать с каталогом из различного про-

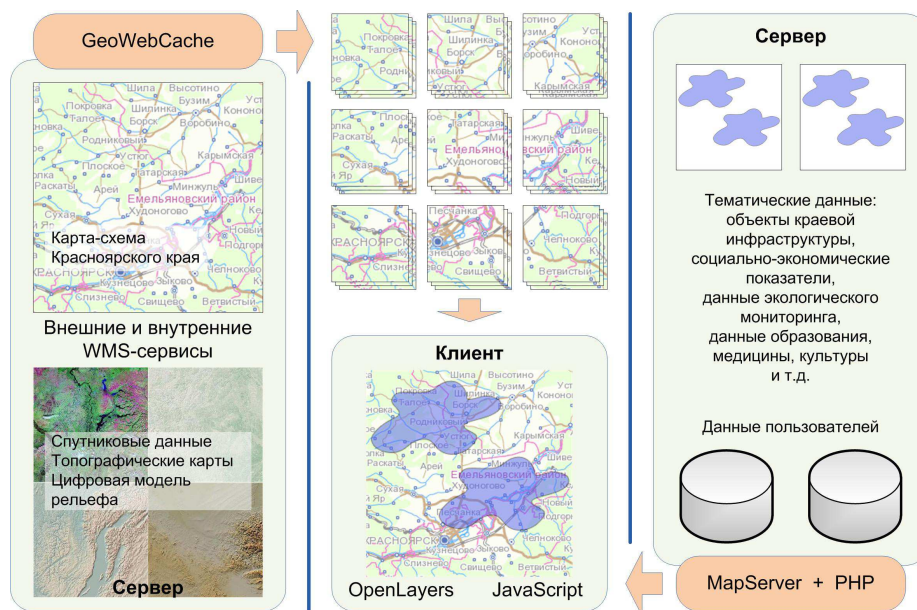


Рис. 1. Формирование картографического веб-приложения.

граммного обеспечения, включая такие известные пакеты, как MapInfo, ESRI ArcGIS, Quantum GIS и другие.

2. Сервис геокодирования. Функции геокодирования позволяют «привязывать» базы данных, которые ведет большинство ведомств, обслуживающих урбанизированные территории и население, к картам территорий. Процесс преобразования адресов пространственных объектов в их географические координаты называется геокодированием. Обратный процесс, преобразование точки на карте в читаемый для пользователя адрес, известен как обратное геокодирование [7].
3. Сервис поиска маршрутов. Поиск маршрута на дорожной сети выполняется по двум указанным точкам – начальной и конечной. Предполагается, что они могут быть заданы как географическими координатами, так и адресом в текстовом виде. В последнем случае используется подсистема геокодирования для определения их координат. Поиск выполняется по дорожному графу, который предварительно подготовлен по базовому слою дорожной сети [7].
4. Сервис построения водотоков вычисляет водоток по графу речной сети из заданной точки. Результат возвращается в виде набора сегментов речной сети, включая пространственные данные. Сервис играет важную роль в задачах по обеспечению безопасности гидротехнических сооружений.
5. Сервис построения тематических карт по территории края позволяет строить различные тематические карты (раскраски, диаграммы) на основе данных из внешних систем. Такие тематические карты могут быть размещены как на сервере геопортала, так и на сайтах органов власти и управления Красноярского края [8].
6. Сервис размещения картографических ресурсов на сайтах других организаций. Картографический сервис работающий в режиме веб-интерфейса с картой Красноярского края предоставляет разработчикам веб-сайтов возможность размещать окно с картой на своих страницах и осуществлять его настройку с помощью разработанного инструментария. Авторы и клиенты могут локально дополнять карту своими данными. Несложные команды позволяют добавить на карту собственные контуры, маркеры, интерактивные под-

сказки, всплывающие окна и тематические раскраски. Данные для отображения можно размещать непосредственно в коде веб-страниц. Предусмотрена возможность создания тематических слоев в режиме просмотра веб-страницы.

- Сервис доступа к каталогу. Базовый доступ к каталогу ресурсов организован через веб-сервис по протоколу SOAP. Программный интерфейс (API) содержит функций для управления объектами каталога.

В региональной информационной системе предусмотрено разделение прав доступа пользователей системы (администратор, оператор, обычный пользователь и т.п.), позволяя одновременно работать нескольким пользователям с различных мест. В качестве основы для такой системы может быть сеть нескольких серверов, функции которых могут быть похожими (для снижения нагрузки на всю систему), либо могут различаться. Один сервер организует доступ к базе данных, другой обеспечивает доступ к хранилищу картографической информации, при этом в качестве хранилища такой информации могут выступать множество серверов распределенных по всему миру. Взаимодействие между ними можно организовывать с использованием встроенных возможностей программного обеспечения MapServer, а также с помощью WMS и WFS технологий. Доступ ко всем сервисам системы осуществляется различными клиентами, каждый из которых выполняет свои определенные функции.



Рис. 2. Форматы данных.

При разработке информационной системы использовались современные программные решения и технологии, что обеспечило доступ к различным форматам пространственных данных (рис. 2). Банк пространственных данных предусматривает возможность хранения следующих типов пространственной информации:

- электронные карты и слои. С помощью библиотек GDAL и OGR появляется возможность для загрузки огромного числа различных растровых и векторных форматов (OGR – PostGIS, ESRI ArcSDE, Oracle Spatial, MySQL, MapInfo и др., GDAL – TIFF/GeoTIFF, EPPL7, MrSID и др.);
- сервисы OGC (WMS и WFS) позволяют подключаться к внешним каталогам пространственным данным;
- предусмотрена возможность загрузки файловых архивов, содержащих картографический материал. Такая возможность позволяет хранить рабочие наборы и картографические слои подготовленных в сторонних программных пакетах (например, ГИС MapInfo или ГИС ESRI ArcInfo);
- набор различных ссылок на внешние электронные ресурсы (например, атласы в Интернет, различные Интернет-ГИС ресурсы и т.п.);
- набор различных ссылок на архивы картографической информации (бумажные карты, компакт диски и т.п.).

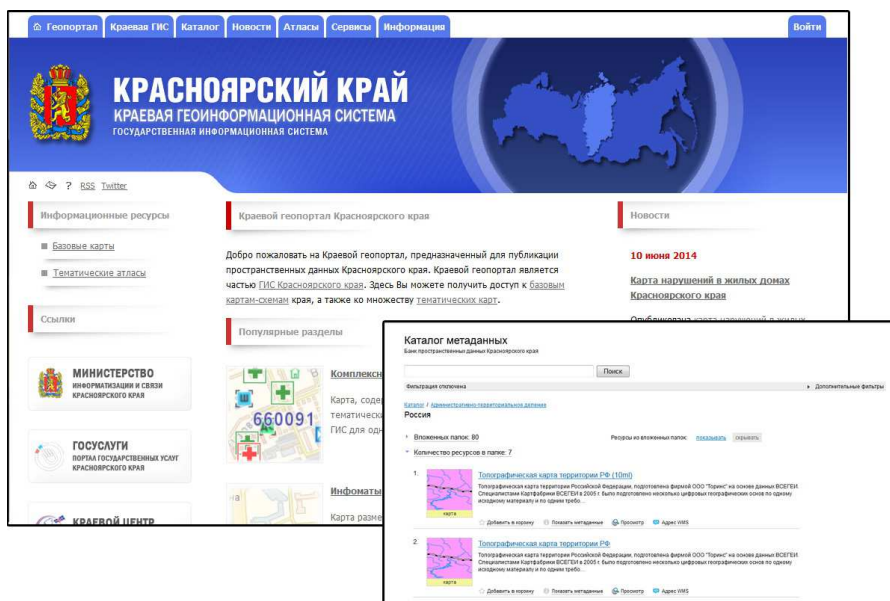


Рис. 3. Интерфейс пользователя.

Для единого описания различной пространственной информации используется общий шаблон метаданных, подготовленный на основе проекта ГОСТа «Географическая информация. Метаданные» [9]. Каждый элемент хранения содержит такую информацию, как название, описание, имя файла, источник, авторские права, система координат, способ приобретения, масштаб первоисточника, топология, информация о пространственно-временных характеристиках, информация об ограничениях.

На рисунке 3 представлен пример с экранными копиями интерфейсов разработанной системы.

## 5 Заключение

Сегодня в результате развития технологий и программного обеспечения получены новые результаты, которые позволили значительно усовершенствовать существующую программно-технологическую платформу для разработки региональных геоинформационных веб-систем. Модульная архитектура позволит усовершенствовать существующую систему в дальнейшем, а также позволит усовершенствовать процессы разработки подобных систем мониторинга для региона, что повысит качество предоставляемых услуг для населения края и качество принимаемых управленческих решений. Рассмотренное решение может быть использовано не только для территории Красноярского края, а ресурсы и инструменты разработанной программно-технологической платформы могут быть использованы при разработке других систем.

Разработанное программное обеспечение и сервисы строились на основе свободно расширяемых технологий и программного обеспечения:

- платформа для публикации картографических данных – MapServer 6.x (<http://www.mapserver.org>);
- система кэширования картографических данных – GeoWebCache (<http://www.geowebcache.org>);
- основной язык разработки – PHP 5.3+ (<http://www.php.net>);



- СУБД – PostgreSQL 9.3+ (<http://www.postgresql.org>) и PostGIS 2.0+ (<http://www.postgis.org>).

## Список литературы

1. Кадочников А.А., Матвеев А.Г., Пятаев А.С., Токарев А.В., Якубайлик О.Э. Программный комплекс «Геопортал ИВМ СО РАН» // Свидетельство о государственной регистрации программы для ЭВМ № 2014612492 от 26 февраля 2014 г.
2. Кадочников А.А. Веб-сервисы и приложения для геоинформационного Интернет-портала ИВМ СО РАН. Материалы Международной конференции «ИнтерКарто-ИнтерГИС-17». Устойчивое развитие территорий: теория ГИС и практический опыт». Белокуриха, Денпасар, 14-19 декабря 2011. – С. 93-97.
3. Якубайлик О.Э., Токарев А.В. Каталог ресурсов для ГИС мониторинга состояния окружающей природной среды в зоне действия предприятий нефтегазовой отрасли // Кузбасс-3: Сборник статей. Отдельный выпуск Горного информационно-аналитического бюллетеня. – М.: издательство «Горная книга», 2009. – С. 215-219.
4. Кадочников А.А. Организация доступа к электронной карте Красноярского края для информационно-аналитических систем с помощью веб-сервисов. // Материалы Международной конференции «ИнтерКарто-ИнтерГИС-18». Устойчивое развитие территорий: теория ГИС и практический опыт» / Редкол.: С.П. Евдокимов (отв. ред.) [и др]. Смоленск, 26-28 июня, 2012 г. Смоленск, 2012. 532 с. – С. 136-140.
5. Tile Map Service Specification. / The Open Source Geospatial Foundation [Электронный ресурс] – URL: [http://wiki.osgeo.org/wiki/Tile\\_Map\\_Service\\_Specification](http://wiki.osgeo.org/wiki/Tile_Map_Service_Specification) (дата обращения: 17.04.2015).
6. OpenGIS Web Map Service (WMS) Implementation Specification. / Open GIS consortium. [Электронный ресурс] – URL: <http://www.opengeospatial.org/standards/wms> (дата обращения: 17.04.2015).
7. Токарев А.В. Построение геопространственных веб-сервисов для задач оперативного мониторинга транспортных средств // Материалы международной конференции ИнтерКарто-ИнтерГИС-18: Устойчивое развитие территорий: теория ГИС и практический опыт / Редкол.: С.П. Евдокимов [и др]. – Смоленск, 2012. – С. 443-448.
8. Кадочников А.А. Модуль визуализации тематических карт на основе аналитических данных регионального уровня для информационно-графических систем. // Свидетельство о государственной регистрации программы для ЭВМ № 2015613033 от 12 января 2015 г.
9. ГОСТ Р 52573-2006. Географическая информация. Метаданные.

# Вычислительная Технология Обработки Данных Комплексного Мониторинга Природных Геообъектов

Михаил Курако<sup>1</sup> and Константин Симонов<sup>2</sup>

<sup>1</sup> Сибирский федеральный университет,

<sup>2</sup> Институт вычислительного моделирования СО РАН [mkurako@gmail.com](mailto:mkurako@gmail.com), [simonovkv@icm.krasn.ru](mailto:simonovkv@icm.krasn.ru)

**Аннотация.** Работа посвящена новому направлению в обработке данных геомониторинга, которое может быть использовано в диагностике сложных природных геообъектов и систем – геометрический анализ визуальных данных, где совместно выполняется вейвлет-преобразование данных для криволинейных объектов и шиарлет-преобразование для линейных объектов. Задаче разделения изображения на морфологически разные составляющие в последнее время уделяют много внимания в связи с её значимостью при решении задач распознавания образов для различных актуальных приложений в науках о Земле. Разрабатываемая вычислительная технология для эффективного решения этой задачи может быть применена к широкому кругу геообъектов, включая исследования, связанные с изысканиями на нефть и газ в сложных геосредах. В результате проведенных исследований разработана вычислительная технология, позволяющая решать задачи обработки данных геомониторинга сложных геообъектов на основе совместного применения вейвлет- и шиарлет-преобразований.

**Ключевые слова:** вейвлет-преобразование, шиарлет-преобразование, морфологический анализ, распознавание образов.

## Введение

В настоящее время современные вычислительные технологии позволяют обеспечивать производство, передачу и хранение больших объемов данных: различных данных и информации в самых разнообразных сферах деятельности (медицина, астрономия, сейсмология, метеорология, управления воздушным движением, интернет-трафик, аудио и видео приложения, цифровые коммуникации и т.д.).

Эти данные требуют эффективного анализа и обработки с целью получения новой информации и знаний. Кроме того, важно не только обеспечить адекватность методики для обработки различных типов данных, но и возможность анализа точности методов для более глубокого понимания основных структур в данных.

Таким образом, задача сводится к разбиению исходных данных (сигналов, изображений) на блоки приемлемых размеров, обработке каждого блока в отдельности некоторым методом и анализу результатов обработки. Для этого «разламывания», чтобы разобраться в объекте исследования, предлагается гармонический анализ.

Для класса данных  $l \subset L^2(\mathbf{R}^d)$ ,  $d \geq 1$  подбирается такое множество анализирующих функций  $(\varphi_i)_{i \in I}$ , где  $I$  – счетное множество индексов, что для всех  $f \in l$  выполняется:

$$f = \sum_{i \in I} c_i(f) \varphi_i.$$

Счётное множество коэффициентов  $c_i(f)$ ,  $i \in I$  является разложением исходного сигнала по базису анализирующих функций и может быть определённым образом проинтерпретировано с целью анализа входных данных. С другой стороны, указанная формула описывает процесс восстановления сигнала по его коэффициентам разложения.

Объектом исследования являются двумерные изображения. Отдельным вопросом анализа изображений является определение фрагментов изображения, имеющих анизотропные характеристики или разрывы (такие как края изображённых объектов или кривые линии на изображении), поскольку традиционные методы обработки изображений нечувствительны к подобного рода характеристикам.

В течение последних двадцати лет были предложены различные методы обработки анизотропных объектов на изображении, такие как направленные вейвлеты, комплексные вейвлеты, контурлеты, кёрвлеты и т. п. В 2006 году предложен несколько иной подход к анализу анизотропных составляющих: шиарлеты. В отличие от вейвлетов или кёрвлетов, шиарлеты строятся в классе аффинных систем, а также обладают возможностью определения направленности благодаря дополнительно введённому параметру сдвига [1,2,3,4,5,6,7].

В свою очередь, шиарлеты обладают набором характеристик, выгодно выделяющих их на фоне остальных методов обработки изображений: конечное число генерирующих функций; оптимальное представление анизотропных характеристик анализируемых данных; быстрая алгоритмическая реализация; единый подход к разложению непрерывных и дискретных данных.

Основными из приложений дискретного шиарлет-преобразования являются алгоритмы решения задач шумоподавления, выделения краёв на изображениях, разделения изображений на объекты различной природы (морфологический анализ) и улучшения качества изображений [8,9,10]. Имеющиеся подходы к анализу изображений могут быть перенесены в пространство больших размерностей (видеоизображения), а также оказаться полезными для решения задач в областях медицины и обработки данных геомониторинга [14].

## 1 Вычислительная методика обработки пространственных данных

Исходя из описанных выше теоретических и методических представлений рассмотрим модификацию метода геометрического анализа визуальных данных, позволяющую решать широкий класс задач обработки сложных изображений экологического мониторинга на основе шиарлет-преобразования.

При этом решаются следующие задачи экологического мониторинга в рамках специализированной информационной системы: разделение точек и кривых на изображениях; выделение контура на изображениях; визуализация данных на основе четырех алгоритмов шиарлет-преобразования.

Предлагается вычислительная методика решения указанных задач, которая состоит из следующих этапов [14]:

1. *подготовительный этап*, когда исходное изображение форматируется под расчетный шаблон и намечается последовательность расчетных процедур для оптимального решения поставленной задачи;
2. *запуск и настройка* алгоритмического обеспечения шиарлет-преобразования, выбор конкретного алгоритма в зависимости от поставленной задачи и от условий яркости и контрастности изображений;
3. *загрузка и обработка исходных изображений* для различных расчетных условий в зависимости от поставленной задачи;
4. *анализ получаемых расчетных изображений* в результате шиарлет-преобразования, контрастирование изображения на основе применения алгоритмов **A**, **B**, **C** и **D**, которые в результате определяются следующим образом: **A** – алгоритм FFST [12,13], **B** – алгоритм Shearlet Toolbox [8], **C** – алгоритм ShearLab [9,10], **D** – алгоритм TGVSHCS [10,11], является аналогом алгоритма **A**.

В результате проведенного исследования в качестве количественного параметра оценки эффективности алгоритмов выбрано среднее время работы алгоритмов. Для сравнения расчеты проводились на изображениях разных размеров (табл. 1).

Из таблицы 1 видно, что с помощью алгоритма **С** расчеты выполняются быстрее, чем на основе алгоритма **А** на изображениях больших размерностей, в то время как алгоритм **А** имеет незначительное преимущество по времени выполнения на изображениях небольшого размера. Изображения размерами более  $512 \times 512$  пикселей анализируются по частям, алгоритм **Д** является самым медленным.

**Таблица 1.** Среднее время работы алгоритмов (в секундах)

Размер изображения в пикселах	Алгоритм <b>А</b>	Алгоритм <b>В</b>	Алгоритм <b>С</b>	Алгоритм <b>Д</b>
$64 \times 64$	0.078	–	0.297	14.218
$128 \times 128$	0.297	–	0.391	58.484
$256 \times 256$	1.297	2.125	1.187	195.937
$512 \times 512$	6.828	10.016	3.578	1312.400

С применением указанных алгоритмов анализировались изображения для ряда смежных областей (снимки распространения пожара, медицинская томография, геоэкология и геодинамика). Проводились исследования указанных типов снимков для различных условий яркости и контрастности. При исследовании возможностей шумоподавления выполнялись оценки для шума Гаусса.

## 2 Задачи обработки данных геомониторинга

*Геометрическое разделение визуальных данных.* В соответствии с проведенным исследованием указанных алгоритмов шпирлет-преобразования предлагаем для решения первой задачи геометрического разделения визуальных данных экологического мониторинга применять алгоритм **С**. Повышение точности разделения оценивается в 5–12% по сравнению с применением кёрвлетов.

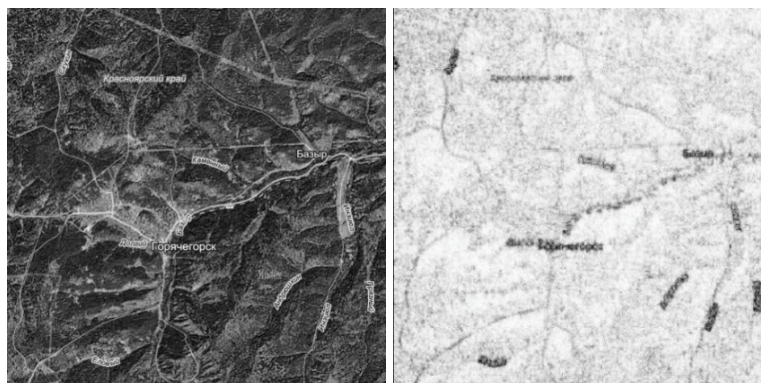
*Выделение контуров объектов на изображениях.* Рассмотрим вторую задачу – выделение контура объектов на изображении. Исследование алгоритма **А** выявило, что контуры объектов можно получить как сумму коэффициентов шпирлет-преобразования фиксированного значения параметра для последнего масштаба и всевозможных значений параметра сдвига. В связи с этим предлагается использовать эту особенность при решении нашей задачи:

$$f_{cont} = \sum_{k=0}^{k_1} \sum_{m=0}^{m_1} sh_{\psi}(f(j^*, k, m)),$$

где  $sh_{\psi}$  ставит в соответствие исследуемой функции  $f$  коэффициенты  $sh_{\psi}(f(j^*, k, m))$ , полученные для последнего масштаба  $j^*$ , всех направлений  $k$  и смещений  $m$ ,  $k_1$  – количество направлений,  $m_1$  – количество смещений.

Результаты решения этой задачи с помощью модифицированного алгоритма FFST показаны на различных данных (рис. 1). Модифицированный алгоритм предлагается применять для выделения контуров.

В таблице 2 приведены результаты соответствующих расчетов для некоторых изображений и сравнение с фильтрами Собела (Sobel) и Превитта (Prewitt). Модифицированный алгоритм сравним по точности с классическими алгоритмами Собела и Превитта.



**Рис. 1.** Выделение контуров объектов по данным геоэкологического мониторинга. Исходное изображение (слева) и результат выделения контуров (справа)

**Таблица 2.** Значения метрики PSNR (в дБ) при решении задачи выделения контуров

Изображение	Модифицированный алгоритм FFST	Алгоритм Собела	Алгоритм Превитта
PA	27.059	27.004	27.004
satg	24.245	24.194	24.194
scene	24.099	24.099	24.099
fire	24.102	24.101	24.101
fire1	24.102	24.102	24.102
fire2	24.102	24.102	24.102

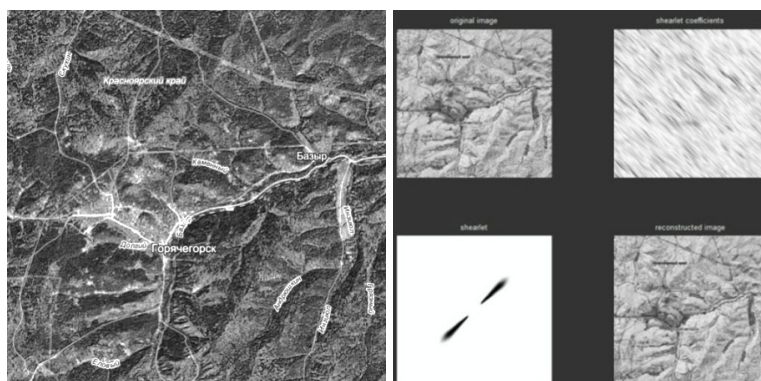
### 3 Сравнительный анализ алгоритмов для решения задачи шумоподавления

В рамках разработанной вычислительной методики выполнен сравнительный анализ алгоритмов дискретного шифротрансформирования для решения базовых прикладных задач специализированной информационной системы: фильтрации визуальных данных и шумоподавления на изображениях. Выполнено исследование алгоритма **A** и на рисунке 2 приведены результаты решения задачи выделения линейных особенностей для визуальных данных экологического мониторинга.

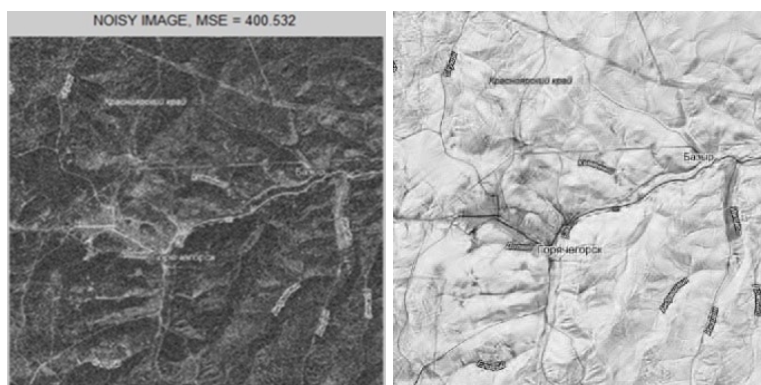
Для решения задачи шумоподавления выполнено исследование алгоритмов **B**, **C** и **D** для снимков из различных предметных областей (распространение пожара, медицинская томография, геоэкология и геодинамика).

Исследовались особенности работы алгоритмов **B**, **C** и **D** для различных условий яркости и контрастности изображений, проводились оценки для шума Гаусса. На рисунках 3-4 приведены решения задачи шумоподавления на основе алгоритмов **B**, **C** и **D**.

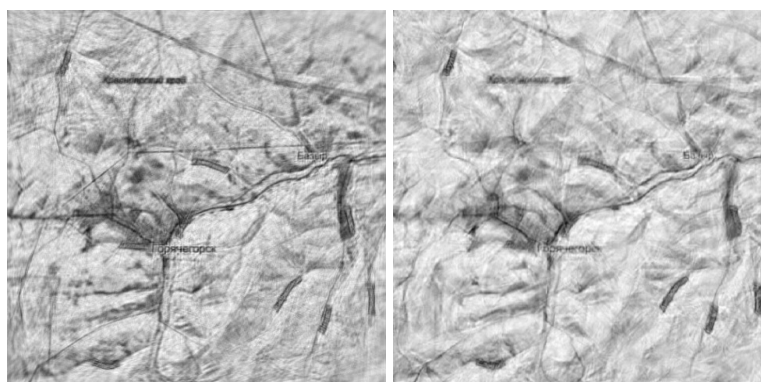
Проведен анализ алгоритмов **B** и **C** с использованием метрики PSNR и оценки визуального восприятия качества для различных изображений (табл. 3). Оценки визуального восприятия качества проводились тремя группами экспертов по 5 человек. Эксперты проводили оценку по 10 бальной шкале. Представлено сравнение алгоритмов **B** и **C** (метрика PSNR) при решении задачи шумоподавления для различных типов изображений.



**Рис. 2.** Решение задачи выделения линейных особенностей на основе алгоритма А. Исходное изображение (слева) и результат обработки (справа). На правом изображении слева направо и сверху вниз: исходное изображение, шярлет-коэффициенты для базисного шярлета, пример базисного шярлета, восстановленное изображение



**Рис. 3.** Решение задачи шумоподавления на основе алгоритма В. Исходное зашумленное изображение (слева) и результат шумоподавления (справа)



**Рис. 4.** Решение задачи шумоподавления на основе алгоритма С (слева) и D (справа)

**Таблица 3.** Значения метрики PSNR (в дБ) и оценок визуального восприятия качества изображений (в баллах)

Изображение	Алгоритм В	Алгоритм С	Визуальные оценки алгоритма В	Визуальные оценки алгоритма С
fire1	35.10	28.63	7.20	6.85
fire2	42.43	33.49	6.15	6.10
tomograf	33.00	27.33	9.20	8.85
satellite	31.44	25.57	8.20	7.85
satellite2	34.95	29.52	7.10	7.05

## Заключение

Таким образом, действие алгоритма **В** эффективнее по сравнению с алгоритмом **С** для всех наборов изображений как на основе количественных показателей (на 22–26%), так и на основе визуального восприятия качества. В тоже время алгоритм **С** превосходит алгоритм **В** по быстройдействию в 1.7–2.6 раза в зависимости от размера изображения.

Как показали исследования, для решения задачи разделения изображения на точки и кривые предлагается использовать алгоритм **С**, повышение точности оценивается в 5–12% по сравнению с применением кёрвлетов.

Решение задачи выделения контура на изображении предлагается выполнять с помощью модифицированного алгоритма **А**, который сравним по точности с классическими алгоритмами Собела и Превитта.

## Список литературы

1. Guo K., Labate D. Optimally Sparse Multidimensional Representation using Shearlets // *SIAM J Math. Anal.* 39 (2007), p. 298-318.
2. Guo K., Labate D., W.-Q Lim. Edge Analysis and Identification using the Continuous Shearlet Transform // *Appl. Comput. Harmon. Anal.* 27 (2009), p. 24-46.
3. Kutyniok G., Labate D. Construction of Regular and Irregular Shearlet Frames // *J. Wavelet Theory and Appl.* 1 (2007), p. 1-10.
4. Kutyniok G., Sauer T. From Wavelets to Shearlets and Back Again // In *Approximation Theory XII. Proceedings of the 12th International Conference, San Antonio, TX, USA, March 4-8, 2007*, p. 201-209.
5. Kutyniok G., Lemvig J., Lim W.-Q. Compactly Supported Shearlets // *Approximation Theory XIII (San Antonio, TX, 2010)*. – Springer, 2010.
6. Kutyniok G., Labate D. Introduction to Shearlets // In *Shearlets. Multiscale Analysis for Multivariate Data*. – Boston, MA: Birkhauser, 2012, p. 1-38.
7. Labate D., Lim W.-Q., Kutyniok G., Weiss G. Sparse Multidimensional Representation using Shearlets // *Wavelets XI (San Diego, CA, 2005)*, p. 254-262, *SPIE Proc.* 5914, SPIE. – Bellingham, WA, 2005.
8. Labate D., Easley G., Lim W. Sparse Directional Image Representations using the Discrete Shearlet Transform // *Applied Computational Harmonic Analysis*, 25 (2008), p. 25-46.
9. Lim W.-Q. The Discrete Shearlet Transform: A New Directional Transform and Compactly Supported Shearlet Frames // *IEEE Trans. Imag. Proc* 19 (2010), p. 1166-1180.
10. Lim W., Kutyniok G., Zhuang X. Digital Shearlet Transforms // *Shearlets: Multiscale Analysis for Multivariate Data*, Preprint, 2010.
11. Zhuang X. University of Osnaabrueck. ShearLab: A Rationally Designed Digital Shearlet Transform. Website: <http://shearlab.org/>.
12. Hauser S. Fast Finite Shearlet Transform: A Tutorial // Preprint of University of Kaiserslautern, 2011.
13. Hauser S. Fast Finite Shearlet Transform. Website: <http://www.mathematik.uni-kl.de/fileadmin/image/hauser/software/FFST.zip>.
14. Simonov K.V., Kirillova S.V., Cadena L. Fast Shearlet Transform Algorithms // *Abstracts of Lecturers and Young Scientists Second China-Russia Conference «Numerical Algebra with Applications»*. – Rostov-on-Don: Southern Federal University Publishing, 2013. – p. 122-123.

# Математическое Моделирование Информационных Процессов в Веб-Пространстве

Ю.И. Шокин<sup>1</sup>, А.Ю. Веснин<sup>2</sup>, А.А. Добрынин<sup>2</sup>,  
О.А. Клименко<sup>1</sup> и Е.В. Рычкова<sup>1</sup>

<sup>1</sup> Институт вычислительных технологий СО РАН, Новосибирск, Россия

<sup>2</sup> Институт математики им. С.Л. Соболева СО РАН, Новосибирск, Россия

**Аннотация.** В статье под веб-пространством понимается множество сайтов в Интернете с установленными между ними гиперссылками. Рассматривается часть веб-пространства, состоящая из сайтов научных организаций и университетов. Для исследования сайтов и связей между ними используется вебметрический анализ, основанный на данных, которые предоставляют поисковые системы. Информационные процессы в веб-пространстве моделируются с помощью ориентированных графов, в которых вершины соответствуют сайтам, а ребра — гиперссылкам. В работе представлены результаты исследования около 400 сайтов научных организаций России, Казахстана, Германии, Сербии, Китая и др.

**Ключевые слова:** веб-пространство, веб-граф, вебметрика, рейтинг.

## 1 Введение

Задача изучения веб-пространств является актуальной в связи с повсеместным распространением сети интернет и ростом объема представленных в ней ресурсов, усилением социальной значимости интернет-технологий. Под веб-пространством понимаются сайты или страницы сайтов в сети интернет, связанные друг с другом гиперссылками. Анализ свойств интернет-пространства как математического объекта впервые был начат в работах Р. Алберта и А.-Л. Барабаша [1]. Так как веб-пространство может представлять собой большую сложно организованную сетевую структуру, то проблематика исследований в этой области включает поиск адекватных представлений такой структуры, методов ее описания, изучение ее устойчивости и декомпозиции, нахождение численных и структурных параметров, характеризующих такую сеть, определение и предсказание изменений этих параметров в ходе эволюции сети. Для изучения содержательных и логических взаимосвязей между объектами интернета удобно использовать их представление в виде веб-графа. Как правило, при построении веб-графа его вершинам соответствуют отдельные страницы сайтов или сайты, рассматриваемые как единое целое. В настоящей работе под веб-графом понимается ориентированный граф, вершины которого соответствуют сайтам, а отношение между сайтами определяется их ссылками друг на друга.

Для изучения веб-пространства используются также методы вебметрики — современного раздела информатики, объектом изучения которого являются информационные ресурсы, структура и технологии интернета. Развитие этого направления началось в 1997 г. после работы Т. Алминда и П. Ингверсена [2]. Методы вебметрики позволяют оценивать эффективность и востребованность интернет-ресурсов с помощью доступных количественных показателей их деятельности. К таким показателям относятся, например, количество посещений, индексы цитируемости, степень связности определенных групп интернет-ресурсов, объемы доступных данных. Одним из направлений исследований, представляющих интерес для работы и развития научных интернет-сообществ, является анализ состояния веб-пространств, порождаемых сайтами университетов и академических научных организаций (см., например, [3],[4],[5],[6],[7],[8],[9],[10],[11],[12],[13]).



В настоящей работе изучаются веб-пространства сайтов научных организаций Сибирского отделения Российской академии наук (СО РАН), научных организаций, объединенных в Общество Фраунгофера в Германии (ОФ), научных учреждений, входящих в Сербскую академию наук и искусств и Zajednice instituta Srbije, университетов Казахстана и некоторых других стран Азии.

## 2 Рейтинг университетов Азии

Рейтинг **Ranking Web** или **Webometrics** является крупнейшим научным рейтингом высших учебных заведений мира [14]. Cybermetrics Lab (Испанский Национальный исследовательский совет, CSIC) начиная с 2004 года каждые шесть месяцев представляет независимое, объективное, свободное, открытое научное исследование для обеспечения надежной, многомерной, обновляемой и полезной информации о представленности университетов всего мира, основанной на их веб-присутствии и влиянии. Cybermetrics Lab занимается разработкой количественных исследований по академической сети начиная с середины девяностых годов.

В 2003 году, после публикации Академического рейтинга университетов мира (ARWU) [15], выполненного Шанхайским университетом Jiatong, Cybermetrics Lab решили принять основные нововведения, предложенные Лю (Liu) и его командой. Рейтинг строится из общедоступных веб-данных, объединенных с другими сведениями в сводный показатель. Первый рейтинг был представлен в 2004 году, затем, с 2008 года новые рейтинги можно видеть на портале [15] дважды в год.

Главной целью рейтинга Webometrics является содействие академическому веб-присутствию, поддержке инициативы Open Access для значительного расширения передачи научных и культурных знаний, созданных в университетском сообществе. Цель рейтинга Webometrics не заключается в оценке сайтов, их дизайна, удобства или популярности их содержания в зависимости от количества посещений и посетителей. Веб-параметры рассматриваются как доверенные представители правильной, всеобъемлющей, глубокой оценки эффективности университетов мира, принимая во внимание их деятельность и результаты, их значение и влияние. Важно понимать, что корректность рейтинга возможна только в том случае, когда присутствие в Интернете является полноценным «зеркалом» университета. Во втором десятилетии XXI века Интернет является ключевым для будущего всех университетов, это уже самый важный научный инструмент коммуникации, будущий канал для дистанционного обучения, открытый форум для общения и универсальная «витрина» для привлечения талантов, средств и ресурсов.

### 2.1 Методология

Webometrics использует **анализ ссылок** для оценки качества, так как это гораздо более мощный инструмент, чем анализ цитируемости [16]. Одним из основных вкладов Шанхайского рейтинга было ввести комбинированный показатель в сочетании с системой нормирования ряда параметров.

Комбинированный показатель рейтинга Webometrics в настоящее время строится следующим образом (в скобках указан процент вклада параметра в итоговое значение):

**Видимость (50 %)**

**Impact Rank** — внешние ссылки. Качество контента сайта оценивается с помощью так называемого «виртуального референдума», учитывающего все внешние входящие ссылки на сайт университета с других сайтов. Эти ссылки показывают институциональный престиж,

академическую эффективность, стоимость информации и полезность услуг, представленных на веб-страницах в соответствии с критериями миллионов веб-редакторов со всего мира. Данные по ссылочной видимости поступают от двух наиболее важных поставщиков этой информации: Majestic SEO [17] и ahrefs [18]. Оба используют свои собственные сканеры, создавая различные базы данных, которые следует использовать совместно для заполнения пробелов или исправления ошибок. Итоговый параметр равен квадратному корню из числа обратных ссылок и количества доменов, содержащих эти обратные ссылки, поскольку важна не только ссылочная популярность, но и ссылочное разнообразие. Максимум из нормированных результатов и является показателем влияния сайта.

#### *Активность (50 %)*

**Presence Rank** — объем сайта (1/3). Общее количество веб-страниц, размещенных на главном домене сайта (включая все поддомены и каталоги) университета и проиндексированных крупнейшим поисковиком Google [19]. Параметр учитывает каждую страницу сайта, различая все форматы, известные Google; учитывает как статические, так и динамические страницы, а также другие загруженные файлы. Невозможно иметь объемное присутствие организации в Интернете без вклада каждого сотрудника, т.к. главные соперники уже могут опубликовать миллионы веб-страниц. Наличие дополнительных доменов, альтернативных доменов на иностранных языках или доменов для маркетинговых целей, не учитывается в этом параметре, т.к. это очень запутывает внешних пользователей.

**Openness Rank** — загруженные файлы (1/3). Глобальные усилия по созданию научно-исследовательских репозиториях проявляются в этом параметре путем подсчета количества загруженных файлов (pdf, doc, docx, ppt), размещенных на веб-сайте в соответствии с требованиями академической поисковой системы Google Scholar [20]. Учитываются все файлы с правильными расширениями имен файлов (например, файлы Adobe Acrobat должны иметь расширение .pdf).

**Excellence Rank** — качество публикаций (1/3). Научные работы, опубликованные в международных журналах с высоким импакт-фактором играют очень важную роль в рейтинге университетов. Использование только общего количества публикаций может ввести в заблуждение, поэтому в рейтинге Webometrics данный параметр ограничивается лишь лучшими публикациями, т.е. теми, которые входят в 10 % наиболее цитируемых статей в соответствующих областях науки. Данные для этого параметра предоставляются группой Scimago [21], в частности, там имеются ненулевые значения для более 5000 университетов (2003–2010 годы).

Важно отметить, что в рейтинге Webometrics приводятся **относительные значения** описанных выше параметров, т.е., чем меньше значение каждого из параметров (Presence Rank, Impact Rank, Openness Rank, Excellence Rank), тем более высокую позицию в итоговом рейтинге занимает данный университет. Как специально отмечено в [16], абсолютные значения параметров недоступны для просмотра.

## 2.2 Анализ рейтинга университетов Азии

Webometrics помимо сводного рейтинга высших учебных заведений всего мира позволяет сделать выборку для отдельных регионов, например, Азии, включающую в себя университеты 46 стран [22]. Анализ того, университеты из каких стран попали в первую сотню рейтинга (рис. 1) показывает, что среди университетов 12 стран, университеты Китая составляют 44 %, университеты Японии, Тайваня, Кореи и Гонконга — 38 %, и 18 % — университеты остальных 7 стран. В табл. 1 приведены университеты Азии, находящиеся в первых 10 позициях рейтинга (по состоянию на январь 2015 года).

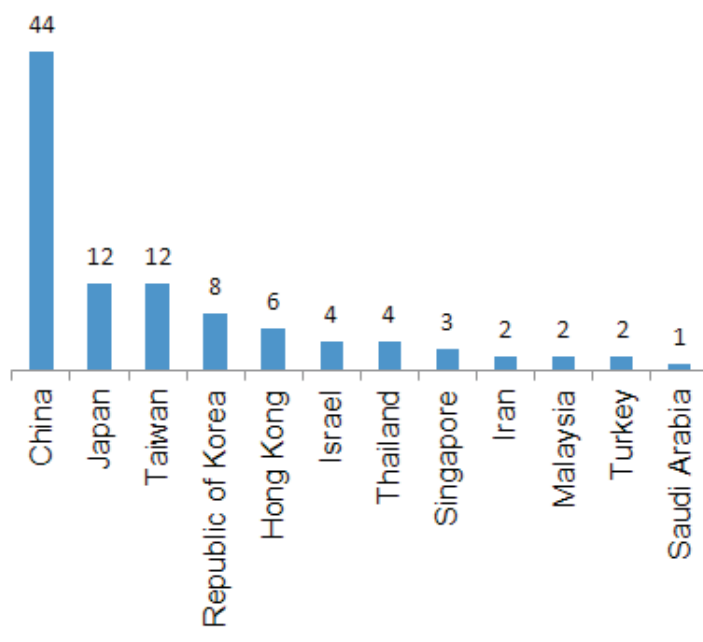


Рис. 1. Университеты Азии (по странам) из первой сотни рейтинга.

В Webometrics, в рамках региона Азия, также представлены рейтинги для каждой страны по отдельности, в частности, рейтинги для Казахстана [23], Киргизии [24] или Узбекистана [25]. В табл. 2 приведены университеты этих трех стран, находящиеся в первых 5 позициях рейтинга в своей стране. Из этой таблицы видно, что первый из казахских университетов, фигурирующий в рейтинге (Eurasian National University L.N. Gumilev) находится лишь в пятой сотне университетов Азии, а Kazakh National University Al Farabi — в седьмой сотне.

Если проанализировать данные табл. 1 и 2, то можно предположить, что университеты, находящиеся на более высоких позициях рейтинга, выигрывают за счет лучших показателей параметров Impact Rank (внешние ссылки) и Excellence Rank (качество публикаций).

**Таблица 1.** Университеты Азии, находящиеся в первых 10 позициях рейтинга.

Asia Rank	World Rank	University, country, web site	Presence Rank	Impact Rank	Openness Rank	Excellence Rank
1	30	National Taiwan University, Taiwan, <a href="http://www.ntu.edu.tw">www.ntu.edu.tw</a>	11	70	19	102
2	47	Peking University, China, <a href="http://www.pku.edu.cn">www.pku.edu.cn</a>	216	63	161	48
3	49	Tsinghua University China, China, <a href="http://www.tsinghua.edu.cn">www.tsinghua.edu.cn</a>	384	44	204	44
4	51	Seoul National University, Korea, <a href="http://www.snu.ac.kr">www.snu.ac.kr</a>	58	59	342	78
5	60	University of Tokyo, Japan, <a href="http://www.u-tokyo.ac.jp">www.u-tokyo.ac.jp</a>	208	109	176	25
6	65	Zhejiang University, (National Che Kiang University), China, <a href="http://www.zju.edu.cn">www.zju.edu.cn</a>	293	53	428	71
7	82	University of Hong Kong, <a href="http://www.hku.hk">www.hku.hk</a>	91	165	74	129
8	83	Shanghai Jiao Tong University, China, <a href="http://www.sjtu.edu.cn">www.sjtu.edu.cn</a>	38	195	140	101
9	87	Kyoto University, Japan, <a href="http://www.kyoto-u.ac.jp">www.kyoto-u.ac.jp</a>	97	260	71	62
10	88	Xiamen University, China, <a href="http://www.xmu.edu.cn">www.xmu.edu.cn</a>	194	73	36	371

**Таблица 2.** Университеты Казахстана, Киргизии, Узбекистана, находящиеся в первых 5 позициях рейтинга (в своей стране).

Asia Rank	World Rank	University, country, web site	Presence Rank	Impact Rank	Openness Rank	Excellence Rank
574	1836	Eurasian National University L.N. Gumilev, Kazakstan, www.enu.kz	1394	2242	203	203
713	2193	Kazakh National University Al Farabi, Kazakstan, www.kaznu.kz	3619	2860	694	3826
857	2528	Kazakh National Agriculture University, Kazakstan, www.kaznau.kz	6250	899	2828	5414
988	2837	Kazakh National Medical University Asfendiyarov, Kazakstan, kaznmu.kz	1948	2536	1827	5414
1034	2936	South Kazakhstan State University M.O. Auevov, Kazakstan www.ukgu.kz	659	3101	2472	5414
1935	5162	Kyrgyz-Russian Slavic University, Kyrgyzstan, www.krsu.edu.kg	5854	8648	2513	4808
2000	5326	Kyrgyz National University, Kyrgyzstan, www.university.kg	3727	4525	12377	4421
2322	6101	American University of Central Asia, Kyrgyzstan, www.auca.kg	5027	9399	2750	5414
2923	7657	Karakalpak State University, Uzbekistan, www.karsu.uz	9147	12445	3326	4421
3028	7924	National University of Uzbekistan, Uzbekistan, www.nuu.uz	10217	11397	7111	3695
3242	8483	Kyrgyz National Agrarian University, Kyrgyzstan, www.knau.kg	8661	5829	13206	5414
3633	9426	Tashkent University of Information Technologies, Uzbekistan, www.tuit.uz	8158	9823	8722	5414
4047	10520	Samarkand Agricultural Institute, Uzbekistan, www.samqxi.uz	7293	12307	13206	4158
4169	10809	Kyrgyz Turkish Manas University, Kyrgyzstan, manas.edu.kg	7077	16073	3039	5414
4210	10915	Samarkand State University, Uzbekistan, www.samdu.uz	5631	14263	13206	3826

Кроме того, повысить позицию также можно за счет активного размещения на сайте достаточно большого количества загруженных файлов (*Openness Rank*), особенно в формате Adobe Acrobat.

### 3 Анализ структуры веб-графов научных организаций методами теории графов

Традиционным подходом к анализу структуры веб-пространств, порождаемых веб-сайтами и гиперсвязями между ними, является использование методов теории графов [27],[28],[29]. Для этого в качестве модели веб-пространства используется веб-граф, в котором вершины соответствуют сайтам, а отношение между сайтами задается наличием ссылок между ними. Дуга графа выходит из вершины  $v$  и входит в вершину  $u$ , если сайт, соответствующий вершине  $v$ , содержит хотя бы одну ссылку на страницы сайта, соответствующего вершине  $u$ . Количество ссылок с одного сайта на другой задает вес соответствующей дуги (ссылки сайта на себя не учитываются). Таким образом, веб-граф является ориентированным графом, в котором любая пара вершин может быть соединена одной дугой или двумя противоположно направленными дугами. Число дуг в кратчайшем ориентированном пути между парой вершин равно наименьшему числу шагов при переходе по ссылкам с одного сайта на другой. При изображении графов и их фрагментов пара противоположно направленных дуг для наглядности будет заменяться одной дугой с двумя стрелками на концах.

Веб-графы научных организаций Общества Фраунгофера (ОФ), СО РАН и научных организаций, входящих в Сербскую академию наук и искусств и *Zajednice instituta Srbije*, обозначим через  $G$ ,  $R$  и  $S$  соответственно.

Веб-граф  $G$  содержит 72 вершины и 321 дугу и отражает связи научных организаций Общества Фраунгофера по состоянию на 8 апреля 2013 г. Список организаций ОФ представлены в [30].

Веб-граф  $R$  состоит из 95 вершин и 949 дуг соответствует состоянию веб-пространства научных организаций СО РАН на 29 октября 2013 г. В этот граф включены все научные организации из информационной системы “Организации и сотрудники СО РАН” [31]. Структура графа  $R$  приводится в [32].

Веб-граф  $S$  имеет 59 вершин и 106 дуг и отражает состояние веб-пространства научных организаций Республики Сербия и *Zajednice instituta Srbije* на 7 апреля 2013 г. Список организаций, соответствующих вершинам графа  $S$ , дан в [33]. Диаграммы всех трех графов представлены в [13].

#### 3.1 Численные характеристики структуры графов

Характеризация структуры графа численными параметрами, проведенная на основе анализа ее локальных фрагментов, полезна при изучении графов большого размера, так как часто не требует трудоемких расчетов. Для описания тех или иных структурных особенностей графов используют инварианты графов, которые, как правило, являются функциями, ставящими в соответствие графу некоторое число. Значение инвариантов зависит только от структуры графа, т.е. на изоморфных графах инвариант всегда принимает одинаковые значения. Далее рассмотрим инварианты, отражающие вклад вершин, дуг и окрестностей вершин графа в формирование его структуры.

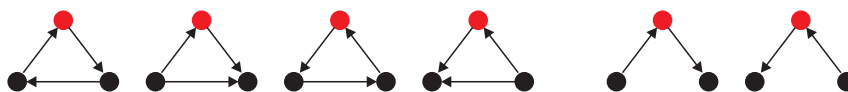
*Индекс вершин графа*  $c_v(H)$ . Этот параметр показывает какая часть сайтов веб-графа включена в информационное взаимодействие с другими сайтами, хотя бы попарное. Пусть ориентированный граф  $H$  имеет  $n$  вершин, и  $k$  из них имеют хотя бы одну исходящую или

входящую дугу. Индексом вершин в графе  $H$  называется величина  $c_v(H) = k/n$ . Близость  $c_v(H)$  к нулю говорит о том, что имеется значительное количество изолированных сайтов, то есть таких, которые не связаны с другими сайтами. Это может иметь место, например, в начальной стадии формирования веб-пространства. Равенство  $c_v(H) = 1$  означает, что каждая организация вовлечена в информационное взаимодействие с другими на уровне веб-сайтов.

*Индекс дуг графа  $c_a(H)$ .* Максимальное число дуг в ориентированном графе  $H$  с  $n \geq 2$  вершинами равно  $n(n-1)$ . Пусть число дуг в графе  $H$  равно  $t$ . Индексом дуг в графе  $H$  называется величина  $c_a(H) = t/n(n-1)$ . В [34] эта величина называется плотностью сети. Индекс дуг показывает какая часть дуг графа участвует в установлении информационного взаимодействия между сайтами. Максимальное значение,  $c_a(H) = 1$ , достигается когда любые две вершины графа  $H$  соединены парой противоположно направленных дуг. В этом случае все сайты ссылаются друг на друга, обеспечивая кратчайший переход с одного сайта на другой.

*Коэффициент кластеризации графа  $cc(H)$ .* Под окрестностью вершины  $v$  будем понимать множество вершин графа, соединенных с  $v$  дугами без учета их ориентации. Пусть  $V_2$  есть множество вершин ориентированного графа  $H$ , окрестность которых содержит не менее чем две вершины. Для вершины  $v$  графа  $H$  обозначим через  $H_v$  ориентированный подграф, порожденный окрестностью вершины  $v$ . Коэффициентом кластеризации вершины  $v$  называется величина  $c_a(H_v)$ , т. е. индекс дуг подграфа, порожденного окрестностью вершины  $v$  [35]. Коэффициент кластеризации графа  $H$  определим как  $cc(H) = \frac{1}{|V_2|} \sum_{v \in V_2} c_a(H_v)$ . Таким образом,  $cc(H)$  показывает как в среднем заполнены дугами окрестности вершин графа.

*Коэффициент транзитивности графа  $\tau(H)$ .* Рассмотрим в графе ориентированные пути длины 2, центральная вершина которых расположена сверху на рис. 2. Обозначим через  $N_\Delta$  число всех ориентированных путей длины 2 в графе  $H$  таких, что концевые вершины  $u$  и  $v$  этих путей соединены дугой без учета ориентации. Все четыре возможные конфигурации показаны слева на рис. 2. Число всех ориентированных путей длины 2, между концевыми вершинами которых нет дуг, обозначим через  $N_\wedge$  (см. две последние конфигурации на рис. 2). Коэффициент транзитивности ориентированного графа  $H$  определяется как  $\tau(H) = N_\Delta/N_\wedge$  [36].



**Рис. 2.** Конфигурации для вычисления коэффициента транзитивности.

*Диаметр графа  $diam(H)$ .* Расстояние между парой вершин в ориентированном графе  $H$  определяется как наименьший по числу дуг путь, соединяющий эти вершины, причем все вершины этого пути разные. Диаметр графа  $H$  определяется как наибольшее расстояние между парами вершин в графе. Таким образом, диаметр веб-графа показывает, какое наибольшее число шагов можно сделать по уникальным сайтам, переходя по ссылкам с сайта на сайт. Значения введенных выше численных параметров для веб-графов научных организаций  $G$ ,  $R$  и  $S$  представлены в табл. 3.

Значения индекса вершин  $c_v$  показывают, что в веб-пространстве научных организаций ОФ все сайты включены в информационное взаимодействие, в то время как некоторые сайты СО РАН и сербских научных организаций не связаны с другими. Значения индекса дуг  $c_a$  для графа  $R$  почти в два-три раза превышает значения для графов  $G$  и  $S$ . Значения

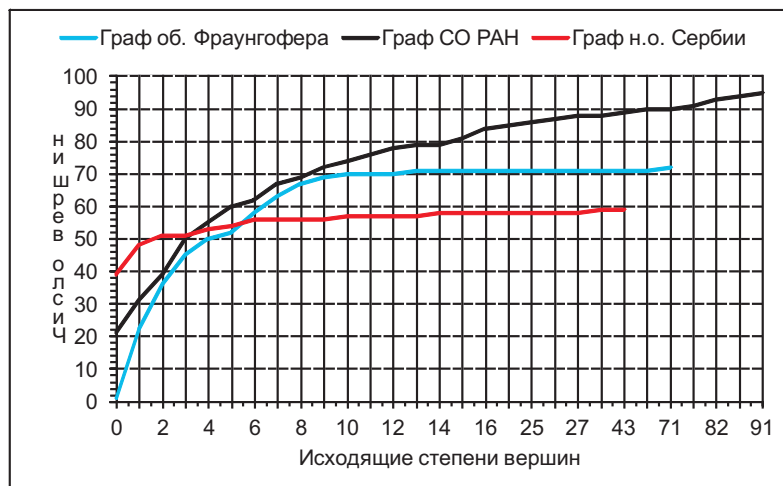
Таблица 3. Численные характеристики веб-графов.

инвариант	граф $G$	граф $R$	граф $S$
Индекс вершин, $c_v$	1,00	0,98	0,95
Индекс дуг, $c_a$	0,06	0,11	0,03
Коэф. кластеризации, $cc$	0,09	0,07	0,03
Коэф. транзитивности, $\tau$	0,10	0,24	0,07
Диаметр, $diam$	2	3	6

коэффициента кластеризации  $cc$  показывают, что в среднем заполнение дугами окрестности вершин во всех графах мало. Величины коэффициента транзитивности для графа  $R$  заметно выше, чем для других графов. Малый диаметр графа  $OF$  объясняется наличием вершины с почти максимально возможными полустепенями. Из этой вершины выходят дуги до всех других вершин, и в нее заходят дуги также из всех вершин, кроме одной.

### 3.2 Веб-коммуникаторы

Естественными характеристиками вершины  $v$  ориентированного графа являются число исходящих из нее дуг  $deg_+(v)$  (полустепень исхода) и число входящих в нее дуг  $deg_-(v)$  (полустепень захода). Вершина  $v$ , для которой  $deg_+(v) = deg_-(v) = 0$ , называется изолированной.

Рис. 3. Распределение вершин в графах  $G$ ,  $R$  и  $S$  по полустепеням исхода  $deg_+$ .

На рис. 3 и 4 приводятся графики функции распределения числа вершин веб-графов  $G$ ,  $R$  и  $S$  по их полустепеням исхода и захода. Функция распределения в точке  $k$  горизонтальной оси равна числу вершин с  $deg_{\pm}(v) = k$ .

Средние полустепени исхода/захода вершин в рассматриваемых графах равны  $avr(G) = 4,46$ ,  $avr(R) = 9,99$  и  $avr(S) = 1,8$  (суммы полустепеней исхода и захода всех вершин графа всегда равны). У трех графов наблюдается сильное различие в числе вершин, из которых не выходит ни одной дуги (1, 21 и 38 вершин). В графе  $G$  полустепень исхода у почти всех вершин ограничена значением 10, а в графе  $S$  — значением 6. В графах  $G$  и  $R$  из вершин с максимальной полустепенью исхода дуги идут почти ко всем вершинам графов, в то время как в графе  $S$  из подобной вершины можно перейти только в 73% вершин. Число вершин, в которые не входит ни одна дуга, во всех графах невелико (0, 2 и 3 изолированные вершины



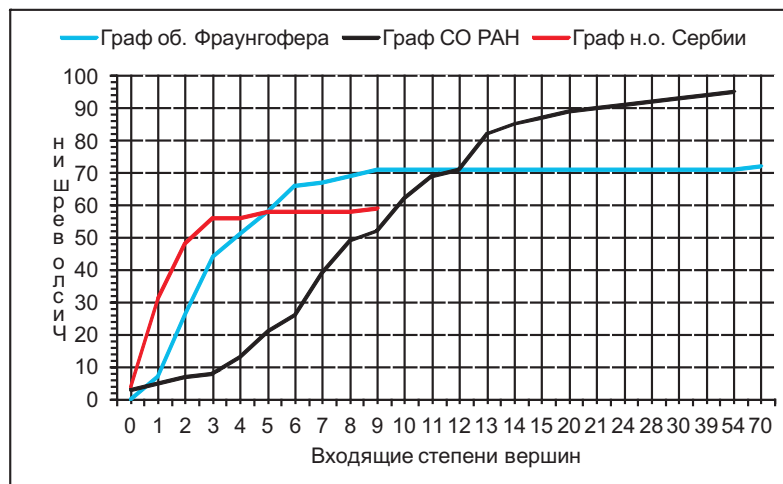


Рис. 4. Распределение вершин в графах  $G$ ,  $R$  и  $S$  по полустепеням захода  $deg_-$ .

в графах  $G$ ,  $R$  и  $S$  соответственно). Полустепени захода вершин в графе  $S$  ограничены значением 9, а подавляющее число вершин имеет полустепень захода не более 3. В графе  $G$  полустепени захода почти всех вершин тоже ограничены значением 9. В графах  $G$  и  $R$  существуют вершины, в которые заходят дуги из почти всех других вершин.

Доступность и использование информационных ресурсов веб-пространства можно характеризовать соотношением между полустепенями исхода и захода вершин веб-графа. Большие полустепени вершины обеспечивают тесные связи соответствующего сайта с остальным веб-пространством. Выделим три типа возможного соотношения числа входящих и исходящих дуг в вершину графа (см. рис. 5).



Рис. 5. Веб-коммуникаторы: индуктор, коллектор и посредник.

Вершины первого типа называют индукторами (мало входящих дуг, много исходящих), второго типа — коллекторами (много входящих дуг, мало исходящих), а третьего типа — посредниками (много как входящих, так и исходящих дуг). Вершины этих типов образует множество веб-коммуникаторов графа.

Коллекторы могут соответствовать сайтам организаций, в которых происходит накопление, хранение и обработка данных. Это могут быть библиотеки, хранилища данных, центры коллективного пользования и обработки данных, справочные ресурсы, журналы. Посредниками могут быть вершины, соответствующие головным сайтам в какой-то области науки, порталам научных центров, сайтам институтов с высокой степенью научной кооперации, официальным сайтам. Индукторами могут являться сайты недавно созданных организаций. Для отнесения вершины графа к веб-коммуникаторам того или иного типа используем численные параметры, характеризующие соотношение между ее полустепенями. Пороговые значения параметров зададим в зависимости от распределения полустепеней вершин в графе. Будем считать, что малые полустепени в веб-коммуникаторах должны быть меньше

средних полустепеней, в большие полустепени должны превышать средние полустепени. Более точно, используем следующие правила для определения веб-коммуникаторов: вершина  $v$  в графе  $H$  является

- индуктором, если  $\deg_-(v) < avr(H)$ ,  $\deg_+(v) > avr(H)$  и  $\deg_+(v)/\deg_-(v) > \Delta_i$ ;
- коллектором, если  $\deg_-(v) > avr(H)$ ,  $\deg_+(v) < avr(H)$  и  $\deg_-(v)/\deg_+(v) > \Delta_c$ ;
- посредником, если  $\deg_-(v) > avr(H)$ ,  $\deg_+(v) > avr(H)$  и  $|\deg_+(v) - \deg_-(v)| \leq \Delta_t$ ,

где  $avr(H)$  – средняя степень вершин в графе  $H$ , а  $\Delta_i$ ,  $\Delta_c$  и  $\Delta_t$  – заданные границы. В табл. 4 представлена информация о веб-коммуникаторах в графах научных организаций РФ, СО РАН и Сербии при  $\Delta_i = \Delta_c = \Delta_t = 2$ . Средние степени вершин показаны в верхней части таблицы. В скобках рядом с числом веб-коммуникаторов указан их процент от числа вершин. В скобках в столбцах таблицы приводятся значения полустепеней веб-коммуникаторов.

**Таблица 4.** Веб-коммуникаторы в графах  $G$ ,  $R$  и  $S$  при  $\Delta_i = \Delta_c = \Delta_t = 2$ .

	Общ. Фраунгофера $avr(G) = 4,5$		Сибирское отделение $avr(R) = 9,9$		Научные орг. Сербии $avr(S) = 1,8$	
	число	( $\deg_-, \deg_+$ )	число	( $\deg_-, \deg_+$ )	число	( $\deg_-, \deg_+$ )
Индуктор	5(7%)	(1,6) (1,7) (2,8) (2,10) (3,13)	1(1%)	(2,43)	2(3%)	(4,1) (14,1)
Посредник	5(7%)	(6,6) (6,7) (7,9) (9,8) (70,71)	7(7%)	(10,10) (11,11) (13,12) (15,13) (14,15)	5(8%)	(2,2) (2,3) (4,2) (5,3) (10,11)
Коллектор	7(10%)	(5,1) (6,1) (8,1) (9,2)	6(6%)	(11,1) (11,2) (10,2) (13,1) (13,2)	1(2%)	(1,5)

В множество веб-коммуникаторов попадают 25 % вершин графа  $G$ , в то время как в графах  $R$  и  $S$  число таких вершин составляет 14 % и 13 % соответственно. При  $\Delta_t = 2$  вершины с большими полустепенями могут не попадать в множество посредников. Например, вершина с полустепенями (91,54), соответствующая сайту ПОРТАЛ СО РАН, не будет посредником, хотя по своему положению в веб-пространстве СО РАН является им. Если при нахождении веб-коммуникаторов не налагать ограничений на разницу между полустепенями вершин (положить, например,  $\Delta_t = 100$ ), то графы  $G$ ,  $R$  и  $S$  будут иметь 7 % (5 вершин), 21 % (20) и 14 % (8) посредников.

### 3.3 Представление структуры графов

Для описания структуры больших веб-пространств часто используется их представление в виде модели “галстук-бабочка” [37]. В веб-графе выделяется максимальная сильно связанная компонента, по отношению к которой классифицируются остальные вершины графа. Подграф называется *сильно связанной компонентой* графа, если между любой парой его вершин

существует ориентированный путь. Проходя по ссылкам сайтов, вершины которых попали в такую компоненту, можно всегда посетить любую вершину компоненты.

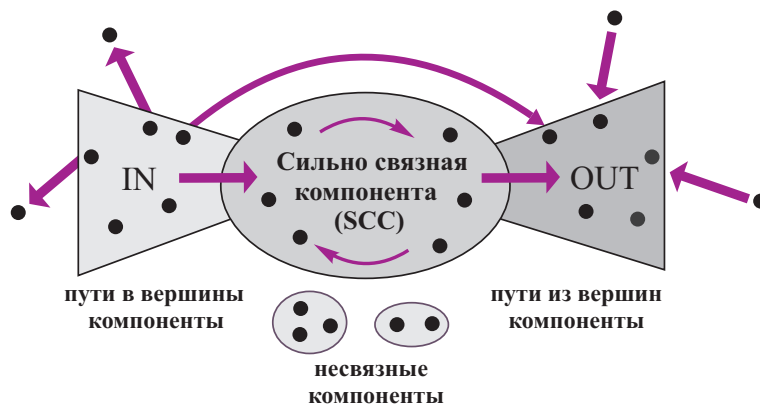


Рис. 6. Представление веб-графа в модели “галстук-бабочка”.

На рис. 6 показано как выглядит веб-граф в таком представлении. Центральную часть образует максимальная сильно связанная компонента (SCC). Левая часть (IN) состоит из вершин, пути из которых ведут в эту компоненту. Правую часть (OUT) образуют вершины, в которые ведут пути из компоненты SCC. Вершины подграфов, называемых отростками (tendrils), связаны путями только с множествами вершин IN и OUT. Эти пути изображены на рис. 6 в виде толстых стрелок, выходящих из IN или входящих в OUT. Некоторые вершины из множества IN могут быть связаны с вершинами из OUT путями, избегающими сильно связанную компоненту SCC. Такие подграфы, называемые туннелями (tubes), изображены на рис. 6 в виде толстой дуги из IN в OUT. Другие подграфы, вершины которых не связаны путями с описанными выше частями графа, образуют отдельные компоненты (показаны внизу на рис. 6).

На рис. 7 показана структура веб-графов  $S$ ,  $R$  и  $G$  научных организаций Сербии, СО РАН и Общества Фраунгофера в виде модели “галстук-бабочка”. В процентах указано количество вершин в частях графов. В графах организаций Сербии и СО РАН несвязные компоненты являются изолированными вершинами. В графе Сербии есть один туннель. У графа Общества Фраунгофера левая часть IN не содержит вершин.

Напомним, что центральная часть бабочки имеет полезное свойство — из любого сайта в SCC всегда можно достичь информационных ресурсов другого сайта в этой части. Размер множества SCC в веб-графах рассматриваемых академических сообществ увеличивается с 29 % до 99 % от всех вершин в графах  $S$  и  $G$ . Большой размер множества SCC в графе Общества Фраунгофера обеспечивается существованием портала, из которого выходят дуги во все другие сайты, и в него также входят дуги из почти всех вершин (кроме одной).

Небольшой размер веб-графа научных организаций Сербии дает возможность наглядно показать расположение его частей в модели “галстук-бабочка” (рис. 8). Вершины сильно связанной компоненты, SCC, изображены красным цветом (17 вершин), вершины множества OUT выделены черным цветом (38 вершин), а единственная вершина части IN показана зеленым цветом. Из нее выходят две дуги — одна в красную вершину из SCC, а другая — в черную вершину из OUT (туннель). Три изолированные вершины образуют несвязанные компоненты (выделены синим цветом).

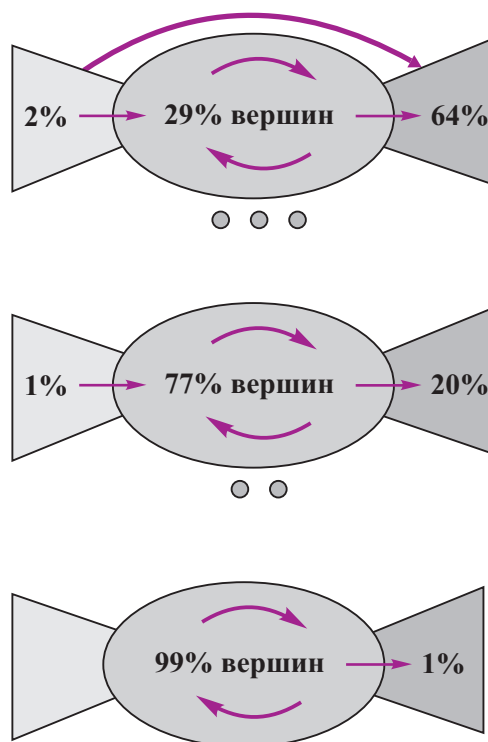


Рис. 7. Представление структуры веб-графов  $S$ ,  $R$  и  $G$ .

### 3.4 Сообщества наиболее тесно связанных организаций

Для изучения организации академических сообществ представляет интерес выявление в веб-графах подмножеств вершин, которые могут быть достижимы друг из друга за небольшое число шагов. Основанием ориентированного графа будем называть неориентированный граф с тем же множеством вершин, в котором пара вершин соединена неориентированным ребром, если эти вершины были соединены в исходном графе одной дугой или двумя противоположно направленными дугами. Таким образом, ребро основания ориентированного графа отражает факт наличия ссылок между парой соответствующих сайтов без указания направления ссылок. Рассмотрим типы подграфов, которые могут образовывать группы тесно связанных друг с другом организаций. Любая пара вершин в таких подграфах соединена хотя бы одной дугой, а типы подграфов определяются ориентацией дуг. Неориентированный подграф называется *полным*, если любая пара его вершин соединена ребром. Полный подграф называется *кликой* графа, если он не содержится ни в каком другом полном подграфе, т.е. является максимальным по включению. Кликку с числом вершин  $k$  будем называть  $k$ -кликкой. Клики основания графа  $H$  порождают в  $H$  ориентированные подграфы, которые будем называть ориентированными кликами. В зависимости от ориентации дуг выделим два типа ориентированных клик:

- ориентированная клика называется *компактной*, если каждая пара ее вершин соединена двумя противоположно направленными дугами. Компактная клика представляет собой оптимальный фрагмент с точки зрения скорости доступа к информационным ресурсам ее вершин. Эволюция веб-графов организаций в какой-либо узкой области науки может приводить к компактной клике;



В графе СО РАН группы близких вершин имеют максимальную мощность 11, в то время как в двух других графах размер таких групп не превышает 6. Разницу между числом полных подграфов и сильных и компактных клик образуют подграфы, в которых есть вершины, не достижимые друг из друга. Компактные клики в графе  $R$  отсутствуют, а в графах  $G$  и  $S$  такие подграфы порождаются некоторыми парами и тройками вершин.

#### 4 Заключение

С помощью вебометрического анализа и методов теории графов изучено веб-пространство, состоящее из сайтов научных организаций и университетов России, Казахстана и других стран. Показано, что авторитетность сайта в академическом сообществе зависит от числа и разнообразия внешних ссылок на сайт. Повысить авторитетность сайта можно за счет размещения на нем презентаций докладов на конференциях и семинарах, полных текстов научных статей, баз данных в предметных областях.

#### Список литературы

1. Albert R., Barabási A.-L. Statistical mechanics of complex networks // *Reviews of Modern Physics*. – 2002. – V. 74, № 1. – P. 47–97.
2. Almind T., Ingwersen P. Infometric analyses on the World Wide Web: Methodological approaches to “webometrics” // *J. Document.* – 1997. – V. 53, № 4. – P. 404–426.
3. Thelwall M., Wilkinson D. Graph structure in three national academic webs: power laws with anomalies // *Am. Soc. Inf. Sci. Technol.* – 2003. – V. 54, № 8. – P. 706–712.
4. Stuart D., Thelwall M., Harries G. UK academic web links and collaboration — an exploratory study // *J. Inf. Sci.* – 2007. – V. 33, № 2. – P. 231–246.
5. Шокин Ю.И., Клименко О.А., Рычкова Е.В., Шабальников И.В. Рейтинг сайтов научных организаций СО РАН // *Вычислительные технологии*. – 2008. – Т. 13, № 3. – С. 128–135.
6. Мазалов В.В., Печников А.А. О рейтинге официальных сайтов научных учреждений северо-запада России // *Управление большими системами*. Выпуск 24. – М.: ИПУ РАН, 2009. – С. 130–146.
7. Шокин Ю.И., Веснин А.Ю., Добрынин А.А., Клименко О.А., Рычкова Е.В., Петров И.С. Исследование научного веб-пространства Сибирского отделения Российской академии наук // *Вычислительные технологии*. – 2012. – Т. 17, № 6. – С. 86–98.
8. Pechnikov A.A., Nwohiri A.M. Webometric analysis of Nigerian university websites // *Webology*. – 2012. – V. 9, № 1. – URL: <http://www.webology.org/2012/v9n1/a95.html>
9. Печников А.А. Применение вебометрических методов для исследования информационного веб-пространства научной организации (на примере Карельского научного центра РАН) // *Труды КарНЦ РАН*. No 1. Сер. Математическое моделирование и информационные технологии. Вып. 4. Петрозаводск: КарНЦ РАН. – 2013. – С. 86–95.
10. Веснин А.Ю., Константинова Е.В., Савин М.Ю. О сценариях присоединения новых сайтов к веб-пространству СО РАН // *Вестник НГУ, серия: информационные технологии*. – 2013. – Т. 11, № 4. – С. 28–37.
11. Shokin Yu.I., Vesnin A.Yu., Dobrynin A.A., Klimenko O.A., Konstantinova E.V., Petrov I.S., Rychkova E.V. Investigation of the academic web space of the Republic of Serbia // *Zbornik radova konferencije MIT 2013*. – Belgrad, Serbia, 2014. – С. 601–607.
12. Шокин Ю.И., Веснин А.Ю., Добрынин А.А., Клименко О.А., Рычкова Е.В. Анализ веб-пространства академических сообществ методами вебометрики и теории графов // *Информационные технологии*. – 2014. – № 12. – С. 31–40.
13. Шокин Ю.И., Веснин А.Ю., Добрынин А.А., Клименко О.А., Рычкова Е.В., Филиппова М.Я. Построение и исследование математической модели веб-пространства // XV Российская конференция с участием иностранных ученых “Распределенные информационные и вычислительные ресурсы — DICR-2014” Новосибирск, Россия, 02.11–05.11.2014): Материалы конференции. – Новосибирск: ИВТ СО РАН, 2014. – Рег. номер 0321500379. – URL: [http://konf.ict.nsc.ru/files/conferences/dicr2014/fulltext/249005/250429/shokin\\_i\\_dr\\_web.pdf](http://konf.ict.nsc.ru/files/conferences/dicr2014/fulltext/249005/250429/shokin_i_dr_web.pdf)
14. Ranking Web of Universities [Электрон. ресурс]. – URL: <http://webometrics.info> (дата обращения январь 2015).

15. Академический рейтинг университетов мира [Электрон. ресурс]. – URL: <http://www.shanghairanking.com/> (дата обращения январь 2015).
16. Ranking Web of Universities. Methodology [Электрон. ресурс]. – URL: <http://webometrics.info/en/Methodology> (дата обращения январь 2015).
17. Majestic SEO – Site Explorer & Backlink Checker [Электрон. ресурс]. – URL: <http://www.majesticseo.com/> (дата обращения январь 2015).
18. Ahrefs – Site Explorer & Backlink Checker [Электрон. ресурс]. – URL: <http://ahrefs.com/> (дата обращения январь 2015).
19. Поисковая система Google [Электрон. ресурс]. – URL: <http://www.google.com/> (дата обращения январь 2015).
20. Академическая поисковая система Google Scholar [Электрон. ресурс]. – URL: <http://scholar.google.com/> (дата обращения январь 2015).
21. Scimago Institutions Rankings [Электрон. ресурс]. – URL: <http://www.scimagoir.com/> (дата обращения январь 2015).
22. Ranking Web of Universities / Asia [Электрон. ресурс]. – URL: <http://webometrics.info/en/Asia> (дата обращения январь 2015).
23. Ranking Web of Universities / Kazakstan [Электрон. ресурс]. – URL: <http://webometrics.info/en/Asia/Kazakstan> (дата обращения январь 2015).
24. Ranking Web of Universities / Kyrgyzstan [Электрон. ресурс]. – URL: <http://webometrics.info/en/Asia/Kyrgyzstan> (дата обращения январь 2015).
25. Ranking Web of Universities / Uzbekistan [Электрон. ресурс]. – URL: <http://webometrics.info/en/Asia/Uzbekistan> (дата обращения январь 2015).
26. Рейтинг сайтов научных организаций СО РАН [Электрон. ресурс]. – URL: <http://www.ict.nsc.ru/ranking> (дата обращения январь 2015)
27. Харари Ф. Теория графов. – М.: Мир, 1973.
28. Емеличев В.А., Мельников О.И., Сарванов В.И., Тышкевич Р.И. Лекции по теории графов. – М.: Наука, 1990.
29. Рейнгольд Э., Нивергельт Ю., Део Н. Комбинаторные алгоритмы. Теория и практика. – М.: Мир, 1980.
30. Связи научных организации Общества Фраунгофера [Электрон. ресурс]. – URL: <http://ousnano.sbras.ru/sitepage.php?PageID=2505> (дата обращения - 27.09.2013)
31. Информационная система “Организации и сотрудники СО РАН” [Электрон. ресурс]. – URL: <http://www.sbras.ru/sbras/db/> (дата обращения январь 2015)
32. Связи научных организаций СО РАН [Электрон. ресурс]. – URL: <http://ousnano.sbras.ru/sitepage.php?PageID=3008> (дата обращения январь 2015)
33. Веб-граф институтов Сербии. [Электрон. ресурс]. – URL: <http://ousnano.sbras.ru/sitepage.php?PageID=2506> (дата обращения январь 2015).
34. Hage P., Harary F. Structural models in anthropology. – Cambridge University Press: Cambridge, UK, 1983.
35. Watts D., Strogatz S. Collective dynamics of small world networks // Nature. – 1998. – V. 393. – P. 440–442.
36. Opsahl T., Panzarasa P. Clustering in weighted networks // Social Networks. – 2009. – V. 31. – P. 155–163.
37. Broder A., Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkins A., Wiener J. Graph structure in the Web // Computer Networks. – 2000. – V. 33, № 1–6. – P. 309–320.

# Использование Разнородных Данных при Сегментации Спутниковых Изображений Высокого Разрешения

Ю.Н. Синявский, И.А. Пестунов, О.А. Дубровская, П.В. Мельников, С.А. Рылов,  
Д.В. Лазарев

Институт вычислительных технологий СО РАН, Новосибирск, Россия  
yorikmail@gmail.com, {pestunov, olga}@ict.nsc.ru

**Аннотация.** Предложена логическая схема единообразного представления разнородных пространственных данных. На ее основе разработана технология сегментации спутниковых изображений высокого пространственного разрешения, которая позволяет учесть всю имеющуюся информацию (спектральные и пространственные признаки, данные полевых наблюдений, тематические карты, базы данных, априорные знания и т.п.). Технология реализована в виде набора стандартизованных веб-сервисов.

**Ключевые слова:** сегментация спутниковых изображений, высокое пространственное разрешение, текстурные и контекстные признаки, обработка разнородных данных, веб-сервисы.

## 1 Введение

В настоящее время в области создания и развития средств и технологий дистанционного зондирования Земли наблюдается стремительный прогресс. Целый ряд спутников (WorldView-2/3, Ресурс-П, GeoEye-1, Pleiades, Kompsat-3 и др.) обеспечивает регулярную поставку мультиспектральных изображений высокого пространственного разрешения (4 м и лучше) [1]. Характерная особенность таких изображений заключается в том, что значительная часть информации, необходимая для их анализа, содержится в пространственных характеристиках (текстура, форма, размер, контекст и т.п.), а также в накопленных базах данных, имеющейся априорной информации и т.п. [2]. Традиционные методы сегментации, учитывающие лишь спектральные признаки, оказываются неэффективными для автоматизированного анализа таких изображений. Получаемые с их помощью картосхемы характеризуются чрезмерной раздробленностью и малоприспособны для дальнейшей интерпретации специалистами предметных областей. Поэтому разработка эффективных методов и технологий обработки и анализа изображений высокого пространственного разрешения, обеспечивающих использование всей доступной информации, является актуальной, но малоизученной проблемой [3].

В работе рассматриваются схема единообразного представления данных и технология сегментации изображений высокого пространственного разрешения, позволяющая при обработке использовать всю доступную разнородную информацию. Технология реализована в виде набора стандартизованных веб-сервисов. Приводятся примеры решения практических задач.

## 2 Логическая схема единообразного представления данных

Логическая схема единообразного представления разнородных данных представлена на рис. 1. В соответствии с этой схемой, все доступные данные используются для формирования набора растровых слоев, которые при дальнейшей обработке интерпретируются как дополнительные признаки.



Полученные слои можно разделить на слои данных и тематические слои. К слоям данных, помимо спектральных каналов исходного изображения, можно отнести построенные ранее тематические карты (растительности, почв, влагосодержания и др.) и карты, формируемые автоматически (среднесуточной температуры, количества осадков, альbedo подстилающей поверхности, высот, экспозиции и ориентации склонов и т.д.). Для формирования тематических слоев (в отличие от слоев данных) необходимо участие эксперта. К ним относятся текстурные и контекстные признаки, полученные с использованием различных процедур фильтрации изображения, а также комплексные спектральные признаки (индексы). К тематическим слоям также относятся бинарные маски, построенные на основе имеющихся априорных знаний и предназначенные для выделения конкретных типов объектов (водной поверхности, теней, растительности, антропогенных территорий и др.). Кроме того, к те-

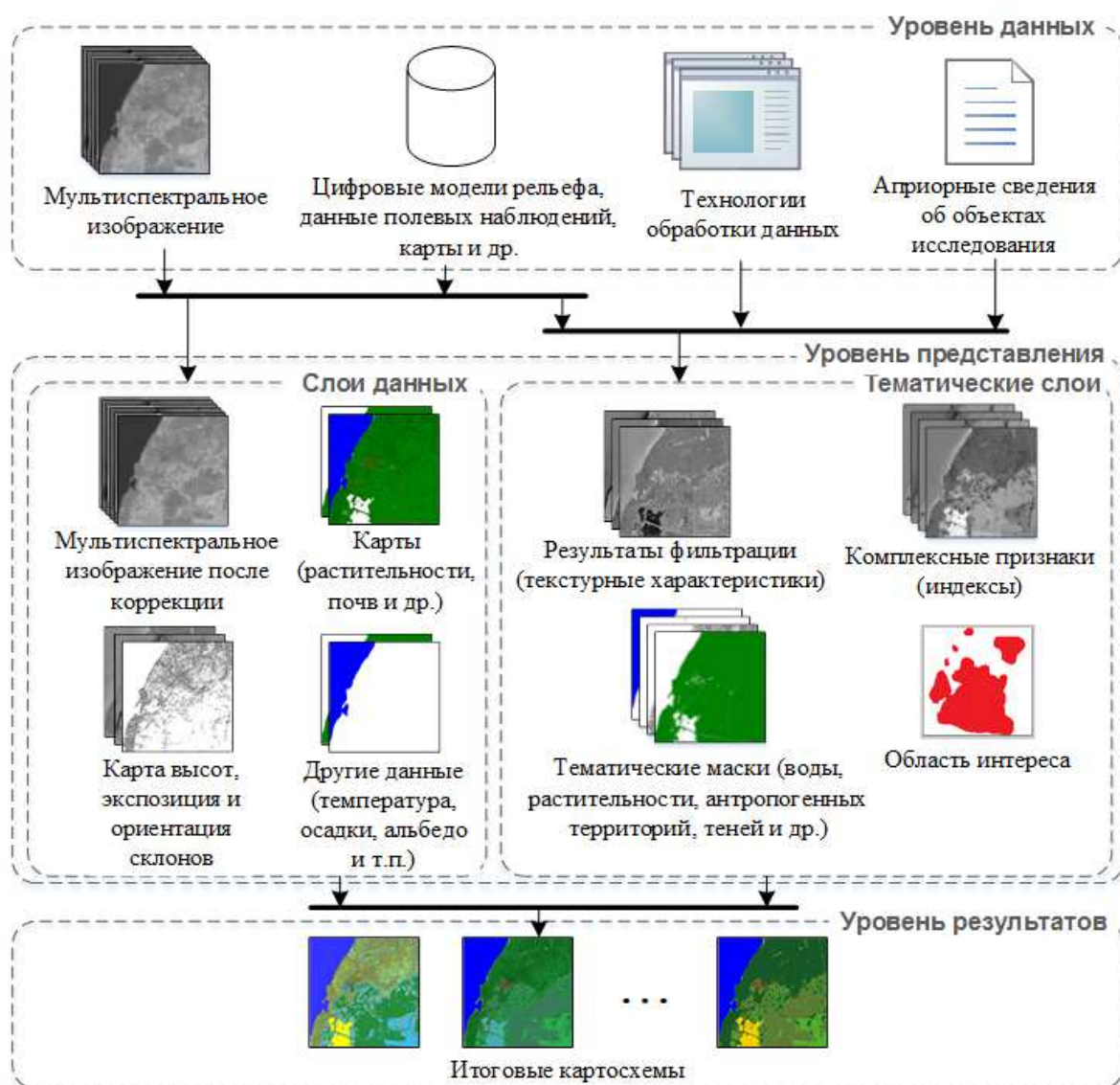


Рис. 1. Логическая схема единообразного представления разнородных пространственных данных.

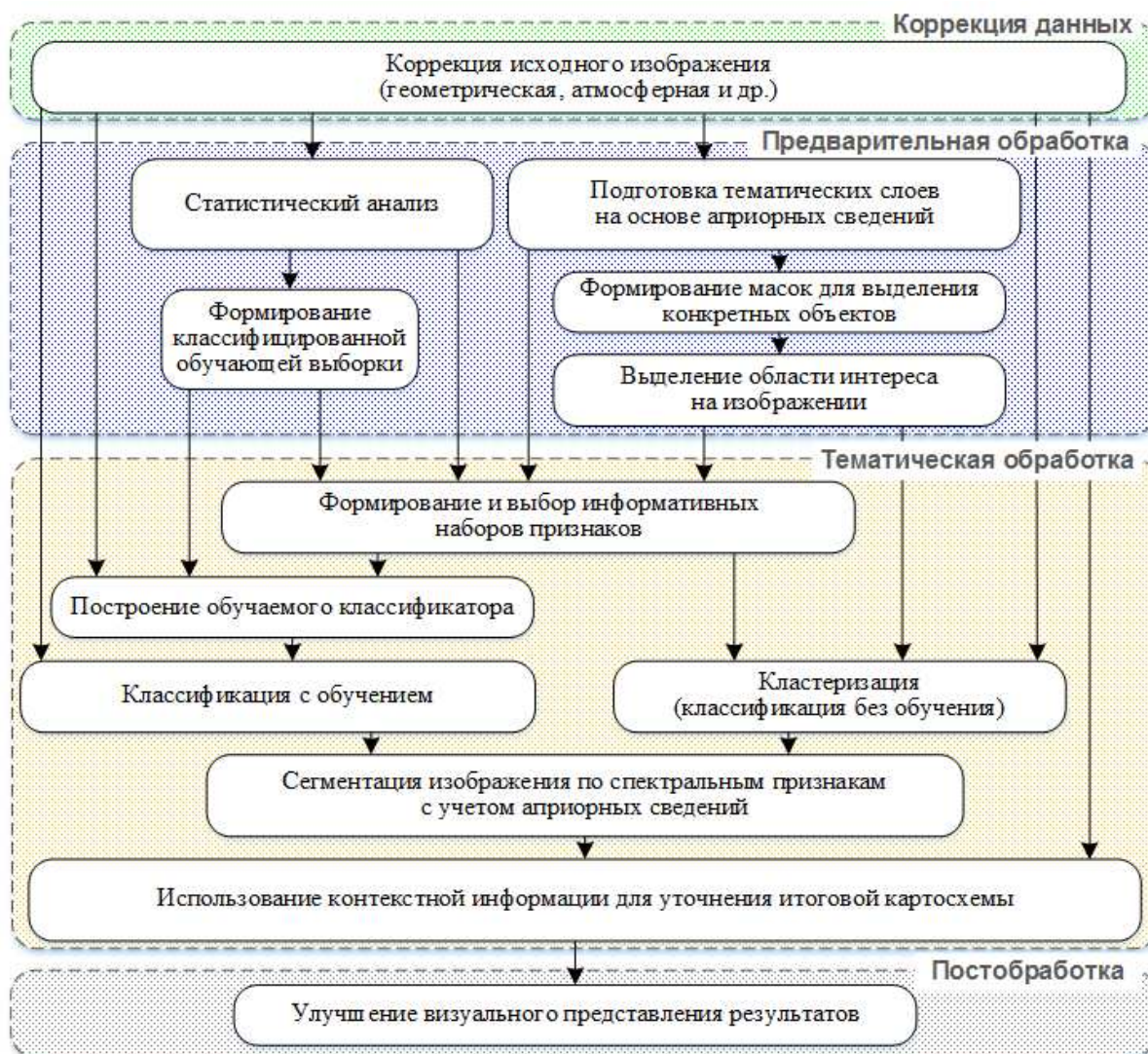


Рис. 2. Поэтапная технология сегментации спутниковых изображений высокого пространственного разрешения.

матическим слоям можно отнести маску, которая позволяет выделить область интереса, определяемую экспертом исходя из особенностей решаемой задачи.

### 3 Технология совместной обработки пространственных данных

Подобное представление разнородной информации позволило разработать технологию сегментации спутниковых изображений высокого пространственного разрешения (рис. 2), предназначенную для исследования и оценки состояния природных объектов.

Разработанная технология включает четыре этапа обработки. На этапе коррекции данных выполняются атмосферная и геометрическая коррекции, а также приведение данных к единой картографической проекции и т.п. Этап предварительной обработки включает в себя формирование слоев данных и тематических слоев, выделение области интереса на

изображении, а также статистический анализ и формирование классифицированной обучающей выборки. На этапе тематической обработки осуществляется выбор информативных наборов признаков и сегментация изображения с помощью алгоритмов классификации с обучением или кластеризации данных [4], а также уточнение результатов с использованием текстурной и контекстной информации. Завершающий этап (этап постобработки) направлен на улучшение визуальных характеристик результирующих картосхем для облегчения их интерпретации (применение различных фильтров, выбор уровня детализации картосхем и т.п.). Для лучшего понимания разработанной технологии рассмотрим пример: разделение типов растительности по снимку, полученному со спутника WorldView-2.

К изображению после первого этапа обработки, представленному на рис. 3, *а*, был применен непараметрический ансамблевый алгоритм кластеризации ЕССА [5]. В результате была построена картосхема (рис. 3, *б*) из 61 кластера, большинство из которых не являются растительностью.

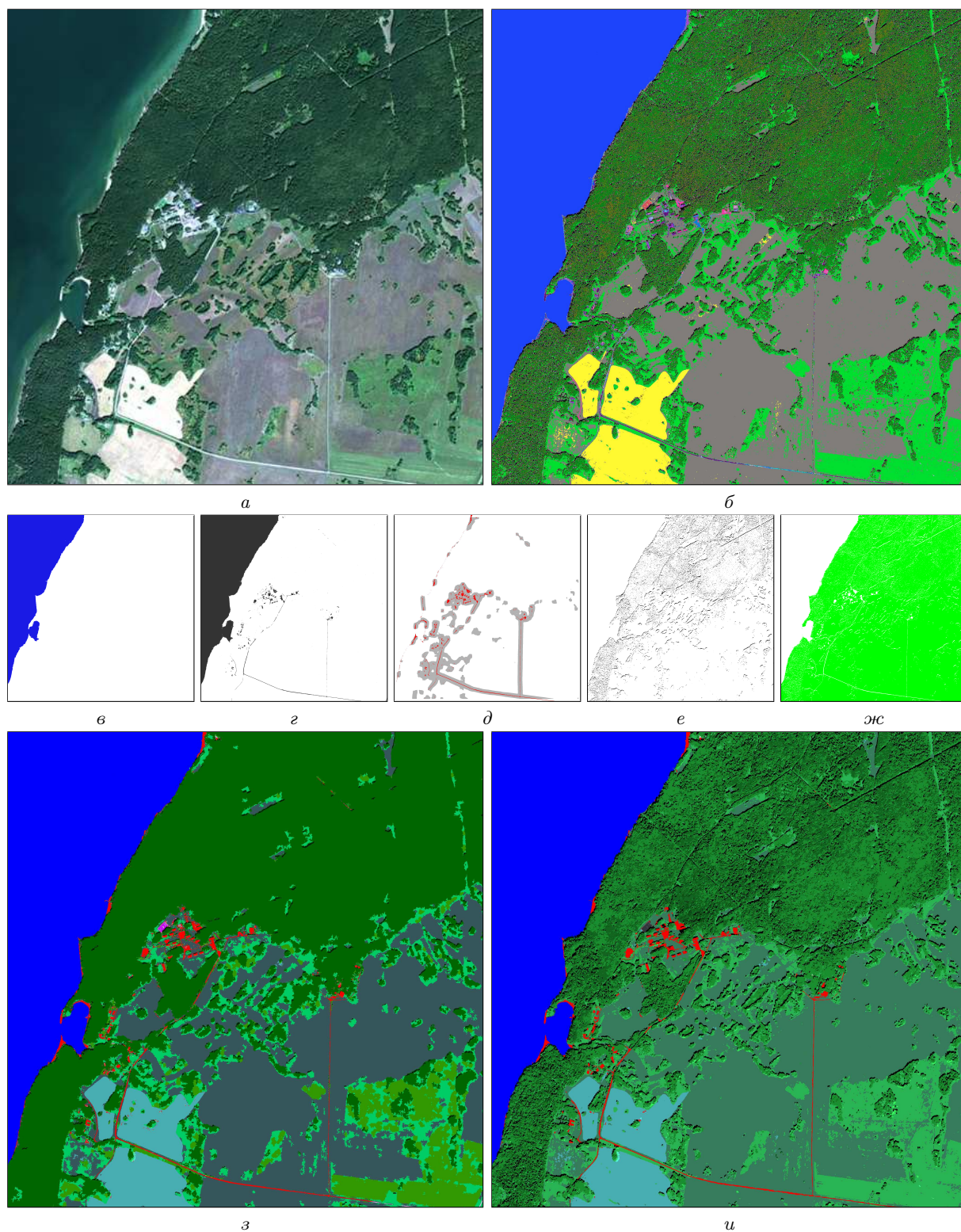
В соответствии с предложенной технологией, на этапе предварительной обработки выполнено формирование четырех тематических масок, которые предназначены для выделения объектов, не представляющих интереса при решении данной задачи. Первая маска (рис. 3, *в*) построена на основе нормализованного разностного водного индекса (NDWI) [6], который реагирует на степень увлажненности территории. Она и позволяет с высокой степенью достоверности выделить поверхность воды (на рисунке выделена синим цветом). Вторая маска (рис. 3, *г*) построена с использованием нормализованного вегетационного индекса (NDVI). Она позволяет выделить территории, не покрытые растительностью (на рисунке выделены серым цветом). Третья маска (рис. 3, *д*) предназначена для выделения на изображении объектов неприродного происхождения (зданий, сооружений, дорог), характерной особенностью которых является наличие прямых границ и углов, а также отсутствие растительности. Для локализации областей, содержащих прямые линии и углы (на рис. 3, *д* выделены серым цветом), к изображению применялись детекторы Харриса [7]. Комбинирование полученной маски с маской территорий, не покрытых растительностью (см. рис. 3, *г*), и результатами непараметрической кластеризации (см. рис. 3, *б*) позволило обнаружить объекты неприродного происхождения, выделенные на рис. 3, *д* красным цветом. Четвертая маска (рис. 3, *е*) построена с использованием метода, предложенного в [8], и предназначена для выделения теней на изображении (на рисунке выделены черным цветом). Полученные маски позволили распознать и исключить из дальнейшего рассмотрения объекты, не представляющие интереса при решении данной задачи, а также выделить область интереса (на рис. 3, *ж* выделены зеленым цветом).

На этапе тематической обработки построенные тематические маски использовались для выделения на изображении трех классов («поверхность воды», «тени» и «объекты неприродного происхождения») и области интереса, которая в дальнейшем была обработана алгоритмом ЕССА для разделения различных типов растительности. Результирующая картосхема была обработана алгоритмом сегментации по текстурным признакам ESEG [9] (рис. 3, *з*) и алгоритмом, учитывающим локальный контекст изображения (соседство пикселей) [10] (рис. 3, *и*). Оба алгоритма позволили выделить 3 класса («поверхность воды», «тени» и «объекты неприродного происхождения») и 18 кластеров, попадающих в область интереса.

### 3.1 Система стандартизованных веб-сервисов

Описанная технология реализована в виде набора стандартизованных веб-сервисов (WPS-процессов). Набор внедрен в сервис-ориентированную геоинформационную систему, создан-





**Рис. 3.** *a* — RGB-композит изображения WorldView-2 (каналы 5, 3 и 2); *б* — результат сегментации по спектральным признакам (61 кластер); *в-е* — тематические маски; *ж* — область интереса; *з* — результат сегментации с учетом текстурных признаков (3 класса и 18 кластеров, попадающих в область интереса); *и* — результат сегментации с учетом информации о соседстве пикселей (3 класса и 18 кластеров, попадающих в область интереса).

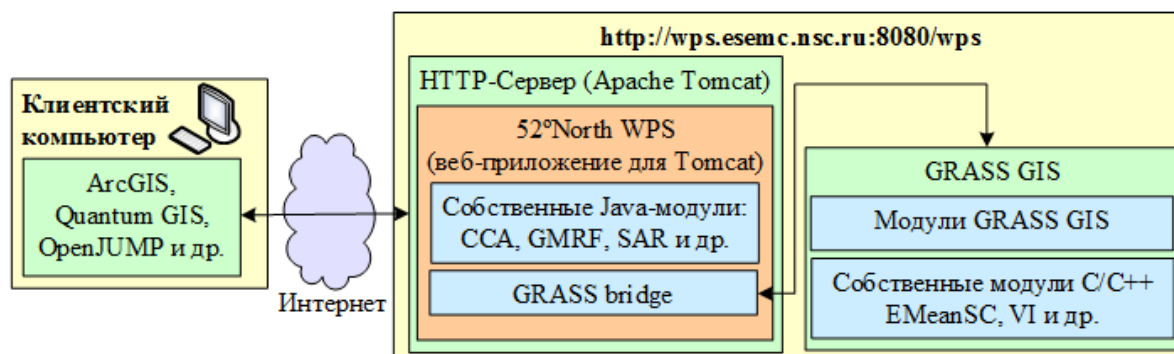


Рис. 4. Структурная схема системы веб-сервисов.

ную в Институте вычислительных технологий СО РАН [11,12,13]. Система обеспечивает простое и оперативное внедрение новых алгоритмов обработки и удобный доступ к ним.

Структурная схема системы сервисов представлена на рис. 4. Ядром системы является WPS-сервер, который создан в рамках проекта 52°North [14] и представляет собой веб-приложение, работающее под управлением контейнера сервлетов Apache Tomcat. Он осуществляет интерпретацию входных и выходных данных согласно спецификации протокола WPS [15] и может включать неограниченное число WPS-процессов. Кроме того, WPS-сервер обеспечивает доступ к открытой геоинформационной системе GRASS GIS. Для реализации системы использованы программные продукты с открытым исходным кодом, распространяемые по лицензии GPL (GNU general public license).

Реализация алгоритмов в виде набора стандартизованных веб-сервисов (WPS-процессов) позволяет использовать предложенную технологию обработки для решения практических задач на стороне пользователя с использованием как свободно распространяемых ГИС-пакетов (QGIS, uDig, OpenJUMP и др.), так и коммерческой геоинформационной системы ArcGIS.

В настоящий момент в систему внедрен набор эффективных непараметрических алгоритмов, созданных в рамках различных проектов и грантов.

#### 4 Решение практических задач

Разработанная технология успешно использована для решения двух задач, связанных с изучением природных процессов и явлений:

1. обнаружение и картирование повреждений кедровых древостоев по данным, полученным со спутника Pleiades;
2. выявление фундаментальных закономерностей формирования растительного покрова степного биома по данным WorldView-2.

Предложенная схема обработки позволила с высокой степенью достоверности выполнить анализ пространственно-временной динамики усыхания кедровых древостоев в горах Кузнецкого Алатау. Выполнены обнаружение и классификация кедрочай по снимку высокого пространственного разрешения, полученному со спутника Pleiades, а также данным Landsat и (Terra+Aqua)/MODIS. Построенные картосхемы позволили выявить очаги и оценить динамику повреждений кедровых древостоев за 2006–2012 годы [10,16].

По снимкам высокого разрешения, полученным со спутника WorldView-2, построены крупномасштабные картосхемы, которые послужили основой для моделирования пространственной организации степной растительности и выявления важных закономерностей формирования растительного покрова горно-степного пояса Южной Сибири. Это позволило провести достоверные наблюдения за мельчайшими изменениями границ сообществ, что имеет большое значение для охраны уникальных реликтовых сообществ и произрастающих в них редких и уникальных видов растений, особенно при исследовании территорий, подверженных сильному антропогенному воздействию [17,18].

## 5 Заключение

Предложены логическая схема единообразного представления разнородных пространственных данных и технология сегментации спутниковых изображений высокого пространственного разрешения, позволяющая учесть всю имеющуюся информацию (спектральные и пространственные признаки, данные полевых наблюдений, тематические карты, базы данных, априорные знания и т.п.). Для этого информация представляется в едином формате (в виде изображений), удобном для дальнейшей совместной обработки. Технология реализована в виде набора стандартизованных веб-сервисов и включена в сервис-ориентированную систему Института вычислительных технологий СО РАН. Разработанная технология используется для решения двух практических задач, связанных с изучением природных процессов и явлений.

Работа выполнена при поддержке Российского фонда фундаментальных исследований (гранты № 14-07-31320-мол\_а, № 13-07-12202-офи\_м) и Российского научного фонда (грант № 14-14-00453).

## Список литературы

1. Дворкин Б.А., Дудкин С.А. Новейшие и перспективные спутники дистанционного зондирования Земли // Геоматика. – 2013. – № 2. – С. 16-21.
2. Dey V., Zhang Y., Zhong M. A review on image segmentation techniques with remote sensing perspective // ISPRS TC VII Symp. – 100 Years ISPRS. – Vienna, Austria, July 5-7 2010. – IAPRS. – Vol. XXXVIII, pt 7A. – P. 31-42.
3. Wang A., Wang S., Lucieer A. Segmentation of multispectral high-resolution imagery based on integrated feature distribution // Intern. J. Remote Sens. – 2010. – Vol. 31, N 6. – P. 1471-1483.
4. Пестунов И.А., Синявский Ю.Н. Алгоритмы кластеризации в задачах сегментации спутниковых изображений // Вестник КемГУ. – Кемерово, 2012. – Т. 52, № 4/2. – С. 110-125.
5. Пестунов И.А., Бериков В.Б., Куликова Е.А., Рылов С.А. Ансамблевый алгоритм кластеризации больших массивов данных // Автометрия. – Новосибирск, 2011. – Т. 47, № 3. – С. 49-58.
6. Черепанов А.С. Вегетационные индексы // Геоматика. – 2011. – № 2. – С. 98-102.
7. Борзов С.М., Пестунов И.А. Сегментация спутниковых изображений высокого разрешения на основе спектральных, текстурных и структурных признаков для анализа ЧС природного и техногенного характера // IV Всерос. конф. «Безопасность и живучесть технических систем»: сборник трудов. – Красноярск, 09-13 октября 2012 г. – Красноярск, 2012. – С. 209-212.
8. Пестунов И.А., Рылов С.А. Метод выделения теней на мультиспектральных спутниковых изображениях высокого пространственного разрешения // Дистанционное зондирование Земли из космоса: алгоритмы, технологии, данные: материалы молодежной школы-семинара. – Барнаул: Алтайский гос. ун-т, 2013. – С. 60-73.
9. Пестунов И.А., Рылов С.А. Алгоритмы спектрально-текстурной сегментации спутниковых изображений высокого пространственного разрешения // Вестник КемГУ. – Кемерово, 2012. – № 4/2 (52). – С. 104-110.
10. Пестунов И.А., Мельников П.В., Дубровская О.А., Синявский Ю.Н., Харук В.И. Обнаружение и картирование повреждений кедровых древостоев по изображениям со спутника Pleiades // Интерэкспо ГЕО-Сибирь-2014. X Междунар. науч. конгр., 8-18 апреля 2014 г., Новосибирск: Междунар. науч. конф. «Экономическое развитие Сибири и Дальнего Востока. Экономика природопользования, землеустройство, лесоустройство, управление недвижимостью»: сб. материалов в 2 т. – Т. 2. – Новосибирск: СГГА, 2014. – С. 359-366.

11. Пестунов И.А., Рылов С.А., Мельников П.В., Синявский Ю.Н. Технология и программный инструментарий для сегментации спутниковых изображений высокого пространственного разрешения // Интерэкспо ГЕО-Сибирь-2013. IX Междунар. научн. конгр.: сборник материалов. – Новосибирск, Россия, 15-26 апреля 2013 г. – Новосибирск: СГГА, 2013. – Т. 1. – С. 202-208.
12. Добротворский Д.И., Куликова Е.А., Пестунов И.А., Синявский Ю.Н. Веб-сервисы для непараметрической классификации спутниковых данных // Интерэкспо ГЕО-Сибирь-2010. VI Междунар. научн. конгр.: сборник материалов. – Новосибирск, Россия, 27-29 апреля 2010 г. – Новосибирск: СГГА, 2010. – Т. 1, ч. 2. – С. 171-175.
13. Жижимов О.Л., Молородов Ю.И., Пестунов И.А., Смирнов В.В., Федотов А.М. Интеграция разнородных данных в задачах исследования природных экосистем // Вестник НГУ. Серия: Информационные технологии. – Новосибирск, 2011. – Т. 9, № 1. – С. 67-74.
14. 52°North project homepage [Электрон. ресурс]. – 2015 – URL: <http://52north.org/communities/geoprocessing/wps/index.html> (дата обращения: 10.04.2015).
15. Web processing service interface standard [Электрон. ресурс]. – 2015. – URL: <http://www.opengeospatial.org/standards/wps> (дата обращения: 12.05.2015).
16. Харук В.И., Пестунов И.А., Дубровская О.А., Мельников П.В., Рылов С.А. Обнаружение и классификация усыхающих кедровых древостоев по спутниковым данным высокого пространственного разрешения // ENVIROMIS-2014: матер. Междунар. конф. и школы молодых ученых по измерениям, моделированию и информационным системам для изучения окружающей среды, 28 июня - 5 июля 2014 г. – Томск: Изд. Томского ЦНТИ, 2014. – С. 178-180.
17. Ermakov N., Larionov A., Polyakova M., Pestunov I., Didukh Y. Diversity and spatial structure of cryophytic steppes of the Minusinskaya intermountain basin in Southern Siberia // Tuexenia. – Göttingen, 2014. – Vol. 34. – P. 431-446.
18. Ермаков Н.Б., Пестунов И.А., Полякова М.А., Дубровская О.А., Рылов С.А., Синявский Ю.Н. Крупномасштабное картографирование структуры степной растительности и выявление сообществ с редкими и уникальными видами растений на территории Южной Сибири с использованием снимков высокого разрешения // Региональные проблемы дистанционного зондирования Земли: матер. междунар. науч. конф. / науч. ред. Е.А. Ваганов; отв. за вып. А.В. Машукова. – Красноярск: Сиб. федер. ун-т, 2014. – С. 224-229.

# Интеграция Географических Метаданных в Современные Системы Организации Цифровых Репозиториях

Д.М. Скачков<sup>1</sup> and О.Л. Жижимов

<sup>1</sup>Институт вычислительных технологий СО РАН, Новосибирск, Россия  
daniil.skachkov@gmail.com, zhizhim@mail.ru

**Аннотация.** В работе рассматривается такая широко известная система для организации цифровых репозиториях, как DSpace. Данная система используется в рамках проектов «Цифровой репозиторий ИВТ СО РАН», «Открытая краеведческая цифровая библиотека Новосибирска» и др. В системах, построенных на базе DSpace, содержится большое количество документов, так или иначе связанных с географической информацией. В то же время, в базовой поставке DSpace отсутствует функциональность, связанная с географическим аспектом информации. В докладе описывается два варианта интеграции географических метаданных в систему DSpace: посредством непосредственного задания координат и посредством ссылок на записи специализированного тезауруса географических названий.

**Ключевые слова:** географические метаданные, географический поиск, DSpace.

## 1 Введение

Системы для организации цифровых репозиториях обладают широкими возможностями по хранению и обработке различной информации как об исключительно цифровых объектах, так и об объектах реального мира. Такие системы включают в себя подсистемы для обработки и визуализации цифровых материалов в различных представлениях, системы индексации и поиска, средства для классификации и каталогизации материалов, инструменты для управления правами доступа и прочие сервисы. Одной из наиболее широко используемых систем для построения цифровых репозиториях является система DSpace [1]. В данной работе будет рассматриваться система DSpace 5 [2], которая является последней.

## 2 Географический поиск

Система DSpace ориентирована на использование текстовых поисковых запросов для описания информационной потребности пользователя. Однако такое представление подходит не для всех типов задач. Одним из примеров такого рода задач является географический поиск. Изначально системы для организации цифровых репозиториях не были рассчитаны на использование географической информации. Однако материалы, хранящиеся в таких системах содержат упоминания географических объектов в том или ином виде.

Приведем пример нескольких записей из Цифрового репозитория ИВТ СО РАН, содержащих упоминания географических объектов.

1. В центре внимания - Байкал
2. Международная конференция "Ультрамафит-мафитовые комплексы складчатых областей докембрия" на Байкале п. Энхалук, 6-9 сент., 2006
3. Микробиологическая оценка состояния бассейна рек Тугнуй - Сухара
4. Минерально-вещественный состав пылеаэрозолей на территории г. Томска



Использование текстовых поисковых запросов для работы с географической информацией не так эффективно, как использование картографических сервисов, таких как Google Maps [3]. Например, если нам нужно будет найти документы, в которых идёт речь о Новосибирской области как географическом регионе, и мы будем использовать текстовый запрос «Новосибирская область», то результаты поиска будут неполными, ведь в Новосибирской области находится множество географических объектов, у которых есть одно или несколько названий, причем эти названия еще и изменялись с течением времени. Поэтому, если мы хотим найти документы, так или иначе, относящиеся к Новосибирской области, то нам понадобится составить огромных размеров поисковый запрос, содержащий названия всех географических объектов, находящихся на территории Новосибирской области, да еще и учитывая исторические изменения этих названий и особенности русского языка. Дела обстоят еще сложнее, если интересующий географический регион не является цельным в административном смысле, например, Байкальская природная территория, которая находится частично в Иркутской области и Республике Бурятия. Если же мы будем использовать картографический сервис, то получим координаты географического региона в виде набора точек и в область поиска будут включены все географические объекты, находящиеся в выбранном регионе. Но получить контур интересующего географического региона недостаточно, ведь целевые информационные системы не являются географическими информационными системами, и никакой поддержки обработки географических координат в них нет.

Есть два возможных варианта решения для задачи географического поиска в такого рода библиотечных системах:

1. непосредственная индексация записей информационных систем географическими координатами;
2. использование тезауруса географических названий [4].

Далее будет рассмотрено применение этих двух подходов к системе DSpace 5.

### 3 Интеграция посредством индексации записей координатами

Для интеграции географического поиска в систему DSpace были выполнены следующие задачи:

1. добавлены описания метаданных в файл конфигурации в директории "config/registries"
2. добавлено поле метаданных в файл конфигурации "input-forms.xml"
3. сконфигурированы поисковые индексы в dspace.cfg
4. добавлены классы для типов данных, относящихся к географическим метаданным
5. модифицированы интерфейсы добавления, просмотра и редактирования метаданных (файлы edit-metadata.jsp, edit-item-form.jsp, файлы из директории search)

#### 3.1 Добавление полей метаданных

Описания добавляемых полей метаданных должны быть помещены в xml файл в директории dspace/config/registries. Имя файла может быть произвольным, например **sbras-geo-types.xml** (данное название будет говорить о том, что в файле собраны типы, относящиеся к схеме sbras-geo). В файле описывается набор полей метаданных с помощью коллекции записей следующего вида:

```

<dc-type>
  <schema>sbras-geo</schema>
  <element>content</element>
  <qualifier>box</qualifier>
  <scope_note>Географические граничные координаты</scope_note>
</dc-type>

```

Здесь указано название специфической схемы - **sbras-geo** - все географические метаданные будут относиться к данной схеме. В поле **element** указано имя элемента, к которому относится данный элемент метаданных. В поле **qualifier** указан квалификатор типа данных. В поле **scope-note** - описание поля.

Чтобы поля метаданных были доступны в формах ввода данных, они должны быть добавлены в файл **input-forms.xml**. Описание поля в данном файле выглядит следующим образом:

```

<field>
  <dc-schema>sbras-geo</dc-schema>
  <dc-element>publicationPlace</dc-element>
  <dc-qualifier>box</dc-qualifier>
  <repeatable>>true</repeatable>
  <label>Место публикации</label>
  <type-bind>Article</type-bind>
  <input-type>location</input-type>
  <hint>Выберите область на карте</hint>
  <required></required>
</field>

```

Здесь указана схема, к которой относится запись. В нашем случае, схема называется **sbras-geo**. Элемент схемы, к которому будет относиться поле метаданных - **publicationPlace**, и квалификатор типа данных - **box**. Далее приводится флаг **repeatable**, указывающий на возможность добавить несколько экземпляров данного поля метаданных к записи в DSpace. Поле **label** указывает заголовок поля в форме ввода. В поле **type-bind** указывается список типов материалов, в которых будет доступно данное поле метаданных, в данном случае "Article" указано в качестве примера, и в данном поле должны быть перечислены все типы материалов, так как любой материал в системе может иметь связь с географическими объектами.

Поле **input-type** указывает тип поля ввода, который будет использоваться для ввода и редактирования поля метаданных. Так как в базовой поставке DSpace не была реализована работа с географическими метаданными ни в каком виде, в формы ввода была добавлена поддержка нового типа поля ввода, который был назван **location**. Данный тип поля ввода позволяет отобразить географическую карту и выбрать на карте замкнутый географический регион. Более подробно поле ввода географического региона будет рассмотрено в подразделе **Модификация интерфейсов** ниже.

В поле **hint** указана подсказка, которая будет отображаться в случае, если поле не заполнено. Флаг **required** указывает, является ли поле обязательным.

Такие записи о метаданных должны быть добавлены для всех желаемых элементов схемы (dc-element), таких как creationPlace, publicationPlace, referencePlace и др.

### 3.2 Конфигурация поискового индекса

Непосредственно механизм поиска по географическим координатам в системе DSpace может быть реализован следующими способами:

1. с помощью непосредственной работы с базой данных PostgreSQL, которая используется в качестве хранилища метаданных в DSpace
2. с помощью платформы поиска Apache Solr [5]

Разработчиками системы рекомендовано использовать Solr для реализации поиска [6], так как он не привязан непосредственно к выбранному хранилищу данных (а DSpace может использовать не только PostgreSQL в качестве хранилища). Поэтому далее будет рассмотрена реализация поиска с применением Solr.

Конфигурация поисковых индексов описывается в файле конфигурации сервиса Solr `solr/search/conf/schema.xml`. Данный файл содержит описания используемых типов данных и полей, для которых будет производиться индексация. Правила описания схемы приводятся в [7].

Для индексации географических метаданных вначале должен быть описан соответствующий тип данных следующим образом:

```
<fieldType name="location_rpt"
  class="solr.SpatialRecursivePrefixTreeFieldType"
  spatialContextFactory="com.spatial4j.core.
    context.jts.JtsSpatialContextFactory"
  units="degrees"/>
```

Затем описывается конкретное поле метаданных, с использованием указанного типа. Пример описания поля:

```
<field name="spatial-search"
  type="location_rpt" indexed="true"
  stored="true" />
```

После данных модификаций схемы поле **spatial-search** станет доступным для поиска.

### 3.3 Модификация интерфейсов

Базовая поставка системы DSpace не поддерживает работу с географическими данными. Поэтому для интеграции работы с географическими метаданными необходимо добавить не только соответствующие поля метаданных, но и доработать пользовательский интерфейс форм ввода данных, модификации данных, поиска. В данной работе рассматривается модификацию форм ввода данных, модификация остальных форм выполняется аналогично.

В список доступных метаданных были добавлены поля с типом данных **location**. Поддержка этого типа должна быть добавлена и в пользовательские интерфейсы. Web-интерфейсы системы DSpace построены на базе технологии Java Server Pages [8]. Формы ввода данных представлена файлом `dspace/src/main/webapp/submit/edit-metadata.jsp`. В нем происходит формирование интерфейсов ввода данных на различных этапах заполнения новой записи. В данный файл в методе проверки и генерации полей ввода был добавлен вызов функции, генерирующей интерфейс для ввода географической информации. Поле ввода географических данных содержит кнопку "Карта" для открытия диалога ввода географического региона, кнопку "Очистить" для очистки значения поля и текстовое

Выберите область на карте

**Место создания**

Выберите область на карте

**Место публикации**

Выберите область на карте

**Другие географические ссылки**

Рис. 1. Поле ввода географических данных

поле с координатами (Рис 1). Координаты в текстовом поле не редактируются, а предназначены только для просмотра. Географический регион выбирается в отдельном диалоге (Рис. 2). Диалог построен с помощью библиотеки для работы с географическими картами OpenLayers [9].

В заполненном виде поле географической информации выглядит как на Рис. 3.

В результате интеграции в системе DSpace была реализована функциональность географического поиска. Однако, такой поиск работает только по записям, в которых заполнены соответствующие поля метаданных.

#### 4 Интеграция с помощью ссылок на записи тезауруса географических названий

Тезаурус географических названий является внешним справочником для системы DSpace, доступ к нему осуществляется по протоколу Z39.50 [10,11]. Суть интеграции с помощью ссылок на записи тезауруса состоит в том, чтобы связать запись из DSpace с одним или несколькими идентификаторами географических объектов из тезауруса. А поиск будет организован с помощью внешней метапоисковой машины, как описано в [12].

Технически привязка записей из DSpace к географическим объектам осуществляется посредством так называемых "Controlled Vocabularies"[13].

Соответствующие настройки должны быть описаны в `dspace/config/dspace.cfg` следующим образом:

```
webui.controlledvocabulary.rgeothes.type = sru
webui.controlledvocabulary.rgeothes.db = rgeothes
webui.controlledvocabulary.rgeothes.url = http://z3950.ict.nsc.ru:210
webui.controlledvocabulary.rgeothes.root = .
webui.controlledvocabulary.rgeothes.title = Retrospective geo-thesaurus
```

В данном примере используется тип доступа SRU. Указаны название базы, адрес доступа и описание словаря, которое будет отображено в диалоге отображения записей тезауруса. Префикс параметров конфигурации `webui.controlledvocabulary.rgeothes` указывает на

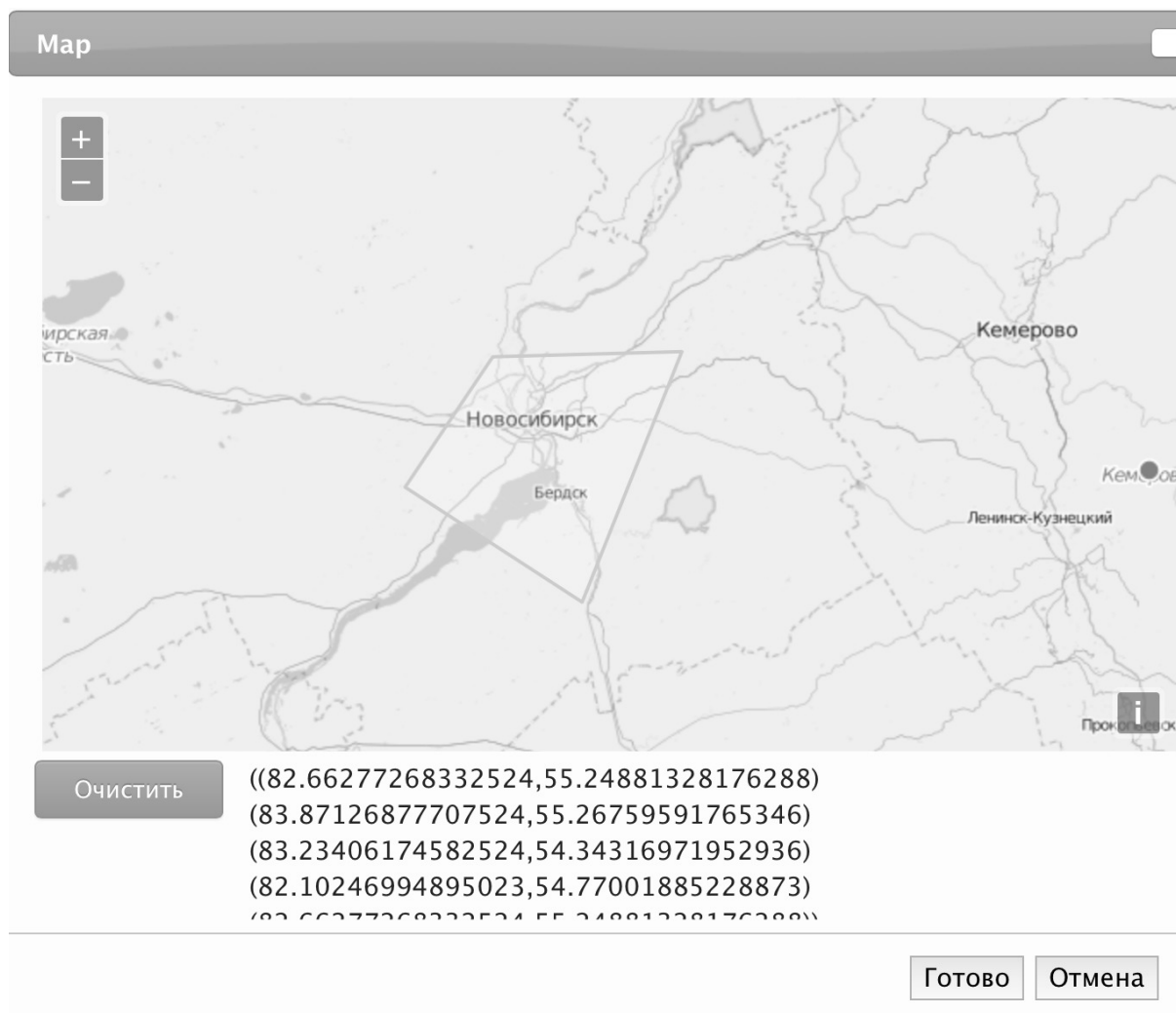


Рис. 2. Диалог выбора географического региона

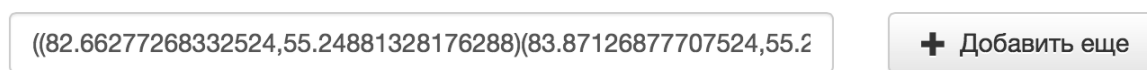


Рис. 3. Заполненное поле географических данных

то, что данный словарь будет использоваться для полей с типом данных **rgeothses**. Соответствующая запись должна быть добавлена в соответствующий xml файл в директории `dspace/config/registries` аналогично примеру, приведенному в разделе 3.

Также должна бть добавлена запись в файл **input-forms.xml** следующего вида:

```
<field>
  <dc-schema>sbras-geo</dc-schema>
  <dc-element>publicationPlace</dc-element>
  <dc-qualifier>rgeothses</dc-qualifier>
  <repeatable>true</repeatable>
```

```

<label>Место публикации</label>
<type-bind>Article</type-bind>
<input-type>twobox</input-type>
<hint>Выберите область на карте</hint>
<required></required>
<vocabulary closed="false">rgeothes</vocabulary>
</field>

```

В данном примере в поле **vocabulary** указано название словаря "rgeothes". Назначение остальных полей совпадает с описанием, приведенным в разделе 3.

## 5 Заключение

В работе были приведены два варианта интеграции географических метаданных в современную систему организации цифровых репозиториях DSpace. В итоге оказалось, что для интеграции с помощью ссылок на описания географических объектов из тезауруса требуется меньше изменений в системе, чем при интеграции с помощью непосредственной привязки географических координат. Более того, при использовании метода привязки посредством ссылок на записи тезауруса географических названий возможно применение подходов автоматической привязки, описанных в [14]. Таким образом этот способ видится авторам более приоритетным и заслуживающим дальнейшего развития.

## Список литературы

1. DSpace is a turnkey institutional repository application [Электрон. ресурс]. – 2015. – URL: <http://www.dspace.org> (дата обращения: 15.05.2015).
2. Release Notes - DSpace 5.x Documentation - DuraSpace Wiki [Электрон. ресурс]. – 2015. – URL: <https://wiki.duraspace.org/display/DSD0C5x/Release+Notes> (дата обращения: 15.05.2015).
3. Google Maps Web APIs — Google Developers [Электрон. ресурс]. – 2015. – URL: <https://developers.google.com/maps/web/?hl=ru> (дата обращения: 15.05.2015).
4. Скачков Д.М., Жижимов О.Л. Технология географического поиска информации в «негеографических» информационных системах // Сборники Президентской библиотеки им. Б. Н. Ельцина / Вып. 4: Научные и организационно-технологические основы интеграции цифровых информационных ресурсов = Scientific, organizational and technological fundamentals of digital information resources integration : сборник научных трудов. - 2013. - 378, [1] с. : ил. - (Серия «Электронная библиотека» / науч. ред. Е. Д. Жабко). - 2013. - С.74-101. - ISBN 978-5-905273-39-1.
5. Apache Solr [Электрон. ресурс]. – 2015. – URL: <http://lucene.apache.org/solr/> (дата обращения: 15.05.2015).
6. Solr - DSpace - DuraSpace Wiki [Электрон. ресурс]. – 2015. – URL: <https://wiki.duraspace.org/display/DSPACE/Solr> (дата обращения: 15.05.2015).
7. SchemaXml - Solr Wiki [Электрон. ресурс]. – 2015. – URL: <http://wiki.apache.org/solr/SchemaXml> (дата обращения: 15.05.2015).
8. Bergsten H., JavaServer Pages. – O'Reilly, 2001. – 684 с. - ISBN 1-56592-746-X
9. OpenLayers 3 - Welcome [Электрон. ресурс]. – 2015. – URL: <http://openlayers.org> (дата обращения: 15.05.2015).
10. Скачков Д.М., Жижимов О.Л. Об интеграции географических метаданных посредством ретроспективного тезауруса // Информатика и ее применения. - 2012. - Т.6. - № 3. - С.42-50. - ISSN 1992-2264.
11. Жижимов О.Л., Скачков Д.М. О профиле доступа к данным тезауруса для ретроспективного геокодирования и географического поиска в электронных библиотеках // XVIII Международная конференция «Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса» - Крым-2011 (Судак, Украина, 04.06 - 12.06.2011): Материалы конференции. - М.: ГПНТБ России, 2011. - ISBN 978-5-85638-150-3. - Гос. регистр. № 0321100651.

12. Скачков Д.М., Жижимов О.Л. Метапоисковая машина для осуществления географическо-временного поиска в документо-ориентированных информационных системах // Вестник Новосибирского государственного университета. Серия: Информационные технологии. - 2014. - Т.12. - № 3. - С.124-131. - ISSN 1818-7900.
13. Use controlled vocabularies (JSP) - DSpace - DuraSpace Wiki [Электрон. ресурс]. – 2015. – URL: <https://wiki.duraspace.org/pages/viewpage.action?pageId=19006518> (дата обращения: 15.05.2015).
14. Барахнин В.Б., Жижимов О.Л., Куперштох А.А., Скачков Д.М., Федотов А.М. Алгоритм извлечения из текстовых документов географических названий, отражающих содержание // Вестник Новосибирского государственного университета. Серия: Информационные технологии. - 2012. - Т.10. - № 1. - С.109-120. - ISSN 1818-7900.

# Системный Подход к Конструированию Интерфейсов Приложений \*

И.Н. Скопин

Институт вычислительной математики и математической геофизики СО РАН, Новосибирск, Россия  
Новосибирский государственный университет, Новосибирск, Россия  
iskopin@gmail.com

**Аннотация.** Изучаются проблемы повышения юзабилити программных систем, связанные с разработкой интерфейсов. Отмечается, что основные интерфейсные ошибки приложений обычно связаны с недостаточной проработанностью разработчиками пользовательской деятельности, для автоматизации которой предназначена предлагаемая программная система. Чтобы преодолеть этот недостаток необходимо до разработки программы проанализировать операционные маршруты пользователей, для автоматизации которых она предназначена. В результате такого анализа до и во время начальной стадии проекта определяются интерфейсные управляющие воздействия, которые направляют систему в желаемое состояние, — так называемый *абстрактный интерфейс*. Абстрактное представление интерфейса отображается в реализации в наиболее подходящие для разных пользователей элементы интерфейса, т.е. конструируется *конкретный интерфейс*. Эта двухступенчатая разработка интерфейса требует рассмотрения автоматизируемой приложением деятельности не изолированно, а в системе других деятельностей, в которых принимает участие пользователь.

**Ключевые слова:** юзабилити, абстрактный интерфейс, конкретный интерфейс, деятельность, полнота поддержки деятельности, интерфейсные стандарты.

## 1 Введение

Интерфейс программной системы — важная составляющая качества, как основы оценки полезности приложения для пользователя. В то же время, при разработке интерфейсу должно уделяться не всегда. В результате в принципе полезные программы часто оказываются неудобными или даже не востребованными пользователями. Интерфейсные ошибки сплошь и рядом «украшают» популярные программы. Положение усугубляется тем, что реальное использование таких программ создает предпосылки формирования стихийных стандартов, закрепляющих и тиражирующих интерфейсные ошибки.

Сведения о том, какая доля трудозатрат при создании системы приходится на интерфейс, довольно разноречивы. Есть мнение, что в среднем она составляет более половины времени реализации проекта. Проверить это трудно, в частности потому, что продуманный интерфейс обычно требует большего времени, чем непродуманный. Разработка интерфейса, основанного на принципах, которые не удовлетворяют пользователя, может усложнять реализацию. Для разных приложений роль интерфейсных характеристик различна, а потому различна трудоемкость интерфейсной части проекта.

В одной из немногих книг [1], посвященных интерфейсам, приводятся рекомендации и предостережения, обоснованные хорошим анализом. Однако автор обсуждает действия пользователя, не связывая интерфейс с функциональностью и архитектурой, что ограничительно. Мы предлагаем восполнить этот пробел, рассматривая интерфейсы в двух аспектах. Во-первых, это абстрактное управление поведением приложения, а во-вторых —

\* Исследования, представленные в статье, поддержаны грантом РФ РНФ № 14 11 00485 "Высокопроизводительные методы и технологии моделирования электрофизических процессов и устройств".



отображение абстрактного управления в конкретные интерфейсные формы. Соответственно, вводятся понятия *абстрактного* и *конкретного интерфейсов* и их полноты. Полнота абстрактного управления — это соответствие функциональности всем элементам деятельности, автоматизируемой приложением, а полнота конкретного интерфейса — отражение в его элементах всех аспектов абстрактного поведения в формах, отвечающих пользовательской потребности автоматизируемой деятельности.

Ключевым положением предлагаемого подхода является анализ всех видов деятельности пользователя, на которые влияет включение приложения в качестве средства или инструмента. Цель анализа — определить место разрабатываемого приложения в контексте работы пользователя и, в частности, обосновать необходимые элементы интерфейса и построить типовые сценарии взаимодействия с приложением. Эти сценарии рассматриваются как основа отображения абстрактного управления в конкретный интерфейс. В результате разработчик имеет возможность обосновать использование активных, пассивных и декоративных элементов интерфейса для тех или иных целей.

## 2 Интерфейс как составляющая качества приложения

Обычные критерии качества программного изделия связываются, в первую очередь, с его функциональностью и производительностью. Что же касается интерфейса, то, признавая существенным его вклад в полезность приложений, равный функциональности и производительности, обычно считают, что качество, привносимое интерфейсом, гарантируется использованием при его реализации сложившихся стандартов и хорошо себя зарекомендовавших библиотек поддержки. Причины тому следующие. Если реализацию функций системы можно спланировать и специфицировать на предпроектной стадии и фиксировать результаты этой работы в виде вполне проверяемых требований, то, за редким исключением, запланировать проверяемые интерфейсные характеристики не удастся. Большинство интерфейсных решений весьма неоднозначны, во многих случаях оценка качества интерфейса субъективна, о приемлемости или непригодности интерфейса часто удается говорить лишь в целом, без обоснованного указания на четкие критерии.

Чаще всего программные проекты организуются лишь для предоставления новых функций и/или более производительных версий существующих систем. Предпроектные требования к интерфейсам сводятся к утверждениям об их соответствии хорошо зарекомендовавшим себя образцам, играющим роль стандарта. Интерфейсные проекты сводятся к предложению стандартных решений и их библиотечных реализаций, которые отражают довольно произвольное мнение о стандартах элементов интерфейса, но не специфику приложений, для которых применение библиотек оправдано.

Характерными примерами поддержки конструирования интерфейсов могут служить библиотеки, использованные для разработки многочисленных интегрированных сред разработки, так называемых IDE-систем (Integrated development environment см.[2]), предназначенных для разработчиков программного обеспечения. Свое начало все они берут от Turbo Vision [3] и послушно наследуют его средства стандартизованного диалога. Однако в такой среде все приложения становятся похожими друг на друга, их специфика теряется. Безусловно, стандартизация элементов интерфейса нужна, но ее роль должна ограничиваться рамками объективно обусловленных эргономических и когнитивных критериев, учитывать складывающиеся у пользователей привычки и предпочтения. Так же нет сомнения в том, что должны быть выработаны унифицированные правила компоновки элементов и технологически выверенные приемы работы пользователя в интерфейсной среде, подобно тому,

как стандартизованы расположение клавиш на пишущей машинке и правила печатания "вслепую".

Унификация — лишь необходимая часть требования максимально полной поддержки конкретной пользовательской деятельности, для которой предназначено приложение. Потребительская ценность программного продукта, его полезность с точки зрения практического использования, которые сегодня принято называть *юзабилити*, предполагают, что поддержка обеспечивается комплексно по следующим трем направлениям:

- *функциональность* — способность программы выполнять требуемые от нее функции (набор задач из области приложения, решению которых способствует данное программное изделие, адаптивность для решения новых задач, уровень автоматизации деятельности пользователя при применении программы и другие подобные характеристики);
- *эффективность* — скоростные характеристики и требования памяти, связанные с выполнением программы (они в конечном итоге определяют границы возможного практического использования программы на данном оборудовании);
- *интерфейс* — средства управления работой программы, благодаря которым осуществляется включение ее в человеко-машинную систему.

Эти составляющие качества программного обеспечения не являются независимыми. К примеру, когда программа мало пригодна для практических целей из-за низкой эффективности (о непригодности вообще речь не идет, если выполнены функциональные требования), правомерна постановка задачи разработки новой программы, для которой уровень эффективности фиксируется как функциональное требование. Аналогичная ситуация возможна с мало приемлемым интерфейсом.

Функциональная составляющая доминирует в оценке приложения: если не выполнены функциональные требования, то говорить об эффективности и интерфейсе не приходится. Функциональные требования — это регламент, фиксирующий *модель вычислений программы*, а саму функциональность естественно отождествлять с такой моделью. В свою очередь, требования эффективности — это ограничения на допустимость реализации, рассматриваемой как *отображение модели вычислений на реальное оборудование*. Наконец, требования к интерфейсу есть не что иное, как спецификации языка управления поведением модели.

Важно отметить, что под языком здесь понимается произвольная знаковая система для передачи информации, включающая не только символьные последовательности, но и визуальные образы (пиктограммы и иные изображаемые знаки), указание картин и их фрагментов с помощью программно-аппаратных средств, задание расцветки изображений, использование меню, транспарантов, табло, обрамлений и др., а также звуковые сигналы и, возможно, речевые средства ввода/вывода. Данный перечень относится к уровню человеко-машинного интерфейса программной системы. Когда речь идет о программах, обеспечивающих доступ к внешнему оборудованию, например, о драйверах, роль языка играют сигналы-сообщения от и для этого оборудования, а также система прерываний. Рассматривая межмодульные интерфейсы, можно заметить, что для окружения, использующего модуль, предоставляемые им средства являются расширением языка реализации системы. Таким образом, всякий раз, когда речь идет об интерфейсе, нужно иметь в виду тот или иной язык, предоставляемый на уровень использования, и как следствие, если есть возможность выбирать язык, критерием выбора его средств должны быть требования со стороны пользователя.

Выделение модели вычислений, ее отображения как реализации в программе и интерфейсного языка управления поведением модели имеет ряд принципиальных следствий.

- (1) При проектировании программной системы эти три составляющие разграничивают требования к программному изделию, что создает предпосылки для независимого построения модели, повышения эффективности реализации и разработки интерфейса.

В реальной практике независимость указанных составляющих относительна, поскольку связи между ними существуют на уровне взаимовлияния. Так, можно запланировать вычисления, которые не реализуемы при заданных требованиях к эффективности. Управляемость поведением модели также имеет свои пределы: нельзя говорить об интерфейсе, если не фиксировать управляемые объекты, их возможное поведение и т. п., нельзя строить интерфейс без учета приемлемых для функционирования системы характеристик эффективности. Тем не менее при заданной функциональности отображение модели на реальную операционную среду можно рассматривать как выбор реализации, удовлетворяющей требованиям эффективности, а проектирование интерфейса — как выбор его языковых средств, адекватных использованию программы. Эти две задачи не только можно, но и целесообразно ставить как независимые, в частности по той причине, что для них различны критерии отбора подходящих решений.

- (2) Явное разграничение трех составляющих качества позволяет ставить задачу разработки серий программ с общей моделью вычислений, предназначенных для различного применения.

Примерами серий такого рода являются системы, учитывающие особенности перерабатываемой информации для повышения эффективности вычислений. Пользователю предоставляется право выбора метода решения задачи, а каждый из методов, если их особенности не проявляются в управлении, — это вариант реализации одной и той же вычислительной модели (если указанное условие не выполняется, т.е. выбор метода является частью управления, то имеет место модель вычислений, включающая с себя варианты методы, а не серия программ-вариантов). В ряде случаев возможен автоматический выбор метода на основе анализа входных данных, и тогда сама серия для пользователя выступает как единая система.

Другой пример серийности на основе общей модели вычислений — различные программы, использование которых связывается с применением оборудования с несопоставимыми интерфейсными возможностями. Выбор варианта программы таких серий целесообразно возлагать на установку системы.

Серийность, учитывающая особенности перерабатываемых данных, отражает неоднозначность критериев эффективности программ. Они не сводятся лишь к требованию оптимизации — минимизации для конкретных вычислений времени выполнения или необходимой памяти. Выбор того или иного алгоритма должен наряду с особенностями обрабатываемых данных учитывать архитектуру вычислительной системы, на которой выполняется программа. Если ставится задача соблюдения оптимальности для разных архитектур, а это характерно для большинства вычислительных библиотек, то серийность часто удается обеспечить за счет компиляции, и тогда можно говорить, что пользователю предлагается единая модель вычислений. В противном случае решение о выборе варианта программы должен принимать пользователь, для которого все такие варианты естественно рассматривать как серию.

Примером серийности в обоих аспектах может служить библиотека Krylov [4], предназначенная для решения сверхбольших систем линейных уравнений. Алгоритмы решения формально одной и той же задачи специализированы для различных особенностей матриц с целью повышения производительности расчетов. Для заданных параметров задачи

в автоматизированном или ручном режиме может быть выбран тот вариант, который дает минимальное время счета.

Еще более неоднозначны интерфейсные критерии, и это обуславливает серии, связанные с вариантами управления. Разные пользователи обладают индивидуальными и групповыми психологическими особенностями восприятия, и, как следствие, для их продуктивной работы требуются различные способы взаимодействия с программной системой. На пользовательские критерии предпочтения влияют и уровень предварительной подготовки, и привычки, и то, что по мере освоения системы требуются различные уровни пояснений, помощи и т.д. В этой связи можно ставить задачу конструирования динамического интерфейса, т.е. явно и/или неявно настраиваемого на особенности пользовательского предпочтения и восприятия информации.

В качестве наглядной иллюстрации этого положения уместно упомянуть о том, как, к примеру, браузер Chrome [5] или YouTube [6] отбирает и упорядочивает информацию для предъявления пользователю с учетом истории взаимодействия с ним. По понятным причинам первая из этих систем делает это на основе более глубокого анализа.

**(3)** При сравнении программных изделий следует разграничивать сопоставление их функциональности, эффективности и интерфейсов.

На практике это разграничение зачастую игнорируется, что снижает полезность информации для пользовательского выбора подходящего изделия. Если при проектировании функциональность является заданным критерием пригодности будущего изделия, то при сравнении программ ее роль двояка.

Во-первых, функциональные характеристики определяют правомерность использования программы для автоматизации конкретной деятельности. Здесь, скорее, следует сравнивать назначение сопоставляемых программных продуктов, а не такие свойства, как применимость их для массового оперирования (эффективность) и удобство для пользователя (интерфейс).

Во-вторых, сравнение эффективности невозможно без выявления точек пересечения в назначении анализируемых программных изделий, т.е. без выделения их функциональной общности. Этот аспект сравнения достаточно ясен, и обычно эффективность проверяется на общих задачах. Тем не менее, иногда не учитывается, что функциональная общность — понятие относительное. Так, сравнение эффективности компилятора, который работает параллельно с пользовательским вводом программы, и автономного компилятора, скорее всего, будет в пользу последнего, а простой редактор программ можно сделать более эффективным, чем текстовый процессор. Разная скорость работы здесь не является критерием качества по той причине, что по внешней схожести применения программ нельзя судить об их функциональности в целом. Другая ситуация, в которой сравнительная эффективность не является критерием качества, — когда у изделия достигнута приемлемая для его модели вычислений эффективность т.е. когда с точки зрения полезности применения нет смысла повышать эффективность вычислений.

Не менее важно разграничение трех составляющих качества для сравнения интерфейсов программ. Проектирование интерфейса в большей степени искусство, чем технология, поэтому для него и критерии качества более субъективны. Но когда сравнение интерфейсов отделено от сопоставления программ по другим составляющим качества, возможна постановка общих вопросов, ответы на которые позволяют сделать оценку интерфейса объективнее.

- (4) При оценке качества программной системы безотносительно к другим аналогичным изделиям нет смысла говорить о хорошем или плохом продукте, если явно не обозначить деятельность человека, в которой изделие должно использоваться. Только с этой точки зрения правомерны утверждения о функциональной пригодности системы, об эффективной или неэффективной ее реализации, об удобствах пользовательского управления.

Уместно разграничивать два рода деятельности, для поддержки которых может использоваться программа: *автоматизируемая деятельность*, существующая и без применения программных средств, и *новая деятельность*, которая была бы невозможна без оцениваемой программы. В первом случае программная система — один из инструментов деятельности, и ее можно сравнивать с позиции заменяемых или дополняемых неавтоматизированных аналогов. Иными словами, здесь аналогами фиксирована функциональная модель, а о качестве изделия можно судить по тому, насколько эффективнее оказалась выполняемая с помощью программы деятельность по сравнению с неавтоматизированным вариантом. Понятно, что имеется в виду эффективность деятельности, а не программы, как воплощения модели. В частности, эффективность такой деятельности существенно зависит от адекватности управляющих воздействий, т. е. от предлагаемого интерфейса. Можно классифицировать возможные интерфейсные средства по степени их влияния на эффективность деятельности, *подбирать интерфейс* под конкретные условия применения программы, анализируя получаемые результаты.

Во втором случае аналогов программного инструмента не существует, и, следовательно, критерии качества меняются. Здесь оценкой программного изделия служат результаты новой деятельности, то, как она встраивается в систему других деятельностей. Таким образом, вклад эффективности и интерфейса в оценку качества программного изделия оказывается менее весом по сравнению с функциональностью. Тем не менее, если изделие претендует на достаточно долгое существование, вопросы эффективности и интерфейса также должны рассматриваться как существенные. Поэтому вместо подбора интерфейсных средств может быть рекомендована стратегия *настройки интерфейса*, осуществляемая пользователем: в ходе эксплуатации программной системы, будут выявлены желательные "точки роста" ее эффективности и дополнительные управляющие воздействия. Возможность и целесообразность такого развития программы надо планировать.

### 3 Задача конструирования интерфейса

Развитые технологии программирования предоставляют разнообразные средства реализации моделей вычислений сложных программных систем и достижения требуемой эффективности. Как правило, они исходят из внешних спецификаций конструируемого программного продукта, в которых можно достаточно точно фиксировать, что должно быть сделано, а следовательно, проверка функциональности и определение достигнутого уровня эффективности в принципе не затруднительны.

Иначе обстоит дело с интерфейсом. Когда интерфейсные критерии удастся описать явно (например, в случае межмодульного интерфейса), его разработка не требует значительных усилий и сводится к поддающимся рационализации согласованиям. Интерфейсные трудности возникают при конструировании пользовательского представления программной системы. Главная причина их в том, что разработчики зачастую слабо представляют себе, кто и как будет применять их изделие. Иногда они даже заявляют примерно следующее: "Вам предоставлена программа с определенным набором средств, а как распорядиться ими — ваша проблема". Это в корне неверное утверждение — забота о технологии применения предлагаемого (программ, оборудования, документации и др.) — существенная задача

ответственных разработчиков, именно они могут и должны точнее других ответить на вопросы о назначении изделия.

Без учета при разработке "портрета" пользователя, без продумывания регламентов и методов применения программы потребительская ценность приложения снижается. Функционально хорошие программы не всегда в состоянии удовлетворить пользователя с его запросами, обусловленными особенностями его конкретной деятельности. Иногда приемлемый для одних целей интерфейс оказывается неудобным для других из-за смещений между назначением и применением программы (соответствующие примеры удачных и неудачных интерфейсных решений можно найти в [7]).

Проблема разработки интерфейса, соответствующего потребности пользователей, в том, что заранее не получается определить истинную потребность. Для ее преодоления мы предлагаем специальный прием, который соответствует хорошо известному шаблону проектирования программных систем MVC (Model-View-Controller) [8]. В соответствии с этим шаблоном при разработке архитектуры и реализации, выделенных в ней компонент, следует представить функциональность приложения как его *модель вычислений* (Model) и *управление* (Control) состояниями модели, понимаемое как комплект воздействий, приводящих к изменению состояния. Информация о состоянии доступна для принятия решений о том, какие воздействия нужно осуществить для перевода системы из текущего состояния в желаемое. Она передается как со стороны модели, так от управления в так называемое *представление* (View) для внешнего использования. Взаимодействие пользователя с приложением осуществляется через представление. Для разных типов пользователей могут требоваться различные представления.

При разработке представлений необходимо понимать, что каждый пользователь понимает модель и управление приложения по-своему. Это связано с тем, что он встраивает приложение в свою деятельность в качестве инструмента. А поскольку эту деятельность пользователя и связанные с ней другие деятельности предсказать невозможно, не стоит надеяться на то, что пользовательское понимание модели и управления будут совпадать с тем, что предлагает программа приложения. Задача каждого из представлений, предлагаемых разным типам пользователей состоит в том, чтобы построить соответствие между моделью и управлением приложения и пользователей. Имея ввиду необходимость поддержки этого соответствия при работе приложения как основы его полезности, или, что то же, юзабилити программной системы в целом, задача конструирования интерфейса разбивается на две части:

- Разработка средств взаимодействия системы с пользователем в виде программных интерфейсов, которые обеспечивают осуществимость выполнения всей функциональности приложения. Комплект таких средств далее называется *абстрактным интерфейсом*.
- Разработка визуальных и/или иных средств предъявления информации для пользователя и задания воздействий на систему, которые реализуются как обращение к элементам абстрактного интерфейса. Комплект таких средств далее называется *конкретным интерфейсом*.

На рис. 1 представлена схема, иллюстрирующая задачу конструирования интерфейса, согласованного с шаблоном проектирования MVC. Шаблон естественным образом задает «водораздел» между абстрактным (относящимся к вычислениям) и конкретным (связанным с деятельностью пользователя) уровнями приложения, к которым относятся две части интерфейсной задачи. На схеме они разделены жирной линией. Модель и управление с точки зрения пользователя обозначены заштрихованными блоками U-Model и U-Control,

соответственно.

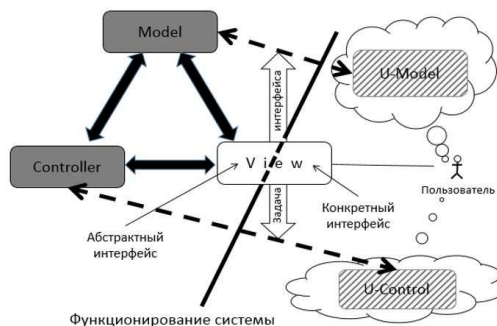


Рис. 1. Model-View-Controller и интерфейс пользователя

Понятия абстрактного и конкретного интерфейсов в точности соответствуют понятиям абстрактного и конкретного синтаксисов из области конструирования компиляторов (см. [9]). И в том, и в другом случае абстрактное является инвариантом множества конкретного: все варианты конкретного вычислительно эквиваленты, т.е. приводят к одним и тем действиям модели и управления. Аналогия распространяется и на разделение задачи конструирования интерфейса на две части: в архитектуре компиляторов ему соответствуют оперирование семантической информацией и синтаксический анализ.

#### 4 Деятельность пользователя и разработка интерфейса приложения

Два аспекта задачи конструирования интерфейса требуют различных методов решения. Разработка конкретного интерфейса исходит из таких понятий, как дизайн экрана, эргономичность, психологические ограничения, использование стандартов и сложившейся практики оформления взаимодействий с приложением. При построении абстрактного интерфейса ситуация иная. Здесь главным является обеспечение автоматизируемой приложением деятельности максимально возможной поддержкой. Понятно, что это только потенциальная возможность, которая может быть начисто уничтожена плохим конкретным интерфейсом. Такого рода примеры, приводит Дж. Раскин в [1]. Он сформулирует ряд универсальных принципов разработки интерфейсов с учетом когнитивных особенностей восприятия информации. В работе [7] мы обсуждали эргономические и психологические ограничения, а также роль привычек и сложившихся стихийно стандартов во влиянии на качество конкретных интерфейсов. По этим причинам ниже обсуждаются вопросы, связанные преимущественно с абстрактными интерфейсами в аспекте его формирования на основе анализа пользовательской деятельности.

##### 4.1 Общие положения деятельностного подхода

Понятие деятельности — ключевое для проектирования программных систем, и это тема отдельного рассмотрения. Здесь мы рассмотрим только те ее аспекты, которые имеют отношение к разработке интерфейсов. И первое, на что стоит обратить внимание, — это *структура деятельности*: из каких элементов она состоит? Естественное ограничение при ответе на этот вопрос — рассматривать только целенаправленные деятельности. Как следствие, появляются следующие *элементы деятельности* (см. рис.2):

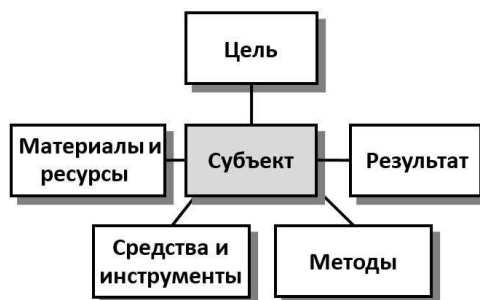


Рис. 2. Элементы деятельности

**Цель** — для чего организуется деятельности, т.е. какие продукты и с каким уровнем качества предписано получить при ее выполнении;

**Результат** — что фактически продуцируется в ходе выполнения деятельности. Результат не всегда соответствует цели. Если он включает все, что предписано целью, деятельность считается *успешной*, иначе — *неуспешной*. В обоих случаях при выполнении деятельности не только те продукты, которые соответствует цели (например, опыт разработчиков, отходы и др.);

**Субъект** — тот, кто выполняет деятельность. Это необязательно индивидуум, поскольку деятельность может выполняться автоматически, и необязательно реальный индивидуум, т.к. деятельность может быть коллективной, и тогда удобно говорить о виртуальном субъекте. Не любой субъект пригоден для выполнения деятельности — он должен обладать вполне определенными атрибутами (например, соответствующей квалификацией), которые необходимы для выполнения деятельности;

**Материалы и ресурсы** — то, из чего продуцируются результаты деятельности. Материалы могут быть как материальные, так и информационные, а ресурсы — расходующимися или не ограниченными (считаются такими);

**Средства и инструменты** — с помощью чего продуцируются результаты деятельности;

**Методы** — указывают на то, как выполнять деятельность для получения целевых результатов, потребляя ее материалы и ресурсы.

Деятельность может находиться в одном из своих *состояний*. Их можно рассматривать в качестве характеристик выполнения деятельности. На множестве состояний определено бинарное отношение возможности перехода из одного состояния в другое. Тем самым задается граф состояний. Переход между состояниями (если он возможен) происходит в результате *действий* субъекта, использующих элементы деятельности. Допускаются изменения состояний и без участия субъекта (пример — конкурентное использование ресурса, общего для разных деятельностей).

В графе состояний выделены *начальное состояние*, в котором деятельность активизируется, и множество *заключительных состояний*, в каждом из которых деятельность может завершиться. Некоторые из заключительных состояний определяются как *целевые* — считается, что в них цели деятельности достигнуты.

Последовательность состояний, которая выстраивается в результате действий субъекта, называется *операционным маршрутом* деятельности. Операционный маршрут называется *полным*, если он ведет от начального состояния в заключительное. Полный операционный маршрут, приводящий к целевому состоянию, считается *успешным*. В отличие от так



называемых *сценариев*, которые для попадания в целевое состояние предписывают субъекту определенные последовательности действий, операционные маршруты выстраиваются субъектом, т.е. допускают свободный выбор очередного состояния, что более точно отражает реальность целенаправленных деятельностей.

Деятельность всегда выполняется в некотором *окружении*, из которого поставляются ее элементы, и в которое передаются ее результаты в качестве элементов других деятельностей из окружения. Связанные этими отношениями деятельности могут группироваться, и такие группы можно трактовать как самостоятельные деятельности со своими окружениями и элементами. Таким образом образуются *система деятельностей*, в которой каждая из деятельностей составляется из других деятельностей, актуализирующих элементы группы.

Системы деятельностей образуются на основе отношения использования продуктов и результатов, передаваемых от одной деятельности другой. Эти и связанные с ними понятия уточняются следующим образом:

**Результат деятельности** — все, что производится в ней, независимо от того, используется это или нет в какой-либо другой деятельности (в последнем случае использование рассматривается вне изучаемой системы);

**Продукт деятельности Дп** — та часть результата Дп, которая используется в другой деятельности Ди в качестве какого-то ее элемента. Множество таких деятельностей  $A = \{D_1, \dots, D_n\}$  очень велико. В частности, оно содержит все варианты использования всех пользователей продукта;

**Целевой (основной) продукт деятельности Дп** — то, что заявлено в Дп как ее цель, т.е. обозначены некие деятельности  $A_T = \{D_{i_i}, i \in \{1, \dots, n\} | D_{i_i} \in A\}$ , о которых предполагается, что они будут использовать продукты Дп;

**Побочный продукт Дп** — та часть результата, не являющаяся целевым продуктом, для которой есть использующая его деятельность  $D_i \in A \setminus A_T$ ;

**Сопутствующий продукт Дп** — тот продукт (возможно, другой деятельности), без которого использование целевого продукта затруднительно (в разной степени).

Рис. 3 иллюстрирует введенные понятия. Светлые стрелки указывают на блоки, раскрывающие результаты деятельностей, которые предоставляют целевые и побочные продукты другим деятельностям (они выделены овалами). Темные стрелки отражают использование продуктов.

Автоматизация деятельности — это подмена каких-то ее элементов программными аналогами. Такая подмена неизбежно влечет изменение и других элементов за исключением ее цели — новый инструмент требует новых методов работы с ним, другой квалификации субъекта и пр. Для эффективного (во всех смыслах) выполнения субъектом деятельности необходимо, чтобы его новые операционные маршруты были бы сходными с ранее реализуемыми маршрутами, в частности, они не должны ломать привычные стандарты взаимодействия с системой. Это справедливо и при конструировании новой деятельности, которая предполагает достижение новых целей. В обоих случаях аналогия помогает освоению автоматизированной деятельности.

Сходство нового операционного маршрута с ранее освоенными пользователями, указывает на то, что деятельность, которую реализует этот маршрут, разбивается на фрагментные деятельности, и некоторые из них соответствуют сформировавшимся у пользователя поведенческим или иным стереотипам. Когда стереотипы отвечают эргономическим и когнитивным ограничениям, производительность труда пользователя повышается. Это наблюдение распространяется и на решение задач конкретного интерфейса. Для разработки абстрактного интерфейса оно может служить рекомендацией для выбора подходящих шаблонов

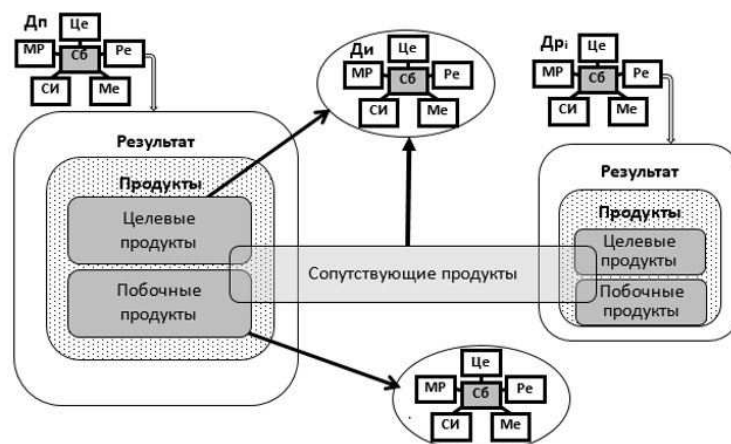


Рис. 3. Результат и продукты деятельности

в качестве атомарных единиц действий, так называемых виджетов — примитивов графического интерфейса пользователя, имеющий стандартный внешний вид и выполняющий стандартные действия [10].

#### 4.2 Конструирование абстрактного интерфейса

Предыдущее обсуждение показывает, что для разработки абстрактного интерфейса необходим анализ операционных маршрутов, при котором решается, что должно предъявляться пользователю в качестве информации и для выбора вариантов действий, но не как это предъявляется: показывается, рассказывается и др. Последнее — задача анализа и формирования представлений конкретного интерфейса, которую нужно решать после того, как приняты решения на абстрактном уровне.

Вариантов операционных маршрутов слишком много, чтобы пытаться охватить их все. Вместо этого целесообразно построить типовые сценарии пользовательской деятельности, которые ограничивают рассмотрение последовательностями состояний деятельности при движении к цели. В отличие от сценариев, предлагаемых пользователю в качестве руководства, аналитические сценарии для каждого состояния сопровождаются информацией о тех состояниях, в которые можно перейти в принципе. С точки зрения отдельного сценария большая часть переходов в такие состояния — это ошибки. Однако не исключено, что другие сценарии организуют последовательности состояний, для которых данные переходы правомерны. Полезное состояние не будет упущено не только, когда другой сценарий уже входит в число типовых, но и случаях, когда его полезность выяснится в дальнейшем, и придется модернизировать систему.

Первый шаг анализа связан с построением и исследованием всех операционных маршрутов аналитических сценариев. В результате для каждого состояния разработчики должны определить:

- Какие элементы деятельности используются в пользовательском действии, как изменяются элементы деятельности;
- Какой результат производится при переходе в новое состояния и в каких деятельности используются продукты;

- В какой программной поддержке нуждаются эти действия.

На основе этой информации выполняется следующий шаг анализа — проверка *функциональной полноты*, т.е. соответствие предлагаемых средств полному набору автоматизируемых функций, необходимых для пользовательских действий. Полнота проверяется исходя из текущего понимания потребностей пользователя, или, что то же, на основе всех действий субъекта на всех проанализированных операционных маршрутах. Если в процессе разработки программного продукта или даже в ходе его эксплуатации выясняется нарушение функциональной полноты, то комплект предлагаемых средств расширяется, и проверка повторяется.

Из функционально полного набор средств выделяются *базовые элементы* модели вычислений и требуемого управления абстрактного уровня. Набор этих элементов должен быть достаточен для реализации всех автоматизируемых функций, что означает требование *реализационной полноты*. Это достигается путем унификации и комбинирования элементов, а при необходимости, их декомпозиции. Реализация базовых элементов осуществляется в рамках принятых для проекта архитектурных решений и требований эффективности с использованием как общесистемных, так и специальных программных средств. В результате работ, связанных с обеспечением реализационной полноты, в архитектуре фиксируется достигнутый уровень функциональных возможностей. В терминах шаблона MVC это означает спецификацию модели, управления и реализационной части представления. Последнее есть основа конструирования конкретного интерфейса, т.е. той части представления, с которой имеет дело пользователь.

Наряду с функциональной и реализационной полнотой имеет смысл рассматривать *интерфейсную полноту*, понимая ее как задание языка управления поведением модели, адекватного восприятию программной системы пользователем. Если задача определения абстрактного интерфейса решена, то определена и семантика такого языка. И тогда для можно подбирать варианты возможного интерфейса, учитывая эргономические и когнитивные ограничения, о которых убедительно говорит Дж. Раскин[1]. Самый главный аспект интерфейсной полноты — поддержка всех сторон автоматизируемой деятельности, в частности, в интерфейсе должны быть распознаваемы все элементы деятельности.

Для каждого варианта конкретного интерфейса, удовлетворяющего ограничениям, в ходе проверки того, достигнута ли поддержка всех сторон автоматизируемой деятельности, выясняется его соответствие абстрактному интерфейсу. В то же время, среди критериев качества не последнее место занимает точность: среди предоставляемых элементов изображений, звуковых и иных сигналов не должно быть ничего лишнего. Это не означает, что в интерфейсе нет места декоративным элементам. Хорошо дозированные такие элементы способствуют комфортным ощущениям пользователя при работе с приложением.

Представленная только что схема применения принципов полноты нуждается в дополнении, когда результаты проверки полноты и точности неудовлетворительны, необходима ревизия проекта. В этом случае следует вернуться к этапу, который стал причиной дефекта. Уточненная схема представлена на рис. 4.

## 5 Заключение

В данной работе представлены те аспекты конструирования программных систем, которые затрагивают конструирование интерфейсов. Как показывает анализ многих программных разработок, часто поддержка деятельности в целом подменяется реализацией требуемых функций без заботы о том, как и для чего пользователь активизирует эти функции. Эта

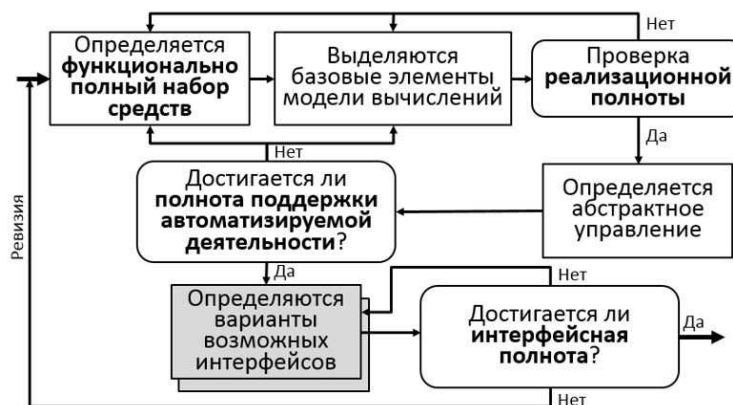


Рис. 4. Схема применения принципа полноты

позиция приемлема, когда в деятельности пользователя используется, к примеру, лишь вычислительный результат, получение которого не требует управления. Если же программа используется как инструмент деятельности, необходимо понимание всех аспектов этой деятельности, т.е. комплексный системный подход к организации предъявляемых средств.

Комплексность автоматизации пользовательской деятельности предполагает, что разработчики при конструировании программ не ограничиваются учетом эргономических, когнитивных и психологических ограничений, а стремятся к достижению соответствия предлагаемых средств структуре деятельности, понимают, что автоматизируемая деятельность есть часть большой системы деятельностей, игнорирование которой приводит к необходимости приспособляться к особенностям программы. На это уже довольно давно указывал Г.П. Щедровицкий в исследованиях по общей теории деятельности и ее приложению к проектировочной деятельности. Его статью [11] можно рассматривать как методологическую основу проведенного в настоящей работе анализа и выдвигаемых принципов конструирования интерфейсов.

Рассмотрение задачи конструирования интерфейсов с позиций теории деятельности приводит к тому, что хорошо себя зарекомендовавший шаблон проектирования MVC должен быть дополнен понятиями абстрактного и конкретного интерфейсов, на основе которых предлагается особый подход к проектированию и реализации интерфейсов. Одним из наиболее важных соглашений этого подхода является принцип полноты поддержки автоматизируемой деятельности, который естественным образом проявляется в функциональную, реализационную и интерфейсную полноту. Было показано, как можно организовать конструирование интерфейсов.

Подход, который мы предлагаем для конструирования интерфейсов и не только их, но и других составляющих разработки программных проектов в ряде аспектов перекликается с тем, что предлагается в документе РМВОК [12]. Однако нельзя не отметить принципиальное отличие: РМВОК ориентирован на процессное представление проектной деятельности, в котором нет места субъектам, разграничению продукта и результата, другим элементам деятельности. В результате проектные деятельности и их части оказываются черными ящиками со входами и выходами, разбавленными инструментами, которые влияют на выполнение процесса. По этой причине концепция деятельностного представления проектов представляется более перспективной тем более, когда речь идет о конструировании интерфейсов.

## Список литературы

1. Раскин Дж. *Интерфейс: новые направления в проектировании компьютерных систем* // Символ-Плюс. 2005. — 272 с. ISBN 5-93286-030-8
2. *Интегрированная среда разработки* // URL: <https://ru.wikipedia.org/wiki/> (дата обращения: 20.04.2015)
3. Фаронов В. В. *Turbo Pascal. Наиболее полное руководство* // BHV-Санкт-Петербург. 2007. — 1054 с. ISBN 5-94157-295-6, CD
4. Бутюгин Д.С., Гурьева Я.Л., Ильин В.П., Первозкин Д.В., Скопин И.Н. *Функциональность и технологии алгебраических решателей в библиотеке Krylov* // Вестник ЮУрГУ, 2013, т. 2, №3. — с. 92 – 105.
5. *Обзор новой версии Google Chrome 26* // URL: <http://www.comss.ru/page.php?id=1394>
6. *YouTube* // URL: <http://www.youtube.com/>
7. Скопин И.Н. *Разработка интерфейсов программных систем* // В сб. Системная информатика. Вып. 6 / Новосибирск: Наука. Сибирская издательская фирма. 1997. — с. 34 – 96.
8. Burbeck S. *Applications Programming in Smalltalk-80(TM): How to use Model-View-Controller (MVC)* // URL: <http://st-www.cs.illinois.edu/users/smarch/st-docs/mvc.html>
9. Ахо А.В., Лам М.С., Сети Р., Ульман Д.Д. *Компиляторы. Принципы, технологии и инструментарий* // Вильямс. 2014. — 1184 с. ISBN 978-5-8459-1932-8.
10. *Что такое виджеты* // URL: <http://widgetok.ru/2009/01/what-is-widgets/>
11. Щедровицкий Г. П. *Автоматизация проектирования и задачи развития проектировочной деятельности* // Разработка и внедрение автоматизированных систем в проектировании (теория и методология). — М.: Стройиздат, 1975. — С. 3 – 177.
12. *Руководство к своду знаний по управлению проектами (Руководство РМВОК®)*. — Project Management Institute, Inc, 2008. — 465 с.

# О Задаче Восстановления и Идентификации Множества Точек Разрыва Геометрических Объектов по Томографическим Данным \*

А.П. Полякова<sup>1,2</sup>, И.Е. Светов<sup>1,2</sup>, Е.Ю. Деревцов<sup>1,2</sup>, М.А. Султанов<sup>3</sup>

<sup>1</sup> Институт математики им. С. Л. Соболева СО РАН, Новосибирск, Россия

<sup>2</sup> Новосибирский государственный университет, Новосибирск, Россия

<sup>3</sup> Международный казахско-турецкий университет им. Х.А. Ясави, Туркестан, Казахстан  
anna.polyakova@ngs.ru, {svetovie, dert}@math.nsc.ru, smurat-59@mail.ru

**Аннотация.** В работе рассматривается задача восстановления множества точек разрыва тензорных полей малого ранга, заданных в единичном круге, по их известным лучевым преобразованиям. Для визуализации множества точек разрыва геометрических объектов используются операторы обратной проекции, действующие на лучевые преобразования, и аппарат тензорного анализа. Математическое описание указанного множества точек требует привлечения аппарата статистического анализа. Задачу определения величины скачка, наряду с методами математического анализа, можно исследовать с привлечением аппарата интегрального оператора Фурье. Разработаны и реализованы некоторые алгоритмы решения поставленных задач.

**Ключевые слова:** томография, тензорное поле, точки разрыва, лучевое преобразование, статистический анализ.

## 1 Введение

Во многих важных, — с точки зрения практики, — естественнонаучных и технических областях объекты исследований математически описываются величинами, терпящими разрыв. Такие объекты часто возникают в задачах, использующих дистанционные методы и, в частности, в задачах томографии. Так, в дефектоскопии обнаружение трещин в промышленных изделиях посредством неразрушающего контроля столь же важная задача, как и определение отклонений внутреннего строения изделия от эталона. Во многих задачах геофизики установление местоположения границ, разделяющих блоки с различными физическими свойствами, является первым этапом в дальнейших исследованиях, направленных на определение физических величин, характеризующих внутреннее строение Земли.

В процессе становления томографии проблема реконструкции объектов с разрывными физическими свойствами выделилась в самостоятельную задачу. Тщательное теоретическое описание объектов с разрывными свойствами в рамках интегральной геометрии посредством обобщенных функций (распределений) было предложено в 1962 году в монографии [1]; дальнейшее развитие теории, уже в томографических постановках, можно найти, например, в работах [2], [3]. В настоящее время развитие теоретических методов исследования геометрических объектов с разрывными свойствами (разрывы самих объектов и разрывы их производных) осуществляется как в традиционных рамках теории распределений, так и, в основном, средствами микролокального анализа. Таким образом, к настоящему времени математический аппарат описания объектов такого рода достаточно хорошо разработан. В то же время алгоритмическая и программная реализация этого аппарата включает в себе

\* Работа осуществлена при частичной финансовой поддержке РФФИ (проект 14-01-31491-мол-а), Министерства образования и науки Республики Казахстан (проект 3630/ГФ4-2015)

определенные сложности, и прежде всего это неудовлетворительная точность восстановления разрывных объектов.

**Восстановление разрывов функции.** Проблема восстановления разрывов функции по ее известному преобразованию Радона как самостоятельная задача была поставлена сравнительно недавно, а известный алгоритм восстановления разрывов был предложен в работе Е. И. Вайнберга с соавторами в 1985 г. [4] (“алгоритм Вайнберга”). Вообще говоря, под термином “восстановление разрывов” логично подразумевать несколько задач. Первая задача состоит в *визуализации разрывов*, и именно эта задача исследуется в подавляющем большинстве работ. Вторая задача заключается в *идентификации разрывов*, т. е. в математическом описании множества точек сингулярного носителя. Третья задача состоит в *определении величины скачка*.

Вернемся к разработанным ранее методам восстановления разрывов функций. Основная идея алгоритма Вайнберга, результатом работы которого является изображение, хорошо “проявляющее” множество точек разрыва функции, состоит в предварительном двойном дифференцировании по переменной  $s$  ( $|s|$  — расстояние от прямой, по которой производится интегрирование, до начала координат) двумерного преобразования Радона, с последующим применением оператора обратной проекции. В дальнейшем такая последовательность действий, приводящая к визуализации множества разрывов, но при этом не позволяющая судить о поведении гладкой составляющей объекта, получила название оператора Вайнберга. Следует подчеркнуть, что результат применения указанного оператора не дает искомую функцию, как это происходит в случае использования формул обращения. Именно, нелокальный псевдодифференциальный оператор, используемый в формулах обращения, заменяется локальным дифференциальным оператором двойного дифференцирования, что существенно упрощает его программную реализацию, но не позволяет восстановить гладкую часть объекта.

Самими авторами применение оператора двойного дифференцирования интерпретировалось как некая фильтрация. В дальнейшем (см., например, [5]) было предложено иное обоснование применения оператора Вайнберга. Именно, этот оператор восстанавливает функцию  $(-\Delta)^{1/2}f$  ( $\Delta$  — оператор Лапласа). Поскольку  $(-\Delta)^{1/2}$  — эллиптический псевдодифференциальный оператор, функция  $(-\Delta)^{1/2}f$  обладает теми же сингулярностями, что и  $f$ . Развитие этой идеи для обращения преобразования Радона  $\mathcal{R}f$  негладкой функции  $f$  состоит в следующем. Прежде всего легко проверяется, что при преобразовании Радона оператор Лапласа  $\Delta$  на функциях  $f(x, y)$  переходит в оператор  $\partial^2/\partial s^2$  на функциях  $(\mathcal{R}f)(\alpha, s)$ . Более подробно, если задана обобщенная функция  $F(r^2)$  ( $r^2 = x^2 + y^2$ ), то при преобразовании Радона  $F(-\Delta)f$  (по определению есть  $F(r^2) * f$ ) перейдет в свертку по  $s$ :  $F(s^2) * (\mathcal{R}f)$ . Пусть теперь задана обобщенная (возможно, негладкая) функция  $f$ , которая представима в виде  $f = (1 - \Delta)^k p$ , где  $p$  обладает необходимой степенью гладкости,  $k$  натуральное. Тогда можно определить преобразование Радона функции  $f$  как свертку  $F((1 + s^2)^k) * (\mathcal{R}p)$ . Далее, вычисляя свертку  $\mathcal{R}f$  с  $(1 + s^2)^{-k}$ , получим достаточно гладкую функцию  $\mathcal{R}p$ . Используя формулу обращения для преобразования Радона, находим функцию  $p$ , а применяя к последней дифференциальный оператор  $(1 - \Delta)^k$ , получаем искомую функцию  $f$ .

Дальнейшее развитие подходов и алгоритмических средств восстановления множества точек разрыва осуществляли, в рамках *локальной томографии*, такие исследователи как А. Фаридани с соавторами [6], [7], А. К. Луис и П. Маасс [8], и многие другие. В этих работах, наряду с оператором Вайнберга, использовался и оператор обращения  $(-\Delta)^{1/2}$ , в сочетании с регуляризацией либо той или иной фильтрацией. Основная цель таких исследований

состояла в восстановлении множества разрывов, а также в возможности определения некоторых усредненных характеристик гладкой составляющей объекта.

В конце 90-х годов Д. С. Аниконовым был предложен иной подход к решению задачи определения множества разрывов функции по лучевым преобразованиям, основанный на теории многомерных сингулярных интегралов [9]. Применяя к лучевому преобразованию оператор обратного проектирования, получаем сингулярный интеграл (с искомой разрывной функцией в подынтегральном выражении) со слабой особенностью. Дифференцирование по пространственным переменным приводит тогда к логарифмическому возрастанию при стремлении точки к линии разрыва. В частности, можно использовать оператор  $|\nabla(\cdot)|$ . Описанный подход теоретически обоснован реализован алгоритмически и программно, в том числе и с включением явления рассеяния в модель среды [10], [11].

Некоторые подходы, позволяющие восстановить не только множество разрывов, но и соответствующие скачки функции по ее преобразованию Радона, предложены в работах [12], [13], [14]. Для решения этой задачи используются методы математического и микролокального анализа, интегральный оператор Фурье.

**Восстановление сингулярного носителя.** Как уже отмечалось, в последние несколько десятилетий задача восстановления множества точек разрыва функции активно исследовалась многими авторами как в России, так и за рубежом. В последние годы, в связи с интенсивным развитием векторной и тензорной томографии, постановка задачи восстановления разрывов были существенно обобщена (см. например [15]), и в настоящее время может трактоваться как задача восстановления сингулярного носителя симметричных тензорных полей по их известным лучевым преобразованиям. Иными словами, речь идет не только о функциях, но и о векторных и тензорных полях, и, кроме того, о восстановлении не только разрывов самих полей, но и их производных. Так, некоторые теоретические подходы определения сингулярного носителя тензорного поля, заданного на римановом многообразии, по известному продольному лучевому преобразованию этого поля, предложены в работе [16]. Описание предлагаемых с целью восстановления сингулярного носителя скалярных и векторных полей алгоритмов, их программных реализаций и тестовых расчетов можно найти в [17], [18], [19].

Обобщение постановки задачи восстановления разрывов функции приводит к необходимости разработки адекватных методов и алгоритмов для реконструкции множества точек сингулярного носителя скалярных, векторных и симметричных 2-тензорных полей по их лучевым преобразованиям, или, шире, по томографическим данным. К настоящему времени в томографии разработано много приближенных методов и огромное число алгоритмов и программных средств, направленных на восстановление внутренних свойств объекта. Чаще всего при этом используются подходы, основанные на формулах обращения, проекционных теоремах, методах прикладного функционального анализа, вариационных и алгебраических методах. Обычно они хорошо себя проявляют при восстановлении объектов с гладкими свойствами, но дают неудовлетворительные результаты для объектов с разрывными характеристиками.

По-видимому, имеется несколько путей восстановления разрывных геометрических объектов, т. е. скалярных, векторных и тензорных полей по их известным лучевым преобразованиям. Опишем кратко, без ограничения общности, эти пути, примененные по отношению к скалярному полю (функции). Первый вариант состоит в применении уже известных в томографии алгоритмических средств, без всякой их модификации. Практика численных экспериментов показывает, что точность восстановления разрывной функции, по сравнению с непрерывными, в 5–10 раз хуже. Поэтому необходим резко возрастающий объем исходных



томографических данных и, следовательно, значительное увеличение времени для достижения приемлемой точности результатов расчетов.

Второй путь заключается в разработке специальных алгоритмических средств, направленных на восстановление именно разрывных функций. Можно следующим образом детализировать поставленную задачу и выделить следующие этапы ее решения: а) визуализация множества точек разрыва исследуемой функции; б) локализация этого множества и его приближенное описание в рамках дискретной модели исследуемого объекта; в) определение, по найденному на предыдущих этапах множеству точек разрыва функции, величины скачка, характеризующего разрыв; г) устраняя разрывы, восстанавливаем более гладкую, уже не обладающую разрывами, функцию любым из общепринятых и хорошо известных в томографии методов; д) по известным величинам скачка восстанавливается исходная разрывная функция.

Имеется и третий путь восстановления разрывных объектов, но он применим лишь к задаче восстановления разрывов (сингулярного носителя) симметричных  $m$ -тензорных полей ранга  $m \geq 1$ . Именно, речь идет о восстановлении потенциалов полей. Известно, что потенциалы обладают на единицу большей гладкостью по сравнению с самими полями, поэтому они по крайней мере непрерывны. Следовательно, их точность восстановления обычными средствами в 5–10 раз лучше, нежели точность восстановления генерируемых этими потенциалами полей.

Остановимся более подробно на втором, самом сложном и требующем привлечения серьезного математического аппарата, пути. Отметим, что в подавляющем большинстве случаев под термином “восстановление разрывов” подразумевается задача визуализации разрывов, т. е. первый из пяти этапов задачи восстановления разрывного объекта.

Задача визуализации разрывов обычно состоит в получении такого изображения, полученного на основе томографических данных и связанного с исследуемым объектом, на котором множество точек, в которых функция терпит разрыв, легко узнаваемо. Это может быть резкая граница в цветных или полутоновых плоских изображениях, либо резкое возрастание значений функции в окрестности точек разрыва, если изображение представляется в форме рельефа. Отметим, что математическая постановка задачи визуализации множества точек разрыва функции или, — в более общей постановке, — сингулярного носителя тензорного поля, является корректной, в отличие от слабо некорректной задачи обращения преобразования Радона. Это вполне ясно из постановки, в которой нелокальный псевдодифференциальный оператор заменяется локальным; это следует также и из результатов численных экспериментов (подробнее см. ниже). Под локализацией (или идентификацией) множества точек разрыва функции понимается его строгое описание в математических терминах. Например, приближенное определение координат точек разрыва и, далее, задание аппроксимаций линий или поверхностей, состоящих из точек разрывов, уравнениями в той или иной форме. Математическое описание указанного множества требует привлечения подходов, весьма отличных от тех, что обычно используются в томографии, поскольку задача по своей сути относится к проблеме распознавания образов (в широком смысле) и, далее, к аппроксимации данных. Прежде всего это статистический анализ данных, описание классов объектов и т. п. Постановка задачи определения величины скачка в точке разрыва ясна и не требует дополнительных пояснений, хотя требует непростого математического аппарата. Так, наряду с использованием уравнения Гамильтона-Якоби и метода стационарной фазы, применяется и аппарат микролокального анализа; в частности, интегральный оператор Фурье. Известно, что преобразование Радона может быть представлено в форме такого оператора, что позволяет стандартными методами установить геометрическую связь между волновым фронтом исходной функции и волновым фронтом ее преобразования Радона.

Таким образом, особенности функции  $f$  могут быть найдены непосредственно из особенностей ее преобразования Радона. Пункты г) и д) тесно связаны. По-существу, здесь требуется построить функцию, устраняющую разрывы, и воспользоваться ею дважды.

Насколько известно авторам, в сформулированной выше детальной постановке задача восстановления разрывной функции по томографическим данным в литературе ранее не встречалась. Часто качественной информации, получаемой в результате реализации пункта а), а именно визуализации разрывов, бывает достаточно для многих практических задач. В некоторые работы рассматривается задача восстановления величины скачка (см. ссылки выше), но при этом, к большому сожалению, убедительных тестовых расчетов не приводится, что связано, по-видимому, со значительной численной неустойчивостью решения этой задачи. Как формулировки, так и решения задач, поставленных в пунктах б), г) и д), в томографической литературе практически отсутствуют.

## 2 Основные томографические операторы и формулы обращения для потенциалов

Введем обозначения для подмножеств плоскости  $\mathbb{R}^2$ , на которой задана прямоугольная декартовой система координат.  $B = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\}$  — круг с границей  $\partial B = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$  — единичной окружностью. Через  $Z = \{(\alpha, s) \in \mathbb{R}^2 : \alpha \in [0, 2\pi], s \in [-1, 1]\}$  обозначен цилиндр  $[-1, 1] \times [0, 2\pi]$ ; через  $\xi \in \partial B$ ,  $\xi = (\cos \alpha, \sin \alpha)$ ,  $\eta := \xi^\perp \in \partial B$ ,  $\eta = (-\sin \alpha, \cos \alpha)$  — единичные векторы. Через  $L_{\xi, s}$  обозначена прямая, задаваемая нормальным уравнением  $x \cos \alpha + y \sin \alpha - s = 0$ ; через  $S^m(B)$  обозначаем множества определенных в  $B$  симметричных, т.е. инвариантных относительно перестановок индексов,  $m$ -тензорных полей.

Введем класс функций, с помощью которых можно описать не только непрерывные или  $C^k$ -гладкие,  $k \in \mathbb{N}$ , но и разрывные векторные и тензорные поля, а также поля с разрывами в производных. Из физических соображений естественно предполагать, что речь идет только о разрывах 1-го рода.

Область  $D \subset \mathbb{R}^2$  такая, что  $\bar{D} \subset \bar{B}$ , состоит из конечного числа непересекающихся подобластей  $\{D_i\}$ ,  $i = 1, \dots, N$ , таких что объединение  $D_0 = \cup D_i$  этих подобластей плотно в  $\bar{D}$ , а их границы гладкие класса  $C^1$ . Отметим, во-первых, что  $D$  может быть многосвязной, а, во-вторых,  $\bar{D}$  может совпадать с  $\bar{B}$ . Нетрудно заметить, что  $\partial D \subset \partial D_0$ , а граница  $\partial D_0$  совпадает с объединением границ  $\cup D_i$  подобластей  $D_i$ ,  $i = 1, \dots, N$ . Важное требование к границам состоит в том, что они не должны содержать прямолинейных участков.

Пусть функция  $\varphi(x, y)$  класса  $C^k$  определена в  $B$ , причем она обращается в 0 на множествах  $\mathbb{R}^2 \setminus B$ ,  $B \setminus D$ , а ее носитель совпадает с замыканием  $D$ ,  $\text{supp } \varphi = \bar{D}$ . В точках  $(x, y) \in D$  область  $D$  функция  $\varphi(x, y)$  бесконечно дифференцируема. В точках  $(x, y) \in \partial D_0$  она непрерывно дифференцируема до  $k$ -го порядка включительно и обращается в 0. В силу своей гладкости в области  $D$  функция  $\varphi$  обладает частными производными любого порядка. Что касается точек, принадлежащих  $\partial D_0$ , то в них все частные производные

$$\frac{\partial^l \varphi}{\partial x^j \partial y^{l-j}}, \quad l = 0, \dots, k, \quad j \leq l,$$

до порядка  $k$  включительно непрерывны, а производные порядка  $k + 1$  терпят разрыв 1-го рода. Таким образом,  $C^{-1}$ -потенциалы описывают разрывные функции, посредством  $C^0$ -потенциалов описываются разрывные векторные поля, и т. д.

Операторы лучевых преобразований  $\mathcal{P}_m^{(j)} : C^k(S^m) \rightarrow C^k(Z)$ ,  $j = 0, \dots, m$ ,  $k \geq -1$ , действующие на симметричное  $m$ -тензорное поле  $w = (w_{i_1, \dots, i_m})$ , переводят его в функции

$g_m^{(j)}(\xi(\alpha), s)$ , определенные в цилиндре  $Z$ , и задаются соотношениями

$$(\mathcal{P}_m^{(j)}w)(\xi, s) = \int_{-\infty}^{\infty} \xi^{i_1} \dots \xi^{i_j} \eta^{i_{j+1}} \dots \eta^{i_m} w_{i_1 \dots i_m} dt. \tag{1}$$

$\xi = (\cos \alpha, \sin \alpha)$ ,  $\eta = (-\sin \alpha, \cos \alpha)$ . При  $j = 0$  это продольное лучевое преобразование  $\mathcal{P}$ , и в подынтегральном выражении участвуют только компоненты направляющего вектора (прямой, вдоль которой производится интегрирование)  $\eta$ . При  $j = m$  это поперечное лучевое преобразование  $\mathcal{P}^\perp$ , в котором участвуют только компоненты нормального вектора  $\xi$ . Полагая  $m = 0$ , получим преобразование Радона функции (скалярного поля). При  $m = 1, j = 0, 1$  получаем продольное и поперечное преобразования векторного поля, соответственно. Если  $m = 2, j = 0, 1, 2$  то имеем продольное, смешанное и поперечное лучевые преобразования симметричного 2-тензорного поля.

Приведем формулу для операторов дифференцирования образа лучевого преобразования (1) по его аргументам  $s$  и  $\alpha$ . Считаем  $(\mathcal{P}_m^{(j)}w)(\alpha, s) \in \mathcal{C}^k(Z)$ ,  $k \geq 0$ , функцией, зависящей от угла  $\alpha$  (так как  $\xi = (\cos \alpha, \sin \alpha)$ )

$$D_s^{(l)}(\alpha, s) := \frac{\partial^l (\mathcal{P}_m^{(j)}w)(\alpha, s)}{\partial s^l}, \quad D_\alpha^{(n)}(\alpha, s) := \frac{\partial^l (\mathcal{P}_m^{(j)}w)(\alpha, s)}{\partial \alpha^n},$$

для  $l, n \leq k + 1$ .

Оператор  $(n, l)$ -углового момента отображает функции  $h(\xi(\alpha), s)$ , определенные в цилиндре  $Z$ , в симметричные  $n$ -тензорные поля, определенные в  $\mathbb{R}^2$ , по правилу

$$(f_n^{(l)})_{i_1 \dots i_n}(x, y) = \frac{1}{2\pi} \int_0^{2\pi} \xi^{i_1} \dots \xi^{i_l} \eta^{i_{l+1}} \dots \eta^{i_n} h(\xi(\alpha), s(x, y, \alpha)) dt.$$

Здесь  $\xi = \xi(\alpha)$ ,  $\eta = \eta(\alpha)$ ,  $s = s(x, y, \alpha) := x \cos \alpha + y \sin \alpha$ . Частным случаем оператора углового момента является оператор обратной проекции. Пусть задан образ  $g_m^{(j)}(\xi(\alpha), s)$  смешанного лучевого преобразования  $(\mathcal{P}_m^{(j)}w)(\xi, s)$  симметричного  $m$ -тензорного поля. Тогда оператор обратной проекции  $(\mathcal{P}_m^{(j)})^\# : \mathcal{C}^k(Z) \rightarrow \mathcal{S}'(S^m)$  задается соотношением

$$(\mu_m^{(j)})_{i_1 \dots i_m}(x, y) = \frac{1}{2\pi} \int_0^{2\pi} \xi^{i_1} \dots \xi^{i_j} \eta^{i_{j+1}} \dots \eta^{i_m} g_m^{(j)}(\xi(\alpha), s(x, y, \alpha)) dt,$$

где  $\mathcal{S}'(S^m)$  – пространство Шварца медленно растущих на бесконечности (обобщенных) симметричных  $m$ -тензорных полей. Иными словами,

$$(\mu_m^{(j)})(x, y) = ((\mathcal{P}_m^{(j)})^\#(\mathcal{P}_m^{(j)}w))(x, y).$$

Операторы дифференцирования  $D_s^{(l)}$  и  $D_\alpha^{(n)}$ , углового момента и обратной проекции действуют на образы лучевых преобразований. Операторы тензорного анализа, которые приводятся ниже, действуют уже на тензорные поля  $(f_n^{(l)})_{i_1 \dots i_n}(x, y)$  или  $(\mu_m^{(j)})_{i_1 \dots i_m}(x, y)$ , являющиеся результатом применения операторов углового момента или обратной проекции.

Обобщениями классических операторов градиента  $\nabla f = (\partial f / \partial x, \partial f / \partial y)$  и ортогонального градиента  $\nabla^\perp f = (-\partial f / \partial y, \partial f / \partial x)$  являются операторы внутреннего дифференцирования  $d$  и  $d^\perp$ . Оператор внутреннего дифференцирования  $d : \mathcal{C}^k(S^m) \rightarrow \mathcal{C}^{k-1}(S^{m+1})$ ,  $k \geq 0$ ,

действует на симметричное  $m$ -тензорное поле  $w$  и дает симметричное  $(m + 1)$ -тензорное поле  $u$  по правилу

$$u_{i_1 \dots i_m j} := (dw)_{i_1 \dots i_m j} = \frac{1}{m+1} \left( \frac{\partial w_{i_1 \dots i_m}}{\partial x^j} + \sum_{k=1}^m \frac{\partial w_{i_1 \dots i_{k-1} j i_{k+1} \dots i_m}}{\partial x^{i_k}} \right).$$

Оператор  $d^\perp : \mathcal{C}^k(S^m) \rightarrow \mathcal{C}^{k-1}(S^{m+1})$ ,  $k \geq 0$ , внутреннего ортогонального дифференцирования действует по правилу

$$v_{i_1 \dots i_m j} := (d^\perp w)_{i_1 \dots i_m j} = \frac{1}{m+1} \left( (-1)^j \frac{\partial w_{i_1 \dots i_m}}{\partial x^{3-j}} + \sum_{k=1}^m (-1)^{i_k} \frac{\partial w_{i_1 \dots i_{k-1} j i_{k+1} \dots i_m}}{\partial x^{3-i_k}} \right).$$

Здесь  $w \in \mathcal{C}^k(S^m)$ ,  $u, v \in \mathcal{C}^{k-1}(S^{m+1})$ ,  $k \geq 0$ , использовались обозначения  $x^1 = x$ ,  $x^2 = y$ . Операторы дивергенции  $\delta$  и ортогональной дивергенции  $\delta^\perp$ ,  $\delta, \delta^\perp : \mathcal{C}^k(S^m) \rightarrow \mathcal{C}^{k-1}(S^{m-1})$ ,  $k \geq 0$ , действуют на симметричные  $m$ -тензорные поля  $w$ ,

$$u_{i_1 \dots i_{m-1}} := (\delta w)_{i_1 \dots i_{m-1}} = \frac{\partial w_{i_1 \dots i_{m-1} 1}}{\partial x} + \frac{\partial w_{i_1 \dots i_{m-1} 2}}{\partial y},$$

$$v_{i_1 \dots i_{m-1}} := (\delta^\perp w)_{i_1 \dots i_{m-1}} = -\frac{\partial w_{i_1 \dots i_{m-1} 1}}{\partial y} + \frac{\partial w_{i_1 \dots i_{m-1} 2}}{\partial x},$$

и дают симметричные тензорные поля  $u, v$  валентности  $m - 1$ .

Дискретные значения образов операторов лучевых преобразований (1) являются исходными данными для задачи восстановления разрывов (сингулярного носителя) симметричных  $m$ -тензорных полей. Все остальные операторы представляют собой математический аппарат, предназначенный для решения поставленной задачи, первый этап решения которой заключается в визуализации разрывов (сингулярного носителя).

**Формулы обращения для потенциалов векторных и симметричных 2-тензорных полей.** Приведем формулы обращения преобразования Радона функций в форме, которая практически без изменений позволяет восстановить потенциалы векторных и симметричных 2-тензорных полей. Формулы обращения, записанные в таком виде хорошо известны особенно в работах прикладного характера,

$$\varphi(x, y) = -\frac{1}{4\pi^2} \int_0^{2\pi} \int_{-\infty}^{\infty} \frac{(\mathcal{R}\varphi)'_s(\alpha, s + x \cos \alpha + y \sin \alpha)}{s} ds d\alpha, \quad (2)$$

где интеграл по  $s$  понимается в смысле главного значения по Коши. Интегрирование по частям внутреннего интеграла приводит к другим разновидностям формул обращения. Так, в одной из них используется вторая производная  $(\mathcal{R}\varphi)''_{ss}$ ,

$$\varphi(x, y) = \frac{1}{4\pi^2} \int_0^{2\pi} \int_{-\infty}^{\infty} (\mathcal{R}\varphi)''_{ss}(\alpha, s + x \cos \alpha + y \sin \alpha) \ln |s| ds d\alpha, \quad (3)$$

а во второй применяется само преобразование Радона  $\mathcal{R}\varphi$ ,

$$\varphi(x, y) = -\frac{1}{4\pi^2} \int_0^{2\pi} \int_{-\infty}^{\infty} \frac{(\mathcal{R}\varphi)(\alpha, s + x \cos \alpha + y \sin \alpha)}{s^2} ds d\alpha.$$

Известно [19], что лучевые преобразования векторных полей связаны с потенциалами, генерирующими эти поля, следующим образом.

– Поперечное лучевое преобразование поля  $u \in \mathcal{C}^k(S^1(B))$ ,  $u = d\psi$ ,  $\psi \in \mathcal{C}^{k+1}(B)$ , связано с преобразованием Радона его потенциала  $\psi$  соотношением

$$(\mathcal{P}^\perp u)(\xi, s) = \frac{\partial}{\partial s} \mathcal{R}\psi(\xi, s).$$

– Продольное лучевое преобразование поля  $v \in \mathcal{C}^k(S^1(B))$ ,  $v = d^\perp \varphi$ ,  $\varphi \in \mathcal{C}^{k+1}(B)$ , связано с преобразованием Радона его потенциала  $\varphi$  соотношением

$$(\mathcal{P}v)(\xi^\perp, s) = \frac{\partial}{\partial s} \mathcal{R}\varphi(\xi, s).$$

– Если  $\varphi = \psi$ ,  $\varphi, \psi \in \mathcal{C}^k(B)$ ,  $k \geq 0$ ,  $u = d\varphi$ ,  $v = d^\perp \varphi$  то  $\langle u, v \rangle = 0$ . Кроме того,

$$(\mathcal{P}^\perp u)(\xi, s) = (\mathcal{P}v)(\xi^\perp, s) = \frac{\partial}{\partial s} \mathcal{R}\varphi(\xi, s).$$

Из этих свойств следует, что потенциалы  $\psi$  или  $\varphi$  векторных полей, в зависимости от типа известных в качестве данных лучевых преобразований — поперечного или продольного, можно найти с помощью модифицированных формул обращения (2), (3). Приведем формулы обращения для потенциала  $\varphi$  соленоидального поля  $v = d^\perp \varphi$ .

$$\varphi(x, y) = -\frac{1}{4\pi^2} \int_0^{2\pi} \int_{-\infty}^{\infty} \frac{(\mathcal{P}w)(\alpha, s + x \cos \alpha + y \sin \alpha)}{s} ds d\alpha,$$

$$\varphi(x, y) = \frac{1}{4\pi^2} \int_0^{2\pi} \int_{-\infty}^{\infty} (\mathcal{P}w)'_s(\alpha, s + x \cos \alpha + y \sin \alpha) \ln(s) ds d\alpha.$$

На основе восстановленного потенциала  $\varphi$  строится соленоидальное поле  $v = d^\perp \varphi$ . Для этого необходимо численно продифференцировать потенциал  $\varphi$  и, тем самым, найти значения компонент  $v_1 = -\frac{\partial \varphi}{\partial y}$ ,  $v_2 = \frac{\partial \varphi}{\partial x}$  поля  $v$ . Аналогичные формулы обращения можно использовать и для восстановления потенциала  $\psi$  потенциального поля  $d\psi$ , а затем найти и значения компонент  $u_1 = \frac{\partial \psi}{\partial x}$ ,  $u_2 = \frac{\partial \psi}{\partial y}$  поля  $u$ .

Отсюда вполне понятен алгоритм восстановления разрывного (с разрывами первого рода) векторного поля. Сначала мы находим потенциал класса  $\mathcal{C}^0$  (непрерывный). После дифференцирования определяем местоположение (например, статистическими методами) множества точек разрыва и величину скачка. В рассуждениях мы пока абстрагируемся, от того, насколько численно неустойчивой может оказаться задача определения величины скачка.

Аналогичные рассуждения и выводы полностью применимы и к симметричным 2-тензорным полям. Действительно, каждое из трех типов таких полей, два (разных) из которых потенциальны и одно соленоидально, однозначно определяется своим потенциалом. Продольное, поперечное и смешанное лучевые преобразования полей связаны с преобразованиями Радона их потенциалов соотношениями, аналогичными соотношениям между лучевыми преобразованиями векторных полей и преобразованиями Радона соответствующих потенциалов. Поэтому потенциалы тензорных полей восстанавливаются точно так же, как и потенциалы векторных, но использовать можно лишь формулу (3).

### 3 Численные эксперименты

**Тест 1:** зависимость точности визуализация линий разрыва векторного поля от уровня шума.

Соленоидальное векторное поле генерируется потенциалом (в полярных координатах)

$$\varphi(x, y) = \begin{cases} r^2, & \text{при } |x| \leq 1/4 \text{ и } |y| \leq 1/4 \\ 4\pi - \phi, & \text{при } r \leq \phi \text{ и } |x| > 1/4 \text{ и } |y| > 1/4 \\ 2\pi - \phi, & \text{при } \phi < r \leq \phi + 2\pi \text{ и } |x| > 1/4 \text{ и } |y| > 1/4 \\ 0, & \text{иначе} \end{cases}$$

Соленоидальное векторное поле задано соотношениями

$$v_1 = \frac{\partial \varphi}{\partial y}, \quad v_2 = -\frac{\partial \varphi}{\partial x}.$$

Потенциал поля и его компоненты приведены на рисунке 1.

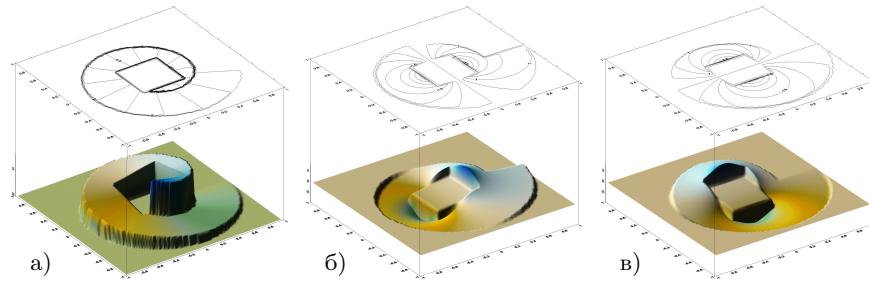


Рис. 1. Потенциал векторного поля (а) и его первая (б) и вторая (в) компоненты.

Визуализация линий разрыва соленоидального векторного поля представлена на рисунке 2. Именно, приведены результаты применения оператора модуля градиента к обратной проекции исследуемого векторного поля, который можно представить схематически следующим образом:

$$g := (\mathcal{P}\varphi)(\xi, s) \rightarrow \psi(x, y) := (\mathcal{P}^\#g)(x, y) \rightarrow \nabla\psi(x, y) \rightarrow |\nabla\psi|(x, y).$$

В исходные данные задачи, т. е. в продольное лучевое преобразование, внесен шум соответствующего уровня. Это 0% (без шума), 5%, 10% и 20%.

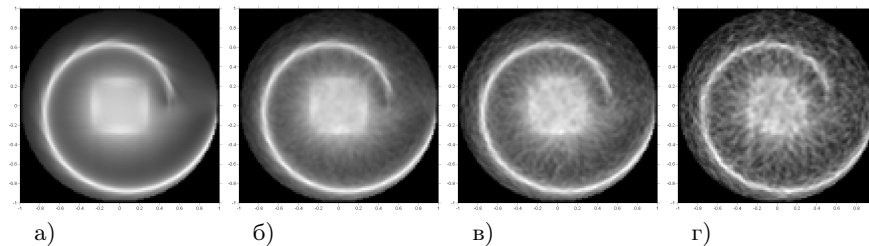


Рис. 2. Визуализация линий разрыва векторного поля при привнесении в исходные данные шума уровней 0% (а), 5% (б), 10% (в) и 20% (г).

Из рисунка видно, что даже очень большой уровень шума в 20% позволяет опознать структуру множества точек разрывов векторного поля.

**Тест 2:** действие операторов Вайнберга и модуля градиента на функции различной степени гладкости.

Оператор Вайнберга, действующий на преобразование Радона функции  $\varphi$ , схематически можно представить следующим образом:

$$(\mathcal{R}\varphi)(\xi, s) \rightarrow g(\xi, s) := \frac{\partial^2(\mathcal{R}\varphi)}{\partial s^2}(\xi, s) \rightarrow (\mathcal{R}^\#g)(x, y).$$

Условно можно сказать, что однократное действие оператора Вайнберга, в определенном смысле (гладкости) эквивалентно двукратному применению оператора модуля градиента.

Элементарный фантом  $\varphi$  представляет собой “ $\lambda$ -параболу” с эллиптическим носителем,

$$\varphi(x, y) = \begin{cases} C(1 - t^2)^\lambda, & t \leq 1, \\ 0, & t > 1, \end{cases}$$

где

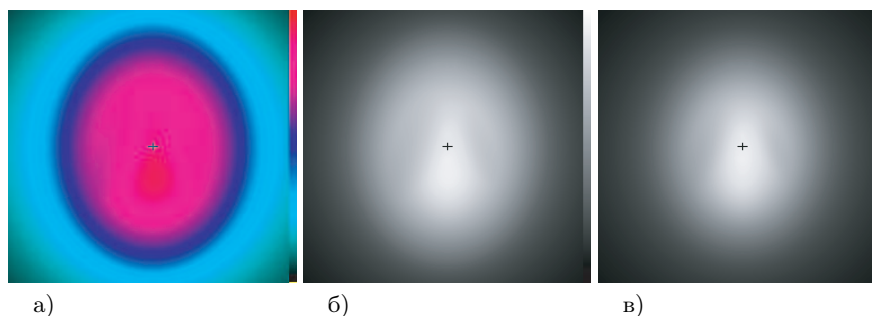
$$t^2 = \frac{((x - x_0) \cos \beta + (y - y_0) \sin \beta)^2}{a^2} + \frac{(-(x - x_0) \sin \beta + (y - y_0) \cos \beta)^2}{b^2},$$

и соответствующим преобразованием Радона,  $\lambda > 0$  [20],

$$(\mathcal{R}\varphi)(s, \phi) = \frac{\sqrt{\pi}abC\Gamma(\lambda + 1)}{|\zeta|\Gamma(\lambda + 3/2)} \left(1 - \frac{(s - s_0)^2}{\zeta^2}\right)^{\lambda+1/2},$$

где  $s_0 = -x_0 \sin \phi + y_0 \cos \phi$ ,  $\zeta^2 = a^2 \sin^2(\phi - \beta) + b^2 \cos^2(\phi - \beta)$ . Из элементарных фантомов составлен сложный фантом, схожий по контурам с фантомом Шеппа-Логана, и условно именуемый “потенциал 2”.

На рисунке 3 продемонстрировано действие оператора обратной проекции, примененного к преобразованию Радона “потенциала 2”, различной степени гладкости, определяемой параметром  $\lambda$ .



**Рис. 3.** Оператор обратной проекции при  $\lambda = 0.5$  (а),  $\lambda = 1$  (б) и  $\lambda = 2$  (в).

На рисунках 4-6 демонстрируются результаты применения операторов Вайнберга (рис. 4), модуля градиента (рис. 5), двукратно модуля градиента (рис. 6), к образу оператора обратной проекции, примененного к преобразованию Радона “потенциала 2”, различной степени гладкости, определяемой параметром  $\lambda$ .

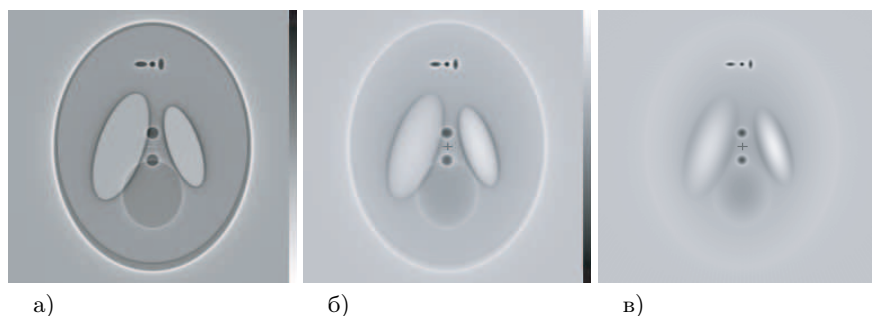


Рис. 4. Применение оператора Вайнберга при  $\lambda = 0.5$  (а),  $\lambda = 1$  (б) и  $\lambda = 2$  (в).

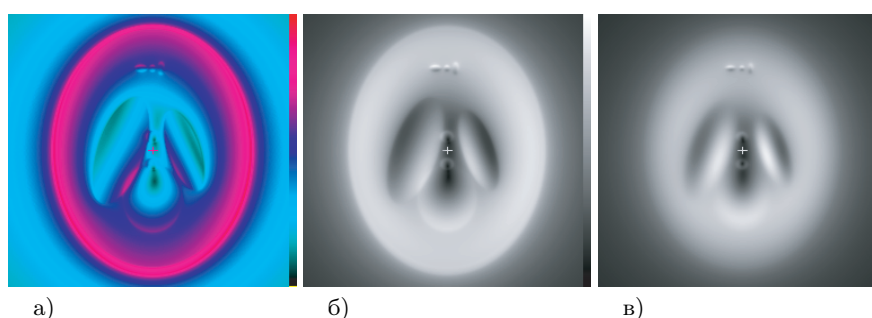


Рис. 5. Применение оператора модуля градиента при  $\lambda = 0.5$  (а),  $\lambda = 1$  (б) и  $\lambda = 2$  (в).

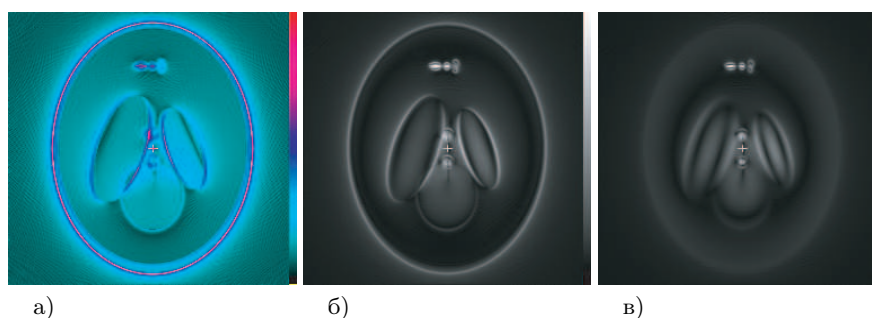


Рис. 6. Двукратное применение оператора модуля градиента при  $\lambda = 0.5$  (а),  $\lambda = 1$  (б) и  $\lambda = 2$  (в).

В следующих экспериментах использовался “параболический фантом Шеппа-Логана”, условно называемый “потенциал 3” и являющийся модификацией “потенциала 2”. На рисунке 7 представлены результаты дважды примененного к преобразованию Радона от “потенциала 3” оператора Вайнберга. Рисунок 8 содержит результаты применения двукратного и трехкратного оператора модуля градиента.

Численные эксперименты показали хорошие результаты визуализации линий разрывов функций и разрывов их производных; линий разрывов векторного поля. Оператор модуля градиента, в целом, оказался более гибким инструментом, нежели оператор Вайнберга, который не отличает разрывов функции и ее производных. Применение двойного дифференцирования, неважно с помощью какого оператора, в некоторых тестах приводило к возникновению на линиях разрывов  $\delta$ -функции и, тем самым, к более сильной сингулярности. Тем не менее, реализация описанных подходов позволила визуально определять и такой тип сингулярности.



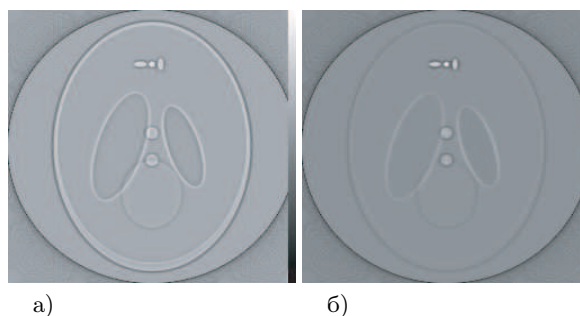


Рис. 7. Двукратное применение оператора Вайнберга при  $\lambda = 0.5$  (а) и  $\lambda = 1$  (б).

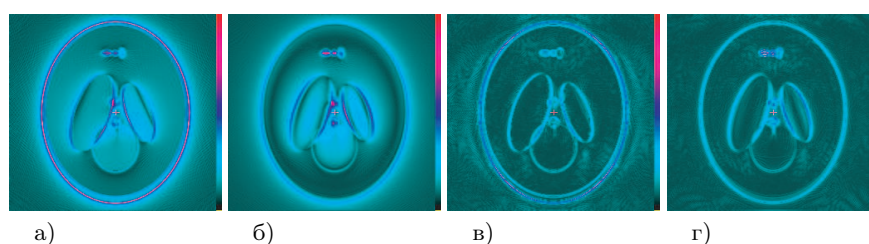


Рис. 8. Применение оператора модуля градиента: двукратное при  $\lambda = 0.5$  (а),  $\lambda = 1$  (б) и трехкратное при  $\lambda = 0.5$  (в),  $\lambda = 1$  (г).

## Список литературы

1. Гельфанд И.М., Граев М.И., Виленкин Н.Я. Обобщенные функции, интегральная геометрия и связанные с ней вопросы теории представлений, выпуск 5. — М.: ГИФМЛ, 1962. — 656 с.
2. Гельфанд И.М., Гончаров А.Б. Восстановление финитной функции, исходя из ее интегралов по прямым, пересекающим данное множество точек в пространстве // ДАН СССР. — 1986. — Т. 290, No. 5 (1986). — С. 1037–1040.
3. Паламодов В.П. Некоторые сингулярные задачи томографии // Математические проблемы томографии / Под ред. И. М. Гельфанда и С. Г. Гиндикина. — М.: Вопросы кибернетики, 1990. — С. 132–140.
4. Vainberg E.I., Kazak I.A., Faingoiz M.L. X-ray computerized back projection tomography with filtration by double differentiation. Procedure and information features // Soviet J. Nondest. Test. — 1985. — Vol. 21. — P. 106–113.
5. Гельфанд И.М., Гиндикин С.Г., Граев М.И. Избранные задачи интегральной геометрии. — М.: Добросвет, 2000. — 208 с.
6. Faridani A., Ritman E.L., Smith K.T. Local tomography // SIAM J. Appl. Math. — 1992. — Vol. 52, No. 2. — P. 459–484.
7. Faridani A., Finch D.V., Ritman E.L., Smith K.T. Local tomography II // SIAM J. Appl. Math. — 1997. — Vol. 57, No. 4. — P. 1095–1127.
8. Louis A.K., Maass P. Contour Reconstruction in 3-D X-Ray CT // IEEE Trans. Med. Imag. — 1993. — Vol. 12, No. 4. — P. 764–769.
9. Михлин С.Г. Многомерные сингулярные интегралы и интегральные уравнения. — М.: ГИФМЛ, 1962. — 256 с.
10. Аниконов Д.С., Ковтанюк А.Е., Прохоров И.В. Использование уравнения переноса в томографии. — М.: Логос, 2000. — 224 с.
11. Аниконов Д.С. Специальная задача интегральной геометрии // Доклады РАН. — 2007. — Т. 415, No. 1. — С. 7–9.
12. Ramm A.G., Katsevich A.I. The Radon transform and local tomography. — CRC Press, Boca Raton, 1996.
13. Quinto E.T. Singularities of the X-ray transform and limited data tomography in  $\mathbb{R}^2$  and  $\mathbb{R}^3$  // SIAM J. Math. Anal. — 1993. — Vol. 24. — P. 1215–1225.
14. Ramm A.G. New methods for finding discontinuities of functions from local tomographic data // J. Inverse and Ill-Posed Problems. — 1997. — Vol. 5, No. 2. — P. 165–175.

15. Деревцов Е.Ю. Некоторые подходы к задаче визуализации сингулярного носителя скалярных, векторных и тензорных полей по томографическим данным // Сиб. Электронные Матем. Известия. — 2008. — Т. 5. — С. 632–646.
16. Sharafutdinov V., Skokan M., Uhlmann G. Regularity of ghosts in tensor tomography // J. Geom. Analysis. — 2005. — Vol. 15, No. 3. — P. 499–542.
17. Derevtsov E.Yu., Pickalov V.V., Schuster T., Louis A.K. Reconstruction of singularities in local vector and tensor tomography // International Conference “Inverse Problems: Modelling and Simulation”, Abstracts. — Fethiye, Turkey, 2006. — P. 38–40.
18. Derevtsov E.Yu., Pickalov V.V., Schuster T. Application of local operators for numerical reconstruction of a singular support of a vector field by its known ray transforms // Journal of Physics: Conference Series. IOP Publishing. — 2008. — Vol. 135. — art. 012035.
19. Деревцов Е.Ю., Пикалов В.В. Восстановление векторного поля и его сингулярностей по лучевым преобразованиям // Сиб. Журн. Вычислительной математики. — 2011. — Т. 14, No. 1. — С. 25–42.
20. Deans S. The Radon Transform and Some of its Applications. — New York, Wiley, 1983. — 294 p.

# Разработка Подсистемы Актуализации Базовых Пространственных Данных по Населенным Пунктам Красноярского Края

А.В. Токарев<sup>1</sup>

<sup>1</sup> Институт вычислительного моделирования СО РАН, г. Красноярск, Россия  
tav@torins.ru

**Аннотация.** Рассматривается задача формирования и актуализации базовых пространственных данных по населенным пунктам региона. Сформулированы основные требования к подсистеме, проанализированы существующие технологии и программные компоненты. Предложена методика актуализации в многопользовательском режиме без прямого подключения к хранилищу пространственных данных. Реализация выполнена на основе программных компонентов с открытым исходным кодом, использовался язык сценариев PHP 5 с фреймворком для разработки веб-приложений Yii. Для хранения данных применялась СУБД PostgreSQL с модулем расширения PostGIS. Решение было успешно внедрено в информационную систему «Банк пространственных данных администрации Красноярского края».

**Ключевые слова:** ГИС, ИПД, актуализация, оцифровка, базовые пространственные данные, геоданные, PostGIS, веб-технологии.

## 1 Введение

По мере широкого распространения геоинформационных технологий во всем мире ценность пространственных данных и осознание их реальной значимости постоянно повышаются, а их использование в разных областях человеческой деятельности расширяется ускоренными темпами. Новые возможности быстрого обмена пространственной информацией, удобного и простого доступа к ней, в особенности по корпоративным и глобальным сетям, обеспечивают принятие более взвешенных решений и более эффективных действий. Постоянный рост инвестиций в ГИС-технологии сопровождается созданием инфраструктур пространственных данных, необходимых для более эффективного управления, устойчивого развития и сохранения нашего мира [1].

В настоящее время многие федеральные органы исполнительной власти, органы исполнительной власти субъектов РФ, органы местного самоуправления, хозяйствующие субъекты активно создают и используют пространственные данные [1,4]. Например, в Красноярском крае, начиная еще с 2006 года, формируется «Банк пространственных данных администрации Красноярского края». Это система информационно-аналитической поддержки деятельности органов государственной власти и местного самоуправления в части решения задач анализа и планирования территориальных аспектов социально-экономического развития [2].

Одной из важных задач является хранение и ведение базовых цифровых картографических материалов территории. Базовые пространственные данные – разрешенные к открытому опубликованию цифровые данные о наиболее используемых пространственных объектах, отличающихся устойчивостью пространственного положения во времени и служащих основой позиционирования других пространственных объектов [3]. На региональном уровне это такие слои, как: железные дороги; автодороги; объекты гидрографии; границы лесных кварталов; границы муниципальных районов; городских округов, городских и сельских поселений; и др.

Обычно формирование базовых пространственных данных происходит поэтапно на основе какой-то первоначальной версии карты. Добавляются новые пространственные объекты, улучшается точность позиционирования и детальность объектов, заполняется атрибутивная информация. Во многих случаях требуется одновременная работа нескольких оцифровщиков над одним массивом данных. Важно обеспечить техническую и информационную поддержку этому процессу.

## 2 Задача актуализации базовых пространственных данных

В работе рассматривается задача формирования и актуализации базовых пространственных данных населенных пунктов некоторого региона. Считаем, что картографический материал – векторный, хранится в «неразрывном» виде и занимают значительный объем на диске. Набор слоев муниципального уровня включает объекты регионального уровня с уточненной метрикой и объекты плана населенного пункта: кварталы, здания, строения; улицы и проезды в населенных пунктах; объекты гидрографии; объекты промышленной, инженерной и социальной инфраструктуры.

Требования к подсистеме актуализации базовых пространственных данных:

- поддержка расширяемого набора слоев разных типов (точки, линии, полигоны) с произвольным набором атрибутивных полей для каждого слоя;
- многопользовательская работа с авторизацией, одновременная оцифровка различных участков карты;
- возможность работы без постоянного подключения к хранилищу пространственных данных;
- хранение истории обновляемых данных, возможность отката к предыдущему состоянию;
- алгоритмы автоматической проверки актуализированных данных (корректность геометрии, правильность заполнения атрибутов, и др.);
- инструменты для ручной проверки качества данных модератором с возможностями обмена сообщениями между участниками работ;
- поддержка различных ГИС пакетов, в которых может выполняться оцифровка объектов (MapInfo, ArcView, QGIS, и др.);
- средства определения и планирования участков карты, нуждающихся в актуализации;
- средства оценки выполненного объема работ каждым оцифровщиком.

## 3 Технологии и программное обеспечение

Были проанализированы существующие технологии и программные компоненты, которые могут быть использованы для решения поставленной задачи.

Открытый геопространственный консорциум (Open Geospatial Consortium, OGS) предложил ряд стандартов, которые упрощают использование пространственных данных в распределенной среде – GML, WMS, WFS, WCS, WFS-T, WPS, и др. В настоящее время они получили широкое распространение и поддержку [6].

Для хранения пространственных данных все чаще используются реляционные СУБД. Это дает следующие преимущества:

- хранение геометрии и атрибутов объекта в виде одного кортежа, что позволяет строить реляционные связи между этими объектами и другими данными БД;
- ускорение пространственных запросов вследствие использования индексов по геометрическим полям;

- широкий набор функций для манипулирования пространственными данными и выявления отношений между объектами в SQL-запросах.

Большинство ведущих разработчиков СУБД создали расширения для поддержки геоданных: Oracle Spatial, Microsoft SQL Server, PostgreSQL/ PostGIS, MySQL Spatial, SQLite/ SpatiaLite. Наиболее развитой из открытых СУБД в мире и являющаяся реальной альтернативой коммерческим базам данных является PostgreSQL – свободно распространяемая объектно-реляционная система управления базами данных. Немаловажным преимуществом является наличие дополнительных модулей, которые облегчают работу с пространственными данными. Расширение PostGIS

(<http://www.postgis.org>) позволяет работать с географическими объектами и функциями в базе данных и соответствует OpenGIS стандартам, разработанным Открытым Географическим Сообществом (OGC). Поддерживается более 300 различных функций для работы с векторными данными, пространственные индексы, популярные обменные форматы и многое другое.

Существует ряд приложений, которые реализуют механизм многопользовательского доступа к пространственным данным независимо от типа СУБД. Среди коммерческих решений, к развитым многофункциональным представителям такого класса программного обеспечения относится ESRI ArcGIS и модуль ArcSDE [5]. Среди приложений с открытым кодом можно выделить MapServer (<http://mapserver.org/>) и GeoServer (<http://geoserver.org/>). Это популярные среды для создания картографических web-сервисов, которые могут быть использованы практически на любых платформах (Windows, Unix, Mac OS, и др.), обладают широкими функциональными возможностями, легкостью интеграции с различными СУБД, реализуют спецификации OGC.

Стоит упомянуть и популярную платформу для совместного создания и использования карт – некоммерческий проект OpenStreetMap (<http://openstreetmap.org/>). В нем используется принцип вики, т. е. каждый зарегистрированный пользователь может вносить изменения в карту. Структура пространственных данных отличается от общепринятой модели в ГИС. Используются такие понятия, как узел (node), линия (way), отношение (relation) и теги (tag), через которые задается атрибутивная информация.

Несмотря на существующие технологии и компоненты, готовых решений для реализации поставленной задачи найдено не было.

#### 4 Методика актуализации

Предлагается методика актуализации карты в многопользовательском режиме без прямого подключения к хранилищу пространственных данных. Основные этапы процесса показаны на рисунке Рис. 1.

1. Менеджер формирует задание на оцифровку участка карты. В задании указывается область в виде прямоугольника или полигона. Сначала задание создается неактивным и считается запланированным. При активации задания и назначении исполнителя выполняется проверка на пересечение области оцифровки с другими активными заданиями. Это необходимо для устранения коллизий, когда одни и те же данные редактируются несколькими исполнителями.
2. Во время очередной итерации сервис обработки создает набор слоев для актуализации. Для этого из каждого слоя общей карты вырезаются объекты по границам задания. Если объект частично попадает в область оцифровки – он обрезается. Полученные слои

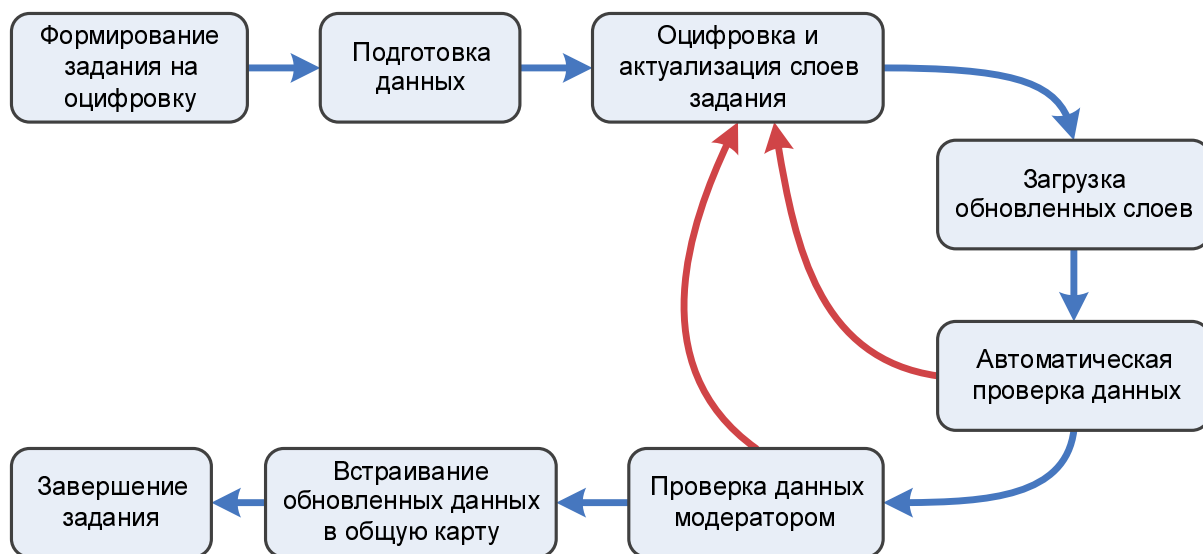


Рис. 1. Схема процесса оцифровки заданий

выгружаются в обменный формат (выбран tab формат MapInfo), добавляется вспомогательный слой с границами задания и проект MapInfo. Файлы подготовленного набора пакуются в zip-архив.

3. Исполнитель скачивает архив задания с сервера и выполняет актуализацию пространственных данных. Работа может выполняться в удобном ГИС пакете благодаря использованию распространенного формата данных.
4. После завершения работы исполнитель загружает результат на сервер в виде zip-архива с аналогичным набором файлов.
5. Сервис обработки выполняет ряд проверок загруженных данных, в том числе:
  - проверка корректности архивного файла и полноты набора файлов для каждого слоя;
  - проверка корректности формата слоев, типа геометрии, набора атрибутов;
  - проверка на выход объектов за границы задания;
  - проверка атрибутов объектов каждого слоя;
  - проверка «сшивки» разрезанных объектов.

Набор проверок можно расширять при необходимости. Если обнаружались ошибки – задание автоматически возвращается исполнителю с набором замечаний.

6. Если данные успешно прошли автоматическую проверку, они отправляются в очередь на проверку модератором. Если у проверяющего появляются замечания – задание отправляется на доработку. Кроме этого, есть возможность отправить комментарии в ленту сообщений задания.
7. Сервис обработки встраивает область с актуализированными слоями в общую карту. Если исходный объект частично пересекал границу области – выполняется обновление части объекта.
8. Выполняется расчет статистики по выполненной работе. Определяется количество добавленных, измененных, удаленных объектов; вычисляется, была ли у объекта изменена геометрия, значения атрибутов или все вместе. На основе полученных показателей в дальнейшем можно оценить трудозатраты оцифровщика.

## 5 Архитектура системы

С технологической точки зрения систему решено строить на основе веб-технологий [6]. Основные ее компоненты:

1. Основная база данных содержит базовые сущности системы, необходимые для ее функционирования. Взаимодействие пользовательского интерфейса и сервиса обработки данных реализовано на уровне базы данных через программный интерфейс в виде представлений и функций.
2. Хранилище пространственных данных построено на основе сервера PostgreSQL с модулем расширения PostGIS. В нем хранятся как основные слои пространственных данных, так и их история изменений. Физически размещено в отдельных схемах основной базы данных. На уровне базы данных реализованы функции вырезки и вставки пространственных данных, функции автоматической проверки корректности.
3. Буферное файловое хранилище используется для передачи пространственных данных от пользователя в пространственную базу данных и обратно.
4. Пользовательский веб-интерфейс с авторизацией является точкой входа в систему для конечных пользователей – администраторов, менеджеров, оцифровщиков. В зависимости от выданных разрешений интерфейс адаптируется и предоставляет необходимые инструменты.
5. Сервис обработки предназначен для выполнения потенциально ресурсоемких задач в отложенном режиме. В настоящее время в нем выполняется подготовка слоев задания и выгрузка в виде файлов, загрузка результата оцифровки в базу данных, автоматическая проверка данных. Сервис запускается с заданным периодом из планировщика задач операционной системы.

Реализация сервиса обработки выполнялась на языке сценариев PHP 5 с использованием фреймворка Yii (<http://www.yiiframework.com>). Это высокоэффективный PHP-фреймворк для разработки веб-приложений реализующий парадигму MVC. Он основан на компонентной структуре и позволяет максимально применить концепцию повторного использования кода, существенно ускоряя процесс веб-разработки.

## 6 Фрагменты базы данных

Основные сущности концептуальной модели базы данных показаны на рисунке 2. Центральной сущностью является проект, который определяет набор слоев, которые будут актуализироваться. Для каждого слоя подготавливается метаописание, включающее в себя наименование таблицы в базе данных, наименование файла для экспорта/импорта, набор его атрибутов.

В рамках проекта работают пользователи, каждому из которых назначена одна или несколько ролей. Роли определяют объем представляемой информации и набор доступных операций в системе.

Модификация пространственных данных в рамках отдельных заданий, которые характеризуются границами области оцифровки. Текущее состояние задания отражается полем статус оцифровки. В ленте сообщений сохраняются замечания менеджера, вопросы и ответы оцифровщика, а также выявленные ошибки после автоматической проверки данных.

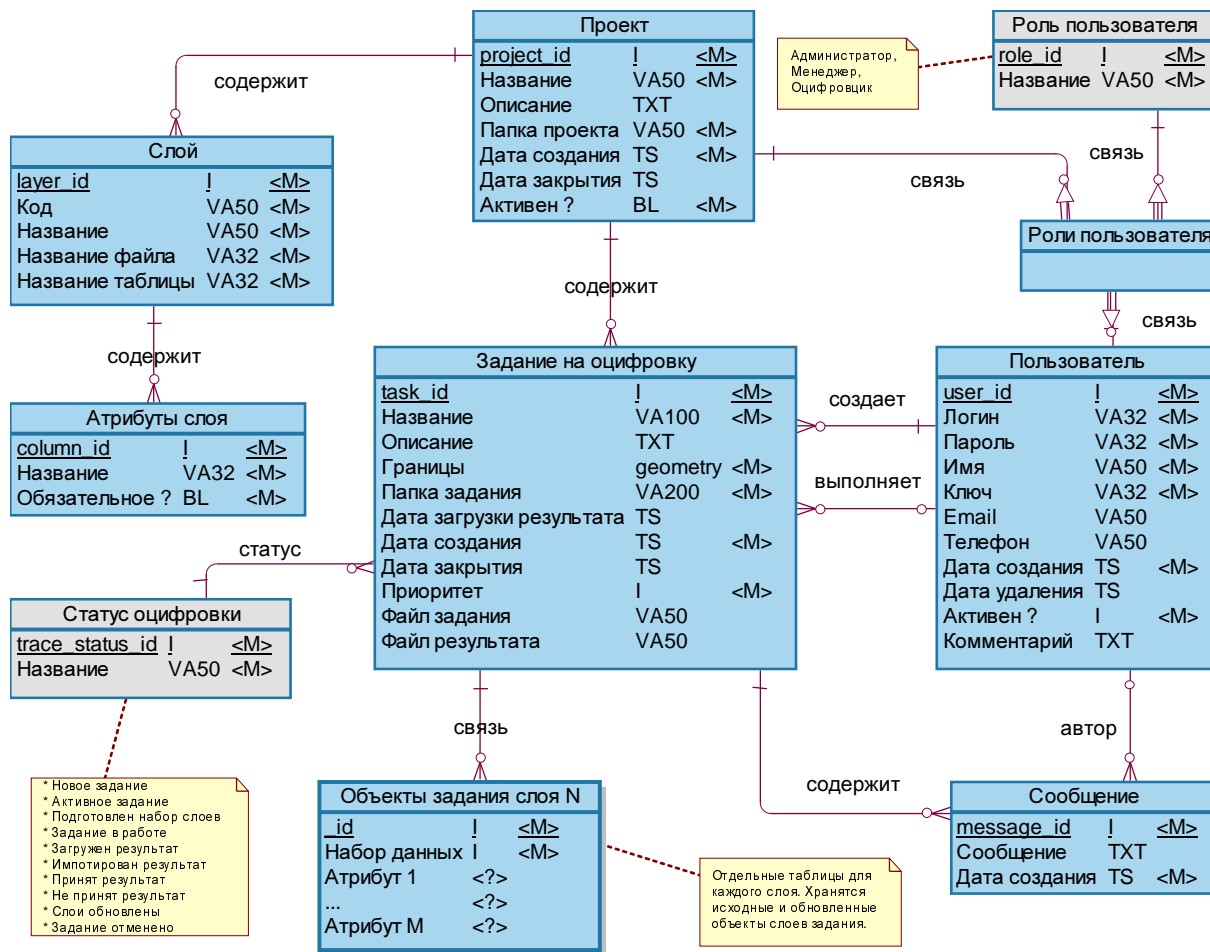


Рис. 2. Фрагмент концептуальной модели базы данных

## 7 Структура хранения пространственных данных

Для того чтобы актуализировать пространственных данные, их необходимо как-то хранить и модифицировать. Традиционные подходы предполагают хранение слоев в виде файлов. Однако такое решение имеет ряд недостатков в контексте поставленной задачи, поэтому решено использовать технологии хранения пространственных данных в реляционной СУБД. Выделены два подмножества данных – слои проекта и слои заданий (рис. 3).

*Слои проекта* содержат текущий слепок пространственных данных. Каждый слой хранится в отдельной таблице. Структура таблицы слоя следующая:

- Глобальный идентификатор объекта (*\_gid*). Это искусственный ключ, генерируется автоматически последовательностью базы данных при первичном наполнении слоя и при добавлении новых объектов. В дальнейшем это поле используется для связывания объектов из слоя заданий и объектов слоя проекта. Если значение *\_gid* в слое задания не пустое – обновляется соответствующий объект в проекте, иначе создается новый и генерируется уникальный идентификатор.
- Поле с геометрией (*geom*) содержит пространственную информацию объекта. Для оптимизации выполнения запросов на поле создается пространственный индекс.
- Пользовательские атрибуты пространственных объектов (атрибут *<...>*). Набор атрибутов произволен и определяется исходными данными.



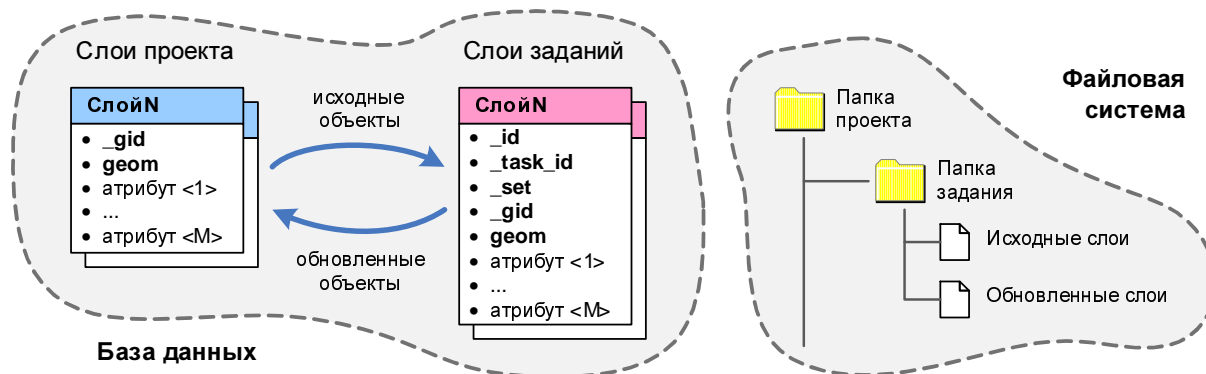


Рис. 3. Структура хранения данных

*Слой заданий* необходимы для хранения наборов пространственных данных для всех заданий. Структура таблиц похожа структуре таблиц для слоев проекта.

- Идентификатор объекта (`_id`), первичный ключ.
- Идентификатор задания (`_task_id`), связывает задание и пространственные объекты.
- Тип набора данных (`_set`). Одна и та же структура данных обеспечивает хранение как исходный набор вырезанных объектов (`_set=1`) для конкретного задания, так и набор обновленных объектов (`_set=2`).
- Глобальный идентификатор объекта (`_gid`).
- Поле с геометрией (`geom`) содержит пространственную информацию объекта.
- Пользовательские атрибуты пространственных объектов (атрибут `<...>`).

Для удобства наборы таблиц слоев проекта и задания имеют одинаковые наименования и разнесены по отдельным схемам базы данных. Передача пространственных данных от пользователя в пространственную базу данных и обратно организовано через файловую систему и выполняется служебным сервисом обработки данных. Для каждого проекта создается индивидуальная корневая папка в файловой системе, в которой размещаются папки заданий. При формировании задания сервис создает архив с исходными данными в этой папке, сюда же попадает архив с обновленными слоями после загрузки результата оцифровки.

## 8 Заключение

Предложенная методика актуализации пространственных данных и разработанные компоненты были успешно внедрены в информационную систему «Банк пространственных данных администрации Красноярского края» [2]. В результате решена задача детальной оцифровки около 1500 малых населенных пунктов на территории края, которая выполнялась в многопользовательском режиме командой из 10 оцифровщиков.

Разработка построена на основе свободно распространяемых технологий и программного обеспечения, что существенно расширяет области ее применения. Прорабатывается встраивание подсистемы актуализации в геоинформационный интернет-портал ИВМ СО РАН (<http://gis.krasn.ru/>).

## Список литературы

1. Радионов Г.П., Загоровский В.И. Инфраструктура пространственных данных Российской Федерации: опыт, технологии, особенности // ArcReview. – 2012. – № 4(63).

2. О.Э. Якубайлик, С.С. Замай, С.А. Артемьев, Ю.В. Глазырин, А.А. Гостева, В.В. Желиховская, А.А. Кадочников, К.В. Мальцев, В.Г. Попов, А.С. Пятаев, А.В. Токарев. Банк пространственных данных администрации Красноярского края // Проблемы информатизации региона. ПИР-2007: Материалы десятой Всероссийской научно-практической конференции. В 2 т. Т. 1. – Красноярск: Сиб. федер. ун-т; Политехн. ин-т, 2007. – С. 188-195.
3. Концепция создания и развития инфраструктуры пространственных данных Российской Федерации – одобрена распоряжением Правительства РФ от 21.08.2006 г. № 1157-р.
4. Иванов К.А. Волонтерские геоинформационные системы в управлении муниципалитетами и регионами / Иванов К.А. [и др.] // Информационное общество. – 2014. – № 3. – С. 10-19.
5. Стрельцов И. Технология ArcSDE – что это? / И. Стрельцов // ArcReview. – 2001. – №4(19). – С. 11–12.
6. О.Э. Якубайлик, А.А. Кадочников, А.В. Токарев. Программно-технологическое обеспечение геопространственных веб-приложений // Инфраструктура научных информационных ресурсов и систем. Сборник избранных научных статей. Труды Четвертого Всероссийского симпозиума (С.-Петербург, 6–8 октября 2014 г.). Под ред. Е.В. Кудашева, В.А. Серебрякова. В 2-х тт. Т. 2. – ISBN 978-5-19601-104-3. – М.: ВЦ РАН, 2014. – С. 107-115.