Технология обработки слабоструктурированных документов

В.Б.Барахнин, А.М.Федотов

Институт вычислительных технологий СО РАН, Новосибирск

Электронные библиографические базы (Current Contents, Zentralblatt MATH, Реферативные журналы) содержат краткие аннотации "бумажных" документов без ссылок на электронные (обычно более подробные, чем аннотация) версии документов;

Обычные ИПС научной тематики (каталоги) работают с документами после непосредственного согласования форматов метаданных;

Системы, использующие концепцию Semantic Web, могут работать только с документами, у которых значения метаданных суть элементы заданных словарей;

Поисковые системы общего назначения работают с любыми документами, но слабо используют анализ метаданных, что приводит к низкой пертинентности найденных документов;

Предлагаемый подход основан на **автоматизированном извлечении метаданных из слабоструктурированных документов** (т.е. документов, у которых значения атрибутов метаданных, как содержательных, так и структурных, не являются элементами заданных словарей), что значительно упрощает процесс работы с информацией, представленной во внешних системах (в т.ч. интернете), включая механизм актуализации информации.

Подходы к обработке слабоструктурированных документов

Алгоритмы обработки слабоструктурированных документов описаны в работах как зарубежных (В.Крещенди, Дж.Мекка, П.Мериальдо, 2001; А.Сауджет, Ф.Азавант, 2001, и др.), так и отечественных авторов (И.Некрестьянов, Е.Павлова, 2002, И.В.Некрасов, В.О.Толчеев, 2005, и др.).

Основная идея таких алгоритмов базируется, как правило, на анализе их html-разметки.

Нерешенные задачи:

1. Из документов извлекаются лишь те данные, которые присутствуют непосредственно в них самих, хотя в удаленных библиографических базах данных зачастую содержатся более подробные описания документов, которые сделаны экспертами, включающие коды классификатора (обычно отсутствующие в самих документах), ключевые слова и др.
2. Координатное индексирование **русскоязычных** документов, как правило, ограничивается **однословными терминами** (что объясняется отсутствием соответствующих алгоритмов ввиду сложности морфологического анализа русских словосочетаний). Весьма характерно замечание И.В.Некрасова и В.О.Толчеева в статье ("Информационные технологии" № 11, 2005), посвященной обзору методов классификации, основанных на индексировании составными ключевыми терминами: "В данной статье будут рассматриваться **англоязычные** библиографические публикации по научно-технической проблематике".

Указанные проблемы значительно снижают качество каталогизации документов, что, в конечном итоге, затрудняет их поиск, а также уменьшает объем данных, используемых для получения новой информации и знаний.

Этапы обработки слабоструктированных документов

0. Создание онтологии (тезауруса) предметной области

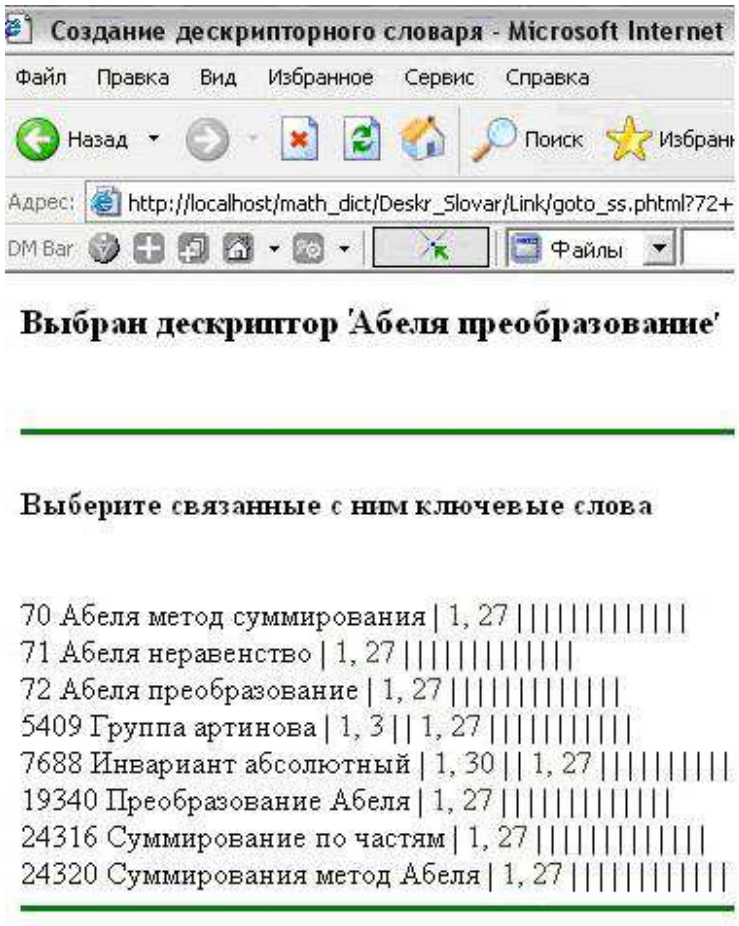
1. Извлечение метаданных из документов

2. Получение недостающих метаданных

3. Индексирование документов

4. Кластеризация документов (для классификации и поиска “по аналогии”

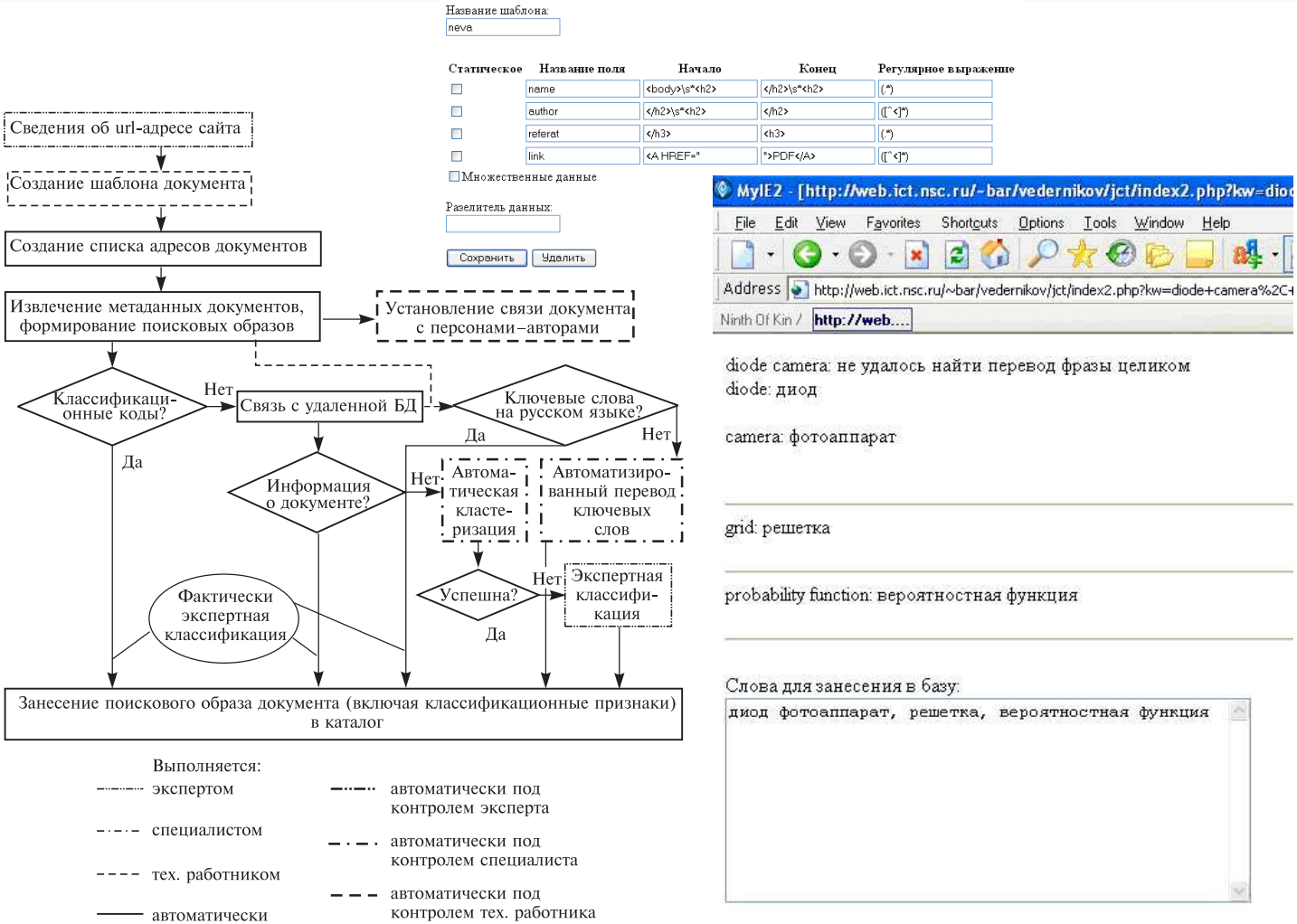
Алгоритм автоматизации создания онтологии на основе предметного указателя энциклопедии



Идея предложенного алгоритма (2003 г.) использовать ссылки на страницы для установления связей между терминами оригинальна.

В работе (2005 г.) создателей РуТез также говорится об использовании предметного указатели энциклопедий, но без механизма автоматизации установления связей.

Алгоритм автоматизации извлечения метаданных и формирования поисковых образов документов



Отличительной особенностью предложенного алгоритма автоматизированного извлечения метаданных от коммерческих пакетов, является **возможность получения недостающих метаданных из удаленных БД**.

Обычно при индексировании русскоязычных научных текстов используется подход, основанный на извлечении одиночных ключевых слов (см., напр., О.В.Пескова, 2008), что упрощает морфологический анализ, но имеет серьезные теоретические недостатки: возможность ложной координации, ложных синтагматических связей и др. (А.И.Михайлов и др., 1968).

Нами предложен алгоритм автоматической индексации текстов с использованием в качестве ключевых слов терминов–словосочетаний (*ключевых терминов*) из заданного лексического словаря.

Отличительной особенностью алгоритма является использование наряду с традиционным индексом

`номер текста - позиция в тексте - номер слова из лексического словаря,`

оригинального индекса

`номер термина - позиция слова в термине - номер слова из лексического словаря.`

Отличия известных аналогов: стимер Яндекса, извлекающий словосочетания, опирается на синтаксис, но не на семантику. С другой стороны, алгоритмы проекта “Микрокосмос” (США), В.А.Тузова, В.А.Фомичева, И.С.Циликова и др., предназначенные для проведения семантического анализа текстов на уровне, близком к восприятию естественно-языковых текстов человеком, весьма сложны в практической реализации.

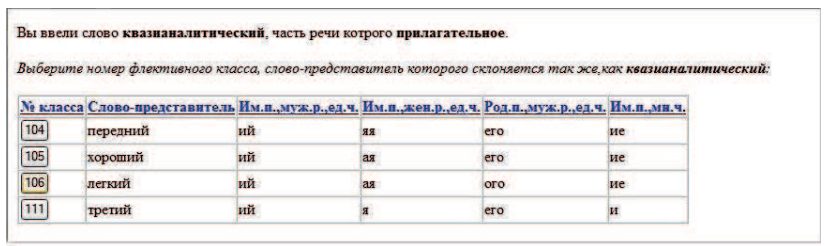
Автоматизация пополнения базового лексического словаря

Для автоматизации работы по пополнению базового лексического словаря было построено веб-приложение, автоматически генерирующее все словоформы заданного слова (существительного или прилагательного) русского языка. В основе работы веб-приложения лежит алгорит Г.Г.Белоногова (1979) разбиения слов языка на флективные классы.

Однако размеры надклассов, на которые разбиты флективные классы, зачастую слишком велики для выбора (сущ.м.р.одуш. – 19, сущ.м.р.неодуш. – 16, сущ.ж.р.одуш. – 8, сущ.ж.р.неодуш. – 12, сущ.ср.р. – 11, прил. – 12).

Предложена модификация алгоритма Г.Г.Белоногова, состоящая в автоматическом анализе окончаний нормализованной словоформы внутри каждого надкласса, что приводит к значительному уменьшению количества элементов, из которых предстоит сделать выбор:

- сущ.м.р.одуш.: $12 + 2 + 2 + 2 + 1$
- сущ.м.р.неодуш.: $10 + 3 + 3$
- сущ.ж.р.одуш.: $4 + 3 + 1$
- сущ.ж.р.неодуш.: $6 + 4 + 2$
- сущ.ср.р.: $5 + 5 + 1$
- прил.: $4 + 4 + 2 + 1 + 1$



Вы ввели слово **квазианалитический**, часть речи котрого **прилагательное**.

Выберите номер флективного класса, слово-представитель которого склоняется так же,как ***квазианалитический:***

№ класса	Слово-представитель	Им.п.,муж.р.,ед.ч.	Им.п.,жен.р.,ед.ч.	Род.п.,муж.р.,ед.ч.	Им.п.,мн.ч.
104	передний	ий	яя	его	ие
105	хороший	ий	ая	его	ие
106	легкий	ий	ая	ого	ие
111	третий	ий	я	его	и

Количество объектов-альтернатив в подавляющем большинстве случаев доведено до рекомендуемого когнитивной психологией ($\sim$ 9 альтернатив). Для сущ.м.р. ситуация неулучшаема.

В алгоритме решения аналогичной задачи из работы Е.А.Каневского (2001) классы словоформ определялись без учета теоретических исследований Г.Г.Белоногова путем непосредственного анализа типов окончаний. Это приводит к появлению более 10 тыс. классов для сущ. и 2,5 тыс. классов для прил.

Мера сходства между документами d_1 и d_2 вычислялась по формуле

$$\mu(d_1, d_2) = \sum \alpha_i \mu_i(d_1, d_2),$$

где i — номер элемента (атрибута) библиографического описания, α_i — весовые коэффициенты, $\sum \alpha_i = 1$, $\mu_i(d_1, d_2)$ — мера сходства по i-му элементу. Если шкалы — номинальные, то мера сходства по i-й шкале определяется следующим образом: если значения i-х атрибутов документов совпадают, то мера близости равна 1, иначе — 0. Если значения атрибутов составные, то $\mu_i = n_{i1}/n_{io}$, где $n_{io} = \max\{n_{io}(d_1), n_{io}(d_2)\}$, $n_{io}(d_j)$ — общее количество элементов, составляющих значение i-го атрибута документа d_j, n_{i1} — количество совпадающих элементов.

Отличительные особенности предложенной методики:

- **использование нескольких шкал**: авторы; ключевые слова (авторские); текст аннотации, из которого извлекаются ключевые термины, — что особенно важно при работе не с полными текстами документов, а с аннотациями (нередко используется только 1 шкала: ключевые слова из текста — Кондратьев, 2006; Пескова, 2008);
- в качестве извлеченных ключевых терминов **рассматриваются словосочетания** (обычно при работе с русскоязычными документами извлекаются одиночные слова);
- используется **апостериорный выбор продукционных правила** для определения весовых коэффициентов при шкалах.

Проведен сравнительный анализ ряда алгоритмов кластеризации: метод клик, метод Роккио, жадный алгоритм, FRiS-алгоритм (Загоруйко, 2007).

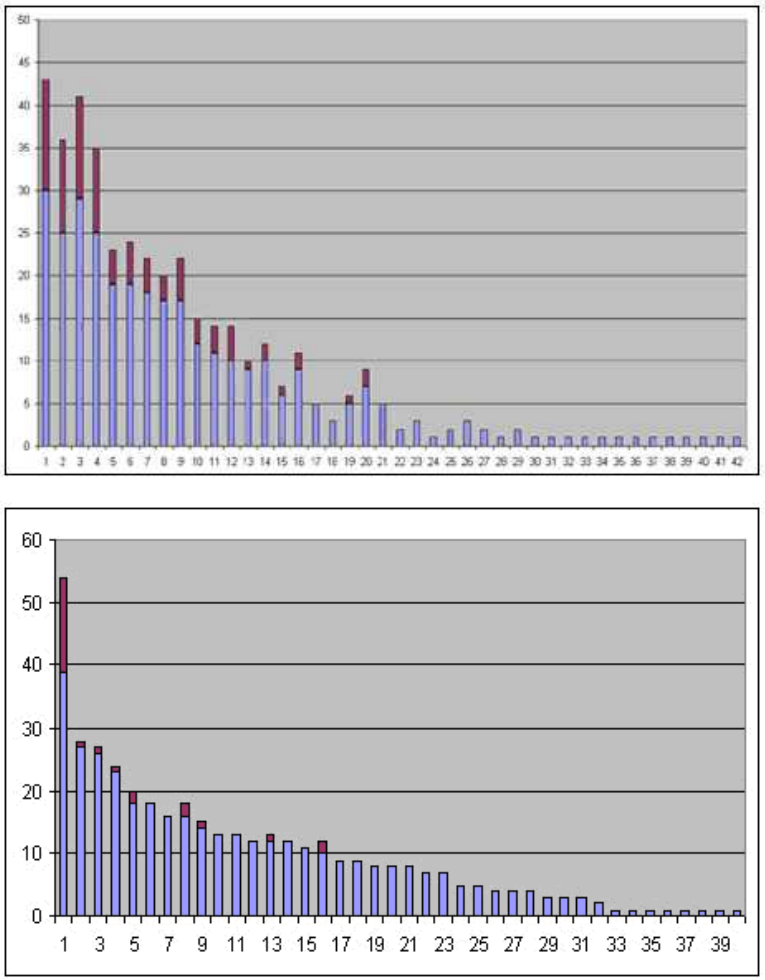
Результаты сравнения алгоритмов кластеризации

Тестирование алгоритма проводилась на электронной БД “Сибирского математического журнала” (порядка 700 записей).

В качестве единственной шкалы для вычисления меры на пространстве документов использовались коды MSC2000 (обычно документу приписано 3 или более кодов). Поскольку совпадение данных кодов для группы документов является объективным критерием совпадения тематики данных документов, то такую меру можно считать образцовой. Если коды классификатора центроида кластера содержались в числе кодов классификатора 2-го уровня данного документа, то мы полагали, что документ был отнесен к кластеру правильно.

Было проведено сравнение классического жадного алгоритма и FRiS-алгоритма. Погрешность классификации в первом случае составила 12 %, во втором – 4 %.

Таким образом **установлено, что среди рассмотренных алгоритмов оптимальным для данной задачи является FRiS-алгоритм**.



Сравнение жадного и FRiS алгоритмов

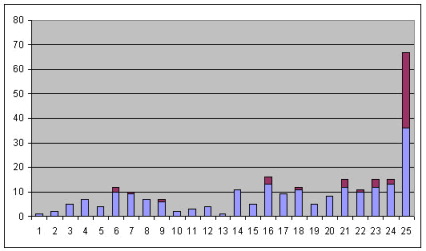
Экспериментальное определение весовых коэффициентов

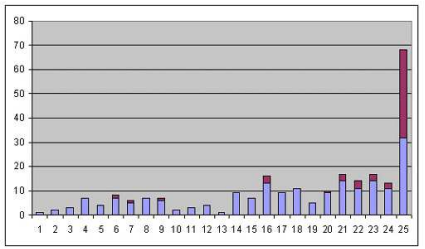
Определим весовые коэффициенты при использовании нескольких шкал: **авторы; ключевые слова (авторские); текст аннотации, из которого извлекаются ключевые термины**.
Коэффициенты определены экспериментально на коллекции статей “Сиб.Мат.Журнала” (250 записей). Критерий выбора коэффициентов — наибольшее сходство с результатом кластеризации по мере, базирующейся на кодах классификатора MSC2000.
В итоге **получены апостериорно выбираемые правила для определения весовых коэффициентов на основании предполагаемой достоверности данных**:

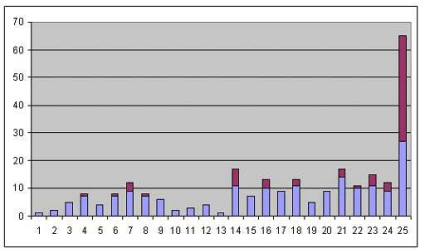
1. Если каждый из документов d_1 и d_2 имеет более 2 авторов и как минимум 2/3 из них совпадают, то коэффициент при атрибуте “авторы” равен 1.
2. Если каждый из документов d_1 и d_2 содержит более 3 ключевых слов и как минимум 3/4 этих слов совпадают, то коэффициент при атрибуте “ключевые слова” равен 1.
3. Если каждый из документов d_1 и d_2 содержит более 4 ключевых терминов в аннотации и как минимум 3/5 этих терминов совпадают, то коэффициент при атрибуте “аннотация” равен 1.
4. Если условия ни одного из правил 1–3 не выполнены, то коэффициент при атрибуте “авторы” равен 0,2, а при атрибутах “ключевые слова” и “аннотация” равен 0,4.

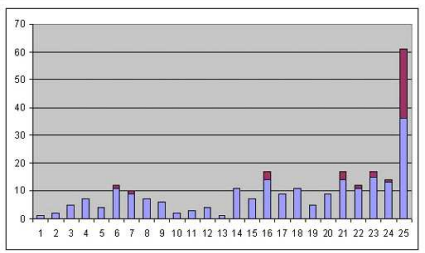
Последний пункт также получен в результате сравнительного анализа ряда наборов коэффициентов:

1) “авторы” – 0,2, “ключевые слова” – 0,4, “аннотация” – 0,4
2) “авторы” – 0,2, “ключевые слова” – 0,6, “аннотация” – 0,2
3) “авторы” – 0,4, “ключевые слова” – 0,4, “аннотация” – 0,2
4) Контрольная кластеризация, основанная на кодах MSC2000





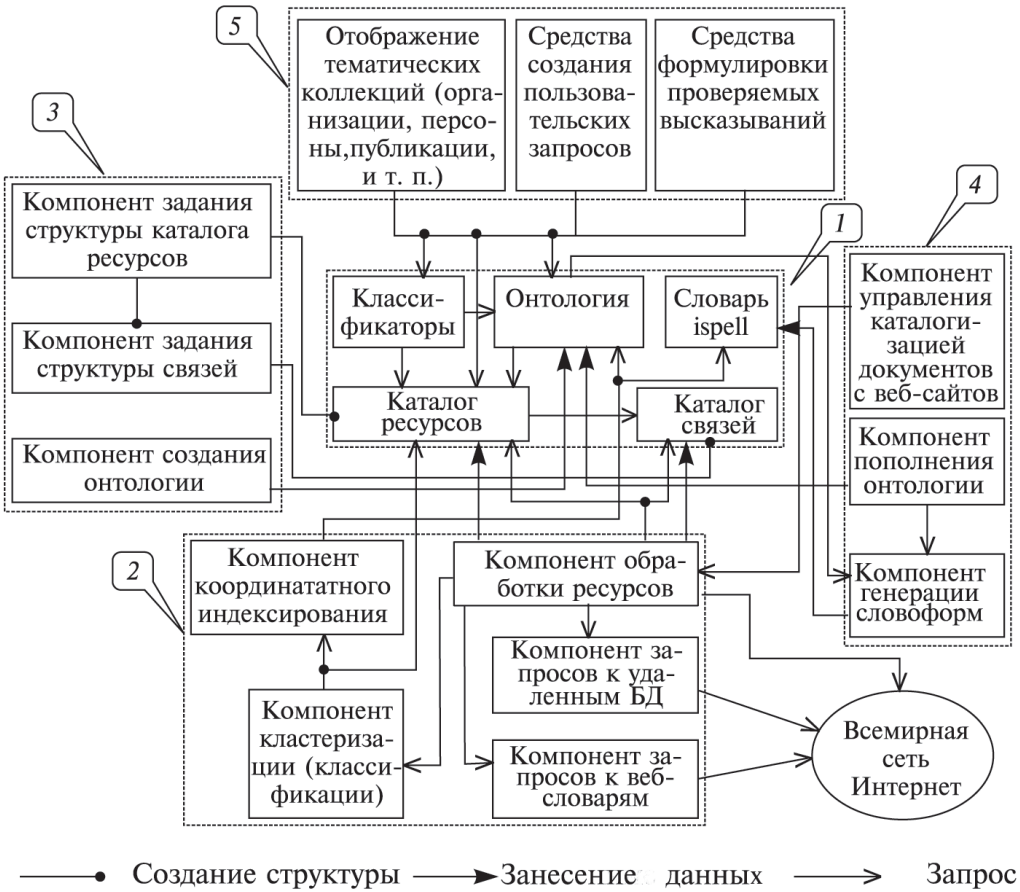




Погрешности:
1) 18%, 2) 23%,
3) 26%, 4) 15%.

Сравнительно высокая погрешность объясняется, во-первых, использованием не полных текстов, а аннотаций, и, во-вторых, наличием в выборке документов из близких разделов 2-го уровня MSC2000 (раздел 76Mxx “Основные методы в механике жидкости” частично поглотил разделы 76Bxx “Несжимаемая невязкая жидкость” и 76Nxx “Сжимаемые жидкости и газовая динамика”).

Функциональная схема программной системы



1 – Хранилище данных
2 – Блок извлечения метаданных из веб-документов
3 – Веб-интерфейс администрирования системы
4 – Веб-интерфейс администрирования данных
5 – Веб-интерфейс пользователя

Спасибо за внимание!