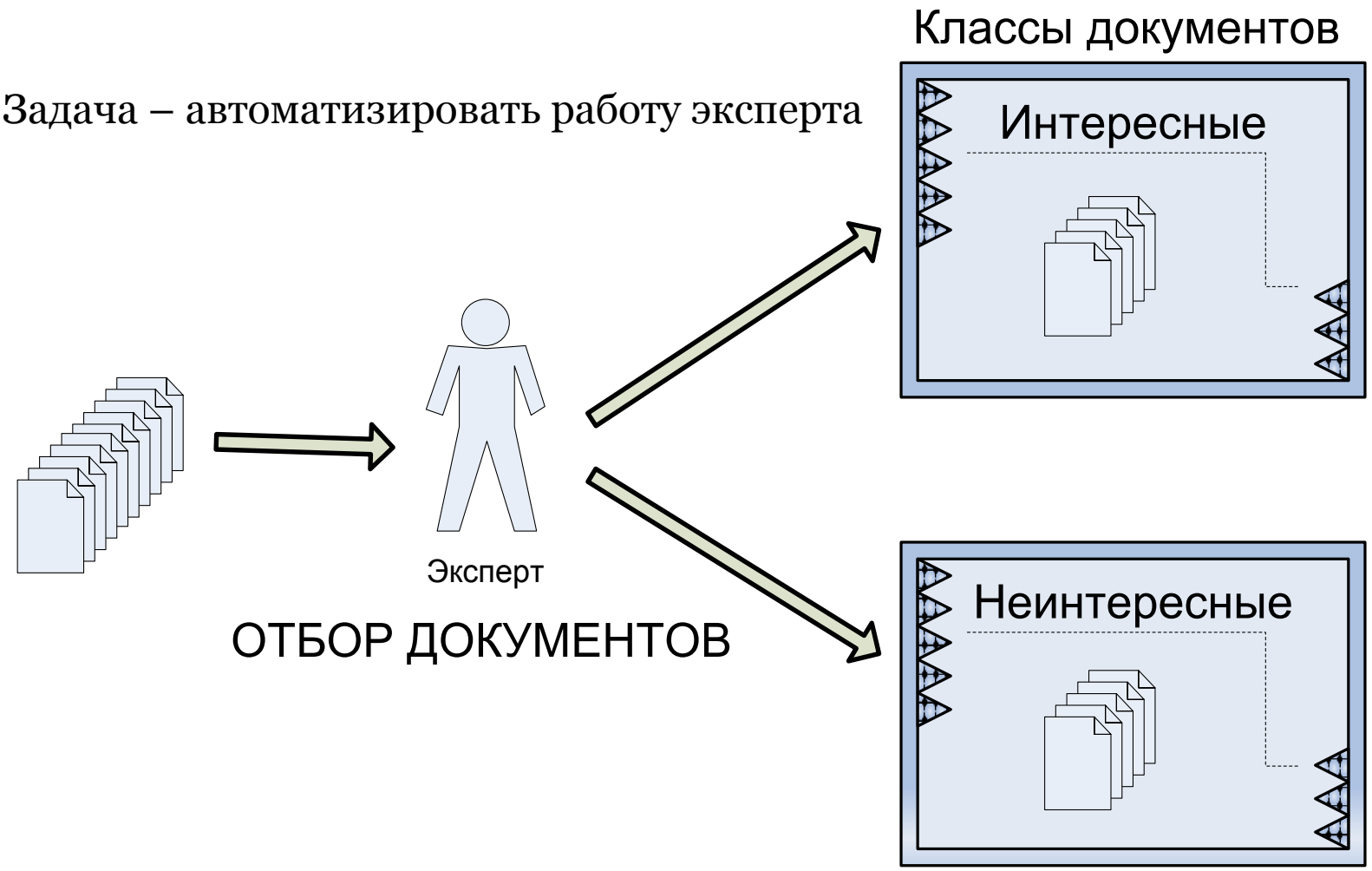
Построение модели поискового образа документа для автоматизации отбора документов

Леонова Ю.В., Федотов А.М.

Институт вычислительных технологий СО РАН, Новосибирск

Отбор документов из поступающего потока документов

Задача – автоматизировать работу эксперта



Способы задания классов

В информационной системе имеется некоторый набор классов, который может быть задан:

- Эмпирически - задание множества классов осуществляется экспертом на основе непосредственного сравнения общих свойств некоторого класса документов.
- Аксиоматически - задание множества классов осуществляется на основе опосредованного анализа некоторого класса документов при помощи ранее выработанных условий или правил, которым должен удовлетворять документ.

Постановка задачи

Необходимо реализовать процедуру автоматизации отнесения произвольного документа к одному или нескольким классам

1. Каждый поступающий в систему документ должен соотноситься с одним или несколькими классами, либо ни с каким классом. Если документ не попадает ни в одну категорию, то он отбрасывается.
2. Процесс отнесения документов должен быть именно автоматическим: отнесение документа происходит без помощи специально обученных экспертов, которые могут определить, попадает тот или иной документ в определенный раздел, никакого подбора правил вручную.

Методы решения

- Для каждой пары <документ, категория> определяется степень принадлежности документа категории, причем эта степень является числом, лежащим в диапазоне от 0 до 1 (чем больше степень принадлежности, тем больше документ подходит этой категории).
- На множестве категорий могут быть заданы теоретико-множественные отношения. Например, множества документов, составляющих категории, могут пересекаться или не пересекаться, т.е. один и тот же документ может принадлежать нескольким категориям. Могут быть заданы составные категории – категория может быть подмножеством другой категории и т.п.

Методы решения. Формализация

- Множество категорий (классов)

$$\mathfrak{C} = \{c_1, \ldots, c_{|\mathfrak{C}|}\}$$

- Множество документов

$$\mathfrak{D} = \{d_1, \ldots, d_{|\mathfrak{D}|}\}$$

- Неизвестная целевая функция – отображение, которая классифицирует документ и выдает степень принадлежности

$$\Phi: \mathfrak{C} \times \mathfrak{D} \to [0,1]$$

Требуется определить данную целевую функцию.

Методы решения. Задание целевой функции

- Наиболее популярный способ – задание меры сходства на множестве документов.

Для каждого класса задается документ-центроид, значение функции – мера сходства между поисковым образом документа и образом центроида класса

Признаковое пространство

- Определение признаков – формирование поискового образа документа (ПОД) – вектор характерных признаков документа, используемый в дальнейшем для принятия решений по работе с документом.
- ПОД представляет собой многомерный вектор в пространстве признаков документа, характеризующих смысловое содержание исходного документа.

Признаковое пространство

- *Признаком* называется отображение $f: \mathfrak{D} \to \mathbb{D}_f$, где $\mathbb{D}_f$ – множество допустимых значений признака.
- Если заданы признаки $f_1, \dots f_n$, то вектор $\mathbf{d} = (f_1(d), \dots, f_n(d))$ называется *признаковым описанием* документа $d \in \mathfrak{D}$.
- Признаковые описания допустимо отождествлять с самими документами.
- При этом множество $\mathfrak{D} = \mathbb{D}_{f_1} \times \dots \times \mathbb{D}_{f_n}$ называют *признаковым пространством*.

Признаковое пространство

Типы признаков:

- *бинарный* признак: $\mathbb{D}_f = \{0,1\}$ (пол человека);
- *ранжированный* признак: $\mathbb{D}_f \in [0,1]$ (мера близости);
- *номинальный* признак: $\mathbb{D}_f$ – конечное множество (место проживания, профессия, работодатель);
- *порядковый* признак: $\mathbb{D}_f$ – конечное упорядоченное множество (образование, занимаемая должность);
- *количественный* признак: $\mathbb{D}_f$ – множество действительных чисел (год издания, возраст, стаж работы);
- *иерархический* признак: $\mathbb{D}_f$ – конечное частично-упорядоченное множество (термины в тезаурусе).

Схемы представления поисковых признаков документа

4 вида признаков:

- Базовые (Dublin Core)
- Основные (ГОСТ)
- Дополнительные
- Сложные (вычислимые)

Схемы представления поисковых признаков документа

- Базовые признаки присутствуют в описании каждого документа, задаются стандартом Dublin Core. Набор элементов метаданных Дублинского ядра (Dublin Core Metadata Element Set; DCMES) состоит из 15 элементов метаданных

Title – название;
Creator – создатель;
Subject – тема;
Description – описание;
Publisher – издатель;
Contributor – внёсший вклад;
Date – дата;

Type – тип;
Format – формат документа;
Identifier – идентификатор;
Source – источник;
Language – язык;
Relation – отношения;
Coverage – покрытие;
Rights – авторские права.

Схемы представления поисковых признаков документа

- В зависимости от типа документа (элемент Type) могут использоваться различные схемы расширения признаков. Например, для документа типа «публикация» для расширения набора поисковых признаков могут использоваться библиографические стандарты и ГОСТы.
- В библиотечном деле составление библиографического описания (совокупности библиографических сведений о документе) выполняется по правилам, установленным ГОСТ 7.1-841 «Библиографическое описание документа».
- В настоящее время насчитывается около 300 поисковых признаков. Поисковые признаки разделяются на два типа: основные и дополнительные.

Схемы представления поисковых признаков документа

- Основные поисковые признаки описывают наиболее полные сведения, необходимые для идентификации и поиска издания – название издания, автор. Некоторая часть этих признаков дублирует схему Dublin Core.
- Дополнительные признаки используются для расширения поиска документа по элементам. Дополнительные признаки содержат указание на классификатор, жанр, язык документа, формат и место хранения и т.п.

Схемы представления поисковых признаков документа

Для поисковой системы большинство дополнительных характеристик несущественны, в отличие от характеристик основного типа. Для реализации эффективного поиска основной интерес представляют классификационные дополнительные признаки, содержащие коды библиотечно-библиографических классификаторов.

Основные библиотечно-библиографические классификаторы:

- система библиотечно-библиографической классификации (ББК);
- универсальная десятичная классификация (УДК);
- десятичная классификация М. Дьюи (ДКД);
- единая классификация литературы для книгоиздания (ЕКЛ).

В основу библиотечно-библиографических классификаций положен тематический принцип.

Схемы представления поисковых признаков документа

- Сложные признаки основаны на использовании вычислимых характеристик текстов, например авторство текста документа. В общем случае текст отображается в вектор вычисленных для него параметров, каждый из которых объективно характеризует некоторый набор особенностей текста. Таким образом, текст графически отображается в некоторую точку n-мерного пространства. При такой формализации автор также может быть представлен в виде аналогичного вектора параметров – этим вектором будет вектор текстов, написанных данным автором.

Другими сложными признаками могут быть:

- пристатейные списки литературы;
- список персон, встречающихся в тексте документа;
- список географических объектов, встречающихся в тексте документа;
- список ключевых слов из тезаурусов и т.д.

Классификация признаков

Построение меры сходства выполняется в зависимости от типа признака

Бинарные

- Creator – создатель;
- Subject – тема;
- Publisher – издатель;
- Contributor – внёсший вклад;
- Language – язык

Классификация признаков

Порядковые

- Date – дата;
- Type – тип;
- Классификаторы – УДК, ДКД (можно находить расстояние)

Ранжированные

- пристатейные списки литературы;
- список персон, встречающихся в тексте документа;
- список географических объектов, встречающихся в тексте документа;

Иерархические

- список ключевых слов из тезаурусов

Шкалирование и метризация пространства признаков

Для применения алгоритмов распознавания и классификации необходимо метризовать пространство признаков. В целях обеспечения однородности и метризуемости пространства признаков необходимо шкалировать это пространство.

Другими словами, для сравнения документов необходимо ввести формальную меру сходства (различия) объектов, в терминах которой и будут сравниваться документы между собой, а точнее – будут сравниваться поисковые образы документов.

Шкалирование и метризация пространства признаков

Среди шкал обычно выделяют 3 основных типа:

1. **количественные**, когда признаки измеряется в некоторой шкале, например, длины, веса, скорости и т.п. и числа, которыми выражаются значения количественных признаков, показывают на или во сколько раз различаются объекты по данному признаку и допускают любые арифметические преобразования с ними – количественная шкала;
2. **классификационные**, когда различные проявления признака можно упорядочить только на уровне наличия или отсутствия, т.е. на уровне классов объектов, например, пол, место жительства, профессия и т.п. – шкала наименований

Шкалирование и метризация пространства признаков

3. **качественные**, когда проявления признаков можно естественным образом упорядочить по их значениям (сила ветра в баллах, оценки на экзаменах или в спортивных соревнованиях и т.п.). При этом градации качественных признаков обычно выражаются упорядоченным рядом слов (плохо, удовлетворительно, хорошо, отлично) или целыми числами (шкала твердости минералов), обычно увеличивающимися с возрастанием степени проявления соответствующего признака. Однако такие числа (точнее цифры-символы) нельзя суммировать или умножать, поскольку эти цифры выражает только место (имя) градации в их последовательности – шкала порядков;

Шкалирование и метризация пространства признаков

Информационная система “Модели изменения биосферы на основе баланса углерода (по натурным и спутниковым данным и с учетом вклада бореальных экосистем)”.

Для метризации пространства признаков необходимо определить наиболее значимые признаки, на основе которых строятся метрики. Качественные и классификационные шкалы строятся на основе классификационных признаков (атрибутивных) – УДК, ББК и т.п. Данные признаки, как правило, имеют большой вес.

Шкалирование и метризация пространства признаков

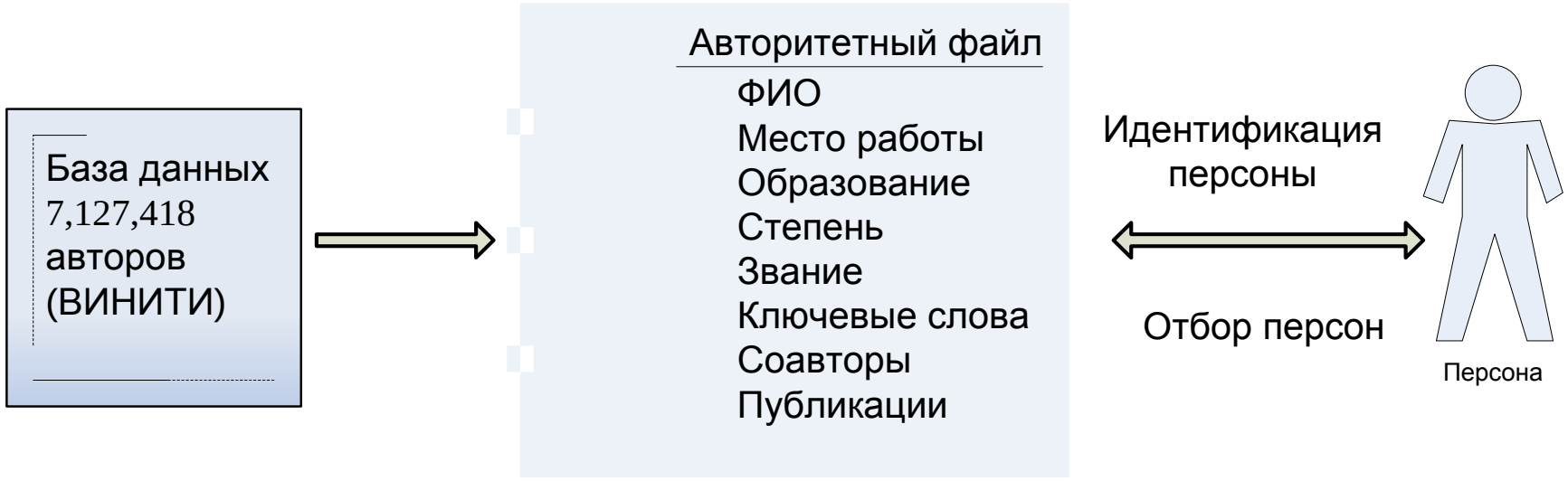
Применение количественных шкал для построения метрик возможно на основе эмпирических данных.

1. Распределение ключевых терминов в тексте документа.
2. Распределение фамилий персон, упоминаемых в документах. Часто коллектив авторов, работающий в определенной области, в текстах статьи ссылается на других персон, работающих в данной области. Поэтому данный признак может использоваться в качестве классификационного.
3. Распределение территорий, упоминаемых в документах. Этот признак, прежде всего, важен для задач, имеющих экологическую тематику. Упоминание территорий в документах привязывает работу к географическому местоположению и таким образом также может являться классификационным признаком.

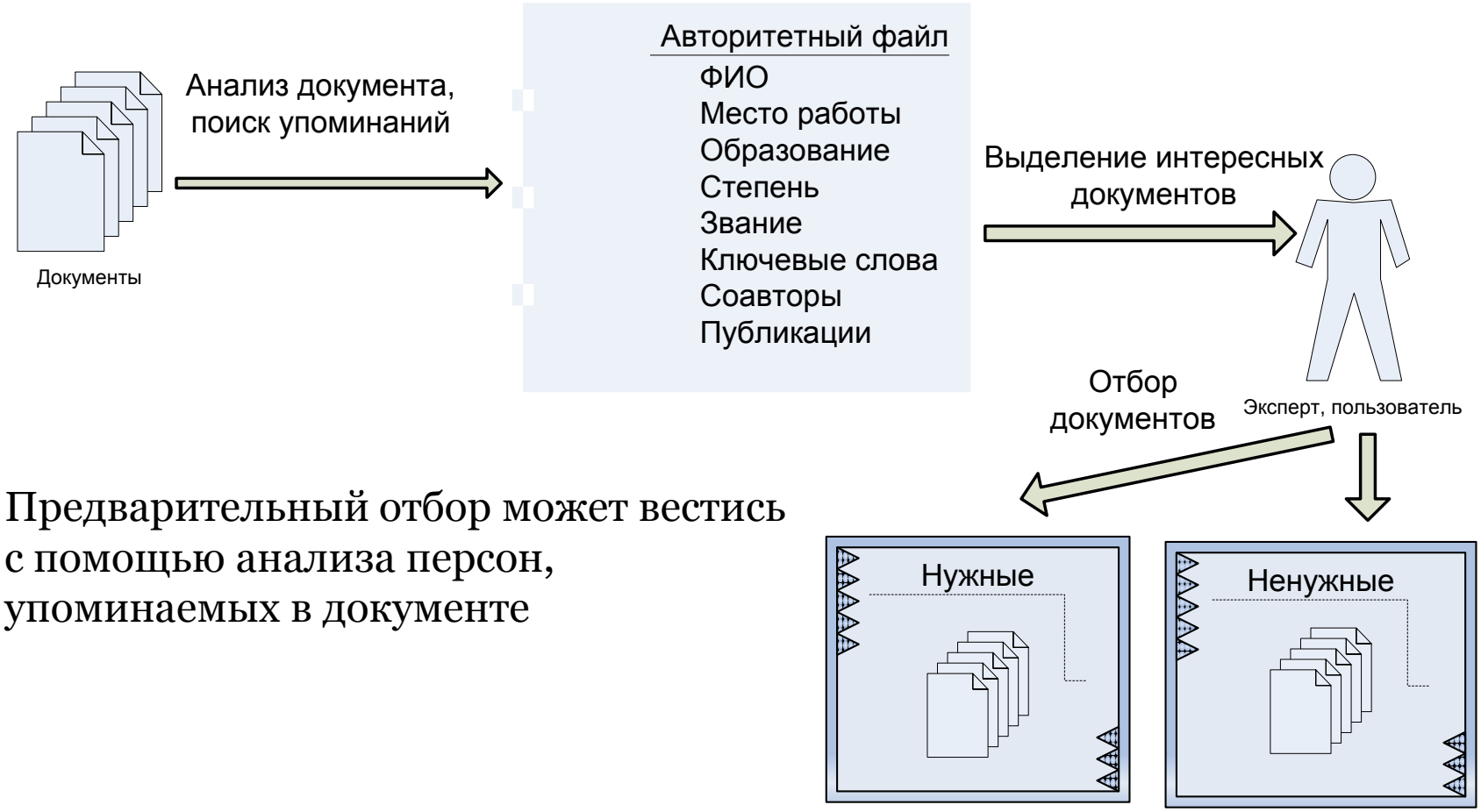
Шкалирование и метризация пространства признаков

- Среди ранжированных признаков выбраны 2: распределения упоминания персон и географических объектов в документах.
- Распределения данных признаков решают задачу отнесения документа к классу, интересующих пользователя документов.
- Для проведения классификации необходимо определять распределение персон, для чего необходимо иметь базу данных персон – авторитетный файл

Построение авторитетного файла



Алгоритм отбора интересных документов



Предварительный отбор может вестись с помощью анализа персон, упоминаемых в документе