Ю. В. Леонова¹, А. М. Федотов^{1,2}, О. А. Федотова^{2,3}

¹ *Институт вычислительных технологий СО РАН
пр. Академика Лаврентьева, 6, Новосибирск, 630090, Россия*

² *Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия*

³ *Государственная научно-техническая библиотека СО РАН
ул. Восход, 15, Новосибирск, 630090, Россия*

juli@ict.nsc.ru

О ПОДХОДЕ К КЛАССИФИКАЦИИ АВТОРЕФЕРАТОВ ДИССЕРТАЦИЙ ПО ТЕМАМ

Рассматривается метод тематической классификации авторефератов диссертаций. Для этого используется специально построенная мера близости документов, учитывающая специфику предметной области. В качестве шкал для определения меры предлагается брать характеристики структурных атрибутов описания авторефератов (научная новизна, положения, выносимые на защиту и т. п.). Значения весовых коэффициентов в формуле для вычисления меры близости определяются предполагаемой апостериорной достоверностью данных соответствующей шкалы.

Ключевые слова: классификация, мера близости, весовые коэффициенты.

Введение

Задача поиска и выделения информации является одной из важнейших задач, возникающих при построении информационных систем. Пользователь ищет не документы как таковые, а сокрытые в них факты или содержимое для удовлетворения своих информационных потребностей. Универсальным подходом, решающим эту задачу, является тематическая классификация документов. К тому же, как было отмечено Дональдом Кнудом (см. [1]), поскольку задачи поиска и классификации документов являются двойственными, то нам достаточно рассмотреть модель классификации документов, наиболее адекватно отражающую особенности работы с информацией.

Наиболее распространенным вариантом классификации библиографических ресурсов является фасетная классификация, теория построения которой формализована индийским библиотекведом Ш. Р. Ранганатаном [2]. Объекты классифицируются одновременно по нескольким независимым друг от друга признакам (фасетам). Применительно к цифровым документам (и электронным ресурсам вообще) в качестве фасетов выступают элементы метаданных, которые включают и ключевые термины.

Кратко фасетная классификация состоит в следующем.

Определяется множество тематических классов документов. Класс имеет несколько фасетов, соответствующих различным аспектам классифицируемого понятия. Из коллекции изучаемых документов выписываются все существенные термины, которые группируются по фасетам, т. е. объединяются в соответствующие классы.

Леонова Ю. В., Федотов А. М., Федотова О. А. О подходе к классификации авторефератов диссертаций по темам // Вестн. НГУ. Серия: Информационные технологии. 2017. Т. 15, № 1. С. 47–58.

Термин, принадлежащий некоторому фасету, называется его фокусом. При индексировании документов их содержание выражается последовательностью фокусов.

В работе предложена формальная модель фасетной классификации, основанная на индексации документов ключевыми терминами, выбираемыми из некоторого словаря. Предложен и апробирован алгоритм классификации, основанный на специально построенной мере близости, учитывающий специфику классификационной модели. В качестве базы для экспериментов выбрана коллекция, состоящая из 4 000 авторефератов. Мы остановили свой выбор на авторефератах диссертаций по следующим причинам: практически одинаковый объем и наличие структуры, позволяющее изучить иерархию фасетов.

Модель классификации

Простейшая формальная модель классификации документов с использованием метаданных (ключевых терминов) документов выглядит следующим образом [3; 4]. Рассмотрим коллекцию документов $D = \{d_i\}$. Любой документ d_i из коллекции D представляется как $d_i = \langle m_i^{j,k} \rangle$, где $m_i^{j,k}$ – значения элементов метаданных T^j ; k – количество значений

(с учетом повторений) соответствующего элемента метаданных в описании документа. Рассмотрим подмножество метаданных T_C , определяющее набор классификационных признаков документов. Для фиксированного элемента метаданных T^j , где $T^j \subset T_C$, заранее определяются подмножества T_i^j множества значений этого элемента метаданных (указанные подмножества могут, вообще говоря, пересекаться). Множество документов разбивается на классы эквивалентности, соответствующие различным значениям или же заранее выбранным подмножествам множества значений этого элемента метаданных.

Будем считать два документа толерантными, если у них совпадает значение хотя бы одного из элементов метаданных, входящих в T_C (напомним, что толерантность – отношение, которое обладает свойствами рефлексивности и симметричности, но, вообще говоря, может не обладать, в отличие от отношения эквивалентности, свойством транзитивности). Каждое такое значение порождает класс толерантности [5].

Рассмотрим всевозможные сочетания значений элементов метаданных, входящих в T_C . Множества документов, обладающие одинаковым набором значений, суть ядра толерантности, которые служат классами эквивалентности на множестве документов. С содержательной точки зрения этой ситуации соответствует вхождение некоторого раздела классификатора в раздел более высокого уровня, когда оба этих раздела учитываются при описании пространства толерантности (разумеется, можно и не учитывать раздел более низкого уровня при определении толерантных элементов, но тогда мы будем иметь дело с пространством толерантности, отличным от первоначального).

Таким образом, поисковое предписание, содержащее подмножества метаданных, определяющего набор классификационных признаков и сочетаний значений этих метаданных при помощи логических операций, определяет конкретное ядро толерантности на множестве документов, которое и выдается пользователю в качестве ответа на его информационный запрос. На множестве классов толерантности также можно, в свою очередь, ввести отношение толерантности, при этом толерантными считаются классы, имеющие хотя бы один общий документ. Такая конструкция оказывается полезной, например, для организации поиска документов «по аналогии».

Мера близости

Предлагаемый подход к построению меры близости (или меры сходства) [6], используемой для классификации документов, основан на понятии толерантности документов [4].

Ограничимся рассмотрением только ключевых терминов, агрегированных по типам признаков. Количественная характеристика меры близости определяется на множестве документов D следующим образом:

$$m: D \times D \rightarrow [0, 1],$$

причем функция m в случае полного сходства принимает значение 1, в случае полного различия – 0. Рассмотрим два документа d_1 и d_2 .

Пусть $T = \{t_i\}_{i=1}^M$ – упорядоченный (каким-либо образом, например, лексиграфически) список ключевых терминов, входящих в оба документа, с учетом повторений (где M – общее количество ключевых терминов). Вычисление меры близости осуществляется по следующей формуле:

$$m(d_1, d_2) = \sum_{i=1}^M \alpha_i m_i(d_1, d_2),$$

где i – номер элемента метаданных (ключевого термина); $m_i(d_1, d_2)$ – мера близости по i -му элементу (иными словами, по i -й шкале), α_i – весовые коэффициенты. Поскольку в описываемой ситуации практически все шкалы – номинальные (состоящие из дискретных текстовых значений), то мера сходства по i -й шкале определяется следующим образом: если значения i -х элементов документов совпадают, то мера близости равна 1, иначе 0.

Весовые коэффициенты должны удовлетворять следующим условиям:

$$\sum_{i=1}^M \alpha_i = 1, \quad \alpha_i = \alpha_j,$$

если значение термина t_i совпадает с значением термина t_j .

Пусть $P = \{p_k\}_{k=1}^N$ – список уникальных ключевых терминов, входящих в оба документа, M_k – число повторений термина p_k . Тогда меру близости можно переписать:

$$m(d_1, d_2) = \sum_{k=1}^N (\alpha_k \cdot M_k) (m_k / M_k),$$

где

α_k – весовой коэффициент, соответствующий значению термина p_k ;

m_k – число совпадений термина p_k в документах d_1 и d_2 .

Мы получаем новые весовые коэффициенты $\beta_k = \alpha_k \cdot M_k$, которые уже характеризуют конкретный ключевой термин.

Не трудно видеть, что

$$\sum_{k=1}^N \beta_k = 1.$$

Отметим, что мы здесь автоматически получаем, что весовой коэффициент пропорционален частоте встречаемости термина. Кроме того, при задании меры можно принять во внимание тот факт, что значения весовых коэффициентов β_k определяются предполагаемой апостериорной достоверностью данных соответствующей шкалы и в определенных случаях один из коэффициентов может быть увеличен с пропорциональным уменьшением остальных. Например, полное (или даже «почти полное») совпадение значений какого-либо атрибута документа d_1 и документа d_2 может быть более весомо в случае, когда количество значений этого атрибута в документе d_1 достаточно велико (по сравнению со случаем, когда документ d_1 имеет всего одно значение).

Построение схемы взвешивания

Документ представляется в виде множества ключевых терминов (термов). Множество всех термов $T = \{t_0, \dots, t_{|T|}\}$. С каждым термом $t_i \in T$ сопоставляется некоторый вес ω_{ij} , $0 \leq \omega_{ij} \leq 1$, характеристика (действительное число) встречаемости слова в документе $d_j \in D$.

Учет разделов. Документ может содержать разделы, имеющие разную значимость с точки зрения вклада в тематическое сравнение двух документов, например, в автореферате «по-

ложения, выносимые на защиту» и «апробация работы». Для учета веса каждого раздела в меру близости добавляются весовые коэффициенты:

$$m(d_1, d_2) = \sum_{r=1}^R \psi_r m_r(d_1, d_2),$$

где R – число разделов; ψ_r – весовой коэффициент, учитывающий априорную значимость r -го раздела, $0 \leq \psi_r \leq 1$, $m_r(d_1, d_2)$ – мера близости по r -му разделу.

Рассмотрим наиболее популярную схему взвешивания TF-IDF.

Схема взвешивания TF (term frequency – частота термина). Каждому встретившемуся в документе термину присваивается вес, который зависит от количества появлений этого термина в данном документе. Таким образом, оценивается важность термина t_i в пределах отдельного документа d_j .

Пусть fr_{ij} – число вхождений термина t_i в документ d_j . Тогда частота термина определяется как

$$TF(t_i, d_j) = \frac{fr_{ij}}{\sum_i fr_{ij}},$$

где $0 \leq i \leq |T|$, $0 \leq j \leq |D|$. Отметим, что эта характеристика уже присутствует при построении меры близости.

Схема взвешивания IDF (inverse document frequency – обратная частота документа) – инверсия частоты, с которой термин встречается в документах коллекции. Учет IDF уменьшает вес широкоупотребительных слов. Для каждого уникального термина в пределах конкретной коллекции документов существует только одно значение IDF:

$$IDF(t_i, D) = \frac{|D|}{|(d_i \subset t_i)|},$$

где, $|D|$ – количество документов в коллекции, $|(d_i \subset t_i)|$ – количество документов, в которых встречается t_i (когда $fr_i \neq 0$), $0 \leq i \leq |T|$.

Априорный вес некоторого термина пропорционален количеству употребления этого термина в документе и обратно пропорционален частоте употребления термина в других документах коллекции. Кроме того, априорный вес термина зависит от экспертной оценки его значимости.

Этапы классификации

На первом этапе автоматической классификации текстов авторефератов выполняется преобразование текста к виду, пригодному для использования алгоритмом классификации. Производится индексация документа, которая определяет содержание документа и выражается последовательностью фокусов, и априорное взвешивание терминов общее для всей коллекции документов.

На втором этапе при помощи обучения на примерах строится классифицирующая функция Φ . Качество классификации зависит как от метода априорного взвешивания терминов документов, так и от алгоритма классификации, применяемого на втором этапе. Методы классификации неспецифичны для задачи классификации текстов и могут использоваться для других задач, например для распознавания образов.

Таким образом, задачу автоматической классификации документов можно разбить на два этапа:

- предварительная обработка документа – текст документа преобразуется в индексированное представление, т. е. выполняется индексирование текста документа;
- классификация – выполняется обучение на множестве документов или классификация документа. Отметим, что для классификации и обучения возможно использование единого метода предобработки текстов.

Предварительная обработка текстов документов

Процесс получения индексированного представления для текста документа называется индексацией документа. Индексация выполняется в два этапа.

1. Извлечение термов (Term extraction) – на этом этапе выполняется поиск и отбор наиболее значимых терминов во всем множестве документов. Результатом данного этапа является множество термов T , используемое для получения весовых характеристик документов.

2. Взвешивание термов (Term weighting) – определяется значимость термина для данного документа. Вес терминов задается специальной весовой функцией.

Рис. 1. Схема выполнения индексации документа

Рассмотрим подробнее этап извлечения терминов (рис. 1).

1. *Графематический анализ*. Отсеиваются все символы, не являющиеся буквами, (например, html-теги и знаки препинания).

2. *Лемматизация*. При построении классификатора текстов не имеет смысла различать формы (спряжение, склонение) слова, поскольку это приводит к чрезмерному разрастанию словаря, увеличивает ресурсоемкость, снижает скорость работы алгоритмов.

Лемматизация заключается в приведении каждого слова к его нормальной форме. В русском языке нормальными формами являются:

- существительные – именительный падеж, единственное число;
- прилагательные – именительный падеж, единственное число, мужской род.

Для лемматизации применяются специальные программы – лемматизаторы, принцип работы которых основан на использовании грамматического словаря со всеми формами слов. Основной недостаток лемматизации – трудоемкость создания грамматических словарей, что обуславливает их неполноту, особенно в области специальной терминологии, представляющей наибольший интерес для тематической классификации.

3. *Уменьшение размерности признакового пространства* (Dimensionality reduction) – процесс сокращения признакового пространства, выполняется удаление слов, не представляющих полезность для классификатора.

- *Отбрасывание стоп-слов*. В первую очередь из вектора терминов, на основе заранее составленного списка, удаляются семантически нейтральные слова, которые часто встречаются в текстах различной тематики (стоп-слова), и их можно отбросить. В качестве стоп-слов могут выступать распространенные предлоги, союзы, суффиксы, причастия, междометия и частицы, некоторые глаголы и прилагательные, местоимения, числительные, наречия. Единого списка русских стоп-слов не существует, обычно

их число невелико – от нескольких сотен до нескольких тысяч. Отбрасывание стоп-слов может заметно сократить длину некоторых обрабатываемых текстов.

Также множество стоп-слов можно пополнить терминами, нетипичными для предметной области, в рамках которой выполняется классификация.

Если текст имеет небольшой размер, то удаления семантически нейтральных слов будет достаточно для сокращения размерности признакового пространства. В текстах большого размера, как правило, имеются слова близкие по смыслу – синонимы, однокоренные слова. Данные термины можно объединять в группы – искусственные термины.

- *Отбрасывание редких слов.* Если слово, встречается в тексте редко, например один раз, то полагаем, что данное слово не имеет большой значимости для данного текста, и его можно отбросить.

4. *Выделение ключевых терминов.* Обычно в качестве терминов используются отдельные слова, встречающиеся в документе. Это может приводить к искажению или потере смысла, который, например, заключается во фразеологизмах, являющихся неделимыми словарными единицами с точки зрения лингвистики. Поэтому при обработке авторефератов вместо отдельных слов выделяют словосочетания (ключевые термины), специфичные для заданной предметной области.

Далее полагаем, что словарь T составлен после предварительной обработки всех документов из множества D и может содержать как отдельные слова, так и фразы. Элементы словаря $t \in D$ будем называть ключевыми терминами.

Описание алгоритма классификации

Алгоритм определения тематической принадлежности текста автореферата основан на четырех процедурах: LemmatizeText, ParseText, FindThemeCover и CalculationRelevance. Рассмотрим их подробнее.

1) LemmatizeText – упрощенный морфологический анализ, или лемматизация, выполняет поиск лексемы из множества T словарных лексем (рис. 2) по префиксу словоформы.

Процедура: LemmatizeText (word).

Вход: word – слово.

Выход: термин в нормальной форме $L \subseteq T$ однословных терминов словаря в нормальной форме, соответствующих word.

2) ParseText – поиск многословных терминов.

Процедура: ParseText(words).

Вход: words – последовательность слов, соответствующих обобщенному шаблону, описывающего словосочетания тезауруса.

Обобщенный шаблон связывает группы слов по общности морфологических признаков, например: [Прил] + [Сущ]

Выход: найденный термин тезауруса.

Первые два этапа формируют терминологическое покрытие текста – упорядоченное множество терминов тезауруса, найденных в данном тексте.

3) FindThemeCover строит тематическое покрытие текста, формирует модель текста и данные для классификации.

Процедура: FindThemeCover(t).

Вход: $t \in T$ – терминологическое покрытие текста с приписанными весами данным терминам в тезаурусе.

Выход: тематическое покрытие $Y \subseteq t \times T$ – отношение, которое связывает темы из T с терминами из t .

4) CalculationRelevance вычисляет релевантности тематик и выбирает наиболее подходящие тематики.

Вход: Y – тематическое покрытие запроса.

Выход: множество $\{\langle \tau_1, v_1 \rangle, \dots, \langle \tau_n, v_n \rangle\}$ тематик $\tau_i \in T$ с приписанными им мерами близости $v_i \in [0, 1]$.

РАСПОЗНАВАНИЕ ИЗОБРАЖЕНИЙ	РЕЧЕВАЯ КОМАНДА	ФОНЕТИЧЕСКАЯ ХАРАКТЕРИСТИКА ЗВУКА РЕЧИ
ОБРАБОТКА ЦИФРОВЫХ ИЗОБРАЖЕНИЙ	РАСПОЗНАВАНИЕ РЕЧИ	АВТОМАТИЧЕСКАЯ ТРАНСКРИПЦИЯ РУССКОЙ УСТНОЙ РЕЧИ
ИКОНИКА	СЛОВАРЬ КОМАНД	АКУСТИЧЕСКИЙ КОРРЕЛЯТ ЗВОНКОСТИ
АВТОМАТИЧЕСКИЙ МЕТОД ОБРАБОТКИ И РАСПОЗНАВАНИЯ ИЗО...	РАСПОЗНАВАНИЕ ГОЛОСОВЫХ КОМАНД	МЕСТОПОЛОЖЕНИЕ ФОРМАНТ
СИСТЕМА ОБРАБОТКИ ИЗОБРАЖЕНИЙ	РЕЧЕВОЙ СИГНАЛ	КАРТОГРАФИРОВАНИЕ
АЛГОРИТМ АНАЛИЗА ИЗОБРАЖЕНИЙ	АЛГОРИТМ РАСПОЗНАВАНИЯ РЕЧИ	ПРОСТРАНСТВЕННО-ВРЕМЕННОЕ МОДЕЛИРОВАНИЕ
АЛГОРИТМ АВТОМАТИЧЕСКОГО АНАЛИЗА ИЗОБРАЖЕНИЙ	КАЧЕСТВО РАСПОЗНАВАНИЯ РЕЧИ	ПРОСТРАНСТВЕННАЯ СТРУКТУРА
РАЗРАБОТКА АЛГОРИТМА АНАЛИЗА ИЗОБРАЖЕНИЙ	ШУМОВАЯ СОСТАВЛЯЮЩАЯ	ГИС-ТЕХНОЛОГИЯ
ПОИСК ИЗОБРАЖЕНИЯ	РАСПОЗНАВАНИЕ СЛИТНОЙ РЕЧИ	ЛАНДШАФТНЫЙ АНАЛИЗ
КОМПЬЮТЕРНОЕ ЗРЕНИЕ	РАСПОЗНАВАНИЕ КЛЮЧЕВЫХ СЛОВ	ЛАНДШАФТНО-ГЕОФИЗИЧЕСКИЙ МЕТОД
ОБРАБОТКА ГРАФИЧЕСКОЙ ИНФОРМАЦИИ	МЕТОД РАСПОЗНАВАНИЯ РЕЧИ	ПРОСТРАНСТВЕННО-ВРЕМЕННАЯ СТРУКТУРА
РАСПОЗНАВАНИЕ ОБРАЗОВ	АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ РЕЧИ	ГЕОИНФОРМАЦИОННЫЙ АНАЛИЗ
МАШИННОЕ РАСПОЗНАВАНИЕ РУКОПИСНЫХ СИМВОЛОВ	СИСТЕМА ГОЛОСОВОГО УПРАВЛЕНИЯ	ВЕКТОРНЫЙ ФОРМАТ
ПОИСК ИЗОБРАЖЕНИЯ ПО СОДЕРЖАНИЮ	РАСПОЗНАВАЕМЫЕ СЛОВА	ЛАНДШАФТНАЯ ДИФФЕРЕНЦИАЦИЯ
НАХОЖДЕНИЕ ИЗОБРАЖЕНИЯ	РЕЧЕВОЕ ВЫСКАЗЫВАНИЕ	ПРОСТРАНСТВЕННОЕ РАСПРЕДЕЛЕНИЕ
МЕДИЦИНСКОЕ ИЗОБРАЖЕНИЕ	ТРАНСКРИПЦИЯ РЕЧИ	ГЕОИНФОРМАЦИОННОЕ МОДЕЛИРОВАНИЕ
ОБРАБОТКА ИЗОБРАЖЕНИЯ	АВТОМАТИЧЕСКАЯ ОБРАБОТКА РЕЧИ	ПРОСТРАНСТВЕННОГО ...
ПОИСК ГРАФИЧЕСКОГО ФАЙЛА ПО СОДЕРЖАНИЮ	МОДЕЛИРОВАНИЕ РЕЧЕВОГО СИГНАЛА	ИНФОРМАЦИЯ ПРОСТРАНСТВЕННОГО ХАРАКТЕРА
ЦИФРОВОЕ ИЗОБРАЖЕНИЕ	МОДЕЛЬ РЕЧЕВОГО СИГНАЛА	МАССИВ ГЕОИНФОРМАЦИИ
СЖАТИЕ ГРАФИЧЕСКИХ ДАННЫХ	ИНТЕРПРЕТАЦИЯ ЗВУКОВ РЕЧИ	ГЕОИНФОРМАЦИОННОЕ РАЙОНИРОВАНИЕ
ИЗВЛЕЧЕНИЕ ИЗОБРАЖЕНИЯ ИЗ БАЗЫ ДАННЫХ	АКУСТИЧЕСКАЯ МОДЕЛЬ РЕЧЕВОГО СИГНАЛА	ГЕОИНФОРМАЦИОННАЯ СИСТЕМА УПРАВЛЕНИЯ
ПИКСЕЛЬ ИЗОБРАЖЕНИЯ	ИНТЕРПРЕТАЦИЯ ЗВУКОВ	МЕТОДЫ ИНТЕРПОЛЯЦИИ ГЕОПРОСТРАНСТВЕННЫХ ДАННЫХ
ДИСКРЕТНОЕ ОТОБРАЖЕНИЕ	АКУСТИЧЕСКИЙ ПРИЗНАК ЗВУКА	ПРОСТРАНСТВЕННО-ВРЕМЕННАЯ СТРУКТУРА
ХАОТИЧЕСКАЯ ТАСОВКА ПИКСЕЛЕЙ	СЕГМЕНТАЦИЯ РЕЧЕВОГО ПОТОКА	
ХАОТИЧЕСКИЙ РАССЕЯННЫЙ ПИКСЕЛЬ	ВЫДЕЛЕНИЕ АКУСТИЧЕСКИХ КОРРЕЛЯТОВ	
ПРОЦЕСС РАСПОЗНАВАНИЯ	АКУСТИЧЕСКИЙ ДИФФЕРЕНЦИАЛЬНЫЙ ПРИЗНАК	

Рис. 2. Список терминов тезауруса темы 4

Тестирование алгоритма классификации. Методика

Из обучающей выборки удаляются документы рубрики, которые присутствуют при тестировании, но не участвуют в обучении.

Варианты исходов для документа:

- 1) «Прав»: документ («Свой») правильно определился в свою рубрику;
- 2) «Чуж»: действительно «Чужой» документ определился как «Чужой»;
- 3) «Ошиб»: документ определился не в свою рубрику;
- 4) «Св_чуж»: «Свой» документ ошибочно определился как «Чужой»;

5) «Чуж_св»: «Чужой» документ ошибочно попал в какую-то рубрику, т. е. ошибочно определен как «Свой».

Исходы 1 и 2 соответствуют правильной работе алгоритма, остальные – ошибочны.

Оценки:

Точность = Прав / (Прав + Ошиб + Чуж_св);

Полнота = Прав / (Прав + Ошиб + Св_чуж).

Практические результаты. В качестве исходных данных для тестирования алгоритма классификации использовались авторефераты диссертаций по 4 темам: «распознавание образов» (тема 1), «распознавание речи» (тема 2), «геоинформационные системы» (тема 3), «онтологии, описание предметной области» (тема 4). В эталонные наборы для каждой тематики вошли по 30 авторефератов.

Формирование списка ключевых терминов (словаря) является отдельной задачей [4]. Например, словарь ключевых терминов может формироваться экспертом на основе его знаний о предметной области. В нашем случае список сформирован на основе текстов эталонных авторефератов, его объем составил 192 ключевых слов.

Классификация проводилась по следующему алгоритму. Первоначально для каждой темы на основе эталонного набора находился центроид – характерный набор ключевых терминов с весами, который потом использовался для сравнения. Далее вычислялась мера близости проверяемого автореферата к центроиду класса (темы).

Результаты тестирования. На вход системе было подано по 4 000 ранее неизвестных текстов авторефератов. Для классификации использовался весь текст автореферата, из которого выделялись значимые ключевые слова. Мера близости рассчитывалась по выделенным ключевым терминам в словаре для каждой темы. При классификации информация выделялась из следующих разделов автореферата:

- актуальность темы исследования;
- цели и задачи; научная новизна;
- объект и предмет исследования;
- теоретическая и практическая значимость работы;
- методология и методы исследования;
- положения, выносимые на защиту;
- степень достоверности и апробация результатов.

Для каждого раздела вычислялась мера близости, итоговая мера близости вычислялась как среднее значение.

Тестирование алгоритма проводилось в трех режимах: 1) классификация по разделам автореферата; 2) классификация по разделам с проверкой многозначности терминов (если термин принадлежит нескольким темам, то проверяется тематическая принадлежность соседних терминов); 3) классификация по всему тексту автореферата без выделения разделов.

Принадлежность автореферата теме определяется по превышению порога близости между тестируемым авторефератом и центроидом темы.

Экспериментально установлено, что если пороговое значение меры близости превышает 0,83, то все отобранные авторефераты относятся к данной теме, при этом число не отнесенных авторефератов, посвященных данной теме, было не выше 5 %. Если пороговое значение меры близости не превышает 0,17, то автореферат не относится к данной теме.

На рис. 3, 4 представлены результаты точности, полноты и F -меры, полученные при тестировании алгоритма. Видно, что наилучшей точностью нахождения чужих документов обладает метод поиска по тексту автореферата, однако полнота и f -мера – наихудшие. Лучшими характеристиками обладает метод поиска по разделам с условием многозначности.

Наихудший параметр точности соответствует Теме 3 – геоинформационные системы, что обусловлено присутствием некоторых терминов, как «пространственное распределение», «пространственная структура» и т. п., в текстах химической направленности. Дополнение словаря химическими терминами позволит повысить точность классификации.

Рис. 3. Нахождение «чужих» документов: a – точность; $б$ – полнота; $в$ – F -мера

Условные обозначения:

- оценка классификации по разделам
- оценка классификации по разделам с условием многозначности
- оценка классификации по тексту

Рис. 4. Нахождение и классификация «своих» документов: a – точность; $б$ – полнота; $в$ – F -мера

Пример авторефератов, создающих шум для Темы 3:

- Петухов, Алексей Сергеевич. Синтез, пространственная структура и свойства семи-членных ацеталей пиридоксина : автореферат дис. ... кандидата химических наук : 02.00.03 / Казан. гос. ун-т. – Казань, 2004.
- Веселовский Александр Владимирович. Компьютерное моделирование активных центров моноаминоксидаз и создание ингибиторов с заданной селективностью : Дис. ... д-ра биол. наук : 03.00.04 : М., 2004.

Заключение

На основании анализа полученных данных, можно сделать следующие выводы. Алгоритм классификации по всему тексту автореферата дает неплохие результаты в случае, когда надо отсеивать «чужие» документы. На практике обычно это и требуется в большинстве случаев. Однако в случае, когда известно, что у каждого документа есть тема, он проигрывает двум другим алгоритмам, выполняющим классификации по разделам автореферата. Алгоритм классификации по разделам с условием многозначности терминов показывает себя не хуже в поиске «чужих» документов, чем алгоритмы классификации по тексту и классификации по разделам. В тестах алгоритм классификации по всему тексту автореферата немного превосходит алгоритм классификации по разделам с условием многозначности терминов. Однако алгоритм классификации по разделам с условием многозначности терминов вырывается вперед по сравнению с другими алгоритмами при поиске «своих» документов.

Список литературы

1. Кнут Д. Искусство программирования / Под ред. Ю. В. Козаченко. М.: Вильямс, 2002. Т. 1: Основные алгоритмы. 720 с. ISBN 5-8459-0080-8.
2. Ранганатан Ш. Р. Классификация двоеточием. Основная классификация / Пер. с англ. М.: ГПНТБ СССР, 1970.
3. Федотов А. М., Барахнин В. Б. Проблемы поиска информации: история и технологии // Вестн. НГУ. Серия: Информационные технологии. 2009. Т. 7, № 2. С. 3–17.
4. Федотов А. М., Барахнин В. Б., Жижимов О. Л., Федотова О. А. Модель информационной системы для поддержки научно-педагогической деятельности // Вестн. НГУ. Серия: Информационные технологии. 2014. Т. 12, № 1. С. 89–101.
5. Шрейдер Ю. А. Равенство, сходство, порядок. М.: Наука, 1971.
6. Воронин Ю. А. Начала теории сходства. Новосибирск: Наука, 1991. 128 с.
7. Леонова Ю. В., Федотов А. М. Извлечение знаний и фактов из текстов диссертаций и авторефератов // Системный анализ и информационные технологии: Тр. V Междунар. конф. Красноярск: ИВМ СО РАН, 2013. Т. 1. С. 232–242.

Yu. V. Leonova¹, A. M. Fedotov^{1,2}, O. A. Fedotova^{2,3}

¹ *Institute of Computational Technologies SB RAS
6 Academician Lavrentiev Ave., Novosibirsk, 630090, Russian Federation*

² *Novosibirsk State University
2 Pirogov Str., Novosibirsk, 630090, Russian Federation*

*State Public Scientific Technological Library SB RAS
15 Voskhod Str., Novosibirsk, 630200, Russian Federation*

juli@ict.nsc.ru

ON THE APPROACH TO THE CLASSIFICATION OF THESIS ABSTRACTS ON THEMES

The method of thematic classification of thesis abstracts is considered in the work. For this purpose, a specially constructed measure of the proximity of documents is used, taking into account the specifics of the subject area. As scales for the definition of a measure, it is suggested to take the characteristics of the structural attributes of the description of the author's abstracts (scientific novelty, provisions to be defended, etc.). The values of the weight coefficients in the formula for computing the proximity measure are determined by the assumed a posteriori reliability of the data of the corresponding scale.

Keywords: Classification of thesis abstracts, weight coefficients, measure of proximity, subject area, model of facet classification, classification algorithm.

References

1. Knut D. The Art of Computer Programming. Ed. by Yu. V. Kozachenko. Moscow, Williams, 2002, vol. 1: Fundamental Algorithms, 720 p. ISBN 5-8459-0080-8. (in Russ.)
2. Ranganatan Sh. R. Classification by colon. The main classification. Trans. with the English. Moscow, State Public Scientific Technical Library of the USSR, 1970. (in Russ.)
3. Fedotov A. M., Barakhnin V. B. The problems of information retrieval: history and technology. *Vestnik NSU. Series: Information technology*, 2009, vol. 7, no. 2, p. 3–17. (in Russ.)
4. Fedotov A. M., Barakhnin V. B., Zhizhimov O. L., Fedotova O. A. Model of an information system for supporting scientific and pedagogical activity. *Vestnik NSU. Series: Information technology*, 2014, vol. 12, no. 1, p. 89–101. (in Russ.)
5. Shreider Yu. A. Equality, similarity, order. Moscow, Nauka, 1971. (in Russ.)
6. Voronin Yu. A. The beginning of the theory of similarity. Novosibirsk, Science. Sib. Branch, 1991, 128 p. (in Russ.)
7. Leonova Yu. V., Fedotov A. M. Extraction of knowledge and facts from the texts of dissertations and abstracts. *V International Conference "System Analysis and Information Technologies" (SAIT-2013): Proceedings of the Conference*. Krasnoyarsk, ICM SB RAS, 2013, vol. 1, p. 232–242. (in Russ.)

For citation:

Leonova Yu. V., Fedotov A. M., Fedotova O. A. On the Approach to the Classification of Thesis Abstracts on Themes. Vestnik NSU. Series: Information Technologies, 2017, vol. 15, no. 1, p. 47–58. (in Russ.)