На правах рукописи

Черняк Екатерина Леонидовна

РАЗРАБОТКА ВЫЧИСЛИТЕЛЬНЫХ МЕТОДОВ АНАЛИЗА НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ АННОТИРОВАННЫХ СУФФИКСНЫХ ДЕРЕВЬЕВ

Специальность 05.13.18 —
«Математическое моделирование, численные методы и комплексы программ»

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

Москва — 2016

Работа выполнена на Департаменте анализа данных и искусственного интеллекта федерального государственного автономного образовательного учреждения высшего образования Национальный исследовательский университет «Высшая школа экономики»

Научный руководитель: доктор технических наук, старший научный сотрудник
Миркин Борис Григорьевич

Официальные оппоненты: **Моттль Вадим Вячеславович,**
доктор технических наук, профессор,
Федеральный исследовательский центр «Информатика и управление» Российской Академии Наук,
ведущий научный сотрудник

Петровский Михаил Игоревич,
кандидат физико-математических наук,
Факультет вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова,
доцент

Ведущая организация: Институт прикладной математики им. М. В. Келдыша РАН

Защита состоится –.–.—- г. в 11 часов на заседании диссертационного совета Д 002.073.04 при Федеральном исследовательском центре «Информатика и управление» Российской Академии Наук по адресу: 117312, Москва, проспект 60-летия Октября, д. 9.

С диссертацией можно ознакомиться в библиотеке Федерального исследовательского центра «Информатика и управление» Российской Академии Наук по адресу: 119333, Москва, ул. Вавилова д.40.

Отзывы на автореферат в двух экземплярах, заверенные печатью учреждения, просьба направлять по адресу: 117312, Москва, проспект 60-летия Октября, д. 9, ученому секретарю диссертационного совета Д 002.073.04.

Автореферат разослан –.–.—-.
Телефон для справок: +7 (499) 135-51-64.

Ученый секретарь
диссертационного совета
Д 002.073.04,
д.т.н., профессор Крутько В. Н.

Общая характеристика работы

Актуальность темы. Проникновение вычислительной техники во все сферы производственной, социальной и политической систем привело к необходимости разработки методов автоматического семантического анализа текстовых документов, размещенных в индивидуальных компьютерах и в интернете. Часть связанных с этим задач хорошо осознана и получает решение в научной и технической литературе. Это, прежде всего, задачи поиска и извлечения информации, категоризации текстов, извлечения ключевых словосочетаний, извлечение фактов и др. Большинство методов решения таких задач основано на предварительной «ручной» разметке текстов (выделение ключевых слов и других данных для обучения). Однако, в связи с наступлением эры глобализации, существует явная потребность в разработке методов, не требующих предварительной разметки текстов. Кроме того, создание корректных и эффективных морфологических и синтаксических парсеров – это трудоемкая задача, решенная не для всех языков. Это делает актуальной задачу разработки методов анализа текстов, не требующих их предварительной разметки.

В большинстве практических задач анализа коллекций текстовых документов, включая задачу информационного поиска, предполагается вычисление оценок релевантности «строка–текст». В качестве текстов, разумеется, выступают те или иные документы, а в качестве строк – ключевые слова и словосочетания, заданные извне или извлеченные из текстовых документов по определённым принципам, или произвольные элементы текста, состоящие из фиксированного количества букв или слов. Мера релевантности должна удовлетворять следующим естественным свойствам:

1. Интуитивная простота (понятные единицы и границы измерения);
2. Независимость от длины текста;
3. Независимость от лексической вариативности текста;
4. Возможность эффективной вычислительной реализации.

Большинство известных мер релевантности основаны на использовании в качестве элементарной единицы текста слова (или его нормальной формы – леммы, или его (псевдо)основы – стема). К этому классу моделей релевантности относятся векторная модель релевантности [Salton, 1988] вероятностная модель релевантности [Robertson, Zaragoza, 2009] языковая модель релевантности на словах или символьных n-граммах [Ponte, Croft, 1998], модель суффиксного дерева [Zamir, Etzioni, 1998]. Эти модели предполагают представление текста в виде неупорядоченного набора слов – «мешка» слов, а также предполагают учет морфологии и синтаксиса языка для идентификации и унификации слов. Существенным недостатком этих моделей можно считать невозможность учесть нечеткие (то есть, с различием на несколько символов) совпадения между строками и текстами. До некоторой степени этот недостаток помогают преодолеть языковая модель релевантности на символьных n–граммах [Ponte, Croft, 1998] и модель суффиксного дерева [Pampapathi и др., 2006]. Однако же, языковая модель

релевантности на символьных n–граммах часто бывает неэффективной с вычислительной точки зрения, поскольку возникающая в ней проблема нулевых вероятностей зачастую решается с помощью вычислительно неэффективных алгоритмов сглаживания, а модель суффиксного дерева, предложенная в [Pampapathi и др., 2006], по определению не удовлетворяет требованиям 3 и 4, сформулированным выше.

Для решения обозначенных выше задач – необходимости предобработки и нечеткости меры релевантности – и с учетом требований 1-4 необходима новая модель совокупности «строка – текст», а также структура данных, позволяющая вычислять нечеткие оценки релевантности.

В данном исследовании предлагается и верифицируется теоретико-множественная модель совокупности «строка – текст», а адекватной структурой данных для вычисления параметров оценки является аннотированное суффиксное дерево.

В теоретико-множественной модели совокупности «строка – текст» текст представляется в виде множества коротких строк, например, последовательных пар или троек слов, а строка S, состоящая из n символов, $S = s_1 s_2 \dots s_n$ – множеством всех подстрок $s_i \dots s_j$, где $i >= 1$, $j <= n$, $i <= j$. Для каждой пары строка – текст несложно найти все возможные общие подстроки, иначе говоря, совпадения. Максимальным совпадением назовем такое совпадение, при добавлении символа в начало или в конец которого, перестает быть совпадением. Допустим, существует совпадение строки с текстом $s_i \dots s_j$. Определим его вероятность, как условную частоту последнего символа s_j : $P(s_i \dots s_j) = P(s_j | s_i \dots s_{j-1})$ (УВС). Вероятностью максимального совпадения тогда является средняя сумма совпадений, в него входящих (СУВС), а полной релевантностью строки тексту – сумма вероятностей максимальных совпадений данному тексту (СУВСС). Для эффективной реализации вычисления оценок релевантности следует использовать аппарат аннотированного суффиксного дерева – структуры данных, которая позволяет вычислять все частоты всех подстрок.

Объект исследования – вычислительные задачи анализа текстовых документов, написанных на естественном языке.

Предмет исследования – вычислительное моделирование текстов как строк символов и задачи их анализа, решаемые путем наложения разных строк друг на друга.

Цель данного диссертационного исследования – разработка оригинальных моделей, методов, алгоритмов и программных комплексов, предназначенных для решения некоторых задач анализа текстовых документов на естественном языке на уровне последовательностей символов.

К **задачам** исследования относятся:

1. Разработка модели представления коллекции текстовых документов строками и ассоциированной с ней функции релевантности;
2. Верификация разработанной модели на реальных задачах анализа коллекций текстовых документов:

a) Рубрикация текстовых документов в соответствии с заданной системой рубрик;
b) Пополнение таксономии с использованием внешней коллекции текстов;
c) Фильтрация коллекции текстовых документов от обсценной лексики.

3. Реализация разработанных моделей и методов в виде комплекса программ.

К **методам**, использованным в исследовании, относятся:

1. Метод Укконена для построения аннотированного суффиксного дерева за линейное время;
2. Метод вычисления релевантности строки тексту с помощью наложения строки на аннотированное суффиксное дерево его представляющее;
3. Методы вычисления релевантности строки тексту, основанные на представлении текстов векторными пространствами и вероятностными моделями.

Научная новизна. В диссертации получен ряд новых научных результатов, которые **выносятся на защиту**:

1. Разработана теоретико-множественная модель совокупности «строка-текст» с методом оценки релевантности строк тексту, основанном на аннотированных суффиксных деревьев. Предложен новый метод вычисления оценок релевантности строки тексту СУВСС, апробированный в работе;
2. Предложен метод рубрикации научных статей с использованием критерия релевантности СУВСС, более точного, чем популярные методы, традиционно используемые в международных публикациях;
3. Разработан метод использования справочных материалов интернета, с учетом наличия в них шумовой компоненты, для пополнения предметных таксономий. Методика апробирована в задачах пополнения таксономий чистой и прикладной математики с использованием русскоязычной Википедии;
4. Показана эффективность использование критерия релевантности СУВСС в классе задач поиска по однословному ключу, в котором полнота важнее, чем точность;
5. Разработаны комплексы программ, реализующие предложенную теоретико-множественную модель совокупности «строка – текст» с использованием критерия релевантности СУВСС, применительно к решению задач в пунктах 2, 3 и 4.

Теоретическая значимость работы заключается в разработке принципиально новых моделей и методов: теоретико-множественной модели совокупности «строка – текст», модели нормированного аннотированного суффиксного дерева с критерием релевантности СУВСС, а также метода построения таблиц релевантности «строка – текст» (РСТ) для применения в конкретных задачах.

Практическая ценность подтверждена экспериментами по сравнительной оценке использования мер релевантности для рубрикации научных статей, результатами расчетов по пополнению таксономий с использованием материалов интернета и результатами решения задач поиска, ориентированных на его полноту. Все разработанные методы реализованы в виде программных комплексов, предназначенных для решения исследовательских и прикладных задач. Разработанные методы и алгоритмы были успешно применены в реальных проектах компании ООО «ФОРС-Центр разработки» (метод фильтрации обсценной лексики использован для анализа и определения тональности текстов в социальных сетях в системе FORSMedia) и «ЕС-Лизинг» (метод рубрикации использован для категоризации проектной документации) и проектах, выполнявшихся по грантам ВШЭ в 2010 – 2015 гг., а также в преподавательской деятельности Департамента анализа данных и искусственного интеллекта Факультета компьютерных наук НИУ ВШЭ.

Достоверность полученных результатов подтверждена строгостью использованных математических моделей и методов, экспериментами по сравнению результатов применения разработанных традиционных методов на конкретных задачах, а также алгоритмической эффективностью программных реализаций.

Апробация результатов работы. Основные результаты работы обсуждались и докладывались на следующих научных конференциях и семинарах:

- 1-ой, 2-ой всероссийских научных конференция “Анализ изображений, сетей и текстов” (АИСТ-2012, АИСТ-2013), Екатеринбург, Россия; темы докладов – “Автоматизация использования таксономий для аннотирования текстовых документов”, “Использование ресурсов интернета для построения таксономии”
- 1-ом семинаре по кластерам, деревьям и порядкам (COT-2013), Москва, Россия; тема доклада – “An AST method for scoring string-to-text similiarity in semantic text analysis”
- 8-ой международной конференции “Диалог” (Диалог-2013), Бекасово, Россия; тема доклада – “Computational refining of Russian-language taxonomy using Wikipedia”
- 3-ей международной научной конференции “Анализ изображений, сетей и текстов” (АИСТ-2014), Екатеринбург, Россия; тема доклада – “Conceptual maps: construction over a text collection and analysis”
- 2-ой международной конференции “Информационные технологии и количественный менеджмент” (ITQM-2014), Москва, Россия; тема доклада – “A method for refining a taxonomy by using annotated suffix trees and Wikipedia recourses”
- 3-ей всероссийской конференции “Искусственный интеллект и естественный язык” (AINL-2014), Москва, Россия; тема доклада – “Создание и визуализация газетного интернет-корпуса”

– 8-ой международной конференции “Веб-поиск и майнинг данных” (WSDM-2015), Шанхай, КНР тема доклада – “An approach to the problem of annotation of research publication”;
– 2-ом международном семинаре по майнингу данных и автоматической обработке текстов (DMNLP-2015) тема доклада – “Some thoughts on using annotated suffix trees for NLP tasks”.

Публикация результатов. Основные результаты работы изложены в 13 научных статьях. 7 статей опубликованы в рецензируемых сборниках трудов международных и всероссийских конференций, 3 статьи опубликованы в журналах из списка ВАК.

Структура диссертации. Диссертация состоит из введения, 6 глав, заключения, и списка литературы, состоящего из 105 наименований.

Во **введении** раскрывается актуальность темы диссертации, формулируются проблемы и задачи исследования, предмет исследования, определяются цели работы, описываются методы исследования, излагаются основные научные результаты, обосновывается теоретическая и практическая значимость работы, даётся общая характеристика исследования.

В **первой главе** приводится обзор основных видов формализации коллекций текстовых документов: векторная модель, языковая модель, модель скрытых тем, модель суффиксного дерева. Вводится теоретико-множественная модель представления коллекции текстовых документов: каждый документ представлен набором всех возможных символьных подпоследовательностей фиксированной длины и короче. Вводится понятие нормированного аннотированного суффиксного дерева, используемого для вычисления частот всех символьных подпоследовательностей текстовых документов в теоретико-множественной модели.

Суффиксное дерево для m-символьной строки S представляет собой ориентированное дерево с корнем, имеющее ровно m листьев, занумерованных от 1 до m. Каждая внутренняя вершина, отличная от корня, имеет не меньше двух детей, а каждая строка помечена непустой подстрой строки S. Никакие две дуги, выходящие из одной и той же вершины, не могут иметь пометок, начинающихся с одного и того же символа. Главная особенность суффиксного дерева заключается в том, что для каждого листа i конкатенация меток дуг на пути от корня к листу i составляет / произносит / кодирует / прочитывает суффикс строки S, который начинается в позиции i, то есть, $S[i : m]$ [Гасфилд, 2003]

Аннотированное суффиксное дерево определяется в [Pampapathi и др., 2006] как суффискное дерево, в котором:
– Символы стоят не на ребрах, а в узлах;
– Каждому узлу соответствует один символ;
– Каждый узел помечен частотой фрагмента, который прочитывает путь от корня до этого узла;
– Опущены терминальные символы и метки листьев, представляющие номер суффикса и входной строки.

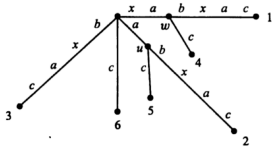

Рис. 1 — Суффиксное дерево для строки $S =$ `xabxac`

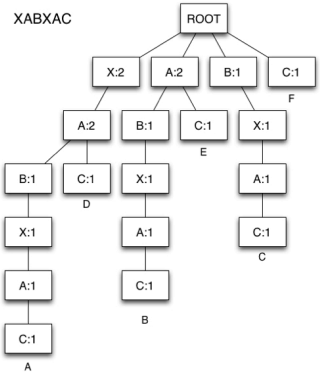

Рис. 2 — Аннотированное суффиксное дерево для строки $S =$ ``xabxac''

Свойство 1. Частота любого узла равна сумме частот его узлов-детей, так как родительский узел соответствует префиксу нескольких суффиксов и его частота складывается из частот этих суффиксов. Отсюда же следует другое свойство АСД.

Свойство 2. Частота родительского узла равна сумме частот листьев, которые он покрывает.

В **первой главе** описывается наивный алгоритм построения АСД для входной строки, являющийся адаптацией классического наивного алгоритма построения суффиксных деревьев [Гасфилд, 2003]. Сначала инициализируется пустая структура, к которой итерационно добавляются суффиксы входной строки. При добавлении каждого суффикса проверяется, совпадает ли префикс текущего суффикса с каким-либо путем от корня в АСД (назовем такой путь совпадением). Если да, то частоты в совпадении увеличиваются на 1, а не совпавший фрагмент суффикса добавляется к последнему узлу в совпадении в качестве потомка с частотами, равными 1. Если совпадения нет, то суффикс добавляется в АСД как цепочка узлов с частотами 1.

Во **второй главе** рассматривается задача вычисления релевантности «строка – текст», являющаяся базовым этапом любой практической задачи обработки и анализа коллекций текстовых документов. Приводится обзор существующих мер релевантности в векторной, вероятностной и языковой моделях, а также их модификации с учетом снижения размерности модели. Общий подход к вычислению релевантности строки тексту заключается в вычислении тех

или иных частотных характеристик коллекции документов и числа совпадений – совпадающих элементов строки и текста. В качестве элемента текста могут выступать слова в неизменном виде, их (псведо)основы, леммы или символьные фрагменты. Утверждается, что все рассмотренные меры релевантности обладают общими недостатками: они не учитывают вложенность совпадений друг в друга и не учитывают возможные нечеткие совпадения. Предлагается использовать в качестве оценок релевантности в теоретико-множественной модели оценки сходства, получаемые по методу нормированного аннотированного суффиксного дерева. Утверждается и демонстрируется, что только такой метод вычисления релевантности позволяет преодолеть сформулированные недостатки других мер релевантности. Вводится понятие нормированного аннотированного суффиксного дерева и связанной с ним естественно интерпретируемой функции релевантности СУВСС (средняя условная вероятность символа в совпадении).

Оценка релевантности строки тексту вычисляется в два этапа: сначала – оценки каждого суффикса входной строки 1, затем – нормированное среднее оценок всех суффиксов входной строки 2. Обозначим $f(\mathtt{node})$ – частота узла $\mathtt{node}$, а $f(\mathtt{node}_{parent})$ – частота родительского узла. Оценка одного суффикса входной строки складывается из условных вероятностей узлов , входящих в совпадение $\mathtt{match}$ префикса данного суффикса с АСД – $\frac{f(\mathtt{node})}{f(\mathtt{node}_{parent})}$), преобразованных при помощи шкалирующей функции и нормированных длинной совпадения $|\mathtt{match}|$.

$$\mathtt{score}(\mathtt{match}(suffix,ast)) = \frac{\sum_{node \in match} \phi(\frac{f(\mathtt{node})}{f(\mathtt{node}_{parent})})}{|suffix|}, \qquad (1)$$

Оценка всей входной строки 2 – это сумма всех нормированных оценок ее суффиксов $suffix$, усредненное длиной строки $match$.

$$\mathtt{relevance}(string,text) = \frac{\sum_{suffix} \mathtt{score}(\mathtt{match}(suffix,ast))}{|string|} \qquad (2)$$

Шкалирующая функция $\phi(x)$ может иметь следующий вид:
– $\phi(x) = 1$ – константа (обозначение – constant);
– $\phi(x) = x$ – линейная (обозначение – linear);
– $\phi(x) = \log \frac{x}{1-x}$ – логистическая (обозначение – logit);
– $\phi(x) = \sqrt{x}$ – корень квадратный (обозначение – root);
– $\phi(x) = x^2$ – квадратичная функция (обозначение – square);
– $\phi(x) = \log(x)$ – логарифмическая функция (обозначение – log);
– $\phi(x) = \frac{1}{1+e^{-x}}$ – сигмоида (обозначение – sigmoid).

Предлагается адаптация алгоритма Укконена [Ukkonen, 1995] для построения АСД за линейное время. На первом шаге этого метода строится непосредственно суффиксное дерево. На втором шаге, обходя дерево снизу и зная, что

частота каждого листового узла равна единице, аннотируем дерево. Использование этого модифицированного метода предполагает добавление терминальных символов к концу строк. Для построения АСД по тексту требуется предварительное представление текста как совокупности строк последовательно идущих слов. Строки строятся следующим образом: первая строка начинается с первого слова в тексте и заканчивается 2м, 3м или 4м словом в тексте; вторая начинается со второго слова и заканчивается соответственно на 3-5 слове. При этом учитываются границы предложений. Такое представление существенно ограничивает глубину и, следовательно, сложность формируемого АСД, зато существенно сокращает вычисления. Количество слов в строке определяется в зависимости от задачи. Таблица релевантности «строка – текст» строится помощью метода нормированного АСД. В этой таблице строки соответствуют отдельным входным строкам, столбцы – отдельным текстам, а элементы – оценки релевантности строк соответствующим текстам.

В **третьей**, **четвертой** и **пятой** главах рассматриваются примеры использования теоретико-множественной модели коллекции текстовых методов и ассоциированного метода вычисления релевантности, использующего нормированное аннотированное суффиксное дерево: задача рубрикации аннотаций научных публикаций темами таксономии, задача пополнения научной таксономии и задача фильтрации обсценной лексики, соответственно.

В **третьей** рассмотрена задача рубрикации научных статей. Задача рубрикации научных статей заключается в категоризации статей в системе рубрик, заданных классификатором или таксономией (корневым деревом тематических единиц) соответствующей области знания или технологии. Для решения задачи категоризации в режиме без учителя мы предлагаем формировать PCT таблицу таксономическая тема – текст, после чего каждый текст категоризуется таксономическими темами, получившими наивысшие оценки релевантности. Автором составлен и проведен эксперимент по сравнению относительных преимуществ использования различных мер релевантности в проблеме рубрикации научных статей. Входные данные эксперимента:

- 5079 аннотации научных статей из журналов ACM, хранящиеся в свободном доступе в электронной библиотеке ACM Digital Library;
- Таксономия ACM CCS 2012, состоящая из 2074 таксономических тем и насчитывающая 6 уровней;
- Авторские темы, приписанные аннотациям научных статей их авторами – 2-3 таксономические темы низших уровней, а также все темы, лежащие на пути от корня до них в дереве таксономии ACM CCS.

В эксперименте рассматриваются наиболее популярные в международной литературе меры релевантности, и производится сравнение результатов их использования с результатами использования введенной нами меры средней условной вероятности символа в совпадении (СУВСС). Полный список рассмотренных мер релевантности приведен в Таблице 1.

Таблица 1 — Обозначения мер релевантности в задаче рубрикации научных статей

Обозначение	Мера релевантности
cosine	Косинусная мера релевантности
LSI_N	Косинусная мера релевантности со снижением до N размерностей методом LSI
LDA_N	Мера релевантности, основанная на ЛРД с N темами
$okapibm25$	Мера релевантности BM25
$Jaccard$	коэффициент Жаккара на множестве буквенных n-грамм
$constant_X$	мера СУВСС с константной шкалирующей функцией и очисткой шума от уровня X
$linear_X$	мера СУВСС с линейной шкалирующей функцией и очисткой шума от уровня X
$square_X$	мера СУВСС с шкалирующей квадратичной функцией и очисткой шума от уровня X
$root_X$	мера СУВСС с линейной шкалирующей функцией корень квадратный и очисткой шума от уровня X
log_X	мера СУВСС с логарифмической шкалирующей функцией и очисткой шума от уровня X
$logit_X$	мера СУВСС с логистической шкалирующей функцией и очисткой шума от уровня X
$sigmoid_X$	мера СУВСС с шкалирующей функцией сигмоид и очисткой шума от уровня X

В эксперименте использовались различные способы предобработки текстов. Косинусная мера релевантности и мера релевантности BM25 предполагают представление текста в виде мешка термов – неупорядоченного набора термов, использование меры Жаккара – в виде множеств n-грамм, использование СУВСС – в виде набора строк.

Для оценки результатов две популярные характеристики точности: `MAP` (Mean Average Precision) и `nDCG` (normalized discounted cumulative gain), а также предложенные автором меры $I(k)$ и $J(k)$, имеющие смысл доли количества статей, верно аннотированных хотя бы одной темой и доли таксономических единиц, формирующих верные аннотации.

Таблица 2 — Результаты эксперимента: меры MAP и nDCG

	MAP_5	MAP_{10}	MAP_{15}	nDCG_5	nDCG_{10}	nDCG_{15}
косинусная мера релевантности						
$words$	0.1775	0.2073	0.2242	0.0478	0.1073	0.1770
$stems$	0.1874	0.2206	0.2368	0.0482	0.1146	0.1806
$lemmas$	0.1970	0.2302	0.2464	0.0478	0.1141	0.1806
$4gram$	0.2202	0.2569	0.2733	0.0516	0.1242	0.1914
мера релевантности BM25						
$words$	0.0294	0.0372	0.0423	0.0062	0.0222	0.0439
$stems$	0.0602	0.0724	0.0789	0.0185	0.0442	0.0709
$lemmas$	0.0455	0.0556	0.0629	0.0127	0.0340	0.0643
$4gram$	0.1247	0.1407	0.1489	0.0309	0.0613	0.0961
мера релевантности, основанная на АСД						
$linear_0$	0.2734	0.3071	0.3221	0.0508	0.1162	0.1786
$linear_1$	0.2742	0.3075	0.3233	0.0500	0.1142	0.1793
$root_0$	0.2854	0.3226	0.3369	0.0534	0.1268	0.1870
$root_1$	0.2826	0.3170	0.3324	0.0548	0.1221	0.1868
$sigmoid_0$	**0.2904**	**0.3258**	**0.3400**	**0.0576**	**0.1264**	**0.1848**
$sigmoid_1$	0.2873	0.3207	0.3359	0.0591	0.1257	0.1874

Использование нечетких мер релевантности, основанных на аннотированных суффиксных деревьях в полтора раза увеличивает точность по сравнению с остальными мерами релевантности, при этом, вид шкалирующей функции (квадратичная, корень квадратный, линейная или сигмоид) не играет особой роли (за исключением логарифмической или логистической). По мерам $nDCG@15$ и $MAP@15$ результаты АСД-релевантности в среднем в полтора раза лучше результатов по мере Жаккара, косинусной мере и мере BM25, в том числе, с учетом снижения размерности. По абсолютным показателям $I(k)$ и $J(k)$, результаты использования АСД, в среднем на 30 - 40 единиц превосходят результаты по мере Жаккара, косинусной мере и мере BM25 , в том числе, с учетом снижения размерности.

Таблица 3 — Результаты эксперимента: меры $I(k)$ и $J(k)$

	I_5	I_{10}	I_{15}	J_5	J_{10}	J_{15}
косинусная мера релевантности						
$words$	333	543	741	239	348	427
$stems$	344	578	766	238	352	439
$lemmas$	350	584	773	253	365	447
$4gram$	387	644	835	276	399	477
мера релевантности BM25						
$words$	53	107	165	44	86	138
$stems$	120	207	285	97	154	197
$lemmas$	87	168	236	72	134	183
$4gram$	237	330	426	194	242	287
мера релевантности, основанная на АСД						
$linear_0$	429	662	839	321	423	491
$linear_1$	427	656	841	322	424	498
$square_0$	401	621	796	306	409	480
$square_1$	406	624	799	308	411	479
$root_0$	451	711	881	328	446	515
$root_1$	452	691	874	332	438	511
log_0	43	117	180	36	98	143
log_1	66	114	136	61	91	105
$logit_0$	103	197	284	84	154	217
$logit_1$	42	66	85	38	55	66
$sigmoid_0$	467	712	878	332	453	509
$sigmoid_1$	468	703	879	337	440	513

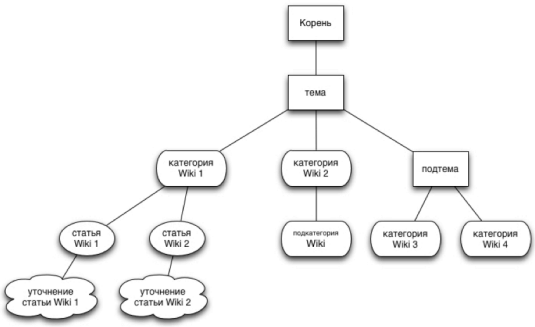

Рис. 3 — Схема пополнения таксономии. В прямоугольниках находятся темы основы таксономии, в скругленный прямоугольниках – достроенные категории и подкатегории Википедии. Листья достроенной таксономии – названия статей Википедии – помещены в овалы. В облачках находятся уточнения листьев.

В **четвертой главе** рассмотрена проблема автоматизации построения таксономий – весьма актуальная как для обработки текстов, так и информационного поиска. Метод построения таксономии, предложенный автором, состоит из двух шагов. На первом шаге задается основа таксономии, два или три уровня, построенных вручную, основываясь на официальных документах и определениях из паспортов специальностей ВАК. Второй шаг заключается в пошаговом пополнении этой основы детализирующими материалами интернета. В качестве таковых рассматриваются фрагменты дерева категорий и статей русскоязычной Википедии. Для соотнесения категорий, названий статей, таксономических тем и статей и очистки дерева категорий Википедии от шума используется мера релевантности СУВСС и аппарат РСТ таблиц. Структура результирующей таксономии представлена на Рис. 3. Метод проиллюстрирован применением к двум областям математики – теории вероятностей и математической статистики (ТВиМС) и численным методам (ЧМ).

Данные, извлеченные из Википедии, необходимо предварительно очистить от шума. Требуется удалить циклы из дерева категорий, если они в нем есть, и оставить в дереве только такие подкатегории и статьи, которые имеют логическую и смысловую связь с родительскими категориями. Второй шаг метода состоит из следующих этапов:

1. Извлечение дерева категорий и статей из Википедии
2. Очистка дерева категорий от нерелевантных статей
3. Очистка дерева категорий от нерелевантных подкатегорий
4. Достраивание категорий Википедии к темам таксономии
5. Формирование промежуточных уровней таксономии
6. Использование названий статей Википедии в качестве листьев в получаемой таксономии
7. Излечение ключевых слов и словосочетаний из статей Википедии для использования их в качестве уточнений листьев.

Таблица 4 — Качество очистки от шума

		эксперты	
		шум	не шум
АСД	шум	731	67
	не шум	21	264

Таблица 5 — Качество достраивания категорий Википедии к темам таксономии и формирования промежуточных уровней

		эксперты	
		родитель	не родитель
АСД	родитель	403	51
	не родитель	95	78

Оценка построенной таксономии ТВиМС была проведена при помощи экспертов. Для них был составлен специальный опросник, в первой части которого экспертов просили определить шумовые темы, а во второй – указать для данной темы родительскую тему. Таким образом, при помощи первой части опросника была проверена точность очистки от шума, а при помощи второй части опросника – точность и корректность достраивания промежуточных уровней таксономии.

В экспертной оценке построенной таксономии ТВиМС участвовали два эксперта. Полученные результаты представлены в Таблицах 4 и 5.

Аккуратность $accuracy$ очистки от шума составляет 0.91, достраивания категорий Википедии к темам таксономии и формирования промежуточных уровней – 0.76.

Достоверность проведенного экспертного оценивания определяется независимо для обеих частей исследования. Согласованность ответов экспертов на вопросы из Части 1 определяется с помощью коэффициента κ Коэна, на вопросы из Части 2 – долей несовпавших ответов. Коэффициент согласованности κ Коэна ответов на вопросы из Части 1 составляет 0.319, т.е., в принципе, ответы экспертов можно считать согласованными. Доля несовпавших ответов на вопросы из Части 2 составляет 12%.

Предложенный метод пополнения таксономии ReTAST-w позволяет построить качественную таксономию: доля полученных ошибок не велика, экспертные оценки, подтверждающие высокое качество таксономии, – достаточно высоки и согласованы на приемлемом статистическом уровне. Фрагменты достроенных таксономий представлены на Рис. 4 и Рис. 5.

В **пятой главе** используется аналогия между задачей поиска по однословному ключу и фильтрации нежелательной (в данном случае – обсценной) лексики.. Утверждается, что несмотря на то, что для решения этих задач могут быть использованы одинаковые методы, оптимизируются разные критерии качества, которые влияют на выбор конкретного метода. Описывается эксперимент по разработке фильтра на основе СУВСС и демонстрируется его эффективность с точки зрения оптимизируемого критерия – полноты, а так же с точки зрения времен-

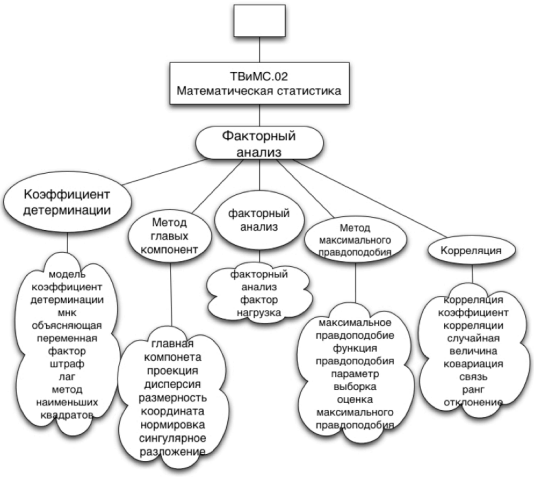

Рис. 4 — Фрагмент достроенной таксономии ТВиМС. В прямоугольниках находятся темы основы таксономии, в скругленных прямоугольниках – достроенные категории и подкатегории Википедии. Листья достроенной таксономии – названия статей Википедии – помещены в овалы. В облачках находятся уточнения листьев.

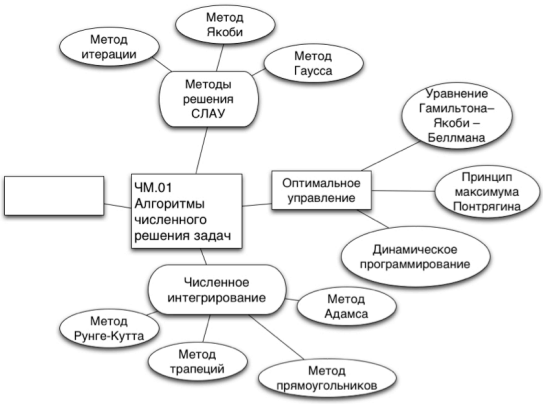

Рис. 5 — Фрагмент достроенной таксономии ЧМ. В прямоугольниках находятся темы основы таксономии, в скругленных прямоугольниках – достроенные категории и подкатегории Википедии. Листья достроенной таксономии – названия статей Википедии – помещены в овалы.

ной сложности. Использование критерия полноты обусловлено тем, что ошибка второго рода – пропуск обсценного слова при фильтрации существенно важнее, чем ошибка первого рода – определение нормативного слова как обсценного.

Для эксперимента по верификации нашего подхода рассматриваются несколько способов фильтрации обсценной лексики:

- Поиск по совпадению: слово t входит в стоп-лист в неизмененной форме
- Поиск по лемме: нормальная форма слова t входит в стоп-лист
- Поиск по основе (стему): основа (стем) слова t входит в стоп-лист
- Поиск по составляющим: найдено такое стоп-слово s, что коэффициент Жаккара между множеством n-грамм, на которые разбивается слово s и множеством n-грамм, на которые разбивается слово t превышает некий заранее заданный порог;
- Поиск по редакционному расстоянию: найдено такое стоп-слово s, что редакционное расстояние Левенштейна (то есть, число операций вставки, удаления и замены символа) [Левенштейн, 1965] между ним и словом t ниже некого заранее заданного порога;
- Поиск с использованием СУВСС: оценка вхождения слова t в АСД, построенное по стоп-листу, превышает некий заранее заданный порог.

В качестве стоп-листа был использован список слов, запрещенных к использованию для наименования ресурсов в доменной зоне “рф”. Стоп-лист содержит 4023 слова, например, таких как “говнецо”, “сиська”, “шалашовка”. Коллекция текстов была составлена и размечена автором исследования самостоятельно. Она состоит из научных статей об этимологии русского мата, текстов произведений Юза Алешковского, Игоря Губермана и Владимира Сорокина, песен групп Ленинград и Красная Плесень, стихотворений Сергея Есенина, Владимира Маяковского и Александра Пушкина, постов Артемия Лебедева в Живом Журнале, статей, опубликованных на портале Луркмор, а так же частушек, анекдотов и пословиц. Общий размер коллекции составляет 294916 словоупотреблений и 60868 словоформ.

По точности лучшим методом фильтрации является поиск совпадения. Худшими фильтрами по точности оказываются фильтры, основанные на расстоянии Левенштейна, но эти же фильтры являются лучшими по полноте. Второе место по полноте занимает фильтр, основанный на АСД: этот фильтр обнаруживает порядка 60% обсценной лексики. Остальные фильтры существенно проигрывают по полноте, но выигрывают по точности. Среди двух использованных лемматизаторов, лучшие результаты достигаются при использовании Mystem. Стемминг позволяет достичь выигрыша порядка 10% по полноте сравнению с лемматизацией при относительно несущественном падении точности. Вычисление меры Жаккара на n-граммах при сравнительно высокой точности приводит к низким значением полноты.

Важным параметром для сравнения фильтров является их вычислительная сложность. Приведем оценки вычислительной сложности каждого фильтра. До-

пустим, что n – это максимум из всех возможных длин слов, m – максимум из длины частотного слова и стоп-листа, $n \ll m$. Тогда:

- Сложность поиска по совпадению (лемме, стему) составляет $O(m)$ – слово (лемма, стем) проверяется на совпадение со словами (стемами) из стоп-листа;
- Сложность попарного вычисления коэффициента Жаккара на множествах n-грамм для слов из частотного словаря и стоп-листа и расстояния Левенштейна составляет $O(n^2 \cdot m^2)$, сложность проверки одного слова – $O(n^2 \cdot m)$;
- Сложность построения АСД с помощью алгоритма Укконена для стоплиста составляет $O(m \cdot n)$, сложность проверки одного слово – $O(n)$.

Таким образом, по общим мерам качества (аккуратности и F_2-мере), выигрывают фильтры, использующие поиск совпадения по леммам. Однако, с учетом важности полноты и эффективности по времени, наилучшими являются фильтры, использующие СУВВС. Они сбалансированы по точности и полноте, имеют аккуратность, сопоставимую с другими рассматриваемыми фильтрами, хотя и невысокое значение F_2-меры из-за низкой точности.

В **шестой главе** приводится описание программных комплексов, реализующих разработанные в исследовании модели и методы, а также решающие некоторые вспомогательные задачи сбора и обработки данных. Программный комплекс EAST реализует предложенный алгоритм построения нормированного аннотированного суффиксного дерева за линейное время, а также выполняет предварительную обработку текстов. Программный комплекс EAST распространяется свободно и доступна как в виде консольного приложения, так и в виде библиотеки для языка Python. Программный комплекс WikiDP позволяет извлекать из Википедии данные различных типов, такие как дерево категорий с корнем в заданном узле и принадлежащие к этом дереву статьи.

В заключении приводятся основные выводы, итоги и результаты работы.

Основные результаты работы

- Предложена теоретико-множественная модель представления коллекций текстовых документов, в которой текст рассматривается как последовательность символов. Представлением текста служат все символьное последовательности фиксированной длины и короче и их частоты;
- Предложено использовать метод нормированного аннотированного суффиксного дерева, который позволяет за линейное от размера текста время найти всего его фрагменты заданной длины и короче, а также вычислить их частоты, для оценки частот теоретико-множественной модели представления коллекций текстовых документов;
- Предложена мера релевантности СУВСС представляет собой среднюю условную частоту символа в максимальном совпадении и позволяет находить оценки релевантности строки тексту, которые
  - не зависят от размера входного текста или коллекции текстов;

  - учитывают нечеткие совпадения между входной строкой и текстом.
- Предложены и верифицированы методы для решения следующих задач:
  1. Метод рубрикации научных статей в соответствии с системой рубрик, заданной таксономией. Экспериментальное сравнение меры релевантности СУВСС с существующими мерами релевантности показывает, что при использовании СУВСС достигаются показатели точности в 1.5 выше, чем при использовании других мер релевантности.
  2. Метод пополнения таксономии предметной области. Экспертное оценивание примеров пополненных таксономий показывает, что метод позволяет построить обширные и качественные таксономии.
  3. Метод фильтрации обсценной лексики. Устанавливается аналогия между очисткой от обсценной лексики и поиском по однословному ключу с поправкой на оптимизируемый критерий. Демонстрируется эффективность метода СУВСС по сравнению со стандартными методами поиска по однословным ключам и редакционному расстоянию по полноте и временной эффективности.
- Разработано два программных комплекса: WikiDP – для загрузки статей и дерева категорий русскоязычной Википедии; EAST – для построения нормированных суффиксных деревьев, вычисления меры релевантности СУВСС и построения таблиц релевантности строк тексту.

Публикации в журналах, входящих **в перечень ВАК**:

1. Черняк Е.Л. Меры релевантности строка-текст в проблеме рубрикации научных статей / Черняк Е.Л., Миркин Б.Г. // Бизнес-информатика. 2014. № 2. С. 51–62. – 1.15 п.л. (личный вклад автора – 0.5 п.л.)
2. Черняк Е.Л. Метод аннотированного суффиксного дерева для оценки степени вхождения строк в текстовые документы / Черняк Е.Л., Миркин Б.Г., Чугунова О.Н. // Бизнес-информатика. 2012. № 2. С. 31–41. – 1 п.л. (личный вклад автора – 0.45п.л.)
3. Черняк Е.Л. Системы автоматической обработки текстов / Черняк Е.Л., Ильвовский Д.А. // Открытые системы. СУБД. 2014. № 1. С. 51–43. – 0.45 п.л. (личный вклад автора – 0.3 п.л.)

Прочие публикации:

1. Chernyak E.L. A Method for Refining a Taxonomy by Using Annotated Suffix Trees and Wikipedia Resources / Chernyak E.L., Mirkin B.G. // 2nd International Conference On Information Technology and Quantitative Management ITQM 2014. Procedia Computer Science. 2014. Vol. 31. P. 193 – 200 – 0.4 п.л. (личный вклад автора – 0.2 п.л.)

2. Chernyak E.L. An AST method for scoring string-to-text similiarity in semantic text analysis / Chernyak E.L., Mirkin B.G. // Springer. 2014. Vol. 92. P. 92 – 96 – 0.35 п.л. (личный вклад автора – 0.2 п.л.)
3. Chernyak E.L. An approach to the problem of annotation of research publications / Chernyak E.L. // Proceedings of 8th ACM International Conference On Web Search and Data Mining. ACM. 2014. Vol. 58. P. 429-434 – 0.4 п.л. (личный вклад автора – 0.4 п.л.)
4. Chernyak E.L. Conceptual maps: construction over a text collection and analysis / Morenko E.N., Chernyak E. L., Mirkin B. G. // Springer International Publishing. 2014. Vol. 439. P. 163-169 (– 0.5 п.л. (личный вклад автора – 0.2 п.л.)
5. Chernyak E.L. Conceptual maps: construction over a text collection and analysis / Chernyak E. L., Mirkin B. G. // Материалы ежегодной конференции Диалог. 2013. Т. 2. Стр. 177 – 185 – 0.5 п.л. (личный вклад автора – 0.25 п.л.)
6. Черняк Е.Л. Аннотированные суффиксные деревья: особенности реализации / Дубов М. С., Черняк Е. Л. // Использование ресурсов Интернета для построения таксономии // Доклады всероссийской научной конференции АИСТ'2013. 2013. Стр. 49 – 57 – 0.45 п.л. (личный вклад автора – 0.2 п.л.)
7. Черняк Е.Л. Использование ресурсов Интернета для построения таксономии / Черняк Е. Л., Миркин Б. Г. // Использование ресурсов Интернета для построения таксономии // Доклады всероссийской научной конференции АИСТ'2013. 2013. Стр. 49 – 57 – 0.35 п.л. (личный вклад автора – 0.2 п.л.)

Черняк Екатерина Леонидовна

РАЗРАБОТКА ВЫЧИСЛИТЕЛЬНЫХ МЕТОДОВ АНАЛИЗА НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ АННОТИРОВАННЫХ СУФФИКСНЫХ ДЕРЕВЬЕВ

Автореф. дис. на соискание ученой степени канд. техн. наук

Подписано в печать ___.___._____. Заказ № _____
Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.
Типография _____