На правах рукописи

Заикин Данила Александрович

ТЕРМИНОЛОГИЧЕСКИЙ ПОИСК В КОЛЛЕКЦИЯХ МАТЕМАТИЧЕСКИХ ТЕКСТОВ

Специальность 05.13.11 –
Математическое и программное обеспечение вычислительных машин,
комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Казань
2014

Работа выполнена на кафедре прикладной информатики института вычислительной математики и информационных технологий Федерального государственного автономного учреждения высшего профессионального образования «Казанский (Приволжский) федеральный университет».

Научный руководитель:	**Соловьев Валерий Дмитриевич**, доктор физико-математических наук, профессор.
Официальные оппоненты:	**Щербаков Андрей Юрьевич**, доктор технических наук, профессор, главный научный сотрудник отдела информационных технологий Федерального государственного бюджетного учреждения науки Вычислительный центр имени А.А. Дородницына Российской академии наук.
	Браславский Павел Исаакович, кандидат технических наук, старший научный сотрудник лаборатории комбинаторной алгебры института математики и компьютерных наук Федерального государственного автономного образовательного учреждения высшего профессионального образования «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина».
Ведущая организация:	Федеральное государственное бюджетное образовательное учреждение высшего образования Московский государственный университет имени М.В. Ломоносова.

Защита состоится «05» марта 2015 года в 15 часов на заседании диссертационного совета Д 002.087.01 при Федеральном государственном бюджетном учреждении науки Институте системного программирования Российской академии наук по адресу: 109004, Москва, ул. Александра Солженицына, дом 25.

С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Института системного программирования Российской академии наук.

Автореферат разослан «03» февраля 2015 года.

Ученый секретарь
диссертационного совета
кандидат физ.-мат.наук /Зеленов С.В./

Общая характеристика работы

Актуальность темы исследования. Взрывообразный рост разнообразных публикаций в сети Интернет приводит к тому, что постоянно повышаются требования к информационно-поисковым системам[1]. Актуальность исследований в области информационного поиска также обусловлена тем, что при поиске информации в сети Интернет число документов, возвращаемых на запрос пользователя, как правило, получается очень большим за счет огромного числа нерелевантных документов, попавших в отклик. Например, в работе Чуна отмечается, что Google, фокусируясь на релевантности первых результатов, мало заботится о числе ответов и качестве низкоранжированных документов[2].

Однако для небольших текстовых корпусов, таких как литература по узкой специальности или архивы статей журналов, подход, ограничивающийся улучшением только первых результатов неприменим по причине малого общего числа документов в отклике[3]. В таком случае часто возникает ситуация, в которой пользователь поисковой системы просматривает все выданные ему результаты. Из-за этого исследователям приходится принимать во внимание точность (отношение числа найденных релевантных документов, к числу документов, возвращенных системой) всей выдачи поисковика, не имея возможности переложить решение проблемы на ранжирование.

В последние годы появились многочисленные поисковые сервисы, стре-

[1]Roberts L. G. Beyond Moore's Law: Internet Growth Trends // Computer. – 2000. – Vol. 33, no. 1. – P. 117-119.

[2]Choon H. D., B. Rajkumar. Guided Google: A Meta Search Engine and its Implementation Using the Google Distributed Web Services // International Journal of Computers and Applications. – Vol. 26. – ACTA Press, 2004. – P. 181-187.

[3]A Scalable Topic-Based Open Source Search Engine / W. Buntine, J. Lofstrom, J. Perkio et al. // Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence. – WI '04. – Washington : IEEE Computer Society, 2004. – P. 228-234.

мящиеся усовершенствовать поисковые технологии, выходя за рамки стандартного поиска по ключевым словам[4]. Разработчики поисковых систем стали использовать более сложные модели представления документов для наиболее эффективного использования имеющихся в этих документах данных.

Одним из направлений подобных исследований является использование специальной лексики (терминов или терминологических словосочетаний, которые в дальнейшем понимаются как синонимы) предметных областей для улучшения качества (здесь и далее в смысле точности и полноты) поиска[5].

Степень разработанности темы исследования. Проблемам поиска по математическим статьям посвящены работы большого количества исследователей. Работы Дж. Мисутка, Л. Галамбо, М. Кохлзасе, И. Сьюкана, П. Либбрехта, Е. Мелиса, М. Адиля, Х.С. Чонга, С.Х. Кияля, В.Д. Соловьева, А.М. Елизарова, Н.Г. Жильцова, О.А. Невзоровой в значительной мере способствовали развитию данной тематики.

Большое внимание к поиску с использованием терминов предметной области в своих работах уделяют Д.В. Джоханссон[5] и Ю.-Х. Лью[6]. В работе Джоханссона термины используются для увеличения весовых коэффициентов слов из контекста этих терминов в ранжирующей функции. Исследования Лью были ограничены использованием индекса словаря MeSH, в который вручную заносится информация о научных статьях по медицинской тематике. Для каждой статьи индекса установлены связи с терминами словаря, что позволило отказаться от процедуры автоматического извлечения терминологических словосочетаний. Однако оба этих подхода ограничиваются исполь-

[4]Koster C. H. A., Seibert O., Seutter M. The PHASAR search engine // Proceedings of the 11th international conference on Applications of Natural Language to Information Systems. – Berlin : Springer-Verlag, 2006. – P. 141-152.

[5]Johannsson D. V. Biomedical information retrieval based on document-level term boosting : Ph. D. thesis / D. V. Johannsson; Norwegian University of Science and Technology. – 2009. – 69 p.

[6]Liu Y.-H. On the potential search effectiveness of MeSH (medical subject headings) terms // Proceedings of the third symposium on Information interaction in context. – New York : ACM, 2010. – P. 225-234.

зованием терминов только для повышения качества ранжирования (точности первых результатов), а задача увеличения точности всей выдачи поисковой системы не рассматривается.

Цели и задачи диссертационной работы: Исследовать способы использования терминов в тексте на качество информационного поиска. Предложить модели и методы организации информационного поиска в научных статьях с использованием специальной лексики предметной области. Разработать на основе предложенных моделей информационно-поисковую систему для работы с архивом статей математического журнала.

Для достижения поставленных целей были решены задачи:

- разработка метода автоматического извлечения терминов для задач информационного поиска;
- разработка алгоритма автоматической генерации приближения к словарю терминов предметной области;
- извлечение метаданных и библиографических ссылок статей и их представление в виде связанных данных;
- разработка модели информационного поиска с использованием информации о терминах в тексте;
- построение прототипа системы терминологического поиска на базе полнотекстовой поисковой платформы;
- разработка функции ранжирования на основе метрики схожести терминологических словосочетаний в тексте и запросе;
- оценка эффективности полученных в ходе исследования алгоритмов;
- оценка качества поиска.

Научная новизна. Предложенный метод поиска отличается от предшествующих подходов к обработке научно-технических текстов тем, что фокусирует внимание на терминологических словосочетаниях статей для увеличения точности выдачи. Использование для ранжирования терминов, выделенных в тексте, позволяет получить результаты по качеству поиска, превосходящие уровень существующих решений поставленной задачи.

Разработанный подход к извлечению терминологии имеет высокий показатель полноты результатов (отношения числа найденных релевантных документов, к числу релевантных документов в базе) для математических текстов, при этом допуская перенос на другие естественно-научные области.

Теоретическая и практическая значимость. Возможность использования разработанных моделей извлечения специальной предметной лексики и моделей использования извлеченных терминов для организации информационного поиска в текстах соответствующей предметной области составляет теоретическую значимость исследования. Проведенные сравнения эффективности поиска могут служить основанием для дальнейших исследований в данной области.

Практическая значимость работы заключается в разработке прототипа информационно-поисковой системы на основе построенных моделей и алгоритма поиска с использованием специальной лексики предметной области. Прототип может быть использован в качестве специализированного поискового сервиса для корпуса текстов узкой предметной области, например архива статей научного журнала.

Методология и методы исследования. Исследования проводились в рамках направлений автоматического извлечения терминов и информационного поиска. Проводились эксперименты на реальных данных, полученные результаты сравнивались по наиболее важным критериям. Использовались общепринятые метрики сравнения эффективности: точность (precision), пол-

нота (recall), сбалансированная F-мера, усредненная средняя точность (mean average precision) и nDCG (normalized discounted cumulative gain).

Положения, выносимые на защиту: В работе получены следующие основные результаты:

1. Разработан алгоритм автоматической генерации приближения словаря русскоязычных терминов предметной области по корпусу статей этой предметной области и общенаучному словнику. Исследовано влияние статистических фильтров на качество автоматического построения словников. Разработан метод автоматического выделения терминологических словосочетаний с использованием данных словников.

2. Разработана модель информационного поиска, использующая специальную лексику для улучшения качества результатов в отклике. Получены оценки сложности алгоритмов информационного поиска.

3. Разработан алгоритм обхода web-интерфейсов научных публикаций для сбора метаданных с последующим их представлением в виде связанных данных. Разработана функция ранжирования, базирующаяся на схожести терминологических словосочетаний в тексте и запросе. Проведено сравнение данной функции с распространенными функциями ранжирования документов применительно к задаче информационного поиска по корпусу научных статей.

4. Реализован прототип информационно-поисковой системы для корпуса математических статей с использованием разработанных методов и алгоритмов[7]. Проведены оценки качества поиска и сложности алгоритмов по времени.

[7]Прототип доступен по адресу: http://searchsh.zapto.org/ Исходные коды доступны по адресу: https://github.com/ksugltronteal/search.sh/

Степень достоверности и апробация результатов. Основные результаты диссертации докладывались на следующих конференциях: Казанская школа по компьютерной и когнитивной лингвистике TEL-2009, TEL-2012 (Казань, 2009, 2012), Интернет и современное общество (Санкт-Петербург, 2010), Научная сессия МИФИ-2011 (Москва, 2010), Системный анализ и семиотическое моделирование SASM-2011 (Казань, 2011), Шестая Российская конференция молодых ученых по информационному поиску RuSSIR-2012 (Ярославль, 2012), Итоговая научная конференция КФУ (Казань, 2010-2012).

Также результаты освещались на научном семинаре Института системного программирования РАН, семинаре «Методы построения информационных систем» МГУ, республиканском научном семинаре КНИТУ «Методы моделирования» и семинарах КФУ «Когнитивное моделирование и компьютерная лингвистика» и «Актуальные информационные технологии».

Публикации. Материалы диссертации опубликованы в 10 печатных работах, из них 3 статьи в рецензируемых журналах в изданиях, рекомендованных ВАК РФ [1–3], 1 статья включена в реферативную базу данных Scopus [4], 5 статей в сборниках трудов конференций [5–9] и 1 тезисы доклада [10]. Список работ приведен в конце автореферата.

В публикации [8] вклад автора заключается в извлечении метаданных статей и обработке библиографических ссылок. В статьях [3, 4] автору принадлежит основополагающий вклад в части построения библиографически связанной коллекции и преобразования метаданных в RDF.

Личный вклад автора. Все представленные в диссертации результаты получены лично автором.

Структура и объем диссертации. Диссертация состоит из введения, 4 глав, заключения и библиографии. Общий объем диссертации 125 страниц, из них 102 страницы текста, включая 21 рисунок. Библиография включает 122 наименования на 16 страницах.

Содержание работы

Во введении обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследований, показана практическая значимость полученных результатов, представлены выносимые на защиту научные положения.

В первой главе рассмотрена задача автоматического извлечения специальной лексики, основные проблемы, с которыми сталкиваются исследователи при ее решении, и базовые характеристики эффективности, на которые обычно опираются при сравнении различных методов решения этой задачи. Проанализированы работы по этой тематике С. Ананиадоу, Е. Брилла, Е. Бонтаса, Ф. Смадья, К. Франци, Л. Хиршмана, А. Моргана, М. Барони, С. Бернардини, Ф. Крауфхаммера, Г. Ненадика, П. Веларди, Ю.М. Логачева, Н.В. Лукашевич, П.И. Браславского. Описаны основные подходы решения задачи автоматического извлечения терминов: словарный, основанный на правилах, статистический, опирающийся на машинное обучение, с использованием поисковых машин интернета.

Рассмотрены все вышеперечисленные группы подходов, системы, их реализующие, проблемы с которыми сталкиваются исследователи и разработчики этих систем, факторы, влияющие на конечную эффективность.

Проанализированы основные области применения терминов для задач информационного поиска: расширение запросов, межъязыковой поиск, классификация текстов, извлечение ключевых фраз. Большое количество работ А. Хотхо, Ю. Ву, К. Ли, Р.С. Бота, Дж. Гринберга по данной тематике указывают на высокую актуальность исследований в этой области.

В конце главы описаны особенности поиска по математическим документам и научным статьям, проведен анализ проблем этих областей и основных подходов к их решению.

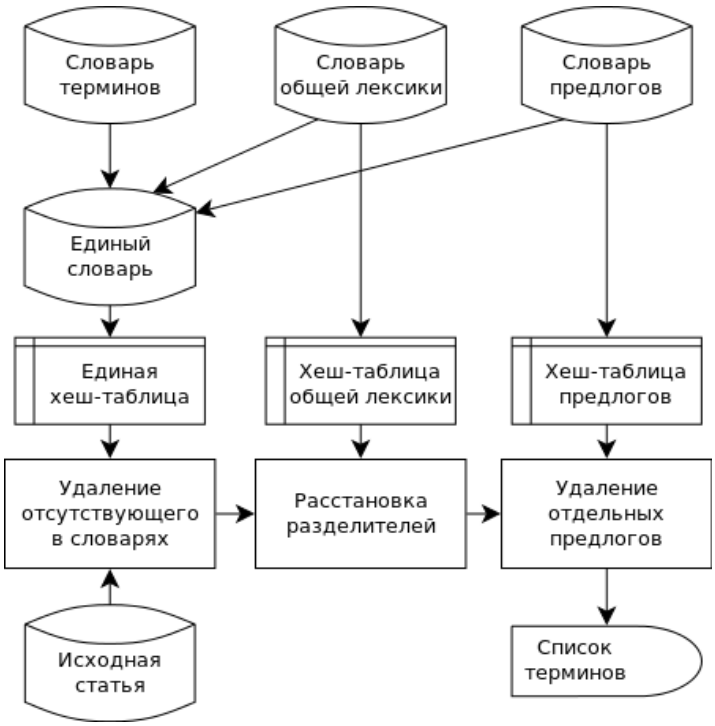

Рис. 1. Схема алгоритма автоматического выделения терминологии

Во второй главе «Автоматическое извлечение терминологии» рассмотрена проблема автоматического извлечения русскоязычной специальной предметной лексики для дальнейшего построения информационно-поисковой системы на основе использования этой специальной лексики.

Проведено исследование ключевых характеристик подходов к извлечению специальной лексики применительно к задаче информационного поиска. В качестве основных параметров выступала полнота результатов. В качестве рабочего, как один из лучших, был выбран подход, основанный на шаблонах, использующих специальные терминологические словари.

Спроектирован и реализован новый подход автоматической генерации словника специальной лексики предметной области на основе словника общенаучной лексики, построенного вручную, и корпуса статей, принадлежащих этой области. Общая схема алгоритма приведена на рисунке 1.

В основе решения лежат хеш-таблицы с открытой адресацией и линейным пробированием. Хеш-таблицы базируются на хеш-функции, основанной

на делении: $H(s) = \left(\sum_{i=1}^{|s|} (C(s[i]) \cdot i)\right) \, mod \, I$, где s – строка, от которой вычисляется значение хеш-функции; $s[i]$ – i-ый символ строки s; $|s|$ – длина строки s; $C(s[i])$ – номер символа $s[i]$; mod – операция вычисления остатка от деления; I – некоторое простое число, равное размеру хеш-таблицы.

Качество словника, сгенерированного автоматически, значительно уступает качеству подходов с использованием ручной работы экспертов. По этой причине результаты подвергаются дополнительной обработке с использованием статистических фильтров. Проведено сравнение различных параметров подобных фильтров с целью максимизации линейной целевой функции: $E(d) = A \cdot P(d) + B \cdot R(d) + C \cdot S(d)$, где $P(d)$ – точность словаря d, $R(d)$ – полнота словаря d, $S(d)$ – размер словаря d, а коэффициенты эмпирически были выбраны $A = 1$, $B = 10$, $C = -0,01$.

В качестве рабочих отбирались фильтры так, что для полученных с их помощью словарей d величина $\Delta E = E(d) - E(d_0) > \varepsilon$, где d_0 – словник, построенный автоматически до применения фильтров, $\varepsilon > 0$ – переменная, добавленная, чтобы исключить из рассмотрения фильтры, дающие малое изменение качества словника. Эмпирически ε была принята равной $0,1$.

Описанный подход к автоматическому извлечению терминов позволяет получать результаты с высокой полнотой, но низкая точность иногда приводит к нежелательному объединению нескольких терминологических словосочетаний в одно. Для решения этой проблемы использована группа синтаксических правил, которая позволяет проводить дополнительное разбиение многословных терминов.

Выделение терминологических словосочетаний в тексте производится по шаблону на основе полученных словников: терминологического (T), общей лексики (G) и связок (P).

В конструкции $g_1 p_1 \ldots p_k t_1 t_2 \ldots t_n q_1 \ldots q_m g_2$

$t_1 t_2 \ldots t_n$ – терминологическое словосочетание, если:

$g_1, g_2 \in G$; $t_1, t_n \in T$; $t_2, \ldots, t_{n-1} \in T \cup P$; $p_1, \ldots, p_k, q_1, \ldots, q_m \in P$; $0 < n$, $0 \leq k$, $0 \leq m$.

Использование автоматической генерации терминологического словника по корпусу статей предметной области позволяет производить перенос метода на другие предметные области без привлечения дополнительного труда экспертов. Ожидается, что примерно равные результаты могут быть получены для корпусов статей естественно-научных текстов.

В третьей главе описана новая модель информационно-поисковой системы с использованием информации о терминологических словосочетаниях в тексте, а также прототип, созданный на основе этой модели.

Модель основана на свободном текстовом поиске (произвольный порядок слов, любая морфологическая форма) в пределах одного термина-словосочетания. Если все слова запроса принадлежат одному терминологическому словосочетанию, то вне зависимости от их порядка и морфологической формы, документ, содержащий этот термин, будет признан релевантным и выдан пользователю. Во всех остальных случаях документ рассматривается как нерелевантный.

Определены множества:

$D = \{d_1, \ldots, d_n\}$ – множество документов; $TT = T_{d_1}, \ldots, T_{d_n} : T_{d_i} = \{t_i^1, \ldots, t_i^{m_i}\}$ – множества терминологических словосочетаний для документов из D; $t_i^j = \{l_{i,1}^j, \ldots, l_{i,p_{i,j}}^j\}$ – множество лексем термина t_i^j.

Для $Q = \{q_1, \ldots, q_k\}$, где q_i – ключевые слова запроса, результат поискового запроса в общем случае формулируется как $S(Q, D) = \{r_1, \ldots, r_z\} \subset D$.

Результат терминологического поиска определен как $S_t(Q, D, TT) = \{r_1, \ldots, r_z\},\ r_i \in D : \exists t \in T_{r_i} \forall q \in Q \exists l \in t : q = l$.

Сравнение качества результатов поискового запроса обычно производится по метрикам Точности ($Precision$), Точности на уровне n ($Precision@n$)

и Полноты (*Recall*)[8]. $Precision = \frac{|D_{rel} \cap D_{retr}|}{|D_{retr}|}$, $Precision@n = \frac{|D_{rel} \cap D^n_{retr}|}{|D^n_{retr}|}$, $Recall = \frac{|D_{rel} \cap D_{retr}|}{|D_{rel}|}$, где D_{rel} – множество релевантных документов в базе, D_{retr} – множество документов, найденных системой, D^n_{retr} – множество из первых n документов, найденных системой.

F-мера – это взвешенное гармоническое среднее точности P и полноты R: $F = \frac{1}{\alpha\frac{1}{P}+(1-\alpha)\frac{1}{R}}, \qquad \alpha \in [0,1]$.

Проверяется гипотеза, что использование терминов предметной области в тексте позволяет улучшить значение сбалансированной ($\alpha = 0,5$) F-меры: $F_1(S_t(Q,D,TT)) > F_1(S(Q,D))$.

Система (прототип) реализована на базе свободной поисковой платформы Solr. Для этих целей разработаны и реализованы в прототипе методы обращения к поисковой платформе (создание и обновление индекса, обработка запроса), а также интерфейс взаимодействия с пользователем. При разработке особое внимание уделено возможности обрабатывать поисковые запросы в реальном времени. Общие схемы организации этапов построения индекса и обработки запроса информационно-поисковой системы приведены на рисунках 2 и 3.

Проведено исследование задачи извлечения метаданных статей и библиографических ссылок из веб-интерфейсов научных коллекций, разработана реализация одного из подходов к ее решению. Полученные данные представлены в формате, соответствующем принятым подходам к публикации данных в сети Интернет – связанным данным (Linked Data). Использована популярная схема AKT Reference Ontology.

Предложена новая функция ранжирования по схожести терминов из запроса и в индексе на основе метрики близости строк N-Gram Distance[9]. Про-

[8]Manning C. D., Raghavan P., Schutze H. Introduction to Information Retrieval. – Cambridge : Cambridge University Press, 2008. – 482 p.

[9]Kondrak G. N-gram similarity and distance // String Processing and Information Retrieval. – Vol. 3772 of Lecture Notes in Computer Science. – Berlin : Springer, 2005. – P. 115-126.

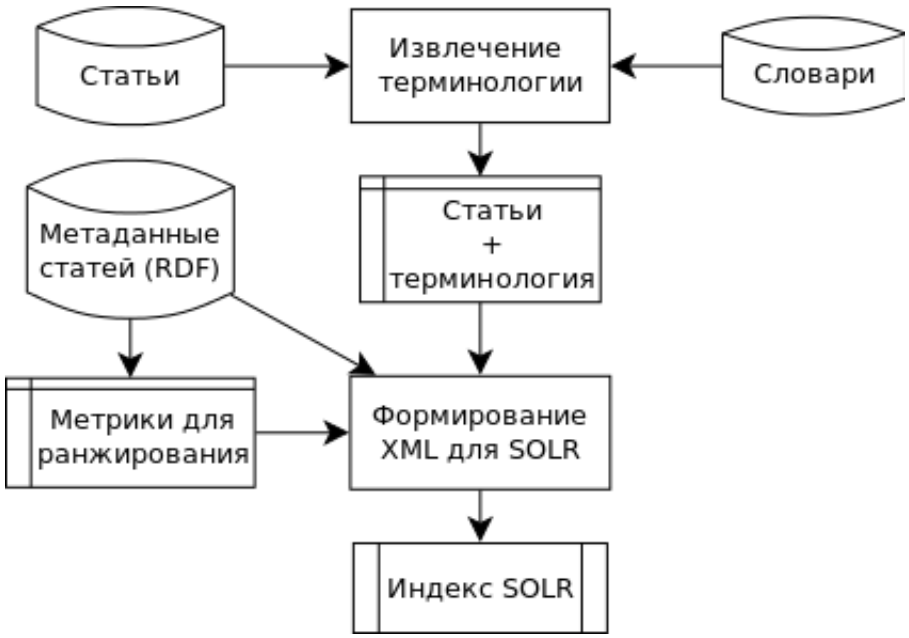

Рис. 2. Схема построения поискового индекса

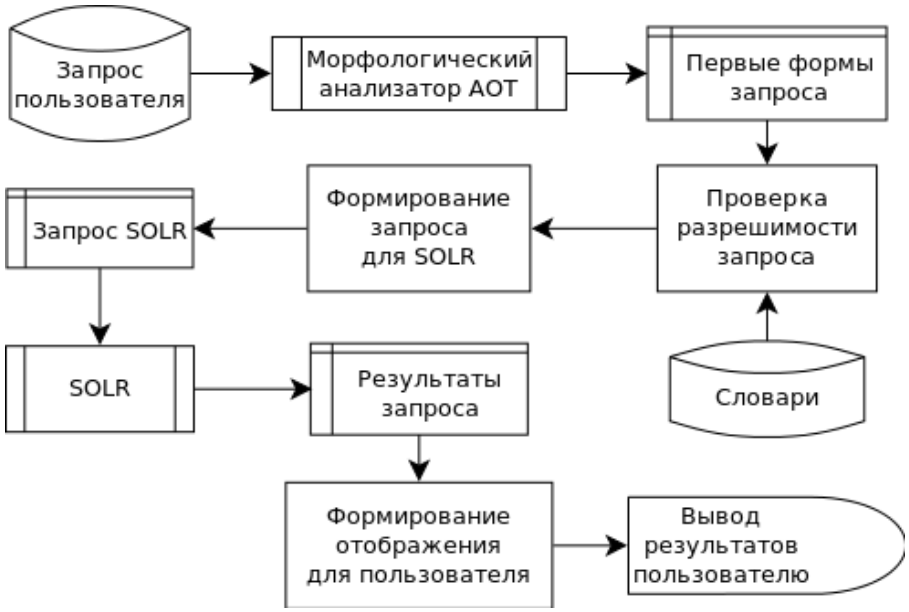

Рис. 3. Схема обработки запроса

ведено сравнение трех различных функций ранжирования:

$Score_1 = F(TF{\cdot}IDF)$ (косинусная близость на базе значения $TF{\cdot}IDF$), $Score_2 = PageRank$, $Score_3 = NGramDist(QT, IT, 2) * \sqrt{|IT|}$,

где $NGramDist(\cdot)$ – метрика N-Gram Distance; QT – термин запроса; IT – термин в индексе; $|IT|$ – длина термина в индексе.

При вычислении $NGramDist$ бралось максимальное значение среди всех удовлетворяющих запросу терминологических словосочетаний документа.

На основании экспериментального сравнения ранжирующих функций выбрана в качестве рабочей и реализована в прототипе функция $Score_3$.

В четвертой главе «Оценки ключевых характеристик информационно-поисковой системы» оценена сложность алгоритмов, используемых в представленном прототипе поисковой системы, а также проведено сравнение качества результатов информационного поиска.

Сложность по времени решения задачи подготовки текстов для загрузки в индекс поисковой системы равняется $O(N+(M+K)logK)$, где N – количество символов в корпусе, M – число статей, а K – число библиографических связей между статьями. Сложность по времени решения задачи подготовки запроса для поисковой платформы равняется $O(N)$, где N – количество символов в запросе.

Ввиду некорректности сравнения разноязычных систем, а также систем с различными индексами, произведено сопоставление качества результатов терминологического поиска с поисковой платформой Solr. В качестве рабочего корпуса выбран архив из 1475 статей журнала «Известия высших учебных заведений. Математика». Все оценки проводились на 100 случайных запросах-терминах, имеющих непустой результат, по крайней мере, в одной из сравниваемых систем. Оценивались метрики точности ($Precision$) и точности на уровне n ($Precision@n$). Обе метрики показали преимущество терминологического поиска над полнотекстовым при условии поиска математических

терминов в тематическом корпусе.

Получены данные о полноте (*Recall*), основываясь на которых можно утверждать, что значение метрики полноты результатов будет выше у полнотекстовой поисковой системы.

Значение интегральной характеристики, сбалансированной F-меры, для терминологического поиска равняется $F_1(S_t(Q, D, TT)) = 0,877$. Полнотекстовый поиск Solr по данным эксперимента показал результат хуже $F_1(S(Q, D)) = 0,806$.

Таким образом, подтверждена гипотеза об улучшении значения F_1-меры в результате использования специальной лексики предметной области в тексте при поиске в корпусе текстов этой предметной области.

Значения вышеописанных характеристик, а также некоторые другие популярные результаты сравнительных оценок приведены в таблице 1.

Таблица 1. Результаты сравнения качества поиска

	$S(Q, D)$	$S_t(Q, D, TT)$
Точность	0,699	0,878
Полнота	0,953	0,877
F_1	0,806	0,877
MAP	0,820	0,843
$nDCG$	0,840	0,948

В Заключении подводится итог разработанной концепции решения задачи предметного информационного поиска в коллекциях научных статей, который опирается на терминологические словосочетания в тексте.

Разработанные в диссертации методы позволяют для предметного корпуса русскоязычных научных текстов автоматически построить словник специальной лексики, соответствующий области этого корпуса, на его основе

извлечь из текста терминологические словосочетания, загрузить их в индекс информационно-поисковой системы и использовать для специализированного тематического поиска.

На основании проведенных экспериментов сделан вывод, что подход к поиску на основе выделения терминов в конкретных предметных областях позволяет увеличить его эффективность.

Список публикаций

1. Заикин Д.А., Соловьев В.Д. Модификация метода поиска по ключевым словам в математических коллекциях // Вестник КГТУ им. А.Н. Туполева. — 2011. — № 1. — С. 136–141.

2. Заикин Д.А. Подход к ранжированию результатов для терминологического поиска // Ученые зап. Казан. ун-та. Серия физ.-мат. науки. — 2014. — Т. 156, № 1. — С. 12–21.

3. Прототип программной платформы для публикации семантических данных из математических научных коллекций в облаке LOD / Невзорова О.А., Жильцов О.А., Заикин Д.А. и др. // Ученые зап. Казан. ун-та. Серия физ.-мат. науки. — 2013. — Т. 154, № 3. — С. 216–232.

4. Bringing Math to LOD: A Semantic Publishing Platform Prototype for Scientific Collections in Mathematics / O. Nevzorova, N. Zhiltsov, D. Zaikin et al. // The Semantic Web - ISWC 2013. — Vol. 8218 of Lecture Notes in Computer Science. — Berlin : Springer, 2013. — P. 369–384.

5. Заикин Д.А. Построение словарей терминов для предметных областей // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2009. — Казань : Отечество, 2010. — С. 71–76.

6. Заикин Д.А., Соловьев В.Д. Сравнение эвристик формирования индекса в терминологическом поиске // Системный анализ и семиотическое моделирование: материалы первой всероссийской научной конференции с международным участивем (SASM-2011). — Казань : Изд-во «Фэн» Академии наук РТ, 2011. — С. 197–204.

7. Заикин Д.А. Метод извлечения метаданных статей из web-интерфейсов научных коллекций в терминах Linked Open Data // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2012. — Казань : Изд-во «Фэн» Академии наук РТ, 2012. — С. 93–101.

8. Khasanshin A., Zaikin D., Zhiltsov N. Indexing mathematical scholarly papers as Linked Open Data // Proceedings of the Sixth Russian Young Scientists Conference in Information Retrieval. — Yaroslavl, 2012. — P. 24–34.

9. Заикин Д.А., Соловьев В.Д. Терминологический поиск в коллекциях математических статей // Интернет и современное общество: Труды XIII Всероссийской объединенной конференции. — СПб. : МПСС, 2010. — С. 80–85.

10. Заикин Д.А., Соловьев В.Д. Новый алгоритм поиска по ключевым словам (на примере коллекции математических текстов) // Научная сессия МИФИ-2011. Труды. — Т. 3. — М. : МИФИ, 2011. — С. 59.