На правах рукописи 

Жильцов Никита Геннадьевич

МЕТОДЫ И АЛГОРИТМЫ ПОИСКА СУЩНОСТЕЙ ПО КЛЮЧЕВЫМ СЛОВАМ В ГРАФОВЫХ БАЗАХ ЗНАНИЙ

Специальность 05.13.11 —
«Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Автореферат

диссертации на соискание учёной степени

кандидата технических наук

Казань — 2016

Работа выполнена на кафедре интеллектуальных технологий поиска Высшей школы информационных технологий и информационных систем ФГАОУ ВО «Казанский (Приволжский) федеральный университет»

Научный руководитель: **Соловьев Валерий Дмитриевич**
доктор физико-математических наук, профессор, ФГАОУ ВО «Казанский (Приволжский) федеральный университет»

Официальные оппоненты: **Кузнецов Сергей Дмитриевич,**
доктор технических наук, профессор,
Институт системного программирования Российской академии наук, главный научный сотрудник

Браславский Павел Исаакович,
кандидат технических наук,
Институт математики и компьютерных наук, ФГАОУ ВО «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина», старший научный сотрудник

Ведущая организация: ФГБОУ ВО «Московский государственный университет имени М.В. Ломоносова»

Защита состоится 23 декабря 2016 г. в 16.00 на заседании диссертационного совета Д 212.081.35 при ФГАОУ ВО «Казанский (Приволжский) федеральный университет» по адресу: 420008 г. Казань, ул. Кремлевская, 35.

С диссертацией можно ознакомиться в библиотеке ФГАОУ ВО «Казанский (Приволжский) федеральный университет» по адресу: 420008, г. Казань, ул. Кремлевская, 35.

Автореферат разослан « » 2016 г.

Ученый секретарь
диссертационного совета
канд.ф.-м.н., доцент Еникеев Арслан Ильясович

Общая характеристика работы

Актуальность темы. Анализ реальных логов современных поисковых машин в Вебе показывает[1,2], что в более чем 75% случаев пользователи запрашивают информацию о конкретных сущностях или группах сущностей, материальных или воображаемых объектах: товарах, людях, организациях, географических местах, событиях, персонажей книг и фильмов и т. п., а также связях между ними. Создание сущностно-ориентированных поисковых приложений – один из актуальных трендов последних лет в индустрии поисковых машин. Примерами таких сервисов являются Google Knowledge Graph[3], Facebook Graph Search[4] и WolframAlpha[5]. Как правило, обработка сущностно-ориентированных поисковых запросов требует проведения подготовительных процедур: извлечение, преобразование, агрегирование и хранение данных из разных источников в виде специального структурированного представления.

В последнее десятилетие формат Resource Description Framework (RDF)[6] стал одним из самых популярных стандартов хранения структурированных данных. Унифицированная модель представления RDF позволяет выражать утверждения о фактах окружающего мира в виде триплетов субъект–предикат–объект, каждый из которых можно интерпретировать как семантическую ссылку в графе между сущностью-субъектом и сущностью-объектом. Благодаря усилиям контент-издателей, разработчиков и исследователей, данные из разнородных источников (например, внутрикорпоративных реляционных баз данных или страниц в Вебе) полуавтоматически переводятся в формат RDF с обогащением семантикой, содержащейся в предметно-ориентированных словарях и онтологиях. Результатом этой работы стало большое «облако» открытых связанных данных (Linking Open Data, LOD cloud)[7], включающее более 300 баз знаний, связанных между собой и находящихся в открытом доступе. Организация поиска по таким ресурсам имеет свои особенности по сравнению с традиционной задачей поиска документов.

[1] Pound J., Mika P., Zaragoza H. Ad-hoc object retrieval in the web of data // Proceedings of the 19th International Conference on World Wide Web. – 2010. – P. 771–780.

[2] Active Objects Actions for Entity Centric Search / T. Lin [и др.] // Proceedings of the 21st WWW. – 2012. – P. 589-598.

[3] https://en.wikipedia.org/wiki/Knowledge_Graph

[4] https://en.wikipedia.org/wiki/Facebook_Graph_Search

[5] http://wolframalpha.com

[6] https://www.w3.org/RDF/

[7] http://lod-cloud.net

Задача поиска сущностей по ключевым словам в RDF-графах была впервые поставлена исследователями из Yahoo! Research[8] и заключается в следующем. Даны граф описаний сущностей G и запрос по ключевым словам Q, отражающий некоторую информационную потребность из заданной классификации. Требуется получить оптимальный отранжированный список идентификаторов сущностей $\pi_{opt} = \{E_1, \ldots, E_k\}$, таких, что $E_i \in G$, и функция потерь L (обычно здесь используются классические меры оценивания поиска) минимальна: $L(\pi_{opt}) = \min_{t \in S_k} L(\pi_t)$, где S_k – число всевозможных перестановок результатов в поисковой выдачи размера k.

В этой же работе выделены основные типы поисковых запросов, каждый из которых, вообще говоря, требует специальных методов разрешения: именные, категорийные, вопросно-ответные (или атрибутивные) и смешанные запросы. А также предложена методология оценивания этой задачи, рассматривающая тестовое множество запросов, тестовую коллекцию данных и сбор размеченных данных средствами краудсорсинга. В продолжении данной работы[9] экспериментально показана стабильность стандартных мер оценивания качества поиска – макроусредненной средней точности (mean average precision, MAP) и нормализованный дисконтированной совокупной выгоды (normalized discounted cumulative gain, NDCG), которые предлагаются как основные целевые меры оценивания для этой задачи и учитывают порядок релевантных результатов в поисковой выдаче, а также дополнительно – уровень релевантности (в случае NDCG).

Большинство предложенных подходов к поиску по ключевым словам в RDF графах использует принцип формирования псевдодокументов, состоящих из набора полей, заполняемых текстовыми утверждениями, упоминающими данную сущность. Далее обычно применяются классические модели ранжирования документов по ключевым словам. В ключевых работах Петера Мики[10] и исследователей из Университета Фрибурга[11] исследовались мо-

[8]Pound J., Mika P., Zaragoza H. Ad-hoc object retrieval in the web of data // Proceedings of the 19th International Conference on World Wide Web. – 2010. – P. 771–780.

[9]Repeatable and reliable search system evaluation using crowdsourcing / R. Blanco [и др.] // Proceedings of the 34th ACM SIGIR Conference. – ACM. – 2011. – P. 923–932.

[10]Blanco R., Mika P., Vigna S. Effective and efficient entity search in RDF data // The Semantic Web – ISWC 2011. – Springer. – 2011. – P. 83–97.

[11]Tonon A., Demartini G., Cudre-Mauroux P. Combining inverted indices and structured search for ad-hoc object retrieval // Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. – ACM. – 2012. – P. 125–134.

дели на основе хорошо известной ранжирующей функции BM25. В работах Кристиана Балога[12,13] исследовалось применение вероятностных языковых моделей.

В научной литературе отмечалась[14,15] необходимость создания новых методов для конкретных типов запросов – вопросно-ответных и категорийных. Также отметим, что предыдущие работы крайне мало использовали выразительную семантическую структуру RDF-графов и, в частности, семантику отношений. Использование такой информации должно было повысить полноту результатов поисковой выдачи за счет расширения контекста для сопоставления описаний сущностей запросу. Также до настоящей работы не была исследована эффективность использования зависимостей между появлениями терминов (например, биграмм) из запроса указанных типов в описаниях сущностей. В то время как модели на основе зависимостей терминов хорошо себя зарекомендовали в задачах стандартного поиска по документам.

Целью диссертационной работы является разработка новых эффективных методов и алгоритмов индексирования графовых баз знаний и поиска сущностей по ключевым словам. Объектами исследований являются графовые базы знаний как модели представления данных в виде графа, в котором вершинами являются описания сущностей (людей, организаций, стран, геолокаций, объекты научного знания и т. д.), ребрами – семантические отношения между сущностями. Примеры графовых баз знаний, выбранных в работе для экспериментов, – граф проекта Linking Open Data в формате RDF, граф набора данных DBpedia в формате RDF и онтологии для представления математических результатов из научных публикаций. Разрабатываемые методы индексирования должны быть применимы как к текстовому содержимому в описаниях сущностей в RDF-данных (например, литералам), так и к полуструктурированным документам и учитывать их логическую структуру. В качестве модели предметной области для извлечения RDF-данных из полуструктурированных документов с учетом структуры выбрана область

[12]Balog K., Bron M., De Rijke M. Query Modeling for Entity Search based on Terms, Categories, and Examples // ACM Transactions on Information Systems. – ACM. – 2011. – Vol. 29. – N. 4. – P. 22:3–22:31.

[13]Neumayer R., Balog K., Norvig K. On the Modeling of Entities for Ad-hoc Entity Search in the Web of Data // In Proceedings of the 34th ECIR. – 2012. – P. 133–145.

[14]Robust question answering over the web of linked data / M.Yahya [и др.] // Proceedings of International Conference on Information and Knowledge Management. – ACM. – 2013. – P. 1107–1116.

[15]Bron M., Balog K., De Rijke M. Example Based Entity Search in the Web of Data. // In Proceedings of the 35th ECIR. – 2013. – P. 392–403.

профессиональной математики. Разрабатываемые методы поиска должны обладать следующими свойствами: они должны быть полностью автоматическими; показатели качества поиска на основе точности и полноты должны превышать аналогичный показатель методов, представленных в современной литературе; методы поиска не должны быть привязаны к предметной области и синтаксису конкретных языков; методы должны обрабатывать основные типы поисковых запросов сущностей; методы должны учитывать как текстовое содержимое описаний сущностей, так и семантические связи между сущностями в графе; методы должны обладать способностью масштабироваться на большие объемы данных (Big Data).

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Разработать подход на основе онтологий для получения структурированного представления в виде графовой базы знаний (RDF-графа) документов математической предметной области.
2. Разработать подход для индексирования текстового содержимого в описаниях сущностей из графовых баз знаний.
3. Разработать метод ранжирования сущностей на основе зависимости появления терминов из запроса по ключевым словам и структуры описаний сущностей.
4. Разработать метод ранжирования сущностей для запроса по ключевым словам с помощью векторного представления, полученного на основе связей в графе.

Основные положения, выносимые на защиту:

1. Предложена модель FSDM[16], как обобщение модели марковских случайных полей и смеси вероятностных языковых моделей, для машинного обучения ранжированию структурированных документов, и разработан алгоритм для обучения параметров модели на основе координатного подъема.
2. Формулируется и доказывается теорема об эффективном вычислении (за линейное время и с линейным пространством по числу вершин в графе) целевой функции тензорного разложения RESCAL.

[16]Fielded Sequential Dependence Model, кратко FSDM

3. Разработан метод ранжирования сущностей с помощью псевдообратной связи по релевантности на основе векторного представления, полученного в результате тензорного разложения RESCAL.
4. Разработано программное обеспечение, и проведены экспериментальные исследования на стандартных открытых тестовых наборах данных, предназначенных для задач сущностно-ориентированного поиска, показывающие эффективность разработанных моделей и алгоритмов.

Научная новизна:

1. Предложена модель ранжирования структурированных документов, обобщающая известные и практически сильные модели ранжирования, – смесь вероятностных языковых моделей (mixture of language models, MLM)[17] и модель последовательной зависимости Мецлера-Крофта (sequential dependence model, SDM)[18]. Таким образом, в отличие от ранее предложенных функций ранжирования, разработанная модель учитывает как зависимость появления терминов из запроса в документе, так и веса полей, в которых встретились термины из запроса, при оценке релевантности документа запросу.
2. Предложен метод ранжирования сущностей, учитывающий скрытую семантическую информацию об отношениях в графе. Таким образом, в отличие от ранее предложенных, разработанная модель позволяет встроить информацию об отношениях между сущностями в графе в ранжирующую функцию для повышения полноты результатов. Также новым результатом является предложенный способ вычисления целевой функции алгоритма RESCAL, что позволяет применять данный алгоритм для совместного разложения разреженных матриц очень большой размерности на сравнительно доступной вычислительной инфраструктуре.

Практическая значимость. Разработанные методы поиска сущностей по ключевым словам могут применяться для повышения точности поис-

[17]Ogilvie P., Callan J. Combining document representations for known-item search // Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. – ACM. – 2003. – P. 143–150.

[18]Metzler D., Croft W.B. A Markov random field model for term dependencies // Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. – ACM. – 2005. – P. 472–479.

ковых систем для поиска в Вебе, а также поиска по документам организации как приложения в сфере корпоративного документооборота. На основе предложенных методов и подходов разработаны следующие программные библиотеки:

- Ext-Rescal[19] – библиотека на языке Python для эффективного вычисления алгоритма разложения разреженных тензоров RESCAL;
- Anduin[20] – библиотека на языке Scala для обработки RDF-графов в парадигме вычислений MapReduce.

Авторская реализация модели FSDM официально включена в версию 3.9 поисковой библиотеки на языке Java с открытым исходным кодом Galago[21], разрабатываемой Университетом Массачусетса Амхерст (США). Имеется соответствующий акт о внедрении.

Достоверность полученных результатов обеспечивается сравнительными экспериментами на достаточно больших наборах данных, которые доступны публично и могут быть использованы сторонними исследователями для воспроизведения результатов. Результаты экспериментов находятся в соответствии с результатами, полученными другими авторами.

Апробация работы. Основные положения докладывались на следующих конференциях:

- седьмой международной конференции по Семантическому Вебу ESWC (2010 г.) в г. Херсониссос (Греция);
- международной конференции в области Интеллектуального Веба WIMS (2011 г.) в г. Согндал (Норвегия);
- конференции молодых ученых в шестой Российской летней школе по информационному поиску RuSSIR (2012 г.) в г. Ярославль;
- двенадцатой международной конференции по Семантическому Вебу ISWC (2013 г.) в г. Сидней (Австралия);
- двадцать второй международной конференции по управлению знаниями и информацией ACM CIKM (2013 г.) в г. Сан-Франциско (США);
- тридцать восьмой международной конференции по информационному поиску ACM SIGIR (2015 г.) в г. Сантьяго (Чили);

[19] https://github.com/nzhiltsov/Ext-RESCAL
[20] https://github.com/nzhiltsov/Anduin
[21] http://lemurproject.org/galago.php

– семинаре по Семантическому представлению математического знания SRMKW в Филдсовском институте Университета Торонто (2016 г.) в г. Торонто (Канада).

Публикации. Основные результаты по теме диссертации изложены в 12 печатных изданиях, 4 из которых изданы в журналах, рекомендованных ВАК [9-12], 8 из которых изданы в журналах, входящих в базу Web of Science и SCOPUS [1-8].

Личный вклад. Автором проведено исследование предметной области, выполнен основной объем теоретических и экспериментальных исследований, изложенных в диссертационной работе, разработаны программные модули на основе созданных методов. В работе [1] А. Котову принадлежит идея использования марковских случайных полей для задачи поиска сущностей, Ф. Николаев внес вклад в получение экспериментальных результатов, а автору принадлежит предложенная модель ранжирования документов FSDM, алгоритм обучения параметров модели и основные эксперименты по сравнительному оцениванию моделей. В работе [2] Е. Агиштейну – предложения по организации процесса краудсорсинга при оценивании задачи, а автору – предложенный метод ранжирования сущностей и проведенные эксперименты. В работе [3] В.Д. Соловьеву принадлежит постановка задачи извлечения логической структуры из математических документов, а автору – предложенная онтологическая модель, метод анализа логической структуры математических документов и проведенные эксперименты. В работах [4-7,9-12] автору принадлежит онтологическая модель логической структуры математического документа, а также архитектура индексирования и семантического поиска. В работе [8] автору принадлежит прототип структурированного интерфейса для поиска по документам.

Содержание работы

Во **введении** обоснована актуальность исследований, проводимых в рамках данной диссертационной работы, сформулирована цель, ставятся задачи работы, сформулирована научная новизна, показана практическая значимость представляемой работы, и раскрыты основные положения, выносимые на защиту.

Первая глава посвящена постановке задачи поиска сущностей по ключевым словам в графовых базах знаний и обзору основных методов и подходов к ее решению. Целью данной главы является анализ эффективности существующих методов индексирования и ранжирования поисковых результатов. В качестве формата представления графов выбран язык RDF.

Графовой базой знаний G будем называть граф, представляющий собой набор описаний и фактов о сущностях, вершинами которого являются идентификаторы сущностей или примитивы (литералы, числа, даты), ребрами – типизированные отношения (бинарные предикаты). Сущности при этом должны быть отнесены к некоторому классу. Набор валидных классов и отношений определяется с помощью онтологий или словарей метаданных.

При поиске сущностей по ключевым словам принято выделять следующие типы запросов, которые соответствуют разным информационным потребностям:

1. *именованные*: объектом информационной потребности является информация о конкретной сущности, например, человеке, организации, вымышленном персонаже или объекте научного знания. Релевантными являются только те сущности, которые подразумеваются с точностью до лексической многозначности. Например, *Владимир Путин* или *президент России в 2003 году*;
2. *категорийные*: интересуемые объекты – сущности определенного класса (например, президенты России или музыкальные группы). Релевантными являются сущности только этого класса или описание этого класса;
3. *атрибутивные*: релевантные объекты – значения атрибутов (примитивов) некоторых сущностей, например, *где родился Владимир Путин?*;
4. *реляционные*: в этом случае ищутся примеры фактов, отражающих связи между запрашиваемыми сущностями, например, *какая связь между Владимиром Путиным и городом Санкт-Петербург.*

Задача сущностно-ориентированного поиска по RDF-графам ставится следующим образом: дан запрос по ключевым словам Q, который имеет тип T, требуется получить отсортированный список идентификаторов сущностей $E_1, \ldots, E_k$, таких, что $E_i \in G$.

Оценивание задачи производится по крэнфилдской методологии[22]: имеется набор тестовых запросов Q, часть сущностей E_i помечена оценками релевантности (обычно бинарными или тринарными) специально обученными людьми – асессорами. Основными мерами оценивания являются точность, макроусредненная средняя точность (mean average precision, MAP) и нормализованная дисконтированная совокупная выгода (normalized discounted cumulative gain, NDCG), широко применяемые в классических задачах информационного поиска по документам. Эти меры считаются на уровне k, то есть для подсчета используется поисковая выдача (отсортированный список результатов) размером k.

Большинство существующих подходов к поиску сущностей в RDF-графах использует принцип формирования псевдодокументов, состоящих из набора полей (multi-fielded documents), заполняемых значениями атрибутов (примитивов) из триплетов, упоминающих данную сущность. Как правило, индексирование производится со специфичными конфигурациями предобработки (сегментирование, стемминг, лемматизация, фильтрация стоп-слов) для разных полей документа. Далее, при поиске по структурированным документам применяются классические модели ранжирования для документов с полями: MLM, BM25F[23], PRMS[24].

Вторая глава посвящена исследованию задачи индексирования структурированных документов и RDF-графов. Описаны:

- новая онтологическая модель логической структуры математического документа Mocassin;
- метод извлечения логической структуры математических документов на основе модели обучения деревьев принятия решений;
- новый подход для индексирования сущностей в графах баз знаний на основе парадигмы MapReduce.

При исследовании подходов к получению структурированного представления документов рассматривается область профессиональной математики, в качестве документов – научные публикации в формате LaTeX. На основе

[22]Voorhees E.M. The philosophy of information retrieval evaluation // Proceedings of Evaluation of Cross-Language Information Retrieval Systems. – 2002. – No. 2406. – LNCS. – P. 355-–370.

[23]Robertson S., Zaragoza H., Taylor M. Simple BM25 extension to multiple weighted fields // Proceedings of the thirteenth ACM international conference on Information and knowledge management. – ACM. – 2004. – P. 42–49.

[24]Kim J., Xue X., Croft W.B. A probabilistic retrieval model for semistructured data // Advances in Information Retrieval. – Springer. – 2009. – P. 228–239.

тестирования существующих моделей представления математических документов в реальных научных коллекциях (прежде всего, онтологий OMDoc и SALT) автором предложена онтология структуры научных публикаций по математике Mocassin, которая содержит типовые концепты и отношения и эффективно извлекаемые из текстов автоматическими методами.

Предлагаемый метод анализа разбивает задачу структурного аннотирования на две подзадачи: определение (классификация) типов сегментов документа в смысле онтологии; классификация семантических отношений между сегментами. Проведена оценка на двух тестовых коллекциях. Первая коллекция представляла собой выборку из 1031 документа англоязычной коллекции arXiv.org. Вторая коллекция содержала 1355 статей журнала "Известия вузов. Математика" на русском языке за период с 1997 по 2009 гг. При этом качество решения первой задачи оценивается на уровне 89–100% по F1-мере в зависимости от типа сегмента, в то время как более сложная задача извлечения и классификации отношений решается на существенно более низком уровне 61–74% по F1-мере. Предложенный метод классификации типов сегментов основан на алгоритме близости строк q-gram. Метод классификации семантических отношений использует тренировочное множество и модель деревьев принятий решений, построенную на основе следующих характеристик ссылочных предложений: структурных – расположение и тип сегментов, и лексических – глагольные фразы. В работах [4,6,9-11] показано, как полученное структурированное представление транслируется в семантический граф в формате RDF, интегрированный с другими данными из облака LOD.

В этой главе также описан масштабируемый подход к индексированию описаний сущностей в RDF-графах произвольных предметных областей. Для формирования документов с полями, удобных для индексирования, автором была предложена модель универсальных метаданных сущностей. Модель состоит из пяти полей, определения которых даны в Таблице 1, где P_{names} – множество предикатов, обозначающих имена, $P_{datatypes}$ – множество предикатов-примитивов, $P_{categories}$ – множество предикатов, обозначающих категории, $E_{sim}(e)$ – множество сущностей, похожих на сущность e.

Для получения универсального представления был разработан алгоритм распределенного индексирования на основе парадигмы MapReduce, который позволяет индексировать RDF-графы произвольного размера при до-

Таблица 1 — **Универсальная модель метаданных для сущностей**

Поле	**Правило фильтрации**
имена	$o : \exists(e, p, o)\&$ $p \in P_{names} = regex(*[name\|label]\$)$
атрибуты	$o : \exists(e, p, o)\&p \in P_{datatypes}\&p \notin P_{names}$
категории	$o : \exists(e_1, p_1, e_2)\&(e_2, p_2, o)\&$ $\&p_1 \in P_{categories}\&p_2 \in P_{names}$
имена похожих сущностей	$o : \exists(e_1, p_1, e_2)\&(e_2, p_2, o)\&$ $e_2 \in E_{sim}(e_1)\&p_2 \in P_{names}$
имена связанных сущностей	$o : \exists(e_1, p_1, e_2)\&(e_2, p_2, o)\&$ $e_2 \notin E_{sim}(e_1)\&p_2 \in P_{names}$

статочном количестве вычислительных машин в кластере Hadoop. Алгоритм состоит из следующих шагов:

1. На подготовительной стадии: RDF-граф в формате N-triples[25] загружается в распределенную файловую систему HDFS – каждый триплет на отдельной строке в файле;
2. На стадии map: реализуется фильтрация RDF-триплетов по правилам из Таблицы 1;
3. На стадии reduce: реализуется агрегирование RDF-триплетов по составному ключу – субъекту и предикату из каждого триплета;
4. На стадии постобработки: объекты из каждого триплета конкатенируются для формирования соответствующих полей псевдодокумента для сущности. Для коллекции полученных структурированных документов строится традиционный полнотекстовый индекс.

Проведены эксперименты по индексированию двух больших RDF-графов – DBpedia 3.7 и BTC 2009, отличающихся по числу уникальных сущностей в 10 раз. Результаты показывают, что за счет параллелизма и увеличения числа map-узлов, удается добиться всего 5-кратного роста по времени для задачи построения структурированных описаний сущностей. При этом время полнотекстового индексирования полученных структурированных описаний растет линейно. Таким образом, основными достоинствами предложенного подхода являются: горизонтальная масштабируемость по числу машин в кластере Hadoop, модульность и эффективность представления данных с точки зрения применения алгоритмов сжатия.

[25] https://www.w3.org/TR/n-triples/

В **третьей главе** представлен Fielded Sequential Dependence Model (FSDM) – новый метод ранжирования структурированных документов на основе случайных марковских полей и смеси вероятностных языковых моделей.

Рассматривается модель случайных марковских полей для информационного поиска, предложенная Метцлером и Крофтом[26]. Согласно ей, совместное вероятностное мультиномиальное распределение терминов из запроса Q и документа D может быть выражено через произведение потенциальных функций, определенных на кликах c в графе вероятностной модели G:

$$P_{G,\Lambda}(Q,D) = \frac{1}{Z_\Lambda} \prod_{c \in C(G)} \psi(c;\lambda_c), \tag{1}$$

где $\psi(\cdot;\lambda_c) = e^{\lambda_c f(c)}$ – фактор потенциалы (потенциалы клик), неотрицательные функции, параметризованные весами $0 \leqslant \lambda_c \leqslant 1, \sum_c \lambda_c = 1$ и характеристической функцией $f(c)$, Z_Λ – нормирующий множитель, которым можно пренебречь при ранжировании. При этом в качестве характеристических функций $f(c)$ выбираются функции от статистик, моделирующих релевантность документа D по отношению к запросу Q:

1. частота терминов из запроса в документе, сглаженная по Дирихле:

$$f_T(q_i, D) = \log P(q_i|\theta_D) = \log \frac{tf_{q_i,D} + \mu \frac{cf_{q_i}}{|C|}}{|D| + \mu}$$

   где q_i – термин из запроса Q, $tf_{q_i,D}$ – частота появления термина q_i в документе D, $|D|$ – длина документа, μ – параметр распределения Дирихле, который обычно выбирается равным средней длине документа в коллекции, cf_{q_i} – частота термина q_i в коллекции и $|C|$ – общее количество терминов в коллекции;

2. частота биграмм из запроса в документе, сглаженная по Дирихле:

$$f_O(q_{i,i+1}, D) = \log \frac{tf_{\#1(q_{i,i+1}),D} + \mu \frac{cf_{\#1(q_{i,i+1})}}{|C|}}{|D| + \mu}$$

[26]Metzler D., Croft W.B. A Markov random field model for term dependencies // Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. – ACM. – 2005. – P. 472–479.

3. частота биграмм из запроса, встречающихся в документе в рамках «окна» фиксированной длины N:

$$f_U(q_{i,i+1}, D) = \log \frac{tf_{\#uwN(q_{i,i+1}),D} + \mu \frac{cf_{\#uwN(q_{i,i+1})}}{|C|}}{|D| + \mu}$$

Обратим внимание, что взятая за основу модель последовательной зависимости (SDM) рассматривает только всевозможные пары последовательностей терминов из запроса длины 2 (биграммы). $\Lambda = (\lambda_T, \lambda_O, \lambda_U)$ – вектор параметров в диапазоне от 0 до 1 – веса, пропорциональные значимости статистик. Обычная схема весов: $\Lambda = (0.8, 0.1, 0.1)$.

Новая модель FSDM для каждой потенциальной функции в формуле (1) рассматривает правдоподобие термина из запроса по отношению к языковой модели документа $P(q_i|\theta_D)$ как смесь языковых моделей полей в документе D:

$$P(q_i|\theta_D) = \sum_j w_j P(q_i|\theta_D^j),$$
$$j = \overline{1, F}, \sum_j w_j = 1, w_j \geqslant 0$$

Тогда характеристическая функция для терминов из запроса выглядит следующим образом:

$$\tilde{f}_T(q_i, D) = \log \sum_j w_j^T P(q_i|\theta_D^j) = \log \sum_j w_j^T \frac{tf_{q_i,D^j} + \mu_j \frac{cf_{q_i}^j}{|C_j|}}{|D^j| + \mu_j}.$$

По аналогии получаются характеристические функции для упорядоченных и неупорядоченных биграмм $q_{i,i+1}$ из запроса. Для ранжирования документов достаточно вычислить апостериорную вероятность:

$$P_\Lambda(D|Q) = \frac{P_{G,\Lambda}(Q, D)}{P_\Lambda(Q)}.$$

Таблица 2 — **Сравнение моделей ранжирования на коллекции DBpedia 3.7. "*"и "†"обозначают статистически значимые улучшения к результатам MLM-CA и SDM-CA, соответственно. Значимость – по рандомизированному тесту Фишера ($\alpha = 0.05$).**

	Все запросы	
	MAP@100	**P@10**
MLM-CA	0.196	0.206
SDM-CA	0.192 (−2.0%)	0.198 (−3.9%)
LM	$0.175^{*}_{\dagger}$ (−10.7%/−8.9%)	$0.182^{*}_{\dagger}$ (−11.7%/−8.1%)
BM25	0.186 (−5.1%/−3.1%)	0.194 (−5.8%/−2.0%)
MLM-tc	0.181 (−7.7%/−5.7%)	$0.185^{*}_{\dagger}$ (−10.2%/−6.6%)
BM25F-tc	0.182 (−7.1%/−5.2%)	0.192^{*} (−6.8%/−3.0%)
PRMS-tc	$0.176^{*}_{\dagger}$ (−10.2%/−8.3%)	0.186^{*} (−9.7%/−6.1%)
PRMS	$0.136^{*}_{\dagger}$ (−30.6%/−29.2%)	$0.136^{*}_{\dagger}$ (−34.0%/−31.3%)
FSDM	$\mathbf{0.231}^{*}_{\dagger}$ (+17.9%/+20.3%)	$\mathbf{0.231}^{*}_{\dagger}$ (+12.1%/+16.7%)

Получается итоговая формула ранжирования для новой модели FSDM:

$$P_\Lambda(D|Q) \overset{rank}{=} \lambda_T \sum_{q_i \in Q} \tilde{f}_T(q_i, D) \; + \lambda_O \sum_{q_{i,i+1} \in Q} \tilde{f}_O(q_{i,i+1}, D) + \lambda_U \sum_{q_{i,i+1} \in Q} \tilde{f}_U(q_{i,i+1}, D). \quad (2)$$

Доказывается, что при $\Lambda = (1.0, 0.0, 0.0)$ FSDM эквивалентна униграммной модели MLM. Также предложен двухступенчатый алгоритм оптимизации весов модели на основе метода координатного подъема (coordinate ascent, CA). Таким образом, предложенная модель достаточно гибка и имеет отдельные смеси языковых моделей с различными параметрами для терминов и биграмм (упорядоченных и неупорядоченных), что позволяет настроить различные весовые схемы для разных типов запросов из классификации.

Экспериментальные исследования на стандартной коллекции для сущностно-ориентированного поиска DBpedia 3.7[27] (485 запросов разных типов, более 3.5 миллионов сущностей) показали (Таблица 2), что модель FSDM имеет значительный прирост по основным мерам оценивания отношению к моделям MLM и SDM, с оптимизированными значения весов по методу координатного подъема (CA). Статистическая значимость проверялась с помощью рандомизированного теста Фишера на уровне $p < 0.05$.

[27]Balog K., Neumayer R. A Test Collection for Entity Search in DBpedia // Proceedings of the 36th ACM SIGIR. – 2013. – P. 737-–740.

Анализ результатов показал, что модель особенно хорошо работает на именованных и вопросно-ответных запросах, показывая прирост по сравнению с SDM на 52% и 6% по MAP, соответственно.

В **четвертой главе** предложен новый подход к применению информации о семантических типах отношений между сущностями в графе для сущностно-ориентированного поиска. В подходе используется векторное представление сущностей в скрытом пространстве, получаемое в результате совместного разложения матриц смежности для предикатов из графа по алгоритму RESCAL, предложенному М. Никелем, В. Треспом и Х.-П. Кригелем[28]. Автором формулируется и доказывается теорема об эффективном вычислении целевой функции RESCAL. Разработан метод ранжирования сущностей с помощью псевдообратной связи по релевантности на основе векторов, полученных в результате разложения RESCAL, который сводится к подсчету косинусной меры векторов сущностей – RESCAL близости.

Дан RDF-граф в виде набора триплетов субъект–предикат–объект $\langle i, k, j \rangle$. Введем тензор $\mathcal{X}$ размерности $n \times n \times m$, где n – количество сущностей, m – количество отношений в графе. При этом $\mathcal{X}_{ijk} = 1$, если в графе существует хотя бы одно k-е отношение между сущностями i и j, и $\mathcal{X}_{ijk} = 0$ в противном случае. Таким образом, каждый k-й фронтальный срез тензора X_k представляет собой матрицу смежности в RDF-графе по k-му предикату. Очевидно, что для реальных RDF-графов тензор $\mathcal{X}$ будет сильно разрежен.

Пусть r – число скрытых факторов, задаваемое как параметр. Тогда ставится следующая задача тензорной факторизации как совместного разложения матриц смежности:

Задача 1 (Задача оптимизации RESCAL)**.** *Пусть A – плотная матрица размерности $n \times r$, состоящая из векторов сущностей в скрытом пространстве, R_k – плотные матрицы размерности $r \times r$, матрицы взаимодействия скрытых факторов. Тогда необходимо определить оптимальные значения*

[28]Nickel M., Tresp V., Kriegel H.-P. A three-way model for collective learning on multi-relational data // Proceedings of the 28th international conference on machine learning (ICML). – 2011. – P. 809–816.

матриц A и R_k, решая следующую задачу оптимизации:

$$\min_{A,R} f(A,R),$$

$$f(A,R) = \frac{1}{2}\left(\sum_k \|X_k - AR_kA^T\|_F^2\right) + \lambda_A\|A\|_F^2 + \lambda_R\sum_k \|R_k\|_F^2$$

$$X_k = AR_kA^T, k = \overline{1,m}.$$

Для ее решения авторами RESCAL[29] предложен итеративный алгоритм чередующихся наименьших квадратов RESCAL, основанный на правилах поочередного обновления матриц A, R_k. Критерием останова алгоритма является условие, при котором величина $\frac{f(A,R)}{\|\mathcal{X}\|_F^2}$ сходится к задаваемому порогу ϵ. Каждая i-я строка a_i в полученной матрице A соответствует векторному представлению i-й сущности в скрытом семантическом пространстве.

Особенности данной модели представления сущностей следующие:

- модель учитывает глобальные зависимости между отношениями;
- сходство сущностей трактуется через структурную эквивалентность: близкие сущности имеют сходные по типу отношения со сходными сущностями;
- субъекты и объекты в исходных триплетах эквивалентны.

Новый результат, выносимый на защиту, – разработан способ вычисления целевой функции алгоритма RESCAL на основе разреженных матриц, который не требует хранения плотной матрицы большой размерности в памяти. Доказывается теорема о временной и пространственной сложности такого способа вычисления целевой функции.

Теорема 1. *Пусть p – число ненулевых значений в тензоре $\mathcal{X}$, а также k – число фронтальных разрезов $X_k = \mathcal{X}_{:,:,k}$ – значительно меньше n и p: $k \ll n, k \ll p$, и r – число скрытых факторов в модели RESCAL – значительно меньше n: $r \ll n$. Существует способ вычисления целевой функции RESCAL, пространственная сложность которого составляет $O(T), T = \max\{rn, p\}$, а временная сложность – $O(rnp)$.*

В доказательстве теоремы при выводе нового способа вычисления целевой функции RESCAL используются альтернативное определение нормы

[29]Nickel M. Tensor factorization for relational learning // Ph.D. thesis, Ludwig-Maximilians-Universitat Munchen. – 2013.

Фробениуса и стандартные свойства оператора взятия следа матрицы – линейность, инвариантность относительно транспозиции и инвариантность относительно циклических перестановок.

Далее, предлагается новый метод ранжирования поисковых результатов, который использует близость векторов сущностей в скрытом семантическом пространстве, полученных на основе алгоритма RESCAL. Пусть дан упорядоченный список сущностей π_Q, полученный для данного запроса Q по некоторому базовому алгоритму ранжирования (например, MLM). Предлагаемый метод состоит из последовательного решения следующих подзадач:

1. Выбрать типы отношений между сущностями в графе для составления тензора $\mathcal{X}$.
2. Решить задачу оптимизации RESCAL.
3. Вычислить значение для ранжирования $S(Q, E_i)$ через косинусную меру между вектором данной сущности и ближайшим вектором сущности из top-K результатов[30] $\pi_Q^{top} = \{E_1, \ldots, E_K\}$: $S(Q, E_i) = \max_{k:E_k \in \pi_Q^{top}} \cos(a_i, a_k)$.

Естественным предположением данного метода является то, что базовая ранжирующая функция дает достаточно качественные результаты на уровне top-K. Метод можно отнести к классу ранжирующих алгоритмов с псевдообратной связью по релевантности.

В виду сильной разреженности тензора, составляемого на основе RDF-графа, ожидается, что предлагаемый метод будет эффективен в качестве дополнительной характеристики в модели машинного обучения ранжированию, а не отдельной ранжирующей функции. Экспериментально была проверена гипотеза о том, что добавление информации о семантике отношений между сущностями приводит к статистически значимому улучшению качества базового метода ранжирования, дающего хорошие результаты в начале списка. Поэтому базовый метод ранжирования, в качестве которого была взята униграммная модель MLM, усиливался набором характеристик: функциями правдоподобия запроса относительно биграммных языковых моделей по отдельным полям документа и характеристики запроса, независящие от документа.

[30] в экспериментах выбирались top-3 результаты

Таблица 3 — **Сравнение моделей GBRT с лексическими и структурными характеристиками для набора SemSearch 2010/11. ** и * обозначают статистическую значимость по парному t-критерию Стьюдента на уровнях значимости** $p < 0.01$ **и** $p < 0.05$

Характеристики	**NDCG@10**	**MAP@10**	**P@10**
Лексические	0.382	0.265	0.539
+ RESCAL близость	**0.409**** (+ 7.1%)	**0.282** (+ 6.4%)	**0.568*** (+ 5.4%)

Далее, на этих характеристиках обучалась модель машинного обучения ранжированию на основе аддитивного ансамбля регрессионных деревьев (gradient boosted regression trees, GBRT). Таблица 3 содержит результаты сравнения данной модели, обученной на наборах характеристик с включением и без RESCAL близости. В качестве тестовой коллекции использовался RDF-граф BTC 2009, содержащий описания более 183 миллионов сущностей. В качестве набора тестовых запросов брались 142 именованных запроса данных из Yahoo! SemSearch Challenge 2010/11[31].

Результаты показывают, что добавление RESCAL близости как характеристики в модель обучения ранжированию ведет к значительным улучшениям по NDCG (с уровнем значимости $p < 0.01$) и P@10, а также к лучшим показателям по MAP. Это подтверждает нашу гипотезу о том, что информация о семантике отношений между сущностями в графе в виде RESCAL близости, позволяет эффективно расширить поисковый контекст.

В **заключении** приведены основные результаты работы, которые состоят в следующем:

1. Предложена модель FSDM – обобщение модели марковских случайных полей и смеси вероятностных языковых моделей – для машинного обучения ранжированию структурированных документов и разработан алгоритм для обучения параметров модели на основе координатного подъема.
2. Формулируется и доказывается теорема об эффективном вычислении (за линейное время и с линейным пространством по числу вершин в графе) целевой функции тензорного разложения RESCAL.

[31] Доступен по адресу: https://github.com/nzhiltsov/YSC-relevance-data

3. Разработан метод ранжирования сущностей с помощью псевдообратной связи по релевантности на основе векторного представления, полученного в результате тензорного разложения RESCAL.
4. Разработано программное обеспечение, и проведены экспериментальные исследования, показывающие эффективность разработанных моделей и алгоритмов для задач поиска сущностей в графовых базах знаний.

Публикации автора по теме диссертации

Статьи в журналах и сборниках, входящих в базы Scopus и WoS

1. Zhiltsov, N. Fielded Sequential Dependence Model for Ad-Hoc Entity Retrieval in the Web of Data / N. Zhiltsov, A. Kotov, F. Nikolaev // Proceedings of the 38th Internationl ACM SIGIR Conference on Research and Development in Information Retrieval. – ACM. – 2015. – P. 253–262. – 0.63 п.л. // 0.5 п.л.

2. Zhiltsov, N. Improving Entity Search over Linked Data by Modeling Latent Semantics / N. Zhiltsov, E. Agichtein // Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. – ACM. – 2013. – P. 1253–1256. – 0.25 п.л. // 0.22 п.л.

3. Zhiltsov, N. Logical Structure Analysis of Scientific Publications in Mathematics / V. Solovyev, N. Zhiltsov // Proceedings of the International Conference on Web Intelligence, Mining and Semantics. – ACM. – 2011. – P. 21:1–21:9. – 0.5 п.л. // 0.44 п.л.

4. Zhiltsov, N. Bringing Math to LOD: A Semantic Publishing Platform Prototype for Scientific Collections in Mathematics / O. Nevzorova, N. Zhiltsov, D. Zaikin, O. Zhibrik, A. Kirillovich, V. Nevzorov, E. Birialtsev // The Semantic Web – ISWC 2013. – Springer. – 2013. – P. 379–394. – 1 п.л. // 0.22 п.л.

5. Zhiltsov, N. OntoMathPro Ontology: A Linked Data Hub for Mathematics / O. Nevzorova, N. Zhiltsov, A. Kirillovich, E. Lipachev // Knowledge Engineering and the Semantic Web. – Springer. – 2014. – P. 105–119. – 0.94 п.л. // 0.25 п.л.

6. Zhiltsov, N. Mathematical Text Collections: Annotation and Application for Search Tasks / O. Nevzorova, E. Birialtcev, N. Zhiltsov // Scientific and

Technical Information Processing. – Allerton Press. – 2013. – Vol. 40. – No. 6. – P. 386–395. – 0.63 п.л. // 0.19 п.л.

7. Zhiltsov, N. Mathematical Knowledge Representation: Semantic Models and Formalisms / A. Elizarov, A. Kirillovich, E. Lipachev, O. Nevzorova, V. Solovyev, N. Zhiltsov // Lobachevskii Journal of Mathematics. – Maik Nauka-Interperiodica Publishing. – 2014. – Vol. 35. – No. 4. – P. 348–354. – 0.63 п.л. // 0.13 п.л.

8. Zhiltsov, N. Publishing Math Lecture Notes as Linked Data / C. David, M. Kohlhase, C. Lange, F. Rabe, N. Zhiltsov, V. Zholudev // The Semantic Web: Research and Applications. – Springer. – 2010. – P. 370–375. – 0.31 п.л. // 0.03 п.л.

Статьи в изданиях, рекомендованных ВАК РФ

9. Жильцов, Н.Г. Коллекции математических текстов: аннотирование и применение в поисковых задачах / О.А. Невзорова, Н.Г. Жильцов, Е.В. Биряльцев // Искусственный интеллект и принятие решений. – 2012. – №3. – С. 51–62. – 0.75 п.л. // 0.31 п.л.

10. Жильцов, Н.Г. Прототип программной платформы для публикации семантических данных из математических научных коллекций в облаке LOD / О.А. Невзорова, Н.Г. Жильцов, Д.А. Заикин, О.Н. Жибрик, А.В. Кириллович, В.Н. Невзоров, Е.В. Биряльцев // Ученые записки Казанского государственного университета. – 2012. – Т. 154. – №3. – С. 216–232. – 1 п.л. // 0.22 п.л.

11. Жильцов, Н.Г. Методы анализа семантических данных математических электронных коллекций / Е.В. Биряльцев, А.М. Елизаров, Н.Г. Жильцов, Е.К. Липачев, О.А. Невзорова, В.Д. Соловьев // Научно-техническая информация. Серия 2: Информационные процессы и системы. – 2014. –№4. – С. 12–17. – 0.38 п.л. // 0.06 п.л.

12. Жильцов, Н.Г. Онтологии математического знания и рекомендательная система для коллекций физико-математических документов / А.М. Елизаров, А.Б. Жижченко, Н.Г. Жильцов, А.В. Кириллович, Е.К. Липачев // Доклады Академии Наук. – 2016. – Т. 467. – №4. – С. 392–395. – 0.25 п.л. // 0.06 п.л.