Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

на правах рукописи

Фролов Дмитрий Сергеевич

Агрегированное представление текстов для задач поиска в коллекциях текстовых документов

РЕЗЮМЕ ДИССЕРТАЦИИ
на соискание учёной степени
кандидата компьютерных наук НИУ ВШЭ

Москва, 2019

Диссертационная работа выполнена в Национальном исследовательском университете «Высшая школа экономики».

Научный руководитель: **Миркин Борис Григорьевич**, д.т.н., Ординарный профессор, Национальный исследовательский университет «Высшая школа экономики»

Тема диссертации

Актуальность исследования

Задачи автоматизации анализа текстов и, прежде всего, задачи их поиска получают всё большую актуальность в связи с современными процессами глобализации и дигитализации общества. Задачи поиска и извлечения текстовых документов широко освещены в современной литературе, в частности, в [3, 16].

Актуальность данной работы определяется постоянно увеличивающимися объемами дигитализованной текстовой информации. Среди первоочередных – необходимость дальнейших продвижений в следующих направлениях:

1) Повышение скорости поиска документов;
2) Повышение качества поиска документов;
3) Обеспечение возможности распределенного хранения документов коллекции;
4) Обеспечение возможности параллельной и неодновременной обработки данных;
5) Автоматизация анализа текстовой информации, включая ее структурирование и интерпретацию.

Переход от совокупностей документов к их **агрегированному представлению** – один из наиболее эффективных способов продвижения в указанных направлениях. Агрегирование, то есть процесс объединения, укрупнения неких объектов по каким-либо общим признакам для получения обобщенных, совокупных показателей, может быть рассмотрено как преобразование исходных данных в новую модель с существенно меньшим числом переменных или ограничений, дающую иное описание изучаемого процесса или объекта.

Многие методы агрегирования используют признаковое описание документов [2, 11], невозможное без существенной предобработки текстов. Существует и другой подход к агрегированию, не требующий предобработки, в котором кодируются не признаки, а произвольные фрагменты текста – так называемый метод аннотированного суффиксного дерева (АСД) [7, 5]. Методы поиска, использующие АСД, могут иметь определенные преимущества, например, при выполнении поиска документов по неточным запросам (возможно, содержащим неточности написания, ошибки). Методы на основе признакового описания сложно адаптируемы к таким ситуациям.

Цель и задачи исследования

Целью исследования является разработка эффективных методов поиска и анализа текстовых данных на основе использования аннотированных суффиксных деревьев для представления коллекций документов, а также программная реализация разработанных методов, их экспериментальная апробация и валидация. Для достижения этой цели планируется решить следующие **задачи**:

1) Разработать эффективный алгоритм поиска и извлечения текстов на основе их представления аннотированными суффиксными деревьями;
2) Разработать варианты этого алгоритма для параллельных и распределенных вычислений, а также неодновременной обработки данных;
3) Адаптировать разработанные алгоритмы для работы с динамически изменяющимися коллекциями документов (случаи вставки, удаления, изменения текста документов);
4) Разработать методику интерпретации результатов информационного поиска с использованием кластерного анализа и таксономического представления предметной области;
5) Разработать математическое обеспечение и провести вычислительные эксперименты по обоснованию эффективности разработанных программ;
6) Провести вычислительные эксперименты по сравнению разработок диссертации с существующими методами;
7) Применить разработанную технологию в задачах реального поиска и извлечения информации.

Объект и предмет исследования

Объектом диссертационного исследования является область информационного поиска. **Предметом исследования** – применение агрегированного представления текста с использованием АСД для задач поиска в коллекциях текстовых документов, структурирование коллекций, интерпретация коллекций текстовых документов с помощью таксономий.

Методы исследования и достоверность результатов

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

1) Метод представления текста в виде аннотированного суффиксного дерева (АСД), а также метод вычисления релевантности строки тексту на основе АСД;
2) Методы индексирования и ранжирования для задач информационного поиска;
3) Метод нечеткой кластеризации FADDIS;
4) Таксономическое представление предметных областей, в особенности, науки о данных.

Достоверность результатов работы подтверждается строгостью математических выкладок, использованных моделей и преобразований, тестированием их программных реализаций, а также тщательной экспериментальной валидацией всех разработанных методов и экспериментами по сравнению предложенных методов с существующими подходами к аналогичным задачам.

Результаты, выносимые на защиту

В ходе работы над диссертацией были получены следующие основные результаты:

1) Разработан новый метод информационного поиска АСДП. Выполнена программная реализация поисковой системы, основанной на данном методе. Эффективность метода доказана вычислительными экспериментами по сравнению АСДП с популярными современными методами поиска, в том числе, и специализированными для задач нечеткого поиска. Сравнение показало качественное преимущество разработанного метода в задачах нечеткого поиска и хороший баланс его качественных характеристик и производительности.
2) Разработан метод интерпретации результатов поиска с помощью формирования и оптимального обобщения нечетких кластеров в таксономии предметной области (ПарГеНМ). Сделана программная реализация разработанного метода. Экспериментальная апробация метода ПарГеНМ проведена на коллекции публикаций издательства Springer в области науки о данных.
3) Разработан метод расширения аудитории интернет-рекламы как задачи информационного поиска на основе оптимального обобщения сегментов

пользователя в таксономии сегментов пользовательских интересов (ОПС). Данное приложение может рассматриваться в качестве примера задачи, в которой эффект оптимального обобщения измерим. Новый метод расширения аудитории рекламных кампаний успешно внедрен на практике.

Научная новизна исследования

Впервые разработана методика информационного поиска для коллекций документов, представленных в виде аннотированных суффиксных деревьев (АСД). Разработан новый метод информационного поиска АСДП. Выполнена программная реализация поисковой системы, основанной на данном методе. Эффективность метода доказана вычислительными экспериментами по сравнению АСДП с популярными современными методами поиска, в том числе, и специализированными для задач нечеткого поиска. Сравнение показало качественное преимущество разработанного метода в задачах нечеткого поиска и хороший баланс его качественных характеристик и производительности.

Разработан и программно реализован метод интерпретации результатов поиска с помощью оптимального обобщения нечетких кластеров в таксономии предметной области (ПарГеНМ). Проведена его экспериментальная апробация и валидация, в том числе, при структурировании и интерпретации научных публикаций в области науки о данных.

Разработан метод расширения аудитории интернет-рекламы на основе оптимального обобщения сегментов пользователя в таксономии предметной области (ОПС). Произведена программная реализация и апробация разработанного метода расширения аудитории ОПС с использованием реальных рекламных кампаний в Интернете. Новый метод расширения аудитории рекламных кампаний успешно внедрен на практике в коммерческой организации.

Теоретическая значимость и практическая ценность

Теоретическая значимость диссертации заключается в том, что в работе предложен метод информационного поиска АСДП и его модификации для параллельных, распределенных и динамических вычислений, а также алгоритм поиска оптимального обобщения нечетких кластеров в таксономии – ПарГеНМ. Кроме того, разработаны методики использования алгоритма оптимального обобщения в

таксономиях ПарГеНМ для задач:

1) Структурирования и интерпретации текстовых коллекций;
2) Повышения эффективности рекламного таргетинга, рассматриваемого как специальный случай задачи информационного поиска.

Практическая значимость работы состоит в следующем:

1) Экспериментально обоснована эффективность метода АСДП в задачах информационного поиска, в том числе, и для нечеткого;
2) Проведен анализ тенденций развития области науки о данных по массиву научных статей, опубликованных в журналах издательства Springer в 1998-2017 годах;
3) Предложена и внедрена методика для эффективного расширения аудитории интернет-рекламы.

Апробация и публикация результатов работы

Результаты работы представлены на следующих конференциях:

1) RuSSIR 2015 (Young Scientists Conference), участие в постер-сессии с докладом «Aggregate Text Representation for Information Retrieval in Collections of Text Documents», 24-28 августа 2015, г. Санкт-Петербург.
2) Летняя школа Факультета компьютерных наук НИУ ВШЭ, участие в постер-сессии с докладом «Using Annotated Suffix Trees for Fuzzy Full Text Search», 27-29 мая 2016, п. Вороново Московской обл.
3) RuSSIR-2016 (Young Scientists Conference), участие в постер-сессии с докладом «Using Annotated Suffix Trees for Fuzzy Full Text Search», 22-26 августа 2016, г. Саратов.
4) 3-й Колмогоровский семинар по компьютерной лингвистике и наукам о языке, участие в постер-сессии с докладом «Annotation of a Document Collection by Finding Thematic Fuzzy Clusters and Parsimoniously Lifting Them in a Domain Taxonomy», 25 апреля 2018, ФКН НИУ ВШЭ, г. Москва.
5) Общемосковский научный семинар «Математические методы анализа решений в экономике, бизнесе и политике», доклад на тему «Рубрикация коллекции документов с помощью формирования тематических нечетких кластеров и их

оптимального подъема в таксономии предметной области», 16 мая 2018, НИУ ВШЭ, Москва.

6) Общемосковский научный семинар «Математические методы анализа решений в экономике, бизнесе и политике», доклад на тему «Обобщение в таксономиях: модель, метод, приложения», 15 мая 2019, НИУ ВШЭ, Москва.

7) IARIA Content 2019, доклад на тему «Method for Generalization of Fuzzy Sets», 5-9 мая 2019, г. Венеция, Италия.

8) ICAISC-2019, участие в постер-сессии с докладом «Method for Generalization of Fuzzy Sets», 16-20 июня 2019, г. Закопане, Польша.

9) IEEE 2019 International Conference on Fuzzy Systems, доклад на тему «Using Taxonomy Tree to Generalize a Fuzzy Thematic Clusters», 23-26 июня 2019, г. Нью-Орлеан, США.

10) World Congress on Global Optimization – 2019, доклад на тему «Globally Optimal Parsimoniously Lifting a Fuzzy Query Set Over a Taxonomy Tree», 8-10 июля 2019, г. Мец, Франция.

11) IHIET-2019, участие в постер-сессии с докладом «A Method for Audience Extending in Programmatic Advertising by Using Parsimonious Generalization of User Segments», 22-24 августа 2019, г. Ницца, Франция.

По теме диссертации опубликованы статьи, перечисленные ниже.
Публикации повышенного уровня:

1) *Frolov D., Mirkin B., Nascimento S., Fenner T.* Using Taxonomy Tree to Generalize a Fuzzy Thematic Cluster // IEEE 2019 International Conference on Fuzzy Systems Proceedings, 2019. (CORE level A)

2) *Frolov D., Mirkin B., Nascimento S., Fenner T.* Method for Generalization of Fuzzy Sets // Rutkowski L., Scherer R., Korytkowski M., Pedrycz W., Tadeusiewicz R., Zurada J. Artificial Intelligence and Soft Computing. ICAISC 2019. Lecture Notes in Computer Science, vol 11508. Springer. C. 273-286. (Web of Science Q4, Scopus Q2)

Публикации стандартного уровня:

3) *Frolov D., Mirkin B., Nascimento S., Fenner T.* Globally Optimal Parsimoniously Lifting a Fuzzy Query Set Over a Taxonomy Tree // Le Thi H., Le H., Pham Dinh

T. Optimization of Complex Systems: Theory, Models, Algorithms and Applications (WCGO). 2019. Advances in Intelligent Systems and Computing, vol 991. Springer, Cham. C. 779-789. (Scopus Q3)

4) *Frolov D., Taran Z., Mirkin B.* A Method for Audience Extending in Programmatic Advertising by Using Parsimonious Generalization of User Segments // Human Interaction and Emerging Technologies (IHIET) 2019. (Scopus Q3)

5) *Frolov D.* Using Annotated Suffix Trees for Fuzzy Full Text Search // Communications in Computer and Information Science. Information Retrieval. 10th Russian Summer School, RuSSIR 2016. Revised Selected Papers. Springer. (Scopus Q3)

6) *Frolov D., Mirkin B., Nascimento S., Fenner T.* Using Domain Taxonomy to Model Generalization of Thematic Fuzzy Clusters // IARIA Content 2019 Proceedings, C. 20-25. (Web of Science)

7) *Frolov D.S.* Annotated suffix tree as a way of text representation for information retrieval in text collections // Business Informatics. 2015. No. 4 (34). P. 63–70. (*Фролов Д.С.* Применение метода аннотированного суффиксного дерева в задачах поиска в коллекциях текстовых документов // Бизнес-Информатика. 2015. №4 (34). С. 63–70.) (Web of Science)

Другие публикации:

8) *Frolov D., Mirkin B., Nascimento S., Fenner T.* Finding an appropriate generalization for a fuzzy thematic set in taxonomy // Working paper WP7/2018/04, Moscow, Higher School of Economics Publ. House, 2018, 60 C.

Объем работы

Диссертация состоит из введения, четырех глав, заключения и четырех приложений. Полный объём диссертации составляет 189 страниц, включая 26 иллюстраций и 30 таблиц. Список использованных источников содержит 163 наименования.

Содержание диссертации

Во **введении** описывается актуальность темы диссертации, формулируется цель и задачи исследования, объект и предмет исследования, описываются методы

исследования, излагаются основные научные результаты, раскрывается теоретическая и практическая значимость работы. Приводятся результаты диссертации, ее апробации и список публикаций автора по теме исследования. Описывается структура диссертации.

В **первой главе** описывается задача информационного поиска, необходимые определения и его основные современные способы.

Цель задач поиска информации состоит в том, чтобы предоставить пользователям множество документов, которые удовлетворят их информационные потребности. При этом такие потребности должны быть сформулированы в форме, «понятной» для механизма поиска. С другой стороны, все множество данных, по которому производится поиск, также должно быть представлено в формате, который позволит механизму поиска быстро идентифицировать потенциально релевантные документы. Очевидно, что в обоих случаях часть информации может быть потеряна в процессе преобразований. Поэтому для разработки моделей поиска важны как организация алгоритмов поискового механизма, так и задача представления информации.

Приводится описание нескольких моделей поиска (Булевой, статистических и других), приводится анализ их главных преимуществ и недостатков. Далее описываются способы представления данных для задач информационного поиска в коллекциях текстовых документов. В настоящее время единой общепринятой классификации подходов к представлению текстов не существует, но можно выделить следующие основные группы [3, 4]: символьные модели, попарное наложение (выравнивание) текстов; различные типы языковых моделей (формирование профилей и скрытых марковских моделей и др.); признаковые описания и модели на их основе; представления фрагментами; векторные представления (эмбеддинги).

В работе описываются перечисленные выше методы, приводятся их достоинства и недостатки. Акцент сделан на агрегированное представление текстов с помощью аннотированных суффиксных деревьев (АСД). Преимущество этого подхода состоит в том, что его можно применять без какой-либо предварительной предобработки данных текстов, в отличие от признакового подхода, который в принципе невозможен без предобработки.

Аннотированное суффиксное дерево (АСД) — это структура данных, используемая для вычисления и хранения всех фрагментов текста совместно с их частотами. Она задается как корневое дерево, в котором каждый узел соответствует одному символу и помечен частотой того фрагмента текста, который кодирует путь от корня до данного узла.

При использовании АСД весь текст разбивается на строки – цепочки символов. Как правило, одна строка формируется из 2-4 подряд идущих слов. Также АСД, построенное для текста, позволяет решать такую задачу, как вычисление релевантности строки тексту, что будет использовано в работе в дальнейшем для решения задачи ранжирования.

Отдельным важным разделом информационного поиска является нечеткий поиск (approximate search, fuzzy search). Это специальный вид поиска, допускающий приблизительное совпадение строк запроса и текстовых документов [1]. Практическую значимость подобных задач трудно переоценить: документам на естественном языке, как и поисковым запросам, свойственно иметь различные виды неточностей – такие как пропуски букв, ошибки, опечатки. Механически существует множество причин, по которым могут образовываться неточности в поисковом запросе и документах коллекции, при которых написание слов отличается от правильного, словарного: от неправильного написания человеком до результатов выполнения каких-либо преобразований текстов. В наши дни в силу большого распространения технологий распознавания текстов (например, с фотографий или сканированных изображений) актуальной является и обработка текстов, содержащих неточности вследствие подобного распознавания – так называемые ошибки OCR (Optical Character Recognition).

В последние годы стремительную популярность приобретает разведочный информационный поиск (exploratory search) [18, 26]. В данное понятие вкладываются такие задачи специализированного поиска, когда пользователь, имеющий информационную потребность:

- Не знаком с предметной областью своей цели (говоря иными словами, нуждается в изучении этой области, чтобы понять, как достигнуть своей цели);
- Не уверен в способах достижения своей цели (в технологии поиска);
- Не может четко сформулировать информационную потребность в понятиях доступной технологии поиска.

Безусловно, задача в такой постановке сильно отличается от классической, когда поиск производится по конкретному поисковому запросу. Разведочный поиск охватывает более широкий класс действий, чем типичный информационный поиск, и включает структурирование данных, анализ, интерпретацию, сравнение результатов. Поэтому для выполнения этого вида поиска часто применяются комбинированные стратегии.

Крайне важными являются задачи выбора методов разведочного поиска – для его ускорения и получения корректных результатов.

Критически важной задачей разведочного поиска является интерпретация результатов. Многие методы, в частности, структурирования коллекций, дают результаты, которые иногда крайне сложно интерпретируются. Классическим примером является интерпретация найденных тем в вероятностном тематическом моделировании, применяемая, например, для анализа скрытых направлений исследований в предметной области. Вариантом решения данной проблемы видится применение подходов, основанных на представлении предметных областей в виде таксономий. Разработке таких подходов посвящена одна из следующих глав диссертации.

Важной подзадачей разведочного поиска является структуризация коллекций: при разведочном поиске важно иметь методы изучения содержимого документов коллекции и изучения ее структуры (например, семантической). Основной группой методов для этого является кластеризация [19, 21]. Кластеризация как задача группировки исходного множества на подмножества-кластеры схожих между собой объектов способна показать внутреннюю структуру коллекции, а изучение отдельных представителей кластеров способно выявить их типичные признаки. Важно отметить, что существующие методы описания кластеров основаны на использовании характеристик того же уровня гранулярности (признаки, слова), которые были использованы для формирования кластеров.

Отдельный ряд методов разведочного поиска предлагает использование таксономий предметной области. Таксономия, как структура понятий и как способ представления знаний, может быть применена для структурирования и интерпретации результатов поиска или отдельных документов, и также допускает понятие обобщения, то есть перехода на крупный уровень гранулярности.

Во **второй главе** приводится описание разработанного нами метода информационного поиска АСДП, использующего представление текстов в виде АСД и обратное фрагментное индексирование [6].

Излагаются результаты экспериментальной апробации данного метода на реальных данных, а также его сравнение с другими методами информационного поиска: как по качественным метрикам, так и по производительности. Описана экспериментальная валидация предложенного метода на реальных данных. Представлено несколько серий экспериментов, в которых разработанный метод сравнивался как на коллекциях

из реальных данных с применением пользовательских запросов, так и на специализированном наборе данных для информационного поиска. Сравнивалось как качество поиска, так и производительность.

В сравнении участвовали метод поиска АСДП (с длиной строки для АСД 3 идущих подряд слова); метод поиска на основе ранжирования BM25 [3]; два метода поиска на основе косинусной близости векторов: слов и 3-х символьных фрагментов, в обоих случаях взвешенных с помощью TF-IDF [3]; а также методы на основе PLSI и LDA (в реализации gensim, для последнего также была реализована модификация, использующая биграммы). Поскольку тестовые коллекции документов содержались в документо-ориентированной базе данных MongoDB, сравнение был включен встроенный механизм полнотекстового поиска этой базы данных. В эксперименте участвовали три коллекции: №1 – коллекция документов на основе каталога интернет-магазина Ozon.ru; №2 – статьи из веб-страниц сайта Habrahabr.ru; №3 – коллекция из специализированного набора данных TREC CAR.

Приведем результаты экспериментального сравнения точности на уровне $N = 10$ документов [3] на коллекции №1 для трех групп запросов (1 – «Название подкатегории» каталога Озон, 2 – «Явные запросы» и 3 – «Неточные запросы»).

Таблица 1. Сравнительная точность рассматриваемых методов поиска на уровне $N = 10$ документов, коллекция №1.

Номер группы запросов	АСДП	BM25	cos + TF-IDF слова	cos + TF-IDF фрагм.	PLSI	LDA	LDA с бигр.	Полнот. поиск MongoDB
1	0.83	0.86	0.79	0.76	0.70	0.85	0.86	0.51
2	0.82	0.84	0.72	0.70	0.68	0.81	0.86	0.52
3	0.77	0.43	0.44	0.65	0.41	0.43	0.55	0.21
Среднее	**0.81**	0.72	0.65	0.70	0.56	0.70	0.76	0.40

Отметим, что полученные оценки точности для метода АСДП, для 3-й группы запросов («Неточные запросы») почти такие же, как и для первых двух, в то время как другие методы, кроме подхода на основе косинусной близости фрагментов, взвешенных по TF-IDF, серьезно проигрывают. Ясно, что это объясняется тем, что в методе АСДП используется фрагментный, а не признаковый подход для представления текста. По результатам сравнения качества и производительности метод АСДП показал один из лучших результатов среди рассматриваемых методов.

Поскольку одним из главных преимуществ АСДП является возможность производить нечеткий поиск, было произведено его экспериментальное сравнение со специальными методами нечеткого поиска. Было выполнено сравнение метода АСДП с популярным алгоритмом, базирующимся на вычислении расстояния Левенштейна (LD-based) с n-граммной обратной индексацией [1], а также с более современными методами хеширования по сигнатурам [1] (SH-based) и поисковой системой Lemur.

Вычислительные эксперименты были проведены с использованием реальных данных. Для того, чтобы провести именно нечеткий поиск, были подготовлены две коллекции и специальные поисковые запросы. Коллекция №1 – известная стандартная коллекция «Reuters-21578», использующаяся для многих задач, в том числе и классического информационного поиска, №2 – коллекция патентов базы USPTO, обе – на английском языке. Все поисковые запросы были получены из строк документов коллекции и последовательностей их слов, заменой оригинальных слов на варианты, заведомо содержащие ошибки.

Итоговые результаты сравнения рассматриваемых четырех методов по материалам проведенных экспериментов приведены в таблице 2.

Таблица 2. Результаты сравнения методов нечеткого поиска (место 1-2 приравнено рангу 1.5).

	Метод АСДП	LD-based	SH-based	Lemur
Качество поиска, место	1-2	1-2	3	4
Производительность, место	3	4	2	1
Суммарный ранг	**4.5**	5.5	5	5

Как видно из таблицы 2, метод АСДП демонстрирует наилучший баланс качества и производительности среди рассматриваемых методов.

По итогу проведенных сравнительных экспериментов, были зафиксированы основные преимущества метода поиска АСДП:

- Позволяет избежать предобработки текстов (лемматизация, стеммирование);
- Допускает тексты с искаженными признаками (ошибки, опечатки в словах);
- Демонстрирует хороший баланс качества и производительности поиска.

Также в главе описана выполненная программная реализация метода АСДП в специальной поисковой системе для операционных систем семейства Ubuntu с консольным интерфейсом.

В **третьей главе** описывается алгоритм оптимального обобщения нечеткого множества (ПарГеНМ) и его применение в задаче разведочного поиска.

Как уже было отмечено ранее, разведочный поиск представляет собой важную и очень объемную проблему анализа текстовых коллекций. Представлен новый подход к задаче, использующий таксономию предметной области, на примере рассмотрения методики обобщения для интерпретации результатов поиска, а затем и демонстрации общей стратегии анализа коллекции документов.

В диссертации рассмотрено моделирование направлений новых исследований в области науки о данных (Data Science) на примере коллекции научных публикаций издательства Springer за 20 лет и таксономии DST [15], разработанной на основе таксономии компьютерных наук Всемирной ассоциацией вычислительной техники (Association for Computing Machinery – ACM [28]).

На примере этой коллекции и таксономии в работе найдены и проанализированы нечеткие кластеры тем таксономии, каждый из которых представляет тенденцию развития, отраженную в коллекции. Основной проблемой является интерпретация такого нечеткого кластера. Как уже указывалось, обычно интерпретация кластеров осуществляется с использованием понятий того же уровня гранулярности (признаки, ключевые слова), которые используются при построении кластеров. В работе же предложено использовать понятие укрупненной гранулярности, для чего воспользоваться таксономией предметной области. Конкретно, речь идет об обобщении понятий таксономии, соответствующих полученному кластеру. Приведем здесь простую иллюстрацию этого подхода на следующем примере (см. рисунок 1).

Рис. 1. Фрагмент таксономии, черные прямоугольники соответствуют элементам, принадлежащим множеству: простой случай (a) и более сложный (b).

На рисунке 1 (a) представлен фрагмент таксономии с листовым кластером,

полностью охватывающим все дочерние элементы родительского узла A. Очевидно, «поднятие» этого кластера в таксономии в вершину A является наиболее естественным обобщением кластера: оно предлагает интерпретацию кластера как всех тем, попадающих в концепцию A. В работе расширен этот подход, чтобы обобщить менее очевидные случаи – такие, как представленный на рисунке 1 (b).

В работе приведена математическая формализация проблемы обобщения как оптимального поднятия нечеткого листового кластера, определяющего множество, на более высокие ранги таксономии и предоставлен рекурсивный алгоритм, приводящий к глобально оптимальному решению задачи.

Для демонстрации задачи рассмотрим четкое множество S, показанное пятью черными прямоугольниками в листьях таксономии на фрагменте дерева на рисунке 1 (b). Если допустить, что множество S может быть обобщено корнем фрагмента, это приведет к тому, что корнем будет покрыто и четыре белых прямоугольника, и они попадут в то же обобщение, что и S, при том, что они не принадлежат множеству S. Такую ситуацию будем далее называть появлением пробелов, здесь их – четыре, а также введена одна головная вершина. Другой возможный вариант подъема – подъем до корня левого поддерева. Видно, что число пробелов резко сократилось – до 1. Однако возник еще один тип неточности подъема: черный прямоугольник справа, принадлежащий S, не покрыт корнем левого поддерева, в которую отображается множество S. Этот тип ошибки будем называть выбросом. При этом варианте подъема появляются три новых объекта: одна головная вершина, один выброс и один пробел. Чтобы определить, какой из вариантов предпочтительнее, мы вводим штрафные веса за появление новых объектов: 1 – штраф за головной узел, λ – штраф за пробел, γ – штраф за выброс. Тогда, например, штраф за второй вариант подъема составит: $1 + \gamma + \lambda$. На основе введенных выше понятий в данной главе определена штрафная функция $p(H)$ для множества головных тем H:

$$p(H) = \sum_{h \in H - I} u(h) + \sum_{h \in H - I} \sum_{g \in G(h)} \lambda v(g) + \sum_{h \in H \cap I} \gamma u(h). \quad (1)$$

Здесь I – множество листовых вершин дерева, $v(g)$ – важность пробела (логично, что пробелы могут быть неодинаково значимы), $G(t)$ – множество пробелов узла t, $V(t) = \sum_{g \in G(t)} v(g)$ – суммарная важность пробелов.

В данной главе предложен алгоритм ПарГеНМ, позволяющий найти оптимальное обобщение нечеткого тематического множества за счет минимизации данной функции штрафа. Перед применением алгоритма необходимо выполнить предварительное

преобразование дерева; для этого нужно сделать следующие шаги.

1) Из дерева нужно удалить все потомки пробелов.

2) Аннотировать все внутренние узлы дерева значениями функции принадлежности, основываясь на значении принадлежности для листьев, например, по следующему правилу: $u(h) = \sqrt{\sum_{t \in \chi(h)} u(t)^2}$ для $h \in T - I$, где $\chi(h)$ – множество потомков узла h.

3) Множества пробелов $G(t)$ и значения суммарной важности пробелов $V(t) = \sum_{g \in G(t)} v(g)$ для последующего использования в (1) необходимо вычислить для каждого внутреннего узла t.

Для каждого узла дерева таксономии t алгоритм ПарГеНМ вычисляет два множества – $H(t)$ и $L(t)$, содержащие такие узлы в $T(t)$, в которых, соответственно, происходят приобретения и потери головных узлов (включая выбросы). Соответствующий штраф обозначается $p(t)$. Выходные данные алгоритма включают значения рассчитываемых множеств в корне, а именно: H – множество головных узлов и выбросов, L - набор пробелов и p – назначенное значение штрафа. Псевдокод алгоритма приведен ниже.

Алгоритм ПарГеНМ

- **INPUT:** u, T
- **OUTPUT:** $H = H(root)$, $L = L(root)$, $p = p(root)$

I **Базовый случай**

Для каждого листа $i \in I$

Если $u(i) > 0$

$H(i) = \{i\}$, $L(i) = \oslash$, $p(i) = \gamma u(i)$

Иначе

$H(i) = \oslash$, $L(i) = \oslash$, $p(i) = 0$

II **Рекурсия**

Если $u(t) + \lambda V(t) \leq \sum_{w \in \chi(t)} p(w)$

$H(t) = \{t\}$, $L(t) = G(t)$, $p(t) = u(t) + \lambda V(t)$

Иначе

$$H(t) = \bigcup_{w \in \chi(t)} H(w),\ L(t) = \bigcup_{w \in \chi(t)} L(w),\ p(t) = \sum_{w \in \chi(t)} p(w)$$

Нетрудно видеть, что алгоритм ПарГеНМ действительно приводит к оптимальному поднятию, как указано в следующей теореме.

Теорема 1. *Любое и-покрытие H, найденное алгоритмом ПарГеНМ, доставляет глобальный минимум функции штрафа $p(H)$ (1).*

После рассмотрения иллюстративных примеров в диссертации продемонстрировано применение ПарГеНМ к задаче разведочного поиска в описанной выше коллекции научных публикаций. Для этого последовательно применены шаги, перечисленные ниже.

- Подготовка коллекции научных статей;
- Подготовка таксономии рассматриваемой предметной области;
- Расчет матрицы значений релевантности между тематическими листьями таксономии и публикациями из коллекции статей;
- Поиск нечетких кластеров по матрице значений релевантности;
- Подъем полученных кластеров в таксономии для их концептуализации с помощью обобщения;
- Построение выводов из обобщений.

После выполнения первых четырех шагов было получено шесть кластеров, три из которых оказались особенно однородными. Основываясь на их содержимом, “Learning”, “Retrieval” и “Clustering”, обозначим их L, R и C соответственно. Пример кластера (“Learning”) приведен в таблице 3.

Полученные кластеры были подняты в таксономии DST с использованием алгоритма ПарГеНМ, со штрафом за пробел $\lambda = 0.1$ и штрафом за выброс $\gamma = 0.9$. Результаты подъема кластера L показаны на рисунке 2. Кластер получил три головных темы: machine learning, machine learning theory, and learning to rank. Они представляют структуру общего понимания “Learning” в соответствии с нашей текстовой коллекцией.

Тематические кластеры, найденные в рассматриваемой коллекции научных публикаций, составляют области будущих разработок. В частности, из головных тем

Таблица 3. Кластер L “Learning”: темы с наибольшими значениями принадлежности кластеру.

$u(t)$	Code	Topic
0.300	5.2.3.8.	rule learning
0.282	5.2.2.1.	batch learning
0.276	5.2.1.1.2.	learning to rank
0.217	1.1.1.11.	query learning
0.216	5.2.1.3.3.	apprenticeship learning
0.213	1.1.1.10.	models of learning
0.203	5.2.1.3.5.	adversarial learning
0.202	1.1.1.14.	active learning

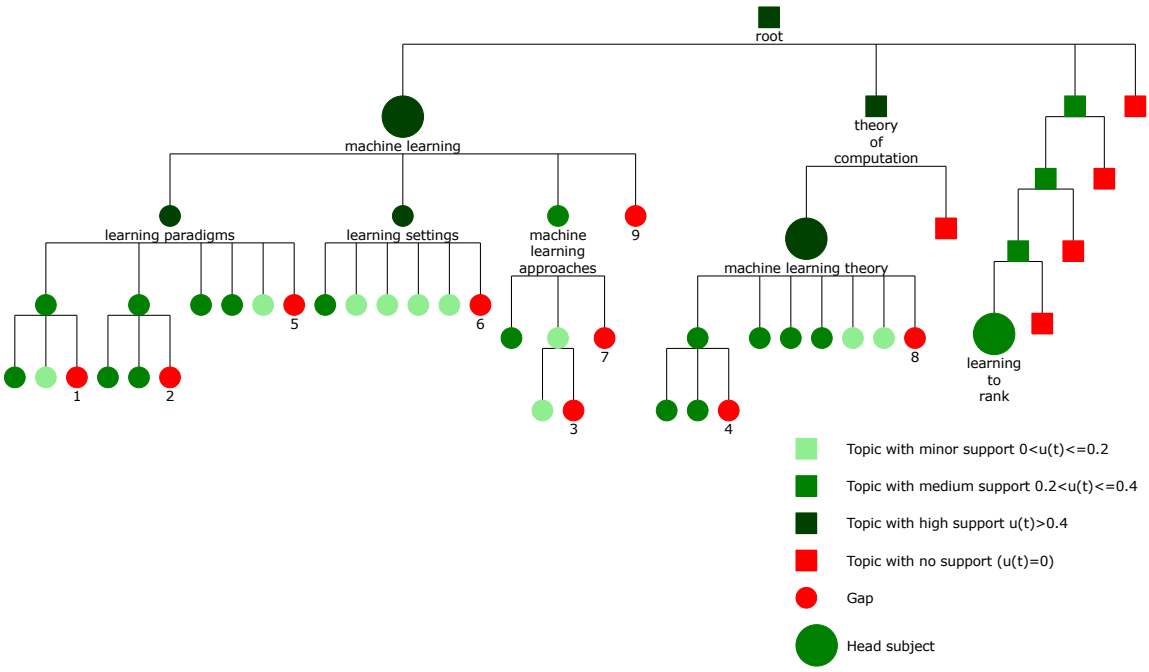

Рис. 2. Результаты подъема кластера L: Learning. Пробелы пронумерованы.

кластера “Learning” хорошо заметно (см. рисунок 2 и комментарии к нему), что основная деятельность здесь все еще сосредоточена на теории и методах, а не на приложениях. Области машинного обучения, ранее сфокусированные в основном на задачах обучения на подмножествах и разбиениях, в настоящее время расширяются в направлении обучения ранжированию и рейтингованию. Приведенное в главе сравнение полученных результатов с популярными методами анализа коллекций – LDA и иерархической кластеризацией UGMA показало значительное преимущество разработанной методики.

Четвертая глава посвящена применению алгоритма ПарГеНМ для задачи расширения аудитории в рекламном таргетинге (Programmatic). В ней рассмотрен

специальный прикладной случай задачи информационного поиска: поиск релевантной аудитории среди множества пользователей для рекламной кампании в сети Интернет. В данном случае в качестве поискового запроса выступают требования к пользователям, которым должна показываться реклама, а документами – сведения о пользователях, накапливаемые рекламной системой.

В последнее время набирает популярность так называемый программатический (Programmatic) подход к интернет-рекламе [25], позволяющий в режиме реального времени показывать пользователям те рекламные креативы, которые соответствуют его интересам, выявляемым по накопленным в рекламных системах сведениям о пользователе. При этом подходе рекламодатель имеет возможность закупать только релевантную аудиторию.

Одним из критериев таргетинга является соответствие интересов пользователя сегментам, установленным для рекламной кампании. Каждому пользователю, о котором рекламная система имеет данные (например, его историю браузинга), сопоставляются несколько сегментов интересов. Сегменты являются элементами какой-либо таксономии. В настоящее время существуют несколько таксономических структуризаций пользовательских интересов. В работе используется одна из наиболее популярных таксономий – таксономию IAB, разработанную Международным Бюро Интернет-Рекламы [29].

Для повышения эффективности рекламных кампаний часто необходимо решать задачу расширения целевой аудитории: не всегда количества пользователей, прямо подходящих под таргетинг, достаточно. Одним из способов решения является техника look-a-like [24], при использовании которой аудитория расширяется за счет подбора пользователей, похожих в какой-то заданной метрике на пользователей из ядра (то есть тех пользователей, которые успешно проходят таргетинг). Другим вариантом является ослабление границ попадания пользователя в сегменты интересов. Также остается вариант закупки дополнительного интернет-трафика, что влечет дополнительные финансовые затраты для рекламной сети.

В данной главе предложен другой вариант расширения целевой аудитории – путем концептуального обобщения пользовательских сегментов. Для этого используется алгоритм ПарГеНМ, «поднимающий» пользовательские сегменты в более высокие ярусы таксономии, за счет чего пользователь приобретает головные сегменты.

Разработанный алгоритм, названный обобщением пользовательских сегментов (ОПС), работает с таксономией пользовательских интересов, в роли которой,

в частности, может выступать уже упомянутая таксономия IAB (либо любая другая индустриальная таксономия). Эта таксономия охватывает традиционные пользовательские интересы, представляя их в виде четырехуровневого корневого дерева, узлы которого помечены темами таксономии.

Благодаря обобщению происходит «интеллектуальное» расширение пользовательской аудитории: алгоритм ПарГеНМ определяет, когда обобщение возможно, исходя из значений принадлежности пользователя к сегментам. В свою очередь, таргетинг производится уже с участием «расширенных» сегментов.

Новый метод был протестирован в трех реальных рекламных кампаниях в компании ООО «Натиматика» (`https://natimatica.com/`). В сравнении участвовали три метода рекламного таргетинга: классический способ programmatic-таргетинга по сегментам, метод, расширяющий целевую аудиторию с использованием алгоритма обобщения пользовательских сегментов (ОПС), метод, расширяющий целевую аудиторию с понижением порогов принадлежности пользователей к сегментам (ПППС).

Основываясь на результатах экспериментов, можно сделать вывод, что использование алгоритма обобщения пользовательских сегментов для расширения аудитории рекламных кампаний позволяет делать существенно большие объемы показов без значительного падения качества аудитории. При этом алгоритм, основанный на понижении порогов принадлежности пользователей сегментам, демонстрирует расширение аудитории с заметным снижением ее качества, что было видно по количеству кликов по рекламным объявлениям и CTR (отношение кликов к показам, click through rate). Отметим, что в этой прикладной задаче эффект обобщения оказался измеряемым.

В качестве меры эффекта естественно взять отношение числа кликов по методу ОПС к числу кликов при классическом таргетинге (см. таблицу 4).

Таблица 4. Эффективность метода ОПС.

Рекламная кампания	Классический таргетинг, число кликов	ОПС-таргетинг, число кликов	Эффективность метода ОПС, %
1	1061	2544	239.8
2	201	367	182.6
3	749	1302	173.8

Конечно, эффект от использования метода оптимального обобщения в общей задаче разведочного поиска можно будет измерить только тогда, когда удастся формализовать процесс оценки уровня знаний.

В **заключении** перечислены основные результаты, достигнутые в ходе работы над диссертацией и выносимые на защиту.

1) Разработан новый метод информационного поиска АСДП. Выполнена программная реализация поисковой системы, основанной на данном методе. Эффективность метода доказана вычислительными экспериментами по сравнению АСДП с популярными современными способами поиска, в том числе и специализированными для задач нечеткого поиска. Сравнение показало качественное преимущество разработанного метода в задачах нечеткого поиска и хороший баланс его качественных характеристик и производительности.

2) Разработан метод интерпретации результатов поиска с помощью оптимального обобщения нечетких кластеров в таксономии предметной области (ПарГеНМ). Сделана программная реализация разработанного метода. Экспериментальная апробация метода ПарГеНМ проведена на коллекции публикаций издательства Springer в области науки о данных.

3) Разработан метод расширения аудитории интернет-рекламы как задачи информационного поиска на основе оптимального обобщения сегментов пользователя в таксономии сегментов пользовательских интересов (ОПС). Данное приложение может рассматриваться в качестве примера задачи, в которой эффект оптимального обобщения измерим. Новый метод расширения аудитории рекламных кампаний успешно внедрен на практике.

В **приложениях** приведены: рассмотренная в работе таксономия науки о данных (целиком), некоторые кластеры, полученные в ходе сравнения разработанного нами метода с конкурентами, листинг программного кода реализации АСД с возможностью вычисления релевантности строки дереву, листинг программной реализации алгоритма ПарГеНМ, листинг программного кода отображения дерева таксономии с результатами ПарГеНМ.

Библиографический список использованной литературы

[1] *Бойцов Л. М.* Классификация и экспериментальное исследование современных алгоритмов нечеткого словарного поиска // Труды Yandex. – 2004. – Т. 6.

[2] *Коршунов А, Гомзин А.* Тематическое моделирование текстов на естественном языке // Труды ИСП РАН. 2012. №1. С.215-244.

[3] *Маннинг К. Д., Рагхаван П., Шютце Х.* Введение в информационный поиск / Маннинг К. Д., Рагхаван П., Шютце Х. – М.: Вильямс, 2011. – 680 с.

[4] *Миркин Б. Г.* Методы кластер-анализа для поддержки принятия решений: обзор / Б. Г. Миркин. – М.: Издательский дом Национального исследовательского университета «Высшая школа экономики», 2011 – 84 с.

[5] *Миркин Б. Г., Черняк Е. Л., Чугунова О. Н.* Метод аннотированного суффиксного дерева для оценки степени вхождения строк в текстовые документы // Бизнес-информатика. 2012. Т. 3. № 21. С. 31-41.

[6] *Фролов Д.С.* Применение метода аннотированного суффиксного дерева для задач поиска в коллекциях текстовых документов // Бизнес-Информатика. 2015. №. 4 (34). С. 63–70.

[7] *Черняк Е. Л., Миркин Б. Г.* Использование ресурсов Интернета для построения таксономии // В кн.: Доклады всероссийской научной конференции АИСТ 2013 / Отв. ред.: Е. Л. Черняк; науч. ред.: Д. И. Игнатов, М. Ю. Хачай, О. Баринова. М. : Национальный открытый университет «ИНТУИТ», 2013. С. 36-48.

[8] *Altevogt P., Nitzsche R.* Method of generating a distributed text index for parallel query processing: пат. 7966332 США. – 2011.

[9] *Ashraf J., Chang E., Hussain O. K., Hussain F. K.* Ontology usage analysis in the ontology lifecycle: A state-of-the-art review // Knowledge-Based Systems, vol. 80, pp. 34-47, 2015.

[10] *Beneventano D., Dahlem N., El Haoum S., Hahn A., Montanari D., Reinelt M.* Ontology-driven semantic mapping // Enterprise Interoperability III, Part IV, Springer, C. 329-341, 2008.

[11] *Blei D.* Probabilistic topic models // Communications of the ACM, 55 (4), C. 77–84, 2012.

[12] *Chernyak E., Mirkin B.* Refining a Taxonomy by Using Annotated Suffix Trees and Wikipedia Resources // Annals of Data Science, 2(1), C. 61-82, 2015.

[13] *Frolov D.S.* Annotated suffix tree as a way of text representation for information retrieval in text collections //Business Informatics. 2015. No. 4 (34). C. 63–70. DOI: 10.17323/1998-0663.2015.4.63.70.

[14] *Frolov D.* Using Annotated Suffix Trees for Fuzzy Full Text Search, in: Communications in Computer and Information Science. Information Retrieval. 10th Russian Summer School, RuSSIR 2016, Saratov, Russia, August 22-26, 2016, Revised Selected Papers. Springer, 2016

[15] *Frolov D., Mirkin B., Nascimento S., Fenner T.* Finding an appropriate generalization for a fuzzy thematic set in taxonomy / Working paper WP7/2018/04, Moscow, Higher School of Economics Publ. House, 2018, 58 p.

[16] *Langville A. N., Meyer C. D.* Google PageRank and beyond: The science of search engine rankings // Princeton University Press, 2011.

[17] *Lloret E., Boldrini E., Vodolazova T., MartÃnez-Barco P., Munoz R., Palomar M.* A novel concept-level approach for ultra-concise opinion summarization // Expert Systems with Applications, 42(20), pp. 7148-7156, 2015.

[18] *Marchionini G.* Exploratory Search: from finding to understanding. Communications of the ACM. 2006, 49(4), p. 41-46.

[19] *Mirkin B. G.* Core Concepts of Data Analysis / B. G. Mirkin. – Springer, 2012. – 416 C.

[20] *B. Mirkin, S. Nascimento, T. Fenner, L.M. Pereira* Building fuzzy thematic clusters and mapping them to higher ranks in a taxonomy, Int. Journal of Software Informatics, vol. 4, no. 3, C. 257-275, 2010.

[21] *Mirkin B., Nascimento S.* Additive spectral method for fuzzy cluster analysis of similarity data including community structure and affinity matrices, Information Sciences, vol. 183, no. 1, C. 16-34, 2012.

[22] *Mueller G., Bergmann R.* Generalization of Workflows in Process-Oriented Case-Based Reasoning // FLAIRS Conference, C. 391-396, 2015.

[23] *Nascimento S., Fenner T., Mirkin B.* Representing research activities in a hierarchical ontology // Procs. of 3rd International Workshop on Combinations of Intelligent Methods and Applications (CIMA 2012), Montpellier, France, August, 28, C. 23-29, 2012.

[24] *Popov A., Iakovleva D.* Adaptive look-alike targeting in social networks advertising // Procedia Computer Science. – 2018. – T. 136. – C. 255-264.

[25] *Sayedi A.* Real-Time Bidding in Online Display Advertising, 2017.

[26] *White R., Roth R.* Exploratory Search: beyond the Query-Response paradigm. San Rafael, CA: Morgan and Claypool, 2009.

[27] *Yuan, Y., Wang, F., Li, J., Qin, R.* A survey on real time bidding advertising. In Service Operations and Logistics, and Informatics (SOLI) // 2014 IEEE International Conference. IEEE. C. 418-423.

[28] The 2012 ACM Computing Classification System [Electronic resource]. 2019 –. – Режим доступа: *http://www.acm.org* /about/class/2012 , свободный. – Загл. с экрана.

[29] IAB Tech Lab Content Taxonomy [Electronic resource]. 2019 –. – Режим доступа: *https://www.iab.com/guidelines/iab-quality-assurance-guidelines-qag- taxonomy/*, свободный. – Загл. с экрана.

[30] OpenRTB Protocol [Electronic resource]. 2019 –. – Режим доступа: *https://www.iab.com/guidelines/real-time-bidding-rtb-project/* , свободный. – Загл. с экрана.