МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра математической кибернетики и компьютерных наук

РАСШИРЕНИЕ ЗАПРОСА С ПОМОЩЬЮ КВАЗИ-RDF-ТРИПЛЕТОВ ПРЕДМЕТНОЙ ОБЛАСТИ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 273 группы
направления 02.04.03 — Математическое обеспечение и администрирование
информационных систем
факультета КНиИТ
Ланкина Ивана Сергеевича

Научный руководитель
доцент, к. ф.-м. н. _____ С. В. Папшев

Заведующий кафедрой
к. ф.-м. н., доцент _____ С. В. Миронов

Саратов 2025

ВВЕДЕНИЕ

Актуальность темы. Актуальной проблемой современного состояния глобального информационного пространства является разработка эффективных методов информационного поиска. Это становится все более распространенной проблемой: просеивание электронной почты организации, изучение газет за десятилетие, характеристика исследований в научной области или обработка текстов вузовских новостных лент.

В огромном объеме информации пользователю нужно искать необходимые данные по запросу. Особенно это актуально для неструктурированных документов, где повышение эффективности поиска релевантной информации требует дополнительного внимания. Одно из направлений исследований связано с методом интерактивного поиска — полученная в ходе исследования информация уточняется и передается повторно на следующую итерацию. Одним из подходов к повышению эффективности интерактивного поиска является метод расширения запроса. В данной связи исследователями предложено два направления развития методов расширения запроса: на основе тезауруса и на основе онтологии.

Тезаурус обычно касается всего языка и поэтому в рамках узкой предметной области он не очень хорошо работает.

Автоматизация онтологического моделирования предметной области представляет собой эффективный способ формирования структурированных знаний, что полезно для таких областей, как информационный поиск, аналитика данных и разработка интеллектуальных систем. К сожалению, разработка онтологии — весьма трудоемкий процесс, требующий привлечения экспертов соотвествующей предметной области. В этой связи разрабатываются методы автоматизации семантического моделирования на основе анализа текстовых данных предметных областей. Для этого широко применяются различные методы лингвистического и статистического анализа текстов, а также методы машинного обучения. Из-за трудоемкости общей задачи автоматического построения онтологии большинство исследователей сосредоточены на конкретных аспектах проблемы или языковой специфике задачи.

Многочисленные исследования в этой области до сих пор в удовлетворительной степени не дают решения ключевых проблем автоматического семантического описания предметной области. Основной проблемой в данной обла-

сти является морфологический анализ и анализ зависимостей в предложении, поэтому проблема автоматизации семантического моделирования предметной области до сих пор остается актуальной.

Многими исследователями признается, что проблема семантического описания предметного поля для корпуса документов во многом зависит от извлечения из текстов так называемых триплетов: субъект-глагол-предмет (Subject-Verb-Object), которые являются основой RDF (Resource Description Framework) описания семантического веба. Данные триплеты очень схожи с теми, которые извлекаются из онтологии. Таким образом на этом пути можно миновать трудоемкую стадию построения онтологии. На основе полученных триплетов можно строить расширенный запрос.

Цель магистерской работы — разработка метода расширения запроса с помощью квази-RDF-триплетов предметной области.

Поставленная цель определила **следующие задачи**:

- подготовка данных предметной области;
- извлечение сущностей предметной области;
- генерация квази-RDF-триплетов;
- реализация модели расширения запроса;
- проведение экспериментов и статистического анализа для проверки качества работы реализованной модели.

Структура и объем работы. Магистерская работа состоит из введения, двух разделов, заключения, списка использованных источников, 2 приложений и цифрового носителя. Общий объем работы — 69 страниц, из них 64 страницы — основное содержание, включая 30 рисунков и 9 таблиц, список использованных источников информации содержит 36 наименований.

1 Методы и алгоритмы препроцессинга текстовых файлов

1.1 Расширение запроса и извлечение информации

Современные информационно-поисковые системы, особенно поисковые системы, должны решать сложную задачу удовлетворения потребностей пользователя, выраженных короткими запросами. Наиболее частые запросы состоят всего из одного, двух или трех слов.

Цель расширения запроса состоит в том, чтобы уменьшить несоответствие между документами и запросами путем расширения запроса с использованием слов или фраз со сходным значением или какой-либо другой статистической связи с набором соответствующих документов.

Расширение запроса переформулирует исходный запрос пользователя, добавляя некоторые дополнительные релевантные термины со схожим значением. Выбор этих терминов расширения играет важную роль в расширении запроса, поскольку только небольшое подмножество терминов-кандидатов расширения действительно имеет отношение к запросу пользователя. Современные коммерческие поисковые системы отлично справляются с интерпретацией этих коротких запросов, однако их результаты могут быть дополнительно улучшены за счет использования дополнительных внешних знаний, полученных путем объединения результатов поиска, для расширения первоначальных запросов.

Извлечение информации направлено на организацию, хранение, извлечение и оценку информации из хранилищ документов, в частности текстовой информации. Методы извлечения информации помогают пользователям находить необходимую им информацию, но явно не возвращают ответы на вопросы. Они информируют о существовании и местоположении документов, которые могут содержать требуемую информацию. Документы, удовлетворяющие требованиям пользователя, называются релевантными документами.

1.1.1 Модель разведочного поиска Exploratory Search

Разведочный поиск — специализация извлечения информации, процесс, характеризующийся неточной и открытой целью поиска информации. В отличие от традиционных задач поиска, когда пользователи имеют четкий, конкретный запрос и ищут один четко определенный ответ, разведочный поиск предполагает поиск более широкого понимания темы или проблемы.

Разведочный поиск работает, позволяя пользователям изучать информацию более гибким и открытым способом.

1.2 Методы на основе ключевых слов

Для обработки текста часто используют методы основанные на ключевых словах, относящиеся к сфере Natural Language Processing (NLP) или, если переводить, обработка естественного языка. Обработка естественного языка представляет собой текстовое кодирование в ту или иную форму. Более формально: кодирование текста — это способ преобразования текста в числовую или векторную форму, сохраняющую значение и связь слов и предложений. Это позволяет машине распознавать структуру и контекст любого текста.

Существует большое количество методов по кодированию текста с использованием NLP основанные на ключевых словах, такие как:

- индексное кодирование;
- мешок слов;
- метрика TF-IDF;
- Word2Vector;
- BERT.

1.3 Использование тезаурусов и онтологий для обработки запросов

В настоящее время множество методов описания контента и обработки запросов при поиске информации основаны на ключевых словах и, следовательно, предоставляют ограниченные возможности для получения концептуальных представлений, связанных с потребностями пользователей и содержимым.

Классические методы поиска, основанные на ключевых словах, имеют следующий ряд существенных ограничений:

- зависимость от формулировки запроса;
- обрабатывают тексты по стандартным алгоритмам без учета смысла;
- не имеют семантического словаря.

В соответствии с определениями стандартов информационно-поисковый тезаурус — это нормативный словарь, явно указывающий отношения между терминами и предназначенный для описания содержания документов и поисковых запросов.

Онтологии представляют собой формальные описания понятий и взаимосвязей между ними. Они играют ключевую роль в семантических веб-приложениях, обеспечивая единое представление знаний о реальных объектах. Это способствует повторному использованию данных и совместимости между различными компонентами системы.

1.4 Анализ тематических лент вузов

В данной работе исследуется набор данных содержащий набор файлов для поддержки и иллюстрации последовательных шагов тематического моделирования для текстовых документов новостной ленты и данных для дальнейших исследований. Набор текстов состоит из 31077 записей новостей сайта sgu.ru.

Для текстов новостей данного набора выполнено тематическое моделирование средствами пакетов Gensim и BigARTM. На основе тематической модели сделан семантический анализ модели новостной ленты и ее тематических рубрик.

1.5 Использование тематического моделирования

В соответствии со своим названием, тематическая модель предоставляет «темы», каждая из которых представляет собой ранжирование всех отдельных слов в электронных письмах по релевантности теме. Если взять N наиболее релевантных слов в каждой теме, то получится набор слов, которые вместе имеют смысл. Одна половина тематической модели связывает темы в беспорядочный «мешок слов», другая — связывает темы с отдельными документами.

Тематическое моделирование — способ построения тематической модели, которая в дальнейшем распределяет документы по темам. Методы тематического моделирования разработаны для более сложного, семантического, анализа текста, что вручную делать неэффективно или долго из-за большого объема данных. Такие методы могут быть предназначены как для коротких, так и для длинных текстов. Они применяются и дают хорошие результаты во многих работах и различных областях анализа текста, таких как латентный семантический анализ, латентное размещение Дирихле, неотрицательная матричная факторизация, латентное семантическое индексирование и другие различные методы NLP. Для того, чтобы качественно распределить документы,

для дальнейшего поиска в данной работе проведено тематическое моделирование выбранного набора документов.

1.6 RDF и RDF-триплеты

Resource Description Framework (RDF) — стандарт семантического веба, разработанный консорциумом W3C и использующийся для описания ресурсов в интернете и связей между ними. RDF как модель предлагает описание данных на основе ориентированных графов, которые позволяют представлять и обмениваться данными во всех предметных областях, то есть RDF-модель является предметно-независимой.

Основная идея RDF-модели связан с совместным использованием разнородных данных и объединением всех данных для конкретной предметной области. Цель RDF — позволить обмениваться данными в Интернете, сохраняя при этом их первоначальный смысл.

В данной работе используется построение триплетов на принципе RDF. RDF-триплет — базовая структура данных в RDF, состоящая из 3-х частей: субъект, предикат, объект.

2 Построение расширения с помощью RDF-триплетов

2.1 Проведение тематического моделирования

Предварительно, проводится тематический анализ предметной области на основе которого слова и статьи относятся к тем или иным темам. Данный тематический анализ делается для оценки качества работы разработанного метода, чтобы понимать к какой теме какие найденные расширением статьи относятся.

Как описывается в разделе 1.4, для изучения берется датасет из 31077 новостей сайта sgu.ru. Для тематического моделирования текста выделяется 6 тем из новостного датасета: «Events and Commemorative Days», «Science and Research», «Learning and Educational Services», «Educational and Scientific Activities», «Students Competitions», «Youth Projects and Innovations». Для удобства, в коде они назывались topic0, topic1, ..., topic5.

2.2 Генерация триплетов предметной области

Процесс автоматизированного онтологического моделирования на основе извлечения информации из текстов обычно включает в себя несколько этапов:

1. Извлечение информации из текста;
2. Структурирование данных;
3. Разработка онтологии;
4. Обучение модели.

В данной работе рассматривается извлечение квази-RDF-триплетов как замена онтологии, поэтому более подробно рассматривается первый пункт данного алгоритма. На этом этапе происходит анализ текстовых данных с применением методов обработки естественного языка и извлечение ключевых терминов, фактов и связей между ними.

Весь процесс и обработка данных производятся на основе инструментария языка Python, молулей collections, operator, itertools; все табличные и матричные операции выполнены с использованием средств модуля Pandas. Текстовые данные очищаются от HTML-тегов и неалфавитных символов. Лемматизация слов производится с использованием морфологического анализатора «pymorphy2» и Python пакета NLTK. Тем самым получается отфильтрованный «мешок слов» по предметной области.

2.2.1 Извлечение биграмм и триграмм из текста

Для создания RDF-триплетов необходимо извлечь из текста так называемые триплеты слов, или триграммы. В качестве предварительного шага для извлечения триплетов производится поиск наиболее встречающихся в наборе данных пар слов, или биграмм. Для семантической опоры триплетов берется порог с количеством встречи биграммы в наборе данных 100 раз. Под данную границу подошло всего 3428 биграмм из имеющихся 1318538. На основе отобранных биграмм генерируется два подмножества биграмм с учетом порядка морфологических характеристик, составляющих биграмму кортежа: подмножество биграмм «Прилагательное-Существительное» и подмножество биграмм «Существительное-Прилагательное».

2.2.2 Генерация RDF-триплетов

На основе отобранных биграмм типа «Прилагательное-Существительное» и «Существительное-Прилагательное» по транзитивной общности элементов кортежей построены 3998 триграмм «Существительное-Прилагательное-Существител из которых лишь 1357 получены методом NLTK последовательным просмотром текста с окном слов 3.

Полученные биграммы, являющиеся, сущностями предметной области с характеризующими их признаками, являются основой для построения триплетов.

На первом этапе триплеты строятся отдельно для каждого предложения. На втором этапе производится агрегация триплетов по всем новостям, при которой частоты триплетов суммировались по всем новостям. В виду вычислительной сложности данной процедуры агрегация осуществляется с добавлением 1000 наиболее частых триплетов по каждой 1000 новостей. Итоговый список составляет 28276 триплетов. Для дальнейшего использования из триплетов оставляются только те, которые упоминаются в тексте минимум 20 раз.

2.2.3 Составление инвертированного списка

Для дальнейшего расширения запроса, который подается на вход, строится так называемый инвертированный список.То есть для каждого слова в подготовленных триплетах необходимо составить список триплетов, в которых оно находится. Тогда при обнаружении биграммы или любого другого

словосочетания из триплета, можно легко найти недостающее слово, сущность или предикат из триграммы, тем самым расширяя запрос.

2.3 Принцип поиска документов для запроса

Перед поиском документов производятся следующие подготовительные действия:

1. Из новости берется только заголовок, так как заголовок лучше всего описывает новость;
2. Лемматизацией выделяются ключевые слова из заголовка;
3. Выделяются все возможные тройки ключевых слов (токенов), которые являются исходными запросами;
4. Проводится расширение для каждого полученного запроса;
5. Определяется порог количества слов, которые входят и в запрос и в содержание документа;

Поиск документов ведется по следующему алгоритму:

1. Для каждого слова из запроса получается его инвертированный список документов, в которых оно присутствует;
2. Все номера документов записываются в один словарь вида «документ — количество слов из запроса»;
3. Если в документе используется слов из запроса больше или равно чем порог, то данный документ попадает в итоговый список;
4. Итоговый список возвращается функцией как результат поиска.

2.4 Проведение экспериментов для статистического анализа

Далее работе проводятся эксперименты для сравнения работоспособности метода расширения с помощью квази-RDF-триплетов. Словарь RDF-триплетов составляется на основе 31077 уникальных новостях сайта СГУ. Для каждого слова составляется инвертированный список с номерами документов, в которых присутствует данное слово.

Эксперименты проводятся по следующей схеме:

1. Выбирается доминантная тема;
2. Выделяется случайная выборка из 50 документов;
3. Для каждого документа проводятся операции описанные в 2.3;
4. Полученные результаты группируются по длине исходного запроса и усредняются.

2.4.1 Общие выводы по экспериментам

Для большего удобства результаты для доминантной темы собраны в таблицах 1 и 2:

Таблица 1 – Проценты доминантной темы для расширения по теме «Science and Research»

Метод	3 слова		4 слова		5 слов	
расширения	до	после	до	после	до	после
RDF-триплеты	15,77	14,15	15,49	13,62	15,13	9,71
Тезаурус	15,77	13,28	15,49	13,33	15,13	12,46

Таблица 2 – Проценты доминантной темы для расширения по теме «Learning and Educational Services»

Метод	3 слова		4 слова		5 слов	
расширения	до	после	до	после	до	после
RDF-триплеты	24,6	23,73	24,43	23,49	31,96	25,52
Тезаурус	24,6	23,78	24,43	23,91	31,96	32,21

Исходя из всех полученных данных по расширению с помощью тезауруса и расширения с помощью RDF-триплетов можно выделить следующие моменты:

— Распределение результатов по темам расширения запроса очень хорошо коррелирует с распределением результатов исходного;
— Для эксперимента по второй темы результаты расширений обоих методов коррелируют с распределением новостей по темам, поэтому для второго эксперимента взята тема с наименьшим коэффициентом корреляции со второй темой;
— Для эксперимента по третьей темы, произошла явная фокусировка поиска на доминантной теме, как для исходного запроса, так и для расширения, однако остальная часть диаграмм все ещё хорошо коррелирует с распределением новостей по темам;
— Из предыдущих пунктов следует, что расширение скорее не следует заданной теме, а равномерно по всем темам;
— Как показано на диаграммах, расширение по тезаурусу в большинстве случаев фокусируется на теме лучше, чем расширение на триплетах.

Несмотря на то, что метод на основе триплетов хорошо коррелирует с распределением исходного и для большинства тем это позволяет им сохра-

нить примерное процентное соотношение по найденным документам, что уже неплохо. И хотя расширение на RDF-триплетах и показало себя чуть хуже расширения на тезаурусе, это компенсируется скоростью работы данного расширения, так как строить квази-RDF-триплеты гораздо быстрее, чем тезаурус. При этом работает расширение на триплетах в разы быстрее расширения на тезаурусе. Также показатели могли оказаться хуже из-за сравнительно малого словаря часто встречаемых триплетов, что напрямую связано с числом документов.

2.5 Результаты и направления дальнейших исследований

Как показано в разделе 2.4, расширение запроса на основе квази-RDF-триплетов удовлетворительно показало себя на случайной выборке в качественном и хорошо в скоростном аспектах. В сравнении с расширением на основе тезауруса разработанный метод показал себя конкурентоспособным, что показано на всех диаграммах в разделе 2.4.

Описанный метод позволяет решать задачи по обработке текста без построения онтологии или тезауруса предметной области, что существенно сокращает время работы подготовительной части к расширению запроса.

Дальнейшие исследования можно проводить в области фокусировки данного метода на доминантной теме.

ЗАКЛЮЧЕНИЕ

Таким образом в работе был разработан метод расширения запроса с помощью квази-RDF-триплетов.

В ходе данной работы были выполнены следующие задачи:

- подготовка данных предметной области;
- извлечение сущностей предметной области;
- генерация квази-RDF-триплетов;
- реализация модели расширения запроса;
- проведен ряд экспериментов и статистический анализ показывающие качество работы реализованной модели.

В ходе проведенных исследований метод показал свою работоспособность и конкурентоспособность в расширении токенов запроса. Был проведен статистический анализ реализованного метода и сравнительный анализ с методом на основе тезауруса.