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

ИНФОРМАЦИОННЫЙ ПОИСК НА ОСНОВЕ СЕМАНТИЧЕСКОГО РАСШИРЕНИЯ ЗАПРОСА

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 451 группы
направления 09.03.04 — Программная инженерия
факультета КНиИТ
Чичакян Амалии Ашотовной

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

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

Саратов 2024

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В условиях постоянно растущих объемов данных и информационного разнообразия, информационный поиск сталкивается с рядом серьезных проблем. Одной из ключевых проблем является трудность нахождения релевантных документов, которые удовлетворяют информационным потребности пользователей. Традиционные методы информационного поиска, основанные на точном совпадении ключевых слов поискового запроса с содержимым целевых документов, часто не справляются с этой задачей, поскольку не учитывают семантические связи между словами. В результате пользователи не могут получить набор документов релевантных исходному запросу.

Данная задача рассматривается в рамках области исследования Information Retrieval (IR) — информационного поиска. Наиболее частым рассматривается итеративный поиск в QA-model.

Основные проблемы, которые возникают в контексте информационного поиска, включают:

- Повышение релевантности результатов поиска. Стандартные методы поиска на основе ключевых слов не всегда способны найти все релевантные документы, особенно если пользователи используют различные термины для описания одних и тех же концептов.
- Нечеткая интерпретация пользовательских запросов. Пользователи могут формулировать свои запросы неоднозначно или использовать термины, не полностью отражающие их информационные потребности.
- Сложность использования семантических связей предметной области. Традиционные методы поиска игнорируют важные семантические связи между терминами, что приводит к неполным результатам.

Для решения этих проблем предложен ряд методов, в частности для повышения эффективности поиска используется метод расширения запроса на основе онтолгии предметной области или тезауруса соответствующего языка. Методы расширения запроса на основе тезауруса, хотя и улучшают полноту поиска, также не учитывают более глубокие семантические связи, что ограничивает их эффективность. Метод расширения на основе онтологии позволяет учитывать более глубокие смысловые связи между терминами, улучшая интерпретацию запросов и повышая избирательность результатов поиска.

Цель данной работы состоит в разработке метода расширения запроса на

основе онтологии заданной предметной области. Для достижения поставленной цели необходимо решить следующие задачи:

- Изучить существующие методы расширения запроса, включая методы на основе тезауруса и онтологий.
- Изучить средства редактирования и хранения онтологий (Protege)
- Реализовать доступ к сущностям онтологии в рамках Protege.
- Реализовать метод семантического расширения запроса на основе онтологии.
- Провести экспериментальное сравнение результатов информационного поиска с использованием методов расширения запроса на основе онтологии и тезауруса.
- Проанализировать полученные результаты и сделать выводы о целесообразности применения семантического расширения запроса на основе онтологии.

1 Теоретические основы информационного поиска и семантического расширения запроса

1.1 Основные понятия информационного поиска

1.1.1 Информационный поиск

Информационный поиск (ИП) — процесс нахождения релевантной информации в данных с целью предоставления наиболее подходящих документов пользователю.

1.1.2 Модели информационного поиска

Существующие классические модели информационного поиска: булева, векторная, вероятностная, например, модель BM25.

1.1.3 Метрики оценки качества

Для оценки качества поиска используют точность, полноту, F-меру.

1.2 Методы расширения запроса

1.2.1 Расширение запроса: понятие и цели

Расширение запроса — это процесс добавления терминов к запросу для улучшения результатов поиска.

1.2.2 Расширение запроса: методы

Существующие подходы к расширению запроса: синонимическое, тезаурусное, онтологическое и другие.

1.2.3 Расширение запроса на основе тезауруса

Тезаурус расширяет запрос, добавляя синонимы и связанные термины. Процесс включает идентификацию ключевых терминов, поиск синонимов и связанных терминов, и формирование расширенного запроса.

1.2.4 Семантическое расширение запроса: теоретические аспекты

Семантическое расширение запроса использует онтологии и семантические сети для учета смысловых связей между терминами. Онтологии представляют знания через концепты и их отношения, семантические сети — через графовые структуры.

1.3 Онтологии и их роль в семантическом расширении запроса

1.3.1 Семантическая сеть

Семантический веб (SW) преобразует весь интернет в глобальную базу данных, расширяя текущие стандарты. Он использует RDF для представления данных в виде тройки (субъект-предикат-объект) и онтологии для моделирования сущностей и их связей. Основные технологии включают RDF Schema (RDFS) и Web Ontology Language (OWL), которые расширяют RDF для более сложных представлений и рассуждений. SPARQL — основной язык запросов для RDF, позволяющий выполнять сложные поисковые запросы и улучшать качество поиска.

1.3.2 Понятие онтологии и ее структура

Онтология представляет собой формальное описание предметной области, включающее концепты, отношения между ними, аксиомы (правила и ограничения) и индивидуумы (конкретные экземпляры). Онтология структурируется в виде графа, где узлы представляют концепты, а ребра отражают отношения между ними.

1.3.3 Разработка и использование онтологий

Разработка онтологий включает несколько этапов: определение предметной области, сбор и анализ данных, формализация онтологии, валидация и тестирование, поддержка и обновление.

Онтологии используются в информационном поиске, семантическом сети, системах искусственного интеллекта.

1.3.4 Примеры онтологий и их применение в информационном поиске

Различные онтологии, применяемые в информационном поиске: WordNet, Gene Ontology и DBpedia. Они способствуют улучшению интерпретации запросов, повышению релевантности результатов и поддержке многоязычного поиска.

2 Расширение запроса на основе онтологии

2.1 Постановка эксперимента

Для оценки эффективности методов расширения запроса на основе онтологии и тезауруса был проведен эксперимент. Эксперимент осуществлялся на корпусе новостей университета.

2.1.1 Цели и задачи эксперимента

Цель эксперимента - сравнить эффективность методов расширения запроса на основе онтологии и тезауруса в информационном поиске. Задачи включают оценку точности и полноты результатов, сравнение степени расширения ответа, анализ зависимости от количества слов в запросе и изучение влияния на распределение тем новостей.

2.1.2 Методика проведения эксперимента

Эксперимент включает следующие этапы: предварительная обработка корпуса новостей университета, разработка методов расширения запроса на основе онтологии и тезауруса, проведение поисковых запросов с использованием различных методов расширения и без них, анализ полученных результатов.

2.1.3 Описание данных и стека технологий

Для проведения эксперимента использовались данные о новостях университета, матрицы Theta и Phi из тематического моделирования, онтология, созданная в Protege, и русскоязычный тезаурус RuWordNet. Использовались библиотеки: supervenn, pymorphy2, python-Levenshtein, gensim и SPARQLWrapper, все на языке программирования Python.

2.2 Реализация методов расширения запроса

Для успешной реализации методов расширения запроса в информационном поиске требуется предварительная обработка текстов.

Пороговая функция играет ключевую роль в отборе новостей на основе длины запроса пользователя: она определяет минимальное количество слов, которое должно быть в документе для его включения в результаты поиска.

2.2.1 Эксперимент по выявлению новостей ядра

В данном разделе описана реализация эксперимента по выявлению новостей ядра. Результаты показывают, что при расширении онтологией количество новостей ядра значительно больше, чем при расширении тезаурусом.

2.2.2 Реализация расширения запроса на основе тезауруса

Этот раздел описывает алгоритм расширения запроса на основе тезауруса. Код осуществляет процесс расширения запроса и сбор статистики для анализа. Результаты показывают, что 3% слов отсутствуют в словаре тезауруса.

2.2.3 Реализация семантического расширения запроса на основе онтологии

Этот раздел описывает алгоритм семантического расширения запроса на основе онтологии. Для каждого слова формулируются и выполняются SPARQL-запросы. Также проводится сбор статистики. Результаты показывают, что 21% слов отсутствуют в онтологии.

2.3 Сравнительный анализ результатов эксперимента

Так как в нашем случае нельзя напрямую посчитать точность и полноту результатов, мы будем использовать распределение тем на основе тематического моделирования для оценки полученных ответов.

2.3.1 Результаты поиска без расширения запроса

В этом разделе описаны результаты поиска без расширения запроса, которые служат базовой линией для сравнения с результатами, полученными при использовании различных методов расширения запроса. Видно, что процентное распределение тем остаётся практически неизменным при увеличении числа слов в запросе, за исключением небольших колебаний.

2.3.2 Результаты поиска с семантическим расширением запроса

В этом разделе представлены результаты поиска с семантическим расширением запроса.

Расширение запроса с использованием онтологий значительно увеличивает количество найденных новостей, особенно для запросов с меньшим числом слов. Процентное распределение тем остаётся стабильным, несмотря на изменение количества слов в запросе. Запросы, для которых тема "Студенческие

соревнования"является доминантой, показывают практически одинаковые результаты, независимо от использования расширения.

2.3.3 Результаты поиска с расширением запроса на основе тезауруса

В этом разделе представлены результаты поиска с расширением запроса на основе тезауруса.

Расширение запроса с использованием тезауруса демонстрирует увеличение количества найденных новостей, но без чёткой закономерности. Распределение тем новостей остаётся практически неизменным при увеличении числа слов в запросе.

2.3.4 Оценка результатов сравнительного анализа

Графики показывают, что расширение запроса, будь то онтология или тезаурус, не меняет существенно процентное распределение тем новостей. В случае с онтологией наблюдается закономерность, выявленная ранее: значительное увеличение количества новостей при меньшем числе слов в запросе и снижение степени расширения с увеличением числа слов.

ЗАКЛЮЧЕНИЕ

В данной дипломной работе была рассмотрена проблема повышения эффективности информационного поиска за счет использования методов семантического расширения запроса. В условиях постоянно растущих объемов данных и увеличивающегося разнообразия информации традиционные методы поиска на основе ключевых слов оказываются недостаточными. Они не учитывают семантические связи между терминами, что приводит к неполным и не всегда релевантным результатам.

Для решения данной проблемы был предложен метод семантического расширения запроса на основе онтологии. Онтологии позволяют моделировать предметные области и учитывать смысловые связи между терминами, что улучшает интерпретацию запросов и повышает релевантность найденных документов.

В рамках работы был разработан и внедрен метод расширения запросов, использующий онтологию новостной ленты сайта университета. Экспериментальные результаты показали, что использование онтологии приводит к более избирательному и точному расширению запросов, чем использование тезауруса. На основе полученных результатов можно рекомендовать внедрение разработанного метода расширения запросов на основе онтологии для улучшения системы поиска в новостной ленте университета.