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

ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ ИСПОЛЬЗОВАНИЯ МЕЛКОЗЕРНИСТОЙ КЛАСТЕРИЗАЦИИ В ЗАДАЧЕ ИНФОРМАЦИОННОГО ПОИСКА

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 411 группы
направления 02.03.02 — Фундаментальная информатика и информационные технологии
факультета КНиИТ
Соломонова Анатолия Владимировича

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

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

Саратов 2025

ВВЕДЕНИЕ

Актуальность работы.

В цифровую эпоху объёмы информации стремительно увеличиваются в экспоненциальной прогрессии. Это создаёт потребность не только в эффективных методах хранения и обработки данных, но и в средствах быстрого и точного поиска необходимой информации. Одной из ключевых задач в этом контексте выступает информационный поиск.

Несмотря на рост мультимедийных данных, таких как изображения, видео и аудио, объём текстовой информации продолжает увеличиваться особенно интенсивно. Поэтому поиск по текстовым коллекциям остаётся одной из центральных проблем в области информационного поиска.

Настоящая работа посвящена повышению эффективности работы информационного-поисковых систем при нечётких условиях поиска.

Цель и задачи бакалаврской работы.

Для решения данной проблемы основной целью работы поставлена разработка метода расширения результатов запроса к поисковой системе на основе мелкозернистой кластеризации набора данных.

Достижение поставленной цели включает в себя следующие задачи:

1. Изучение способов повышения эффективности информационного поиска на основе методов определения семантической близости текстовых документов.
2. Разработка метода мелкозернистой кластеризации набора данных.
3. Разработка метода расширения результатов запроса на основе теории грубых множеств.
4. Выбор средств реализации задач.
5. Программная реализация вопросно-ответной системы на основе разработанных методов.
6. Оценка эффективности информационного поиска методом расширения результатов запроса на основе тематической модели для набора данных.

Структура и объём работы.

Для решения поставленных задач выполнена выпускная квалификационная работа, включающая в себя введение, 5 основных глав, заключение, список использованных источников из 20 наименований и 6 приложения. Работа изложена на 54 страницах.

Первая глава «Информационный поиск» посвящена изучению понятия информационного поиска, а также используемых инструментов и подходов для его реализации.

Во второй главе «Вопросно-ответные системы и проблема формулировки запроса» рассматривается подход к реализации информационного поиска в виде вопросно-ответной системы, а также рассматривается проблема неопределённости и многозначности формулировки запроса на естественном языке и способы её решения.

Третья глава «Мелкозернистая кластеризация» посвящена подходу к кластеризации данных. В ней предоставлены описание, преимущества и недостатки подхода мелкозернистой кластеризации для использования в задаче информационного поиска, а также предлагается алгоритм реализации.

Четвёртая глава «Программная реализация вопросно-ответной системы» содержит описание используемых технологий, а также подробное описание процесса реализации, включая мелкозернистую кластеризацию данных и вопросно-ответную систему.

В пятой главе «Оценка релевантности результатов поиска» проводится тематическое моделирование набора данных и строится имитационная модель для оценки эффективности исследуемого в работе подхода к расширению результатов поисковой системы.

Выпускная квалификационная работа заканчивается заключением, списком использованных источников, а также приложениями А-Е с наборами данных, расширенными результатами моделирования системы и кодом.

Основное содержание работы

Информационный поиск (англ. Informational Retrieval, IR) представляет собой задачу выявления релевантных документов из коллекции данных в ответ на пользовательский запрос. Цель информационного поиска — предоставить пользователю документы, наиболее подходящие по содержанию к той информации, в которой он заинтересован.

Ключевыми компонентами системы информационного поиска являются:

- **пользовательский запрос**, описывающий интерес или проблему пользователя;
- **коллекция документов**, по которой осуществляется поиск;
- **модель сопоставления** запросов и документов (включая метрики оценки сходства).

Результатом работы такой системы обычно является ранжированный список документов, упорядоченный по убыванию степени релевантности запросу. Релевантность может оцениваться с использованием различных моделей и метрик: от простых лексических (например, TF-IDF, BM25) до сложных нейросетевых (на основе языковых моделей и эмбеддингов).

Инструменты информационно-поисковых систем. Эффективный поиск требует инструментов для представления, индексирования и сопоставления текстов. Современные информационно-поисковые системы используют как классические алгоритмы, так и методы машинного обучения и обработки естественного языка, однако последние требуют значительных вычислительных мощностей.

Ключевым элементом в классических алгоритмах является инвертированный индекс, позволяющий быстро находить документы по словам. В некоторых системах применяются сразу два индекса вместе: один хранит списки документов, другой — позиции слова в тексте.

Для представления текста используются частотные модели, например, «мешок слов» и TF-IDF, оценивающий важность термина с учётом его частоты в документе и в коллекции. Эти подходы не учитывают порядок слов и их семантические связи, однако дают хорошие результаты при работе с техническими и тематически однородными текстами.

Метод TF-IDF остаётся одним из основных для оценки релевантности: он придаёт больший вес словам, часто встречающимся в конкретном документе,

но редко — в остальных.

Кроме того, для корректной работы всех вышеперечисленных компонентов важна предобработка текста, приводящая его к представлению, понятному компьютеру:

- приведение к одному регистру;
- удаление пунктуации и стоп-слов;
- лемматизация или стемминг;
- токенизация — разделение текста на слова или смысловые единицы.

Вопросно-ответные системы и проблема формулировки запроса. Вопросно-ответные системы (англ. Question Answering, QA) являются одним из направлений информационного поиска, нацеленным на предоставление пользователю ответа на вопрос, сформулированный им на естественном языке. Такие системы стремятся минимизировать участие пользователя в процессе уточнения формулировки запроса и автоматизировать интерпретацию и поиск релевантной информации.

Одной из ключевых проблем поисковых систем, принимающих запрос на естественном языке является неопределённость и многозначность формулировки запроса. Пользователи часто формулируют вопросы неструктурированно, используют синонимы, сокращения и могут не попасть по ключевым словам тех документов, которые они хотят найти.

Данная проблема может быть решена с помощью расширения ответа системы, которое можно разделить на два различных подхода:

- **Расширения запроса**: на этапе обработки запроса к запросу добавляются новые слова, семантически или тематически близкие к заданным пользователем. Например, синонимы и гипонимы.
- **Расширение результата**: на этапе формирования результата множество найденных системой релевантных документов дополняется новыми, близкими по содержанию с уже имеющимися. Например, на основе кластеризации данных.

В основе расширения результатов на основе кластеризации лежит теория грубых множеств — она основывается на понятии приближений множества с использованием доступной информации о подобии объектов. Основным понятием теории являются аппроксимации: нижняя и верхняя. Нижнее приближение множества содержит все объекты, которые однозначно могут быть отнесены к

интересующему нас классу, а верхнее — те, которые могут к нему принадлежать.

На основе её принципа можно считать, что если кластер содержит хотя бы один точно релевантный документ, то остальные элементы кластера можно считать возможно релевантными и также добавить в ответ системы.

Мелкозернистая кластеризация является подходом к кластеризации, при котором учитываются более мелкие детали и получаются более мелкие кластеры, что позволяет избежать переизбытка документов в ответе системе при использовании в задаче расширения результата.

В сравнении с классическими прямыми методами кластеризации, данный подход имеет ключевые преимуществ для использования в рассматриваемой задаче:

1. **Учёт деталей**: позволяет учитывать более мелкие и незначительные различия между объектами, что может быть полезно при работе с данными, где малейшие отличия имеют значение.
2. **Точность**: благодаря более дательному анализу может дать более точное представление о взаимосвязях и структуре данных, особенно когда важны тонкости.
3. **Равномерность**: позволяет обеспечить более равномерное и мелкое распределение данных по группам благодаря большему количеству кластеров меньшей размерности.

Из недостатков подхода можно выделить вычислительную сложность при обработке крупного набора данных. Данный недостаток может компенсироваться тем, что кластеризация происходит не во время работы системы, а перед её запуском, а её результаты можно сохранять и подгружать при повторных использованиях.

Реализация мелкозернистой кластеризации. Для выполнения кластеризации текстовых данных применялось их векторное представление с использованием взвешивания TF–-DF. Для последующей обработки высокоразмерных и разреженных векторов использовалось снижение размерности методом сингулярного разложения (Truncated SVD), известного как скрытый семантический анализ (LSA), с последующей нормализацией. На полученные данные последовательно применялся разработанный в работе алгоритм мелкозернистой кластеризации, с базового методом в виде алгоритма KMeans. Кластеризация прово-

дилась итеративно: на каждой итерации крупные кластеры (с числом элементов свыше 75) делились на подмножества. При этом малочисленные кластеры (менее 25 элементов) объединялись с ближайшими по размеру, что позволило избежать избыточной фрагментации.

Такой подход позволил получить иерархическую структуру кластеров. В результате обработки корпуса из 31077 статей было сформировано 644 конечных кластера. Несмотря на вычислительную затратность в виде около 20 минут выполнения, реализованный метод позволил достичь высокой степени тематической детализации.

Программная реализация вопросно-ответной системы.

Реализация вопросно-поисковой системы разделяется на несколько модулей:

1. Лемматизация текстовых данных;
2. Построение обратного индекса;
3. Функция поиска релевантных документов по запросу с расширением результатов.

Используемые среда и технологии. В ходе работы использовалась среда Jupyter Notebook, обеспечивающая интерактивную разработку, пошаговую отладку и визуализацию результатов.

Для реализации алгоритмов кластеризации применялась библиотека scikit-"learn, включающая метод кластеризации K-Means, а также средства визуализации.

Дополнительно использовались следующие библиотеки:

— Pandas — для обработки табличных данных;
— pymorphy3 и NLTK — для лемматизации и предварительной обработки текстов;
— Matplotlib — для построения графиков диаграмм.

Подготовка данных для использования вопросно-ответной системой

Планируемая вопросно-ответная система в настоящей работе будет строиться на основании табличных данных, содержащих набор из 31077 новостных статей с сайта Саратовского государственного университета за период 2010–2022 годов.

Лемматизация текстовых данных. Для компьютерной обработки текстовой информации с определением смыслового содержания важно привести

данные к наиболее подходящему виду, чтобы одинаковые слова с разными склонениями приводились к одному, а слова, не несущие смысловой нагрузки, не учитывались для сравнения содержания различных текстов. Такая обработка называется лемматизацией, при которой из текста удаляются заданные стоп-слова и знаки пунктуации, а оставшиеся слова приводятся к своей начальной форме и называются токенами.

Для реализации лемматизации в настоящей работе используются библиотека NLTK для удаления знаков пунктуации, а также библиотека pymorphy3 для приведения слов к начальной форме. В качестве стоп-слов приводится несколько прописанных списков: расширенный список стоп-слов, изначально полученный из библиотеки NLTK через $stopwords.words('russian')$, собственный список, а также список имён, которые могут встретиться в текстах.

Обратный индекс. Для соотношения конкретных слов к текстам необходима информация о том, в каких документах встречается данное слово, включая наличие не только в основном тексте, но и в заголовке. С этой целью строится обратный индекс — словарь вида «слово: список документов, содержащих данное слово». В настоящей работе обратный индекс представляется с помощью класса и для эффективности повторных запусков сохраняется в новый лист таблицы с набором данных.

Общая схема работы реализованной вопросно-ответной системы. Разработанная система поиска с расширением результатов на основе мелкозернистой кластеризации функционирует по следующей последовательности шагов:

1. **Получение пользовательского запроса**. Система принимает пользовательский запрос на естественном языке.
2. **Предобработка запроса**. Запрос проходит лемматизацию и токенизацию: удаляются знаки препинания и стоп-слова, а слова приводятся к начальной форме. Результатом является список список токенов, используемых для поиска.
3. **Поиск по обратному индексу**. Система использует обратный индекс, чтобы найти все документы, содержащие хотя бы один из полученных токенов.
4. **Определение релевантных кластеров**. Для каждого найденного документа определяется, в каком кластере он находится. На этом этапе система формирует множество кластеров, содержащих хотя бы один релевант-

ный документ.

5. **Расширение результатов**. В итоговый результат добавляются все документы из найденных кластеров. Таким образом, результат поиска дополняется документами, близкими по тематике к уже найденным.
6. **Формирование итогового списка**. Полученные документы объединяются в итоговый ответ системы, который возвращается пользователю.

Оценка релевантности результатов поиска.

Для оценки улучшения общей релевантности ответа системы после расширения результатов запроса необходим дополнительный способ оценки релевантности документов, независимый от реализованной кластеризации. Наиболее подходящим для этой цели является проведение тематического моделирования набора данных для разбиения статей на несколько тематик, по которым и будет строиться оценка релевантности.

Тематическое моделирование. Перед тем, как присвоить каждому документу его тематику, следует определиться, на сколько категорий нужно разделить тексты и обозначить их названия для лучшего понимания.

После ручной оценки примерного количество различных тематик используемого набора статей, можно выделить следующие:

1. History of SSU and memorable days;
2. SSU Club Events;
3. Educational and scientific activities;
4. Research activities;
5. Sport;
6. Youth projects and competitions.

Для построения тематической модели использовалась библиотека BigARTM, предоставляющая реализацию метода аддитивной регуляризации тематических моделей (ARTM). Эта библиотека позволяет настраивать множество параметров модели, в том числе разреженность, сглаживание и количество тем, а также обеспечивает высокую производительность на больших текстовых коллекциях. В результате моделирования набор данных был разделён на вышеупомянутые категории размерами 4427, 3885, 5794, 5741, 4523 и 6787 статей, соответственно. Также в таблицу с набором данных добавлен новый столбец «$Topic\_Prediction$» со значениями из списка тематик, соответствующими каждому документу.

Имитационная модель поиска. Для более эффективной и автоматизированной оценки была построена имитационная модель поиска по статьям одной определённой темы.

Модель представляет собой цикл, отправляющий 200 запросов по одной заданной теме к вопросно-ответной системе и сохраняющий информацию о том, сколько суммарно статей соответствующей темы встретилось в результатах поисков. Тексты запросов будем собирать из случайных слов лемматизированного заголовка одной из статей интересуемой темы.

Визуализация результатов. Для наглядной оценки эффективности расширения результатов вопросно-ответной системы были построены гистограммы распределения тем статей в выборках до и после применения механизма расширения. На рисунках 1, 2 и 3 представлены результаты тематического анализа по двум различным темам: «История университета», «Клубная деятельность» и «Образовательные и научные активности», соответственно.

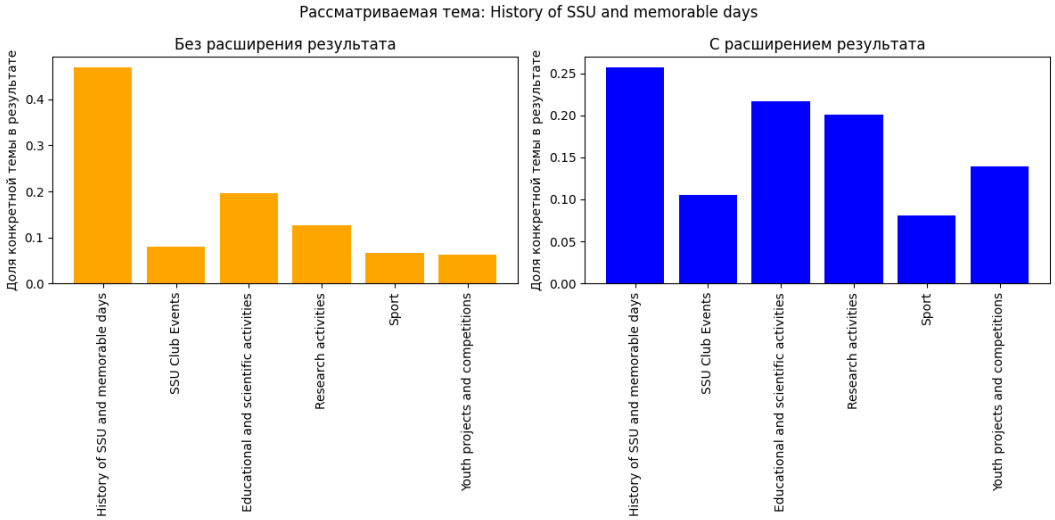

Рисунок 1 – Распределение тем в результатах по запросу «History of SSU and memorable days»

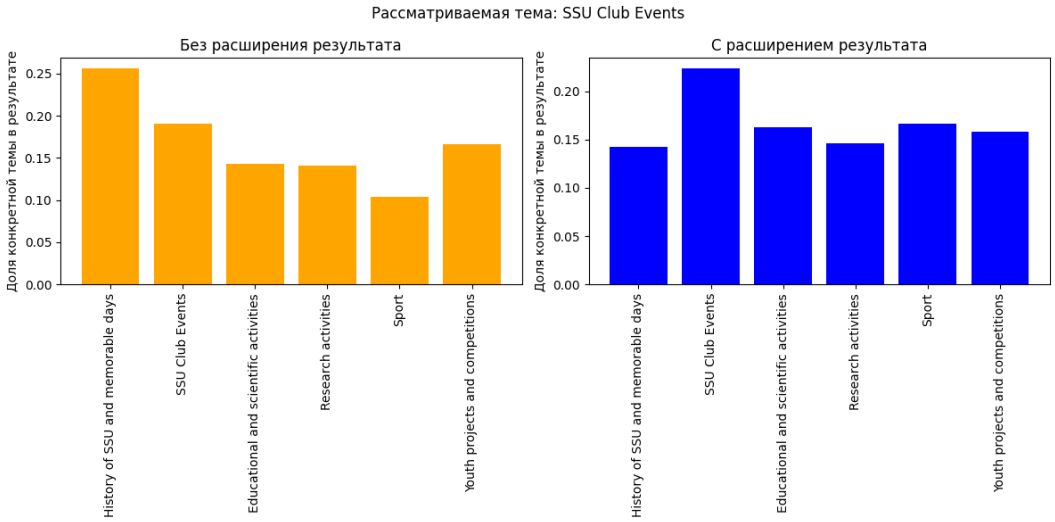

Рисунок 2 – Распределение тем в результатах по запросу «SSU Club Events»

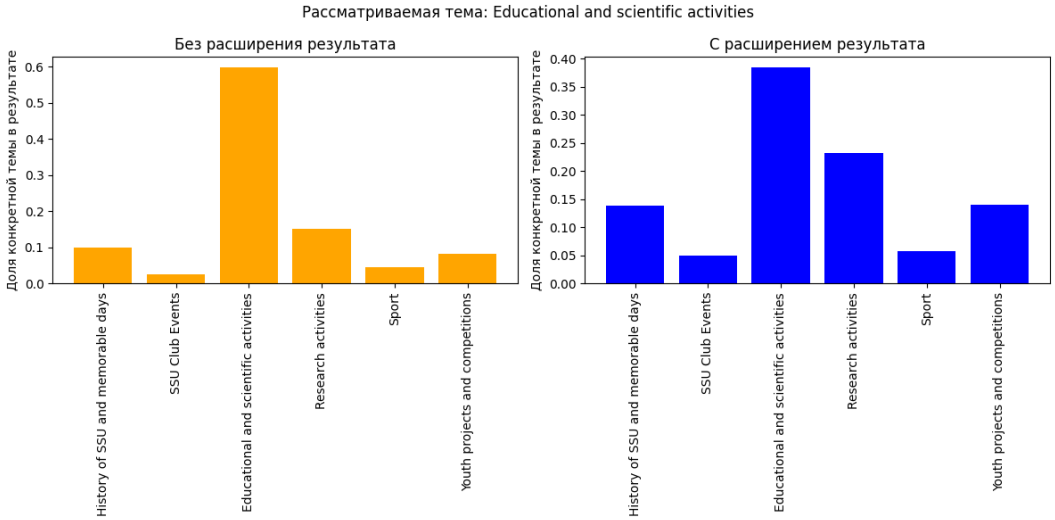

Рисунок 3 – Распределение тем в результатах по запросу «Educational and scientific activities»

Графики показывают, что расширение увеличивает число найденных документов, включая не только по целевой теме, но и по остальным — возможно, из-за семантической близости или неочевидной классификации. При этом целевая тематика остаётся или становится преобладающей: например, на рисунке 2 она занимает лидирующую долю только после применения расширения.

Таким образом, визуализация подтверждает, что расширение действительно способствует увеличению полноты выборки по запрашиваемой теме, хотя и может сопровождаться ростом количества нерелевантных текстов.

ЗАКЛЮЧЕНИЕ

Выполненная работа была направлена на повышение эффективности информационно-поисковых систем в условиях нечёткой формулировки запроса. Предложенный и реализованный подход основан на расширении результатов поиска на основе мелкозернистой кластеризации, что позволило учитывать семантическую близость документов и повысить полноту выдачи без существенного снижения её тематической точности.

Разработанная вопросно-ответная система использует обратный индекс для первоначального поиска, после чего применяет механизм расширения результатов за счёт включения всех документов из кластеров, где были найдены релевантные тексты. Это позволило преодолеть ограничения, связанные с неопределённостью и многообразием формулировки запроса.

Оценка эффективности разработанного подхода была проведена с использованием тематического моделирования на основе библиотеки BigARTM, а также имитационной модели, производящей серию поисковых запросов по заданным темам. Визуализация результатов показала, что применение механизма расширения увеличивает общее количество документов в ответе по всем тематикам. При этом наибольшее увеличение наблюдается именно по теме исходного запроса. Благодаря этому, несмотря на общий рост долей других тематик, рассматриваемая тема сохраняет или занимает преобладающую долю в ответах системы. Это указывает на то, что расширение способствует улучшению полноты результатов без существенного снижения тематической направленности.

Таким образом, можно сделать вывод, что применение мелкозернистой кластеризации в задаче расширения результатов информационного поиска является перспективным направлением, позволяющим повысить релевантность и полноту поиска в условиях нечёткой или неполной формулировки запроса.