МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

РАЗРАБОТКА РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ НА ОСНОВЕ АЛГОРИТМА ПОИСКА ПАТТЕРНОВ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 248 группы
направления 09.04.03 — Прикладная информатика

механико-математического факультета
Богоявленского Виталия Георгиевича

Научный руководитель
Зав. кафедрой, д. ф.-м. доцент _____ С. П. Сидоров

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2026

Общая характеристика работы

Актуальность темы исследования. Современные рекомендательные системы являются неотъемлемой частью информационных платформ, цифровых библиотек, образовательных сервисов и научных репозиториев. Увеличение объёма цифрового контента приводит к существенному усложнению процессов поиска релевантной информации. Особенно остро данная проблема проявляется в научной среде, где количество публикуемых статей ежегодно возрастает, а исследователю требуется оперативно находить материалы, соответствующие тематике текущей работы и области научных интересов.

Существующие рекомендательные системы преимущественно основываются на методах коллаборативной фильтрации, статистического анализа текстов либо на применении нейросетевых моделей и эмбеддингов. Несмотря на высокое качество рекомендаций, подобные подходы обладают рядом ограничений. Коллаборативные методы существенно зависят от объёма пользовательских данных и подвержены проблеме холодного старта. Методы, основанные на эмбеддингах и трансформерных архитектурах, требуют значительных вычислительных ресурсов и зачастую характеризуются низкой интерпретируемостью результатов. В задачах анализа научных публикаций прозрачность формирования рекомендаций имеет особую значимость, поскольку исследователю важно понимать, какие тематические зависимости повлияли на выбор рекомендуемых материалов.

В связи с этим актуальной задачей является разработка рекомендательной системы, способной учитывать скрытые тематические взаимосвязи между объектами текстового корпуса, обеспечивать адаптацию к пользовательским предпочтениям и одновременно сохранять интерпретируемость формируемых рекомендаций.

Объектом исследования являются рекомендательные системы, функционирующие на основе анализа текстовых данных.

Предметом исследования выступают методы поиска тематических паттернов в текстовых коллекциях и алгоритмы построения рекомендаций на основе анализа совместной встречаемости признаков.

Целью работы является разработка рекомендательной системы для подбора научных публикаций на основе алгоритма поиска паттернов, использу-

ющего матрицы соупоминаний и граф совместной встречаемости признаков.

Для достижения поставленной цели были сформулированы следующие задачи:

1. провести анализ существующих методов поиска частых наборов элементов, графовых подходов и алгоритмов рекомендательных систем;
2. исследовать особенности применения методов анализа паттернов к слабоструктурированным текстовым данным;
3. разработать алгоритм формирования пользовательского паттерна на основе матриц совместной встречаемости признаков;
4. реализовать рекомендательную систему для подбора научных публикаций;
5. выполнить сравнительный анализ разработанного подхода с TF-IDF и Sentence-BERT моделями;
6. провести экспериментальную оценку качества рекомендаций с использованием метрики Recall@K и моделирования пользовательского взаимодействия.

Методы исследования. В работе использовались методы интеллектуального анализа данных, теория графов, методы поиска частых наборов элементов, алгоритмы обработки естественного языка, статистические методы анализа текстов, а также методы построения рекомендательных систем.

Научная новизна работы заключается в разработке алгоритма рекомендаций, основанного на анализе матриц соупоминаний и графа совместной встречаемости признаков, позволяющего формировать интерпретируемые тематические паттерны пользовательских интересов без необходимости полного переобучения модели при обновлении пользовательского профиля.

Практическая значимость работы определяется возможностью применения разработанной системы для анализа научных публикаций, образовательных платформ, тематических каталогов и иных текстовых коллекций большого объёма. Реализованный подход обеспечивает адаптивное обновление пользовательских предпочтений и может использоваться в интерактивных рекомендательных сервисах.

Структура работы. Магистерская диссертация состоит из введения, двух разделов, заключения, списка использованных источников и приложений.

В первом разделе рассматриваются теоретические основы поиска паттернов и построения рекомендательных систем. Выполнен анализ алгоритмов Apriori, Eclat, FP-Growth, графовых методов и probabilistic graphical models. Рассмотрены подходы к анализу совместной встречаемости признаков и методы формирования пользовательских паттернов на основе графовых структур. Также приведено теоретическое описание предлагаемого метода поиска паттернов на основе матриц соупоминаний.

Во втором разделе описывается практическая реализация рекомендательной системы для подбора научных публикаций. Представлена характеристика используемого корпуса данных, рассмотрен применяемый технологический стек и описаны этапы построения системы. Выполнено сравнение разработанного подхода с TF-IDF и Sentence-BERT моделями. Приведены результаты экспериментальной оценки качества рекомендаций с использованием Recall@K и моделирования пользовательского взаимодействия посредством LLM-агентов.

Содержание первого раздела

Первый раздел магистерской работы посвящён исследованию теоретических основ поиска паттернов в транзакционных и текстовых данных, анализу существующих методов интеллектуального анализа информации, а также разработке подхода к формированию рекомендаций на основе матриц соупоминаний и графа совместной встречаемости признаков.

В начале раздела рассматривается задача поиска паттернов как одна из фундаментальных задач интеллектуального анализа данных. Показано, что в рамках анализа текстовых коллекций под паттерном может пониматься устойчивое сочетание признаков, совместно встречающихся внутри множества объектов корпуса. Подобные структуры позволяют выявлять скрытые тематические зависимости, обнаруживать повторяющиеся контексты и формировать пользовательские интересы на основе совокупности взаимосвязанных признаков.

В работе формализована постановка задачи поиска наиболее частого набора элементов фиксированной мощности. Рассматривается множество элементов $E = \{e_1, e_2, \ldots, e_{|E|}\}$ и база объектов $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$, где каждый объект представляет собой подмножество элементов множества E. Для множества элементов $X \subseteq E$ вводится понятие поддержки:

$$\operatorname{supp}(X) = \left|\{S \in \mathcal{S} \mid X \subseteq S\}\right|.$$

Поддержка отражает количество объектов, содержащих рассматриваемый набор элементов. На основе данного определения формулируется задача поиска множества X^* мощности n, обладающего максимальной поддержкой:

$$X^* \in \arg \max_{X \subseteq E,\, |X|=n} \operatorname{supp}(X).$$

Отмечается, что подобная постановка широко применяется в задачах поиска ассоциативных правил, тематического анализа текстов, рекомендательных систем и биоинформатики.

Далее в разделе проводится анализ существующих алгоритмов поиска частых наборов элементов. Рассматриваются классические методы семейства

Apriori, основанные на генерации и проверке кандидатов. Показано, что алгоритм Apriori использует анти-монотонное свойство поддержки, позволяющее отсекать нечастые множества ещё на ранних этапах вычислений. Несмотря на формальную простоту и интерпретируемость, данный подход характеризуется существенным ростом количества кандидатов при увеличении размерности пространства признаков.

Особое внимание уделяется вертикальным представлениям данных, используемым в алгоритмах Eclat, dFIN и negFIN. Рассматриваются структуры TID-list, Nodeset и DiffNodeset, позволяющие эффективно вычислять поддержку наборов элементов посредством пересечения множеств идентификаторов транзакций. Показано, что подобные методы демонстрируют высокую производительность при обработке плотных данных, однако требуют значительных объёмов оперативной памяти при работе с крупными текстовыми коллекциями.

Отдельно исследуются методы роста шаблонов, включая алгоритм FP-Growth. В работе рассматривается структура FP-tree и механизм построения условных деревьев шаблонов. Отмечается, что отказ от явной генерации кандидатов позволяет существенно сократить вычислительные затраты по сравнению с Apriori-подходами.

В рамках анализа существующих решений также рассматриваются методы поиска максимальных частых наборов элементов, включая алгоритмы MaxMiner, MAFIA и CarpenterMax. Подобные методы ориентированы на поиск максимально информативных паттернов без необходимости перечисления всех частых множеств.

Дополнительно исследуются подходы, основанные на ограничениях мощности множества и условиях поддержки. Рассматривается возможность постановки задачи поиска паттернов в виде задачи целочисленного линейного программирования. Для бинарных переменных x_e и y_j приводится следующая формулировка:

$$\max \sum_{j=1}^{m} y_j,$$

при ограничениях:

$$y_j \leq x_e,$$

$$\sum_{e \in E} x_e = n.$$

Показано, что подобные постановки позволяют получать точные решения, однако вычислительная сложность резко возрастает при увеличении размерности данных.

Существенная часть первого раздела посвящена графовым методам анализа транзакционных и текстовых данных. Рассматривается представление базы объектов в виде двудольного графа, связывающего элементы и транзакции. Показано, что подобная модель позволяет использовать аппарат теории графов для поиска устойчивых структур совместной встречаемости.

В работе подробно анализируются методы поиска максимальных клик и биклик, а также подходы, основанные на Formal Concept Analysis. Отмечается, что плотные подграфы естественным образом интерпретируются как устойчивые тематические паттерны внутри корпуса данных.

Дополнительно рассматриваются вероятностные графические модели, байесовские сети и методы анализа марковских оболочек. Показано, что использование условных зависимостей между признаками позволяет уменьшать размерность пространства признаков и выделять наиболее значимые взаимосвязи внутри данных. Вместе с тем построение подобных моделей требует существенных вычислительных ресурсов и большого объёма обучающих данных.

После анализа существующих подходов в первом разделе формулируется предлагаемый метод поиска паттернов на основе матриц соупоминаний. В отличие от классических алгоритмов поиска частых наборов элементов, рассматриваемый подход ориентирован преимущественно на анализ текстовых коллекций и слабоструктурированных данных.

В рамках предложенного метода каждый объект корпуса представляется множеством признаков:

$$S_j \subseteq E.$$

Для множества признаков строится матрица совместной встречаемости:

$$C = (c_{ij}),$$

где элемент матрицы определяется следующим образом:

$$c_{ij} = |\{S \in \mathcal{S} : e_i \in S \wedge e_j \in S\}|.$$

Элемент c_{ij} отражает количество объектов, содержащих одновременно признаки e_i и e_j. На основе матрицы соупоминаний формируется взвешенный граф совместной встречаемости:

$$G = (E, E_G, w),$$

где вершины соответствуют признакам, а веса рёбер определяются частотой совместного появления элементов:

$$w(e_i, e_j) = c_{ij}.$$

Для сокращения количества слабых связей вводится пороговая фильтрация. Ребро сохраняется только при выполнении условия:

$$w(e_i, e_j) \geq \beta.$$

Полученная структура позволяет интерпретировать плотные компоненты графа как тематические паттерны, устойчиво встречающиеся внутри текстового корпуса.

Далее в разделе рассматривается процесс формирования пользовательского паттерна. Пользовательский профиль описывается как динамически изменяемый граф предпочтений, формируемый на основе положительно оценённых объектов. Для множества понравившихся пользователю документов строится локальная матрица совместной встречаемости, после чего выполняется выделение наиболее плотных тематических компонент.

Показано, что подобный подход позволяет учитывать не отдельные ключевые слова, а устойчивые сочетания признаков, характеризующие интересы пользователя. Для учёта отрицательных реакций используется механизм штрафов, уменьшающий веса нежелательных связей внутри пользовательского графа.

Завершается первый раздел сравнительным анализом рассмотренных методов. Выполняется сопоставление классических алгоритмов поиска частых наборов элементов, статистических методов ранжирования, embedding-based моделей и графовых подходов.

Показано, что методы TF-IDF характеризуются низкой вычислительной сложностью, однако практически не учитывают структуру совместной встречаемости признаков. Embedding-модели обеспечивают высокое качество рекомендаций за счёт анализа скрытых семантических зависимостей, но требуют значительных вычислительных ресурсов и обладают низкой интерпретируемостью.

Предлагаемый метод на основе матриц соупоминаний занимает промежуточное положение между классическими алгоритмами поиска паттернов и современными embedding-based подходами. Использование графа совместной встречаемости позволяет учитывать структуру тематических связей внутри корпуса, обеспечивать интерпретируемость рекомендаций и поддерживать инкрементальное обновление пользовательского профиля без полного переобучения модели.

В завершение раздела делается вывод о целесообразности применения предложенного подхода при разработке рекомендательной системы для подбора научных публикаций.

Содержание второго раздела

Во втором разделе магистерской работы рассматриваются практические аспекты разработки рекомендательной системы для подбора научных публикаций на основе пользовательских интересов. Практическая часть исследования включает этапы подготовки набора данных, реализации алгоритмов рекомендаций, проектирования пользовательского интерфейса, а также проведения сравнительного анализа разработанных методов.

В качестве источника данных использовался открытый архив научных публикаций arXiv, содержащий материалы по различным научным направлениям. Для проведения экспериментов был сформирован поднабор публикаций, включающий статьи из областей компьютерных наук, машинного обучения, обработки естественного языка, анализа данных и смежных тематик. Предварительная обработка текстов выполнялась средствами языка программирования Python с использованием специализированных библиотек для анализа данных и обработки текстовой информации.

В работе была реализована система предварительной обработки текстов научных публикаций, включающая токенизацию, лемматизацию, удаление стоп-слов и выделение значимых терминов. Для решения данных задач использовались библиотеки spaCy, NumPy, pandas и NLTK. Применение указанных инструментов позволило сформировать унифицированное текстовое представление документов и подготовить данные для последующего построения рекомендательных моделей.

Практическая часть исследования основывается на сравнении трёх подходов к построению рекомендаций научных публикаций. Первый подход базируется на использовании статистической модели TF-IDF, широко применяемой при анализе текстовых коллекций и поиске релевантных документов. Данный метод позволяет оценивать важность терминов внутри документа относительно всей коллекции публикаций. Для вычисления векторных представлений текстов и оценки косинусного сходства документов использовались инструменты библиотеки scikit-learn

Второй реализованный подход основан на использовании современных языковых моделей семейства BERT. В частности, в исследовании применялась архитектура Sentence-BERT, предназначенная для построения семанти-

ческих векторных представлений текстов. Использование эмбеддингов позволило учитывать скрытые семантические связи между публикациями и повысить качество рекомендаций по сравнению с классическими статистическими методами анализа текста.

Третий подход представляет собой разработанный алгоритм рекомендаций на основе матриц со-упоминаний терминов. Предлагаемый метод базируется на построении графовой структуры совместной встречаемости ключевых понятий в пользовательских предпочтениях. В отличие от классических алгоритмов поиска частых наборов, таких как Apriori, FP-Growth и PrefixSpan, разработанный подход ориентирован не только на поиск повторяющихся элементов, но и на выявление устойчивых тематических взаимосвязей между терминами внутри текстовых коллекций.

Теоретические предпосылки графового представления паттернов опираются на исследования в области поиска максимальных клик, анализа графовых структур, а также методов поиска частых наборов элементов. Использование матриц со-упоминаний позволяет формировать тематические паттерны, отражающие устойчивые связи между терминами в пользовательских предпочтениях. Данный подход ранее исследовался автором в рамках бакалаврской работы, а также в публикации, посвящённой алгоритму поиска паттернов на основе матриц со-упоминаний.

В практической части исследования особое внимание уделяется сравнению разработанного алгоритма с существующими методами анализа текстовых данных. В качестве базовых алгоритмов для сопоставления рассматривались TF-IDF, TextRank, а также модели распределённых представлений текстов. Проведённый анализ показал, что использование графовой модели совместной встречаемости терминов позволяет формировать более устойчивые тематические паттерны, особенно в задачах рекомендаций научных публикаций, где существенную роль играют скрытые взаимосвязи между терминами предметной области.

Разработка программной реализации рекомендательной системы выполнялась с использованием библиотеки Streamlit, предназначенной для создания интерактивных веб-приложений на языке Python. Реализованное приложение обеспечивает взаимодействие пользователя с рекомендательной си-

стемой посредством оценки предлагаемых публикаций. Пользователь может положительно или отрицательно оценивать рекомендуемые статьи, после чего система динамически перестраивает профиль интересов и формирует новые рекомендации.

Важной частью практического исследования являлась оценка качества разработанной рекомендательной системы. Для анализа эффективности использовалась метрика Recall, позволяющая оценить полноту релевантных рекомендаций среди найденных системой публикаций. Дополнительно применялся подход с использованием LLM-агентов, моделирующих поведение различных категорий пользователей. В рамках исследования были сформированы несколько виртуальных пользовательских профилей, соответствующих различным уровням подготовки: бакалавру, магистранту, аспиранту, преподавателю и пользователю без специализированной подготовки.

Результаты проведённых экспериментов показали, что разработанный алгоритм рекомендаций на основе матриц со-упоминаний обеспечивает более устойчивое тематическое соответствие рекомендуемых публикаций пользовательским интересам по сравнению с традиционными статистическими подходами. При этом отмечается, что применение современных эмбеддинговых моделей позволяет получать высокую семантическую близость рекомендаций, однако графовый подход демонстрирует большую интерпретируемость формируемых тематических паттернов.

Таким образом, практическая часть исследования включает полный цикл разработки рекомендательной системы: от подготовки текстовых данных и реализации алгоритмов до экспериментальной оценки качества рекомендаций и анализа пользовательских сценариев взаимодействия с системой.

Заключение

В рамках выполненной магистерской работы были исследованы методы поиска паттернов в текстовых данных и подходы к построению рекомендательных систем для научных публикаций. Основное внимание в работе уделялось разработке алгоритма рекомендаций, основанного на использовании матриц со-упоминаний терминов и графового представления пользовательских интересов.

В теоретической части исследования был проведён анализ существующих подходов к поиску частых наборов и последовательных паттернов. Рассмотрены классические алгоритмы интеллектуального анализа данных, включая Apriori, FP-Growth и PrefixSpan, а также методы графового анализа и вероятностного моделирования. Дополнительно были исследованы современные подходы к построению рекомендательных систем, основанные на статистических моделях, семантических эмбеддингах и методах обработки естественного языка.

В практической части работы была разработана рекомендательная система для подбора научных публикаций на основе пользовательских предпочтений. Реализация системы выполнена с использованием языка программирования Python и современных библиотек обработки текстовых данных. В качестве источника данных использовалась коллекция научных публикаций arXiv, содержащая статьи по различным направлениям компьютерных наук и анализа данных.

В ходе исследования были реализованы и сравнены три подхода к формированию рекомендаций: TF-IDF модель, рекомендательная система на основе Sentence-BERT и предложенный алгоритм на основе матриц со-упоминаний. Проведённое сравнение показало, что статистические методы обеспечивают высокую скорость вычислений и простоту реализации, однако обладают ограниченной способностью учитывать скрытые тематические взаимосвязи между публикациями. Использование эмбеддинговых моделей позволяет учитывать семантическую близость текстов, но снижает интерпретируемость результатов.

Разработанный алгоритм на основе матриц со-упоминаний продемонстрировал возможность формирования устойчивых тематических паттернов

пользовательских интересов за счёт анализа совместной встречаемости терминов. Применение графовой модели позволило повысить интерпретируемость рекомендаций и учитывать структурные связи между ключевыми понятиями предметной области.

Для оценки качества разработанной рекомендательной системы использовались метрики полноты рекомендаций, а также сценарии взаимодействия с виртуальными пользовательскими агентами, моделирующими различные категории пользователей. Полученные результаты показали, что предложенный алгоритм обеспечивает более устойчивое тематическое соответствие рекомендаций пользовательским интересам по сравнению с базовыми статистическими подходами.

Практическая значимость работы заключается в возможности применения разработанного алгоритма в системах интеллектуального анализа научных публикаций, образовательных платформах и специализированных рекомендательных сервисах. Предложенный подход может быть использован для построения персонализированных систем поддержки научного поиска и навигации по большим текстовым коллекциям.