МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»**

Кафедра дифференциальных уравнений и математической экономики

Разработка рекомендательной системы на примере научных текстов

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы

направления 09.03.03 – Прикладная информатика

механико-математического факультета

Володькиной Ольги Николаевны

Научный руководитель
профессор, д.э.н., профессор _____ В.А. Балаш

Заведующий кафедрой
зав. кафедрой, д.ф.-м.н., доцент _____ В.С. Рыхлов

Саратов 2026

Введение. Выпускная квалификационная работа посвящена разработке прототипа рекомендательной системы поиска научных текстов по ключевым словам, принадлежащих заданной тематике.

В отличие от существующих систем, предложенный подход позволяет не только осуществлять поиск, но и проводить анализ ключевых слов и их отбор по новизне.

Актуальность выбранной темы обусловлена тем, что обычный поиск по словам перестает быть эффективным. Он требует точного формулирования запросов и, как правило, не учитывает смысловые связи между работами, семантику текста и контекст исследований. Для решения этой проблемы была разработана концепция рекомендательных систем, которые заменяют классические поисковые инструменты.

В отличие от традиционных поисковых систем, рекомендательные инструменты способны анализировать содержание научных статей, выявлять скрытые взаимосвязи (например, цитирования, общие методологии) и предлагать материалы, которые могут быть интересны исследователю, даже если он ранее о них не знал (если система персонализирована).

Новизна работы заключается в создании рекомендательной системы, основанной на аналитическом подходе, которая обеспечивает комплексный анализ ключевых слов и научных статей, выходя за рамки простого поиска релевантных документов.

Целью выпускной квалификационной работы является разработка рекомендательной системы с поиском и анализом научных статей по компьютерным наукам на основе методов машинного обучения.

Для достижения поставленной цели необходимо выполнить следующие задачи:

1. Изучить основные характеристики рекомендательных систем.
2. Провести разведочный анализ данных.
3. Найти ключевые слова для научных статей.
4. Разработать рекомендательную систему.
5. Разработать поиск статей по названию, авторам и категориям.
6. Разработать комбинированный поиск.
7. Разработать визуальную статистику для научных статей.

8. Разработать систему анализа ключевых слов.
9. Разработать интеграцию с открытыми API с сервисами Crossref и OpenAlex.
10. Реализовать графический интерфейс для рекомендательной системы.

Объектом исследования являются научные статьи по теме «компьютерные науки»(computer science), представленные в базе данных arXiv. Предмет исследования — методы машинного обучения, применяемые для анализа данных, поиска ключевых слов и рекомендации научных текстов.

Работа состоит из введения, трех разделов, заключения, списка использованных источников и двух приложений.

В первом разделе рассмотрены ключевые аспекты, включая определение рекомендательных систем, классификацию их видов, а также анализ специализированных сервисов, предназначенных для поиска и рекомендаций научных публикаций. Особое внимание уделено вопросу выбора и обоснования оптимального вида рекомендательной системы, применимого в рамках работы.

Во втором разделе проводится анализ базы данных arXiv, а также рассматриваются используемые в работе инструменты и методы классификации ключевых слов. Особое внимание уделяется выбору метода извлечения ключевых слов. Кроме того, определяются метрики для оценки работы рекомендательной системы и формируются рекомендации для статьи.

В третьем разделе представлено описание работы рекомендательной системы, реализованной с помощью Streamlit, включающей механизмы поиска, семантического анализа ключевых слов и визуализации данных. В разделе также отражены результаты практического применения системы, что позволяет оценить её эффективность.

В заключении сформулированы основные выводы, полученные в результате выполнения работы.

В приложениях представлен программный код, использующийся для разработки рекомендательной системы на языке Python.

Описание структуры. Работа состоит из введения, трех разделов, заключения, списка использованных источников и двух приложений.

В первом разделе рассмотрены ключевые аспекты, включая определение рекомендательных систем, классификацию их видов, а также анализ специализированных сервисов, предназначенных для поиска и рекомендаций научных публикаций.

Рекомендательная система представляет собой комплекс алгоритмов, программных решений или сервисов, который анализирует предпочтения, историю действий или поведение пользователя, чтобы предсказать и предложить ему наиболее релевантный контент или товары.

На сегодняшний день существует множество рекомендательных систем для научных работ. В рамках данной работы были рассмотрены следующие сервисы: Google Scholar, Microsoft Academic и Connected Papers.

1. Google Scholar — наиболее распространённая бесплатная поисковая система по научным публикациям. В качестве рекомендательного механизма используется функция «похожие статьи», основанная на совпадении ключевых слов, соавторстве и сходстве списков литературы. Система предоставляет базовую аналитику (количество цитирований, индекс Хирша), но не имеет полноценной визуализации связей между работами и не позволяет анализировать временную динамику терминов.
2. Microsoft Academic — высокотехнологичная аналитическая система, чьи наработки легли в основу современных открытых платформ. Её ключевым преимуществом была графовая структура знаний с более чем 540000 научных концепций. Рекомендации формировались на основе глубокого семантического анализа (NLP), а анализ ключевых слов реализовывался через шестиуровневую иерархию, что позволяло учитывать смысловые связи, а не формальное совпадение терминов. Однако система прекратила своё существование в виде веб-интерфейса в 2021 году.
3. Connected Papers — визуальный инструмент, строящий интерактивный граф связей на основе заданной исходной статьи, где узлы — это работы, а рёбра — общие цитирования и сходство текстов. Это позволяет быстро выявлять ключевые публикации («предшественников» и «производные» работы). Основным недостатком является отсутствие персо-

нализации и аналитики временной динамики терминов, а также ограниченность функционала бесплатной версии.

В результате анализа сделан вывод, что ни один из существующих сервисов не предоставляет комплексного решения, объединяющего анализ научных статей, анализ терминов и отслеживание их динамики.

В зависимости от методов анализа данных и формирования рекомендаций выделяют следующие типы рекомендательных систем:

1. Системы, основанные на популярности (popularity-based recommender systems).
2. Системы, основанные на контенте (content-based).
3. Системы коллаборативной фильтрации (collaborative filtering).
4. Гибридные системы, сочетающие различные подходы.
5. Современные системы на основе глубокого обучения, обучения с подкреплением и графовых нейронных сетей.

Наиболее подходящими являются системы, основанные на контенте. Задача сводится к поиску документов, содержание которых наиболее близко к заданному образцу, что позволяет применять контентные методы даже при отсутствии данных о пользователе.

Во втором разделе проводится анализ базы данных, а также рассматриваются используемые в работе инструменты и методы классификации ключевых слов.

В качестве источника данных использован «arXiv». Для разработки системы отобраны только статьи по компьютерным наукам. Исходный набор данных содержал 13 полей, из которых после фильтрации оставлены только значимые для рекомендательной системы: уникальный идентификатор, название, категории, аннотация и авторы.

Код написан на языке Python. Для создания веб-приложения используется Streamlit. Обработка данных выполняется с помощью библиотек pandas и numpy. Для машинного обучения и обработки естественного языка применяются scikit-learn (TfidfVectorizer, NearestNeighbors), а также re для очистки текста. Для стилизации пользовательского интерфейса и обеспечения адаптивного отображения элементов используется CSS (Cascading Style Sheets). Визуализация реализована с помощью matplotlib, seaborn и WordCloud. Для

хранения метаданных используется SQLite3. Интеграция с внешними API (Crossref, OpenAlex) выполнена через библиотеку requests. Вспомогательные функции обеспечивают библиотеки json, datetime, hashlib, collections, os, sys, logging.

Для каждой научной статьи необходимо определить ключевые слова. Это необходимо для разработки системы, которая будет анализировать основные термины, проводить дальнейший анализ статей и выявлять их ключевые аспекты. Для извлечения ключевых слов применяется статистический метод на основе TF-IDF (Term Frequency - Inverse Document Frequency) с улучшениями, такими как адаптивный порог редкости терминов, анализ терминов различной длины, приоритизация терминов из словаря предметной области.

Разработанная система реализует контент-ориентированный подход. Для каждой статьи создаётся объединённое текстовое поле (название и аннотация). Преобразование текста в числовые векторы выполняется с помощью TF-IDF с ограничением в 5000 признаков. Поиск рекомендаций основан на косинусном расстоянии между векторными представлениями статей. Для оценки схожести используется косинусная близость, и рекомендации ранжируются по её убыванию.

Поскольку система не предсказывает рейтинги, метрики RMSE и MAE не применимы.

Для контентных систем ключевыми метриками являются:

1. Точность (precision) — доля релевантных документов среди рекомендованных.
2. Полнота (recall) — доля релевантных документов, найденных системой.
3. F1-мера — гармоническое среднее точности и полноты.

Для учёта порядка рекомендаций используются метрики precision@k, recall@k и F1@k, оценивающие качество первых k элементов списка.

Косинусное сходство: среднее значение 0.3861, медиана 0.3736, интерквартильный размах [0.3279, 0.4310], диапазон [0.1838, 1.0000]. Precision@1 = 0.9011 – первый рекомендованный материал оказывается релевантным в 90% случаев. Precision@k практически не меняется с ростом k, что свидетельствует о равномерной плотности релевантных статей в топ-10. Recall@10 = 1.0 – все релевантные статьи попадают в топ-10. F1@10 = 0.9273 – высокий пока-

затель баланса между точностью и полнотой. Покрытие системы составляет 100%.

В третьем разделе представлено описание работы рекомендательной системы, реализованной в формате веб-приложения с боковой навигационной панелью, обеспечивающей доступ к шести функциональным разделам:

1. Раздел «Начальная страница» отображает ключевые метрики базы данных (общее количество статей, количество категорий в области компьютерных наук, число публикаций за последние годы, среднее количество ключевых слов на статью). На этой странице реализован поиск по названию.
2. Раздел «Поиск статей» предоставляет три вкладки: по авторам, по категориям, комбинированный поиск.
3. Раздел «Рекомендации» предлагает два метода выбора статьи: поиск по названию или ввод идентификатора публикации. После выбора система выводит список семантически близких статей с указанием степени схожести.
4. Раздел «Статистика» содержит метрические карточки, гистограмму распределения публикаций по годам, топ-10 категорий компьютерных наук, а также показатели по извлечённым ключевым словам (общее количество, уникальные ключевые слова, среднее количество на статью, количество терминов).
5. Раздел «Анализ слов» включает:
   (a) новые слова по годам;
   (b) топ слов по годам;
   (c) связанные термины: поиск терминов, совместно встречающихся с заданным ключевым словом, с расчётом процента совместной встречаемости и классификацией силы связи;
   (d) тепловая карта: визуализация частот ключевых слов по годам.
6. Раздел «Поиск в других источниках» — интеграция с внешними научными базами данных Crossref и OpenAlex через их публичные API с возможностью настройки количества результатов (от 1 до 20) и выбора источников.

Реализация рекомендательной системы представлена на рисунке 1.

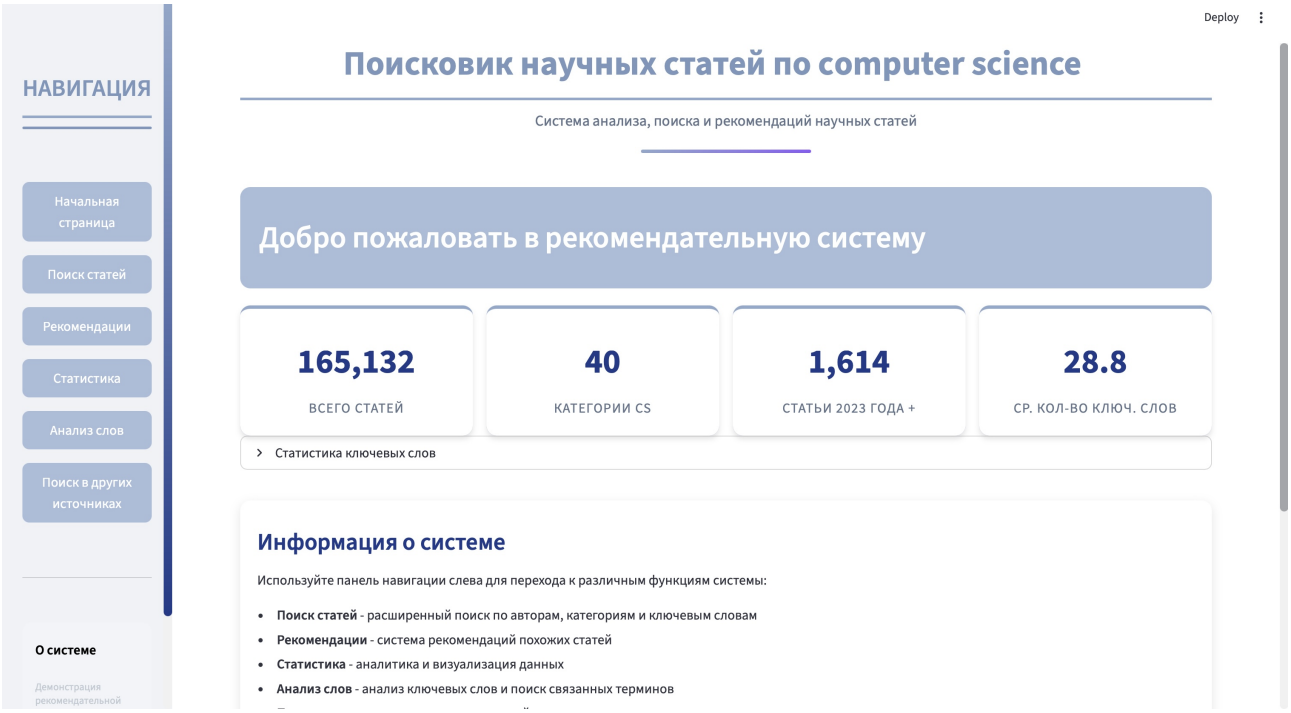

Рисунок 1 — Начальная страница

Вкладка «Рекомендации» представлена на рисунке 2.

Рисунок 2 — Начальная страница вкладки «Рекомендации»

Реализация выдачи рекомендаций для статьи показана в соответствии с рисунком 3.

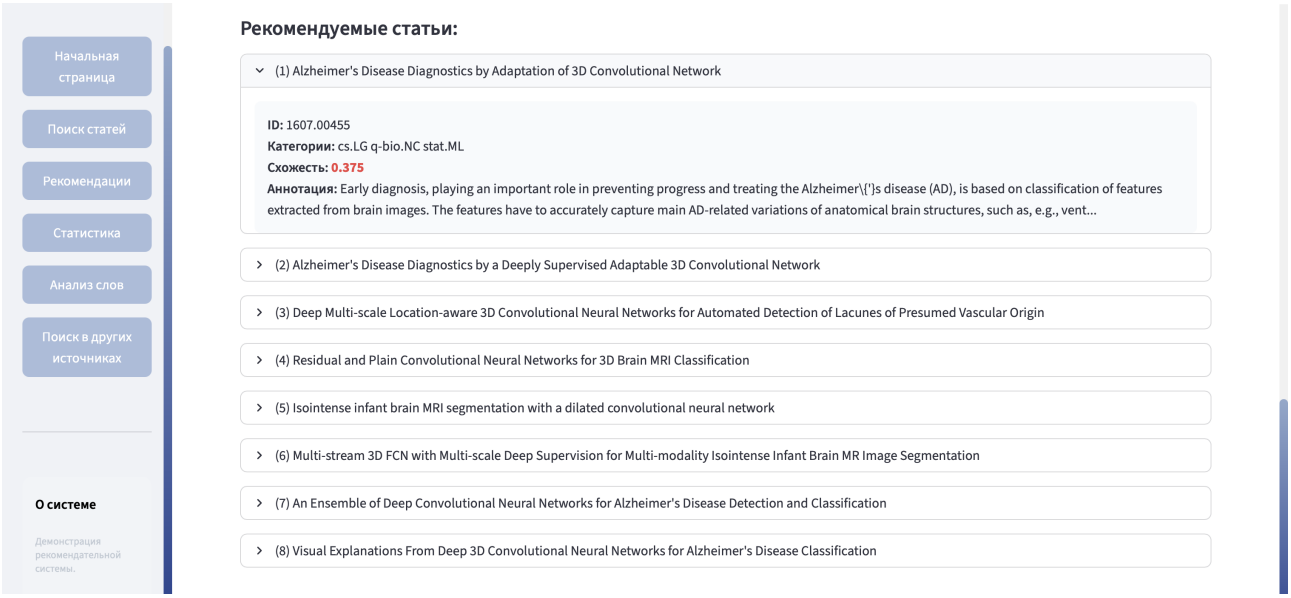

Рисунок 3 — Рекомендации для статьи «3D regression neural network for the quantification of enlarged perivascular spaces in brain mri»

В разделе «Анализ данных» присутствует вкладка «Новые слова по годам», которая позволяет просматривать и регулировать список слов, впервые упомянутых в выбранном году.

Примеры ключевых слов, впервые появившихся в 2011 году, представлены на рисунке 4.

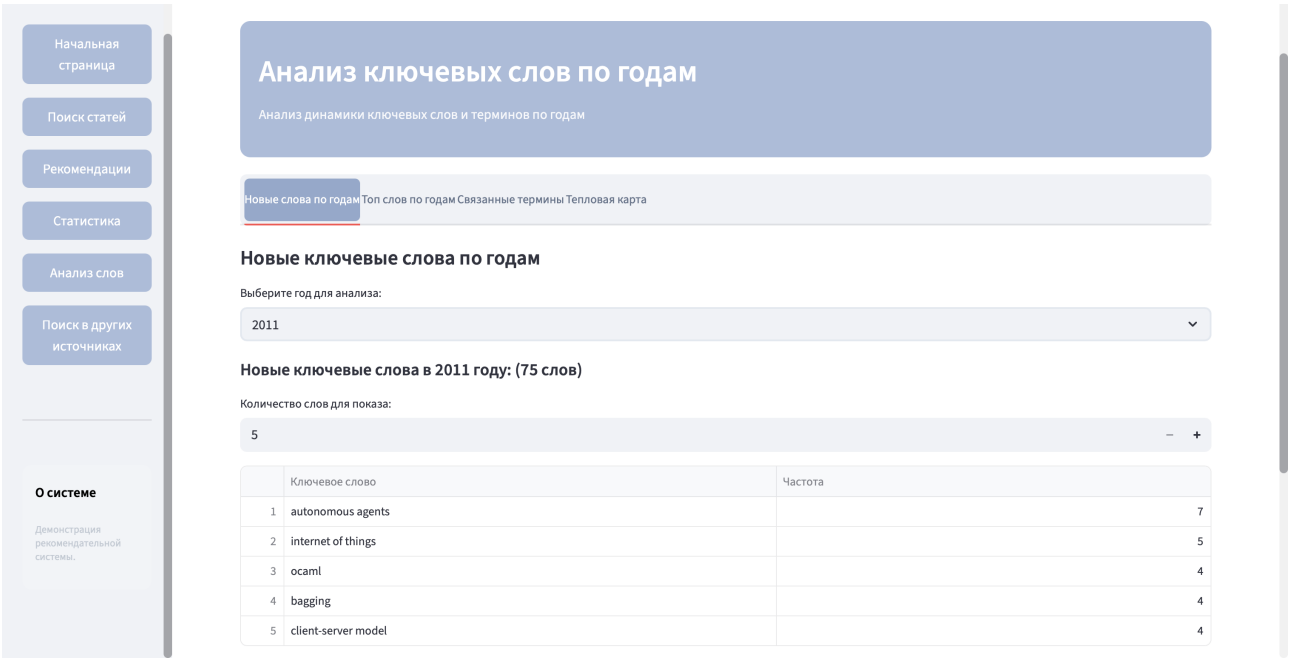

Рисунок 4 — Список ключевых слов, появившихся в 2011 году

В разделе также есть вкладка «Связанные термины». Она позволяет пользователю ввести ключевое слово или выбрать один из наиболее часто встречающихся терминов, чтобы найти связанные с ним понятия. Эти термины ранжируются по значимости: от наиболее релевантных до наименее значимых.

Результат работы программы продемонстрирован на рисунке 5.

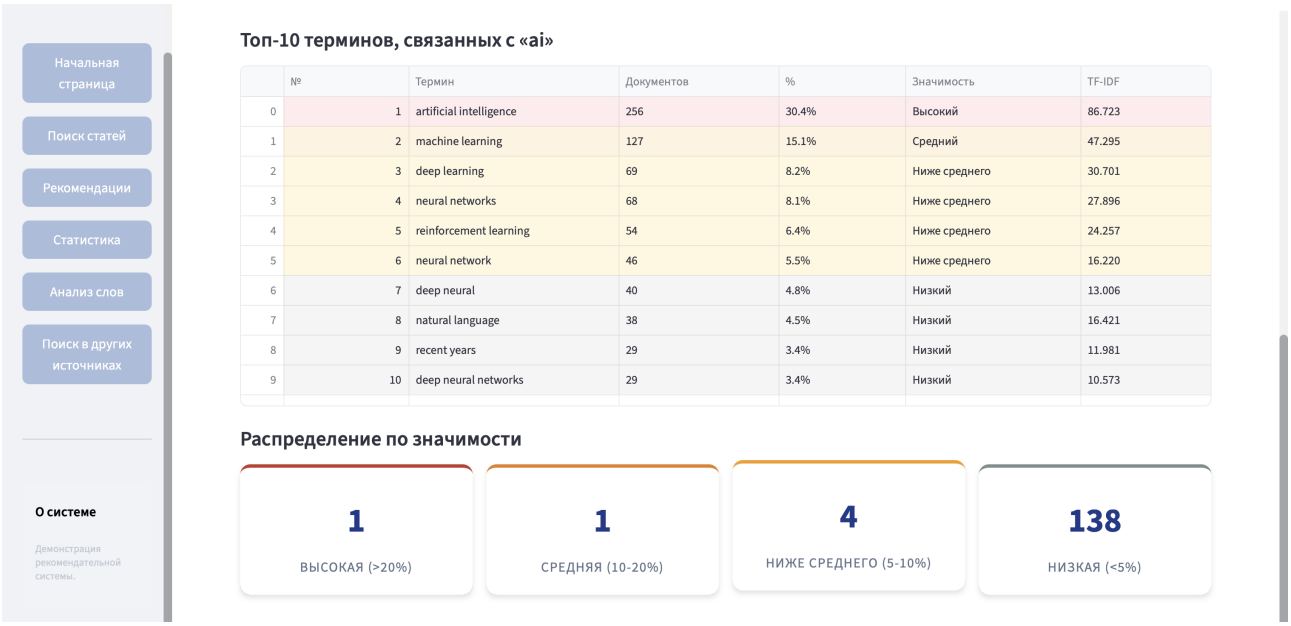

Рисунок 5 — Список связанных терминов для AI

В приложениях представлен программный код, использующийся для разработки рекомендательной системы на языке Python.

Заключение. В ходе выполнения выпускной квалификационной работы была достигнута поставленная цель: разработана рекомендательная система для поиска и анализа научных статей по компьютерным наукам, основанная на методах машинного обучения. Для достижения цели были решены ранее поставленные задачи:

1. Изучены основные характеристики рекомендательных систем и механизмы работы существующих систем (Google Scholar, Microsoft Academic, Connected Papers), выявлены их преимущества и недостатки.
2. Проведен разведочный анализ данных, на основе которого созданы визуализации, отражающие структуру научных публикаций.

3. Реализован механизм выделения ключевых слов на основе TF-IDF, что позволило перейти от лексического поиска к семантическому анализу текстов статей.
4. Разработаны рекомендательная система и её модули, включающие поиск по названию, авторам и категориям, а также комбинированный поиск.
5. Создана система анализа ключевых слов и визуальной статистики, обеспечивающая исследователю инструменты для оценки динамики научных направлений.
6. Реализована интеграция с открытыми API (Crossref, OpenAlex), что позволяет пополнять базу системы актуальными данными.
7. Разработан графический интерфейс, объединяющий все созданные модули в единый удобный инструмент.

Система может быть применена исследователями и студентами, занимающимися научной деятельностью в области компьютерных наук и смежных дисциплин. Система позволяет находить релевантные публикации за счет семантического поиска, а также анализировать эволюцию терминов и ключевых научных направлений, что существенно сокращает время на первичную обработку информации.

Разработанный подход может быть масштабирован на другие предметные области при условии наличия соответствующей выборки и ключевых слов.

Достигнутые результаты свидетельствуют об эффективности предложенного подхода, однако разработанная система имеет значительный потенциал для дальнейшего развития.

Текущая реализация рекомендательного механизма базируется на традиционных подходах, что открывает возможности для внедрения более совершенных архитектур, способных значительно повысить качество персонализации и точность рекомендаций.

В качестве наиболее перспективных направлений разработки планируется применить следующие подходы:

1. Графовые нейронные сети (GNN), потому что в основе анализа связей между научными статьями уже лежит графовое представление (цитирования, авторство, семантическое сходство). Применение GNN позво-

лит моделировать сложные многоуровневые взаимосвязи между публикациями, авторами и тематическими категориями. Графовые нейронные сети способны эффективно обучаться на разреженных данных, распространяя информацию по графу и выявляя скрытые паттерны, которые невозможно обнаружить при изолированном рассмотрении отдельных статей. Их внедрение позволит системе формировать рекомендации на основе анализа не только самой статьи, но и её окружения, например, круга авторов.

2. Нейросетевые методы на основе трансформеров. В отличие от используемого метода TF-IDF, модели на основе BERT (Bidirectional Encoder Representations from Transformers) способны учитывать контекст употребления терминов, понимать многозначность понятий и выявлять смысловые связи, которые могут поначалу показаться неочевидными. Это позволит осуществлять более качественный анализ содержания статей и ключевых слов, а также повысить релевантность рекомендаций.

Следует отметить, что для развития системы необходимо значительно увеличить вычислительные мощности. В отличие от текущей реализации, способной функционировать на ограниченных ресурсах, модели на основе трансформеров и графовых нейронных сетей требуют наличия GPU-вычислений и больших объемов размеченных данных для эффективного обучения.