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

РАЗРАБОТКА АГЕНТНОЙ СИСТЕМЫ НАУЧНОГО ПОИСКА ПО КОРПУСУ ARXIV

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 273 группы
направления 02.04.03 — Математическое обеспечение и администрирование информационных систем
факультета КНиИТ
Чернигина Михаила Андреевича

Научный руководитель
доцент _____ Б. А. Филиппов

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

Саратов 2026

ВВЕДЕНИЕ

Актуальность темы. Научная коммуникация всё в большей степени опирается на открытые цифровые архивы публикаций. Одним из наиболее известных ресурсов такого типа является arXiv, содержащий миллионы препринтов по физике, математике, компьютерным наукам, статистике, биологии, экономике и другим направлениям. Быстрый рост числа публикаций делает задачу поиска релевантных научных работ всё более сложной: исследователю необходимо не только найти документы по ключевым словам, но и сопоставить несколько работ, выделить методы, результаты и ограничения, а также сформировать обоснованный ответ на исследовательский вопрос.

Классические поисковые системы возвращают ранжированный список документов по одному запросу. Такой подход удобен для простых информационных потребностей, но ограничен для сложных научных вопросов: пользователь может не знать точной терминологии, а релевантные публикации могут описывать близкие идеи разными словами. Семантический поиск частично решает эту проблему, но сам по себе также остаётся механизмом поиска по запросу, а не полноценным средством анализа найденных работ.

В последние годы широкое распространение получили большие языковые модели, которые способны анализировать текст, объяснять научные идеи и формировать развёрнутые ответы. Многие существующие веб-интерфейсы взаимодействия с большими языковыми моделями удобны для пользователя, но имеют существенные ограничения в задачах научного поиска: они не всегда имеют доступ к нужному корпусу публикаций, могут неявно использовать непроверенные источники, а также не предоставляют воспроизводимого механизма поиска, отбора и анализа научных работ. Поэтому актуальной становится задача построения агентной системы, в которой языковая модель использует полнотекстовый и семантический поиск как инструменты для работы с научным корпусом.

Цель магистерской работы — разработка агентной системы научного поиска по корпусу arXiv.

В рамках достижения цели система рассматривается как средство поиска и формирования ответов на исследовательские вопросы с использованием полнотекстового и семантического поиска. Качество ответов разработанной системы сравнивается с результатами, получаемыми через существующие веб-

интерфейсы взаимодействия с большими языковыми моделями.

Поставленная цель определила следующие **задачи**:

- исследовать методы полнотекстового, семантического и агентного поиска научных публикаций;
- разработать архитектуру системы, объединяющей поисковую инфраструктуру и агентный слой;
- реализовать или интегрировать подсистему сбора и подготовки корпуса научных публикаций arXiv;
- построить полнотекстовый и семантический поисковые индексы;
- реализовать агентный модуль, способный декомпозировать пользовательский вопрос, выполнять итеративный поиск и анализировать найденные публикации;
- сформировать бенчмарк вопросов по научным публикациям;
- разработать методику экспериментального сравнения качества ответов;
- провести оценку качества ответов по критериям релевантности, фактологической точности, полноты и корректности использования источников.

Объектом исследования являются системы поиска и анализа научных публикаций. **Предметом исследования** являются методы построения гибридного и агентного поиска, объединяющего полнотекстовые индексы, семантические представления документов и большие языковые модели.

Основная **гипотеза работы** состоит в том, что агентная система, использующая полнотекстовый и семантический поиск как инструменты, способна повышать качество ответов на исследовательские вопросы за счёт поиска релевантных источников, декомпозиции запроса и итеративного анализа найденных публикаций.

Методологические основы работы составляют методы информационного поиска и ранжирования документов, семантический поиск на основе векторных представлений, генерация с поисковым дополнением, а также подходы к построению агентных систем на основе больших языковых моделей. Теоретической базой послужили работы К. Маннинга, П. Рагхавана и Х. Шютце, С. Робертсона и Х. Сарагосы, Н. Реймерса и И. Гуревич, П. Льюиса и соавторов, А. Асай и соавторов, а также работы по агентным системам и инструментальному взаимодействию.

Практическая значимость работы состоит в разработке архитектуры и

прототипа исследовательского помощника по научным публикациям. Система не только возвращает найденные документы, но и сохраняет ход агентной сессии: поисковые запросы, вызовы инструментов, найденные источники, использованные фрагменты и снимки контекста. Это делает результат проверяемым и пригодным для анализа ошибок. Дополнительно подготовлен бенчмарк ArXiv-CS ResearchQA, который может использоваться для дальнейшего сравнения систем научного поиска и ответов на исследовательские вопросы.

Структура и объём работы. Магистерская работа состоит из введения, двух разделов, заключения, списка использованных источников и двух приложений. Общий объём работы составляет 74 страницы, включая 10 рисунков и 3 таблицы; список использованных источников содержит 25 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел **«Методы научного поиска и агентных систем»** посвящён анализу предметной области и методов, на которых строится разрабатываемая система.

В начале раздела рассматриваются особенности научных публикаций как источника знаний. Показано, что исследователь часто стремится не просто найти документ по ключевым словам, а восстановить состояние предметной области, сопоставить методы или подтвердить конкретное утверждение. Отдельно отмечается проблема лексического разрыва: одна и та же идея может описываться разными терминами. Затем рассматривается arXiv как корпус научных публикаций: его преимуществами являются устойчивые идентификаторы, доступность метаданных и наличие PDF-версий большинства публикаций, а трудностями — большой объём корпуса и ошибки извлечения текста из PDF.

На основе анализа предметной области сформулированы требования к системе научного поиска. Она должна поддерживать полнотекстовый поиск по названиям, аннотациям и полным текстам публикаций; семантический поиск, позволяющий находить близкие по смыслу работы; проверяемость ответа через сохранение источников и промежуточных шагов; итеративность, необходимую для сложных исследовательских вопросов. Поэтому поисковая инфраструктура в работе рассматривается не только как самостоятельный механизм выдачи документов, но и как инструмент агентной системы.

Далее описаны классические методы информационного поиска: индексация текстовых документов, обратный индекс, ранжирование документов и метрики качества поиска. Показано, что BM25 остаётся сильным и интерпретируемым базовым методом для полнотекстового поиска, но лексический поиск ограничен при различии терминологии запроса и документа. Поэтому отдельно рассмотрен семантический поиск: векторные представления текста, поиск ближайших векторов, приближённые индексы ближайших соседей и гибридный поиск. Сделан вывод, что семантический поиск полезен для обзорных и сравнительных вопросов, но должен дополняться анализом найденных источников.

Завершающая часть первого раздела посвящена агентным системам на основе больших языковых моделей. Рассмотрены агентный цикл, инструментальное взаимодействие, декомпозиция и переформулирование запросов, а также подход с участием пользователя. Для научного поиска агентный цикл позво-

ляет не генерировать ответ сразу, а сначала собрать доказательную базу из найденных публикаций.

В результате первого раздела обоснована необходимость гибридного агентного подхода. Полнотекстовый поиск обеспечивает точные совпадения по терминам, семантический поиск расширяет область поиска за счёт смысловой близости, а агентный слой организует итеративный исследовательский процесс и формирует ответ с опорой на источники.

Второй раздел **«Разработка и оценка агентной системы»** посвящён архитектуре, реализации и экспериментальной проверке разработанной системы.

Система строится как совокупность поисковой инфраструктуры и агентного слоя. Поисковая инфраструктура отвечает за получение публикаций, извлечение текста, хранение метаданных, построение полнотекстового и семантического индексов. Агентный слой использует эти возможности как инструменты: анализирует вопрос пользователя, формирует поисковые запросы, обращается к индексам, выбирает источники и составляет ответ с указанием найденных материалов. В архитектуре выделены три контура: подготовка корпуса, поисковая инфраструктура и агентный слой.

Подсистема сбора и хранения публикаций сохраняет идентификатор arXiv, название, аннотацию, дату публикации, авторов, категории, ссылку на документ и извлечённый текст. Полный текст разбивается на фрагменты, пригодные для поиска и передачи в контекст агента. Полнотекстовый поиск используется для точного поиска терминов, методов, датасетов и аббревиатур; семантический поиск дополняет его в случаях, когда релевантный документ не содержит точных слов из запроса. В системе предусматриваются разные уровни представления: название и аннотация для первичного отбора, фрагменты полного текста — для уточнения ответа.

Агентный слой является надстройкой над поисковой инфраструктурой. Его задача состоит в управлении исследовательским процессом: агент получает вопрос пользователя, выделяет ключевые понятия, формирует один или несколько поисковых запросов, вызывает инструменты поиска, анализирует найденные публикации и решает, нужны ли дополнительные итерации. Среда выполнения агента хранит состояние сессии: исходный вопрос, историю действий, вызовы инструментов, полученные наблюдения, выбранные источники и черновые выводы. Это делает многошаговый поиск проверяемым и воспроизводимым.

При реализации поисковой инфраструктуры использованы локально разворачиваемые компоненты SearXiv. Серверные компоненты реализованы на Rust; для хранения метаданных и состояния задач используется PostgreSQL, для серверного API — Actix Web, для доступа к базе данных — sqlx. Полнотекстовый поиск строится на Tantivy, извлечение текста из PDF выполняется через Poppler. Получение данных из arXiv организовано как обработка задач по датам, что позволяет запускать несколько клиентов параллельно и продолжать обработку после сбоев.

Агентный слой реализуется на основе Curi — локальной среды выполнения исследовательского агента. Он обеспечивает управляемый доступ к инструментам, сохранение сессий, инспекцию контекста и воспроизводимый журнал действий. Взаимодействие с языковыми моделями выполнено через абстракцию поставщика модели. Для агента определён набор предметных инструментов: полнотекстовый, семантический и гибридный поиск, получение карточки публикации, чтение релевантных фрагментов и сохранение выбранного источника. Агент работает только через эти операции, что позволяет связать итоговые утверждения с конкретными источниками.

Пользовательский интерфейс поддерживает командный, веб- и настольный режимы. Во всех режимах пользователь может видеть не только итоговый ответ, но и ход его получения: выполненные запросы, найденные публикации, использованные фрагменты и снимки контекста.

Для экспериментальной оценки был подготовлен бенчмарк ArXiv-CS ResearchQA. Источником данных является arXiv Computer Science за период с 2020 по 2025 год; в итоговую версию вошло 38 публикаций. Бенчмарк содержит 50 вопросов: 17 фактологических, 17 сравнительных и 16 обзорных. Часть вопросов явно называет публикации, которые требуется сопоставить, а часть требует самостоятельно найти подходящие источники по описанию метода, задачи или научной проблемы.

Ответы систем оценивались моделью-судьёй по фиксированной рубрике. На вход оценщику передавались вопрос, тип вопроса, эталонный ответ, подтверждающие фрагменты и ответ тестируемой системы. Использовались четыре критерия: релевантность, фактологическая точность, полнота и корректность использования источников. По каждому критерию выставлялась оценка от 0 до 2 баллов, поэтому максимальный суммарный балл за один ответ составлял 8.

В эксперименте сравнивались три режима. Первый режим — большая языковая модель без внешнего поиска. Он показывает, какого качества можно ожидать от ответа, основанного только на параметрических знаниях модели и формулировке вопроса. Второй режим — существующий веб-интерфейс взаимодействия с большой языковой моделью с включённым веб-поиском. Третий режим — агентная система Curi, получающая доступ к SearXiv через зарегистрированный инструмент поиска.

Режим без внешнего поиска получил средний результат 64,3% от максимальной оценки. Наиболее сильным критерием оказалась релевантность, а наиболее слабым — полнота. Это подтверждает, что модель способна давать связный ответ по теме, но без доступа к текстам публикаций часто не может подтвердить детали и полностью раскрыть вопрос.

Веб-интерфейс с поиском получил 87,8%. По сравнению с режимом без внешнего поиска улучшились все критерии, особенно полнота ответа и использование источников. На фактологических вопросах веб-поиск часто позволял найти конкретную публикацию и извлечь недостающие детали. На обзорных вопросах результат оказался ниже, поскольку внешняя поисковая система иногда находила тематически близкие, но не совпадающие с эталонными источники.

Агентный режим Curi получил 93,5% и показал наибольший средний результат среди трёх режимов. Особенно высокие значения получены на фактологических и сравнительных вопросах: 100,0% и 99,3% соответственно. Для обзорных вопросов средний результат ниже и составляет 80,5%. Полученный результат показывает, что агентный режим особенно полезен тогда, когда нужно найти конкретную публикацию, проверить несколько деталей и собрать ответ из источников.

Для оценки устойчивости различий были рассчитаны доверительные интервалы методом повторной выборки. Парная разность между Curi и веб-интерфейсом составила 5,8 процентного пункта с интервалом [0,5; 11,0] процентного пункта, что указывает на устойчивость наблюдаемого преимущества в рамках выбранной выборки.

Анализ ошибок показал, что наиболее сложными остаются обзорные вопросы, где нужно сопоставить несколько работ и точно определить, какие источники соответствуют формулировке вопроса. Ошибки Curi чаще возникали на этапе выбора или проверки источника: агент мог сформулировать слишком

широкий запрос, выбрать тематически близкую, но неподходящую публикацию или неверно интерпретировать аббревиатуру. Эти случаи показывают направления дальнейшей работы: улучшение проверки соответствия источника вопросу, контроль уровня обобщения и более явное планирование обзорных ответов.

Таким образом, во втором разделе описаны разработанная архитектура, реализационные решения и экспериментальная оценка. Основной самостоятельный результат работы состоит в проектировании и реализации агентного контура поверх поисковой инфраструктуры, подготовке бенчмарка ArXiv-CS ResearchQA и проведении сравнительного эксперимента, показывающего преимущество агентного поиска в выбранной постановке.

ЗАКЛЮЧЕНИЕ

В работе рассмотрена задача научного поиска по корпусу arXiv и разработан подход, в котором классическая поисковая инфраструктура дополняется агентным слоем на основе большой языковой модели. В отличие от обычной поисковой выдачи, такая система не ограничивается возвратом списка документов: агент анализирует вопрос, формирует поисковые запросы, читает найденные материалы и синтезирует итоговый ответ с опорой на источники.

В теоретической части были рассмотрены основные методы научного поиска: полнотекстовый поиск, обратный индекс, ранжирование, семантический поиск, генерация с поисковым дополнением, а также агентные системы, использующие инструменты. Было показано, что для научного поиска важно объединять точный поиск по терминам, работу с семантически близкими формулировками и возможность итеративно уточнять запрос. Эти требования обосновывают архитектуру, в которой поисковая система выступает инструментом агента, а не самостоятельной конечной точкой взаимодействия.

В практической части описана архитектура системы: подсистема подготовки корпуса, поисковая инфраструктура и агентный слой Curi. Поисковая часть хранит метаданные и тексты публикаций, строит полнотекстовые и векторные индексы, поддерживает поиск по названиям, аннотациям и фрагментам полного текста. Агентный слой организует цикл рассуждения и действий, регистрирует инструменты, сохраняет сессии и позволяет анализировать ход поиска после получения ответа.

Для оценки качества был создан бенчмарк ArXiv-CS ResearchQA из 50 вопросов, сгенерированных на основе выбранных публикаций вместе с эталонными ответами. Вопросы были разделены на фактологические, сравнительные и обзорные. Оценка проводилась по четырём критериям: релевантность, фактологическая точность, полнота и корректность использования источников; максимальная оценка за ответ составляла 8 баллов.

Эксперимент показал, что режим без внешнего поиска набрал 64,3% от максимальной оценки. Это подтверждает, что большая языковая модель может формировать связный и релевантный текст, но без доступа к источникам часто не хватает фактических деталей и полноты. Веб-интерфейс с поиском повысил результат до 87,8%: доступ к внешним источникам в выбранной постановке улучшил фактологическую точность, полноту и использование источников.

Агентный режим Curi получил 93,5% и показал наибольший средний результат среди трёх режимов. Особенно высокие значения получены на фактологических и сравнительных вопросах: 100,0% и 99,3% соответственно.

Тем самым экспериментальная проверка дала свидетельства в пользу основной гипотезы работы: агентная система, использующая поиск как инструмент, способна повышать качество ответов на исследовательские вопросы в выбранной постановке. При этом эксперимент также выявил ограничения подхода. На обзорных вопросах результат ниже: 80,5% для Curi. Ошибки чаще связаны не с отсутствием связного ответа, а с выбором тематически близкого, но неподходящего источника, неверным уровнем обобщения или смешением одноимённых терминов из разных областей. Следовательно, дальнейшая работа должна быть направлена на улучшение проверки найденных источников, контроль соответствия источника вопросу и более явное планирование обзорных ответов.

Практическая значимость работы состоит в том, что предложенная архитектура может использоваться как основа для исследовательского помощника по научным публикациям. Система сохраняет не только итоговый ответ, но и промежуточные действия агента, что важно для воспроизводимости научного поиска и последующей проверки результата. При этом вклад локального семантического индекса и вклад внешнего поиска в проведённом прогоне отдельно не изолировались; это остаётся задачей для дальнейшей оценки.

Отдельные части магистерской работы к моменту подготовки автореферата не были опубликованы.

Основные источники информации

1. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008.
2. Robertson S., Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond // Foundations and Trends in Information Retrieval. 2009. Vol. 3, no. 4. P. 333–389.
3. Vaswani A., Shazeer N., Parmar N. et al. Attention is All You Need // Advances in Neural Information Processing Systems. 2017. Vol. 30.
4. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. 2019. P. 4171–4186.
5. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of EMNLP. 2019. P. 3982–3992.
6. Lewis P., Perez E., Piktus A. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 9459–9474.
7. Karpukhin V., Oguz B., Min S. et al. Dense Passage Retrieval for Open-Domain Question Answering // Proceedings of EMNLP. 2020. P. 6769–6781.
8. Yao S., Zhao J., Yu D. et al. ReAct: Synergizing Reasoning and Acting in Language Models // Proceedings of ICLR. 2023.
9. Schick T., Dwivedi-Yu J., Dessì R. et al. Toolformer: Language Models Can Teach Themselves to Use Tools // Advances in Neural Information Processing Systems. 2023. Vol. 36. P. 68539–68551.
10. Asai A., Wu Z., Wang Y. et al. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection // Proceedings of ICLR. 2024.