Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК [004.8: 004.738.1]-047.44

Кулевич
Алексей Олегович

Система искусственного интеллекта для анализа данных тематических сайтов с использованием RAG технологий, LLM и графа знаний

АВТОРЕФЕРАТ
на соискание степени магистра

по специальности 7-06-0612-02 – Информатика и технологии программирования

Научный руководитель
Батура М.П.
доктор тех. наук, профессор

Минск 2026

ВВЕДЕНИЕ

В современном информационном обществе объемы данных, публикуемых на тематических сайтах, растут стремительными темпами. Научные порталы, специализированные ресурсы и образовательные платформы содержат огромные массивы информации, которая требует интеллектуальной обработки для извлечения полезных знаний. Традиционные методы анализа данных часто оказываются недостаточно эффективными для извлечения структурированной информации и формирования релевантных ответов на пользовательские запросы, особенно при работе с разнородными и распределенными источниками данных.

Современные подходы в области искусственного интеллекта, такие как Retrieval-Augmented Generation (RAG) и большие языковые модели (LLM), в сочетании с графовыми базами данных, открывают новые возможности для интеллектуального анализа информации. Использование этих технологий позволяет не только систематизировать и хранить данные в виде графа знаний, но и формировать семантические связи между объектами, обеспечивая контекстный поиск и генерацию осмысленных ответов.

Применение систем на основе RAG и LLM для анализа данных тематических сайтов имеет ряд значительных преимуществ:

– позволяет автоматизировать сбор и обработку больших массивов информации с различных источников, включая научные публикации, блоги и специализированные базы данных;

– обеспечивает построение графа знаний, в котором сохраняются семантические связи между объектами, что упрощает поиск информации и выявление скрытых взаимосвязей;

– интеграция LLM с графовой базой данных позволяет генерировать ответы на запросы пользователей с учетом контекста и структурированных знаний;

– системы легко масштабируются и могут адаптироваться к новым источникам данных, что обеспечивает их эффективность при росте объема информации;

– интеллектуальные алгоритмы способны выявлять скрытые паттерны и закономерности, которые трудно обнаружить с помощью традиционных методов анализа.

Анализ больших объемов информации с тематических сайтов требует эффективных методов обработки, структурирования и представления данных. Целью работы является создание интеллектуальной системы, способной автоматизировать сбор и обработку информации, формировать взаимосвязанные модели знаний и предоставлять пользователю инструменты для получения аналитических результатов.

Разработка такой системы способствует повышению эффективности анализа информации, облегчает доступ к знаниям и открывает новые возможности для исследовательской и практической деятельности в области обработки данных.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Назначение, цель и задачи исследования

Назначением магистерской диссертации является создание системы искусственного интеллекта для анализа данных тематических сайтов с использованием технологий RAG, больших языковых моделей и графа знаний. Система должна обеспечивать автоматизированный сбор данных, их обработку, хранение в графовой базе и предоставление пользователю ответов на основе интеллекАтуального анализа.

Целью магистерской диссертации является разработка системы интеллектуального анализа данных, которая с применением RAG-технологий, LLM и графа знаний позволит формировать тематические графовые базы данных, извлекать и структурировать знания из различных источников, использовать LLM для генерации ответов с учетом семантических связей и контекста, предоставлять пользователю удобный интерфейс для интерактивной работы с системой.

Для достижения поставленной цели необходимо решить следующие задачи:

1 Выполнить анализ предметной области и существующих решений в области анализа данных с применением RAG и LLM.

2 Изучить методы построения графа знаний и интеграции его с технологиями RAG.

3 Определить архитектуру системы и структуру графовой базы данных.

4 Разработать алгоритмы автоматизированного сбора и предобработки данных с тематических сайтов.

5 Исследовать возможности применения библиотек LangChain, LlamaIndex и LangGraph для построения RAG-системы.

6 Сформулировать функциональные требования к системе и интерфейсу пользователя.

7 Разработать модуль интеграции LLM с графовой базой данных и системой RAG.

8 Реализовать продукт для анализа данных и генерации ответов на основе извлеченных знаний.

9 Обеспечить визуализацию данных и результатов анализа в удобной форме.

Объектом исследования являются методы и технологии интеллектуального анализа данных с использованием RAG, LLM и графовых баз знаний.

Предметом исследования является методы и алгоритмы интеграции RAG-технологий, больших языковых моделей и графовых структур для построения систем искусственного интеллекта анализа данных.

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность использования алгоритмов RAG и больших языковых моделей в сочетании с графовыми базами данных для построения интеллектуальной системы анализа данных тематических сайтов, способной

автоматически структурировать знания и генерировать релевантные ответы на пользовательские запросы.

Личный вклад соискателя

В процессе написания диссертации соискатель проанализировал предметную область и существующие решения в области интеллектуального анализа данных, разработал методологию построения RAG-системы на основе инфологической модели и GraphReader-агента, исследовал методы построения графа знаний и интеграции RAG и LLM, разработал и протестировал GraphReader-агента для работы с графовой базой данных и LLM, реализовал программную платформу для анализа данных и генерации ответов, включая веб-интерфейс для пользователей, провёл экспериментальное тестирование системы и оценку её эффективности. Результаты, приведённые в диссертации, получены соискателем лично. Вклад научного руководителя М. П. Батуры, заключается в формулировке целей и задач исследования и консультаций по теме.

Опубликованность результатов диссертации

По теме диссертации опубликовано 5 работ, из них 1 статья в рецензируемом журнале ВАК, 4 работы в сборниках трудов и материалов конференций.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, трех глав, заключения, списка использованных источников, списка публикаций автора и приложений. Первая глава посвящена анализу предметной области, описанию технологий LLM, RAG, графовых баз данных и систем анализа данных. Вторая глава содержит методологию построения инфологической модели графовой базы знаний, разработку GraphReader-агента и построение RAG-системы. Третья глава описывает архитектуру программного обеспечения, реализацию модулей системы и тестирование разработанного решения

Общий объем работы составляет 95 страниц, из которых основного текста – 59 страниц, 40 рисунков на 16 страницах, список использованных источников из 30 наименований на 2 страницах и 1 приложение на 18 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении диссертационной работы определена предметная область исследования, обоснована актуальность выбранной темы, сформулированы цель и задачи работы, а также определены объект и предмет исследования. Обоснована практическая значимость разработки интеллектуальной системы анализа данных, основанной на интеграции технологий Retrieval-Augmented Generation (RAG), больших языковых моделей (LLM) и графовых баз знаний. Определены основные направления применения результатов исследования в научной, образовательной и корпоративной среде.

В первой главе выполнен обзор предметной области и теоретических основ построения интеллектуальных систем анализа данных. Рассмотрены современные методы обработки естественного языка (NLP), а также принципы функционирования и применения больших языковых моделей для извлечения, обобщения и интерпретации знаний. Проанализированы архитектурные подходы к построению систем Retrieval-Augmented Generation, обеспечивающих интеграцию генеративных моделей с внешними источниками информации. Особое внимание уделено графовым базам знаний, включая платформу Neo4j, и их роли в структурировании, хранении и семантическом анализе данных.

В главе также рассмотрены embedding-модели как инструмент векторного представления текстовой информации, обеспечивающий поиск по смысловому сходству. Проведён сравнительный анализ существующих решений в области интеллектуального анализа данных и обоснован выбор гибридной архитектуры, сочетающей LLM, графовые структуры и retrieval-механизмы. По результатам анализа показана целесообразность использования комплексного подхода для повышения точности, объяснимости и адаптивности интеллектуальных систем.

Во второй главе разработана авторская методология построения интеллектуальных систем анализа данных, основанная на интеграции RAG-технологий, графа знаний и больших языковых моделей. Представлен метод построения инфологической модели графовой базы знаний, включающей основные сущности предметной области, типы связей и атрибуты, отражающие структуру анализируемых данных.

Предложен метод построения и функционирования GraphReader-агента, обеспечивающего интеллектуальное взаимодействие языковой модели с графом знаний. Агент реализует итеративный механизм анализа, включающий планирование стратегии поиска, последовательный обход графа и выбор релевантных узлов в соответствии с контекстом пользовательского запроса. Такой подход позволяет формировать интерпретируемые и контекстно обоснованные ответы, а также снижать вероятность ошибок и генерации недостоверной информации.

В главе также представлена схема интеграции RAG-подхода с инфологической моделью графа знаний и описан процесс обработки пользовательских запросов, включающий этапы формирования контекста, анализа связей между сущностями и генерации итогового ответа.

Разработанная методология обеспечивает модульность, масштабируемость и возможность адаптации системы к новым предметным областям.

В третьей главе представлена практическая реализация предложенной методологии и архитектуры интеллектуальной системы анализа данных ScienceChat. Разработан программный модуль работы с графовой базой данных на платформе Neo4j, обеспечивающий семантический поиск и структурированный доступ к данным. Реализован GraphReader-агент с использованием библиотеки LangGraph, организующий взаимодействие между языковой моделью и графом знаний.

Подробно описаны этапы функционирования GraphReader-агента, включая формирование плана анализа, определение релевантных узлов графа, последовательное чтение и анализ текстовых фрагментов, извлечение и проверку атомарных фактов, расширение контекста и финальный синтез ответа. Также описан процесс интеграции RAG-компонента с большими языковыми моделями с использованием библиотек LangChain и LlamaIndex, что позволило объединить преимущества структурированного поиска и генеративного анализа текста. Проведено тестирование системы на тематических данных, показавшее эффективность предложенного решения по показателям релевантности, полноты и объяснимости результатов.

В результате проведённого исследования создана полнофункциональная интеллектуальная система анализа данных тематических сайтов ScienceChat, обеспечивающая:

– автоматизированный сбор и структурирование информации в виде графа знаний;

– интеллектуальное взаимодействие между большими языковыми моделями и графовой базой данных посредством GraphReader-агента;

– семантический поиск и генерацию контекстно обоснованных ответов;

– модульную и масштабируемую архитектуру с возможностью адаптации к различным предметным областям.

В заключении подведены итоги работы, сформулированы основные научные и практические результаты, подтверждающие достижение поставленной цели. Показано, что разработанная методология и реализованная на её основе система интеллектуального анализа данных позволяют повысить эффективность обработки, структурирования и интерпретации информации.

Результаты исследования могут быть использованы при разработке аналитических систем, рекомендательных сервисов, систем мониторинга информационных потоков, а также корпоративных систем поддержки принятия решений.

ЗАКЛЮЧЕНИЕ

В данной диссертационной работе разработан и теоретически обоснована методология построения интеллектуальной системы анализа данных тематических сайтов с использованием технологий Retrieval-Augmented Generation (RAG), больших языковых моделей (LLM) и графовых баз знаний. Работа носит комплексный характер: в ней проведён как анализ современных методов обработки текстовой информации, так и практическая реализация предложенного подхода, ориентированного на повышение точности, прозрачности и адаптивности интеллектуальных систем.

В первой главе проведён системный анализ ключевых технологий, лежащих в основе построения современных RAG-систем. Подробно рассмотрены принципы функционирования больших языковых моделей, их роль в извлечении, обобщении и интерпретации знаний, а также особенности их интеграции с внешними источниками данных. Проанализированы методы обработки естественного языка (NLP), позволяющие эффективно извлекать смысловые зависимости и контекстные связи, что является критически важным при анализе сложных тематических текстов. Отдельное внимание уделено графовым базам знаний, в частности Neo4j, и их преимуществам в структурировании данных и поддержке семантического поиска. Также выполнен анализ embedding-моделей как инструмент векторного представления текста, обеспечивающий релевантность и масштабируемость поиска. На основе проведённого анализа показано, что именно комбинация LLM, графовых структур и retrieval-механизмов формирует основу интеллектуальных систем нового поколения, способных не только извлекать информацию, но и рассуждать о ней.

Во второй главе предложена Методология построения Retrieval-Augmented Generation (RAG) - системы с использованием GraphReader-агента и инфологической модели основаной на интеграции трёх взаимосвязанных компонентов: графовой базы данных, формализованной инфологической модели, GraphReader-агента и генеративной языковой модели (LLM). Предложен метод построения инфологической модели графовой базы знаний, определяющей структуру хранения и представления информации в системе и метод построения и функционирования GraphReader-агента – ключевого компонента, обеспечивающего взаимодействие между языковой моделью и графом знаний. В отличие от прямого поиска, агент использует стратегию итеративного планирования и осмысленного обхода графа: он формирует план действий, определяет релевантные узлы, последовательно анализирует их и принимает решение о завершении поиска или его расширении. Такая организация позволяет адаптировать поведение системы под различные типы запросов – от простого поиска фактов до комплексных объяснительных вопросов. В работе показано, что использование графовых структур повышает точность поиска, сокращает количество нерелевантных фрагментов и снижает вероятность генерации неточных ответов языковой моделью.

В третьей главе представлена реализация интеллектуальной системы ScienceChat архитектура, которой разработана на основе предложенной методологии. Разработан модуль работы с графовой базой данных на базе Neo4j, обеспечивающий хранение, обновление, семантический поиск и структурированный доступ к данным. Реализован GraphReader-агент с использованием LangGraph, а также тщательно проработаны все этапы его функционирования: планирование стратегии поиска, чтение текстовых фрагментов, выбор соседних узлов, принятие решений о завершении поиска и финальный синтез ответа. Отдельное внимание уделено построению системы промптов, управляющих поведением агента на каждом этапе. Интеграция с LLM через RAG-подход позволила объединить точность структурированного поиска и гибкость генеративной модели, что значительно расширяет возможности системы по обработке сложных информационных запросов. Приведены примеры, демонстрирующие работу созданной интеллектуальной системы ScienceChat.

В ходе выполнения диссертационной работы все поставленные задачи были реализованы. Был проведён комплексный анализ предметной области и определены оптимальные принципы интеграции RAG, LLM и графовых структур, разработана инфологическая модель графовой базы знаний, обеспечивающая хранение и взаимосвязь тематических данных, реализован GraphReader-агент, позволяющий осуществлять интеллектуальный семантический поиск и взаимодействие между LLM и графом знаний, создана архитектура программной системы, включающая модули сбора, анализа и генерации ответов на основе извлечённых знаний.

Результатом работы стала архитектура интеллектуальной системы ScienceChat, способная анализировать текстовую информацию, адаптироваться к различным сценариям и формировать понятные ответы. Архитектура модульная: возможно добавление новых узлов, обновление промптов или изменение стратегий обхода графа без переобучения, что обеспечивает устойчивость системы к изменениям предметной области.

Практическая значимость методологии заключается в широком спектре применения: корпоративные интеллектуальные системы, платформы поддержки решений, системы мониторинга информации и контент-аналитические сервисы. Подход обеспечивает прозрачность логики получения ответа и повышенную достоверность за счёт интеграции структурированных знаний и генеративных возможностей LLM.

Таким образом, в диссертации предложена и реализована комплексная методология построения интеллектуальной системы анализа данных с использованием RAG, LLM и графа знаний. Проведено как теоретическое обоснование архитектуры, так и практическая реализация, подтверждающая её эффективность и гибкость. Полученные результаты формируют основу для дальнейших исследований в области гибридных архитектур искусственного интеллекта, а также могут служить базой для разработки прикладных решений анализа больших объёмов неструктурированных данных, мониторинга информационных потоков и интеллектуальной поддержки принятия решений.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1-А. Кулевич, А. О. Система комплексного анализа данных тематических сайтов ИСКАД ИИ = System of complex data analysis of thematic sites ISCAD IS / И. И. Пилецкий, М. П. Батура, Н. А. Волорова [и др.] // Информатика. – 2024. − Т. 21, № 1. – С. 105–120.

2-А. Кулевич, А. О. Методология построения системы анализа данных с использованием RAG технологий и графовой базы данных / А. О. Кулевич // Компьютерные системы и сети : сборник статей 60-й научной конференции аспирантов, магистрантов и студентов, Минск, 22–26 апреля 2024 г. / Белорусский государственный университет информатики и радиоэлектроники. – Минск, 2024. – С. 573–574.

3-А. Кулевич, А. О. Архитектурное решение построения RAG-системы анализа данных с использованием библиотеки LangChain и графовой базы данных Neo4j / М. П. Батура, А. О. Кулевич // Информационные технологии и системы 2024 (ИТС 2024) = Information Technologies and Systems 2024 (ITS 2024) : материалы международной научной конференции, Минск, 20 ноября 2024 г. / Белорусский государственный университет информатики и радиоэлектроники ; редкол. : Л. Ю. Шилин [и др.]. – Минск, 2024. – С. 47–48.

4-А. Кулевич, А. О. Реализация GraphReader-агента для улучшения технологии RAG с использованием Neo4j, LangChain и LangGraph / А. О. Кулевич // Компьютерные системы и сети : материалы 61-й научной конференции аспирантов, магистрантов и студентов, Минск, 22–26 апреля 2025 г. / Белорусский государственный университет информатики и радиоэлектроники. – Минск, 2025. – С. 461–463.

5-А. Кулевич, А. О. Система искусственного интеллекта анализа данных тематических сайтов с использованием RAG технологий, LLM и графа знаний / М. П. Батура, А. О. Кулевич // Информационные технологии и системы 2025 (ИТС 2025) : материалы Международной научной конференции, Минск, 19 ноября 2025 / Белорусский государственный университет информатики и радиоэлектроники ; редкол.: Л. Ю. Шилин [и др.]. – Минск, 2025. – С. 37–38.