

Разработка метода построения графов знаний по данным слабоструктурированных источников для семантического поиска в специализированных областях знаний.

Специальность: 1.2.1. — Искусственный интеллект и машинное обучение

Аспирант:

Ларин А.

Руководитель:

Жукова Н.А, д.т.н., доцент

Санкт-Петербург, 2026

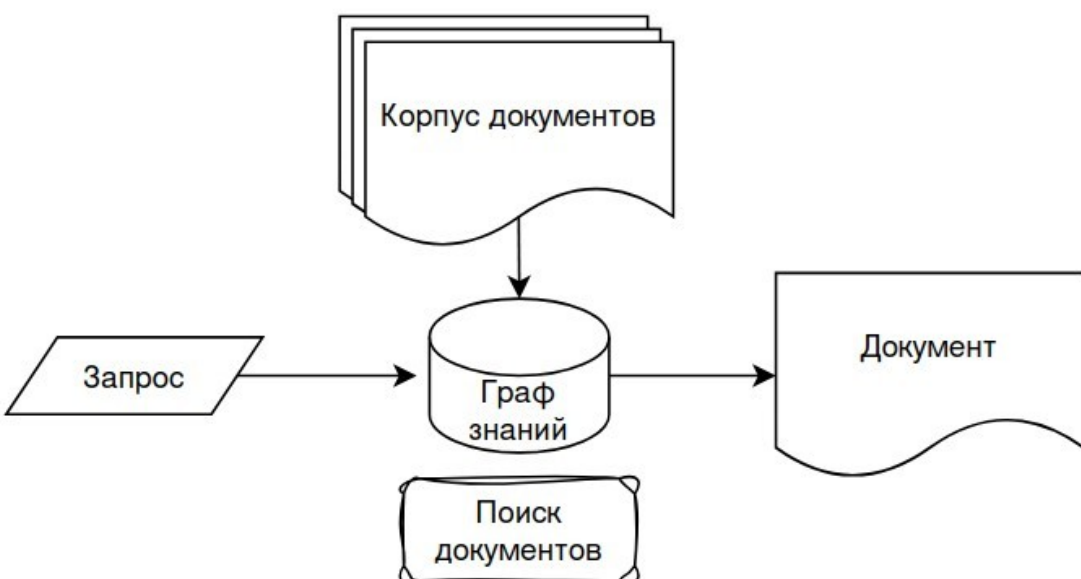
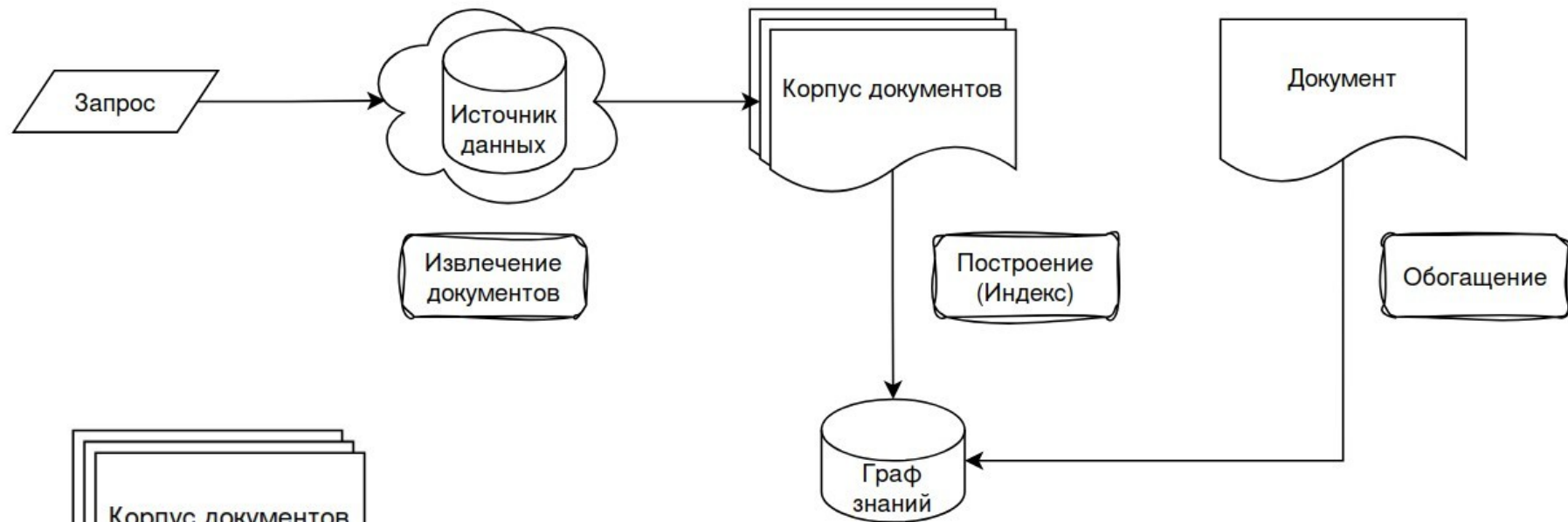
Актуальность

- Обзоры литературы требуют больших временных затрат и ручного анализа.
- Существующие поисковые системы оперируют документами, зачастую неструктурированными
- Последние исследования показывают, что SKG перспективны. Но методы наполнения и извлечение недоисследованы
- **Новизна:**
 - Переход от документов к графам знаний
 - Разработка методов наполнения, обновления, извлечения
 - Расширение на кроссдомен

Цель и задачи

Задачи:

- Проведение обзора
- Определение источников данных
- Эксперименты по построению онтологий
- Эксперименты по методам наполнения графа
- Задача поиска. Метрика сходства на графе



Обзор

- Kaplan A. и др. A Classification Framework for Scientific Documents to Support Knowledge Graph Population. Gesellschaft für Informatik, Bonn, 2025.
- Zhang Z. и др. A Graph Based Document Retrieval Method // 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD)). 2018.
- Ryen V., Soylu A., Roman D. Building semantic knowledge graphs from (semi-) structured data: a review // Future Internet. MDPI, 2022. T. 14, № 5. С. 129.
- Khalid S. и др. Comprehensive review of academic search systems: evolution, analysis, and future research directions // Soc. Netw. Anal. Min. 2025. T. 15, № 1. С. 66.
- Meng Z. и др. Domain Ontology-Driven Knowledge Graph Generation from Text // ACM Trans. Probab. Mach. Learn. 2025. T. 1, № 4. С. 21:1-21:18.

Данные

- Trec Covid - датасет статей по covid для IR
- BrIR - сборник различных датасетов для IR
- ACL-OCL - датасет статей основанный на acl-anthology с полнотекстами
- Paperswithcode-data - аннотированная база papers with code
- ORKG – нет датасета, но есть API

Ontology learning

- Identify the domain-relevant terms
 - A raw list of candidate terms
- Extract taxonomic relationships
 - Terms hierarchy
- Non-taxonomic relation extraction
 - (Subject, Predicate, Object) triples
- Manual review
- Bulk ontology learning
- Alignment and refinement
-

paper:hasResearchProblem X

paper:hasMethod X

paper:hasModel X

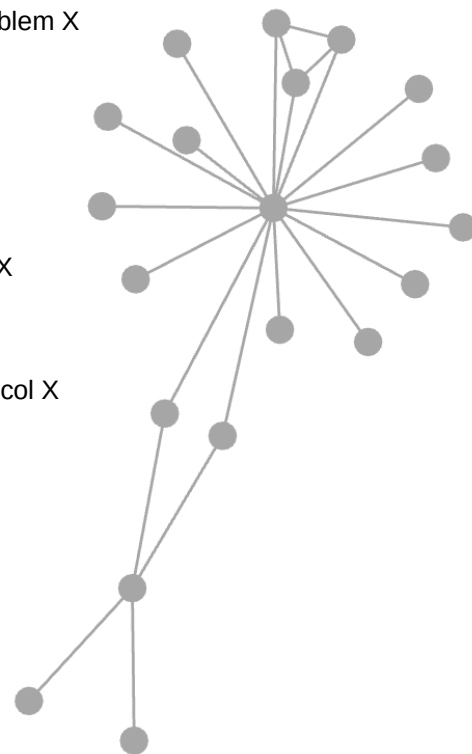
paper:hasDataset X

paper:usesFramework X

paper:hasEvaluation X

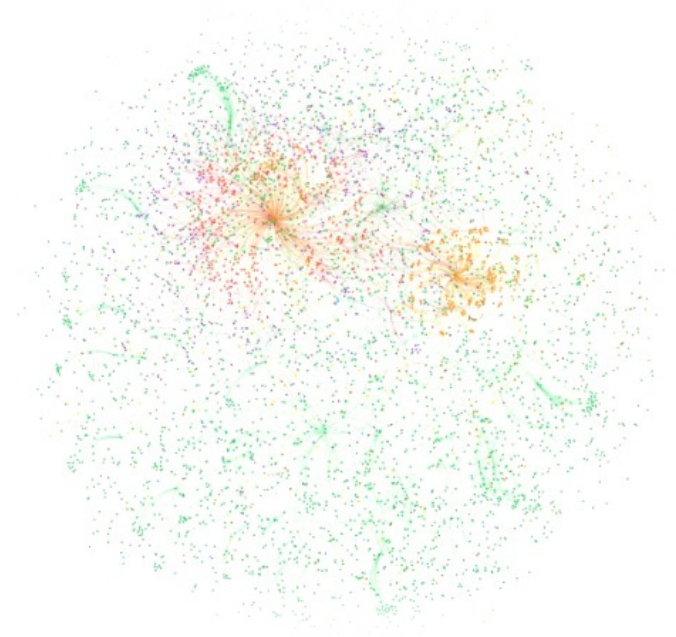
paper:hasTrainingProtocol X

paper:hasHardware X



Knowledge graph population

- Guided extraction with prompting
 - given ontology extract instance-level data
- LLM-based triple extraction
 - set of candidate triples
- Grounding (entity and relation linking)
 - Link entity in the KG with identifier
- Manual review



Результаты

- Выступление на конференции, публикация тезисов
 - Конгресс молодых учёных, ИТМО

Шиян А.С., Ларин А. (науч. рук. Жукова Н.А.) Повышение точности распознавания болезней растений компактной vlm с помощью базы знаний // Сборник тезисов докладов конгресса молодых ученых. Электронное издание. – СПб: Университет ИТМО, [2026]. URL: <https://kmu.itmo.ru/digests/article/16507>

Написана статья (ещё не подана):

Shiyan A. S., Larin A., Baimuratov I., Zhukova N. Knowledge-Augmented Hybrid Classification Framework for Unseen Compositions.

Вышла статья*

Shiyan A. S., Larin A., Baimuratov I. R., Zhukova N. A. Methodology for Automatic Information Extraction and Fusion for the Ontology of Plant Diseases // PRIA Journal Special Issue. XXVII International Conference "Analytics and Data Management in Data-Intensive Domains" (DAMDID/RCDL-2025), October 29–31, 2025, St. Petersburg, Russian Federation. Selected Papers. Section Papers—Data Representation and Modeling. Knowledge Bases. 2025. DOI: 10.1134/S1054661826700355.

Проблемы

- Качественные аннотированные данные
- Универсальные бенчмарки
- Масштабирование на большие объёмы (шумные графы)

Планы

- Провести эксперименты
- Интеграция с ORKG
- Подбор бенчмарков
- Методы ретрива
- Публикация