УДК 004.8:004.738.5

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА КОРПОРАТИВНОГО ПОИСКА С ИСПОЛЬЗОВАНИЕМ ГИБРИДНОГО ПОДХОДА

Тимаева А. А.
Научный руководитель – канд. техн. наук, научный сотрудник Шматков В. Н.
Университет ИТМО
alina.timayeva97@mail.ru

Введение

В условиях цифровой трансформации стремительный рост объёма неструктурированных корпоративных данных обусловливает повышенные требования к эффективности поисковых инструментов. Традиционный поиск работает по принципу точного совпадения слов и не понимает смысла запроса. Ситуация изменилась с развитием NLP и появлением трансформерных моделей вроде BERT и SBERT [1]. Семантический поиск через векторные представления текста научился улавливать смысл. Но и он не идеален – уступает лексическому поиску, когда речь идет о конкретных фактах, аббревиатурах или цифрах, где важно строгое совпадение. Анализ современных решений, как российских, так и зарубежных, показывает, что всё больше внимания привлекают гибридные подходы. В отечественных работах, например у Наумова М.Д. и Кореевой Е.Б., подтверждается, что комбинация классических методов (вроде TF-IDF) с нейросетевыми эмбеддингами даёт хорошие результаты при оценке смысловой близости текстов [2]. В международном контексте эта логика развивается в сторону универсальных моделей вроде SPLADE (Formal T. и соавторы), которые совмещают разреженное лексическое представление с расширением запросов [3]. Исследователи сходятся в том, что эффективны не просто способы объединения ранжированных списков, но и дообучение моделей для переранжирования результатов. Однако многие коммерческие и opensource-продукты до сих пор остаются либо строго лексическими, либо строго векторными, а попытки их совместить часто требуют сложной настройки. Целью данной работы является разработка готовой к внедрению интеллектуальной системы корпоративного поиска, в которой за счет гибридизации удастся добиться значимого роста качества и релевантности выдачи.

Основная часть

Предлагаемое решение представляет собой гибридную поисковую систему, в которой семантический и лексический подходы не просто сменяют друг друга, а работают согласованно, на одном конвейере обработки запроса. При индексации система обрабатывает документы разных форматов (PDF, DOCX, TXT) и разбивает их на фрагменты. Разбиение адаптивное: сохраняются смысловые границы, используется перекрытие, чтобы не терять контекст. Для каждого фрагмента параллельно строятся два представления: векторное – с помощью легковесной нейросетевой модели rubert-tiny2, и лексическое – для поиска по BM25. Когда приходит запрос, оба движка запускаются одновременно. Векторный находит фрагменты, близкие по смыслу. Лексический отдаёт приоритет точным совпадениям терминов. Два ранжированных списка затем объединяются, оценки релевантности нормализуются и суммируются с весами. Эмпирика подсказывает, что эффективно работает соотношение 0,6 в пользу семантики и 0,4 – в пользу лексики. Чтобы повысить качество верхних позиций выдачи, мы применяем двухэтапную постобработку. Первый этап – переранжирование по принципу Learning-to-Rank, но в упрощённом варианте. Финальная оценка фрагмента корректируется с учётом

дополнительных признаков: места в документе (введение или заключение дают больше веса), типа текстового блока (заголовок или основной текст), даты публикации и того, насколько полно запрос покрыт лексически. Второй этап – диверсификация. Чтобы выдача не состояла из десятка фрагментов одного и того же документа, алгоритм ограничивает их количество. Так пользователь сразу видит информацию из разных источников, а не упирается в один. С архитектурной точки зрения система построена на современных, экономичных и производительных технологиях. Ядро системы реализовано на Python с использованием асинхронного фреймворка FastAPI. Это обеспечивает высокую пропускную способность. В качестве единого хранилища данных и векторов применяется PostgreSQL с расширением pgvector, который упрощает развертывание и сопровождение по сравнению со связкой из отдельной векторной и документной баз данных. Предложенная архитектура минимизирует эксплуатационные затраты и обеспечивает линейную масштабируемость за счет контейнеризации (Docker) и балансировки нагрузки.

Выводы

Разработанная система поиска заметно выигрывает у базовых решений. На реальном корпусе корпоративных документов точность на первых десяти позициях оказалась на 35-40% выше, чем при использовании одного лишь векторного поиска. По фактологическим запросам прирост достигает 40-60%, по смысловым – 30-50%. При этом система выдерживает до 50 запросов в секунду при среднем времени отклика 120-150 миллисекунд. Практическая ценность системы заключается в ее готовности к внедрению в корпоративную ИТ-инфраструктуру. Она предоставляет современный веб-интерфейс и реализует ролевую модель доступа, обеспечивая безопасность данных. Система может быть успешно внедрена в организациях любого масштаба для организации поиска по внутренним базам знаний, технической документации, архивам отчетов и регламентов. Предложения по внедрению включают пилотную интеграцию в одном из департаментов организации для оценки реального эффекта на производительность сотрудников, последующее тиражирование на всю компанию и настройку под конкретные доменные области, такие как обучение специализированных эмбеддинг-моделей на корпоративной лексике. Дальнейшие направления развития связаны с интеграцией больших языковых моделей (LLM) для генерации кратких ответов на основе найденных фрагментов и реализации полностью адаптивного гибридного ранжирования, динамически подбирающего веса методов в зависимости от типа и сложности входящего запроса.

Литература

1. Devlin J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). – Minneapolis, 2019. – P. 4171–4186. – URL: https://aclanthology.org/N19-1423.pdf (дата обращения: 04.02.2026).
2. Наумов М. Д., Кореева Е. Б. Гибридный подход к оценке смысловой близости текстовых данных с использованием tf-idf и нейросетевых эмбеддингов // Вестник науки. 2026. №1 (94). URL: https://cyberleninka.ru/article/n/gibridnyy-podhod-k-otsenke-smyslovoy-blizosti-tekstovyh-dannyh-s-ispolzovaniem-tf-idf-i-neyrosetevyh-embeddingov (дата обращения: 04.02.2026).
3. Formal T. et al. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking // Proceedings of SIGIR. 2021. P. 2288-2292. – URL: https://arxiv.org/abs/2107.05720 (дата обращения: 04.02.2026).