**Научная статья**
**УДК 004.678:004.421:378**

# ПРИМЕНЕНИЕ LLM В СИСТЕМЕ ИНФОРМАЦИОННОГО ПОИСКА НА БАЗЕ ЖУРНАЛА «ЭКОНОМИКА СЕЛЬСКОГО ХОЗЯЙСТВА РОССИИ»

**Солопенко Виктория Владимировна[1]**

**Научный руководитель – Демичев Вадим Владимирович[2]**



**[1,2] ФГБОУ ВО Российский государственный аграрный университет – МСХА имени К.А. Тимирязева; Россия, Москва**
**[1]студентка 4 курса института экономики и управления АПК, solopenko2005@yandex.ru**
**[2]к.э.н, доцент,demichev_v@rgau-msha.ru**

***Аннотация**: В статье рассматривается интеграция больших языковых моделей (LLM) в задачи информационного поиска как закономерный этап эволюции поисковых систем. Проведен сравнительный анализ классических алгоритмов ранжирования TF-IDF и BM25, представлена их программная реализация с оптимизациями (нормализация релевантности, динамическая фильтрация стоп-слов). На основе данных, полученных посредством индексации официального журнала «Экономика сельского хозяйства России», состоящего из 28032 страниц (16668 уникальных лемм), выявлены ограничения лексического подхода: высокая скорость обработки запроса (300 мс) достигается ценой «слепоты» к семантическим связям и контекстуальным синонимам. Систематизированы основные архитектурные паттерны интеграции LLM: плотный поиск, гибридный поиск и двухэтапный поиск с реранкингом. Для каждого паттерна проанализированы преимущества и ограничения в контексте создания локальной академической поисковой системы с поддержкой русского языка. Обоснована целесообразность применения гибридного подхода с настраиваемым коэффициентом балансировки между лексической и семантической компонентами, а также перспективы использования fine-tuned LLM для узкоспециализированных предметных областей.*

***Ключевые слова**: информационный поиск, большие языковые модели, LLM, TF-IDF, BM25, плотный поиск, гибридный поиск, реранкинг, семантический поиск, лемматизация, академическая поисковая система, русскоязычная морфология.*

***Для цитирования:** Солопенко Виктория Владимировна ПРИМЕНЕНИЕ LLM В СИСТЕМЕ ИНФОРМАЦИОННОГО ПОИСКА НА БАЗЕ ЖУРНАЛА «ЭКОНОМИКА СЕЛЬСКОГО ХОЗЯЙСТВА РОССИИ» / Солопенко Виктория Владимировна // Агрофорсайт. 2026. № 3— Саратов: ООО «ЦеСАин»,2026. – 1 электрон. опт. диск (CD-ROM). – Загл. с этикетки диска.*

# APPLICATION OF LARGE LANGUAGE MODELS (LLMs) IN AN INFORMATION RETRIEVAL SYSTEM BASED ON THE JOURNAL “ECONOMICS OF AGRICULTURE IN RUSSIA”

**Victoria V. Solopenko[1]**

**Supervisor: Vadim V. Demichev[2]**

**[1,2] Russian State Agrarian University – Moscow Timiryazev Agricultural Academy; Moscow, Russia**
**[1] Fourth-year student of the Institute of Economics and Management in the Agro-Industrial Complex, solopenko2005@yandex.ru**
**[2] Candidate of Economic Sciences, Associate Professor, demichev_v@rgau-msha.ru**

**Abstract:** The article examines the integration of large language models (LLMs) into information retrieval tasks as a natural stage in the evolution of search systems. A comparative analysis of classical ranking algorithms TF-IDF and BM25 is conducted, and their software implementation with optimisations is presented (relevance normalisation, dynamic stop-word filtering). Based on data obtained through indexing the official journal *“Economics of Agriculture in Russia”*, which contains 28 032 pages (16 668 unique lemmas), the limitations of the lexical approach are identified: the high query processing speed (300 ms) comes at the cost of “blindness” to semantic relationships and contextual synonyms. The main

architectural patterns for LLM integration are systematised: dense retrieval, hybrid search, and two-stage search with re-ranking. For each pattern, the advantages and limitations are analysed in the context of developing a local academic search system with Russian language support. The feasibility of applying a hybrid approach with a configurable balancing coefficient between lexical and semantic components is justified, along with the prospects of using fine-tuned LLMs for highly specialised subject domains.

**Keywords:** information retrieval, large language models, LLM, TF-IDF, BM25, dense retrieval, hybrid search, re-ranking, semantic search, lemmatisation, academic search system, Russian morphology.



Основной задачей поисковых систем была и остается удовлетворение информационных потребностей человека. Эта парадигма изначально основывалась на принципах точного совпадения терминов: алгоритмы BM25 и TF-IDF повсеместно использовались в поисковых системах, стремясь через индексацию и релевантность выдать слова, соответствующие запросу пользователя. В современных же реалиях данный подход является лишь основой для формирования более продвинутых поисковых движков, где активно внедряются большие языковые модели (Large Language Model).

Данная статья посвящена интеграции LLM, анализу основных архитектурных паттернов, вызовам, с которыми сталкиваются разработчики, а также адаптации понятия «релевантности» в эпоху генеративного ИИ [1].

**Классические алгоритмы поиска TF-IDF и BM25**

*TF-IDF (Team Frequency – Inverse Document Frequency)* – самый популярный метод определения веса слов в документе относительно коллекции, используется как базовый эталонный алгоритм в поисковых системах.

TF – отношение числа вхождений некоторого слова к общему количеству лемм документа. Таким образом оценивается важность слова $t_i$ в пределах конкретного документа (1).

$$tf(t,d) = \frac{n_t}{\sum_k n_k}, \quad (1)$$

где nt – число вхождений некоего слова t в документ,

$\sum_k n_k$ – общее число слов в данном документе.

IDF – инверсия частоты, с которой некоторое число встречается в документах коллекции. Учет IDF уменьшает вес слов, которые широко упоминаются в пределах отдельной коллекции документов (2).

$$idf(t,D) = \frac{|D|}{|\{d_i \in D \mid t \in d_i\}|}, \quad (2)$$

где

$|D|$ - число документов в коллекции;

$|\{d_i \in D \mid t \in d_i\}|$ - число документов из коллекции D, в которых встречается t (когда $n_t \neq 0$)

Алгоритм TF-IDF отличается простотой интеграции в поисковые системы, высокой скорости вычисления, интегрируемостью. Однако, при использовании на практике данного метода появляется «эффект насыщения», связанный с линейностью TF-IDF.

BM25 (Best Matching 25) – семейство функций ранжирования, разработанное в рамках проекта Okapi, считается современным классиком и стандартом индустрии поисковиков: используется в Lucene, Elasticsearch, PostgreSQL.

Данный алгоритм представляет собой современные TF-IDF-подобные функции ранжирования, которые широко используются на практике в поисковых системах.

Пусть дан запрос Q, содержащий слова $q_1, \ldots, q_n$, тогда функция BM25 дает следующую оценку релевантности документа D запросу Q (приведена упрощенная формула расчета релевантности (3)):

$$score(D,Q) = \sum_{i=1}^{n} idf(q_i) * \frac{f(q_i,D)*(k_1+1)}{f(q_i,D)+k_1*(1-b+b\frac{|D|}{avgdl})}, \quad (3)$$

где

$f(q_i, D)$ есть частота слова $q_i$ в документе D;
|D| есть длина документа (количество слов в нем);
avgdl – средняя длина документа в коллекции;
$k_1$ и b – свободные коэффициенты, обычно их выбирают как $k_1$ = 2.0 и b = 0.75.



На основе вышеописанных алгоритмов был создан поисковый движок для поиска научной информации в академической среде. Расчет релевантности был произведен с оптимизацией метода TF-IDF.

Основные вычисления ранга леммы (неизменяемой части слова) производятся в Листинге 1:

**Листинг 1 – Основной алгоритм TF-IDF с нормализацией**

```
for (Page page : pages) {
    double relevance = 0;
    for (Lemma lemma : lemmas) {
        Float rank = indexRepository.findRankByPageAndLemma(page, lemma);
        if (rank != null) {
            relevance += rank; // <- Суммирование рангов (TF * IDF)
        }
    }
    relevanceMap.put(page, relevance);
}
```

Далее после подсчета «сырой релевантности» применяется нормализация относительно максимума (Листинг 2).

**Листинг 2 – Нормализация релевантности**

```
if (maxRelevance > 0) {
    for (Map.Entry<Page, Double> entry : relevanceMap.entrySet()) {
        entry.setValue(entry.getValue() / maxRelevance);
    }
}
```

Это классический подход в информационном поиске: самая релевантная страница получает 1.0, остальные же получают значения от 0 до 1 пропорционально их релевантности.

Оптимизация базовых принципов расчета ранга слов представлена в Листинге 3.

**Листинг 3 – Фильтрация редких/частых слов**

```
double lemmaFrequencyRatio = (double) lemmaObj.getFrequency() / totalPages;
if (lemmaFrequencyRatio <= threshold) {// threshold = 1.0
    filteredLemmas.add(lemmaObj); // Оставляем редкие и средние слова
} else {
    // Исключаем слишком частые слова (стоп-слова по частоте)
}
```

Слова, встречающиеся на каждой странице (отношение частоты > 1.0), бесполезны для поиска. Представлена динамическая фильтрация стоп-слов на основе реальных данных [2].

**Ограничения классического подхода: анализ на основе экспериментальных данных**

Представленные выше алгоритмы TF-IDF и BM25, а также их программная реализация (Листинги 1-3) демонстрируют зрелость и практическую применимость классического подхода к информационному поиску. Нормализация релевантности относительно максимума и динамическая фильтрация стоп-слов на основе частотного анализа позволяют добиться приемлемого качества ранжирования в большинстве стандартных сценариев.

Однако при ближайшем рассмотрении выявляются фундаментальные ограничения, заложенные в самой парадигме лексического совпадения. Даже с учетом морфологического анализа (лемматизации) и эвристик, подобных фильтрации частотных слов (Листинг 3), классические методы остаются «слепыми» к семантическому контексту. Они оперируют словами как дискретными единицами, но не способны уловить смысловые связи между ними [3].

Для количественной оценки этих ограничений был проведен эксперимент на базе разрабатываемой поисковой системы, ориентированной на академическую среду. В качестве полигона исследований выступил журнал «Экономики сельского хозяйства России» (архив номеров за 2022–2025 гг., общим объёмом 2 380 статей), что позволило обеспечить репрезентативность предметной выборки и достаточную глубину для тестирования как лексических, так и семантических алгоритмов поиска.

После этапа индексации и лемматизации общий объем обработанных данных составил 16668 уникальных лемм. Эксперимент ставил целью не только продемонстрировать «слепые зоны» классического подхода, но и установить количественные метрики его работы, которые в дальнейшем послужат базой для сравнения с методами, использующими большие языковые модели [4].

Лемматизация запроса (первые 10 строк таблицы с леммами) представлена в Таблице 1 при классическом методе индексации коллекций.

**Таблица 1 - Лемматизация запроса (индексируемый вид для классической системы)**

| id<br>integer | site_id<br>integer | lemma<br>character varying (255) | frequency<br>integer |
|---|---|---|---|
| **1** | 1 | апк | 3223 |
| **2** | 1 | прогнозирование | 1597 |
| **3** | 1 | производство | 4550 |
| **4** | 1 | исследование | 1836 |
| **5** | 1 | экономический | 3925 |
| **6** | 1 | журнал | 13934 |
| **7** | 1 | публикация | 13946 |
| **8** | 1 | хозяйство | 14937 |
| **9** | 1 | сельскохозяйственный | 4763 |
| **10** | 1 | эффективность | 3384 |

Как демонстрируют данные Таблицы 1, классический подход к индексации, основанный на алгоритмах BM25 и TF-IDF, успешно решает задачу частотного анализа: леммы «журнал» (частота 13934) и «публикация» (частота 13946) ожидаемо доминируют в выборке, что характерно для академической направленности коллекции документов, где значительную долю составляют периодические издания в области агропромышленного комплекса.

Однако, именно данный результат и выявил ключевое ограничение лексического подхода: при различных запросах семантически релевантный запрос не будет найден, поскольку использует иной лексический вокабуляр [5].



Экспериментальное измерение времени обработки запроса классическим движком составило 300 мс. Данный показатель является эталонным с точки зрения производительности и достигается благодаря прямой работе с инвертированным индексом, содержащим 16668 уникальных лемм. Однако цена этой скорости – концептуальная «слепота» к семантическим связям и контекстуальным синонимам.

**Архитектурные паттерны интеграции LLM**

Выявленное противоречие между высокой скоростью лексического поиска и его семантической ограниченностью формирует запрос на новые архитектурные решения. Современные поисковые системы всё чаще обращаются к гибридным архитектурам, где классические алгоритмы (BM25, TF-IDF) выполняют роль первичного фильтра, а большие языковые модели обеспечивают смысловую глубину ранжирования [6].

В зависимости от требований ко времени отклика, вычислительным ресурсам и точности поиска сложилось три основных архитектурных паттерна интеграции LLM: плотный поиск, гибридный поиск и двухэтапный поиск с реранкингом. Рассмотрим каждый из них в контексте задачи создания академической поисковой системы, работающей с русскоязычными текстами.

Плотный поиск

В основе этого подхода лежит переход от разреженных векторных представлений к плотным эмбеддингам. Документы и запросы проецируются в единое векторное пространство с помощью модели-энкодера (например, семейство BERT-моделей). Релевантность определяется как близость векторов (косинусное расстояние), что позволяет находить семантических соседей, даже если они не содержат точных терминов запроса [7].

Для русскоязычных академических систем перспективно использование моделей типа ruBERT или RuRoBERTa, которые могут быть дообучены на корпусе научных текстов. Однако, несмотря на высокую семантическую точность, плотный поиск предъявляет повышенные требования к вычислительным ресурсам и требует эффективных алгоритмов приближенного поиска ближайших соседей для обеспечения приемлемой скорости работы.

Гибридный поиск

На практике наиболее актуальным решением часто становится комбинация лексического и семантического подходов – гибридный поиск. Результирующая релевантность вычисляется как взвешенная сумма оценок (4).

$$Score = \alpha * \text{BM25\_score} + (1 - \alpha) * \text{Dense\_score}, \quad (4)$$

где α – настраиваемый коэффициент, позволяющий балансировать между точным совпадением терминов и смысловым сходством.

Данный паттерн особенно уместен в системах с уже реализованным классическим поисковым ядром, где лемматизированный индекс служит естественной основой для лексической компоненты [8].

Варьируя коэффициент α, можно адаптировать систему под специфику коллекции: для точных наук (где важна терминологическая строгость) увеличить вес BM25, для гуманитарных дисциплин – усилить семантическую составляющую.



## Двухэтапный поиск с реранкингом

Данный паттерн решает проблему вычислительной сложности больших языковых моделей.

Архитектура строится в два этапа:

1. **Поиск –** быстрый метод (BM25 или dense retrieval) отбирает ограниченное множество кандидатов (обычно top-50–100).

2. **Переранжирование – б**олее тяжелая, но точная модель (Cross-Encoder) попарно оценивает каждый кандидат, выдавая финальную ранжировку.

Cross-encoder, в отличие от bi-encoder, обрабатывает пару (запрос, документ) совместно, что позволяет модели уловить тонкие нюансы их взаимосвязи, но делает невозможным предварительное индексирование. Это ограничение нивелируется первым этапом, ограничивающим количество обрабатываемых пар [9].

Для академической поисковой системы, ориентированной на специфику агропромышленных журналов, реранкинг открывает широкие перспективы: именно на этом этапе может применяться fine-tuned LLM под конкретную предметную область (например, в области агроинженерии, растениеводства, экономики АПК), существенно повышая точность выдачи для узкоспециализированных запросов без необходимости перестраивать всю поисковую инфраструктуру.

Таким образом, выбор конкретного архитектурного паттерна должен определяться требованиями к точности поиска, доступным вычислительным ресурсам и спецификой предметной области. Для разрабатываемой локальной поисковой системы академической направленности наиболее целесообразным представляется внедрение гибридного подхода с последующей имплементацией реранкинга для решения узкоспециализированных задач, что полностью соответствует дорожной карте проекта и позволит обеспечить баланс между скоростью работы и семантической точностью поиска [10].

**Список литературы:**

1. Google Академия [Электронный ресурс] // Google. — URL: https://scholar.google.com.ru/ (дата обращения: 10.03.2026).
2. Elasticsearch vs Sphinx [Электронный ресурс] // Хабр. — URL: https://habr.com/ru/articles/590181/ (дата обращения: 15.03.2026).
3. Рыжов, С. С. Применение эффективных алгоритмов оптимизации поисковых движков / С. С. Рыжов // INTERNATIONAL INNOVATION RESEARCH : сб. ст. X Междунар. науч.-практ. конф., Пенза, 07 августа 2017 г. — Пенза : Наука и Просвещение (ИП Гуляев Г. Ю.), 2017. — С. 108–110. — EDN ZCDDGZ.
4. Сохина, С. А. Алгоритмы выполнения поисковых запросов в сети интернет. Ранжирование результатов / С. А. Сохина, С. А. Немченко // Программная инженерия: современные тенденции развития и применения (ПИ-2021) : сб. материалов V Всерос. науч.-практ. конф., Курск, 15–16 марта 2021 г. — Курск : Юго-Западный государственный университет, 2021. — С. 135–138. — EDN JMVWLE.

5. Уколова, А. В. Разработка информационной системы учёта и обработки данных с поддержкой проведения статистического анализа на C++ / А. В. Уколова, Д. В. Быков // Цифровые технологии анализа данных в сельском хозяйстве. — Москва : Научный консультант, 2022. — С. 61–127. — EDN FQKUBE.

6. Миронов, А. И. Создание частичного индексирования таблицы для оптимизации поисковых запросов / А. И. Миронов, В. И. Мунерман // Современные информационные технологии и ИТ-образование. — 2022. — Т. 18, № 3. — С. 558–565.

7. Куприяновский, В. П. Умная инфраструктура, физические и информационные активы, Smart Cities, BIM, GIS и IoT / В. П. Куприяновский, В. В. Аленьков, И. А. Соколов [и др.] // International Journal of Open Information Technologies. — 2017. — Т. 5, № 10. — С. 55–86. — EDN ZISODV.

8. Азарова, А. Н. Научный обзор исследований в области системного подхода / А. Н. Азарова, И. И. Шанин // Актуальные направления научных исследований XXI века: теория и практика. — 2020. — Т. 8, № 3. — С. 84–89.

9. Агеев, А. И. Искусственный интеллект и экономика: возможности и вызовы / А. И. Агеев, И. В. Ильин. — М. : ИНФРА-М.

10. Ловина, В. В. Исследование средств оптимизации системы продвижения сайтов / В. В. Ловина // Проблемы современной науки и образования. — 2016. — № 18 (60). — С. 33–35.

**Referenes**

1.Google Scholar. (n.d.). Retrieved March 10, 2026, from https://scholar.google.com.ru/

2.Habr. (n.d.). *Elasticsearch vs Sphinx*. Retrieved March 15, 2026, from https://habr.com/ru/articles/590181/

3.Ryzhov, S. S. (2017). Application of effective algorithms for optimising search engines [Primenenie effektivnykh algoritmov optimizatsii poiskovykh dvizhkov]. In *INTERNATIONAL INNOVATION RESEARCH: Proceedings of the 10th International Scientific and Practical Conference* (pp. 108–110). Penza: Nauka i Prosveshchenie (IP Gulyaev G. Yu.). EDN ZCDDGZ.

4.Sokhina, S. A., & Nemchenko, S. A. (2021). Algorithms for executing internet search queries. Ranking results [Algoritmy vypolneniya poiskovykh zaprosov v seti internet. Ranzhirovanie rezul'tatov]. In *Software Engineering: Modern Trends in Development and Applications (PI-2021): Proceedings of the 5th All-Russian Scientific and Practical Conference* (pp. 135–138). Kursk: Southwestern State University. EDN JMVWLE.

5.Ukolova, A. V., & Bykov, D. V. (2022). Development of an information system for data accounting and processing with support for statistical analysis in C++ [Razrabotka informatsionnoy sistemy ucheta i obrabotki dannykh s podderzhkoy provedeniya statisticheskogo analiza na C++]. In *Digital Data Analysis Technologies in Agriculture* (pp. 61–127). Moscow: Nauchnyy Konsul'tant. EDN FQKUBE.

6.Mironov, A. I., & Munerman, V. I. (2022). Creating partial table indexing to optimise search queries [Sozdanie chastichnogo indeksirovaniya tablitsy dlya optimizatsii poiskovykh zaprosov]. *Modern Information Technologies and IT Education*, *18*(3), 558–565.

7.Kupriyanovsky, V. P., Alenkov, V. V., Sokolov, I. A., et al. (2017). Smart infrastructure, physical and information assets, Smart Cities, BIM, GIS and IoT [Umnaya infrastruktura, fizicheskie i informatsionnye aktivy, Smart Cities, BIM, GIS i IoT]. *International Journal of Open Information Technologies*, *5*(10), 55–86. EDN ZISODV.

8.Azarova, A. N., & Shanin, I. I. (2020). Scientific review of research in the field of systems approach [Nauchnyy obzor issledovaniy v oblasti sistemnogo podkhoda]. *Actual Directions of Modern Scientific Research: Theory and Practice*, *8*(3), 84–89.

9.Ageev, A. I., & Ilyin, I. V. (n.d.). *Artificial intelligence and the economy: Opportunities and challenges*. Moscow: INFRA-M.

10. Lovina, V. V. (2016). Research on optimisation tools for website promotion systems [Issledovanie sredstv optimizatsii sistemy prodvizheniya saytov]. *Problems of Modern Science and Education*, *18*(60), 33–35.