

УДК 004.891.2,004.032.26

DOI: 10.18413/2518-1092-2026-11-3-0-4

Левина Т.М.
Беляков Я.В.

ОПТИМИЗАЦИЯ ГИБРИДНОГО АЛГОРИТМА ПОИСКА В ТЕКСТОВЫХ КОРПУСАХ НА ОСНОВЕ МНОГОУРОВНЕВЫХ МЕТРИК РЕЛЕВАНТНОСТИ

Институт нефтепереработки и нефтехимии ФГБОУ ВО УГНТУ в г. Салавате,
ул. Губкина, 22Б, г. Салават, 453250, Россия

e-mail: belyakov.yaroslav@bk.ru

Аннотация

Актуальность исследования обусловлена экспоненциальным ростом объемов неструктурированных текстовых данных и возрастающими требованиями к скорости и релевантности их обработки в таких прикладных областях, как анализ новостных потоков, интеллектуальный поиск и обработка естественного языка (NLP). Существующие алгоритмы поиска часто сталкиваются с проблемами низкой точности из-за морфологической сложности языков (например, русского), наличия большого количества шумовых слов и необходимости учета семантической близости. Цель работы – разработка и исследование гибридного алгоритма поиска, сочетающего метрики лексического соответствия, морфологического анализа (стеминг) и взвешенного ранжирования для повышения полноты и точности поиска в текстовых корпусах. Научная новизна заключается в предложенной многоуровневой метрике релевантности, которая интегрирует подсчет точных совпадений, анализ основ слов с использованием оптимизированного стеминга для русского языка и временные характеристики документов. Методология включает разработку алгоритма, его программную реализацию и сравнительный анализ эффективности на реальном наборе данных, а именно новостей. В ходе эксперимента подтверждено, что гибридный подход позволяет достичь более высокой точности (на 15-20%) по сравнению с базовыми методами, особенно при работе с разносклоняемыми формами слов и составными запросами. Выводы работы указывают на перспективность дальнейшего развития предложенного подхода за счет интеграции семантических векторных моделей и глубокого обучения для учета контекстной близости понятий.

Ключевые слова: информационный поиск; стеминг; морфологический анализ; метрики релевантности; ранжирование; алгоритм оптимизации; обработка естественного языка

Для цитирования: Левина Т.М., Беляков Я.В. Оптимизация гибридного алгоритма поиска в текстовых корпусах на основе многоуровневых метрик релевантности // Научный результат. Информационные технологии. – Т.11, №3, 2026. – С. 42-48. DOI: 10.18413/2518-1092-2026-11-3-0-4

Levina T.M.
Belyakov Ya.V.

OPTIMIZATION OF A HYBRID SEARCH ALGORITHM IN TEXT CORPORA BASED ON MULTI-LEVEL RELEVANCE METRICS

Institute of Oil Refining and Petrochemistry
Federal State Budgetary Educational Institution of Higher Education USPTU in Salavat,
22B Gubkin St., Salavat, 453250, Russia

e-mail: belyakov.yaroslav@bk.ru

Abstract

The relevance of this research stems from the exponential growth of unstructured text data and the increasing demands on the speed and relevance of its processing in such applied areas as news analysis, intelligent search, and natural language processing (NLP). Existing search algorithms often face problems with low accuracy due to the morphological complexity of languages (e.g., Russian), the presence of a large number of noise words, and the need to account for semantic similarity. The goal of this study is to develop and evaluate a hybrid search algorithm that combines lexical match metrics, morphological analysis (stemming), and weighted ranking to improve the completeness and accuracy of searches in text corpora. The scientific innovation lies in the

proposed multi-level relevance metric, which integrates exact match counting, word stemming analysis using optimized stemming for the Russian language, and document temporal characteristics. The methodology includes algorithm development, its software implementation, and a comparative analysis of its effectiveness on a real-world dataset, namely, news. The experiment confirmed that the hybrid approach achieves 15-20% higher accuracy than basic methods, especially when working with inflected word forms and compound queries. The study's findings suggest potential for further development of the proposed approach by integrating semantic vector models and deep learning to account for the contextual proximity of concepts.

Keywords: information retrieval; stemming; morphological analysis; relevance metrics; ranking; optimization algorithm; natural language processing

For citation: Levina T.M., Belyakov Ya.V. Optimization of a Hybrid Search Algorithm in Text Corpora Based on Multi-Level Relevance Metrics // Research result. Information technologies. – Т.11, №3, 2026. – P. 42-48. DOI: 10.18413/2518-1092-2026-11-3-0-4

ВВЕДЕНИЕ

Современная цифровая эпоха характеризуется лавинообразным ростом объемов текстовой информации. Новостные агрегаторы, архивы документов, базы знаний и системы поддержки принятия решений требуют эффективных механизмов для быстрого и точного поиска необходимых сведений. Классические методы поиска, основанные на точном лексическом соответствии (keyword search), демонстрируют недостаточную эффективность при работе с естественными языками, обладающими богатой морфологией, таким как русский язык. Проблема усугубляется наличием синонимии, омонимии и необходимостью фильтрации нерелевантной информации (стоп-слов).

Эволюция технологий поиска прошла путь от простого инвертированного индекса до сложных систем, учитывающих семантику и контекст. Однако, несмотря на успехи глубокого обучения, гибридные подходы, комбинирующие лингвистические методы и статистические метрики, остаются наиболее практичным и эффективным решением для многих прикладных задач, особенно в условиях ограниченных вычислительных ресурсов или специализированных предметных областей.

Основная проблема заключается в разработке алгоритма, который не только находит документы, содержащие ключевые слова запроса, но и ранжирует их по степени релевантности, учитывая морфологические вариации, значимость терминов и актуальность документа во времени.

ОСНОВНАЯ ЧАСТЬ

Проблема информационного поиска традиционно занимает центральное место в компьютерной лингвистике и науке о данных. Классические модели, включая булеву, векторную и вероятностную модель BM25, подробно исследованы в фундаментальных трудах по информационному поиску [4, 5]. Следует отметить, что модель BM25 продолжает оставаться эталоном в области полнотекстового поиска, хотя ее эффективность существенно зависит от качества токенизации и лемматизации входных текстов.

Для языков с развитой морфологической системой, к которым относится русский язык, этап приведения слов к нормальной форме представляется особенно значимым. Алгоритм стемминга Портера для русского языка [6] длительное время оставался доминирующим методом, несмотря на потенциально избыточное или недостаточное отсечение. Современные подходы демонстрируют тенденцию к использованию морфологических анализаторов, таких как MyStem [7], и методов машинного обучения [3], что, однако, сопровождается возрастанием вычислительной сложности.

Проблема ранжирования результатов поиска находит широкое освещение в научной литературе. Алгоритм PageRank [8] заложил фундамент для ссылочного ранжирования в веб-пространстве. В контексте полнотекстового поиска активно применяются метрики, основанные на частоте терминов и обратной частоте документа [9]. Важно подчеркнуть, что классический TF-IDF не учитывает позиционные характеристики терминов, их взаимную близость и временные параметры документов.

В последние годы наблюдается растущий интерес к гибридным системам, интегрирующим символьные и нейросетевые подходы [10]. Трансформерные архитектуры и модели типа BERT [11] демонстрируют впечатляющие результаты в области семантического анализа, однако их практическое применение для поиска по крупным документальным коллекциям остается ограниченным вследствие высокой вычислительной стоимости. В этой связи актуальной научной задачей становится разработка облегченных гибридных методов, использующих глубокое обучение для тонкой настройки релевантности поверх результатов, полученных более эффективными лингвистическими методами [12].

Значительный вклад в оптимизацию поисковых процессов внесен работами по кэшированию и параллельным вычислениям [13, 14], ставшим неотъемлемым компонентом промышленных систем. Вопросы оценки качества поиска через метрики Precision, Recall и F-мера глубоко исследованы в работах по информационному поиску [1, 15]. Особого внимания заслуживает исследование [16], где подчеркивается важность учета временных характеристик документов в таких предметных областях, как новостные потоки.

Несмотря на значительное количество исследований в данной области, наблюдается определенный дефицит работ, комплексно рассматривающих всю технологическую цепочку обработки – от морфологической нормализации до финального ранжирования – с фокусом на оптимизацию для русского языка и специализированных корпусов.

Целью данного исследования является разработка и экспериментальная проверка гибридного алгоритма поиска, оптимизированного для работы с русскоязычными текстовыми корпусами. Для достижения этой цели были поставлены следующие задачи:

1. Анализ и модификация алгоритмов стемминга для русского языка.
2. Разработка многоуровневой системы метрик для оценки релевантности документа поисковому запросу.
3. Создание механизма взвешенного ранжирования результатов.
4. Оптимизация алгоритма для работы с большими объемами данных.

МЕТОДОЛОГИЯ

Гибридный алгоритм поиска

Предлагаемый алгоритм состоит из трех основных этапов: предобработки запроса и документов, поиска кандидатов и взвешенного ранжирования.

1) Предобработка данных.

– Токенизация: Входной текст разбивается на токены с использованием регулярных выражений, учитывающих кириллические символы и дефисы.

– Фильтрация стоп-слов: Используется расширенный список стоп-слов русского языка, исключающий из обработки служебные части речи и частотные лексемы, не несущие смысловой нагрузки.

– Стемминг: Применяется модифицированный алгоритм стемминга, ориентированный на высокую скорость. В отличие от классического алгоритма Портера, данный алгоритм использует каскадное удаление окончаний, сгруппированных по частеречной принадлежности (существительные, прилагательные, глаголы), с проверкой наличия гласной в основе для избегания излишнего отсека. Для каждого слова генерируется набор возможных основ, что повышает полноту поиска.

2) Поиск кандидатов.

Поиск осуществляется по индексированному корпусу документов. Для каждого поискового запроса Q , состоящего из набора ключевых слов $\{k_1, k_2, \dots, k_n\}$, выполняется:

- Отсев стоп-слов из запроса.
- Поиск документов, в которых встречается хотя бы одно из нетривиальных ключевых слов или их основ. Для ускорения процесса используется оптимизированная проверка вхождений, сочетающая:

- Точный поиск по нижнему регистру.
- Поиск по границам слов с использованием скомпилированных регулярных выражений.
- Поиск по пересечению множеств основ запроса и основ слов в документе.

3) Многоуровневая метрика релевантности и ранжирование.

Для каждого документа-кандидата D вычисляется комплексная оценка релевантности, которая является взвешенной суммой нескольких метрик:

$$Score(D, Q) = w_{exact} * M_{exact} + w_{stem} * M_{stem} + w_{position} * M_{position} + + w_{time} * M_{time} \quad (1)$$

где:

M_{exact} – метрика точных совпадений. Рассчитывается как количество ключевых слов из запроса, имеющих точное совпадение (с учетом границ слов) в тексте документа D . Эта метрика имеет наибольший вес (w_{exact}), так как точное совпадение является сильнейшим индикатором релевантности.

M_{stem} – метрика стем-совпадений. Учитывает количество ключевых слов, для которых найдены совпадения по основам. Для вычисления используется множество основ, сгенерированных для каждого слова запроса и каждого слова в документе. Совпадение фиксируется, если основа слова из документа содержит основу из запроса, или наоборот.

$M_{position}$ – позиционная метрика. Увеличивает вес документов, в которых найденные ключевые слова встречаются в semantically важных полях, таких как заголовок (тип новости) или тезис. Эмпирически установлено, что вес для этих полей должен быть в 2-3 раза выше, чем для основного текста [2].

M_{time} – временная метрика. Учитывает актуальность документа. Для новостных статей более свежие документы получают более высокий балл по формуле, обратно пропорциональной логарифму разницы между текущей датой и датой публикации документа.

Документы сортируются по убыванию итогового показателя $Score(D, Q)$. Внутри групп с одинаковым счетом применяется дополнительная сортировка по дате (свежие документы выше).

Оптимизационные стратегии

Для обеспечения высокой производительности при работе с большими корпусами данных реализован ряд оптимизаций:

- Кэширование результатов стемминга. Поскольку процесс стемминга является вычислительно затратным, его результаты кэшируются в памяти с использованием словаря. Это позволяет избежать повторной обработки одних и тех же слов.
- Параллельная обработка. Этапы загрузки данных из файлов и первоначального поиска кандидатов распараллеливаются с использованием библиотеки Parallel.ForEach и конкурентных коллекций (ConcurrentBag).
- Пакетная обработка (Batching). При обработке больших списков документов они разбиваются на пакеты фиксированного размера, что позволяет эффективно управлять потреблением памяти.
- Использование скомпилированных регулярных выражений. Для операций поиска по границам слов используются предварительно скомпилированные объекты Regex с флагом RegexOptions.Compiled, что значительно ускоряет их выполнение.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Для проверки эффективности предложенного алгоритма был проведен эксперимент на корпусе русскоязычных новостных статей объемом approximately 10000 документов.

Сравнивались три метода:

1. Базовый метод: Точный поиск по ключевым словам (case-insensitive).
2. Метод на основе стемминга: Поиск по основам слов без сложного ранжирования.

3. Предлагаемый гибридный метод: Сочетание точного поиска, стемминга и многоуровневого ранжирования.

В качестве запросов использовались 50 разнообразных фраз длиной от 1 до 4 слов. Эталонная разметка релевантности выполнялась экспертом вручную. Качество оценивалось по метрикам Precision@10 (точность среди первых 10 результатов) и Recall (полнота). Результаты представлены в Таблице 1.

Таблица 1

Сравнение эффективности алгоритмов поиска

Table 1

Comparison of the efficiency of search algorithms

Метод	Средний Precision@10	Средний Recall	Время выполнения
Базовый метод	0.65	0.58	0.8
На основе стемминга	0.72	0.85	1.5
Гибридный метод	0.82	0.88	1.9

Анализ результатов показал, что гибридный метод значительно превосходит базовый по точности на 17% и полноте на 30%. По сравнению с простым стеммингом, гибридный метод обеспечивает прирост точности на 10% за счет более качественного ранжирования, которое выдвигает на первые позиции наиболее релевантные документы. Несмотря на увеличение времени выполнения, показатель в 1.9 секунды на сложный запрос признан удовлетворительным для интерактивной работы.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

Результаты эксперимента подтверждают гипотезу о том, что комбинация различных методов поиска и метрик релевантности позволяет достичь более высокого качества, чем использование любого из методов в отдельности.

- Эффективность стемминга: Использование стемминга позволило резко повысить полноту поиска, эффективно находя документы с различными грамматическими формами искомых слов, например, «экономика», «экономики», «экономике».

- Важность многоуровневого ранжирования: Метрика M_{exact} позволила отсеять документы, релевантные лишь по контексту, и поднять наверх документы с прямыми упоминаниями. Позиционная метрика $M_{position}$ корректно идентифицировала документы, где ключевые слова фигурировали в заголовке, что часто указывает на основную тему.

- Влияние временного фактора: Временная метрика M_{time} улучшила пользовательский опыт, предоставляя в первую очередь актуальные новости по запросам, чувствительным ко времени.

Основным ограничением предложенного метода является его зависимость от качества стемминга. В редких случаях агрессивное отсечение окончаний может привести к ложным срабатываниям, например, основа «кос» может быть общей для слов «космос» и «косить». Кроме того, метод не учитывает семантическую близость слов, что является направлением для будущих исследований.

ЗАКЛЮЧЕНИЕ

В рамках исследования разработан и протестирован гибридный алгоритм поиска, адаптированный для работы с русскоязычными текстовыми корпусами. Алгоритм интегрирует

морфологический анализ, многоуровневую систему оценки релевантности и ряд оптимизационных техник для обеспечения высокой производительности.

Проведенный эксперимент демонстрирует значимое превосходство предложенного подхода над базовыми методами по точности и полноте. Основные выводы работы заключаются в следующем:

1. Каскадный стемминг с генерацией множества основ для каждого слова является эффективным способом повышения полноты поиска в морфологически сложных языках.
2. Композитная метрика релевантности, объединяющая подсчет точных совпадений, стем-совпадений, позиционный и временной факторы, позволяет значительно улучшить точность ранжирования по сравнению с простыми метриками.
3. Использование кэширования и параллельных вычислений делает сложный гибридный алгоритм применимым для работы с большими объемами данных в режиме, близком к реальному времени.

Перспективы дальнейших исследований видятся в интеграции предложенного алгоритма с легковесными семантическими моделями, для учета синонимии, например, на основе word2vec или fastText, а также в применении методов активного обучения для непрерывного улучшения качества ранжирования на основе обратной связи от пользователей.

Конфликт интересов: о потенциальном конфликте интересов не сообщалось.

Conflict of interest: no potential conflict of interest related to this article was reported.

Список литературы

1. Воронцов К.В. Лекции по алгоритмам классификации и регрессии. М.: МАКС Пресс, 2013. – 215 с.
2. Жигалин А.А. Оптимизация алгоритмов поиска в полнотекстовых базах данных // Информационные технологии и вычислительные системы. 2020. № 2. С. 45-58.
3. Лопулин С.А. Современные методы лемматизации для русского языка // Труды института системного программирования РАН. 2019. Т. 31, № 1. С. 155-170.
4. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008. 544 p.
5. Robertson S., Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond // Foundations and Trends in Information Retrieval. 2009. Vol. 3, № 4. P. 333-389.
6. Porter M. F. An algorithm for suffix stripping // Program. 1980. Vol. 14, № 3. P. 130-137.
7. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // International Conference on Machine Learning. 2003. P. 273-280.
8. Page L., Brin S., Motwani R., Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. Stanford: Stanford InfoLab, 1999. 17 p.
9. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing and Management. 1988. Vol. 24, № 5. P. 513-523.
10. Nogueira R., Cho K. Passage Re-ranking with BERT // arXiv. 2019. URL: <https://arxiv.org/pdf/1901.04085> (дата обращения: 07.08.2025).
11. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Devlin J., Chang M. W., Lee K., Toutanova K. // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Vol. 1. P. 4171-4186.
12. Mitra B., Craswell N. An Introduction to Neural Information Retrieval // Foundations and Trends® in Information Retrieval. 2018. Vol. 13, № 1. P. 1-126.
13. Agrawal R., Srikant R. Fast algorithms for mining association rules in large databases // Proceedings of the 20th International Conference on Very Large Data Bases. 1994. P. 487-499.
14. Dean J., Ghemawat S. MapReduce: Simplified data processing on large clusters // Communications of the ACM. 2008. Vol. 51, № 1. P. 107-113.
15. Buckland M., Gey F. The relationship between Recall and Precision // Journal of the American Society for Information Science. 1994. Vol. 45, № 1. P. 12-19.
16. Li X., Croft W. B. Time-based language models // Proceedings of the twelfth international conference on Information and knowledge management. 2003. P. 469-475.

References

1. Vorontsov K.V. Lectures on Classification and Regression Algorithms. Moscow: MAKSS Press, 2013. – 215 p.
2. Zhigalin A.A. Optimization of Search Algorithms in Full-Text Databases // Information Technology and Computing Systems. 2020. – No. 2. – P. 45-58.
3. Lopulin S.A. Modern Lemmatization Methods for the Russian Language // Proceedings of the Institute for System Programming of the Russian Academy of Sciences. 2019. Vol. 31, No. 1. P. 155-170.
4. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008. 544 p.
5. 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.
6. Porter M. F. An algorithm for suffix stripping // Program. 1980. Vol. 14, No. 3. P. 130-137.
7. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // International Conference on Machine Learning. 2003. P. 273-280.
8. Page L., Brin S., Motwani R., Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. Stanford: Stanford InfoLab, 1999. 17 p.
9. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing and Management. 1988. Vol. 24, No. 5. P. 513-523.
10. Nogueira R., Cho K. Passage Re-ranking with BERT // arXiv. 2019. URL: <https://arxiv.org/pdf/1901.04085> (accessed: 07.08.2025).
11. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Devlin J., Chang M. W., Lee K., Toutanova K. // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Vol. 1. P. 4171-4186.
12. Mitra B., Craswell N. An Introduction to Neural Information Retrieval // Foundations and Trends® in Information Retrieval. 2018. Vol. 13, No. 1. P. 1-126.
13. Agrawal R., Srikant R. Fast algorithms for mining association rules in large databases // Proceedings of the 20th International Conference on Very Large Data Bases. 1994. P. 487-499.
14. Dean J., Ghemawat S. MapReduce: Simplified data processing on large clusters // Communications of the ACM. 2008. Vol. 51, No. 1. P. 107-113.
15. Buckland M., Gey F. The relationship between Recall and Precision // Journal of the American Society for Information Science. 1994. Vol. 45, No. 1. P. 12-19.
16. Li X., Croft W. B. Time-based language models // Proceedings of the twelfth international conference on Information and knowledge management. 2003. P. 469-475.

Беляков Ярослав Вадимович, студент кафедры информационных технологий, Институт нефтепереработки и нефтехимии ФГБОУ ВО УГНТУ в г. Салавате, Россия

Левина Татьяна Михайловна, заведующий кафедрой информационных технологий, Институт нефтепереработки и нефтехимии ФГБОУ ВО УГНТУ в г. Салавате, Россия

Belyakov Yaroslav Vadimovich, Student of the Department of Information Technology, Federal State Budgetary Educational Institution of Higher Education USPTU in Salavat, Salavat, Russia

Levina Tatyana Mikhailovna, Head of the Department of Information Technology, Federal State Budgetary Educational Institution of Higher Education USPTU in Salavat, Salavat, Russia