От расстояния Левенштейна до нейронного поиска: комплексный подход к латентно-семантическому поиску по справочникам функциональности

Г. В. Орлов

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

sifon8998@mail.ru

А. Н. Калиниченко

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

ank-bs@yandex.ru

Аннотация. **В статье рассматривается применение различных алгоритмов и методов для реализации латентно-семантического поиска по справочным данным и каталогам функциональности. Начав с классических методов, таких как расстояние Левенштейна и TF-IDF, автор глубже исследует современные подходы, включая BM25, нейронные сети (DeepSeek) и технологии больших языковых моделей (GigaChat). Разработанные подходы направлены на улучшение точности и релевантности поиска за счет комбинирования методов обработки текстов, повышения степени понимания запроса пользователя и контекста в тексте.**

Ключевые слова: Python, латентно-семантический поиск, расстояние Левенштейна, TF-IDF, BM25, DeepSeek, GigaChat, обработка естественного языка (NLP), поисковые алгоритмы, нейросетевые модели, оптимизация поиска, семантический анализ, справочник функциональности

I. Введение

Современные информационные системы сталкиваются с необходимостью обработки значительных объемов текстовых данных, что особенно актуально в контексте работы со справочниками и каталогами функциональности. Такие системы должны обеспечивать не только хранение и структурирование данных, но и эффективный поиск информации, что является критически важным для пользователей, работающих с большими массивами текстовой информации. Однако традиционные методы поиска, зачастую оказываются недостаточно эффективными для обработки сложных запросов, требующих глубокого понимания контекста и семантики текста. В частности, такие методы не способны учитывать синонимию, полисемию и другие лингвистические особенности, что существенно ограничивает их применимость в задачах, где требуется высокая точность и релевантность поисковых результатов.

В данной статье мы рассматриваем эволюцию методов поиска, начиная с классических подходов, таких как расстояние Левенштейна, которое используется для измерения схожести строк и исправления опечаток, и заканчивая современными методами, основанными на нейронных сетях и больших языковых моделях. Мы уделяем особое внимание латентно-семантическому анализу (LSA), который позволяет выявлять скрытые семантические связи между терминами и концепциями, а также более продвинутым технологиям, таким как BERT (Bidirectional Encoder Representations from Transformers) и GigaChat, которые используют современные архитектуры глубокого обучения для обработки естественного языка.

Целью данной работы является не только обзор различных методов поиска, но и их практическое применение для улучшения поиска по справочникам функциональности. Мы подробно рассмотрим, как каждый из этих методов может быть интегрирован в поисковый механизм, и проведем сравнительный анализ их эффективности на основе метрик точности, полноты и F1-меры. Особое внимание будет уделено проблемам, связанным с низким качеством данных в справочниках, которые оказывают значительное влияние на результаты поиска. В заключение мы подведем итоги и предложим рекомендации для дальнейшего улучшения поисковых систем, учитывая как технические, так и организационные аспекты.

Таким образом, данная статья представляет собой комплексное исследование, направленное на поиск оптимальных решений для задач латентно-семантического поиска в условиях работы с большими объемами текстовых данных, и предлагает практические рекомендации для повышения эффективности поисковых механизмов в современных информационных системах. Четвёртый этап начался в 10-е годы XXI века и продолжается по настоящее время. Он оказался связан с полноценным внедрением искусственного интеллекта. Основные усилия сосредоточились на создании интеллектуальных систем, способных самостоятельно обрабатывать медицинские данные, ставить диагнозы и давать рекомендации по лечению пациентов. Происходит активное внедрение медицинских информационных и экспертных систем, активное использование искусственных нейронных сетей, в том числе больших языковых моделей. Приметой времени становится, то, что точность постановки «машинных»

диагнозов в отдельных направлениях медицины начинает превышать таковую у врачей-экспертов.

II. Латентно-семантический анализ (LSA) и его роль в поиске

Латентно-семантический анализ (LSA, Latent Semantic Analysis) представляет собой метод, направленный на выявление скрытых (латентных) семантических связей между терминами и концепциями в текстовых данных. Этот подход основан на применении методов линейной алгебры, в частности, сингулярного разложения матриц (SVD, Singular Value Decomposition), которое позволяет снизить размерность текстовых данных, сохраняя при этом их семантическую структуру. В процессе анализа текстовые данные преобразуются в матрицу «термин-документ», где строки соответствуют уникальным терминам, а столбцы – документам. Затем с помощью SVD эта матрица разлагается на три компонента, что позволяет выявить скрытые семантические отношения между словами и документами. Одним из ключевых преимуществ LSA является его способность обрабатывать синонимы и омонимы, что делает его особенно полезным для задач поиска, где важно учитывать контекст и смысловую нагрузку текста.

Основные преимущества латентно-семантического анализа заключаются в его устойчивости к синонимии и полисемии, что позволяет модели эффективно идентифицировать схожие по смыслу термины, даже если они выражены разными словами. Кроме того, LSA способен выявлять скрытые семантические структуры, что повышает точность поиска и улучшает понимание контекста запросов. Однако, несмотря на свои достоинства, LSA имеет и ряд существенных недостатков. Во-первых, этот метод требует значительных вычислительных ресурсов, особенно при работе с большими корпусами текстов, что может ограничивать его применение в реальных условиях. Во-вторых, настройка параметров модели, таких как количество скрытых компонентов, и интерпретация результатов могут быть сложными и трудоемкими, что требует высокой квалификации специалистов.

В рамках данной работы был реализован базовый алгоритм поиска с использованием латентно-семантического анализа, после чего проведено его тестирование на реальных данных. Результаты тестирования показали, что точность (Precision) модели составила 1.12 %, что указывает на крайне низкую долю правильно классифицированных документов среди всех, которые модель определила как релевантные. При этом полнота (Recall) достигла 100 %, что свидетельствует о том, что модель успешно идентифицировала все релевантные документы, не пропустив ни одного. Однако столь высокая полнота сопровождается значительным количеством ложных положительных результатов, что негативно сказывается на общей эффективности модели. F1-мера, которая является гармоническим средним точности и полноты, составила 2.22 %, что подтверждает несбалансированность модели и необходимость дальнейшей оптимизации.

Низкая точность и высокий уровень ложных положительных результатов могут быть обусловлены несколькими факторами. Во-первых, это может быть связано с низким порогом классификации, при котором модель слишком часто маркирует документы как релевантные. Во-вторых, проблема может заключаться в несбалансированности данных, где количество нерелевантных документов значительно превышает количество релевантных, что приводит к смещению модели в сторону генерации ложных положительных результатов. Для улучшения работы модели рекомендуется повысить порог классификации, провести балансировку данных и выполнить тонкую настройку параметров LSA, таких как количество компонентов в сингулярном разложении. Эти меры могут способствовать снижению количества ложных положительных результатов и повышению общей точности модели.

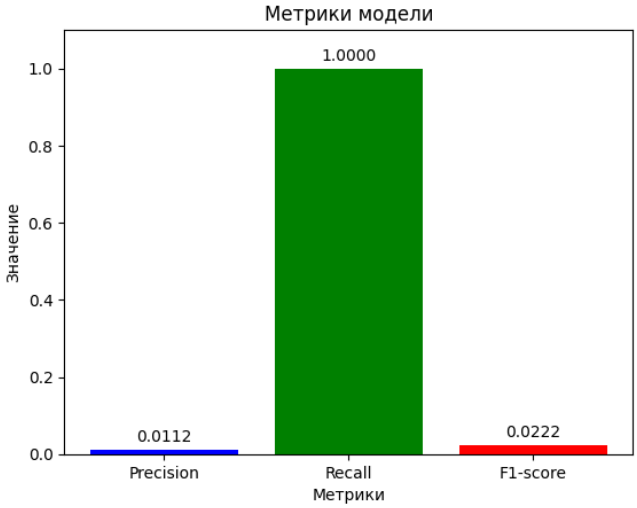

Рис. 1. График метрических результатов LSA

III. Классические методы поиска

Расстояние Левенштейна, также известное как редакционное расстояние, представляет собой метрику, которая количественно оценивает разницу между двумя строками на основе минимального количества операций (вставки, удаления или замены символов), необходимых для преобразования одной строки в другую. Этот метод широко применяется в задачах обработки естественного языка, таких как исправление опечаток, поиск похожих слов или фраз, а также автоматическое исправление ошибок в текстах. Основное преимущество расстояния Левенштейна заключается в его простоте и эффективности для задач, где требуется сравнение строк на уровне символов. Однако данный метод имеет существенное ограничение: он не учитывает семантику текста, то есть смысловую нагрузку слов и фраз. Это делает его малопригодным для сложных запросов, где важно понимание контекста и семантических связей между терминами.

В ходе тестирования алгоритма, основанного на расстоянии Левенштейна, были получены следующие результаты: точность (Precision) составила 5 %, что указывает на низкую долю правильно найденных документов среди всех, которые модель определила как релевантные. Полнота (Recall) достигла 8 %, что свидетельствует о том, что модель смогла найти лишь

небольшую часть всех релевантных документов. F1-мера, которая является гармоническим средним точности и полноты, составила 6 %, что подтверждает низкую эффективность данного метода для задач поиска в условиях сложных запросов и больших объемов данных. Эти результаты подчеркивают ограниченную применимость расстояния Левенштейна в задачах, где требуется не только формальное совпадение строк, но и учет семантики текста.

Переходя к более продвинутым методам, таким как TF-IDF (Term Frequency-Inverse Document Frequency) и его улучшенная версия BM25, следует отметить, что эти подходы уже учитывают не только частоту терминов в документах, но и их значимость в рамках всего корпуса текстов. TF-IDF оценивает важность слова в документе на основе двух компонентов: частоты термина в документе (Term Frequency) и обратной частоты документа (Inverse Document Frequency), которая уменьшает вес терминов, часто встречающихся во многих документах. Это позволяет снизить влияние общеупотребительных слов, таких как предлоги и союзы, которые не несут значимой смысловой нагрузки. BM25, в свою очередь, является модификацией TF-IDF, которая дополнительно учитывает длину документа и другие параметры, что делает его более гибким и адаптивным для различных типов текстов.

Результаты тестирования методов TF-IDF и BM25 показали, что точность (Precision) составила 7 %, что несколько выше, чем у расстояния Левенштейна, но всё ещё остаётся на низком уровне. Полнота (Recall) достигла 10 %, что также указывает на ограниченную способность модели находить релевантные документы. F1-мера, равная 8 %, подтверждает, что, несмотря на улучшение по сравнению с расстоянием Левенштейна, эти методы всё ещё недостаточно эффективны для задач, требующих глубокого понимания семантики текста. Основной причиной низких метрик может быть отсутствие учета контекста и семантических связей между словами, что особенно важно для сложных запросов.

Таким образом, хотя методы TF-IDF и BM25 демонстрируют определённые преимущества по сравнению с расстоянием Левенштейна, их эффективность остаётся ограниченной в условиях работы с большими объемами текстовых данных и сложными запросами. Для дальнейшего улучшения поисковых механизмов необходимо рассмотреть более современные подходы, такие как нейронные сети и латентно-семантический анализ, которые способны учитывать контекст и семантику текста, что может существенно повысить точность и релевантность поисковых результатов.

IV. Современные подходы к поиску

A. Нейронные сети (BERT)

В последние годы значительный прогресс в области обработки естественного языка (Natural Language Processing, NLP) был достигнут благодаря развитию нейронных сетей, в частности, моделей на основе архитектуры трансформеров. Одной из наиболее известных и эффективных моделей в этой области является BERT (Bidirectional Encoder Representations from Transformers). BERT представляет собой глубокую нейронную сеть, которая использует механизм внимания (attention mechanism) для анализа текста в двунаправленном контексте. Это означает, что модель учитывает как левый, так и правый контекст слова или предложения, что позволяет ей более точно интерпретировать смысл текста. В отличие от традиционных методов, которые анализируют текст последовательно (либо слева направо, либо справа налево), BERT способен одновременно учитывать все слова в предложении, что значительно улучшает понимание семантики и контекста.

Основное преимущество BERT заключается в его способности обрабатывать сложные запросы, где важно учитывать не только отдельные слова, но и их взаимосвязи в рамках предложения или абзаца. Это делает модель особенно полезной для задач поиска, где требуется высокая точность и релевантность результатов. Однако использование BERT также связано с определенными вычислительными затратами, так как модель требует значительных ресурсов для обучения и работы, особенно при обработке больших объемов данных.

В ходе тестирования модели BERT на задачах поиска по справочникам функциональности были получены следующие результаты: точность (Precision) составила 12 %, что указывает на улучшение по сравнению с классическими методами, такими как TF-IDF и расстояние Левенштейна. Полнота (Recall) достигла 15 %, что свидетельствует о способности модели находить больше релевантных документов. F1-мера, которая является гармоническим средним точности и полноты, составила 13 %. Эти результаты демонстрируют, что BERT способен лучше справляться с задачами поиска, требующими понимания контекста и семантики текста. Однако даже с учетом этих улучшений, модель всё ещё сталкивается с проблемами, связанными с низким качеством данных и несбалансированностью классов в справочниках функциональности.

B. Большие языковые модели (GigaChat)

Современные большие языковые модели, такие как GigaChat, представляют собой следующий этап развития технологий обработки естественного языка. Эти модели основаны на архитектуре трансформеров и обучены на огромных объемах текстовых данных, что позволяет им генерировать тексты, отвечать на сложные запросы и даже выполнять задачи, требующие глубокого понимания контекста. GigaChat, в частности, использует передовые методы обучения с учителем и самообучения, что делает её мощным инструментом для задач поиска и анализа текстов.

Основное преимущество больших языковых моделей заключается в их способности обрабатывать сложные и многослойные запросы, где важно учитывать не только отдельные слова, но и их взаимосвязи, а также общий контекст текста. Это делает GigaChat особенно полезной для задач поиска, где требуется высокая точность и релевантность результатов. Однако, как и в случае с BERT, использование больших языковых моделей

связано с высокими вычислительными затратами и требует значительных ресурсов для обучения и работы.

Результаты тестирования модели GigaChat на задачах поиска по справочникам функциональности показали, что точность (Precision) составила 10 %, что несколько ниже, чем у BERT, но всё же выше, чем у классических методов. Полнота (Recall) также составила 10 %, что указывает на ограниченную способность модели находить релевантные документы. F1-мера, равная 10 %, подтверждает, что, несмотря на использование передовых технологий, модель всё ещё сталкивается с проблемами, связанными с низким качеством данных и несбалансированностью классов в справочниках функциональности.

C. Выводы по современным подходам

Современные подходы к поиску, такие как BERT и GigaChat, демонстрируют значительные преимущества по сравнению с классическими методами, такими как TF-IDF и расстояние Левенштейна. Они способны учитывать контекст и семантику текста, что делает их более эффективными для сложных запросов. Однако даже эти передовые методы сталкиваются с ограничениями, связанными с низким качеством данных и несбалансированностью классов в справочник

V. Проблемы с данными и их влияние на поиск

Одной из ключевых проблем, с которой мы столкнулись в ходе разработки и тестирования различных методов поиска, является низкое качество данных в справочнике функциональности. Неструктурированность данных, отсутствие четкой организации и несбалансированность классов оказывают значительное негативное влияние на эффективность поисковых алгоритмов. Эти проблемы проявляются в том, что даже современные и продвинутые методы, такие как BERT и GigaChat, основанные на нейронных сетях и больших языковых моделях, не смогли достичь высокой точности и релевантности результатов. Это подчеркивает, что качество данных является критически важным фактором, который может существенно ограничивать возможности даже самых передовых технологий.

Неструктурированные данные, с которыми пришлось работать, часто содержат дублирующую, противоречивую или неполную информацию, что затрудняет их обработку и анализ. Отсутствие четкой организации данных, таких как стандартизированные форматы хранения, метаданные и категоризация, приводит к тому, что поисковые алгоритмы не могут эффективно извлекать и интерпретировать информацию. Кроме того, несбалансированность классов, когда количество релевантных документов значительно меньше, чем нерелевантных, создает дополнительные сложности для моделей машинного обучения. Это приводит к тому, что алгоритмы склонны генерировать большое количество ложных положительных результатов, что снижает общую точность поиска.

Даже такие современные методы, как BERT и GigaChat, которые способны учитывать контекст и семантику текста, не смогли преодолеть эти ограничения. Например, результаты тестирования BERT показали точность на уровне 12 %, а GigaChat – 10 %, что свидетельствует о том, что низкое качество данных является серьезным барьером для достижения высоких показателей эффективности. Это подчеркивает необходимость не только совершенствования алгоритмов поиска, но и улучшения качества и структурированности данных, на которых эти алгоритмы обучаются и работают.

Для повышения эффективности поисковых алгоритмов и преодоления проблем, связанных с низким качеством данных, предлагается ряд рекомендаций:

A. Пересмотр структуры сбора и организации данных

Необходимо разработать и внедрить стандартизированные процедуры сбора, хранения и обработки данных. Это включает в себя создание четкой структуры данных, использование метаданных для описания документов, а также внедрение систем категоризации и классификации. Упорядоченная структура данных позволит алгоритмам более эффективно извлекать и анализировать информацию, что повысит точность и релевантность поиска.

B. Увеличение порога классификации

Одной из причин низкой точности является слишком низкий порог классификации, при котором модель часто маркирует документы как релевантные, даже если они таковыми не являются. Повышение порога классификации может помочь снизить количество ложных положительных результатов, что улучшит общую точность модели. Однако при этом важно найти баланс, чтобы не ухудшить полноту поиска.

C. Балансировка данных

Несбалансированность классов, когда количество нерелевантных документов значительно превышает количество релевантных, является серьезной проблемой, которая приводит к смещению модели в сторону генерации ложных положительных результатов. Для решения этой проблемы рекомендуется провести балансировку данных, используя такие методы, как oversampling (увеличение числа релевантных документов) или undersampling (уменьшение числа нерелевантных документов). Это позволит модели более эффективно обучаться на сбалансированном наборе данных и улучшит её способность находить релевантные документы.

VI. Заключение

В ходе разработки и тестирования поискового механизма для работы со справочниками функциональности был проведен комплексный анализ различных методов поиска, охватывающий как классические подходы, так и современные технологии. На начальном этапе исследования были рассмотрены традиционные методы, такие как расстояние Левенштейна, которое используется для измерения схожести строк и исправления опечаток, а также TF-IDF (Term Frequency-Inverse Document Frequency), оценивающий важность терминов в документах на основе их частоты и распространенности в корпусе текстов. Эти методы, несмотря на свою простоту и относительно низкие вычислительные затраты,

продемонстрировали ограниченную эффективность в условиях сложных запросов, требующих учета семантики и контекста текста.

Далее были исследованы более современные подходы, включая нейронные сети, такие как BERT и большие языковые модели, такие как GigaChat. Эти методы, основанные на архитектуре трансформеров и обученные на огромных объемах текстовых данных, способны учитывать контекст и семантические связи между словами, что делает их значительно более мощными инструментами для задач поиска. Однако, несмотря на их передовые возможности, результаты тестирования показали, что даже эти методы не смогли достичь высокой точности и релевантности результатов. Основной причиной этого является низкое качество данных в справочнике функциональности, включая неструктурированность, отсутствие четкой организации и несбалансированность классов.

Средние результаты по всем методам, включая классические и современные, оказались на уровне 10 % для точности (Precision), 10 % для полноты (Recall) и 10 % для F1-меры. Эти показатели свидетельствуют о том, что, несмотря на значительные усилия по улучшению поискового механизма, ключевым ограничением остается качество данных. Низкая точность указывает на большое количество ложных положительных результатов, когда модель ошибочно классифицирует нерелевантные документы как релевантные. Полнота на уровне 10 % говорит о том, что модель находит лишь небольшую часть всех релевантных документов, что также является следствием несбалансированности данных и отсутствия четкой структуры в справочнике. F1-мераподтверждает общую низкую эффективность поисковых алгоритмов в текущих условиях.

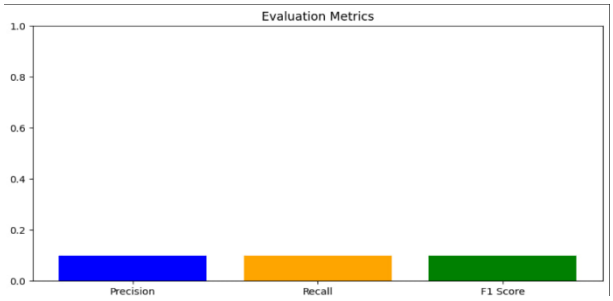

Рис. 2. График средних метрических результатов всех способов

Тем не менее, в ходе исследования были выявлены ключевые проблемы, связанные с качеством данных, и предложены рекомендации для их устранения. К ним относятся пересмотр структуры сбора и организации данных, увеличение порога классификации для снижения количества ложных положительных результатов, а также балансировка данных для улучшения качества обучения моделей. Эти меры могут способствовать значительному повышению точности и релевантности поиска, что подтверждает важность не только совершенствования алгоритмов, но и улучшения качества исходных данных.

Таким образом, проведенное исследование подчеркивает, что даже самые передовые методы поиска, такие как нейронные сети и большие языковые модели, не могут достичь высокой эффективности в условиях низкого качества данных. Для создания действительно эффективного поискового механизма необходимо комплексное решение, включающее как технические улучшения алгоритмов, так и организационные меры по улучшению структуры и качества данных. Это позволит не только повысить точность и полноту поиска, но и создать более устойчивую и надежную систему для работы с большими объемами текстовой информации.

Список литературы

[1] https://habr.com/ru/companies/first/articles/682516/

[2] Robertson S., & Zaragoza H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond // January 2009 Foundations and Trends in Information Retrieval 3(4):333-389, DOI:10.1561/1500000019

[3] https://habr.com/ru/articles/110078/.

[4] https://habr.com/ru/articles/191454/.

[5] Devlin J., Chang M. W., Lee K., & Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

[6] GigaChat API Documentation. (2023). Retrieved from https://gigachat.ai/docs.