DOI 10.37882/2223-2966.2025.01-2.12

ДВУХЭТАПНЫЙ ИНФОРМАЦИОННЫЙ ПОИСК С ИСПОЛЬЗОВАНИЕМ COLBERT

TWO-STEP INFORMATION RETRIEVAL USING COLBERT

O. Maksimenko
Yu. Boronina

Summary. The article discusses the implementation of information search using both well-known ranking algorithms and language models. The goal is to improve the search quality by using a language model. One of them, ColBERT, is a fast and accurate search model that provides scalable BERT-based search in large text collections in tens of milliseconds. The model was trained on test data, and then it was tested using metrics used to evaluate search quality.

Keywords: data, language model, information search, digitalization, ranking.

Максименко Олег Евгеньевич
Аспирант, НИЯУ МИФИ, Москва
skipperprivate@yandex.ru
Боронина Юлия Сергеевна
Аспирант, НИЯУ МИФИ, Москва
Borona05@mail.ru

Аннотация. В статье рассматривается реализация информационного поиска с использованием как общеизвестных алгоритмов ранжирования, так и языковых моделей. Цель — повысить качество поиска за счет использования языковой модели. Одна из них, ColBERT — это быстрая и точная поисковая модель, обеспечивающая масштабируемый поиск на основе BERT в больших текстовых коллекциях за десятки миллисекунд. Были проведены обучение модели на тестовых данных, а затем она была протестирована с использованием метрик, применяемых при оценке качества поиска.

Ключевые слова: данные, языковая модель, информационный поиск, цифровизация, ранжирование.

Введение

Поиск информации всегда в наше время стал важной частью жизни любого человека в связи с бурным развитием средств связи и коммуникации. Интернет позволяет получить доступ к огромному объему информации, размещенной на различных ресурсах. Однако процесс их нахождения по имеющемуся у пользователя запросу — это нетривиальный процесс, учитывая тот факт, что запрос пишется на естественном языке, зачастую без соблюдения какой-либо четкой структуры и правил. Этот процесс называется информационным поиском. Можно утверждать, что информационный поиск — это процесс поиска в большой коллекции некоего неструктурированного материала, удовлетворяющего информационные потребности [1]. Качество поиска оценивается релевантностью полученных документов на запрос пользователя. Актуальность исследования состоит в удовлетворении потребностей пользователей в повышении точности поиска.

Реализация может быть произведена самыми различными способами, начиная с функций ранжирования и заканчивая большими языковыми моделями. Программу или совокупность программ, позволяющих пользователю получить доступ к необходимой информации, называется поисковым движком или поисковой системой.

Чтобы согласовать эффективность и контекстуализацию в IR, мы предлагаем ColBERT, модель ранжирования, основанную на контекстуализированном позднем взаимодействии с BERT. Как следует из названия, ColBERT предлагает новую парадигму позднего взаимодействия для оценки релевантности запроса q и документа d. При позднем взаимодействии q и d отдельно кодируются в два набора контекстуальных вложений, а релевантность оценивается с помощью дешевых и удобных для сокращения вычислений между обоими наборами, то есть быстрых вычислений, которые позволяют ранжировать без исчерпывающей оценки каждого возможного кандидата.[7]

Литературный обзор

Книга «Введение в информационный поиск» представляет собой введение в область информационного поиска и охватывает основные концепции и методы, используемые для поиска информации в больших объёмах данных. Она может быть полезна для исследователей и специалистов, работающих в области информационного поиска, а также для студентов, изучающих эту тему.

Статья Витухновской А.А. посвящена обучению технологиям и стратегиям информационного поиска. Автор рассматривает дифференциальные признаки информационно-поисковых систем и их роль в обучении информационному поиску. Статья может быть полезна исследователям и специалистам, работающим в области информационного поиска, а также преподавателям и студентам, изучающим эту тему.

Информационный поиск актуален в современном мире по нескольким причинам:

Быстрый доступ к информации. В эпоху цифровизации информация обновляется и растёт с невероятной скоростью. Информационный поиск помогает быстро находить актуальные данные, новости, исследования и другие материалы.

Принятие обоснованных решений. Информационный поиск позволяет получать достоверные и разнообразные источники информации, что способствует принятию более обоснованных и информированных решений в различных областях деятельности.

Образование и саморазвитие. Благодаря информационному поиску люди могут получать доступ к образовательным ресурсам, научным публикациям, онлайн-курсам и другим материалам для обучения и развития.[6]

Профессиональная деятельность. Специалисты в разных областях используют информационный поиск для нахождения новых идей, методов работы, исследований и других материалов, которые помогают им развиваться профессионально и повышать качество своей работы.

Материалы и методы

Перед первым этапом ранжирования текстовые документы и текстовые запросы проходят предварительную обработку. Она заключается в отсеивании данных, малозначимых для поиска (знаки препинания, числа и т.д). Далее текст подвергается стеммингу, приходя таким образом в вид, пригодный для дальнейшей работы.[5]

Вторым этапом идет ранжирование с помощью семейства функций BM25. Существует несколько подвидов этой функции — на основе метрики Precision была выбрана BM25L. В результате работы функции отбирается топ N документов, соответствующих запросу.[9]

Заключительный этап ранжирования проходит с использованием модели ColBERT. Она работает на ограниченном наборе из уже N отсортированных документов по каждому из запросов.

В качестве датасет был использован CISI, в котором данные были собраны Центром изобретений и научной информации («CISI») и состоят из текстовых данных о 1460 документах и 112 связанных с ними запросах. Его назначение — использовать для построения моделей поиска информации, в которых данный запрос возвращает список документов, имеющих отношение к запросу. Файл «CISI.REL» содержит правильный список (т.е. «золотой стандарт» или «обоснование») соответствия запроса документу, и вашу модель можно сравнить с этим «золотым стандартом», чтобы увидеть, как она работает. Файл CISI.ALL состоит из 1460 «документов», каждый из которых имеет уникальный идентификатор (.I), название (.T), автора (.A), аннотацию (.W) и список перекрестных ссылок на другие документы (.X). Это набор данных для обучения моделей при использовании в сочетании с запросами (CISI.QRY). В CISI.QRY содержатся 112 запросов, каждый из которых имеет уникальный идентификатор (.I) и текст запроса (.W). CISI.REL содержит сопоставление идентификатора запроса (столбец 0) с идентификатором документа (столбец 1). Запрос может быть сопоставлен нескольким идентификаторам документа. Этот файл содержит «основную информацию», которая связывает запросы с документами, которая используется для обучения и тестирования алгоритма.

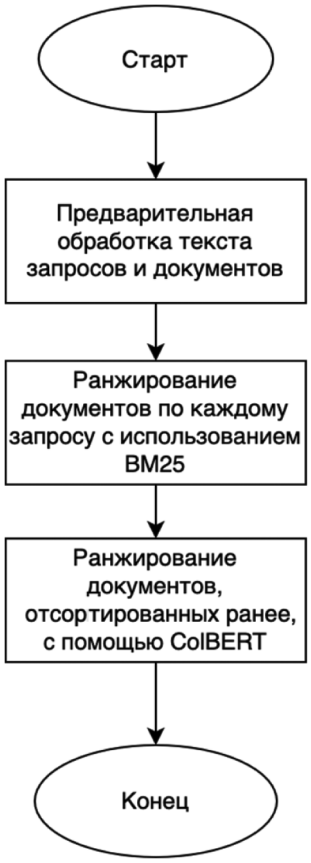

Рис. 1. Алгоритм

Результаты

Первый этап ранжирования

Для первого этапа ранжирования выбрано семейство функций BM25 [3].

Таблица 1.

Результаты первого этапа ранжирования

Функция	BM25	BM25+	BM25L
Precision	0.4	0.43	0.44

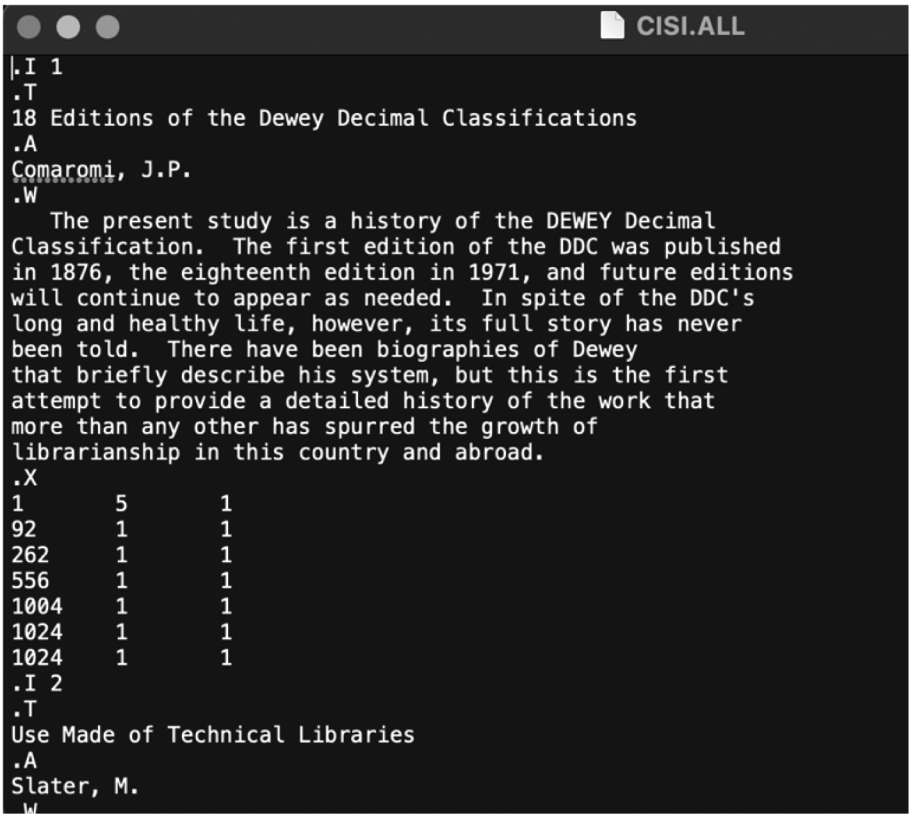

Рис. 2. Файл CISI.ALL

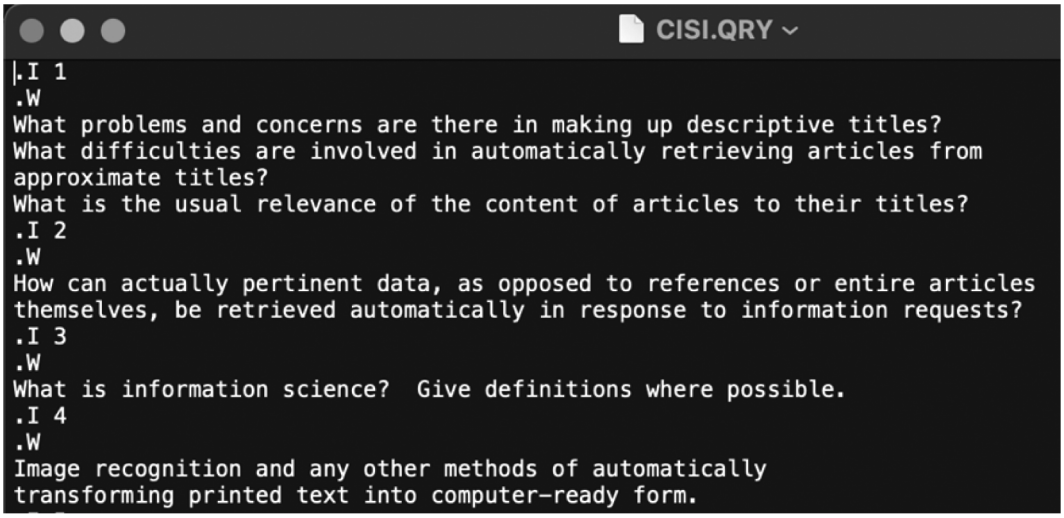

Рис. 3. Файл CISI.QRY

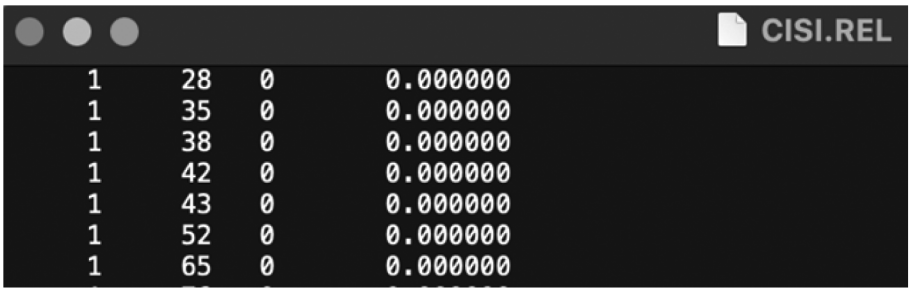

Рис. 4. Файл CISI.REL

Второй этап ранжирования

Для следующего этапа ранжирования была выбрана модель ColBERTv2. Она полагается на детальное контекстуальное взаимодействие в конце: он кодирует каждый фрагмент в матрицу вложений на уровне токенов (показано выше синим цветом) [4]. Затем во время поиска он вставляет каждый запрос в другую матрицу (показана зеленым цветом) и эффективно находит фрагменты, которые контекстуально соответствуют запросу, используя масштабируемые операторы векторного подобия (MaxSim).

Таблица 2.

Результаты второго этапа ранжирования

Значение N	Precision
5	0.62
10	0.55
20	0.5

Обсуждение

«Взаимодействие» относится к процессу оценки релевантности между запросом и документом путем сравнения их представлений. «Запоздалое взаимодействие» — это суть ColBERT. Этот термин основан на архитектуре модели и стратегии обработки, в соответствии с которой взаимодействие между запросом и представлением документа происходит на поздней стадии процесса, после того как они были независимо закодированы. Это контрастирует с моделями «раннего взаимодействия», в которых вложения запросов и документов взаимодействуют на более ранних этапах, как правило, до или во время их кодирования моделью. Более поздние модели взаимодействия, такие как ColBERT, оптимизируют эффективность и масштабируемость, позволяя предварительно вычислять представления документов и используя более легкий этап взаимодействия в конце, который фокусируется на уже закодированных представлениях. Такой выбор конструкции сокращает время поиска и требования к вычислительным ресурсам, что делает его более подходящим для обработки больших коллекций документов.

Заключение

Использование 2 видов ранжирования в поиске оправдано возрастающей точностью поиска. Более простые функции следует использовать на первом этапе, чтобы отсеять самые нерелевантные варианты, при этом не затратив значительных вычислительных мощностей. На втором этапе, имея набор документов, предварительно отсортированный, имеет смысл использовать более ресурсозатратные способы ранжирования, например языковые модели. Одной из таких моделей является ColBERT, которая сама является оптимизацией аналогичных решений с «ранними» вычислениями.

ЛИТЕРАТУРА

1. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. — М.: Вильямс, 2011. — С. 23.
2. А.В. Назаров. — М.: Издательство ИПЦ «Маска», 2019. — 480 с. — ISBN 978-5-6043822-8-2.
3. Витухновская А.А. Обучение технологии и стратегии информационного поиска на основе дифференциальных признаков информационно-поисковых систем// Информационное общество. 2013. № 1–2. С.69–79.
4. Тультаева И.В. Возможности применения поисковых систем для оптимизации сбора информации в сети Интернет // Бизнес. Образование. Право. 2014. № 4. С. 111–117.
5. Карпович С.Н. Многозначная классификация текстовых документов с использованием вероятностного тематического моделирования ml-PLSI / С.Н. Карпович // Труды СПИИРАН. — СПб.: СПб ФИЦ РАН, 2016
6. Галеева И.С. Интернет как инструмент библиографического поиска. М.: Профессия, 2007.С. 256.
7. Авдюшина, А.Е. Применение методов географического информационного поиска для анализа новостных данных // А.Е. Авдюшина, Ю.А. Королёва, Т.А. Маркина и др. // Информационные и математические технологии в науке и управлении. — 2024. № 1 (33). — 154–165. — https://www.imt-journal.ru/archive/public/article?id=363 (дата обращения: 25.12.2024). doi: 10.25729/ESI.2024.33.1.014
8. Калмыков, М.А. История развития поисковых систем и алгоритмы их работы / М.А. Калмыков, О.В. Медникова. — Текст: электронный // Научные известия. — 2022. — № 28. — С. 33–35. — URL: www.elibrary.ru/download/elibrary_49030696_47350442.pdf (дата обращения: 22.12.2024).
9. Мизаев, М.М. Поисковые сайты и технологии поиска информации / М.М. Мизаев, М.С.М. Мусаева. — Текст: электронный // Экономика: вчера, сегодня, завтра. — 2022. — Т. 12, № 4-1. — С. 203–209. — URL: www.elibrary.ru/download/elibrary_49185941_75206277.pdf (дата обращения: 02.12.2024).
10. Федотов, А.М. Проблемы поиска информации: история и технологии. — Текст: электронный / А.М. Федотов, В.Б. Барахнин // Вестник Новосибирского государственного университета. Серия: Информационные технологии. — 2009. — Т. 7, № 2. — С. 3–17. — URL: www.elibrary.ru/download/elibrary_12916756_44907357.pdf (дата обращения: 22.12.2024).

© Максименко Олег Евгеньевич (skipperprivate@yandex.ru); Боронина Юлия Сергеевна (Borona05@mail.ru)
Журнал «Современная наука: актуальные проблемы теории и практики»