УДК 004.912

ИСПОЛЬЗОВАНИЕ АЛГОРИТМОВ ПОИСКА ПО КОНТЕКСТУ ДЛЯ РЕШЕНИЯ ПРОБЛЕМЫ НЕСОВПАДЕНИЯ СЛОВАРЕЙ ЗАПРОСОВ И ДОКУМЕНТОВ В ИНФОРМАЦИОННОМ ПОИСКЕ

Горшков А.Д., Абрамович Р. К., Добрынин В.Ю. (Университет ИТМО)
Научный руководитель – доцент, кандидат технических наук, Платонов А.В.
(Университет ИТМО)

Введение. В настоящий момент существующие системы информационного поиска используют двухэтапную архитекту. Специальные алгоритмы, модели машинного обучения, нейронные сети, в том числе и трансформеры используются в таких системах с целью извлечения наибольшего количества документов, которые удовлетворяют запросу пользователя. При поиске документов по запросу необходимо учитывать не только слова (токены), но и их контекст. В данной работе рассмотрены способы решения проблемы несовпадения словарей запроса и документа с помощью алгоритмов, учитывающих контексты слов.

Основная часть

В современных системах информационного поиска для нахождения релевантных документов используются методы представления токенов документа и запроса в виде векторов (эмбеддингов) в одном векторном пространстве [1]. В таком пространстве эмбединги документов, которые удовлетворяют запросу пользователя, находятся ближе к эмбеддингу запроса, чем остальные документы. Документы ранжируются по рангу релевантности от наиболее релевантного к наименее релевантному.

Однако объемы данных увеличиваются с каждым годом, по этой причине перебор и поиск документов исключительно по векторам занимал бы большое количество времени и сильно нагружал бы систему. Для решения данной проблемы используется обратный индекс. Обратный индекс – это структура данных, в которой каждому токену соответствуют документы, в которых этот токен встречается.

С помощью данного решения можно существенно увеличить скорость поиска информационной системы.

Описанный выше подход решает одну из самых важных проблем информационного поиска – проблему производительности [2]. Однако существует высокая вероятность того, что пользователь не всегда использует в своем запросе именно те токены, которые встречаются в документах. Пользователь может использовать синонимы, область и интересующая сфера может быть размытой, а также сам пользователь может не до конца понимать какой результат он хочет получить.

Для решения вышеперечисленных проблем используются методы контекстного поиска. Такие методы рассчитаны на расширение границ запросов пользователя и помогают сделать поиск более гибким. Как говорилось выше, для поиска релевантных документов по запросу используется поиск ближайшего элемента в векторном пространстве. Для того, чтобы преодолеть проблему различных контекстов используются модели нейронных сетей, которые преобразуют весь документ в единый вектор. Аналогичным образом производится преобразование запроса. Затем для

вектора запроса отбираются наиболее близкие вектора документов. Такой способ называется “поиском по глобальному контексту”.

Другие модели нейронных сетей рассматривают определенную область вокруг выбранного слова как локальный контекст (окно токенов) [3]. Только этот локальный контекст кодируется в эмбеддинг и используется для вычисления релевантности.

Несмотря на преимущества обоих способов у них существуют и некоторые недостатки. Так метод глобального контекста не учитывает связь отдельных токенов запроса и документа, а успех от применения метода локального контекста сильно зависит от ключевого токена в окне.

Выводы. В данном тезисе рассмотрены алгоритмы, призванные решить проблему несовпадения словарей запросов и документов. Решение данной проблемы позволяет повысить качество поиска. Так, метод локального контекста был выбран для решения проблемы несовпадения словарей в исследуемом, на данный момент, алгоритме. Обратный индекс используется для повышения производительности и снижения потребляемой памяти. Результаты исследования будут описаны в дальнейших работах.

Список использованных источников:

1. Qi, Y., Zhang, J., Xu, W. et al. Salient context-based semantic matching for information retrieval // EURASIP J. Adv. Signal Process. 2020, 33.

2. Bhaskar Mitra, Eric Nalisnick, Nick Craswell, and Rich Caruana. 2016. A dual embedding space model for document ranking // arXiv preprint arXiv:1602.01137. – 2016.

3. Hongyin Tang, Xingwu Sun, Beihong Jin, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. Improving Document Representations by Generating Pseudo Query Embeddings for Dense Retrieval // arXiv preprint arXiv:2105.03599. – 2021.