УДК 004.422.81

АЛГОРИТМЫ ВЕКТОРНОГО ПОИСКА В ЗАДАЧЕ ТЕКСТОВОГО ПОИСКА

Бабаянц А.А. (Университет ИТМО), **Томилов Н.А.** (Университет ИТМО), **Туров В.П.** (Университет ИТМО)

Научный руководитель – кандидат технических наук, доцент Платонов А.В. (Университет ИТМО)

Введение. Количество данных, генерируемых человечеством, стремительно растет. Генерируемые данные имеют множество форматов, включая текстовый. Классический полнотекстовый поиск предполагает построение обратного индекса, однако это требует точного поискового запроса. Алгоритмы с применением глубокого обучения способны формировать векторные представления, позволяющие отображать семантическую составляющую документов, представленных в виде плотных векторов (dense embedding), что позволяет находить релевантные документы не только по точному совпадению, но и по близким по смыслу словам [1]. Однако при таком подходе теряется возможность построения инвертированного индекса, в связи с чем возникает необходимость либо полного перебора векторных представлений всех документов, либо применения алгоритмов приближенного поиска, увеличивающих скорость поиска в ущерб его точности. Существует множество таких алгоритмов, каждый из которых имеет достоинства и недостатки.

Основная часть. В докладе предлагается сравнение алгоритмов класса приближенного поиска ближайших соседей на основе векторных представлений, полученных из текстового датасета MS MARCO [2].

Для каждого документа из датасета было получено несколько векторных представлений (эмбеддингов) с помощью нескольких языковых моделей [3], использующих разные метрики расстояния, такие как евклидово расстояние, косинусное расстояние и скалярное произведение. Ввиду большего количества комбинаций параметров, полученные базы данных занимают значительный объем доступного дискового пространства. Для решения данной проблемы было проведено «сжатие» путем кластеризации векторных представлений и поиска десяти центроидов, после чего полученные эмбеддинги были проиндексированы различными алгоритмами векторного поиска, включая HNSW [4] и Annoy [5].

Каждый из алгоритмов был протестирован на скорость (latency) и точность (recall) поиска относительно полного перебора всех векторных представлений, что позволяет оценить выигрыш в скорости и потери точности при использовании алгоритмов векторного поиска. Помимо этого также была измерена точность поиска по текстовым документам относительно датасета MS MARCO, что позволяет оценить применимость самого подхода с применением векторных представлений документов по сравнению с применением Elasticsearch [6].

Выводы. В данной работе произведено сравнение нескольких алгоритмов неточного векторного поиска в применении к задаче полнотекстового поиска. Полученные результаты будут использованы для дальнейшей работы в области улучшения алгоритмов векторного поиска.

Список использованных источников:

1. Маннинг К.Д., Шютце Х. Введение в информационный поиск. – М.: Вильямс, 2020. – 528 с.

2. MS MARCO [Электронный ресурс]. – Режим доступа: https://microsoft.github.io/msmarco/ (дата обращения: 30.01.2023).
3. Pretrained Models - Sentence Transformers [Электронный ресурс]. – Режим доступа: https://www.sbert.net/docs/pretrained_models.html (дата обращения: 30.01.2023).
4. Yu. A. Malkov, D. A. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/1603.09320 (дата обращения: 24.01.2022).
5. spotify/annoy: Approximate Nearest Neighbors in C++/Python optimized for memory usage and loading/saving to disk [Электронный ресурс]. – Режим доступа: https://github.com/spotify/annoy (дата обращения: 24.01.2022).
6. What is Elasticsearch? [Электронный ресурс]. – Режим доступа: https://www.elastic.co/what-is/elasticsearch (дата обращения: 30.01.2023).