Построение поисковой системы, учитывающей контекстное вхождение общих между запросом и документами слов

И. В. Тарлинский[1]

Представлен подход к построению поисковой системы, учитывающей как контекст, так и вхождение общих между запросом и документами слов.

Ключевые слова: поисковые системы.

На сегодняшний день одним из самых эффективных способов построения поисковой системы является отображение поисковых единиц в векторное пространство $V = \mathbb{R}^n$ фиксированной размерности n. Такой подход применяется к данным совершенно разной модальности: картинкам, текстам, аудио-дорожкам и т.д. После вложения данных в евклидово пространство, поиск осуществляется по вектору во вложенном конечномерном пространстве V, используя для этого различные варианты алгоритма «ближайшего соседа» [1].

Очевидно, что при таком «общем» подходе к проектированию системы точность любого поиска определяется всего двумя параметрами: качеством вложенных векторов и точностью алгоритма поиска в пространстве V. Алгоритмы «поиска ближайшего соседа» инварианты относительно модальности данных и могут быть исследованы отдельно от самой задачи поиска [1].

В свою очередь задача информационного поиска часто решается с учетом структуры самих данных.

Так, например, в реализации поиска по текстовым документам хорошо себя зарекомендовал алгоритм ранжирования BM25 (англ. «Best Match») [3], использующий в основе своей работы структуру обратного инвертированного индекса, сопоставляющую каждому слову документы, в которых это слово встречалось. При таком подходе сложность обработки запроса растет нелинейно по количеству документов и коррелирована с распределением слов по документам, что очень эффективно и быстро. В то же время, огромным недостатком этого подхода является неспособность понимать «контекст» запроса и выдавать релевантные документы,

[1] *Тарлинский Игорь Викторович* — DL NLP Engineer, Myna Labs, e-mail: itarlinskiy@gmail.com.

Tarlinskiy Igor Viktorovich — DL NLP Engineer, Myna Labs.

не содержащие общих слов с запросом. Кроме того, часто бывает и обратная ситуация, при которой выдаются документы, имеющие общие слова с запросом пользователя, но даже близко не похожие на ответ. Особенно сильно вышеупомянутая проблема характерна для морфологически сложных языков, в которых нередко встречаются омонимы.

С другой стороны, для учета контекста документов, с целью различать слова омонимы и сделать поиск более осмысленным существуют не одна языковая модель, позволяющая моделировать представление текстов, учитывая контекст. Одной из последних моделей стала нейронная сеть BERT [2], отображающая документы d_i, длиной не более $d_i \leq 512$ слов в $\mathbb{R}^n$, $n = 768$ (базовая версия). Однако нетрудно заметить, что какой бы точной и качественной не была языковая модель, с ростом количества данных, поиск «по вектору» будет все менее и менее точным. Помимо ограничений, накладываемых на точность, скорость поиска теперь пропорциональна количеству документов. Возникает вопрос: можно ли совместить «контекстность» и точность нейронного подхода с эффективностью структуры обратного инвертированного индекса?

В докладе будет представлен подход, позволяющий по запросу Q рассматривать только документы $D = \{d_i\}_{i=1}^{n}$, имеющие с запросом общие слова, что позволит выиграть в скорости, но в то же самое время, учитывать контекст, в котором то или иное слово встречается в документе. В предложенной модели семантика и контекстное представление будут основаны на уникальном способе моделирования текстовых данных нейронной сетью BERT [2]. Расскажем про преимущества и недостатки такого подхода, в частности: что делать, если в запросе есть слово, которого не было ни в одном документе.

Отдельное спасибо моему научному руководителю к.ф.-м.н. Крейнес Е.М. за помощь при подготовке данного доклада, советы и консультации по исследованиям в области обработки текстовых данных на естественном языке.

Список литературы

[1] Andoni A., Indyk P., Razenshteyn I., “Approximate Nearest Neighbor Search in High Dimensions”, 2018, 27 pp., arXiv: 1806.09823

[2] Devlin J., Chang M.-W., Lee K., Toutanova K., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, 2018, 16 pp., arXiv: 1810.04805

[3] Pérez-Iglesias J., Pérez-Agüera J. R., Fresno V., Feinstein Y. Z., “Integrating the Probabilistic Models BM25/BM25F into Lucene”, 2009, 6 pp., arXiv: 0911.5046

The search system with both exact lexical matching and contextualized word representations common in between query and documents

Tarlinskiy I. V.

An approach to building a search engine that takes into account both the context and the occurrence of common words between the query and documents is presented.

Keywords: search engines.

References

[1] Andoni A., Indyk P., Razenshteyn I., "Approximate Nearest Neighbor Search in High Dimensions", 2018, 27 pp., arXiv: 1806.09823

[2] Devlin J., Chang M.-W., Lee K., Toutanova K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2018, 16 pp., arXiv: 1810.04805

[3] Pérez-Iglesias J., Pérez-Agüera J. R., Fresno V., Feinstein Y. Z., "Integrating the Probabilistic Models BM25/BM25F into Lucene", 2009, 6 pp., arXiv: 0911.5046