Мультиметрические методы информационного поиска

Сендерович Никита Леонидович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: senderovich.nikita@yandex.ru

В данной работе произведено исследование методов поиска информации в корпусе коротких текстов по пользовательским запросам. Рассматриваются свойства различных функций сходства между текстами и методы их комбинирования, позволяющие компенсировать недостатки отдельно взятых таких функций за счёт более полного использования метрической информации.

Цель работы — разработка, теоретическое исследование и эмпирическое сравнение моделей информационного поиска, основанных на обучении комбинаций функций сходства различных типов.

Задача информационного поиска рассматривается как задача ранжирования коротких корпусных текстов на основе рассчитанных значений близости каждого текста заданному запросу. Способ комбинирования функций сходства определяется формулой ранжирования. Проведено эмпирическое сравнение способов подбора оптимальной формулы с помощью методов машинного обучения разных типов: логистической регрессии, градиентного бустинга, случайных лесов. Также проанализированы модели ранжирования на основе отдельно взятых функций сходства.

В ходе исследования программно реализованы методы расчёта функций сходства, основанные на использовании различных внутренних представлений текстов. Используются как классические методы расчёта близости [1], так и современные способы, основанные на нейросетевых моделях языка [2].

Результаты работы моделей валидированы на двух наборах данных: открытом наборе Web Answer Passages (`https://ciir.cs.umass.edu/downloads/WebAP/`), а также специально собранном и экспертно размеченном корпусе научных публикаций. Оба набора состоят из множества вопросов, для каждого из которых в сопутствующем корпусе текстов размечены предложения, являющиеся ответом на заданный вопрос.

Путём измерения показателей качества моделей на тестовых выборках показано, что модель логистической регрессии в совокупности со специально подобранным набором функций сходства позво-

ляет достичь наилучших результатов поиска.

Реализован программный стенд для анализа результатов работы моделей, представляющий собой вопросно-ответную систему: по заданному пользователем вопросу с помощью настроенной модели в корпусе отыскиваются релевантные документы и отрывки текста, содержащие ответ.

По результатам исследования подана публикация.

Литература

1. Gomaa, W. H., Fahmy, A. A. A survey of text similarity approaches. // In International Journal of Computer Applications, 2013, 68(13).
2. Kusner M. J., Sun Y., Kolkin N. I., Weinberger K. Q. From word embeddings to document distances // In Proceedings of the 32nd International Conference on Machine Learning, 2015, P. 957–966.