УДК 004.912

Использование обучаемых разреженных векторных представлений документов в задаче информационного поиска

Добрынин В.Ю. (Университет ИТМО)
Научный руководитель – к.т.н., доцент Платонов А.В.
(Университет ИТМО)

Современные решения задачи информационного поиска имеют двухэтапную архитектуру. Долгое время на первом этапе использовались модели, основанные на терминах, имеющие недостаток в виде несовпадения словарей запроса и документа. Для помещения запроса и документа в одно семантическое пространство применяются их разреженные векторные представления. В работе рассмотрены алгоритмы решения задачи поиска, использующие разреженные векторные представления, а также их преимущества и недостатки.

Введение. Задача удовлетворения информационной потребности пользователя является центральной в информационном поиске. Объем данных увеличивается с каждым годом, как и темпы роста, в связи с этим необходимо улучшать скорость и качество их обработки. Основным драйвером развития указанной области выступает веб-поиск, который стал незаменимым инструментом для получения информации.

Ввиду требования искать наибольшее количество релевантных документов за короткий промежуток времени, современные поисковые системы имеют двухэтапную архитектуру. На первом этапе из большого массива документов извлекаются кандидаты с использованием менее ресурсоемкой модели, в основе которой лежит обратный индекс. На втором этапе используются более точные, но медленные модели, чтобы сократить набор документов, полученных на первом этапе, и оставить наиболее подходящие. Этот подход широко используется как в академической, так и в промышленной сфере. Более того, он достиг передовых результатов. В течение длительного промежутка времени внимание исследователей было приковано ко второму этапу, но в последнее время появляется все больше статей, в которых исследуются способы извлечения большего количества релевантных документов на первом этапе, этому посвящена и данная работа.

Долгое время для первого этапа поиска использовались модели основанные на терминах, показавшие высокую эффективность за счет простой логики и использования обратного индекса. Однако этот подход имеет недостатки в виде несовпадения словаря запроса/документа и отсутствия учета порядка терминов в документах. Таким образом, модели на основе терминов могут потерять часть релевантных документов в самом начале. Чтобы избавиться от этих ограничений, были предложены различные решения, такие как словари синонимов или расширение запросов и документов, но это не привело к значительным результатам.

Для решения этой проблемы можно обучить нейронную сеть, которая будет кодировать запрос и документ так, чтобы они находились в одном семантическом пространстве. Чаще всего такие пространства являются плотными представлениями, но они алгоритмически сложны, поскольку при поиске необходимо перебирать векторы документов. В связи с этим было принято решение о представлении документов и запросов в векторных пространствах большей размерности (разреженных представлениях) с дальнейшим их применением к обратному индексу.

Основная часть. Существует множество методов поиска и постоянно появляются новые. В работе рассмотрены несколько алгоритмов, появление которых характеризует развитие использования разреженного представления.

SNRM

В основе модели лежит нейронная сеть, архитектура которой построена таким образом, что эмбеддинги запросов и документов имеют общие параметры и одно семантическое пространство. Из полученных разреженных представлений строится инвертированный индекс, который используется для получения документов на запрос. Однако у SNRM есть недостаток в виде потери интерпретируемости исходных терминов.

SparTerm

Модель состоит из двух компонентов: предиктора важности и стробирующего контроллера. Предиктор генерирует плотный вектор, представляющий семантическую важность каждого элемента словаря. Контроллер генерирует двоичный вектор с целью контроля того, какие термины должны появляться в окончательном разреженном представлении. Модель конвертирует исходный текстовый отрывок/документ в разреженное представление. У этой модели также есть ограничения: поскольку механизм стробирования изучен заранее, это не позволяет модели изучить оптимальную стратегию разреживания, также модель не ограничивает доминирование некоторых терминов.

SPLADE

Этот подход основан на модели SparTerm, он позволяет улучшить ее и избавиться от недостатков. За счет добавления логарифмического насыщения в формулу оценки важности, модель ограничивает доминиривованние определенных терминов и естественным образом обеспечивает разреженность в представлениях, это позволяет не использовать стробирование, как в предыдущей модели. Также для повышения эффективности обучения используются отрицательные выборки. Таким образом, SPLADE является простым и эффективным алгоритмом для задач информационного поиска. Более того, позднее авторы представили обновленную версию SPLADE v2.

Выводы. В результате проведенного обзора были рассмотрены некоторые из алгоритмов, которые используют обучаемые разреженные векторные представления в задаче информационного поиска. Разработан стенд для проверки указанных алгоритмов на качество и скорость поиска, задержку при индексации, объем занимаемой памяти индексом. Следующим шагом исследования является работа над улучшением существующих решений.

Добрынин В.Ю. (автор) Подпись

Платонов А.В. (научный руководитель) Подпись