Разведочный поиск на основе тематического моделирования

Еремеев Максим Алексеевич[1]
Янина Анастасия Олеговна[2]

1: *Студент, факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия*
2: *Аспирант, факультет инноваций и высоких технологий МФТИ, Москва, Россия*

E-mail: maks5507@yandex.ru, yanina@phystech.edu

Классические техники информационного поиска ориентированы на решение задач поиска по четко сформулированному запросу (known-item search). Техника тематического разведочного поиска (exploratory search), представленная в работе, применяется, когда запрос пользователя нельзя представить в кратком виде, например, при изучении неизвестной научной области.

В основе алгоритма тематического поиска лежит теория вероятностного тематического моделирования (probabilistic topic modeling), которая позволяет описывать документы в коллекции текстов дискретным распределением на множестве тем. Такое распределение представляет из себя нормированный неотрицательный вектор, который называют тематическим. Метод аддитивной регуляризации тематических моделей (ARTM), базирующийся на максимизации линейной комбинации логарифма правдоподобия и нескольких регуляризаторов, позволяет эффективно строить различные тематические модели, в том числе иерархические, которые рекурсивно делят темы на подтемы. Иерархическая тематическая модель используется для получения нескольких векторных представлений документа по темам разных уровней иерархии.

В работе исследованы два алгоритма поиска релевантных документов из сформированной коллекции текстов, когда поисковым запросом является один или несколько текстов объема примерно на один лист А4. Если запрос представлен одним документом, то первый алгоритм предполагает построение тематических векторов документов коллекции и запроса, а затем поиск ближайших векторов документов коллекции к тематическому вектору запроса. Если же докумены в запросе несколько, то предлагается построить взвешенную сумму их тематических векторов, получив тематический профиль запроса и применить алгоритм для поиска по запросу из одного документа.

Второй алгоритм предполагает получение двух тематических векторов (нижнего и верхнего уровня иерархии) документа запроса с помощью двухуровневой иерархической тематической модели. Так, с помощью тематического вектора запроса нижнего уровня, предлагается выделять наиболее вероятные темы, а затем искать близкие к запросу документы коллекции по векторам верхнего уровня только среди документов коллекции, принадлежащих наиболее вероятным темам запроса.

В данной работе предложено несколько реализаций алгоритмов разведочного поиска, основанных на построении инвертированного и векторных поисковых индексов, проведено сравнение данных подходов. Также разработана архитектура, позволяющая обрабатывать несколько запросов параллельно. Написанная на Python, реализация поддерживает как запуск в формате библиотеки, так и в виде API на удаленном сервере.

Эксперименты проводились на коллекциях научных, курсовых и дипломных работ ресурса Научный Корреспондент (примерно 7 тыс. документов) и технических статей ресурса Хабрахабр (около 100 тыс. документов). Была проведена предобработка коллекции, отброшены слишком короткие и слишком длинные документы, затем каждый текст был представлен как набор слов и наиболее частых биграмм. На данных коллекциях были построены иерархические тематические ARTM модели, запущена реализация алгоритма тематического поиска, исследовано качество поиска и скорость работы представленной архитектуры.

Проведенные исследования показывают, что разведочный поиск на основе тематических моделей демонстрирует более высокие результаты по сравнению с другими подходами к решению задачи.

Литература

1. Yanina A., Golitsyn L., Vorontsov K. Multi-objective topic modeling for exploratory search in tech news // Communications in Computer and Information Science, 2017, V. 789, AINL-6: Artificial Intelligence and Natural Language Conference, St. Petersburg, Russia, 2017, P. 181–193.
2. Chirkova N., Vorontsov K. Additive regularization for hierarchical multimodal topic modeling // Journal Machine Learning and Data Analysis, 2016, V. 2, № 2, P. 187–200.
3. Vorontsov K., Potapenko A. Additive regularization of topic models // Machine Learning, 2015, V. 101, № 1, P. 303–323.