Упрощение моделей информационного поиска

Бабанин И. М., Кузнецов М. П.

В работе исследуется вопрос об упрощении моделей информационного поиска для достижения их интерпретируемости и приемлемой сложности. Предлагается приблизить базовый алгоритм с высоким качеством (метод градиентного бустинга решающих деревьев) с помощью комбинаций некоторых элементарных функций. В результате ожидается получить функцию ранжирования простой структуры, но сравнимого качества.

Ключевые слова: *Информационный поиск, генерирование функций, метод градиентного бустинга решающих деревьев.*

Введение

В работе рассматривается задача ранжирования в информационном поиске, где необходимо упорядочить множество результатов (документов, изображений или других файлов) по релевантности относительно запроса пользователя. Ранжирование имеет прикладное применение в поисковых и рекомендательных системах, социальных сетях. Используемые модели информационного поиска очень сложны, поэтому предлагается рассмотреть возможность приближения некоторой модели более простой, но без большой потери качества.

В работе [1] рассматривается подход к поиску функций ранжирования вида композиции некоторых элементарных математических функций. В результате исследования были получены функции, показывающие более высокие результаты на некоторых наборах данных, чем новейшие функции ранжирования. Таким образом можно ожидать, что можно аппроксимировать некоторую сложную модель с помощью функций такого вида.

В работе предлагается приблизить метод градиентного бустинга решающих деревьев [2] [3], котороый широко используется в существующих поисковых системах, на данных из TREC (trec.nist.gov) наборов данных с помощью линейной комбинации композиций элементарных базовых функций, построенных пользуясь идеями из [1], которые были бы корректны и отвечали эвристическим ограничениям на функции ранжирования в информационном поиске [4]. Оптимальные функции для линейной комбинации находятся с помощью жадного алгоритма таким образом, чтобы минимизировать расстояние до приближаемого метода на наборах данных. Также необходимо по возможности уменьшить количество функций в аппроксимирующей линейной комбинации для достижения большей простоты структуры, что достигается с помощью анализа распределения корреляции между значениями приближаемого метода и значениями построенных простых функций на наборах данных.

Литература

[1] Goswami, P., Moura, S., Gaussier, E., Amini, M., Maes, F.: Exploring the space of ir functions. In: Advances in Information Retrieval, pages 372–384. Springer (2014)

[2] Friedman, J.: Greedy Function Approximation: A Gradient Boosting Machine. In: The Annals of Statistics, Vol. 29, No. 5 (2001)

[3] Xu, J., Li, H.: AdaRank: A Boosting Algorithm for Information Retrieval. In: SIGIR (2007)

[4] Fang, H., Tao, T., Zhai, C.: A formal study of information retrieval heuristics. In: Proceedings of the 27th ACM SIGIR Conference (2004)