Создание ранжирующих моделей для систем информационного поиска. Алгоритм прогнозирования структуры локально-оптимальных моделей *

Лепехин М. Н., Кулунчаков А. С., Стрижов В. В.
lepehin.mn@phystech.edu

В данной работе исследуются различные методы построения нелинейных моделей для задач регрессии. В качестве возможных моделей для задач регрессии рассматривается множество моделей, представимых в виде суперпозиции более простых моделей, предлагаемых экспертами. В этой работе предлагается использование метода для прогнозирования структуры модели, которая будет представлена алгоритмом как последовательность вершин при обходе в глубину синтаксического дерева. Качество спрогнозированной ранжирующей модели проверяется как на синтетических данных, так и на основе данных из текстовой коллекции TREC. При помощи заданных метрик происходит сравнение результатов, полученных на различных данных, с уже известными моделями.

Ключевые слова: *ранжирующие системы, информационный поиск, локальнооптимальная модель, временные ряды, структурное обучение.*

1 Введение

В данной работе решается задача ранжирования текстов по поисковым запросам. Задача ранжирования актуальна для современных поисковых систем. Несмотря на то, что уже существует большая коллекция алгоритмов, решающих эту задачу и оптимизирующих некоторую заданную метрику, часто качество их работы оказывается недостаточно высоким.

Одна из наиболее важных проблем, возникающих при решении задач ранжирования текстов, - проблема переобучения. В [1, 2] приведены примеры решений этой задачи, учитывающих особенности поисковых запросов, которые столкнулись с проблемой переобучения.

Существуют модели высокого качества, позволяющие относительно эффективно бороться с переобучением. Примеры таких моделей есть в работе [3]. Они строятся как суперпозиции математических примитивов. Лучшие по качеству модели из [3] превосходят на коллекциях TREC модели из [1, 2]. Однако, для них характерная другая проблема - высокая структурная сложность. Структурной сложностью модели называется число элементов грамматики, необходимых для её описания. С возрастанием структурной сложности повышается трудность изучения модели.

Кроме того, существуют решения задачи ранжирования с использованием генетического алгоритма. Его основное преимущество - гибкость порождаемых им моделей. В работах [4, 5] эта гибкость позволила перейти к представлению порождённых моделей деревом их синтаксического разбора. Однако после 30-40 итераций генетического алгоритма сложность порождаемых моделей резко возрастает, не давая при этом существенного повышения качества.

В данном работе предлагается учесть проблемы, возникающие при решении задачи ранжирования текстов при помощи описанных выше алгоритмов. Для этого предлагается

*Работа выполнена при финансовой поддержке РФФИ, проект №00-00-00000. Научный руководитель: Стрижов В. В. Задачу поставил: Эксперт Кулунчаков А. С. Консультант: Консультант Кулунчаков А. С.

изменить реализацию генетического алгоритма, добавляя каждое фиксированное число эпох случайную модель в популяцию.

При исследовании текстов в задачах ранжирования будут использованы основные характеристики слов текста - tf (частоты слов в документе) и idf (числа документов, в которых слово встречается).

В данной работе предлагается развить идею пердсказания промежуточной мета-модели. Для этого рассматривается кластеризация документов по значениям tf-idf, посчитанным по корпусу текста. При этом предполагается, что текст разбивается на кластеры таким образом, что внутри каждого кластера документы будут близки друг к другу при ранжировании. Мета-модель предлагается строить как линейную комбинацию моделей, построенных для каждого кластера.

Литература

[1] *Salton, Gerard and McGill, Michael J.* Introduction to Modern Information Retrieval // McGraw-Hill, Inc. New York, NY, USA, 1986

[2] *Ponte, Jay M. and Croft, W. Bruce* A Language Modeling Approach to Information Retrieval // In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 275–281. ACM

[3] *P. Goswami, S. Moura, E. Gaussier, M.-R. Amini, F. Maes* Exploring the space of ir functions // ECIR'14, 2014, pp. 372–384.

[4] *Fan, Weiguo and Gordon, Michael D. and Pathak, Praveen* Personalization of Search Engine Services for Effective Retrieval and Knowledge Management // In Proceedings of the twenty first international conference on Information systems (ICIS '00). Association for Information Systems Atlanta, GA, USA, 20-34.

[5] *Fan, Weiguo and Gordon, Michael D. and Pathak, Praveen* A Generic Ranking Function Discovery Framework by Genetic Programming for Information Retrieval // Inf. Process. Manage. 40, 4 (May 2004), 587-602.