

# Создание ранжирующих моделей для систем информационного поиска.\*

*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 суперпозиций более простых функций, предлагаемых экспертами. Структура суперпозиции функций определяется с помощью обхода в глубину её синтаксического дерева. Предлагается метод прогнозирования такой структуры. Подход к порождению структуры оптимальной модели проверяется как на синтетических данных, так и на основе данных из текстовой коллекции TREC. На этих данных качество построенных суперпозиций алгебраических функций сравнивается с результатами уже известных моделей.

**Ключевые слова:** *ранжирующие системы, информационный поиск, временные ряды, структурное обучение, суперпозиция моделей.*

## 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 для решения этой задачи, часто качество их работы оказывается недостаточно высоким. Например, в работах [1, 2] приведены примеры решений задачи ранжирования, столкнувшиеся с проблемой переобучения.

Существуют алгоритмы построения моделей высокого качества, позволяющие относительно эффективно бороться с переобучением. Их примеры есть в работе [3]. Они строятся как суперпозиции элементарных функций. Под элементарными функциями в данном случае подразумеваются показательная, логарифмическая и степенная функции. Лучшие по качеству композиции алгебраических функций из [3] превосходят на коллекциях TREC модели из [1, 2]. Но проблема такого метода построения моделей в том, что он использует полный перебор. Из-за этого время построения даже относительно простых функций резко возрастает.

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 суперпозиций. 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 суперпозиций примитивов 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 предсказания 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.