

# Ранжирующие модели для систем информационного поиска. Прогнозирование структуры локально-оптимальных моделей.

Поповкин Андрей Алексеевич  
Романенко Илья Игоревич

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам  
(практика, В. В. Стрижов)/Группа 594, весна 2018

## Цель исследования

Исследовать возможность порождения ранжирующей функции при помощи генетического алгоритма и при помощи нейронной сети, сравнить полученные результаты с результатами сообщества.

## Проблема

Сложно исследовать пространство существенно нелинейных функций.

- ❶ *Kulunchakov A. S., Strijov V. V.* Generation of simple structured IR functions by genetic algorithm without stagnation // <http://strijov.com/papers/Kulunchakov2014RankingBySimpleFun.pdf>
- ❷ *Salton, Gerard and McGill, Michael J.* Introduction to Modern Information Retrieval // McGraw-Hill, Inc., New York, NY, USA, 1986
- ❸ *Gordon, M.* Probabilistic and Genetic Algorithms in Document Retrieval // Commun. ACM 31, 10 (October 1988), 1208-1218.

# Постановка задачи

Пусть  $\mathbf{C} = \{d_i\}$  - коллекция документов,  $\mathbf{Q}$  - пользовательских запросов,  $q = \{w_j\}$ . Определена функция релевантности  $r(d, q) \rightarrow \{0, 1\}$ .

Рассматриваются характеристики пары документ-слово:  $(d, w, \mathbf{C}) \rightarrow (\text{tf}, \text{idf})$ .

$$\text{idf}(w, \mathbf{C}) = \frac{\text{count}(w, \mathbf{C})}{|\mathbf{C}|}$$

$$\text{tf}(w, d, \mathbf{C}) = \text{freq}(w, d) \cdot \log\left(1 + \frac{\text{size}_{\text{avg}}}{\text{size}(d)}\right)$$

$\mathcal{T}$  - множество суперпозиций функций от  $\text{tf}$ ,  $\text{idf}$ . Будем аппроксимировать функцию  $r(d, q)$ , как функцию  $f(d, q) = \sum_{w \in d} f'(\text{tf}, \text{idf})$ , где  $f' \in \mathcal{T}$ .

$$f^* = \text{argmax}_{f \in \mathcal{T}} (\text{MAP}(f, \mathbf{C}, \mathbf{Q}) - \|f\|^2)$$

Качеством аппроксимационной функции будем считать MAP.

$$\text{MAP}(f, \mathbf{C}, \mathbf{Q}) = \frac{1}{|\mathbf{Q}|} \cdot \sum_{q \in \mathbf{Q}} \text{AvgP}(f, q, \mathbf{C})$$

$$\text{AvgP}(f, q, \mathbf{C}) = \frac{\sum_{i=0}^{|C_q|} \text{PrefSum}(r(d_i, q), k) \cdot r(d_i, q)}{\sum_{d \in C_q} r(d)}$$

## Цель эксперимента

Получение результатов, сравнимых с предыдущими работами в этой сфере. Улучшение этих результатов.

## Используемые данные

Коллекция TREC (датасеты 5-8).

<https://trec.nist.gov/data.html>

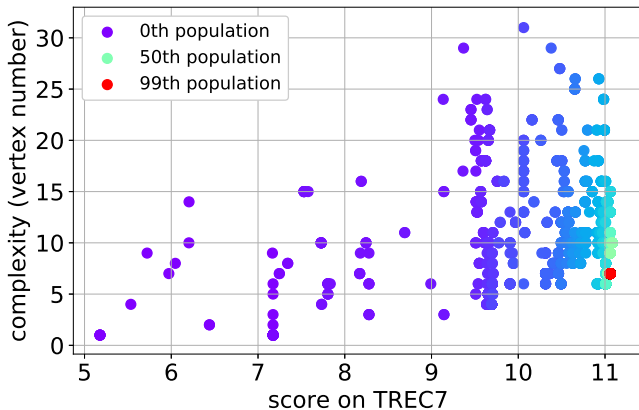
Используется генетический алгоритм с регуляризацией по числу узлов в дереве.

Со следующими процедурами:

- 1 мутация — замена произвольной вершины на заново сгенерированную.
- 2 crossover — обмен местами двух произвольных вершин деревьев.

# Вычислительный эксперимент

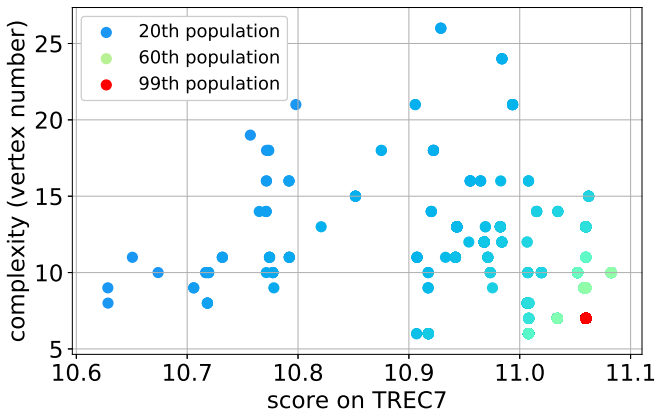
Сложность моделей в зависимости от целевой метрики.



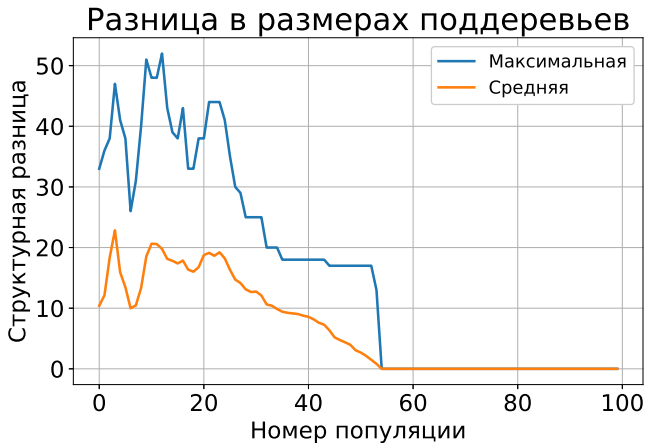


# Вычислительный эксперимент

Сложность моделей в зависимости от целевой метрики, начиная с 20-ого поколения



Метрика максимального общего поддерева для определения стагнации.



Экспертные функции:

$$f_1 = e^{\sqrt{\log(1 + \frac{tf}{idf})}}$$

$$f_2 = \sqrt[4]{\frac{tf}{idf}}$$

$$f_3 = \sqrt{idf} + \sqrt{\frac{tf}{idf}}$$

Найденные наилучшие функции :

$$h_5^* = \log\left(1 + \frac{\log(1 + \log(1 + \log(1 + \sqrt{\log(1 + tf) - \sqrt{idf}})))}{2 \cdot idf}\right)$$

$$h_6^* = \sqrt{\frac{\sqrt[4]{tf}}{2 \cdot idf}}$$

$$h_7^* = \frac{\sqrt[8]{\log(1 + tf)}}{idf}$$

Результаты при сравнении на корпусах TREC-5, TREC-6, TREC-7.

Superposition	TREC-5	TREC-6	TREC-7
Функции сообщества			
$f_1$	8.785	13.715	10.038
$f_2$	8.908	13.615	9.905
$f_3$	8.908	13.615	9.905
Найденные наилучшие функции			
$h_5^*$	<b>9.537</b>	13.762	10.584
$h_6^*$	8.903	<b>13.967</b>	10.771
$h_7^*$	8.526	13.424	<b>11.060</b>

- 1 Для каждого корпуса получили функцию наилучшим образом ранжирующую документы для данного запроса.
- 2 Улучшили существующие результаты сообщества.
- 3 Весь код доступен в репозитории на GitHub