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

Поповкин Андрей Алексеевич
Романенко Илья Игоревич

Московский физико-технический институт
Сколковский институт науки и технологий

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

Цель исследования

Цель исследования

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

Проблема

Сложно исследовать пространство существенно нелинейных функций.

Список литературы

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

Постановка задачи

Пусть $C = \{d_i\}$ - коллекция документов, Q - пользовательских запросов, $Q_i = \{w_j\}$. Определена функция релевантности $r(d, q) \to \{0, 1\}$.
Рассматриваются характеристики пары документ-слово: $(d, w, C) \to (tf, idf)$.

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

$$tf(w, d, C) = freq(w, d) \cdot log(1 + \frac{size_{avg}}{size(d)})$$

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

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

Функционал качества

Качеством аппроксимационной функции будем считать MAP.

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

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

Теоретическая часть

To be continued

Цели эксперимента

Цель эксперимента

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

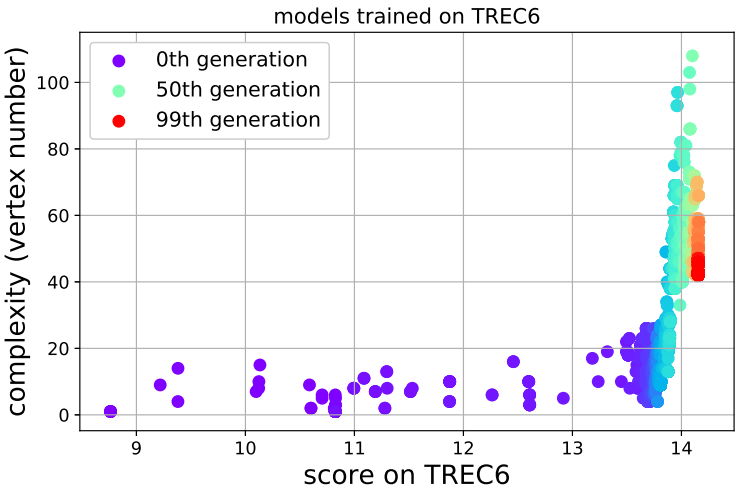
Используемые данные

Коллекция TREC (датасеты 5-8).
`https://trec.nist.gov/data.html`

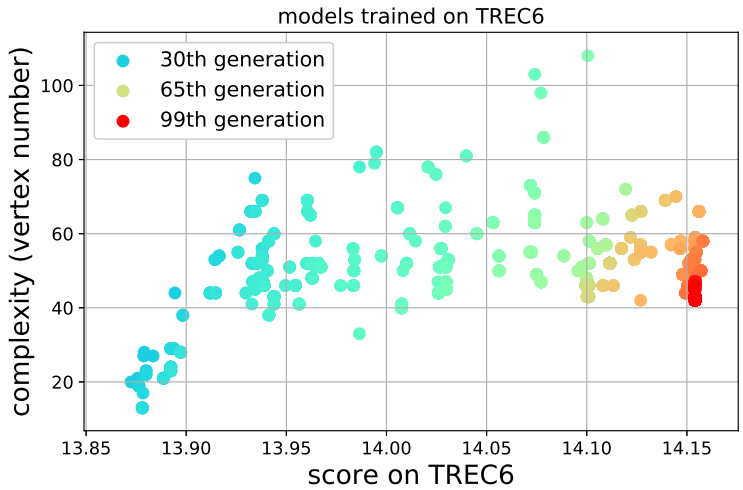
Вычислительный эксперимент

Используется генетический алгоритм с регуляризацией по числу узлов в дереве.

Вычислительный эксперимент



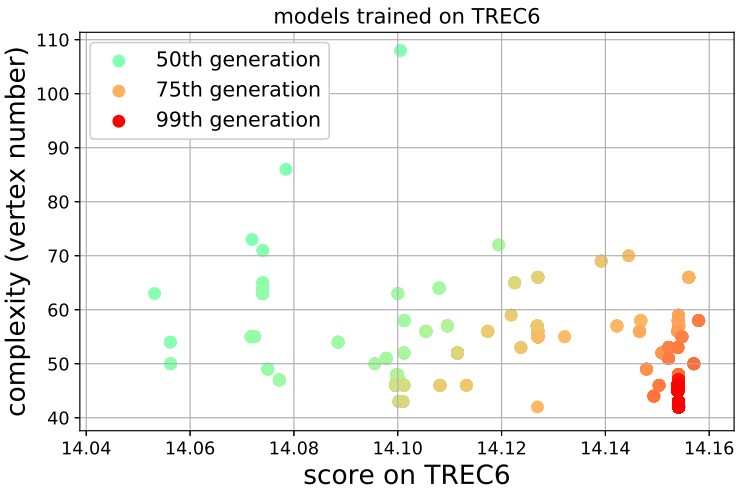
(a) Сложность моделей



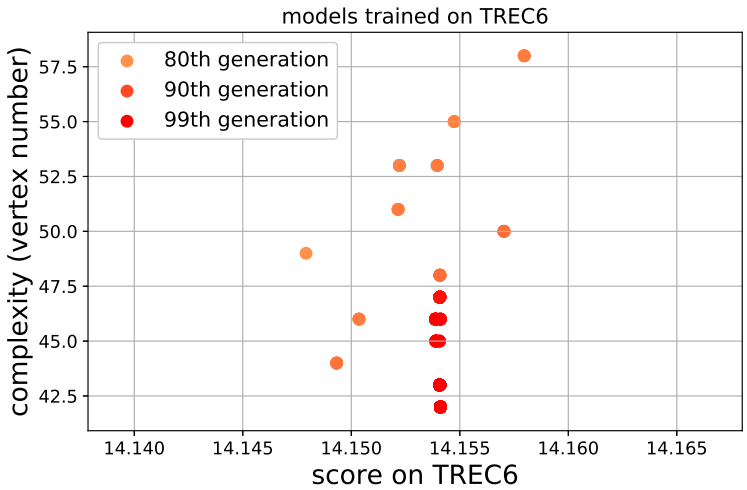
(b) Сложность моделей с 30-го поколения

Рис.: Зависимость сложности модели от целевой метрики

Вычислительный эксперимент



(a) Сложность моделей с 50-го поколения



(b) Сложность моделей с 80-го поколения

Рис.: Зависимость сложности модели от целевой метрики

Вычислительный эксперимент

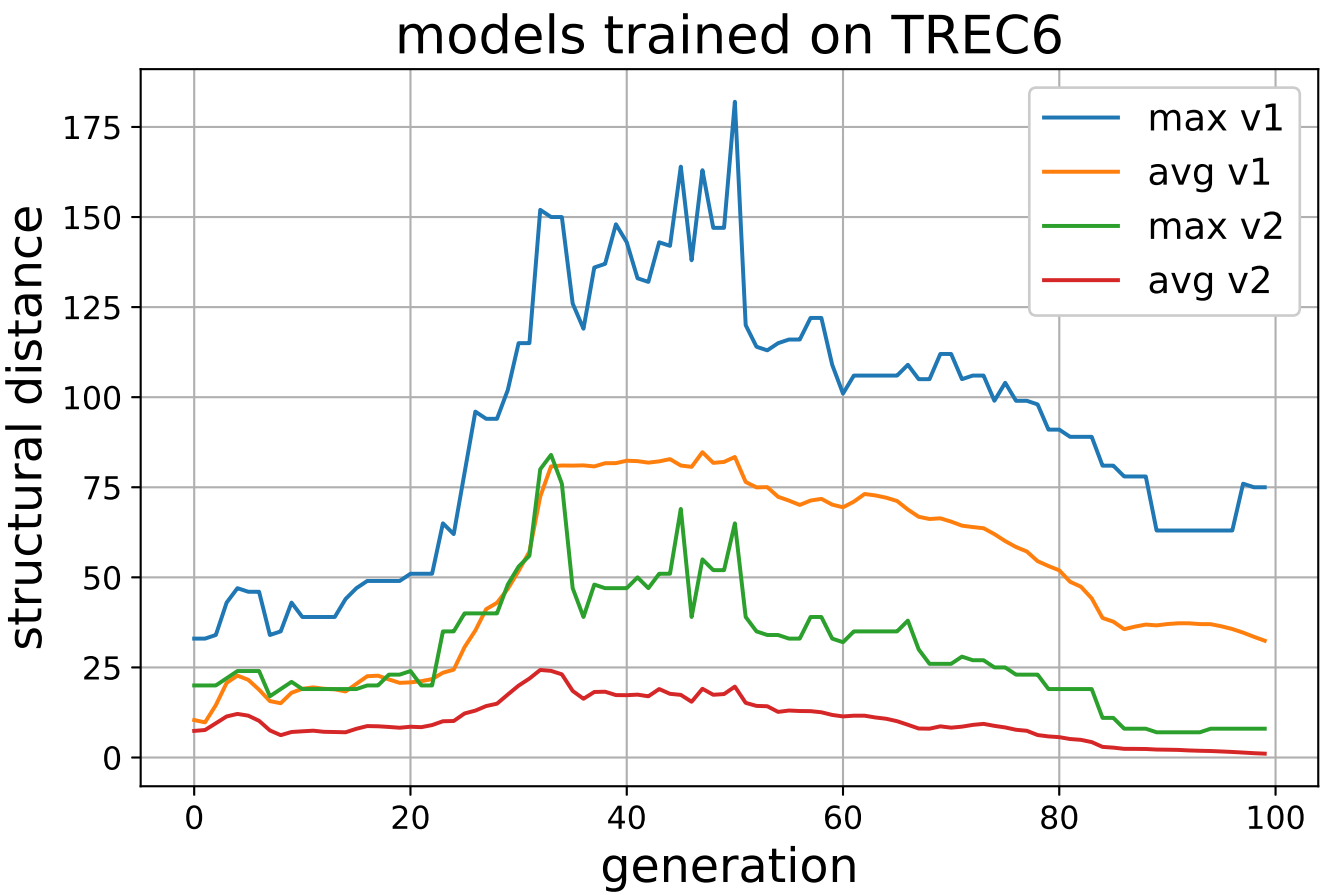


Рис.: Сравнение метрик (Левенштейн и максимальное общее поддерево) для определения стагнации

Заключение

Для каждого датасета получили функцию наилучшим образом, ранжирующую документы для данного запроса. В результате опыта для каждого датасета получили практическую границу максимально возможной точности модели при фиксированной сложности.

Вычислительный эксперимент

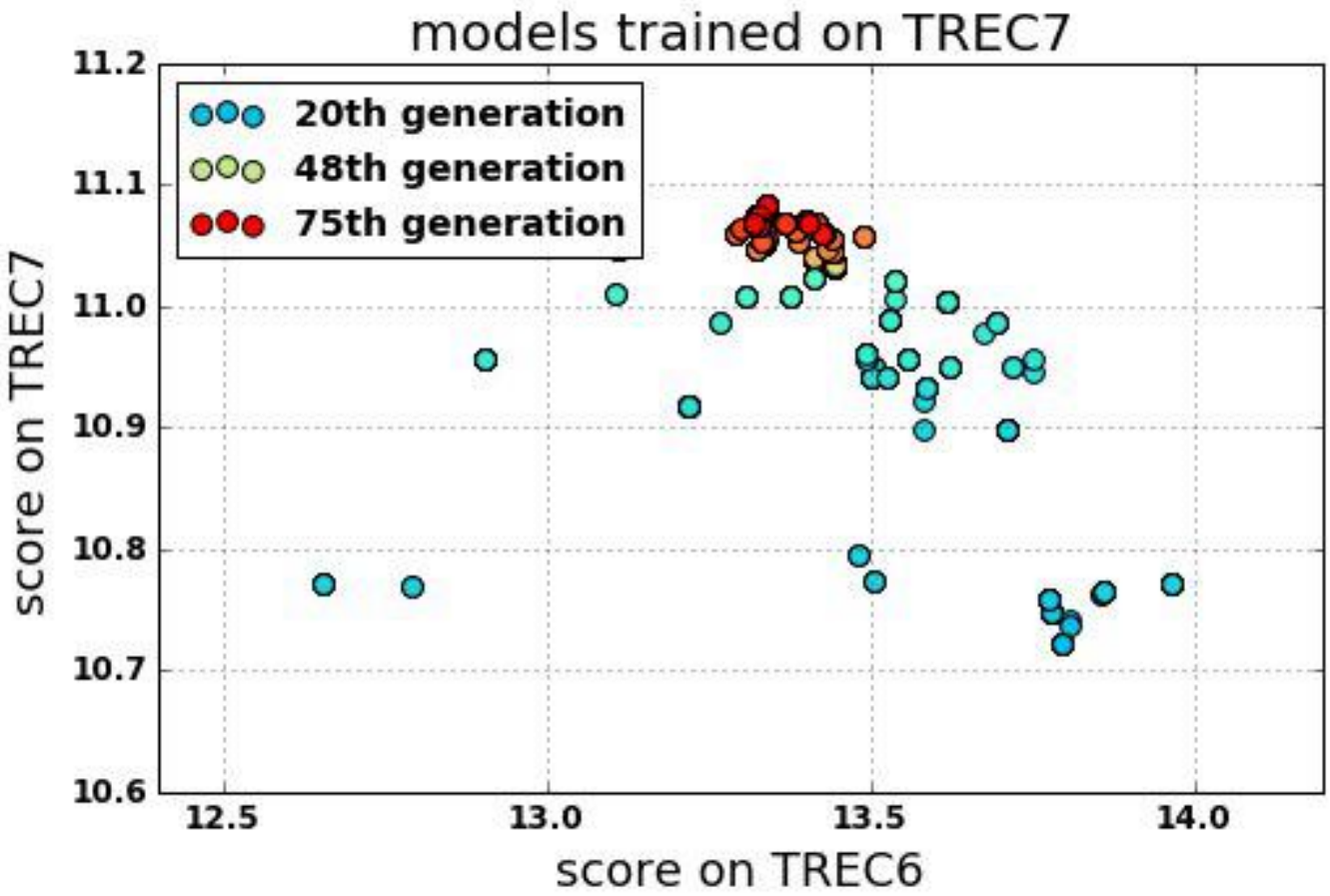


Рис.: Переобучение модели под фиксированный датасет