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

Поповкин А. А., Романенко И. И.
romanenko.ii@phystech.edu, popovkin.aa@phystech.edu
Московский физико-технический институт (государственный университет), Москва

В этой работе рассматриваются методы порождения нелинейных моделей для задач регрессии. Рассматривается пространство моделей, представимых в виде суперпозиции элементарных математических функций, предлагаемых экспертами. Предлагается алгоритм, определяющий структуру оптимальной модели для задачи регрессии. Алгоритм представляет структуру модели как последовательности вершин при DFS обходе синтаксического дерева разбора суперпозиции. Проводится проверка работы алгоритма на синтетическом примере временного ряда. Итоговый алгоритм применяется для практической задачи создания ранжирующих моделей для систем информационного поиска на примере датасетов TREC. Приводится сравнительный анализ полученных результатов с уже известными моделями, предложенными сообществом.

Ключевые слова: *информационный поиск, генетические алгоритмы, нейросети.*

Введение

Рассмотрим state-of-art задачу, когда при ранжировании используются 2 основные характеристики текста — частота слова в документе — (tf) и количество документов, в которых встречается слово — (idf). С начала развития области были изобретены различные модели, решающие данную задачу [4, 5, 6]. При их построении учитывались особенности запросов пользователей, однако эти модели сталкивались с проблемой переобучения.

Модели высоко качества так же были найдены с помощью программного перебора. Как описано в работе [7] такие модели могут быть рассмотрены как суперпозиции математических функций от базовых характеристик текста. Для моделей вводились структурная характеристика — число элементов грамматики, используемых для их описания, накладывались структурные и целевые ограничения. Лучшие модели, описанные в работе [7], превосходят по качеству на коллекциях TREC модели из работ [4, 5, 6]. В работе [7] рассматривались модели с ограничением на глубину дерева разбора. Однако более детальное исследование пространства структурно сложных суперпозиций - нетривиальная задача.

Одним из подходов к поиску оптимальной модели путем программного перебора является генетический алгоритм. Генетический алгоритм основан на идее итеративного отбора моделей, их скрещивания и мутаций. Первые попытки его использования были произведены в статьях [9, 10]. Производились поиски наилучших параметров генетического алгоритма. Как показано в статьях [11, 12], оптимальный выбор операции кроссовера может существенно улучшить порождаемые модели.

Основным преимуществом генетического алгоритма является его гибкость к используемым данным, что позволило в работе [13, 14] перейти к представлению ранжирующей модели ее деревом синтаксического разбора (обсудить). Однако базовый генетический алгоритм, описанный в работах [13, 14], подвержен стагнации и после 30-40 итераций мута-

Работа выполнена при финансовой поддержке РФФИ, проект Научный руководитель: Стрижов В. В. Задачу поставил: Кулунчаков А. О. Консультант: Кулунчаков А. О.

ций и кроссовера сложность порождаемых функций значительно возрастает и изменения в популяции становятся незначительными.

Улучшения этого метода описаны в статье [8], благодаря использованию регуляризации при оценивании порождаемых моделей, удается добиться улучшения разнообразия порождаемых фукнций (обсудить), что ведет к повышению качества итоговой модели.

В работе [15] рассматривается другой подход к аналитическому программированию. Предлагается использовать принципы глубокого обучения, разбивая задачу по уровням абстракции. В качестве промежуточного уровня предложенно использовать матрицу вероятностей переходов в дереве разбора суперпозиции, полученную обучением нейронной сети. Далее, итоговая модель строится итеративным жадным алгоритмом. (этот абзац лучше как-нибудь прочитать Вадиму Викторовичу, я слабо ориентируюсь в работе Варфоломеевой)

Предлагается альтернативная реализация генетического алгоритма со следующими изменениями <тут наши предложения по генетике>.

В качестве основной новации рассматривается развитие идеи предсказания промежуточной мета-модели <тут наши предложения>

Работа построена следующим образом. <Описание структуры>

Теоретическое введение

Дана коллекция текстовых документов $C = \{d_i\}$ и пользовательских запросов Q, каждый из которых представляет из себя множество слов $q = w_i$. Дана функция $r(d, q) \to \{0, 1\}$, определенная экспертами, и показывающая, является ли данный документ d релевантным для запроса q (1 — является).

Рассмотрим две характеристики пары документ-слово: $(d, w, C) \to (tf, idf)$. Определенных следующим образом:

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

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

где $count(w, C)$ — количество документов $d \in C$ содержащих слово w, $freq(w, d)$ — частота слова w в документе d, $size(d)$ — количество слов в d, а $size_{avg}$ среднее количество слов в документах из коллекции C.

Положим f — суперпозиция математических функций от аргументов tf и idf. Назовем моделью — дерево синтаксического разбора данной суперпозиции и рассмотрим множество всех таких деревьев $\mathcal{T}$.

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

Качеством аппроксимационной функции будем считать MAP (mean average precision).

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

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

C_q — множество документов коллекции, размеченных для запроса q. $d_{(i)}$ — i-ый документ из C_q в ряду, упорядоченному по убыванию значения $f(d_{(i)}, q)$. $PrefSum(r(d_{(i)}, q))$ — сумма первых k элементов ряда $r(d_{(i)}, q)$.

Нашей задачей является нахождение ранжирующей функции

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

где $P(f)$ — штрафная функция, ограничивающая структурную сложность суперпозиции f.

Заключение

Литература

[1] *Porter M. F.* Readings in Information Retrieval // Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997, Ch. An Algorithm for Suffix Stripping, Pp. 313–316.

[2] *Metzler, Donald and Croft, W. Bruce* A Markov Random Field Model for Term Dependencies // Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05, ACM, New York, NY, USA, 2005, pp. 472–479.

[3] *Amati, Gianni and Van Rijsbergen, Cornelis Joost* Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness // ACM Trans. Inf. Syst. 20 (4) (2002) pp. 357–389

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

[5] *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.

[6] *Clinchant, Stéphane and Gaussier, Eric* Information-based Models for Ad Hoc IR // In Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 234–241. ACM.

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

[8] *Kulunchakov A. S., Strijov V. V.* Generation of simple structured IR functions by genetic algorithm without stagnation // `http://strijov.com/papers/Kulunchakov2014RankingBySimpleFun.pdf`

[9] *Goldberg, David E.* Genetic Algorithms in Search, Optimization and Machine Learning // Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[10] *Koza, John R.* Genetic Programming: On the Programming of Computers by Means of Natural Selection // MIT Press, Cambridge, MA, USA, 1992.

[11] *Vrajitoru, Dana* Crossover Improvement for the Genetic Algorithm in Information Retrieval // Inf. Process. Manage. 34, 4 (July 1998), 405-415.

[12] *Gordon, M.* Probabilistic and Genetic Algorithms in Document Retrieval // Commun. ACM 31, 10 (October 1988), 1208-1218.

[13] *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.

[14] *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.

[15] *Варфаломеева А. А.* Методы структурного обучения для построения прогностических моделей // http://www.machinelearning.ru/wiki/images/f/f2/Varfolomeeva2013Diploma.pdf