HokieGo at 2017 PM Task: Genetic Programming based re-ranking method In Biomedical Information Retrieval

Junyan Wu, Xiaofu Ma, Weiguo Fan

[1]Virginia Tech

{wujy128, xfma, wfan}vt.edu

Abstract. *This paper summarizes our efforts on TREC 2017 Precision Medicine Track. We present a genetic programming based learning-to-rank algorithm. We perform two training experiments on 2014 and 2016 TREC CDS data and apply the pre-trained model as re-ranking method to improve the performance. In addition, two utility functions, CHK and FFP4, have been used for the training optimization.*

1. Introduction

The focus of the TREC 2017 is to treat the cancer patients in a better way by providing clinicians with the useful and precise medicine-related information. The major goal is to retrieve existing knowledge in the scientific literature and to understand the coherence of the patients and the experimental treatments. Specifically, the two challenges to be resolved are (1) how to retrieve biomedical articles for corresponding patents with the article abstract information, and (2) how to retrieve the suitable clinical trials for the corresponding patients.

The basic idea of our methodology is to apply genetic programming (GP) method on the clinical decision support task. Our approach is based on listwise learning method. By using the utility function to project the ground truth label into a linear space, the algorithm will find the best ranking function to better fit the document ranking order.

2. Methodology

2.1. Query Expansion

We first identify the UMLS concepts and extract the synonym from the “disease” and “other” text by MetaMap. To reduce the recall from MetaMap, we select certain semantics types of concepts that are relevant for the task according to the previous research [4]. We replace the age information by certain description terms, and add a certain amount of gender related synonyms into the query based on the research [1].

2.2. Query Simplification

According to previous research [5], removing the negation can significantly improve the precision. The presence of negation in medical records will significantly decrease the performance according to previous research works [3]. Thus, we remove the negations from the query.

2.3. Baseline Retrieval

After testing bunch of ranking functions on previous TREC-CDS data, we found bm25fcomb ranking function could achieve the best retrieval performance. We thus extract top 5000 most relevant documents by bm25fcomb ranking function on GALAGO platform.

2.4. Genetic Programming Based Re-Ranking Method

In GP, terminals are leaf nodes of a tree data structure which are essentially weighting features used in term weighting to capture lexical statistics. Specifically, terminals were chosen after examining various term weighting and ranking formulas.
Functions in GP are the operations such like "+", "-", "x", "/", "log" that are used to produce other trees by combining terminals and/or sub-trees. The initial population generation is a population set contains individuals which represent document term weighting formulas. We generate the initial set of trees which are constrained to have a maximum depth of four levels. The generating method we used is the ramped half-and-half method. It stipulates that half of the randomly generated trees must be generated by a random process which ensures all branches of the maximum initial depth. The remaining randomly generated trees require branches whose lengths do not exceed this depth. These constraints have been found to generate a good initial sample of trees. Then, a fitness function is used to measure the effectiveness of a ranking function represented by an individual tree is for ranking.
We use implement Reproduction and Crossover operations in our method. Reproduction copies the top rate trees in the current generation to the next generation directly without executing any genetic transformation. The reproduction rate is generally 0.1 or less. Crossover brings variations by creating trees that differ from their parents. For crossover, a method called tournament selection is used in our method. Tournament selection first selects, with replacement, k (we use 6) trees randomly from the current generation. The two trees with the highest fitness are paired to exchange sub-trees.
We stop the GP discovery process after as less as 100 generations for the following reasons. First, the simulation is highly computationally intensive. Second, our pilot experiments with sample queries indicate that 100 generations is a sufficient period to generate high-performance trees.

3. Experiments and Evaluation

3.1. Document and Query Preprocessing

We use MetaMap to perform query expansion, and add synonyms of medical concepts into the queries. We also define 4 age ranges: blow 12 as child, between 12 and 18 as teenager, between 18 and 80 as adult, above 80 as elder. We add synonyms coordinated to each age range into the queries. Furthermore, we also add gender related words into queries.

3.2. Experiement and Evaluation

The terminals we have used are the basic IR features calculated from both query and document. Details are shown in Table 1.

Because the relevance labels of each document are 0, 1, 2, an utility function has to be applied to convert the predict values into the ground truth labels. According to the previous GP based listwise learning to rank study [2], we choose FFP4 (Fig 1) and CHK (Fig 2) as our utility functions.

Table 1. Terminals used in the GP algorithm

Terminal	Meaning
tf	Number of occurrences of a term in a document
tf_max	Maximum tf in a document
tf_avg	Average tf in a document
tf_doc_max	Maximum tf in the document collection
df	Number of unique documents have a term
df_max	Maximum df for a given query
N	Total number of documents in the documents collection
length	Length of a document
length_avg	Average length of a document in the collection
R	Constant number randomly generated by the GP system
n	Number of unique terms in a document

$$\text{Fitness}_{\text{FFP4}} = \sum_{i=1}^{|D|} r(d_i) \times k_8 \times k_9^i$$

Figure 1. ffp4 utility function

$$\text{Fitness}_{\text{CHK}} = \frac{1}{|D|} \sum_{i=1}^{|D|} \left(r(d_i) \times \sum_{j=i}^{|D|} \frac{1}{j} \right)$$

Figure 2. chk utility function

The top 5000 document samples that coordinate to a query have been splitted into 2450 training samples, 1500 validation samples and 1050 test samples. The data was trained by 100 generations with 1000 populations per generation. The performances during generations are shown in Figures 3 and 4. The training experiments have been repeated for 5 times.

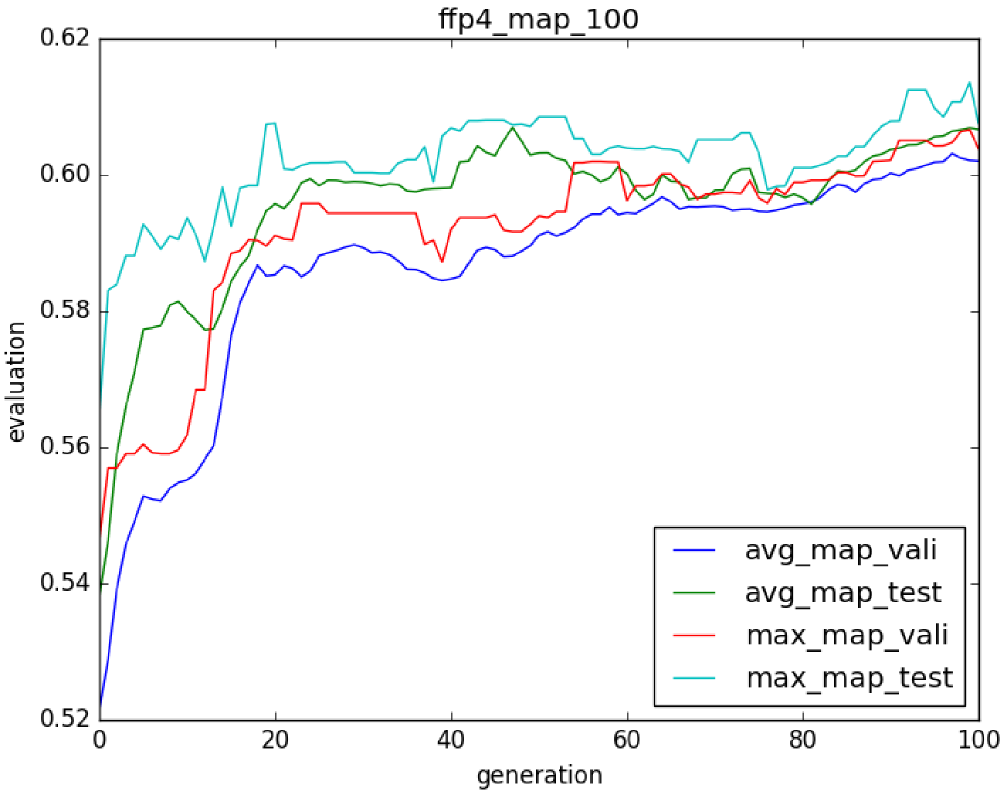

Figure 3. Mean average percision of 100 generations with ffp4 utility function

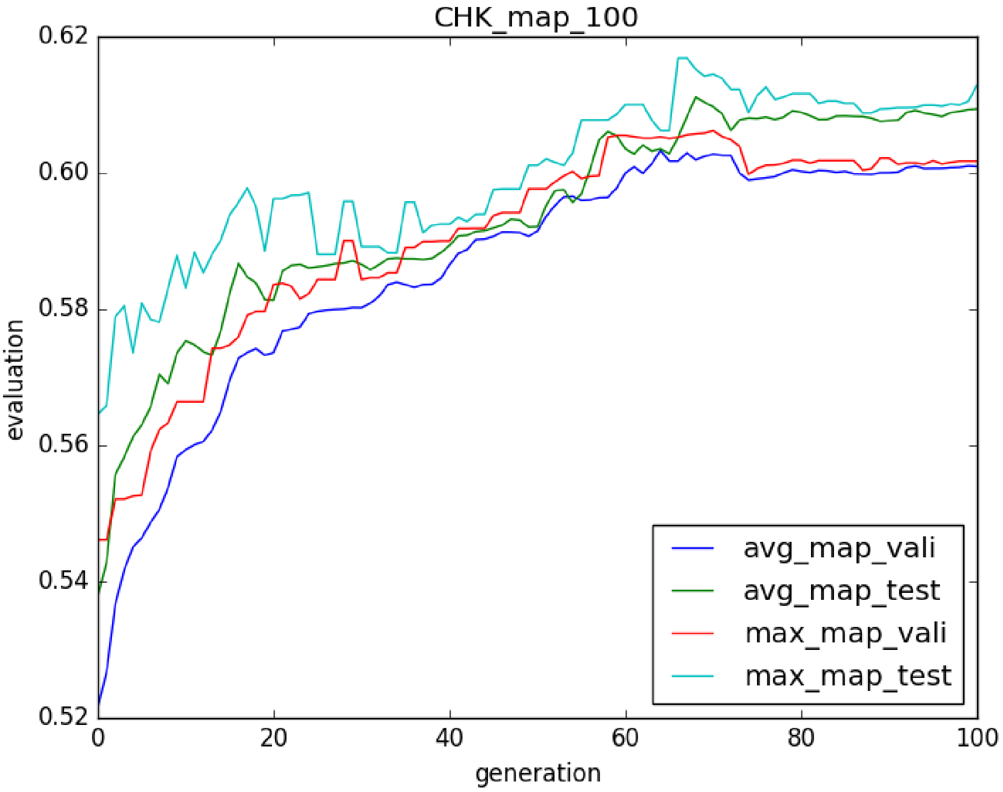

Figure 4. Mean average precision of 100 generations with chk utility function

The results show that the performance of CHK utility function grows more smoothly than ffp4.

4. Conclusion and Future Works

This paper presents our proposed genetic programming based re-ranking algorithm on TREC 2017 Precision Medicine Track. We apply the pre-trained model as re-ranking method and use two utility functions, CHK and FFP4, in our training optimization. Through the evaluation based on two training experiment sets on 2014 and 2016 TREC CDS data, we verify the improved performance of the proposed method. Future investigation on the performance tuning and theoretically boundary analysis are considered as our next steps.

References

[1] E. D'hondt, B. Grau, S. Darmoni, A. Névéol, M. Schuers, and P. Zweigenbaum. Limsi@ 2014 clinical decision support track. Technical report, NATIONAL CENTER FOR SCIENTIFIC RESEARCH ORSAY (FRANCE) COMPUTER SCIENCES LAB FOR MECHANICS AND ENGINEERING SCIENCES, 2014.

[2] W. Fan, E. A. Fox, P. Pathak, and H. Wu. The effects of fitness functions on genetic programming-based ranking discovery for web search. *Journal of the Association for Information Science and Technology*, 55(7):628–636, 2004.

[3] L. Kuhn and C. Eickhoff. Implicit negative feedback in clinical information retrieval. *arXiv preprint arXiv:1607.03296*, 2016.

[4] J. Palotti and A. Hanbury. Tuw@ trec clinical decision support track 2015. Technical report, Vienna University of Technology Vienna Austria, 2015.

[5] K. Roberts, M. Simpson, D. Demner-Fushman, E. Voorhees, and W. Hersh. State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the trec 2014 cds track. *Information Retrieval Journal*, 19(1-2):113–148, 2016.