Expert Systems With Applications 85 (2017) 221–230

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Generation of simple structured information retrieval functions by genetic algorithm without stagnation

A.S. Kulunchakov[a,*], V.V. Strijov[b]

[a] Moscow Institute of Physics and Technology, Russia
[b] Computing Centre of the Russian Academy of Sciences, Russia

ARTICLE INFO

Article history:
Received 15 May 2016
Revised 5 May 2017
Accepted 7 May 2017
Available online 8 May 2017

Keywords:
Information retrieval
Genetic programming
Ranking function
Evolutionary stagnation
Overfitting

ABSTRACT

This paper investigates an approach to construct new ranking models for Information Retrieval. The IR ranking model depends on the document description. It includes the term frequency and document frequency. The model ranks documents upon a user request. The quality of the model is defined by the difference between the documents, which experts assess as relative to the request, and the ranked ones. To boost the model quality a modified genetic algorithm was developed. It generates models as superpositions of primitive functions and selects the best according to the quality criterion. The main impact of the research if the new technique to avoid stagnation and to control structural complexity of the consequently generated models. To solve problems of stagnation and complexity, a new criterion of model selection was introduced. It uses structural metric and penalty functions, which are defined in space of generated superpositions. To show that the newly discovered models outperform the other state-of-the-art IR scoring models the authors perform a computational experiment on TREC datasets. It shows that the resulted algorithm is significantly faster than the exhaustive one. It constructs better ranking models according to the MAP criterion. The obtained models are much simpler than the models, which were constructed with alternative approaches. The proposed technique is significant for developing the information retrieval systems based on expert assessments of the query-document relevance.

1. Introduction

In Manning, Raghavan, and Schutze (2008) Information retrieval is defined as finding documents of an unstructured nature, usually text that satisfies an information need from within large collections. An IR system stores text archives as a collection. To retrieve documents relevant to a query, one needs a rank estimation procedure called *ranking model*. It is defined on pairs *document-query*. For each pair it returns relevance of the *document* to the *query*. Goswami, Moura, Gaussier, Amini, and Maes (2014) define IR ranking models as functions of two basic features of these pairs: term frequency (*tf*) and document frequency (*idf*). In this paper ranking models are constructed considered as mathematical functions defined on tf-idf features. Instead of enlarging the set of features to provide better performance (Yea, Huangb, & Lina, 2011), current paper use the same tf-idf features to make further comparison consistent.

An information to retrieve is specified by a query, which is first preprocessed with same preprocessing steps as documents. The query terms are searched within the collection terms. Relative documents retrieved from the collection. These documents are ranked according to the ranking function and returned to the user. To evaluate the performance of an IR system a group of experts assess the ranked documents. The experts make a set of queries. For each query an expert makes an assessment of relevance of ranked documents. It gives relevance of a document to a query for query-document pairs. The main problem of the IR system constructing is how to discover a ranking function, which returns the most related documents to each query from a large and diverse test set queries. Developing new term-document scoring functions that outperform already existing traditional scoring schemes is one of the most acute and demanded research area in the theoretical information retrieval (Datta, Varma, C., & Singh, 2017; Vanopstal, Buysschaert, Laureys, & Stichele, 2013) with many applications in the expert systems(Kauer & Moreira, 2016; Tu & Seng, 2009).

The Text REtrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense, was started in 1992 as part of the TIPSTER Text program. For each TREC, National Institute of Standards and

* Corresponding author.
E-mail addresses: kulu-andrej@yandex.com (A.S. Kulunchakov), strijov@gmail.com (V.V. Strijov).

http://dx.doi.org/10.1016/j.eswa.2017.05.019
0957-4174/

Technology (NIST) provides a test set of documents and questions. Participants run their own retrieval systems on the data, and return to NIST a list of the retrieved top-ranked documents. NIST pools the individual results, judges the retrieved documents for correctness, and evaluates the results. Thus each TREC consists of a collection of documents, user queries and judgments for a subset of a collection Each TREC is associated with this triplet. Each triplet has a collection of nearly 500 000 documents. 50 queries to the collection and 2000 judgments for each query in average. The number specified after the name “Trec” denotes the year of the creation of the TREC.

The ranking models in Porter (1997), Metzler and Croft (2005), Amati and Van Rijsbergen (2002), Clinchant and Gaussier (2010) and Ponte and Croft (1998) are derived on some theoretical assumptions. These assumptions This allow to build ranking models without an IR collection, but these assumptions are not often met. For example, the derived ranking models are not optimal according to mean average precision (Manning et al., 2008) on TREC collections (Goswami et al., 2014). Moreover, the quality of these models significantly differs on various the collections (Goswami et al., 2014).

High-performing ranking models are also discovered by automatic procedures. The paper (Goswami et al., 2014) exhaustively explores a set of IR ranking models represented as superpositions of expert-given grammar elements. The grammar is an expert-given set of primitive mathematical functions, where variables are *tf-idf* features (Salton & McGill, 1986). The exhaustive algorithm explores the set of superpositions, which consists of at most 8 grammar elements. The best explored ranking functions in Goswami et al. (2014) are better in average on TREC collections than ones in Porter (1997), Metzler and Croft (2005), Amati and Van Rijsbergen (2002), Clinchant and Gaussier (2010) and Ponte and Croft (1998). Moreover, these functions are guaranteed to have simple structure. However, this algorithm has high computational complexity (Goswami et al., 2014). Therefore, an exploration of more complex superpositions is an intractable problem.

Another approaches to improve IR expert systems include various genetic algorithms: search for an optimal document indexing (Gordon, 1988; Valizadegan, Jin, Zhang, & Mao, 2009), clustering documents according to their relevance to queries (Gordon, 1991; Raghavan & Agarwal, 1987), tuning parameters of queries (Petry, Buckles, Sadasivan, & Kraft, 1994; Yang, Korfhage, & Rasmussen, 1992), facilitate automatic topic selections (Chiu, Pan, & Yu, 2009), search for key words in documents (Chen, 1995) and optimal coefficients of a linear superposition of ranking models (Billhardt, Borrajo, & Maojo, 2002; Pathak, Gordon, & Fan, 2000). Genetic algorithms are applied to select features in image retrieval and classification (Lina, Chenb, & Wua, 2014). Genetic algorithms are used to generate ranking functions represented as superpositions of grammar elements (Fan, Gordon, & Pathak, 2000, 2004; Koza, 1992). These procedures significantly extend the set of ranking superpositions considered in Goswami et al. (2014). However, the *basic* algorithms in Fan et al. (2000), 2004) produce superpositions with significant structural complexity after 30–40 iterations of mutations and crossovers (Koza, 1992). The basic algorithms do not control the structural complexity of generated superpositions and do not solve a problem of evolutionary stagnation, when a population stops to change.

The problem of evolutionary stagnation appears when a majority of stored superpositions have similar structure and high quality. Next crossover operations constructs superpositions, which are similar to the stored ones. The mutation operation constructs a superposition, which is unlikely to have as high quality as the stored superpositions. This superposition highly probably will be eliminated. Therefore the population will pass to the next iteration without changes. The genetic algorithm stops actual generation.

Table 1
Strengths and weaknesses comparison of the algorithms for IR ranking.

Strengths	Weaknesses
Fan et al. (2000), 2004)	
Large feasible set of ranking functions	Complicated final superpositions
Fast convergence to a local optimum	Does not provide global optimum in the feasible set of functions
	Have not been tested on different datasets to show uniform improvement on them
Goswami et al. (2014)	
Provides global optimum with respect to the feasible set	Small feasible set of ranking functions
Compact final ranking functions	
Have been tested on different datasets and uniform improvement over existing approaches was shown	
Robertson and Zaragoza (2009)	
Theoretically justified	Is not uniformly good over different datasets
Simple and compact explicit expression	
The proposed model generation algorithm	
Large feasible set of ranking functions	Does not provide global optimum in the feasible set of functions
Fast convergence to a local optimum	
Compact final ranking functions	
Have been tested on different datasets to show uniform improvement on them	

To outperform the ranking functions found in Goswami et al. (2014), one needs to extend the set of superpositions considered there. To perform it, a modified genetic algorithm is proposed. It detects evolutionary stagnation and replaces the worst stored superpositions with random ones. This detection is implemented with a structural metric on superpositions. Regularizers solve the problem of overfitting. They penalize the excessive structural complexity of superpositions. The paper analyzes various pairs regularizer-metric and chooses the pair providing a selection of better ranking superpositions. All strengths and weakness of compared approaches are summarized in Table 1. The novelty of the proposed algorithms is the solution of the problem of stagnation in the consequent model generation procedure. It brings variety in the generated models and makes the search procedure faster. The significance of the proposed approach is the next level of quality in the ranking functions, which outperforms the exhaustive search.

The paper (Goswami et al., 2014) uses TREC collections to test ranking functions. To make the comparison of approaches consistent, the present paper also use these collections. The collection TREC-7 (trec.nist.gov) is used as the train dataset to evaluate quality of generated superpositions. The collections TREC-5, TREC-6, TREC-8 are used as test datasets to test selected superpositions.

2. Problem statement

There given a collection C consisting of documents $\{d_i\}_{i=1}^{|C|}$ and queries $Q = \{q_j\}_{j=1}^{|Q|}$. For each query $q \in Q$ some documents C_q from C are ranked by experts. These ranks g are binary

$$g : Q \times C_q \to \mathbb{Y} = \{0, 1\},$$

where 1 corresponds to relevant documents and 0 to irrelevant.

To approximate g, superpositions of grammar elements are generated. The grammar $\mathfrak{G}$ is a set $\{g_1, \ldots, g_m, x_w^d, y_w\}$, where each g_i