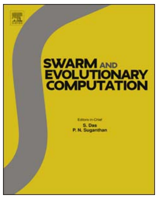
Swarm and Evolutionary Computation xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Swarm and Evolutionary Computation

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



Rank fusion and semantic genetic notion based automatic query expansion model

Jagendra Singh*, Aditi Sharan

Jawaharlal Nehru University, India

ARTICLE INFO

Keywords:
Query expansion term selection
Information retrieval
Semantic filtering
Rank combination
Score combination
Genetic algorithm

ABSTRACT

Query expansion term selection methods are really very important for improving the accuracy and efficiency of pseudo-relevance feedback based automatic query expansion for information retrieval system by removing irrelevant and redundant terms from the top retrieved feedback documents corpus with respect to user query. Individual query expansion term selection methods have been widely investigated for improving its performance. However, it is always a challenging task to find an individual query expansion term selection method that would outperform other individual query expansion term selection methods in most cases. In this paper, first we explore the possibility of improving the overall performance using individual query expansion term selection methods. Second, we propose a model for combining multiple query expansion term selection methods by using rank combination approach, called multiple ranks combination based query expansion. Third, semantic filtering is used to filter semantically irrelevant term obtained after combining multiple query expansion term selection methods, called ranks combination and semantic filtering based query expansion. Fourth, the genetic algorithm is used to make an optimal combination of query terms and candidate term obtained after rank combination and semantic filtering approach, called semantic genetic filtering and rank combination based query expansion. Our experimental results demonstrated that our proposed approaches achieved significant improvement over each individual query expansion term selection method and related state-of-the-art approaches.

1. Introduction

Retrieving relevant documents that can fulfill user need is one of the major challenges in information retrieval system. One of the most feasible and successful technique to handle this problem is Pseudo-Relevance Feedback (PRF) based Query Expansion (QE), where in PRF based query expansion some top documents retrieved in first iteration are used to expand the original user query. In order to consider the above problem, there is a need of automatic PRF based query expansion techniques that can automatically reformulate the original user query. In last some years, it has been observed that the volume of data available online has dramatically increased while the number of query terms searched remained very less. According to the authors in [1], the average query length was 2.30 words, the same reported ten years after in Van Rijsbergen [2]. While there has been a slight increase in the number of long queries (of five or more words), the most prevalent queries are still those of one, two, and three words. In this situation, the need and the scope of PRF based Automatic Query Expansion (AQE) have increased. In this work, our focus is on PRF Based Automatic Query Expansion (PRFBAQE). The PRFBAQE is an effective technique for boosting the overall performance in Information Retrieval (IR). It assumes that top-ranked documents in the first-pass retrieval are relevant, and then uses feedback documents as a source for selecting adding potentially related terms. Although PRF based query expansion has been shown to be effective in improving IR performance [3–10] in a numbers of IR tasks.

The main problem of automatic query expansion is that it cannot work efficiently due to the inherent sparseness of the user query terms in the high dimensional corpus. Another problem is that not all the terms of top retrieved documents (feedback documents) are important for the query expansion. Some of the query expansion term may be redundant or irrelevant. Some may even misguide the result, especially when there are more irrelevant query expansion terms than relevant ones. Query expansion selection aims to remove redundant and irrelevant terms from the term pool (top retrieved documents as feedback documents for selecting user query expansion term), and the selected query expansion term set should contain sufficient and reliable information about the original document. Thus, query expan-

* Corresponding author.
E-mail addresses: jagendrasngh@gmail.com (J. Singh), aditisharan@gmail.com (A. Sharan).

http://dx.doi.org/10.1016/j.swevo.2017.09.007
Received 29 June 2015; Received in revised form 30 November 2016; Accepted 9 September 2017
2210-6502/ © 2017 Elsevier B.V. All rights reserved.

sion term selection should not only reduce the high dimensionality of the feedback document corpus (term pool), but also provide a better understanding of the documents, in order to improve the automatic query expansion result. Feedback based different query expansion term selection methods have been widely used in the automatic query expansion, and it has been reported that query expansion term selection methods can improve the efficiency and accuracy of information retrieval model.

Traditional query expansion term selection methods for automatic query expansion are either corpus statistics based or term association based, depending on the used algorithm in the retrieval model. Term association based term selection methods, such as Mutual information [11] and Co-occurrence information [2,12] estimate the goodness of each term based on the occurrence of terms in feedback documents (term pool). Corpus statistics based query expansion term selection methods, such as Kullback-leibler divergence [13], Binary independence model [14] and Robertson selection value [15] estimate the goodness of each term based on the distribution of terms across the corpus and using the query term informations present in feedback documents. He et al. [16] proposed an improved approach of the classical Okapi-BM25 model by utilizing the term proximity evidence. Bache et al. [17] presented that some scoring functions possess a likelihood property, which means that the scoring function indicates the likelihood of matching when compared to other retrieval tasks. This is potentially more useful than pure ranking even though it cannot be interpreted as an actual probability. Pedronette et al. [18] presented a novel approach for the re-ranking problem. It relies on the similarity of top-*k* lists produced by efficient indexing structures, instead of using distance information from the entire collection. Singh et al. [19] presented contextual window based approach to select the query context related terms from top feedback documents. Wu [20] addressed deficiencies in current information retrieval models by integrating the concept of relevance into the generation model using various topical aspects of the query.

Most of study on the query expansion term selection focused on the performance improvement of individual query expansion term selection methods. However, it remains as a challenge to develop an individual query expansion term selection method that would outperform other methods in most cases. Moreover, as multiple query expansion term selection methods are available; it is natural to combine them for better performance by taking advantage of their individual strength. In the past, experiments of combining multiple query term selection methods have been conducted, but no theoretical analysis has been done. Combinations of two uncorrelated and high-performing query expansion term selection methods have been tested [21]. After combining expansion terms from different term selection methods, it became compulsory to check the semantic meaning of selected expansion terms with the user query for to avoid query drifting problem. For this purpose, we use the concept of semantic similarity with the help of Word2Vec [22].

For this purpose, we use the concept of Word2Vec based semantic similarity with the help of Word2Vec. After applying Word2Vec based semantic filtering, a refined set of additional expansion terms obtained. Even After applying semantic similarity concept, there are a large set of expansion terms, but we need some selected combination of expansion terms, now we apply genetic algorithm for finding optimal combination of expansion term and query te0rms. Some work has been done for using a genetic algorithm for information retrieval and query expansion. Most of the work has been done to tune the weights of query terms or matching functions. Pathak [23] have used a genetic algorithm for improving the efficiency of matching function of an information retrieval system. Araujo [24] have used a genetic algorithm for query expansion based on stemming and morphological variations. The author in [25] presents a new method for query reweighting to deal with document retrieval. The proposed method uses genetic algorithms to reweight a user's query vector, based on the user's relevance feedback, to improve the performance of document retrieval systems. Some related studies [26–35] showing the motivational results achieved with the help of genetic algorithm and other evolutionary approaches.

In this research, we investigated a new approach of rank combination to combine multiple query expansion term selection methods. The rank combination is a method to analyze the combine multiple scoring systems. The rank combination has been applied to a variety of domains such as information retrieval, recommendation system, expert system and many more. In this paper, we studied the use of the rank combination, semantic filtering based rank combination and score combination of four traditional term selection methods: Kullback-Leibler Divergence (KLD), Co-occurrence Information (Co-occurrence), Binary Independence Model (BIM) and Robertson Selection Value (RSV). After it we use semantic filtering to filter irrelevant terms from term collection, and then the genetic algorithm is used to optimize expansion terms combination, finally some selected query expansion terms are used to reformulate the original query. Our experimental results with real data sets demonstrated that combining multiple query expansion term selection methods could improve the performance of AQE for retrieval model in term of the average precision, recall and F-measure values of the results.

The major contributions of this work are summarized as follows:

- ➢ First, we presents KLD, Co-occurrence, BIM and RSV term selection methods for pseudo-relevance feedback based automatic query expansion, with this, the experimental analysis of all these term selection methods are presented with evaluation parameter score.
- ➢ Second, we combine the ranked list of query expansion terms suggested by different expansion term selection methods discussed in Step 1; here we combine these ranks with the help of most popular rank aggregation methods such as Borda, Condorcet, Reciprocal, and Sumscore.
- ➢ Third, we proposed Word2Vec based semantic filtering approach that is used to filter the irrelevant and redundant expansion terms with context to user query obtained from Step 2; for this purpose Word2Vec based semantic similarity module is used to find semantic similarity.
- ➢ Fourth, we proposed a genetic algorithm to get an optimal combination of expansion terms obtained from Step 3 with user query terms, with the help of this approach we found a set of best performing query expansion terms.
- ➢ Conduct Paired *t*-test between our proposed approaches and other's model considered as the baseline model.

The organization of this paper is as follows. In Section 2, we briefly introduce four individual query expansion term selection methods. Section 3 explained our proposed model and its algorithm with rank aggregation, semantic filtering and with genetic algorithm based approach. Section 4 presents the experimental results of different query expansion term selection methods are compared and with each other, next in this section our proposed approaches results are presented and compared or analyzed with baseline approaches in terms of the precision, recall and F-measure on both FIRE and TREC datasets. Finally, Section 5 presents conclusion and future research directions.

2. Query expanding term selection methods

2.1. Kullback-Leibler divergence based query expansion

The Kullback-Leibler Divergence (KLD) [13] is well-known in information theory [12]. KLD based approach has been used in natural language and speech processing applications based on statistical language modelling in information retrieval. KLD can be used as a term-scoring function that is based on the differences between the

Download English Version:

https://daneshyari.com/en/article/6903206

Download Persian Version:

https://daneshyari.com/article/6903206

Daneshyari.com