Information Processing and Management 52 (2016) 478–489

Contents lists available at ScienceDirect

Information Processing and Management

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

A query term re-weighting approach using document similarity

Payam Karisani [a,*], Maseud Rahgozar [a], Farhad Oroumchian [b]

[a] Database Research Group, Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University of Tehran, Iran
[b] University of Wollongong, Dubai

ARTICLE INFO

Article history:
Received 27 May 2015
Revised 19 September 2015
Accepted 23 September 2015
Available online 10 November 2015

Keywords:
Text retrieval
Query term re-weighting
Document similarity
Query expansion

ABSTRACT

Pseudo-relevance feedback is the basis of a category of automatic query modification techniques. Pseudo-relevance feedback methods assume the initial retrieved set of documents to be relevant. Then they use these documents to extract more relevant terms for the query or just re-weigh the user's original query. In this paper, we propose a straightforward, yet effective use of pseudo-relevance feedback method in detecting more informative query terms and re-weighting them. The query-by-query analysis of our results indicates that our method is capable of identifying the most important keywords even in short queries. Our main idea is that some of the top documents may contain a closer context to the user's information need than the others. Therefore, re-examining the similarity of those top documents and weighting this set based on their context could help in identifying and re-weighting informative query terms. Our experimental results in standard English and Persian test collections show that our method improves retrieval performance, in terms of MAP criterion, up to 7% over traditional query term re-weighting methods.

1. Introduction

The traditional computer-based IR is concentrated on techniques that improve the performance of retrieval systems. Examples of such techniques are probabilistic or language modeling (Craswell, Robertson, Zaragoza, & Taylor, 2005; Zaragoza, Craswell, Taylor, Saria, & Robertson, 2004), personalized search (Croft, Cronen-Townsend, & Lavrenko, 2001; Sieg, Mobasher, & Burke, 2007), query classification (Kang & Kim, 2003), and query modification (Lavrenko & Croft, 2001; Lee, Croft, & Allan, 2008). Query modification techniques are a group of models that try to improve the retrieval performance by improving the original user query. There are two main classes of query modification methods. The first class is called query expansion in which the system reformulates the user query (Lavrenko & Croft, 2001; Lee, Croft, & Allan, 2008) by adding extra terms and re-weighting the query terms. The second class however, concentrates only on re-weighting the query terms (Bendersky & Croft, 2008; Robertson & Jones, 1976).

In this paper, we propose an approach to query modification through query term re-weighting. We use automatic feedback to retrieve the first set of relevant documents, and then we extract the information which is needed for assigning a meaningful weight to each query term. Our experimental results in English and Persian languages indicate that our method outperforms traditional query term re-weighting approaches.

The rest of this paper is organized as follows: Section 2 provides an overview of the related studies. Section 3 presents our approach to query term re-weighting in detail. Section 4 reports our results, i.e., Section 4.1 explains our experimental setup,

* Corresponding author. Tel.: +98 2182089718.
E-mail addresses: p.karisani@gmail.com (P. Karisani), rahgozar@ut.ac.ir (M. Rahgozar), oroumchian@acm.org (F. Oroumchian).

http://dx.doi.org/10.1016/j.ipm.2015.09.002
0306-4573/

Sections 4.2 and 4.3 present our results in English and Persian data sets, and Section 4.4 discusses the method. Finally, Section 5 concludes the paper.

2. Related work

Substantial amount of work has been done (Bendersky & Croft, 2008; Lavrenko & Croft, 2001; Lee, Croft, & Allan, 2008; Robertson & Jones, 1976) in English Information Retrieval. Several research studies have influenced our work in one way or another. Lee, Croft, and Allan (2008) propose a method based on local clustering hypothesis. The cluster hypothesis states that a group of similar documents tend to be relevant to the same query. Using a K-NN method they cluster the top retrieved documents, and rank the clusters based on the likelihood of generating the query. Then using the relevance model (Lavrenko & Croft, 2001) they extract the new terms for expansion from the documents which belong to the top clusters. In their method, the documents which appear in several clusters are called dominant. Their hypothesis is that these documents have a good representation of the topics of the query. Because they appear multiple times in the clusters, they can contribute more to the expansion process and improve the precision. Liu, Natarajan, and Chen (2011) use local clustering to propose a novel method for query suggestion. Based on the number of clusters which exist in the top documents, their goal is to suggest a diversified set of expanded queries to the user. Their assumption is that this set of queries will cover all the topics which are related to ambiguous user queries. The result of each query in the set, when is ran against the collection, should be the corresponding cluster with the highest precision and recall. They prove that this problem is NP-hard and try to propose two algorithms which predict the queries. While our method like these methods tries to extract the information which the top documents carry, there are still some differences. First, we do not add new terms to the query. The information which is extracted is used to re-weigh the original query terms. Second, our approach to extract the information is different. We do not cluster the top documents; instead, we treat each one as a single entity which carries information.

One of the first studies on query term re-weighting has been carried out by Robertson and Jones (1976). Their approach is based on the probabilistic retrieval model. The main idea of the probabilistic model is that there is a set of documents which exactly contains all the related documents. Using the properties of this set we could retrieve the related documents. Because we do not have access to the set we try to guess the properties. Thus an initial guess is made about the weights of the query terms to retrieve the first set of documents. In the next step, using an incidence contingency table over the top documents the weights of the query terms are refined to retrieve the final set. Here we do not use probabilistic framework, and we also try to use the information which the top documents carry in relation to each other. There is no such a step in the Robertson's model.

Bendersky and Croft (2008) propose a framework to discover key concepts in verbose queries. First, they propose a model based on language modeling approaches to incorporate concept weights into the retrieval process. Then they define a function which estimates the membership of terms in the set of related concepts to the query. The normalized version of this function is used in their retrieval process. To evaluate the value of this function they use a machine learning approach. In their method concepts are mapped to a feature vector. The values of the vector are several query-dependent and query-independent features. One of their most effective features is the Weighted Information Gain (Zhou & Croft, 2007) which we discuss in Section 4. Here we also focus on short queries. Besides, we directly map terms to the corresponding weights because we only use one resource, which is the top documents.

Recently many studies have been conducted in Persian text retrieval. Saboori, Bashiri, and Oroumchian (2012) investigated the role of query term re-weighting using vector space model (Salton, Wong, & Yang, 1975). Hakimian and Taghiyareh (2008) tried optimizing the parameters of Local Context Analysis (Xu & Croft, 2000). The role of N-gram based vector space model and Local Context Analysis approach has been studied in Aleahmad, Hakimian, Mahdikhani, and Oroumchian (2007).

In this research, we demonstrate that query term re-weighting can be useful even in short queries—those with about three terms. Furthermore, we propose a straightforward, yet effective method for estimating the importance of query terms. An immediate impact of our work would be achieving a higher performance in document retrieval through emphasizing those terms in more elaborate weighting schemes.

Our main motivation for doing this research was the amount of work which has been carried out in this area about verbose queries. Much research has concentrated around long queries, since it is intuitive to assume identifying and eliminating less influential terms in long queries could boost the performance. However, there are not many research studies that specifically investigate the role of keyword detection in short queries. Therefore, it was felt that such an effort is needed to understand the contribution connections of terms in all kinds of queries. Apart from this aim, other requirements of our work are simplicity and robustness in order to make our methods suitable for real world scenarios. We achieve simplicity by only using attributes that readily available at run time. The robustness of our method comes from the fact that we do not rely on a single evidence to assign our weights; instead, we use several filters and steps to ensure the effectiveness of the process.

3. Proposed term re-weighting method

In this section, we present our term re-weighting method. First, we use the original user's query to retrieve the initial relevant documents; then we assign a weight to each relevant document which defines the importance of that document to the user's information need. Finally, we modify the weight of each query term based on their occurrence in these weighted documents. Our method can be categorized as one of the local feedback query modification methods. Local feedback query modification methods