Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 315–326
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Document reranking by term distribution and maximal marginal relevance for chinese information retrieval

Lingpeng Yang, Donghong Ji *, Munkew Leong

Institute for Infocomm Research, Media Understanding, 21 Heng Mui Keng Terrace, Singapore 119613, Singapore

Received 25 May 2006; accepted 25 July 2006

Abstract

In this paper, we propose a document reranking method for Chinese information retrieval. The method is based on a term weighting scheme, which integrates local and global distribution of terms as well as document frequency, document positions and term length. The weight scheme allows randomly setting a larger portion of the retrieved documents as relevance feedback, and lifts off the worry that very fewer relevant documents appear in top retrieved documents. It also helps to improve the performance of maximal marginal relevance (MMR) in document reranking. The method was evaluated by MAP (mean average precision), a recall-oriented measure. Significance tests showed that our method can get significant improvement against standard baselines, and outperform relevant methods consistently.

Keywords: Relevance feedback; Term extraction; Term weighting; Maximal marginal relevance; Chinese information retrieval

1. Introduction

How to further improve the rankings of the relevant documents after an initial search has been extensively studied in information retrieval. Such studies include two main streams: automatic query expansion and automatic document reranking. While the assumption behind automatic query expansion is that the high ranked documents are likely to be relevant so that the terms in these documents can be used to augment the original query to a more accurate one, document reranking is a method to improve the rankings by re-ordering the position of initial retrieved documents without doing a second search. After document reranking, it is expected that more relevant documents appear in higher rankings, from which automatic query expansion can benefit.

Many methods have been proposed to rerank retrieved documents. Lee, Park, and Choi (2001) proposes a document reranking method based on document clusters. They build a hierarchical cluster structure for the whole document set, and use the structure to rerank the documents. Balinski and Danilowicz (2005) proposes a document reranking method that uses the distances between documents for modifying initial relevance

* Corresponding author.
E-mail addresses: lpyang@i2r.a-star.edu.sg (L. Yang), dhji@i2r.a-star.edu.sg (D. Ji), mkleong@i2r.a-star.edu.sg (M. Leong).

0306-4573/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.07.011

weights. Luk and Wong (2004) uses the title information of documents to rerank documents, while Crouch, Crouch, Chen, and Holtz (2002) uses the un-stemmed words in queries to re-order documents. Xu and Croft (1996, 2000) makes use of global and local information to do local context analysis and then use the information acquired to rerank documents. Qu, Xu, and Wang (2000) uses manually built thesaurus to rerank retrieved documents, and each term in a query topic is expanded with a group of terms in the thesaurus. Bear et al. (1997) uses manually crafted grammars for topics to re-order documents by matching grammar rules in some segment in articles. Kamps (2004) proposes a reranking method based on assigned controlled vocabularies. Yang, Ji, and Tang (2004, 2005) use query terms which occur in both query and top $N(N \Leftarrow 30)$ retrieved documents to rerank documents.

One problem in automatic document reranking (also for query expansion) is how many top documents are regarded as relevance feedback in the first retrieval results, which is also faced by most methods mentioned above (Crouch et al., 2002; Kamps, 2004; Lee et al., 2001; Luk & Wong, 2004; Yang et al., 2004, 2005). Usually, a pre-defined smaller number of the documents (say top 10–30) are considered. However, in the cases that very few relevant documents fall within the range, the method will fail. On the other hand, if a larger scope (say 500, 1000) is considered, many irrelevant documents will come inside, and the noisy terms will dominate.

Another problem is that most methods mentioned above do not consider correlation between query terms. Mitra, Singhal, and Buckley (1998) uses maximal marginal relevance (MMR) to adjust the contribution of relevant terms. They argue that usually a document covering more aspects of a query should get higher score, which can be captured somehow by word correlation. The new score for a document is computed by summing the idf (inverse document frequency) of each query word where each word is normalized by correlation probability based on a large number of retrieved documents (say top 1000 documents). It is reported that their method achieves better result in reranking top 50–100 documents. But we find that within top initially retrieved documents, some really relevant terms do appear in larger portion of the documents, which will be unexpectedly assigned lower scores by idf scheme.

In this paper, we propose a new term weighting scheme to deal with the two problems mentioned above. First, we consider document rankings, i.e., document positions in the ranking list, in the weighting scheme of the terms. Intuitively, a term gets a lower document frequency when occurring in a lower-ranking document, and a higher document frequency when occurring in a higher-ranking document (in contrast, the usual way for document frequency is that a document gets 1 count no matter where the document is located in the list). In this way, we can randomly choose a larger number of the documents as relevance feedback, without any worry about the irrelevant documents inside. Furthermore, we do not need to worry about the cases that top documents only contain very few relevant documents, since we can randomly set a larger scope as relevance feedback.

Second, the weighting scheme incorporates both local (feedback) and global distribution of the terms, and we use it to replace the idf scheme in MMR. If a term occurs in feedback documents more frequently than in the whole collection, it tends to have more contribution to document reranking; otherwise, it will be a noise.

Our method does not use word but uses the key terms extracted from queries and top retrieved documents. One motivation of this choice is that terms (including multi-word units) usually contain more complete information than individual words, and have more potential for improving the performance of information retrieval. Another motivation of this method is specifically for Chinese language information retrieval, where a word segmentation module is usually needed, which, however, generally requires some manual resources and suffers from the problem of portability. An automatic term extraction module could be a good alternative.

The rest of this paper is organized as the following. In Section 2, we describe key term extraction from documents. In Section 3, we talk about term weighting. In Section 4, we specify how to rerank the documents based on the key terms and their weighting together with MMR based on term correlation. In Section 5, we evaluate the method on NTCIR-3 CLIR Chinese SLIR document collection and give some analysis. In Section 6, we present the conclusion and future work.

2. Term extraction

Term extraction concerns the problem of what is a term. Intuitively, key terms in a document are some word strings which are conceptually prominent in the document and play main roles in discriminating the document from other documents.