Expert Systems with Applications 39 (2012) 2710–2719

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

A multi-level matching method with hybrid similarity for document retrieval

Haijun Zhang, Tommy W.S. Chow *

Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong

ARTICLE INFO

Keywords:
Document retrieval
EMD
Multi-level matching
Hybrid similarity
Multi-level structure

ABSTRACT

This paper presents a multi-level matching method for document retrieval (DR) using a hybrid document similarity. Documents are represented by multi-level structure including document level and paragraph level. This multi-level-structured representation is designed to model underlying semantics in a more flexible and accurate way that the conventional flat term histograms find it hard to cope with. The matching between documents is then transformed into an optimization problem with Earth Mover's Distance (EMD). A hybrid similarity is used to synthesize the global and local semantics in documents to improve the retrieval accuracy. In this paper, we have performed extensive experimental study and verification. The results suggest that the proposed method works well for lengthy documents with evident spatial distributions of terms.

1. Introduction

Collection and searching of documents has become an integral part of most people's lives because of the easy access to the internet. Document retrieval (DR) refers to finding similar documents for a given user's query that can be ranged from a full description of a document to a few keywords for different searching purposes. Most of the widely used retrieval approaches are still keywords based searching methods, in which untrained users provide search engine several keywords to find relevant documents. Another approach is to use a query document for searching similar ones. Apparently, using an entire document as a query will likely to deliver better retrieval results compared with using just few keywords, but the entire document based approach is more complex and computationally demanding (Chow & Rahman, 2009). In addition, using an entire document is undoubtedly more flexible to users, because it is straightforward and quite often a query may be in paragraphs, sections, and even a whole chapter. Thus, in this paper we propose a DR using the entire document based approach. Also, the corresponding data structure is developed accordingly.

Existing DR systems use statistical models and natural language processing (NLP) approaches with different document representations to facilitate text data mining. Currently, most document representation schemes are based on vector space, latent semantic space and language models. The vector space model (VSM) (Salton & Buckley, 1988; Salton & McGill, 1983), which usually uses *tf.idf* for term weighting, constructs a basic vocabulary of "words" or "terms" for feature description. The term frequency (*tf*) is the number of occurrences of each term. The inverse document frequency (*idf*) is a function of the number of document where a term appears. A term weighted vector is then constructed for each document using *tf* and *idf*. Similarity between two documents is measured using 'cosine' distance or any other distance functions (Zobel & Moffat, 1998). Several dimensionality reduction methods that map documents into latent semantic space are proposed to extract the statistical structure of documents in an attempt to relieve the computational burden of VSM when computing the similarity between two documents.

Latent semantic indexing (LSI) (Deerwester & Dumais, 1990) maps the incident matrix of documents and terms into a latent representation by employing a linear projection to compress the feature vector of the VSM. LSI is not only a widely used technique for performing feature compression, it is also useful for encoding the semantics (Berry, Dumais, & O'Brien, 1995). Hofmann (1999) later proposed a probabilistic latent semantic indexing (PLSI) approach to reveal the statistical properties of latent semantics. PLSI defines a proper generative model to represent each word in a document as a sample from a mixture distribution and develop factor representations for mixture components. A brief overview of other probabilistic models, such as latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003), exponential family harmonium (EFH) (Welling, Rosen-Zvi, & Hinton, 2004) and rate adapting Poisson (RAP) model (Gehler, Holub, & Welling, 2006), can be referred to (Zhang, Chow, & Rahman, 2009). Apart from these probabilistic models, language model (Ponte & Croft, 1998), which is an alternative to VSM, has become increasingly popular. The relevance of a document to a given query is ranked by statistical techniques and the underlying language model. Erkan (2006) also introduced a language model-based document representation using random walks for document clustering. In addition to these techniques that aim

* Corresponding author.
E-mail address: eetchow@cityu.edu.hk (T.W.S. Chow).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.08.128

to reduce the feature dimension, other attempts such as clustering methods and new system structures were suggested to facilitate DR. Rooney, Patterson, Galushka, and Dobrynin (2006) introduced a clustering method for large document corpus. This method narrows the searching scope by comparing a query to a group of documents that are clustered according to the semantic relevance. Apart from these clustering techniques, a new file structure (Du, Ghanta, Maly, & Sharrock, 1989) was also suggested to support the retrieval process.

Despite the above progresses, most reported techniques are largely based on typical *tf* information from the "bag of words" model. They use the flat feature representation, which is a function of *tf*. As this feature extraction scheme is only a rough representation of a document, it often results in losing certain important semantic information. For example, two documents containing similar term frequencies may be contextually different when the spatial distribution of terms are different, i.e., *school*, *computer*, and *science* means very different when they appear in different parts of a document compared with the case of "*school of computer science*" that appear together. Thus, only relying on the *tf* information is not the most effective way to account contextual similarity that includes word inter-connections and spatial distributions of words. The semantics may be very different irrespective of whether or not the spatial distributions of terms are to be considered. By considering the patterns of the query term occurrence, Park, Ramamohanarao, and Palaniswami (2005) proposed a spectral-based DR approach. They suggested that documents can be considered relevant if they contain query terms that follow a similar positional pattern. But this method is more applicable to keywords based query only. To reflect the subtopic structure of a document, Kim and Kim (2004) introduced a passage-based text categorization model. It segments a test document into several passages, assigns categories to each passage, and merges passage categories into document categories. In Kim and Kim (2004), it suggested that the location of a passage can determine its degree of contribution to the test document. To improve the categorization performance of document, Xue and Zhou (2009) proposed to represent the compactness of the appearances of a word and the position of the first appearance of the word by the distribution features. In our recent work, we proposed a new document representation using the *tf* and term connections. These features are extracted from different term graphs using weighted feature extraction method (Chow, Zhang, & Rahman, 2009). We also developed a new dual wing harmonium model (DWHM) integrating the *tf* features and term connection features into a low dimensional semantic space. It is noted that the DWHM approach does not increase the computation burden of DR (Zhang et al., 2009).

In this paper, DR is performed in a way of comparing the entire document via a multi-level matching (MLM) method. The multi-level structure of the document data includes document level and paragraph level that correspond to the global semantics and local semantics of a document, respectively. This data structure that we used to represent the spatial distributions of terms enables us to generate a signature for document matching. More importantly, this multi-level-structured representation can accurately reflect the underlying semantics that the traditional flat feature structure is unable to handle. In our proposed method, the matching between documents is transformed into an optimization procedures using Earth Mover's Distance (EMD). Document matching is conducted via linear programming that finds the optimal distance between documents. We use a hybrid similarity to synthesize the global and local semantics to improve the retrieval accuracy. Preliminary theoretical results and analysis on the relationship of the global similarity and local similarity are presented in later section of this paper. In application level, we developed a two-step retrieval system to reduce the computational burden, and to facilitate practical applications for large data corpus. In this paper, we conducted intensive study on a new dataset including lengthy documents, which were collected from the internet (Chow & Rahman, 2009; Zhang et al., 2009). The proposed method is compared with other two methods (VSM and LSI). The presented results show that the proposed method performs well for lengthy documents with evident spatial distributions of terms. This paper also investigates the sensitivity of parameter setting on the results. Our results indicate that using the multi-level-structured representation together with MLM can provide a general framework for DR.

The contribution of this paper is twofold. First, we propose a multi-level-structured representation to express more semantic information of the term inter-connections and spatial distribution of a document. Second, an MLM method incorporate with EMD distance solved by linear programming is introduced to find the optimal similarity between documents. A hybrid similarity including the global and local information is then used to enhance the retrieval accuracy. Experimental results corroborate that our proposed method works well for lengthy documents. Our proposed two-step retrieval system can serve as a general and computationally efficient solution for DR. It is worth pointing out that the motive of this work focuses on large document size problem, because it useful to real world DR. The methodology described in this paper is a major improvement compared to the traditional methods such as VSM (Salton & Buckley, 1988; Salton & McGill, 1983) and LSI (Deerwester & Dumais, 1990), because our method considers document partition, dimension reduction, and many-to-many matching.

The remaining sections of this paper are organized as follows. Multi-level feature extraction is introduced in Section 2. Multi-level-structured signatures of document are generated for later MLM. In Section 3, multi-level matching scheme with hybrid similarity is proposed by using EMD. Analysis on the interrelations of global similarity and local similarity is presented. Section 4 introduces the general framework of DR system together with implementation details. Extensive experimental results followed by discussions are presented in Section 5. The paper ends with conclusions in Section 6.

2. Multi-level-structured feature extraction

2.1. Document partition

In order to represent the spatial distributions of terms and include more semantic information, we propose a multi-level-structured representation that only consists of text content. To extract a multi-level structure, a document can be partitioned into sections. Sections are further partitioned into paragraphs. For simplicity without loss of generality, we only segment a document into paragraphs that form a two-level structure consisting of document level and paragraph level. We only consider HTML documents in this paper. In HTML format document, paragraphs can be easily identified using HTML tags. Before document segmentation, we firstly filter out the formatted text that appears within the HTML tags. The text is not accounted for in word counts or document features. A document is segmented into a number of paragraphs blocks using the HTML tags: "<p>", "<br\>", "<li>", "</td>", etc. In order to control the number of paragraphs, we merge the subsequent blocks to form a new paragraph until the total number of words of the merged blocks exceeds a paragraph threshold value τ_p. We also include the study of the value of τ_p in the experimental section. We set the minimum threshold for the total number of words in a paragraph to 30; otherwise the new block is merged with the previous paragraph. In this way, the blocks that contain