Pattern Recognition 42 (2009) 2950 – 2960

Contents lists available at ScienceDirect

Pattern Recognition

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

A new dual wing harmonium model for document retrieval

Haijun Zhang, Tommy W.S. Chow*, M.K.M. Rahman

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

ARTICLE INFO

Article history:
Received 14 October 2008
Received in revised form 16 February 2009
Accepted 15 March 2009

Keywords:
Dual wing harmonium
Term connection
Graph representation
Document retrieval
Multiple features

ABSTRACT

A new dual wing harmonium model that integrates term frequency features and term connection features into a low dimensional semantic space without increase of computation load is proposed for the application of document retrieval. Terms and vectorized graph connectionists are extracted from the graph representation of document by employing weighted feature extraction method. We then develop a new dual wing harmonium model projecting these multiple features into low dimensional latent topics with different probability distributions assumption. Contrastive divergence algorithm is used for efficient learning and inference. We perform extensive experimental verification, and the comparative results suggest that the proposed method is accurate and computationally efficient for document retrieval.

1. Introduction

The rapid development of Internet has made massive amount of document data available and easy access to people's lives, which leads to a growing demand of higher accuracy and speed for document retrieval. Document retrieval refers to finding similar documents for a given user's query. A user's query can be ranged from a full description of a document to a few keywords. Most of the extensively used retrieval approaches are keywords-based searching methods, e.g. www.google.com, in which untrained users provide a few keywords to the search engine finding the relevant documents in a returned list. Another type of document retrieval is to use a query document to search similar ones. Using an entire document as a query performs well in improving retrieval accuracy, but it is more computationally demanding compared with the keywords-based method. Most existing document retrieval systems only use term frequency as feature units to build statistical models and develop natural language processing (NLP) approaches for document retrieval [1]. Usually the connections among terms are overlooked which results in losing important semantic information of documents. To exploit rich information in documents and enhance the performance of relevant data mining, it is often necessary to model more features extracted from documents into a lower dimensional semantic space.

Vector space model (VSM) [2], the most popular and widely used term frequency (*tf*)–inverse-document-frequency (*idf*) scheme, uses a basic vocabulary of "words" or "terms" for feature description. The term frequency is the number of occurrences of each term, and the inverse-document-frequency is a function of the number of document where a term took place. A term weighted vector is constructed for each document using *tf* and *idf*. Similarity between two documents is then measured using "cosine" distance or any other distance functions [3]. Thus, the VSM scheme reduces arbitrary length of term vector in each document to fixed length. But a lengthy vector is required for describing the frequency information of terms, because the number of words involved is usually huge. This causes a significant increase of computational burden making the VSM model impractical for large corpus. In addition, VSM scheme reveals little statistical structure about a document because of only using low level document features (i.e. term frequency).

To overcome the shortcomings of VSM, researchers have proposed several dimensionality reduction methods with low dimensional latent representations to capture document semantics. Latent semantic indexing (LSI) [4], an extension from VSM model, maps the documents and terms to a latent space representation by performing a linear projection to compress the feature vector of the VSM model into low dimension. Singular value decomposition (SVD) is employed to find the hidden semantic association between term and document for conceptual indexing. In addition to feature compression, LSI model is useful in encoding the semantics [5]. A step forward in probabilistic models is probabilistic latent semantic indexing (PLSI) [6] that defines a proper generative model of data to model each word in a document as a sample from a mixture distribution and develop factor representations for mixture components. Chien and Wu [7] further developed an adaptive Bayesian PLSI for incremental learning and corrective training that was designed to retrieve relevant documents in the presence of changing domain or

* Corresponding author. Tel.: +852 27887756; fax: +852 27887791.
E-mail address: eetchow@cityu.edu.hk (T.W.S. Chow).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.03.021

topics. By realizing overfitting problems and the lack of description at the level of documents in PLSI, Blei et al. [8] introduced an extension in this regard, latent Dirichlet allocation (LDA). LDA is viewed as a three-level hierarchical Bayesian model, in which each document is modeled as a finite mixture over an underlying set of topics. Using probabilistic approach is able to provide an explicit representation of a document. Compared with LDA, exponential family harmonium (EFH) model [9] is an alternative two-layer model using exponential family distributions and the semantics of undirected models for document retrieval. EFH is able to reduce the feature dimension significantly using a few latent topics (or hidden units) to represent a document. But EFH is only practical for term observations with very few states (e.g. binary). Gehler et al. [10] then developed a rate adapting Poisson (RAP) model that follows the general architecture of EFH. RAP model couples latent topics to term counts using a conditional Poisson distribution for observed count data and conditional binomial distribution for latent topics involving a weight matrix, respectively. Xing et al. [11] and Yang et al. [12] developed dual wing harmonium (DWH) and hierarchical harmonium (HH) to model associated data from multiple sources jointly for the special applications in video classification. In their DWH model, the authors directly treated the term counts via Bernoulli distribution whose rates are determined by the combination of latent topics and the whole image color histogram via a multivariate Gaussian distribution whose mean is determined in the same way.

In all the above mentioned approaches, it is noticed that they use independent word as feature unit. These feature extraction schemes are a rough representation of a document. For example, two documents containing similar term frequencies may be contextually different when the spatial distribution of terms is very different, i.e. *school*, *computer*, and *science* mean very different when they appear in different parts of a document compared to the case of *school of computer science* that appear together. In addition, with the evolution of natural language, there are increasing combinatorial words emerged such as *computer network*, *neural network*, and *complex network*. Thus, using only term frequency information from the "bag of words" model is not the most effective way to account contextual similarity that includes the word inter-connections and spatial distribution of words throughout the document. The semantics may be very different whether considering the term connections or not. To address these shortcomings and improve the retrieval accuracy, first, we in this paper introduce undirected graph for document representation that resulting in more semantic information to be included. Term frequency features and vectorized graph connectionists are then extracted from each document by weighted feature extraction method. Motivated by ideas in Ref. [11], we then develop a new dual wing harmonium to generate distributed latent representations of documents with modeling multiple features jointly. We model term counts (term frequency features) with a conditional Poisson distribution and term connection features with a conditional Bernoulli distribution, respectively. Latent topics are treated as a conditional binomial distribution involving weighted matrixes and multiple features. DWH in this paper is an extension of RAP [10] model with combining multiple features into document latent representation framework without increasing computation burden. The performance of DWH model is investigated in the applications of document retrieval. We show the superiority of DWH for retrieval accuracy compared to RAP model and the recently proposed LDA [8]. We also investigate the influence of number of latent topics and different learning methods for DWH inference. Therefore, the contribution of this paper is twofold. First, we propose a multiple feature extraction framework for representing a document combined with traditional term counts feature and term connection feature extracted from graph. Multiple features are able to express more semantic information of the term inter-connections and spatial distribution throughout document. Second, a new DWH model is developed to project multiple features to low dimensional latent representations capturing the semantics hidden in documents. These latent topics are then applied to document retrieval with promising results.

The remaining sessions of this paper are organized as follows. Multiple features extraction framework is introduced in Section 2. In Section 3, a new DWH model is described in details with brief introduction to EFH and RAP models. Section 4 introduces contrastive divergence (CD) algorithm for DWH learning and inference, and summarizes the implementation framework for document retrieval system. Extensive experimental results followed by discussions are presented in Section 5. The paper ends with conclusions and future work propositions in Section 6.

2. Multiple features extraction framework

In this section, we describe multiple features (terms and term connections) extraction framework to extract more information from each document for better document analysis.

2.1. Graph representation of document

In our work, we use undirected graph to represent each document in corpus. It is worth mentioning that graph representation for document is not new. An interesting application of graph representation describing words links with a perspective of evolving complex network for human language study can be found in Refs. [13,14]. In Refs. [15,16], different directed graphs with a few most frequent terms as nodes were defined to represent a document, k-nearest neighbor algorithm (k-NN) with different graph matching distances based on maximum common subgraph was applied to web document classification. Graph matching can be accomplished in polynomial time making it impractical for large datasets. Apart from the computation time limitation, there may be difficulties in finding maximum common subgraph (subgraph isomorphism) between two documents. Although it is quite straightforward to apply directed graph to express the semantics using terms in sequence appearing in the document, in many cases the sequence of terms is convertible with conveying the same semantics for human language. For example, "*computer science*" can be expressed as "*science of computer*", which delivers the same meaning. Thus, in this paper we use undirected graph for representation of each document.

First, we remove the stop words (set of common words such as "in", "the", and "are", etc.) which deliver little discriminate information. Then, we use the rest of the terms to form an undirected graph. An undirected graph G for a document is denoted by $G=(V,E,\phi,\theta)$, where V represents a set of vertices (i.e. terms), E is a set of edges or connections between terms, $\phi : V \rightarrow L_V$ assigns an attribute (i.e. term frequency) to each vertex of V, similarly, $\theta : E \rightarrow L_E$ assigns an attribute (i.e. term connection frequency) to each edge of E. For example, Fig. 1 illustrates how such a graph would look like for a sentence "*we found it significantly more expensive for sending money to Mexico, but slightly less for sending money to the United Kingdom*". Note that we use only a single vertex for each term even if a term appears more than once in the document. In an early implementation, we used a single vertex to represent a term chain consisting of two and three words that appear together throughout the document, but later found that using only a single vertex for each term is sufficient and improves the performance of our application. Each vertex is labeled with term frequency measure that indicates how many times the related term appears in the document. Similarly, each edge is labeled with term connection frequency measure that indicates how many times the connected terms appear together in the document.