Expert Systems with Applications 37 (2010) 4403–4412

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A novel dual wing harmonium model aided by 2-D wavelet transform subbands for document data mining

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

Keywords:
Dual wing harmonium
2-D wavelet
Term association
Graph representation
Document data
Multiple features

ABSTRACT

A novel dual wing harmonium model that integrates multiple features including term frequency features and 2-D wavelet transform features into a low dimensional semantic space is proposed for the applications of document classification and retrieval. Terms are extracted from the graph representation of document by employing weighted feature extraction method. 2-D wavelet transform is used to compress the graph due to its sparseness while preserving the basic document structure. After transform, low-pass subbands are stacked to represent the term associations in a document. 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 in document classification and retrieval, and comparative results suggest that the proposed method delivers better performance than other methods.

1. Introduction

In this paper, we consider the problem of modeling document data using multiple features. The evolution of human languages leads to a growing demand of extracting more features from documents to express rich information and different meanings of term combinations. Another demand is to find low dimensional semantic expressions of documents with integrating multiple features while preserving the essential statistical relationships between terms and documents, which is useful for facilitating processing of large corpora and dealing with data mining tasks such as classification, retrieval, summarization and plagiarism detection.

Vector space model (VSM) (Salton & McGill, 1983), the most popular and widely used *tf–idf* scheme, uses a basic vocabulary of "words" or "terms" for feature description. The term frequency (*tf*) is the number of occurrences of each term, and the inverse-document-frequency (*idf*) is a function of the number of document where a term took place. A term weighted vector is then constructed for each document using *tf* and *idf*. Similarity between two documents is then measured using 'cosine' distance or any other distance functions (Zobel & Moffat, 1998). 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). Latent semantic indexing (LSI) (Deerwester & Dumais, 1990), 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 (Berry, Dumais, & O'Brien, 1995). A step forward in probabilistic models is probabilistic latent semantic indexing (PLSI) (Hofmann et al., 1999) 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 (2008) 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 topics. By realizing overfitting problems and the lack of description at the level of documents in PLSI, Blei, Ng, and Jordan (2003) 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

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

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.11.088

LDA, exponential family harmonium (EFH) model (Welling, Rosen-Zvi, & Hinton, 2004) 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). By following the general architecture of EFH, Gehler, Holub, and Welling (2006) then developed a rate adapting Poisson (RAP) model that 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, Yan, and Hauptmann (2005) and Yang et al. (2008) 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.

These approaches only use independent word as feature unit, and these feature extraction schemes are a rough representation of a document. However, in real applications, it is important to consider the document structure and term associations in each document. For example, two documents containing similar term frequencies may be contextually different when the spatial distribution of terms are very different, i.e., *school*, *computer*, and *science* means very different when they appear in different parts of a document compared to the case of *school of computer science* that appear together. 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. By realizing this problem, Chow and Rahman (2009) introduced a tree structure and used multilayer self-organizing map (SOM) for document retrieval and plagiarism detection with promising results. In this paper, we try to use graph, wavelet compression and statistical data reduction with multiple features to improve document data mining performance. First, we introduce undirected graph for document representation that results in more semantic information to be included. Terms are extracted by using weighted feature extraction method. Each document graph is then compressed by employing 2-D wavelet transform. We use stacked low-pass subbands with preserving document structure as term associations features. Motivated by ideas in reference (Xing et al., 2005), we then develop a novel dual wing harmonium (DWH) to generate distributed latent representations of documents with modeling multiple features jointly. We model term counts (term frequency, TF) with a conditional Poisson distribution and wavelet transform (WT) features with a conditional multivariate Gaussian 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 (Gehler et al., 2006) model with combining multiple features into document latent representation framework. The performance of DWH model is investigated in the applications of document classification and retrieval. We report accuracy results comparing with RAP model and traditional LSI. We also investigate the influence of the number of latent topics, different inference methods and normalization parameter for balancing weights of TF feature and WT feature. Therefore, the contribution of this paper is twofold. First, we propose a multiple feature extraction framework for representing a document combined with traditional TF feature and WT feature extracted from graph compression using 2-D wavelet transform. Multiple features are able to express more semantic information of the terms associations 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 classification and 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 algorithm for DWH learning and inference. Application results together with discussions are presented in Section 5. The paper ends with conclusions and future work propositions in Section 6.

2. Multiple features extraction framework

2.1. TF feature

First, extract all the words from all documents except for stop words (set of common words such as "in", "the", "are", etc.) which deliver little discriminate information in a database and apply stemming algorithm to each word. Here, Porter stemming algorithm (Porter, 1980) is applied to extract stem of each word, and stems are used as basic features instead of original words. Thus, "send", "sent" and "sending" are all considered the same word. Store the stemmed words together with the information of term frequency f_t (the number of times that a term appears in one document) and the document-frequency f_d^t (the number of documents where a term appears). Then, construct the vocabulary based on TF features. We use a term-weighting measure in calculating the weight of each word, which is similar to VSM (Salton & Buckley, 1988)

$$W_t = \sqrt{f_t} \times idf, \tag{1}$$

where the inverse-document-frequency $idf = \log_2\left(\frac{N}{f_d^t}\right)$, and N is the total number of documents in the corpus. Then, the words are sorted in descending order according to the weights and the first n words are selected to construct the vocabulary. The choice of n depends on the database.

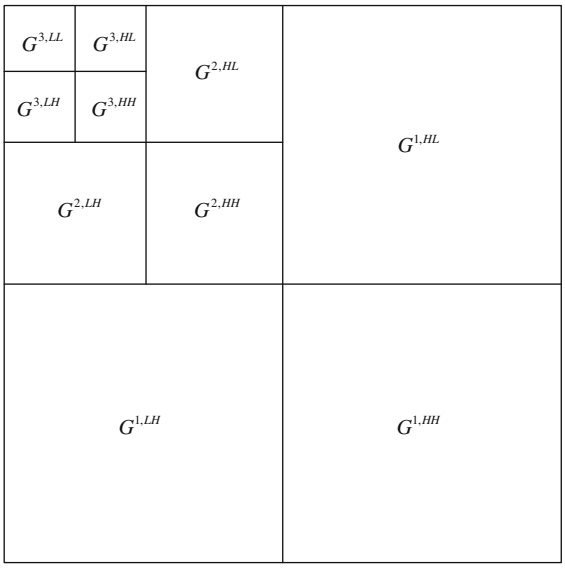

Fig. 1. Three-scale 2-D wavelet graph decomposition.