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.

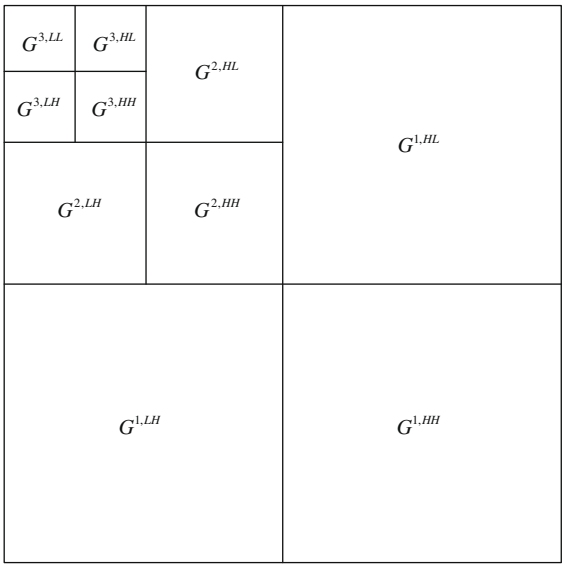


Fig. 1. Three-scale 2-D wavelet graph decomposition.

2.2. Term graph

In our work, we use above extracted terms to construct undirected graph to represent the term associations in each document. 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 Dorogovtsev and Mendes (2001) and I Cancho and Sole (2001). In Schenker, Last, Bunke, and Kandel (2003, 2004), 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 data sets. 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.

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 associations between terms, $\phi : V \to L_V$ assigns an attribute (i.e. term frequency) to each vertex of V, similarly, $\theta : E \to L_E$ assigns an attribute (i.e. term association frequency) to each edge of E. Note that we use only a single vertex for each term even if a term appears more than once in the document. 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 association frequency measure that indicates how many times the connected terms appear together in the document. Here, "connected" means that two terms are adjacent to each other without distinguishing the term sequence.

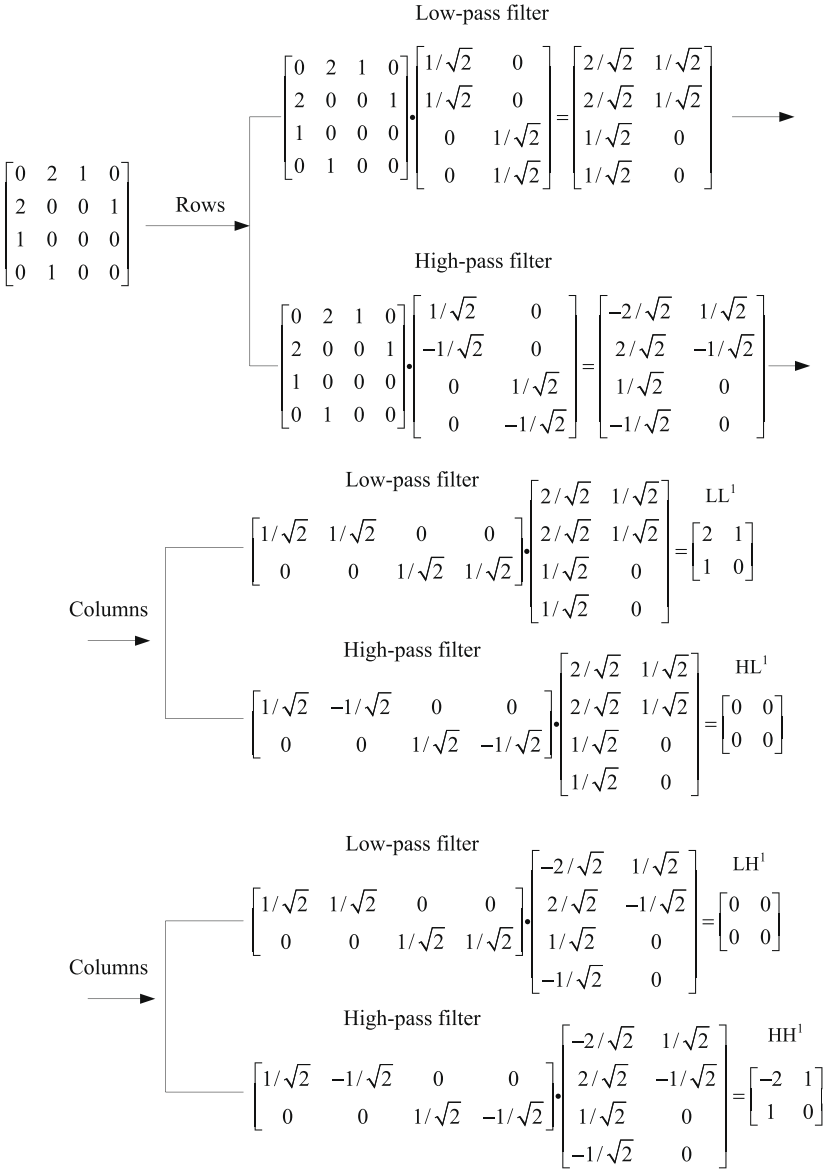


Fig. 2. One-scale complete wavelet transform process.

Actually, two terms also can be treated as "connected" if they appear together in a sentence, a paragraph or even a page, which depends on the applications or the datasets.

2.3. 2-D wavelet transform and WT feature

2-D wavelet transform, an effective tool for signal multiresolution analysis, has been widely used in image processing (Lewis & Knowles, 1992; Mallat, 1989; Ouhsain & Ben Hamza, 2009; Sengur, 2008) and complex network analysis (Fan & Wang, 2005). To perform wavelet transform of documents, we first build the adjacent matrix for the term graph of each document. The adjacent matrix A^l $(l = 1, 2, \ldots, N)$ for term graph G^l is denoted by $A^l = [a^l_{ij}]_{n\times n}$ where a^l_{ij} represents the term association frequency between term i and term j in document l. Then, we use the hierarchical wavelet decomposition (Mallat, 1989) for A^l. A high-pass filter (H) and a low-pass filter (L) are applied to the adjacent matrix A^l in both the horizontal and vertical directions, and the filter outputs subsampled by a factor of two, generating three orientation selective high-pass subbands, i.e. HH, HL, and LH, and a low-pass subband LL. The process is then repeated on the LL subband to generate the next level of the decomposition, etc. (Fan & Wang, 2005; Lewis & Knowles, 1992). Fig. 1 illustrates the block diagram of the decomposition in this way. As seen in Fig. 1, three levels of decomposition lead to ten subbands. Because wavelet transform has good energy compaction property, i.e. most of the system energy is usually concentrated in the lowest frequency subband while with preserving little or no energy in the high frequency subbands, it is efficient to use it to transform these sparse adjacent matrixes into more compact ones whilst preserving the basic graph structures. By performing the recursive filtering process, we are able to complete graph compression with multiresolution analysis.

To perform the 2-D wavelet transform, we here use Haar wavelets functions that consist of a short positive pulse followed by a short negative pulse. The multiresolution version of the Haar wavelet transform is an averaging and difference process for the graph. The first level of the multiresolution analysis splits the original matrix up into a low frequency part (averaging process) and three high frequency parts (difference process), which are horizontal, vertical, and diagonal details (Fan & Wang, 2005; Lewis & Knowles, 1992). The high frequency parts are also called the Haar wavelet coefficients. This process can be repeated as many times as desired. We will now provide a simple example of how we can use the Haar wavelet that has scaling function (low-pass filter) $[1/\sqrt{2}\,1/\sqrt{2}]$ and wavelet coefficients (high-pass filter) $[1/\sqrt{2} - 1/\sqrt{2}]$ to compress the adjacent matrix while providing us with the different levels of resolution of a graph. For example, if the adjacent matrix is $A^l = [0210; 2001; 1000; 0100]$ that illustrates the associations among four terms, we first use the Haar wavelet transform for each rows of A^l and we obtain the wavelet decomposition in horizontal direction $\widetilde{A}^l$=$[2/\sqrt{2}\,1/\sqrt{2} - 2/\sqrt{2}\,1/\sqrt{2}; 2/\sqrt{2}\;1/\sqrt{2}\,2/\sqrt{2} - 1/\sqrt{2}; 1/\sqrt{2}\,0\,1/\sqrt{2}\,0; 1/\sqrt{2}\,0 - 1/\sqrt{2}\,0]$. We then use the Haar wavelet transform for each columns of $\widetilde{A}^l$ and we get the wavelet decomposition in vertical direction $A^l_1 = [2100; 1000; 00 - 21; 0010]$ whose low-pass subband LL = $[21; 10]$ can be as the input of next level decomposition. The one-scale complete wavelet transform process is summarized in Fig. 2. The multiscale wavelet decomposition provides a natural way to compress the term graph. After the multiscale transform, we process the low frequency parts as WT features which preserve the basic graph structure and qualitatively include most of the system energy.

By performing a r-scale wavelet transform of A^l, we denote the lowest resolution subband as $G^{r,LL} = A^l_r = [a^l_{r,ij}]_{n_r\times n_r}$ $(n_r = n/2^r)$. Without loss of generality, we assume that 2^r is a factor of n. Then we denote $Y = (y_1, \ldots, y_{n_r})$ as WT features of the document and $y_j \in R^{n_r}$ is a n_r-dimensional vector that represents the j-th row of A^l_r. So Y is a stacked vector constructed by A^l_r. By performing the 2-D wavelet transform, the compression rate of the original adjacent matrix is 4^r. The choice of r depends on the database. In this way, we achieve high computational efficiency.

3. DWH model for document data

The original harmonium model based on harmonium theory (Smolensky, 1986) refers to a family of bipartite undirected graphical models. Fig. 3(a) illustrates the bipartite topology of a harmonium that consists of two layers of nodes. Nodes $X = \{X_i\}$ at the bottom layer represent the observed data and nodes $H = \{H_k\}$ at the top layer denote the latent topics (or hidden units) of the data. For document data, X can represent TF feature (i.e. term counts) of each document, and H represent resultant discriminator by projecting higher dimensional TF feature into low dimensional semantic space. One of advantages of harmonium model is that the nodes within the same layer are conditionally independent given the nodes in the other layer, which facilitates the generation of harmonium distribution based on two between-layer conditional distributions $p(x|h)(p(x|h) = \prod_i(x_i|h))$ and $p(h|x)(p(h|x) = \prod_j p(h_j|x))$.

3.1. EFH model and RAP model

EFH model introduced by Welling et al. (2004), a special class of harmonium models in exponential family, can be understood as an undirected probability model that combines latent topics in the log-probability domain. The conditional distributions at two layers and the joint distribution (harmonium random field) are in the following way (Welling et al., 2004; Yang et al., 2008).

$$p(x|h) = \prod_i (x_i|h) \propto \prod_i \exp\left\{\left(\theta_i + \sum_j W_{ij} g(h_j)\right) f(x_i)\right\} \tag{2}$$

$$p(h|x) = \prod_j p(h_j|x) \propto \prod_j \exp\left\{\left(\eta_j + \sum_i W_{ij} f(x_i)\right) g(h_j)\right\} \tag{3}$$

$$p(x,h) \propto \exp\left\{\sum_i \theta_i f(x_i) + \sum_j \eta_j g(h_j) + \sum_{ij} W_{ij} f(x_i) g(h_j)\right\} \tag{4}$$

where $\{f(x_i)\}$ and $\{g(h_j)\}$ are the sufficient statistics of node $\{x_i\}$ and $\{h_j\}$. $\{\theta_i\}$, $\{\eta_j\}$ and $\{W_{ij}\}$ are the parameters, they can be

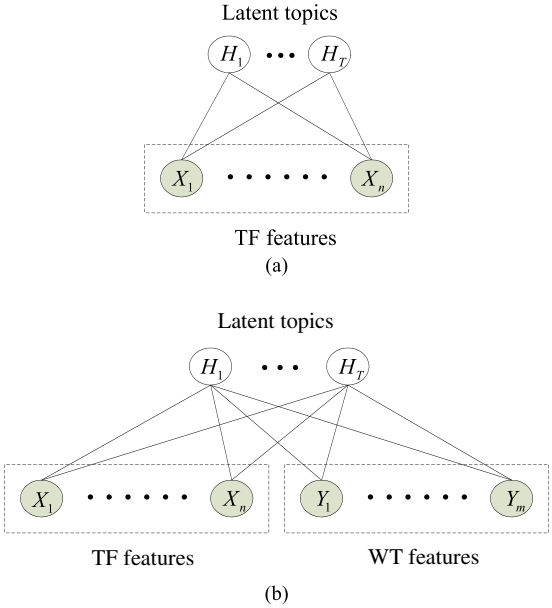


Fig. 3. Topologies of different harmonium models: (a) basic harmonium (b) DWH.

identified by learning algorithm. In above distributions the global partition function are not explicitly shown, which makes the harmonium learning more difficult. From the distributions, we can see that the data nodes the term $\{W_{ij}\}$ couples the data nodes x to the latent topics h. Through learning and inference, latent topics h will be harmonized with the observed data x so that h capture the semantics in x.

To generate a component-wise nonlinear projection from input space to output latent space, Gehler et al. (2006) extended the EFH model to RAP model that is a more general topology of the exponential family harmonium. RAP model couples latent topics to term counts using a conditional Poisson distribution involving a single weight matrix. They used conditional Poisson distribution for the TF feature and conditional binomial distribution for the latent topics as follows (Gehler et al., 2006):

$$p(x|h) = \prod_i \left(Poisson_{x_i}\left(\alpha_i + \sum_k W_{ik}h_k\right)\right) \tag{5}$$

$$p(h|x) = \prod_k \left(Binomial_{h_k}\left(\sigma\left(\tau_k + \sum_i W_{ik}x_i\right), M_k\right)\right) \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function, α_i is the log mean rate of the conditional Poisson distribution for term i, $\tau_k = \log(p_k/(1-p_k))$ (p_k is the probability of success), and M_k is the total number of samples for the conditional binomial distribution for topic k. The joint distribution over (x,h) can be expressed as

$$p(x,h) \propto \exp\left\{\sum_i(\alpha_i x_i - \log(\Gamma(x_i))) + \sum_k(\tau_k h_k - \log(\Gamma(h_k)) - \log(\Gamma(M_k - h_k))) + \sum_{ik} W_{ik}x_i h_k\right\} \tag{7}$$

where $\Gamma(\cdot)$ is the Gamma function. The marginal probability of nodes x is given by

$$p(x) \propto \exp\left\{\sum_i(\alpha_i x_i - \log(\Gamma(x_i))) + \sum_k\left(M_k \log\left(1 + \exp\left(\sum_i W_{ik}x_i + \tau_k\right)\right)\right)\right\} \tag{8}$$

RAP models the behavior that the values of the variables at the opposite layer shift the canonical parameters of the variables at the corresponding layer. The variation of $\{\tau_k\}$ decides the impact on the Poisson rate $\{\alpha_i\}$ with rate adapting property.

3.2. DWH model

Motivated by reference Xing et al. (2005) using DWH modeling the video data, we will present a new DWH model for document data in this section. Fig. 3(b) shows the architecture of DWH for document data that consists of two wings at the bottom layer. One wing represents the observed TF features $\{X_i\}$, and the other denotes the sampled WT features $\{Y_i\}$. Note that WT features $\{Y_i\}$ are linearly normalized corresponding to the TF features with normalization weight C that is used to balance the contributions between both types of features. The normalization is performed as follows:

$$Y_i(s) = C \times \frac{\mathrm{WTF}(s)}{\sum_{v=1}^{n_r \times n_r}\mathrm{WTF}(v)} \times \sum_{q=1}^{n}\mathrm{TF}(q) \tag{9}$$

where $\mathrm{WTF}(\cdot)$ represents the original WT feature units of one document and $\mathrm{TF}(\cdot)$ denotes the TF feature units of one document. Thus DWH integrates TF and WT features as low level features into latent topics as high level features to represent document semantics. These two types of features interact with each other through with the weighted matrixes.

In our DWH, we use conditional Poisson distribution for the TF feature like RAP model as follows:

$$p(x_i|h) = Poisson\left(x_i|\alpha_i + \sum_k W_{ik}h_k\right) \tag{10}$$

For WT feature, the WT feature y_j of the jth row of the lowest frequency subbands admits a conditional multivariate Gaussian distribution as follows:

$$p(y_j|h) = Gaussian\left(y_j|\Lambda_j(\beta_j + \sum_k U_{jk}h_k), \Lambda_j\right) \tag{11}$$

where both β_j and $\{U_{jk}\}$ are n_r-dimensional vectors, and so $\beta = (\beta_1, \ldots, \beta_{n_r})$ is a stacked vector with dimension $(n_r)^2$ and $U = [U_{jk}]$, a matrix of size $(n_r)^2 \times K$ where K is the total number of the latent topics, represents the weighted matrix coupling the WT features to the latent topics. Note that Λ_j is a covariance matrix with size $n_r \times n_r$, which, for simplicity, is set to identity matrix. Finally, the latent topics $\{H_k\}$ follow the conditional binomial distribution depending on a weighted combination of the TF features x and WT features Y in the following way:

$$p(h_k|x,Y) = Binomial\left(h_k|\sigma(\tau_k + \sum_i W_{ik}x_i + \sum_j U_{jk}y_j), M_k\right) \tag{12}$$

We then define the following joint distribution to be consistent with above conditional distributions

$$p(x,Y,h) \propto \exp\left\{\sum_i(\alpha_i x_i - \log(\Gamma(x_i))) + \sum_j \beta_j y_j - \sum_j \frac{y_j^2}{2} + \sum_k(\tau_k h_k - \log(\Gamma(h_k)) - \log(\Gamma(M_k - h_k))) + \sum_{ik} W_{ik}x_i h_k + \sum_{jk} U_{jk}y_j h_k\right\} \tag{13}$$

The marginal distribution over (x,Y) can be expressed as follows by marginalizing out the latent topics h in Eq. (13)

$$p(x,Y) \propto \exp\left\{\sum_i(\alpha_i x_i - \log(\Gamma(x_i))) + \sum_j \beta_j y_j - \sum_j \frac{y_j^2}{2} + \sum_k\left(M_k \log\left(1 + \exp\left(\sum_i W_{ik}x_i + \sum_j U_{jk}y_j + \tau_k\right)\right)\right)\right\} \tag{14}$$

The detailed derivation of Eq. (14) can be found in the Appendix. Likewise, in Eqs. (13) and (14) the global partition function is not explicitly shown.

From above probability distributions, we see that DWH model in this paper is an extension of RAP model. It inherits rate adapting property that is not only determined by TF features but also influenced by WT features. Thus the learned latent topics will capture more semantic information from documents to perform document data mining tasks.

3.3. Learning and inference

The parameters of DWH model including $\{\alpha_i\}$, $\{\beta_j\}$, $\{\tau_k\}$, $\{W_{ik}\}$ and $\{U_{jk}\}$ can be learned by maximizing the likelihood of the document data according to Eq. (14). Due to the complexity of the model, it is extremely difficult to obtain closed-form solution to the optimization problem. Thus we have to perform stochastic gradient ascent on the log-likelihood of data in iteration. The learn-

ing rules can be derived from log-likelihood of Eq. (14) in the following way:

$$\delta\alpha_i = \langle x_i \rangle_{\tilde{p}} - \langle x_i \rangle_p \tag{15}$$

$$\delta\beta_j = \langle y_j \rangle_{\tilde{p}} - \langle y_j \rangle_p \tag{16}$$

$$\delta\tau_k = M_k\left(\langle \sigma(\bar{h}_k + \tau_k)\rangle_{\tilde{p}} - \langle \sigma(\bar{h}_k + \tau_k)\rangle_p\right) \tag{17}$$

$$\delta W_{ik} = M_k\left(\langle x_i \sigma(\bar{h}_k + \tau_k)\rangle_{\tilde{p}} - \langle x_i \sigma(\bar{h}_k + \tau_k)\rangle_p\right) \tag{18}$$

$$\delta U_{jk} = M_k\left(\langle z_j \sigma(\bar{h}_k + \tau_k)\rangle_{\tilde{p}} - \langle z_j \sigma(\bar{h}_k + \tau_k)\rangle_p\right) \tag{19}$$

where $\bar{h}_k = \sum_i W_{ik} x_i + \sum_j U_{jk} y_j$, $\langle\cdot\rangle_{\tilde{p}}$ represents expectation under empirical distribution (i.e. data average), and $\langle\cdot\rangle_p$ denotes the expectation under model distribution of the harmonium at the current values of the parameters. However, due to the presence of global partition function in the log-likelihood of Eq. (14), it is hard to directly estimate the model expectation $\langle\cdot\rangle_p$. There are many approximate inference methods to estimate this expectation such as contrastive divergence (CD) learning (Hinton, 2002; Welling & Hinton, 2002), mean field (MF) approximation (Xing, Jordan, & Russell, 2003), and Langevin method (Murray & Ghahramani, 2004). CD learning algorithm is proposed to approximate exact gradient ascent search. MF is an alternative method that approximates the model distribution through a factorized form as a product of marginal distributions over clusters of variables (Xing et al., 2003; Xing et al., 2005). With inheriting all the proposal moves of Langevin Monte Carlo method, the Langevin approach uses noisy steepest ascent to avoid local optima as well as taking advantage of the gradient information (Murray & Ghahramani, 2004). In this section we only introduce the details how to use CD learning algorithm for DWH training. We also compare the performance of different algorithms for learning and inference in application examples.

In each step of gradient ascent, CD starts from a separate Gibbs sampler defined by Eqs. (10)–(12) at a data-case, runs it for only a few steps and then uses these samples to approximate the model expectation $\langle\cdot\rangle_p$ together with computing the gradient through Eqs. (15)–(19). It has been proved that the parameters through this learning process will converge to the maximum likelihood estimation (Welling & Hinton, 2002). The whole learning procedures are described as follows.

CD learning procedure for DWH model:

Initialize the parameters $\{\alpha_i\}, \{\beta_j\}, \{\tau_k\}, \{W_{ik}\}$ and $\{U_{jk}\}$
Loop until convergence (by setting thresholds)
(1) Sample the latent topics given the input data using Eq. (12)
(2) Resample the corresponding TF data-case given the sampled values of the latent topics using Eq. (10)
(3) Resample the corresponding WT data-case given the sampled values of the latent topics using Eq. (11)
(4) Compute the data averages and sample averages in Eqs. (15)–(19)
(5) Update the parameters using the gradient ascent rules in Eqs. (15)–(19)
End Loop
Return $\{\alpha_i\}, \{\beta_j\}, \{\tau_k\}, \{W_{ik}\}, \{U_{jk}\}$

After learning and inference, all the document data can be mapped to low dimensional latent representations, and then DWH model is ready to perform various document data mining tasks.

4. Relationship to video DWH

Building on the early work of Welling et al. (2004), Xing et al. (2005) and Yang et al. (2008) developed the DWH models for video data (VDWH) with applications to classification, retrieval, image annotation and video classification tasks by projecting two types of features in video shot, text and image, into low dimensional latent space and with promising results. In fact, our DWH model for document data (DDWH) is an analogy to VDWH, and there is a deep connection between DDWH and VDWH. The TF features in DDWH are analogous to the text features in VDWH, and the WT features in DDWH can be seen as the color histogram features in the keyframe of each video shot in VDWH. In the video shot, each keyframe is evenly divided into a grid of fixed size rectangular regions, and each region is represented by the color histogram vector. Likewise, for document data, the term graph also can be treated as an image that describes the associations among the terms in the document. The low frequency subbands generated from the 2-D wavelet transform are histograms of the compressed regions of the document. Therefore, DDWH and VDWH are able to be handled in a union frame because of their deep connections.

5. Applications

In this section, we evaluate DWH-WTF model on different datasets and for different data mining tasks, including classification and retrieval. Here, we use the name DWH-WTF to denote our algorithm since the DWH is aided by wavelet transform features.

5.1. Document classification

For document classification, we compiled the documents referenced by the Open Directory Project[1] on health conditions and diseases. Knowing more details, readers are referred to the website.[2] The original collection contained 5916 documents in 21 top-level classes. For each top-level class, we first moved all the documents in its sub-class to the top-level class and removed all the sub-classes. We then removed all the short documents with the size less than 50 terms. Thus, 4527 documents were left with us in 19 classes that had more than 60 files. The details of this dataset were summarized in Table 1. The parameters of DWH-WTF algorithm in the simulation were set as follows. The number of selected terms$n$ was equal to 2048. The number of wavelet transform scales$r$ was set to 6, so the length of WT feature vector was 1024. The normalization weight$C$ was set to 0.2. The learning rate and the momentum term to speed up the convergence in DWH-WTF model were set to 0.01 and 0.95, respectively. The DWH-WTF based on 200,000 learning iterations using gradient ascent on mini-batches of 100 random training samples per iteration. To perform classification task, we use DWH-WTF to project each data point into a lower dimensional latent space. We then held out 90% of the entire data corpora for training purpose and 10% for testing the performance. We used MATLAB*Arsenal*[3] toolbox to learn a KNN classifier on the training data.

In order to investigate the performance of harmonium model using multiple features, we first compared the DWH-WTF model with the RAP model (Gehler et al., 2006) using single features (i.e. term counts). Table 2 summarized the accuracy performance of DWH-WTF and RAP model with the number of latent topics from 5 unto 25 at increments of 5. It is observed that both methods perform better and better with the increase of the number of latent topics from 5 to 15, and the accuracy results of them deteriorate in

[1] http://www.dmoz.org/.
[2] http://www.di.uniba.it/~malerba/software/webclass/WebClassIII.htm.
[3] http://www.informedia.cs.cmu.edu/yanrong/MATLABArsenal/MATLABArsenal.zip.

Table 1
Details of data set for document classification.

Class	Number of articles	Topic
1	89	Allergies
2	661	Cancer
3	499	Cardiovascular disorders
4	101	Communication disorders
5	163	Digestive disorders
6	218	Endocrine disorders
7	76	Wounds and injuries
8	136	Sleep disorders
9	110	Skin disorders
10	198	Respiratory disorders
11	62	Nutrition and metabolism disorders
12	998	Neurological disorders
13	322	Musculoskeletal disorders
14	311	Infectious diseases
15	186	Immune disorders
16	85	Genitourinary disorders
17	91	Genetic disorders
18	98	Food and water borne
19	123	Eye disorders

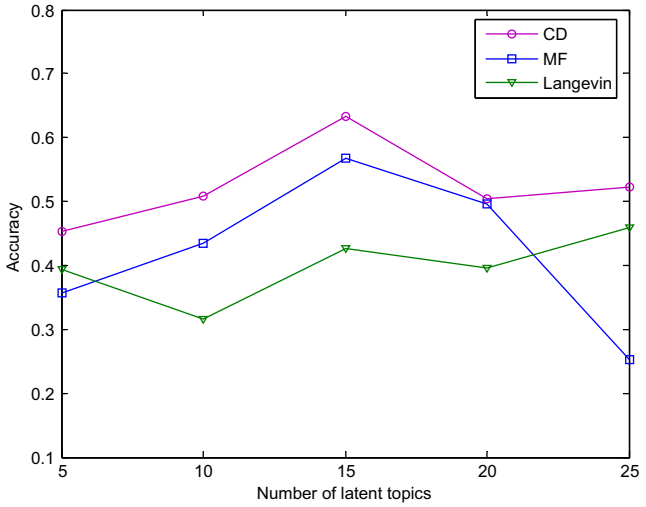


Fig. 4. Classification results of DWH-WTF using different learning methods.

Table 2
Classification accuracy of different models (%).

Method	Number of latent topics				
	5	10	15	20	25
DWH-WTF	45.35	50.89	63.27	50.44	52.21
RAP	38.27	50.22	55.31	43.36	45.35

Table 3
Classification accuracy of different models.

Method	DWH-WTF	RAP	LSI
Accuracy (%)	63.27	55.31	58.62

a small significant rate when continuing to increase the number of latent topics. Thus, using 15 latent topics seems to be a good choice of the number of latent topics that delivers superior performance for both DWH-WTF and RAP. DWH-WTF consistently performs better than RAP, which is believed that by including WT features to harmoniums indeed captures additional discriminate factors for the classification task. We then compared the accuracy results of harmonium models with that of LSI model (Deerwester & Dumais, 1990) under 15 latent topics summarized in Table 3. It is also noted that DWH-WTF delivers about 5% improvement of classification accuracy compared with LSI.

We also studied the effect of different learning approaches for DWH-WTF inference based on classification results. Fig. 4 together with Table 4 shows the accuracy results of DWH-WTF model implemented using different approximate inference methods with different numbers of latent topics. From Fig. 4, it is observed that Langevin learning method performs worst compared with other two approaches when the number of latent topics is between 10 and 20. Contrastive divergence (CD) learning delivers significantly better results than Langevin and mean field (MF) sampling with 10 latent topics and 15 latent topics. In Table 4, it is observed that CD learning provides at least 6% improvement of accuracy compared with MF, and provides about 20% improvement of accuracy compared with Langevin with 15 latent topics. Therefore, in our study CD appears to be the best choice for the learning and inference of DWH-WTF model in terms of classification performance.

Table 4
Classification accuracy of DWH-WTF using different learning methods (%).

Method	Number of latent topics learning				
	5	10	15	20	25
CD	45.35	50.89	63.27	50.44	52.21
MF	35.62	43.36	56.86	49.56	25.22
Langevin	39.38	31.64	42.70	39.60	46.02

5.2. 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 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. In this study, we used an entire document as a query.

To perform document retrieval, the document database, "Html_CityU1", which consists of 26 categories (Chow & Rahman, 2009), was used in this application example. Each category includes 400 documents making a total number of 10,400 documents. The corpus was split into a training set and a test set that was used for queries. 1040 test documents were randomly selected from the 26 categories, i.e. 26×40. The remaining 9360 documents were used for training. In order to provide a more real-life testing platform, we established this database consisting of documents with size ranged from few hundred words to over 20 thousand words. For each category, 400 documents were retrieved from "Google" using a set of keywords. Some of the keywords are shared among different categories, but the set of keywords for a category is different from that of other categories. The details of this dataset were summarized in Table 5. The database can be found online at www.ee.cityu.edu.hk/~twschow/Html_CityU1.rar for other researchers.

The parameters of DWH-WTF algorithm in this section were set as follows. The number of selected terms$n$ was equal to 4096. The number of wavelet transform scales$r$ was set to 7, so the length of WT feature vector was 1024. The normalization weight$C$ was set to

Table 5
Details of data set for document retrieval.

Category	Number of articles	Keywords
1	400	Bank + Money + Transfer
2	400	Bush + Iraq + War + Saddam
3	400	Cook + Food + Instruction
4	400	Cosmetic + Beauty + Fashion
5	400	Cricket + Batsman + Bowler
6	400	Dog + Pet + Home + Guard
7	400	Face + Image + Recognition
8	400	Garden + Flower
9	400	Hitler + Germany + Nazis
10	400	Html + Syntax + Tag
11	400	Java + Jsp + Servlet
12	400	Law + Thermodynamics
13	400	Middle + East + Crisis
14	400	Mongolian + Invasion
15	400	Mountain + Hiking + Trail
16	400	Mountain + Skating + Ice
17	400	Nuclear + Energy + Physics
18	400	Plant + Care + Cultivation
19	400	River + Dam + Hydropower
20	400	River + Water + Irrigation
21	400	Second + Two + World + War
22	400	Sleep + Medical + Problem
23	400	Sports + Football + News
24	400	Visual + Basic
25	400	Visual + c++
26	400	Web + Search + Engine

0.3. The learning rate and the momentum term to speed up the convergence in DWH-WTF model were set to 0.01 and 0.95, respectively. The DWH-WTF based on 200,000 learning iterations using gradient ascent on mini-batches of 200 random training samples per iteration. To quantify the retrieval results, we used averaged precision and recall values for each query document from the test set. The precision and recall measure are defined as follows:

$$Precision = \frac{No.\ of\ correctly\ retrieved\ documents}{No.\ of\ total\ retrieved\ documents} \tag{20}$$

$$Recall = \frac{No.\ of\ correctly\ retrieved\ documents}{No.\ of\ total\ documents\ in\ relevant\ category} \tag{21}$$

Based on above precision and recall measures, to evaluate the effect of different numbers of latent topics, a measure named "area under the precision-recall curve" (AUC) as a function of the number of latent topics can be simply defined as follows:

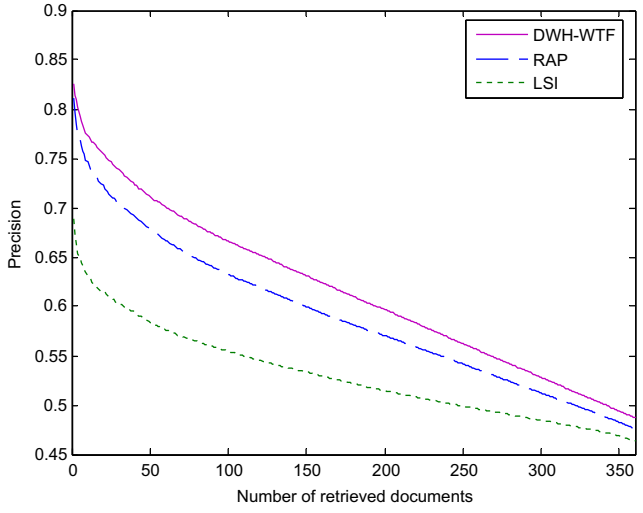
$$\mathrm{AUC}(L) = \sum_{i_A=2}^{n_{\max}} \frac{(P_L(i_A) + P_L(i_A - 1)) \times (R_L(i_A) - R_L(i_A - 1))}{2} \tag{22}$$

where L represents the number of latent topics, $n_{\max}$ denotes the maximum number of retrieved documents, $P_L(i_A)$ and $R_L(i_A)$ represent the precision and recall values with i_A documents retrieved corresponding to the number of latent topics L.

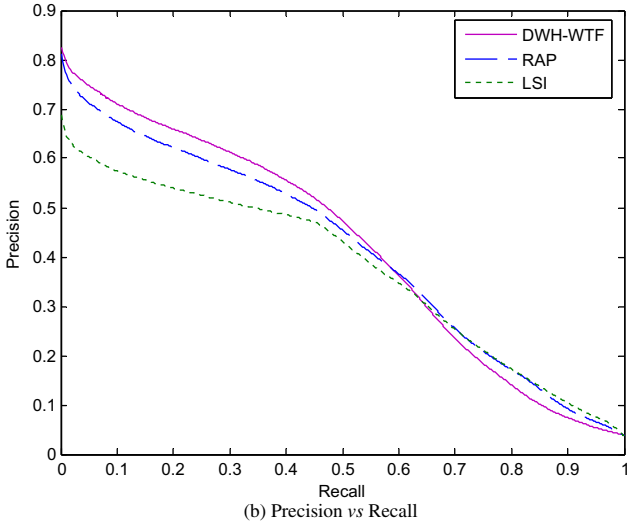
Table 6 summarized the AUC values of DWH-WTF and RAP model with the number of latent topics from 5 unto 25 at increments of 5 in order to evaluate the performance of harmonium model using multiple features. It is observed that both methods perform better and better with the increase of the number of latent topics from 5 to 10, and the AUC results of them deteriorate in a small significant rate when continuing to increase the number of latent topics. Based on this observation, using 10 latent topics seems to be a good choice of the number of latent topics that delivers superior performance for both DWH-WTF and RAP. DWH-WTF performs better than RAP except for the case of using 5 latent topics. We then compared the performance of different models under 10 latent topics. Fig. 5(a) shows the precision results when the retrieved documents, the most similar training documents from the dataset for every query, vary from 1 to 360. It is observed that DWH-WTF delivers the best precision results than other models, and RAP model (Gehler et al., 2006) performs better than LSI model (Deerwester & Dumais, 1990). As shown in Fig. 5(b) for the sketch of the relationship between precision and recall, DWH-WTF exhibits significantly superior performance compared with other models when the recall value is smaller than about 0.65. While the recall value being in the range from 0.65 to 1.0, LSI and RAP performs better than DWH-WTF in a small rate. This interesting observation is caused by the overlapping of topics combinations and adding terms associations is not helpful for enhancing the performance when retrieving so many documents. In Table 7, we quantitatively listed the precision and recall results of different models under 10

Table 6
AUC values of different models.

Method	Number of latent topics				
	5	10	15	20	25
DWH-WTF	0.33	0.42	0.40	0.32	0.30
RAP	0.35	0.41	0.36	0.29	0.26



(a) Precision *vs* Number of retrieved documents



(b) Precision *vs* Recall

Fig. 5. Performance of different models based on retrieval results.

Table 7
Retrieval results of different models.

Method	Number of retrieved documents							
	1	10	40	360	1	10	40	360
	Precision (%)				Recall (%)			
DWH-WTF	82.60	77.37	72.36	48.69	0.23	2.15	8.04	48.69
RAP	81.06	74.62	69.19	47.61	0.23	2.07	7.69	47.61
LSI	68.94	63.07	59.38	46.32	0.19	1.75	6.60	46.32

latent topics with the number of retrieved documents from 1 to 360. DWH-WTF provides about 3% improvement of the precision in average compared with RAP model with only TF features. It is also believed that the improvement is resulted by including the WT features to harmoniums. Harmonium models are able to deliver at least 10% improvement of the precision compared with LSI model. Similar results are also shown in the recall results among different models.

We also studied the effect of different learning approaches for DWH-WTF inference based on AUC results. Fig. 6 together with Table 8 shows the AUC values of DWH-WTF model using different approximate inference methods with different numbers of latent topics. From Fig. 6, it is observed that Contrastive divergence (CD) learning delivers better results than Langevin and mean field (MF) methods. MF method performs better and better with the increase of the number of latent topics. In Table 8, it is observed that CD and Langevin obtain similar AUC values under 10 latent topics, whilst MF and CD deliver similar AUC results under 25 latent topics. Therefore, CD appears to be the best choice for the learning and inference of DWH-WTF model in terms of AUC results based on our dataset.

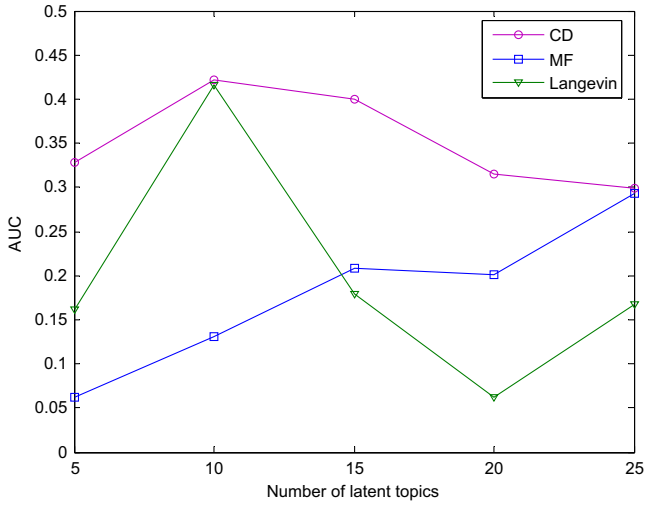


Fig. 6. AUC results of DWH-WTF using different learning methods.

Table 8
AUC values of DWH-WTF using different learning methods.

Learning method	Number of latent topics				
	5	10	15	20	25
CD	0.33	0.42	0.40	0.32	0.30
MF	0.06	0.13	0.21	0.20	0.29
Langevin	0.16	0.42	0.18	0.06	0.17

6. Conclusions

A novel dual wing harmonium (DWH) model using multiple features is proposed for modeling the document data with applications to classification and retrieval tasks. This DWH model integrates multiple document features into low dimensional semantic space with few latent topics for document representation. First, TF features are extracted from documents by using weighted feature extraction method. We then formed an undirected graph to represent each document based on extracted terms as nodes. In order to improve the computational efficiency, term graph compression is conducted by employing 2-D wavelet transform. Thus, the term graph becomes more compact while preserving the basic graph structure. Multiple features that consist of TF features and WT features constructed by stacked low-pass subbands are then as inputs of DWH model. DWH model extends the basic RAP model to two wings by using different conditional probability distributions. It does not only include the properties of RAP, but also contains capability to capture terms associations information. Two application examples corroborate the efficiency of our method. Our future work will include enriching our database to support the assumption on terms associations and further studying the stability of inference algorithms of harmonium models.

Acknowledgments

The authors would like to express our thanks to Peter V. Gehler for his patient explanation to their work. Also, we must thank Xiyuan Lu for discussions on wavelets and thank Jun Yang for sending original codes of their work.

Appendix A

This is to the derivation of the marginal distribution over (x, Y) in DWH model. We defined the joint distribution over (x, Y, h) as mentioned in Section 3.2 in the following way,

$$p(x,Y,h) \propto \exp\left\{\sum_i (\alpha_i x_i - \log(\Gamma(x_i))) + \sum_j \beta_j y_j - \sum_j \frac{y_j^2}{2} + \sum_k (\tau_k h_k - \log(\Gamma(h_k)) - \log(\Gamma(M_k - h_k))) + \sum_{ik} W_{ik} x_i h_k + \sum_{jk} U_{jk} y_j h_k\right\}$$

On the other hand, the latent topics $\{H_k\}$ follow the conditional binomial distribution depending on a weighted combination of the TF x and binary TCF Y as follows,

$$\begin{aligned} p(h_k|x,Y) &= Binomial\left(h_k|\sigma\left(\tau_k + \sum_i W_{ik}x_i + \sum_j U_{jk}y_j\right), M_k\right) \\ &= (M_k h_k)\left(\sigma\left(\tau_k + \sum_i W_{ik}x_i + \sum_j U_{jk}y_j\right)\right)^{h_k} \\ &\quad \times \left(\sigma\left(\tau_k + \sum_i W_{ik}x_i + \sum_j U_{jk}y_j\right)\right)^{(M_k - h_k)} \end{aligned}$$

where,

$$\binom{M_k}{h_k} = \frac{\Gamma(M_k)}{\Gamma(h_k)\Gamma(M_k - h_k)} = \frac{M_k!}{h_k!(M_k - h_k)!}$$

According to the definition of conditional probability distribution, we are ready to derive the marginal distribution over (x, Y) as

$$
\begin{aligned}
p(x,Y) &= \frac{p(x,Y,h)}{p(h|x,Y)} = \frac{p(x,Y,h)}{\prod_k p(h_k|x,Y)} \\
&\propto \frac{\exp\left\{\sum_i(\alpha_i x_i - \log(\Gamma(x_i))) + \sum_j \beta_j y_j - \sum_j \frac{y_j^2}{2} + \sum_k h_k\left(\tau_k + \sum_i W_{ik}x_i + \sum_j U_{jk}y_j\right) - \sum_k(\log(\Gamma(h_k)) + \log(\Gamma(M_k - h_k)))\right\}}{\prod_k\left(\binom{M_k}{h_k}\left(\sigma\left(\tau_k + \sum_i W_{ik}x_i + \sum_j U_{jk}y_j\right)\right)^{h_k}\left(\sigma\left(\tau_k + \sum_i W_{ik}x_i + \sum_j U_{jk}y_j\right)\right)^{(M_k - h_k)}\right)} \\
&= \exp\left\{-\sum_k \log(\Gamma(M_k))\right\} \times \exp\left\{\sum_i(\alpha_i x_i - \log(\Gamma(x_i))) + \sum_j \beta_j y_j - \sum_j \frac{y_j^2}{2} + \sum_k\left(M_k \log\left(1 + \exp\left(\sum_i W_{ik}x_i + \sum_j U_{jk}y_j + \tau_k\right)\right)\right)\right\} \\
&\propto \exp\left\{\sum_i(\alpha_i x_i - \log(\Gamma(x_i))) + \sum_j \beta_j y_j - \sum_j \frac{y_j^2}{2} + \sum_k\left(M_k \log\left(1 + \exp\left(\sum_i W_{ik}x_i + \sum_j U_{jk}y_j + \tau_k\right)\right)\right)\right\}
\end{aligned}
$$

which is exactly consistent with Eq. (14).

References

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review, 37*(4), 573–595.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Chien, J. T., & Wu, M. S. (2008). Adaptive Bayesian latent semantic analysis. *IEEE Transaction on Audio, Speech, and Language Processing, 16*(1), 198–207.

Chow, T. W. S., & Rahman, M. K. M. (2009). Multi-layer SOM with tree structured data for efficient document retrieval and plagiarism detection. *IEEE Transactions on Neural Networks, 20*(9), 1385–1402.

Deerwester, S., & Dumais, S. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science, 41*(6), 391–407.

Dorogovtsev, S. N., & Mendes, J. F. F. (2001). Language as an evolving word web. *Proceedings of the Royal Society B: Biological Sciences, 268*(1485), 2603–2606.

Fan, J., & Wang, X. F. (2005). A wavelet view of small-world networks. *IEEE Transaction on Circuits and Systems—II: Express Briefs, 52*(5), 238–241.

Gehler, P., Holub, A., & Welling, M. (2006). The rate adapting Poisson model for information retrieval and object recognition. In *Proceedings of the 23rd international conference on machine learning, Pittsburgh, PA.*

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation, 14*, 1771–1800.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international SIGIR conference*.

I Cancho, R. F., & Sole, R. V. (2001). The small-world of human language. *Proceedings of the Royal Society B: Biological Sciences, 268*(1482), 2261–2265.

Lewis, A. S., & Knowles, G. (1992). Image compression using the 2-D wavelet transform. *IEEE Transactions on Image Processing, 1*(2), 244–250.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11*(7), 674–693.

Murray, I., & Ghahramani, Z. (2004). Bayesian learning in undirected graphical models: Approximate MCMC algorithms. In M. Chickering & J. Halpern (Eds.), *Proceedings of the 20th annual conference on uncertainty in artificial intelligence* (pp. 392–399). Banff, Canada: AUAI Press.

Ouhsain, M., & Ben Hamza, A. (2009). Image watermarking scheme using nonnegative matrix factorization and wavelet transform. *Expert Systems with Applications, 36*(2), 2120–2128.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Salton, G., & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513–523.

Salton, G., & McGill, M. (Eds.). (1983). *Introduction to modern information retrieval*. McGraw-Hill.

Schenker, A., Last, M., Bunke, H., & Kandel, A. (2003). Classification of web document using a graph model. In *Proceedings of the 7th international conference on document analysis and recognition (ICDAR'03)*.

Schenker, A., Last, M., Bunke, H., & Kandel, A. (2004). Classification of web documents using graph matching. *International Journal of Pattern Recognition and Artificial Intelligence, 18*(3), 475–496.

Sengur, A. (2008). Wavelet transform and adaptive neuro-fuzzy inference system for color texture classification. *Expert Systems with Applications, 34*(3), 2120–2128.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. *In Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1: Foundations*. Cambridge, MA: MIT Press.

Welling, M., & Hinton, G. E. (2002). A new learning algorithm for mean field Boltzmann machines. In *ICANN '02: Proceedings of the international conference on artificial neural networks* (pp. 351–357). London: Springer-Verlag.

Welling, M., Rosen-Zvi, M., & Hinton, G. (2004). Exponential family harmoniums with an application to information retrieval. *Advances in neural information processing systems* (Vol. 17, pp. 1481–1488). Cambridge, MA: MIT Press.

Xing, E., Yan, R., & Hauptmann, A. (2005). Mining associated text and images with dual-wing harmoniums. In *Proceedings of the conference on uncertainty in artificial intelligence*.

Xing, E., Jordan, M., & Russell, S. (2003). A generalized mean field algorithm for variational inference in exponential families. In *Uncertainty in artificial intelligence (UAI2003)* (pp. 583–591). San Francisco, CA: Morgan Kaufmann Publishers.

Yang, J. et al. (2008). Harmonium models for video classification. *Statistical Analysis and Data Mining, 1*, 23–37.

Zobel, J., & Moffat, A. (1998). Exploring the similarity space. *ACM SIGIR Forum, 32*(1), 18–34.