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.

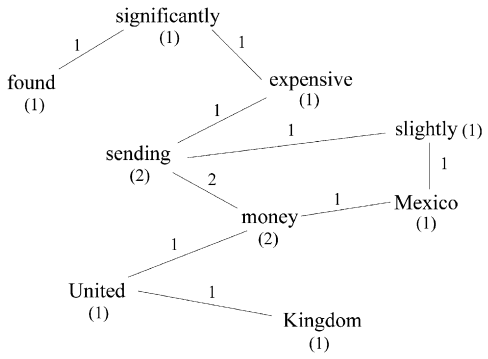

Fig. 1. Undirected graph as an example: "we found it significantly more expensive for sending money to Mexico, but slightly less for sending money to the United Kingdom". (Here, "we", "it", "more", "for", "to", "but", "less", "for", "the" are stop words that are removed.)

Here, "connected" means that two terms are adjacent to each other without distinguishing the term sequence.

2.2. Term-frequency (TF) feature extraction

First, extract all the words from all documents except for stop words in a database and apply stemming algorithm to each word. Here, Porter stemming algorithm [17] 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 and the document-frequency f^t_d. 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 [18]

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

where the inverse-document-frequency $idf=\log_2(N/f^t_d)$, 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.

2.3. Term connection frequency (TCF) feature extraction

Feature extraction of TCF is based on the word vocabulary, which is constructed in Section 2.2. We use terms in the word vocabulary to build an undirected graph for each document. Based on graph representation, if we directly use graph matching methods to calculate the semantic similarity like Ref. [16], much time and storage space will be wasted for large datasets because the adjacent matrix of each document is so sparse. The adjacent matrix $A^l(l=1,2,\ldots,N)$ for graph G^l is denoted by $A^l=[A^l_{ij}]_{n\times n}$ where $A^l_{ij}=f^{l,tc}_{ij}$ represents TCF between term i and term j in document l. Then, we calculate the total TCF adjacent matrix for all the documents (i.e. $A=\sum_{l=1}^{N}A^l$). We also store the document frequency $f^{tc}_{d,ij}$ for term connection between term i and term j in the database. We then use the same weighting measure to calculate the weight of each term connection for a pair of terms.

$$W^{tc}_{ij}=\sqrt{f^{tc}_{ij}}\times idf^{tc}_{ij} \tag{2}$$

where the inverse-document-frequency $idf^{tc}_{ij}=\log_2(N/f^{tc}_{d,ij})$. Then, we sort the term connections by using the weights in descending order and select the first m term connections.

2.4. Summary on multiple features extraction framework

Multiple features extraction aims to provide inputs with more semantic information for the DWH model. The overall procedure of extracting multiple features is summarized as follows.

(1) Extract words from all the documents in the corpus excepting for stop words and apply stemming to each word. Calculate the weight of each word according to Eq. (1), and select the first n words to construct TF-based vocabulary.
(2) Build graph for each document using selected words as nodes and calculate the total adjacent matrix A. Select the first m term connections (or the indexes of edges in graph) based on Eq. (2) to construct TCF-based vocabulary.
(3) Calculate TF histogram and TCF histogram for each document. Each element of the histogram indicates the number of times that the corresponding term or term connection appears in a document.
(4) Save the multiple features (TF and TCF) for each document as inputs for DWH model.

3. Dual wing harmonium model for document data

The original harmonium model based on harmonium theory [19] refers to a family of bipartite undirected graphical models. Fig. 2(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 the 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 $p(x|h)$ $(p(x|h)=\Pi_i(x_i|h))$ and $p(h|x)$ $(p(h|x)=\Pi_j p(h_j|x))$.

3.1. EFH model

EFH model introduced by Welling et al. [9], 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 [9,12]:

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

$$p(h|x)=\Pi_j p(h_j|x)\propto\Pi_j\ \exp\left\{\left(\eta_j+\sum_i W_{ij}f(x_i)\right)g(h_j)\right\} \tag{4}$$

$$p(x,h)\propto\ \exp\left\{\sum_i\theta_i f(x_i)+\sum_j\eta_j g(h_j)+\sum_{ij}W_{ij}f(x_i)g(h_j)\right\} \tag{5}$$

where $\{f(x_i)\}$ and $\{g(h_j)\}$ are the sufficient statistics of nodes $\{x_i\}$ and $\{h_j\}$. $\{\theta_i\}$, $\{\eta_j\}$ and $\{W_{ij}\}$ are the parameters, they can be identified by learning algorithm. In the above distributions the global partition function is 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.

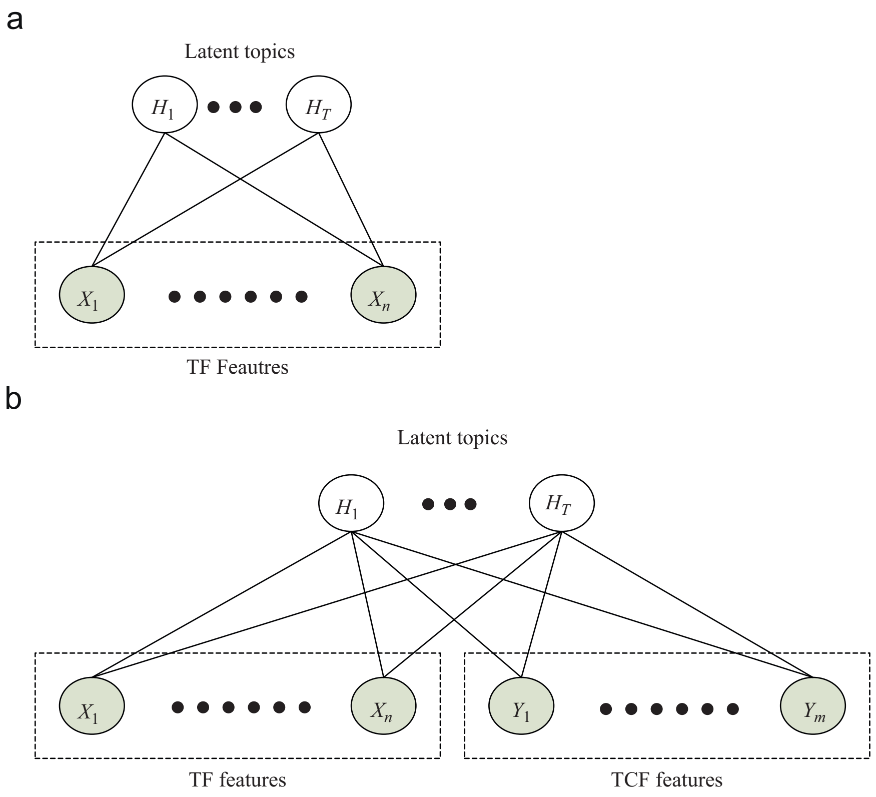

Fig. 2. Topologies of different harmonium models: (a) basic harmonium and (b) DWH.

Through learning and inference, latent topics h will be harmonized with the observed data x so that h capture the semantics in x.

3.2. RAP model

To generate a component-wise nonlinear projection from input space to output latent space, Gehler et al. [10] 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 [10].

$$p(x|h) = \Pi_i \left(Poisson_{x_i} \left(\alpha_i + \sum_k W_{ik} h_k \right) \right) \tag{6}$$

$$p(h|x) = \Pi_k \left(Binomial_{h_k} \left(\sigma \left(\tau_k + \sum_i W_{ik} x_i \right), M_k \right) \right) \tag{7}$$

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

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

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

$$p(x) \propto \exp \left\{ \sum_i (\alpha_i x_i - \log(\Gamma(x_i))) + \sum_k \left(M_k \log \left(1 + \exp \left(\sum_i W_{ik} x_i + \tau_k \right) \right) \right) \right\} \tag{9}$$

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.

3.3. DWH model

For video and image applications, Xing et al. [11] and Yang et al. [12] proposed a DWH for the fusion of features from multiple data sources including text features and image features. In their DWH model, the authors directly treated the term counts via Bernoulli distribution and the whole image color histogram via a multivariate Gaussian distribution, and then multiple features were projected into the latent space with low dimension. This new fusion strategy performs well for image annotation and video classification. Motivated by Ref. [11] using DWH modeling the video data, we will present a new DWH model for document data in this section. Fig. 2(b) shows the architecture of DWH for document data that consists of two wings at the bottom layer. One wing represents the observed TF feature $\{X_i\}$, and the other denotes the sampled TCF feature $\{Y_i\}$. Thus DWH integrates TF and TCF 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 TCF feature, we first binarize the TCF feature extracted in Section 2.3 indicating the presence or absence of connection between two terms. Through weighted selecting method for TCF, it is enough to discriminate the intro-information among different classes of documents. We do not put strong emphasis on this issue because in our applications this binary state of term connection still captures the similarity approximate to use counts of term connection. Then we use conditional Bernoulli distribution for the binary TCF feature as follows:

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

where $\{U_{jk}\}$ represents the weighted matrix coupling the TCF feature to latent topics. Finally, the latent topics $\{H_k\}$ follow the conditional binomial distribution depending on a weighted combination of the TF x and binary TCF Y in the following way:

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

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

$$\begin{aligned} p(x,Y,h) \propto \exp\Bigg\{ & \sum_i(\alpha_i x_i - \log(\Gamma(x_i))) + \sum_j \beta_j y_j \\ & + \sum_k(\tau_k h_k - \log(\Gamma(h_k)) - \log(\Gamma(M_k - h_k))) \\ & + \sum_{ik} W_{ik}x_i h_k + \sum_{jk} U_{jk}y_j h_k \Bigg\} \end{aligned} \tag{13}$$

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

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

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 the 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 TCF features. Thus the learned latent topics will capture more semantic information from documents to perform document data mining task.

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

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

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

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

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

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

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 learning [20,21], mean field (MF) approximation [22], and Langevin method [23]. 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 the clusters of variables [11,22]. 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 [23]. 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 Section 6.

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 [21]. 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 TCF 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 ascentrules 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. Here, we summarize the whole implementation framework for document retrieval system as an application example shown in Fig. 3.

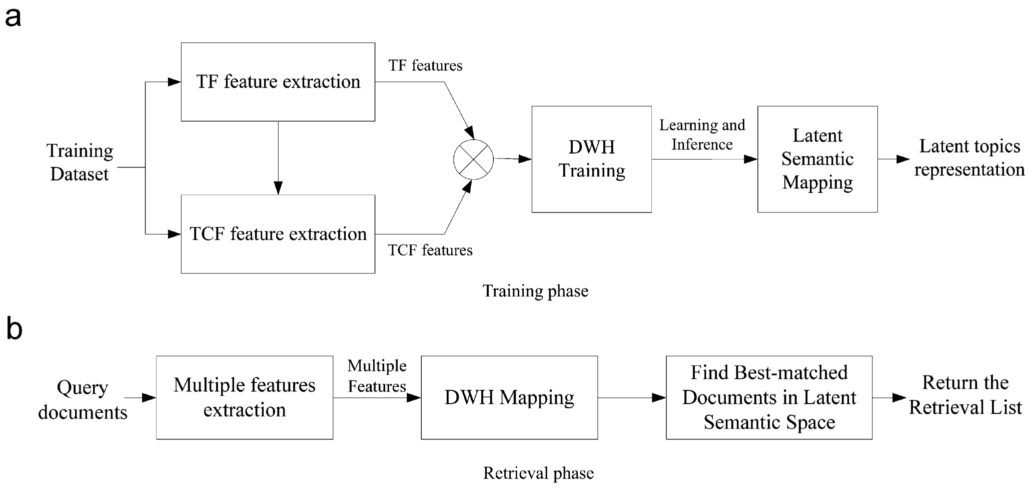

Fig. 3. Implementation framework for document retrieval system. (a) Training phase. (b) Retrieval phase.

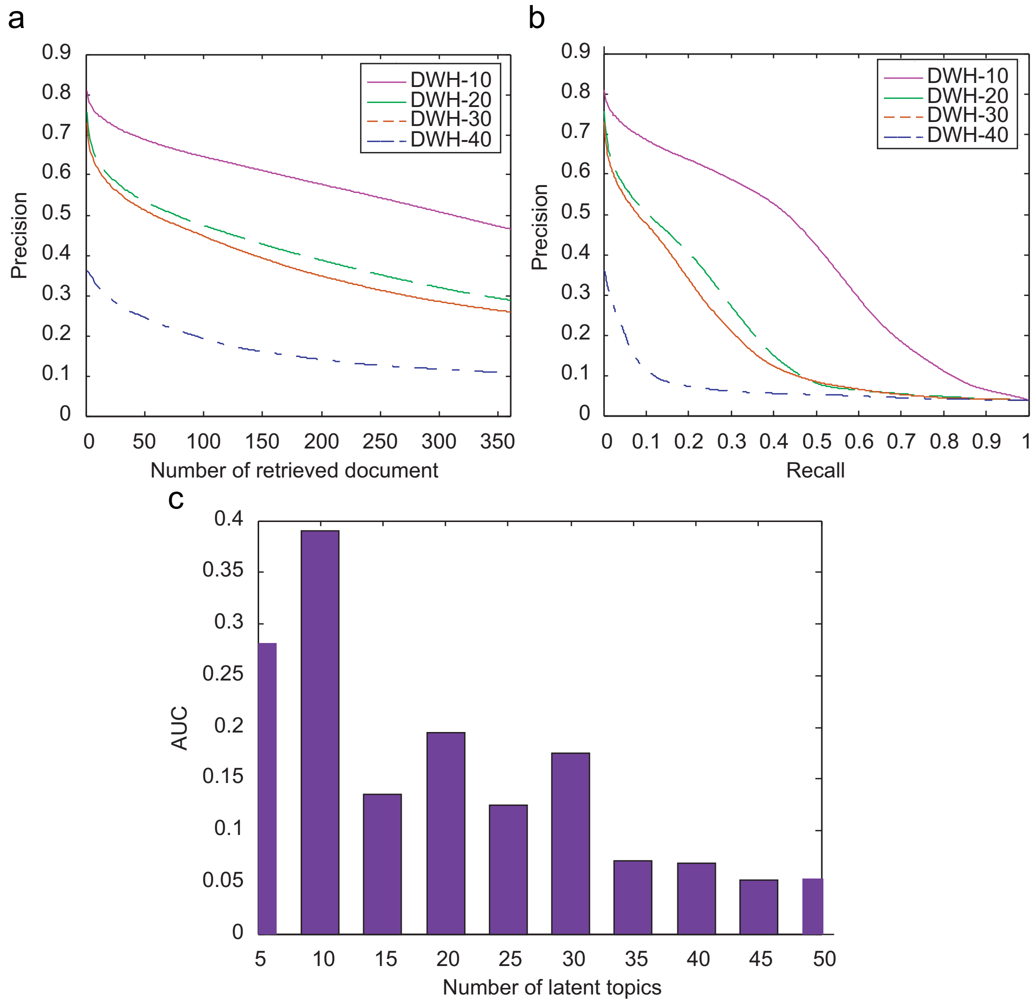

Fig. 4. Performance of DWH model based on retrieval results. (a) Precision vs number of retrieved document. (b) Precision vs recall. (c) AUC vs number of latent topics.

5. Experimental results and discussions

5.1. Database and experimental setup

In this study, the document database, "Html_CityU1", which consists of 26 categories [1], were used for all simulations. 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 is used for query; 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 docu-

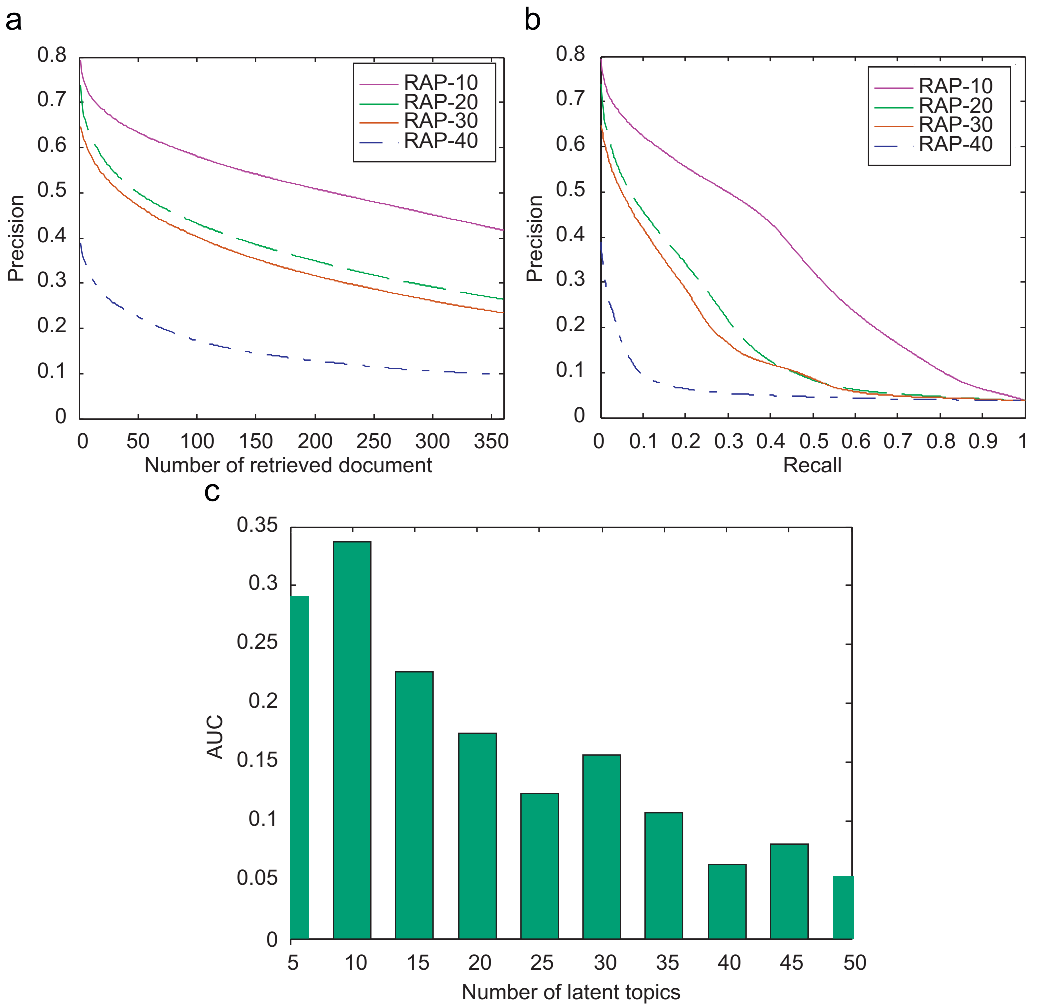

Fig. 5. Performance of RAP model based on retrieval results. (a) Precision vs number of retrieved document. (b) Precision vs recall. (c) AUC vs number of latent topics.

ments 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 database can be found online at "www.ee.cityu.edu.hk/~twschow/Html_CityU1.rar" for other researchers. After DWH training, the test set was used to verify the performance of this work. All the simulations were performed on a PC with Intel Core-2 2.13 GHz and 2 GB memory. The feature extraction programs were written in Java programming language, and all the document retrieval programs were tested in Matlab 7.1.

5.2. Comparative study on retrieval performance

In this section based on the above dataset we extensively compare DWH model to RAP and LDA on retrieval performance. Parameters in the simulation are set as follows. Both number of selected terms n and that of term connections m were equal to 4000. The learning rate and the momentum term to speed up the convergence in DWH model were set to 0.01 and 0.95, respectively. The DWH based on 1000 learning iterations using gradient ascent on mini-batches of 100 random training samples per iteration. All the above parameters were found delivering good performance. It was also noticed that a mild deviation from these settings would not have noticeable effect on the overall performance. In order to delete the effect of sampling randomness on results, harmonium models (i.e. DWH and RAP) were run for 20 times independently in training phase. The details of RAP and LDA can be found in Refs. [10,8]. 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 influence 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:

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

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.

First, we summarize the retrieval performance of DWH, RAP and LDA with different measures in Figs. 4–6, which leads to the best

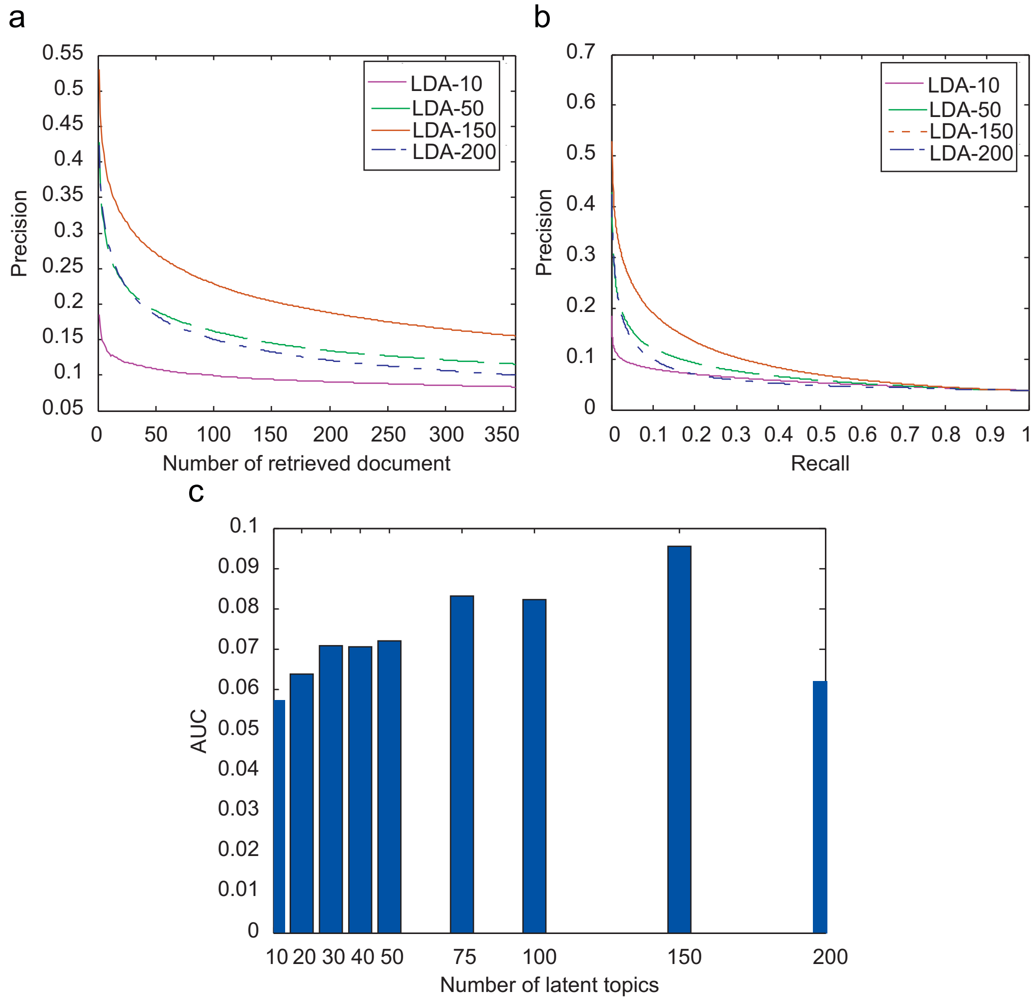

Fig. 6. Performance of LDA model based on retrieval results. (a) Precision vs number of retrieved document. (b) Precision vs recall. (c) AUC vs number of latent topics.

choice of the number of latent topics for document retrieval. Several interesting observations can be found in Figs. 4–6. For DWH model with the number of latent topics from 10 unto 40, Fig. 4(a) shows the precision results when the retrieved documents, the most similar training documents from the datasets for every query, vary from 1 to 360. It is observed that using 10 latent topics deliver the better precision results, and using 20 and 30 latent topics exhibits similar performance. It is also noted that the performance has been significantly deteriorated when the number of latent topics becomes 40. Similar results are shown in Fig. 4(b) for the sketch of the relationship between precision and recall. In order to study the effect of the number of latent topics thoroughly, we scan the number of latent topics from 5 to 50 at increments of five in Fig. 4(c) with AUC evaluation measure. It is found that using 10 latent topics performs better than other number of latent topics. The performance degrades slightly when the number of latent topics is in the range of 15–30. Number of latent topics from 35 to 50 significantly deteriorates the retrieval results compared to 10 latent topics. The performance of RAP model with different number of latent topics is summarized in Fig. 5. Similarly, RAP model with 10 latent topics delivers better results. Similar tendency of increasing the number of latent topics in RAP is shown in Fig. 5(a) and (b) compared with DWH model. In Fig. 5(c), it is observed that the performance does not deteriorate in a significant rate with the increase of the number of latent topics. Fig. 6 shows the performance of LDA model with the number of latent topics from 10 to 200. LDA with 150 latent topics exhibits better results, and its performance improves with increasing the number of latent topics from 10 to 150. The performance deteriorates significantly when continuously increasing number of latent topics from 150 to 200.

In summary, harmonium models (DWH and RAP) deliver better results with only 10 latent topics to capture the semantics in our datasets, whil LDA has to use more latent topics to represent the semantics hidden in documents. Therefore, harmonium models are computationally efficient because they are able to use fewer latent topics to represent the semantics of documents. These interesting observations show that harmonium models are more efficient than other counterpart probability models in the capability of mapping low level document features to high level semantic latent topics. We then summarized the retrieval results of different models based on their best choices of number of latent topics in Table 1 together with Fig. 7. In Fig. 7, we see that DWH model consistently delivers better performance than the other two models. Also, harmonium models perform significantly better than LDA model that uses more latent topics. In Table 1, we list the precision and recall values quantitatively with the number of retrieved documents from 1 to 360. DWH provides about 5% improvement of precision compared with RAP model with only TF features. It is believed that the improvement is resulted by including the TCF features to harmoniums. Harmonium models are able to deliver at least 35% improvement of precision

Table 1
Retrieval results of different models with different latent topics.

Method	Latent topics	No. of retrieved documents							
		Precision (%)				Recall (%)			
		1	10	40	360	1	10	40	360
DWH	10	0.8087	0.7507	0.6996	0.4654	0.0022	0.0209	0.0777	0.4654
RAP	10	0.7942	0.7121	0.6468	0.4166	0.0022	0.0198	0.0719	0.4166
LDA	150	0.5298	0.3675	0.2850	0.1551	0.0015	0.0102	0.0317	0.1551

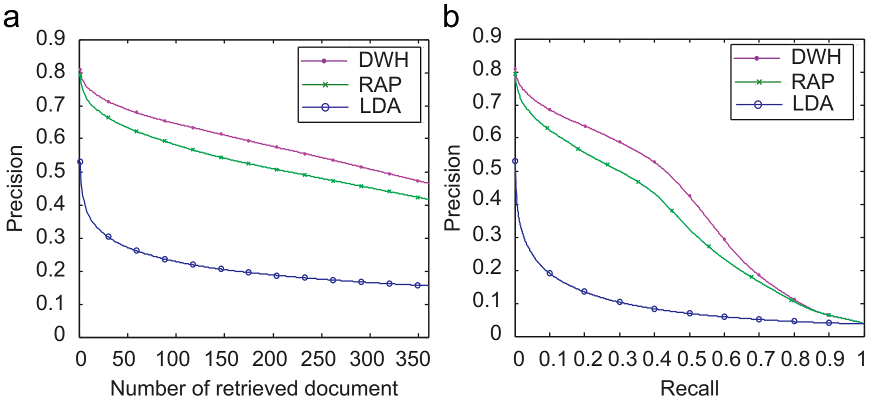

Fig. 7. Comparative retrieval results among different models. (a) Precision vs number of retrieved document. (b) Precision vs recall.

Table 2
Retrieval results of harmonium models with different features.

Method	Feature	No. of retrieved documents							
		Precision (%)				Recall (%)			
		1	10	40	360	1	10	40	360
DWH	TF + TCF	0.8087	0.7507	0.6996	0.4654	0.0022	0.0209	0.0777	0.4654
RAP	TF	0.7942	0.7121	0.6468	0.4166	0.0022	0.0198	0.0719	0.4166
RAP	TCF	0.4346	0.3123	0.2336	0.0970	0.0012	0.0087	0.0260	0.0970

compared with LDA model with 150 latent topics. Similar results are also shown in the recall results among different models. In order to show the significance of TCF features based on our dataset, we also summarized the retrieval results of harmonium models with different features in Table 2. It is observed that DWH model is able to automatically combine these multiple features with better performance.

Our comparative study and analysis indicate that the superior performance delivered by DWH model is attributed to the co-existence of the basic harmonium properties and integration of multiple features which includes more semantic information from documents. Conditional probability independence in two between-layer units and efficient inference contribute the promising performance compared to other counterpart probabilistic models like LDA.

5.3. Comparative study on different algorithms for DWH learning

This section studies the effect of different learning approaches for DWH inference based on retrieval results. Fig. 8 together with Table 3 shows the AUC values of DWH model implemented using different approximate inference methods with the number of latent topics from 5 to 50 at increments of five. From Fig. 8, Langevin and contrastive divergence learning methods perform similarly except for the case of using 10 latent topics. CD learning delivers significantly better results than Langevin and mean field sampling with 10 latent topics. On the other hand, MF performs slightly better than Langevin and CD methods with increasing the number of latent topics from 15 to 40. In Table 3, it is observed that CD learning provides

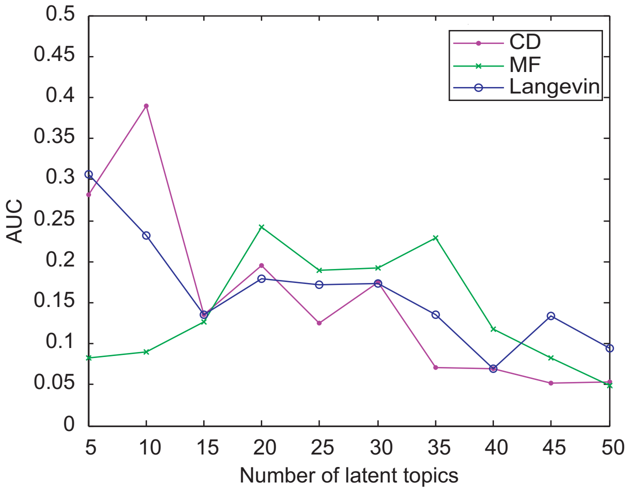

Fig. 8. AUC performance of different learning methods for DWH training.

Table 3
AUC results of different learning methods for DWH training.

	Latent topics									
Methods	5	10	15	20	25	30	35	40	45	50
CD	0.2810	0.3896	0.1349	0.1947	0.1243	0.1751	0.0702	0.0690	0.0515	0.0537
MF	0.0820	0.0893	0.1258	0.2414	0.1886	0.1921	0.2285	0.1172	0.0819	0.0492
Langevin	0.3066	0.2310	0.1356	0.1795	0.1711	0.1738	0.1355	0.0691	0.1332	0.0946

30% improvement of AUC compared with MF, and provides 15% improvement of AUC compared with Langevin with 10 latent topics. Therefore, in our study CD appears to be the best choice for the learning and inference of DWH model in terms of retrieval performance.

6. Conclusion

A new dual wing harmonium model for document data is proposed for the application to document retrieval. This DWH model integrates multiple document features into low dimensional semantic space with few latent topics for document representation. First, we formed an undirected graph to represent each document, term frequency features and term connection frequency features through vectorizing graph connectionists are extracted from documents by using weighted feature extraction method. These multiple features that consist of more semantic information hidden in documents 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 term connection semantic information with significant improvement of accuracy for document retrieval. The experimental results corroborate that the proposed approach is accurate and computationally efficient for document retrieval. Our future work will include finding more efficient methods of mining term connections from documents, and we also need to work on the inference algorithms to enhance the learning efficiency 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 Jun Yang for sending original codes of their work and thank anonymous reviewers for helpful comments to improve the standard of this paper.

7. Appendix

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 3.3 in the following way:

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

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:

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

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

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

which is exactly consistent with Eq. (14).

References

[1] T.W.S. Chow, M.K.M. Rahman, Multi-layer SOM with tree structured data for efficient document retrieval and plagiarism detection, IEEE Transactions on Neural Networks, to appear.

[2] G. Salton, M. McGill (Eds.), Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

[3] J. Zobel, A. Moffat, Exploring the similarity space, ACM SIGIR Forum vol. 32 (1) (1998) 18–34.

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

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

[6] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International SIGIR Conference, 1999.

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

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

[9] M. Welling, M. Rosen-Zvi, G. Hinton, Exponential family harmoniums with an application to information retrieval, Advances in Neural Information Processing Systems, vol. 17, MIT Press, Cambridge, MA, 2004, pp. 1481–1488.

[10] P. Gehler, A. Holub, M. Welling, The rate adapting Poisson model for information retrieval and object recognition, in: Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006.

[11] E. Xing, R. Yan, A. Hauptmann, Mining associated text and images with dual-wing harmoniums, in: Proceedings of the Conference on Uncertainty in Artificial Intelligence, 2005.

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

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

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

[15] A. Schenker, M. Last, H. Bunke, A. Kandel. Classification of web document using a graph model, in: Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'03), 2003.

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

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

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

[19] P. Smolensky, Information Processing in Dynamical Systems: Foundations of Harmony Theory, 1986, pp. 194–281.

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

[21] M. Welling, G.E. Hinton, A new learning algorithm for mean field Boltzmann machines, in: ICANN '02: Proceedings of the International Conference on Artificial Neural Networks, Springer, London, 2002, pp. 351–357.

[22] E. Xing, M. Jordan, S. Russell, A generalized mean field algorithm for variational inference in exponential families, in: Uncertainty in Artificial Intelligence (UAI2003), Morgan Kaufmann Publishers, San Francisco, CA, 2003, pp. 583–591.

[23] I. Murray, Z. Ghahramani, 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, AUAI Press, Banff, Canada, 2004, pp. 392–399.

About the Author—HAIJUN ZHANG received his B.Eng. degree in the Department of Civil Engineering and Master degree in the Department of Control Theory and Engineering from Northeastern University, Shenyang, P.R. China in 2004 and 2007, respectively. He worked as a Research Assistant at City University of Hong Kong in April–September 2007. He is currently working toward the Ph.D. degree at City University of Hong Kong, Hong Kong. His research interests are evolutionary optimization, neural network, pattern recognition and their applications.

About the Author—TOMMY W.S. CHOW (IEEE M'93–SM'03) received the B.Sc (First Hons.) and Ph.D. degrees from the University of Sunderland, Sunderland, UK. He joined the City University of Hong Kong, Hong Kong, as a Lecturer in 1988. He is currently a Professor in the Electronic Engineering Department. His research interests include machine fault diagnosis, HOS analysis, system identification, and neural network learning algorithms and applications.

About the Author—M.K.M. RAHMAN received his B.Eng. degree in the Department of Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology in 2001, and received Ph.D. degree at City University of Hong Kong in 2007. He is currently a Research Fellow in City University of Hong Kong. His research interests are structural data processing, neural network, pattern recognition and their applications.