Multidimensional Latent Semantic Analysis Using Term Spatial Information

Haijun Zhang, John K. L. Ho, Q. M. Jonathan Wu, *Senior Member, IEEE*, and Yunming Ye

***Abstract*—In this paper, we consider the problem of in-depth document analysis. In particular, we propose a novel document analysis method, named multidimensional latent semantic analysis (MDLSA), which enables us to mine local information efficiently from a document with respect to term associations and spatial distributions. MDLSA works by first partitioning each document into paragraphs and building a term affinity graph, which represents the frequency of term cooccurrence in a paragraph. We then conduct a 2-D principal component analysis to achieve an optimal semantic mapping. This analysis involves finding the leading eigenvectors of the sample covariance matrix of a training set to characterize the lower dimensional semantic space. A hybrid document similarity measure is designed to further improve the performance of this framework. Our algorithm is examined in two document applications: retrieval and classification. Experimental results demonstrate that the proposed technique outperforms current algorithms with respect to accuracy and computational efficiency.**

***Index Terms*—Dimensionality reduction, multidimensional, principle component analysis (PCA), semantic analysis, term association.**

I. Introduction

WE are investigating the potential of an in-depth document analysis by using term spatial information and dimensionality reduction techniques. The evolution of human languages has been expedited by the use of the Internet. We see a growing demand for semantic representation that includes the term associations and spatial distributions. Another demand is to find low-dimensional semantic expressions of documents, while preserving the essential statistical relationships between terms and documents. Some usages of low-dimensional representation are extremely useful for facilitating the processing of large document corpora and the handling of various data mining tasks, such as classification, retrieval, plagiarism, etc. However, the main challenge for document analysis is knowing how to locate the low-dimensional space with the fusion of local information, which conveys term associations and spatial distributions, in a unified framework.

Here, we introduce a new model for in-depth document analysis, named multidimensional latent semantic analysis (MDLSA). It starts by partitioning each document into paragraphs and establishing a term affinity matrix. Each component in the matrix reflects the statistics of term cooccurrence in a paragraph. It is worth noting that the document segmentation can be implemented in a finer manner, for example, partitioning into sentences. Thus, it allows us to perform an in-depth analysis in a more flexible way. We then conduct a 2-D principal component analysis (2DPCA) [25] with respect to the term affinity matrix. This analysis relies on finding the leading eigenvectors of the sample covariance matrix to characterize a lower dimensional semantic space. According to our empirical study, we find that using only a 1-D projection to represent each document is sufficient to achieve marked results. Moreover, a hybrid document similarity measure is designed to further improve the performance of this framework. In comparison with the traditional "Bag of Words" (BoW) models such as the latent semantic indexing (LSI) and the principal component analysis (PCA), MDLSA aims to mine the in-depth document semantics, which enables us to not only capture the global semantics at the whole document level, but also to deliver the semantic information from local data-view regarding the term associations at the paragraph level. We conduct extensive experimental verifications including document retrieval and classification. The results corroborate that the proposed technique is accurate and computationally efficient for performing various document applications.

The remaining sections of this paper are organized as follows. A brief overview of related works is given in Section II. A brief description of feature extraction procedures from global data-view is presented in Section III. We build a word affinity graph in Section IV. We illustrate the details of the MDLSA algorithm in Section V. A hybrid similarity measure is designed in Section VI. We evaluate the performance of MDLSA on retrieval and classification in Section VII. We provide discussions based on the experimental results in Section VIII. We conclude this paper in Section IX.

Manuscript received September 25, 2011; revised February 12, 2012 and July 4, 2012; accepted October 11, 2012. Date of current version November 18, 2013. This work was supported in part by the National Natural Science Foundation of China under Grant 61073195, the National Key Technology R&D Program of MOST China under Grant 2012BAK17B08, the Shenzhen Strategic Emerging Industries Program under Grant ZDSY20120613125016389, the Shenzhen Foundation Research Fund under Grant no. JCY20120613115205826, the Canada Chair Research Program, the Natural Sciences and Engineering Research Council of Canada, and the City University of Hong Kong under Grant 9220053. This paper was recommended by Associate Editor Z. Zdrahal of the former IEEE Transactions on Systems, Man and Cybernetics, Part C.

H. Zhang and Y. Ye are with the Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China, and also with the Shenzhen Key Laboratory of Internet Information Collaboration, Shenzhen 518055, China (e-mail: aarhzhang@gmail.com; yym@hitsz.edu.cn).

J. K. L. Ho is with the Department of Mechanical and Biomedical Engineering, City University of Hong Kong, Kowloon, Hong Kong (e-mail: mejohnho@cityu.edu.hk).

Q. M. J. Wu is with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON N9B 3P4, Canada (e-mail: jwu@uwindsor.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCC.2012.2227112

II. Related Work

This section briefly reviews the previous work. It involves the BoW models, which consider only the term frequency (*tf*) information, and the models that take the term associations into account.

A. BoW Model

The last two decades have witnessed the rapid development of the BoW model representing a document from a lengthy vector to a low-dimensional semantic expression. The earliest work on document modeling is the vector space model (VSM) [1], which usually uses the *tf-idf* scheme for term weighting. A vocabulary of terms (or words) is first constructed for feature description. The term frequency (*tf*) is the number of occurrences of each term. The inverse document frequency (*idf*) is a function of the number of documents, where a term appears. A weighted term vector is then formulated to represent each document. Similarity between two documents is measured by using the *cosine* distance or other advanced distance functions. The beauty of the VSM is the capability of reducing the arbitrary length of each document to a fixed length by a term vector. Nevertheless, a lengthy vector is required to describe the frequency information of terms because the number of words involved is usually a huge amount. Not to mention, the VSM reveals little statistical property of a document due to using only low-level document features (e.g., *tf*). To overcome these shortcomings, researchers have proposed several dimensionality reduction methods by using low-dimensional latent representations to capture document semantics. For instance, the LSI [2] maps documents associated with terms onto a latent space, by performing a linear projection: singular value decomposition, which is capable of compressing the lengthy feature vector into a lower dimensional domain, while preserving the essential statistics. The PCA, an alternative to the LSI, is a traditional linear technique that is able to project the high-dimensional term vectors to a lower dimensional space, by finding the solution of an eigenvalue problem. This problem usually involves the calculation of the sample covariance matrix of a training set. Moreover, there has been growing interest in developing low-dimensional representations through subspace learning techniques, which have been successfully used in the fields of image processing, computer vision, and face recognition. An excellent overview of these techniques can be referred to [3]. It is worth pointing out that these locality preserving methods can also be applied to document representations. Such an attempt has been experimented with by Cai *et al.* [4]. Besides these dimensionality reduction techniques, a step forward to statistical models is the probabilistic latent semantic indexing (PLSI) [5], which defines a proper generative model to sample each word from a mixture distribution and develop factor representations for the mixture components. A brief overview of related statistical models, such as the latent Dirichlet allocation [6], the generalized Dirichlet multinomial distributions [7], the exponential family harmonium [8], and the rate adapting Poisson (RAP) model [9], can be referred to [10]. Indeed, most reported techniques aforementioned are largely based on typical *tf* information with respect to the BoW model. They all use a flat feature representation by formulating a function of *tf*. This representation scheme is only a rough description of a document. As a result, some useful semantic information will have been overlooked because two documents containing similar term frequencies may be contextually different, when the spatial distribution of terms is different. For example, *school*, *computer*, and *science* mean very different terms when they appear in different parts of a document in comparison to the case of *school of computer science* that appear together. Thus, solely relying on the *tf* information from the BoW model is not a promising way to discriminate contextual similarity, because we must consider both the *tf* and the word interconnections and spatial distributions throughout the document.

B. In-Depth Document Analysis

Recently, the problem of in-depth document analysis has been investigated by many researchers. In the Web documents that are using graph matching, different directed graphs with a few most frequent terms attributed as nodes are defined to represent each document [11]. Although it is quite successful to enhance the classification accuracy, the graph matching process involved in the proposed approach must be accomplished in polynomial time. For large datasets, it may need the approximate techniques, for example, compressive sampling [45], to tackle the computational constraints. Fuketa *et al.* [12] introduced a field understanding method using field association words for document classification. Others used either bigrams [13] or term association rules [14] to enhance the classification accuracy. To reflect the subtopic structure of a document, Kim and Kim [15] introduced a passage-based text categorization model. It segments a test document into several passages, assigns categories to each passage, and merges passage categories into document categories. Meanwhile, Xu and Zhou [16] proposed a model, regarding both the compactness of the appearances of a word, and the position of the first appearance of the word by distributional features. However, finding an optimal combination for different information sources extracted by those two methods [15], [16] is still a practical issue. Another interesting study of considering the term associations is the spectral-based approach reported by Park *et al.* [17]–[19]. They took the patterns of query term occurrence into account, while suggesting that documents containing the query terms, which follow a similar positional pattern are supposed to be more relevant. The approach does yield impressive results to enhance the text retrieval performance. However, it is only applicable to the case of a few keywords as a query. Making the term spectrum contribute in a more general document application, which relies on between-document similarity, remains unclear.

A flexible multilayer self-organizing map (MLSOM) [20] is designed to process generic tree-structured data, such as document data, which can be hierarchically represented as document-pages-paragraphs. It is suggested that MLSOM is computationally efficient to the handling of a large dataset for retrieval and plagiarism detection, while capturing the semantics from the spatial distributions of terms. However, MLSOM may be parameter sensitive for a specific dataset, as a result of the setting size of SOM, the training steps, and the selection of training samples. The updating issue is yet another intractable task, even if the MLSOM could be implemented online. Another work has been focused on the multiple features (MF) extraction schemes by using different word graphs [21]. Term connection frequency (TCF) is extracted from each document by employing different feature extraction methods. In a later study, two dual-wing harmonium models have been developed to generate the latent representations of documents by jointly modeling MF [10], [22]. In the latest work [23], a multilevel matching (MLM) strategy was designed for retrieval and plagiarism detection. MLM employs the Earth Mover's Distance (EMD) [24] to fuse the semantics from paragraphs. Despite promising performance on retrieval and plagiarism detection, the major drawback of MLM lies in the computational burden of calculating the EMD. The time cost increases exponentially, as the number of paragraphs or sections increases as well. As a result, calculating the EMD based on sentences becomes impossible. Thus, the document segmentation stops appropriately at only paragraph level.

On the other hand, many existing papers deal with computational models for lexical cooccurrence [35], [36], context vectors [37], and term distribution in the context of coherence metrics [38], [39]. One of the most interesting works is explicit semantic analysis (ESA) [40], which attempts to classify a given document with respect to a set of explicitly given external categories. In this sense, the ESA is explicit in comparison with the LSI [2], which aims to represent documents with latent topics. It is noted that the ESA relies on the external categories to evaluate the semantic relatedness, while our model, i.e., MDLSA, extracts term associations within a document collection. We may combine the semantics from the ESA and the LSI for a more effective document representation.

C. Features of Our Model

The major features of our model include two parts: 1) word affinity graph; and 2) MDLSA. The first part relies on the term association matrix. The MDLSA works by finding an optimal mapping from the original term association space, which is large, to a low-dimensional semantic space.

With respect to previous works, we clarify that our approach is most related to the 2DPCA [25], the LSI [2], and the MLM method [23], but the document modeling proposed here is considerably different. Compared with the 2DPCA [25], MDLSA can be seen as an extension of the 2DPCA, which has been successfully applied in face recognition [25]. To our knowledge, this is the first report to use the power of 2DPCA for documents. In addition, MDLSA projects a word affinity graph onto a reduced semantic space from two directions, while 2DPCA only projects an image matrix from a single direction. Compared with the LSI [2] and the MLM method [23], MDLSA considers the joint modeling of term frequencies and term associations in a principled manner, and it provides us a more accurate representation of document. The main difference between MLM and MDLSA lies in the representation of the term spatial distribution. The MLM method relies on the many-to-many matching of paragraphs by solving the linear programming, while MDLSA considers the term associations from a dimensionality reduction viewpoint. It is worth pointing out that MLM requires a large amount of computational time in practice.

III. Extracting Global Features

In this section, we introduce the common procedures of document feature extraction, such as preprocessing, vocabulary construction, forming a weighted term vector, which is regarded as a global representation of a document, and dimensionality reduction.

A. Vocabulary Construction

First, we introduce the common document feature extraction procedures. The preprocessing works by first separating the main text contents from documents, for example, HTML-formatted documents. We then extract words from all the documents in a dataset and apply stemming to each word. Stems are often used as basic features instead of original words. Thus, "program," "programs," and "programming" are all considered as the same word. We remove the stop words (a set of common words like "a," "the," "are," etc.) and store the stemmed words together with the information of the *tf*, f_u^t (the frequency of the uth word in all documents), and the document frequency, f_u^d (the number of documents the u-th word appears). Forming a histogram vector for each document requires the construction of a word vocabulary each histogram vector can refer to. Based on the stored *tf* and document frequency, we use the well-known *tf-idf* term-weighting measure to calculate the weight of each word

$$w_u = f_u^t \cdot idf \tag{1}$$

where idf denotes the inverse-document-frequency that is given by $idf = \log_2(n/f_u^d)$, and n is the total number of documents in a dataset. It is noted that this term-weighting measure can be replaced by other feature selection criteria [26]. The words are then sorted in descending order according to their weights. The first m words are selected to construct the vocabulary M. According to the empirical study [21], [23], using all the words in the dataset to construct the vocabulary is not necessarily expected to deliver the improvement of performance because some words may be noisy features for some topics. We have conducted detailed experiments to evaluate the performance in terms of different options of the vocabulary size, i.e., the value of m (see Section VII).

B. Term Weighting

After the vocabulary construction, each document can be represented by a column vector $x_i = [w_1, w_2, \ldots, w_m]^T$, which is associated with the terms in the vocabulary. To reduce the impact of certain documents and term properties from affecting the semantic analysis, term-weighting schemes are usually adopted [19], [27]. The weighting schemes investigated in this study were

$$\text{NORM: } w_u = \left(\frac{f_{u,i}}{W_i}\right) \log\left(n/f_u^d\right) \tag{2}$$

BD − ACI − BCA:

$$w_u = \left(\frac{1+\log(f_{u,i})}{(1-s)+sW_i/\bar{W}_i}\right)\log\left(1+f_u^m/f_u^d\right) \tag{3}$$

AB − AFD − BAA (Okapi):

$$w_u = \left(\frac{f_{u,i}}{f_{u,i}+\tau_i/\bar{\tau}_i}\right)\log\left(1+n/f_u^d\right) \tag{4}$$

BI − ACI − BCA:

$$w_u = \left(\frac{1+\log(f_{u,i})}{(1-s)+sW_i/\bar{W}_i}\right)\left(1-\frac{n_u}{\log_2(n)}\right) \tag{5}$$

Lnu.ltu (SMART):

$$w_u = \left(\frac{(1+\log(f_{u,i}))/(1+\log(\bar{f}_{u,i})}{(1-s)+s\tau_i/\bar{\tau}_i}\right)\log\left(n/f_u^d\right) \tag{6}$$

where $f_{u,i}$ is the term frequency of the uth word associated with the ith document, f_u^d is the document frequency of term u, f_u^m is the largest f_u^d for all u, W_i is the document vector l_2 norm, i.e., $W_i = \|x_i\|_2$, $\bar{W}_i$ is the average W_i in the entire dataset, τ_i and $\bar{\tau}_i$ are the number of unique terms in document i and the average unique terms, respectively, s is a slope parameter (set to 0.7 [19], [28]), and n_u is a noise measure of term u [27], [28]. The NORM weighting was recently used in [20], [21], and [23]; the other four schemes, which are well-known weighting methods, were used in [19] and [28].

C. Dimensionality Reduction

A document set can be represented by $X = [x_1, x_2, \ldots, x_n] \in R^{m\times n}$, which is a rectangular matrix of terms and documents. The desire of latent semantic analysis is to produce a set Y, which is an accurate representation of X, but resides in a lower dimensional space. Y is of dimension d, with $d \ll m$, and it is produced by the form

$$Y = V_g^T X \tag{7}$$

where V_g is an $m \times d$ linear transformation matrix. Thus, it is straightforward to replace each document x_i by its projection $y_i = V_g^T x_i$ such that we can make between or within comparisons facile in the lower dimensional latent semantic space. There are a number of ways to accomplish this projection. The transformation matrix V_g can be obtained by traditional techniques such as the PCA, the LSI, or other dimensionality reduction approaches [3]. In this study, we use the classical PCA to determine the matrix V_g. The PCA is a well-known technique in the category of dimensionality reduction. In the PCA, the determination of V_g is given by maximizing the variance of the projected vectors, which is in the format of

$$\max_{V_g} \sum_{i=1}^{n} \|y_i - \frac{1}{n}\sum_{i=1}^{n} y_i\|_2^2. \tag{8}$$

It has been shown that the matrix V_g is the set of eigenvectors of the sample covariance matrix associated with the d largest eigenvalues. Keep this in mind, as we will use this set of global representations $\{y_1, y_2, \ldots, y_n\}$ to formulate a hybrid similarity of two documents (see Section VI).

IV. Word Affinity Graph

This section introduces a scheme to produce an in-depth document representation. First, we segment each document into paragraphs. Second, we build a word affinity graph, which describes the local information of each document.

A. Document Segmentation

As we mentioned before, the major drawback of the traditional modeling methods such as the PCA and the LSI is that they lack the description of term associations and spatial distribution information over the reduced space. In this study, we propose a new document representation that contains this description. First, each document is segmented into paragraphs. Since we only considered the HTML documents in this paper, a Java platform was developed to implement the segmentation. For the HTML format document, we can use the HTML tags to identify paragraphs easily. Before document segmentation, we first filter out the formatted text that appears within the HTML tags. The text is not accounted for in word counts or document features. The overall document partitioning process can be summarized as follows [20], [23].

1) Partition a document into blocks using the HTML tags: "<p>," "<br\>," "<li>," "</td>," etc.
2) Merge the subsequent blocks to form a new paragraph until the total number of words of the merged blocks exceeds a paragraph threshold (set at 50).
3) The new block is merged with the previous paragraph if the total number of words in a paragraph exceeds the minimum threshold (set at 30).

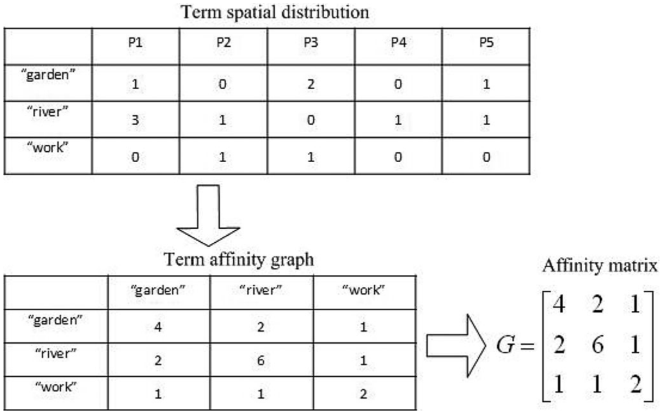
For the HTML documents, it is noted that there is no rule for minimum/maximum number of words for paragraphs [20]. Setting a threshold for word counts, however, still enables us to control the number of paragraphs flexibly in each document and remove the blocks, which contain only a few words (e.g., titles), by being attached to the real paragraph blocks. It is worth pointing out that we are able to further partition each paragraph into sentences by marking periods (the tag "\.") to form a finer structure such that more semantics can be included.

B. Word Affinity Graph

Building a word affinity graph for each document is to represent the frequency of term cooccurrence in a paragraph. Consider a graph denoted by a matrix $G_i \in R^{m\times m}$, in which each element $g_{i,u,v}$ $(u, v = 1, 2, \ldots, m)$ is defined by

$$g_{i,u,v} = \begin{cases} F_{u,v} \cdot \log_2(n/DF_{u,v})/\|G_i\|_2, u \neq v \\ f_u^t \cdot \log_2(n/f_u^d)/\|G_i\|_2, u = v \end{cases} \tag{9}$$

where $\|.\|_2$ is the Frobenius norm, $F_{u,v}$ is the frequency of the cooccurrence in a paragraph associated with the terms u and v in the ith document, $DF_{u,v}$ is the document frequency that the terms u and v coappear in a document, and notations of f_u^t and f_u^d are as described in (1). Note that if we do not consider term



Term spatial distribution

	P1	P2	P3	P4	P5
"garden"	1	0	2	0	1
"river"	3	1	0	1	1
"work"	0	1	1	0	0

Term affinity graph

	"garden"	"river"	"work"
"garden"	4	2	1
"river"	2	6	1
"work"	1	1	2

Affinity matrix

$$G = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

Fig. 1. Example of establishing a word affinity graph. The top table shows the term spatial information distributed over five paragraphs. Here, we assume that three words, i.e., "garden," "river," and "work," are selected, and the document is partitioned into five paragraphs. The second table shows the term affinity graph transmitted from the first table. The diagonal elements represent the term frequency in this document, and the off-diagonal elements represent the term cooccurrence. Note that here we do not consider any weighting scheme with respect to document frequency [as shown in (9)].

cooccurrence in paragraphs, i.e., let $g_{i,u,v} = 0$ (for $u \neq v$), the affinity graph G_i becomes a diagonal matrix with the elements corresponding to the global feature vector x_i shown in (2) (the NORM weighting). By definition, the graph G_i is a symmetric matrix. This graph contains the local semantic information of a document in a way that we can design an efficient semantic representation including term interconnections and distributions in a unified framework. For clarity, Fig. 1 gives us an example of establishing a word affinity graph.

V. Multidimensional Latent Semantic Analysis

This section presents a new model, MDLSA, which considers word affinity graphs and maps them onto a low-dimensional latent semantic space. First, we introduce the objective of the MDLSA model. Second, we learn a semantic subspace by using the 2DPCA algorithm. Third, we further process and select the semantic projections. We summarize the MDLSA algorithm in the end.

A. Semantic Projection

Despite the capability of delivering more semantics, a word affinity graph is usually of large size and sparseness. It is computationally demanding if we simply rely on these graphs to make between or within comparisons. Besides, assembling the similarity between two matrices is another demanding issue. On the other hand, without further processing, these graph representations contain a large quantity of noises, which spread out the original term distributional space. As a result, these noises cause degradation of performance. Therefore, it is important to design an efficient dimensionality reduction technique, which is able to compress the graph in a principled manner and form an accurate representation in a lower dimensional space. The proposed MDLSA model is just this. Given a word affinity graph G of size $m \times m$ (see Section IV-B), the goal of MDLSA is to produce a projection $\tilde{Z}$ of size $d \times d$ ($d \ll m$) resided in a lower dimensional semantic space. We then use a matrix Z of size $d \times k$ ($k \leq d$), which is constructed by a smaller number of columns of $\tilde{Z}$. In linear algebra, the projection $\tilde{Z}$ can be obtained by

$$\tilde{Z} = V^T G V \tag{10}$$

where V is an $m \times d$ linear transformation matrix, as mentioned in (7). The problem comes to finding an optimal transformation V for this dimensionality reduction.

B. Learning a Semantic Subspace

To acquire the optimal transformation matrix V, we use the 2DPCA method [25], which has been successfully implemented in a face recognition system.

For completeness, the process of calculating the matrix V is summarized here, and the details can be found in the article reported by Yang *et al.* [25]. Let $\{G_1, G_2, \ldots, G_n\}$ be a set of training documents. By representing the word affinity graph G_i associated with the ith document, the graph covariance (or scatter) matrix C can be written by

$$C = \frac{1}{n} \sum_{i=1}^{n} (G_i - \bar{G})^T (G_i - \bar{G}) \tag{11}$$

where $\bar{G}$ denotes the average graph of all the training samples. Similar to PCA, 2DPCA introduces this total scatter of the projected samples to measure the discriminatory power of a transformation matrix V. In fact, the total scatter of the samples in a training set can be characterized by maximizing the criterion [25]

$$J(v) = v^T C v \tag{12}$$

where v is a unitary column vector, which is called the optimal mapping axis by maximizing the above quantity. In general, it is not sufficient to have only one optimal mapping axis. It is required to find a set of mapping axis, $v_1, v_2, \ldots, v_d$, subject to the orthogonal constraints and maximizing the criterion $J(V)$ by the form [25]

$$\begin{gathered} \{v_1, v_2, \ldots, v_d\} = \arg\max_{v} J(v) \\ \text{subject to } v_j^T v_l = 0 (j \neq l, j, l = 1, 2, \ldots, d). \end{gathered} \tag{13}$$

According to linear algebra, the optimal mapping axes, $v_1, v_2, \ldots, v_d$, are the orthogonal eigenvectors of C associated with the first largest d eigenvalues. If we denote these mapping axes by $V = [v_1, v_2, \ldots, v_d]$, the projection $\tilde{Z}$ of a word affinity graph G will be acquired easily by the product of the resulting matrices, as shown in (10). Here, note that we take advantage of the symmetry of the affinity graph G. If the graph G is asymmetric, the transformation shown in (10) will be the same as the bidirectional PCA [29].

C. Selection of the Semantic Projections

Actually, we can use another matrix Z of size $d \times k$ ($k \leq d$), which is a submatrix of $\tilde{Z}$, to represent the original graph G for optimal approximate fit by discovering lower dimensional

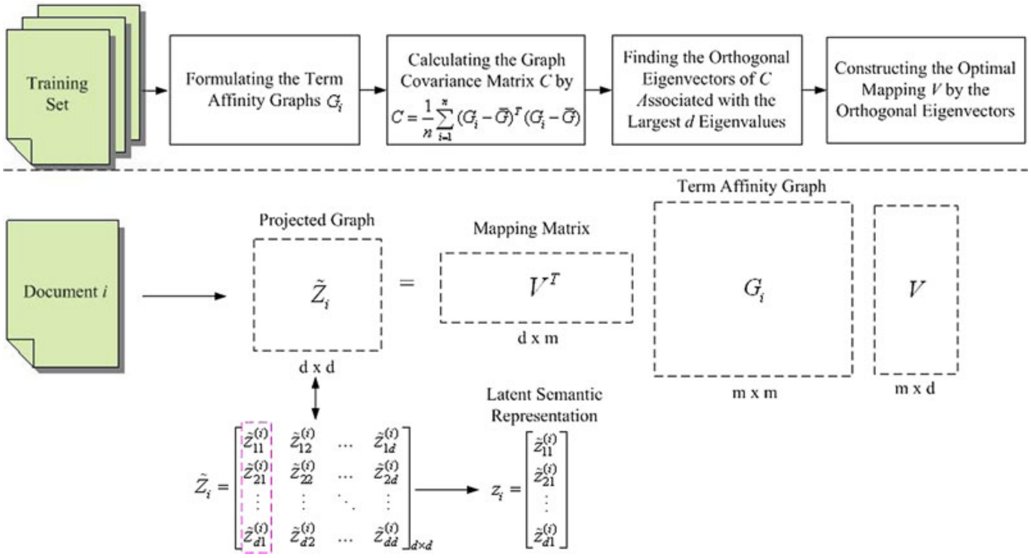

Training Set
Formulating the Term Affinity Graphs G_i
Calculating the Graph Covariance Matrix C by $C=\frac{1}{n}\sum_{i=1}^{n}(G_i-\bar{G})^T(G_i-\bar{G})$
Finding the Orthogonal Eigenvectors of C Associated with the Largest d Eigenvalues
Constructing the Optimal Mapping V by the Orthogonal Eigenvectors
Document i
Projected Graph
$\tilde{Z}_i$
=
Mapping Matrix
V^T
Term Affinity Graph
G_i
V
d x d
d x m
m x m
m x d
Latent Semantic Representation
$\tilde{Z}_i = \begin{bmatrix} \tilde{z}_{11}^{(i)} & \tilde{z}_{12}^{(i)} & \cdots & \tilde{z}_{1d}^{(i)} \\ \tilde{z}_{21}^{(i)} & \tilde{z}_{22}^{(i)} & \cdots & \tilde{z}_{2d}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{z}_{d1}^{(i)} & \tilde{z}_{d2}^{(i)} & \cdots & \tilde{z}_{dd}^{(i)} \end{bmatrix}_{d\times d}$
$z_i = \begin{bmatrix} \tilde{z}_{11}^{(i)} \\ \tilde{z}_{21}^{(i)} \\ \vdots \\ \tilde{z}_{d1}^{(i)} \end{bmatrix}$

Fig. 2. Description of the MDLSA Algorithm.

space. In practice, only using the first column of $\tilde{Z}$ is sufficient to achieve remarkable results. Thus, the matrix Z is of size $d \times k$ (here, $k = 1$) and turns out to be a column vector like y_i produced by the traditional PCA corresponding to the global feature x_i. We also conducted an empirical study on the selections of value of k (see Section VII). To avoid confusion, in the following context, let z_i be the first column of $\tilde{Z}_i$, which denotes the projection matrix of the ith affinity graph G_i. Alternatively, the local information from the ith training document can be represented by the column vector z_i. This is a very promising property of MDLSA by delivering three important advantages. First, in comparison with 2DPCA [25], it does not need an assembled metric to conduct direct matrix comparison such that MDLSA is easier to make between comparisons. Second, much less time is required to compare two documents because MDLSA does not need the many-to-many matching compared with the MLM method [23]. Third, MDLSA contains local semantic information of documents compared withthe PCA and the LSI [2].

D. Algorithm Details

The overall procedure of the MDLSA algorithm is summarized as follows.

Input: The training set, the vocabulary M, and the dimension of the reduced space d.

Output: Latent semantic representations $\{z_i\}$ for training samples and z_t for a new test sample.

1) Input the training set, the vocabulary M, and the dimension of the reduced space d.
2) Partition each document into paragraphs and form the affinity graphs $\{G_1, G_2, ..., G_n\}$.
3) Solve the eigenvalue problem as shown in (13), and construct the mapping V whose column vectors are taken from the eigenvectors associated with the d largest eigenvalues.
4) Calculate the projected graphs $\tilde{Z}_i = V^T G_i V$, as shown in (10).
5) Select the first column of $\tilde{Z}_i$ to represent the ith training sample denoted as z_i.
6) Given a new affinity graph G_t associated with a new testing document, repeat Steps 4 and 5, map it onto the subspace, and achieve the latent semantic expression z_t.

For clarity, we visually describe the algorithm in Fig. 2.

VI. Hybrid Similarity Measure

Many document applications rely on the calculation of similarity between two documents. In order to further improve the performance of our framework, we develop a hybrid similarity measure to synthesize the information from a global data-view and local data-view.

In this study, we have extracted two sets of features from each document: a feature vector x_i containing global information (i.e., *tf*) and an affinity graph G_i delivering local information (i.e., term associations). We then use dimensionality reduction techniques to map these features onto the latent semantic space, which is of lower dimension. Intuitively, combining these two information sources may bring performance gain. Therefore, we design a hybrid similarity associated with both the global and local information. Given two documents p and q, let y_p be the latent representation of document p associated with the global feature x_p, and z_p the latent representation of document p produced from the local source G_p. Likewise, let y_q be the latent representation of document q associated with the global feature x_q, and z_q the latent representation of document q produced from the local source G_q. We work by a combined similarity measure in the form, which involves the *cosine* distance criterion

$$S(p,q) = \mu S_g(p,q) + (1-\mu) S_l(p,q)$$
$$S_g(p,q) = \frac{y_p \cdot y_q}{\|y_p\|_2 \|y_q\|_2}, S_l(p,q) = \frac{z_p \cdot z_q}{\|z_p\|_2 \|z_q\|_2} \tag{14}$$

where $S_g(p,q)$ represents the global similarity, $S_l(p,q)$ denotes the local similarity, and $\mu(0 \leq \mu \leq 1)$ is a weight parameter used to balance the importance of the global and local similarity.

TABLE I
NAMES OF EXPERIMENTAL METHODS USING PREWEIGHTING

Value of *	Label	Description
{MDLSA-Hybrid, PCA, LSI, VSM}	*-NORM	NORM pre-weighting
	*-BD-ACI-BCA	BD-ACI-BCA pre-weighting
	*-AB-AFD-BAA	AB-AFD-BAA pre-weighting
	*-BI-ACI-BCA	BI-ACI-BCA pre-weighting
	*-SMART	Lnu.ltu (SMART) pre-weighting

Method names are of the form *-{NORM, BD-ACI-BCA, AB-AFD-BAA, BI-ACI-BCA, SMART}.

TABLE II
DESCRIPTIONS OF PARAMETERS INVOLVED

Notation	Description
k	The number of columns selected from the projected matrix $\tilde{Z}$
μ	The weight used to balance the importance of the global and local similarity
m	The vocabulary size
d	The dimension of projected features

Thus, the system provides users flexibility to select the value of μ to balance this hybrid measure according to their expectations. In this study, we also include the effect study of the parameter μ in experiments (see Section VII). Note that the local similarity $S_l(p, q)$ is associated with the features produced by only the MDLSA method, while the global similarity $S_g(p, q)$ relies on the features obtained by the PCA.

VII. EXPERIMENTS

In this section, we evaluate the performance of MDLSA on two document applications: retrieval and classification. We use two implementations: MDLSA, which only measures the local similarity, and MDLSA-Hybrid, which is based on both the global and local similarity (see Section VI). These two algorithms are compared with MLM-Hybrid [23], MLM-Local [23], MF [21], TCF [21], PCA, LSI [2], VSM [1], RAP [9], PLSI [5], and direct graph matching (DGM). MLM-Local only uses the similarity associated with paragraph-level matching, while MLM-Hybrid relies on the similarity, which is produced by both document-level features and paragraph-level features. The contributions of these two features to the similarity measure are balanced by a weight parameter, as shown in (14). The details of the MLM methods can be found in [23]. The TCF method works by only using the feature represented by term connection. The MF approach is based on both TF features and TCF features that are weighted by a parameter similar to the case of MDLSA-Hybrid and MLM-Hybrid. See [21] for the details of MF and TCF. The PCA and the LSI perform on only *tf* features. RAP and PLSI, which are statistical methods, use only tf features without any term-weighting schemes. DGM, which is similar to the method described in [11], was tested on only the YahooScience set due to its heavy computational burden. But the results have clearly demonstrated that the MDLSA outperforms the DGM by a significant amount. The VSM regarded as a baseline method is investigated by without any data reduction operations. The details of the VSM and the LSI can be found in [1] and [2], respectively. As we investigated many weighting schemes, as shown in (2)–(6), the methods relying on these preweights were listed in Table I. For clarity, we also listed the notations of the parameters involved in this study, as shown in Table II. All the experiments were performed on a PC with Intel(R) Core(TM) i7 CPU 860@ 2.80 GHz and 6.00-GB memory. The feature extraction programs were written in Java programming language. The document retrieval and classification programs were tested on MATLAB 7.5.0 (R2007b).

TABLE III
DETAILS OF THE DATASETS

	CityU1	YahooScience	WebKB4
Class	26	6	4
Number of Documents	10400	861	4171
Maximal Number of words in Each Document	363068	36318	57267
Average Number of Words in Each Document	1849	913	290
Number of Maximal Paragraphs	2368	427	529
Average Number of Paragraphs	20.34	10.96	4.17

A. Document Retrieval

In this section, we conducted a large scale of experiments to show the retrieval performance of our proposed approach. Intuitively, MDLSA-related methods are more effective on large size of documents because the spatial distributions of terms will become conspicuous in a lengthy document. To provide a more real-life experiment, Chow and Rahman [20] collected a dataset, CityU1, with 26 categories consisting of documents with the size ranging from few hundred words to over 200 thousand words. This dataset is selected because it features in including many lengthy documents. Each category includes 400 documents making a total number of 10 400 documents. 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.[1] This dataset has been used in [10] and [20]–[23]. The distribution details of this dataset were summarized in Table III. The dataset was divided equally among ten folds. We held out 90% of the data corpora as a candidate set and 10% as a test set that is used for query. We performed tenfold cross validation, and the results were averaged over each query, then over the ten folds. The query in this study is the whole document. The relevant documents are the ones that belong to the corresponding category. First, we introduce the performance metrics used in this study. Second, we present the comparative results with respect to retrieval performance and query time performance. Third, we include the study on the influence of different parameters involvedin the algorithms.

1) Performance Metrics: To quantify the retrieval results, we used averaged precision and recall values [9], [10] for each query document. The precision and recall measures are defined as follows:

$$Precision = \frac{\text{No. of correctly retrieved documents}}{\text{No. of retrieved documents}} \tag{15}$$

[1] www.ee.cityu.edu.hk/∼twschow/Html_CityU1.rar

TABLE IV
COMPARATIVE RESULTS OF DIFFERENT RETRIEVAL METHODS

Method	AUC(%)	No. of Retrieved Documents					
		10	40	360	10	40	360
		Average Precision (%)			Average Recall (%)		
MDLSA-Hybrid-NORM	73.97	87.64	85.88	75.32	2.43	9.54	75.32
MLM-Hybrid	72.64	87.99	85.65	73.75	2.44	9.52	73.75
MDLSA-Hybrid-BI-ACI-BCA	70.88	87.74	85.68	72.17	2.44	9.52	72.17
MDLSA-Hybrid-AB-AFD-BAA	70.83	87.62	85.56	72.44	2.43	9.51	72.24
MDLSA-Hybrid-SMART	70.78	87.85	85.67	72.16	2.44	9.52	72.16
MF	70.54	85.00	82.61	73.24	2.36	9.18	73.24
MLM-Local	70.32	88.52	85.90	71.08	2.46	9.54	71.08
MDLSA-Hybrid-BD-ACI-BCA	70.29	87.56	85.33	71.71	2.43	9.48	71.71
MDLSA	69.84	87.52	85.49	71.21	2.43	9.50	71.21
PCA-NORM	68.26	82.78	80.13	71.85	2.30	8.90	71.85
PCA-BI-ACI-BCA	66.78	86.20	83.26	69.00	2.39	9.52	69.00
PCA-SMART	66.11	86.09	82.83	68.59	2.39	9.20	68.59
PCA-AB-AFD-BAA	65.72	84.95	82.02	68.45	2.36	9.11	68.45
PCA-BD-ACI-BCA	64.00	83.87	81.10	67.09	2.33	9.01	67.09
VSM-NORM	63.20	77.01	77.78	66.83	2.14	8.64	66.83
PLSI	62.20	73.34	76.08	67.56	2.04	8.45	67.56
VSM-BI-ACI-BCA	58.49	80.68	78.29	61.72	2.24	8.70	61.72
VSM-SMART	57.27	79.32	77.18	60.89	2.20	8.58	60.89
VSM-BD-ACI-BCA	57.06	78.95	76.85	60.78	2.19	8.54	60.78
LSI-BD-ACI-BCA	52.23	80.15	76.68	58.07	2.23	8.52	58.07
LSI-BI-ACI-BCA	46.94	80.22	76.28	52.82	2.23	8.48	52.82
LSI-AB-AFD-BAA	46.73	79.35	74.77	53.10	2.20	8.31	53.10
TCF	46.58	73.09	69.02	54.24	2.03	7.67	54.24
VSM-AB-AFD-BAA	46.05	75.97	72.14	51.06	2.11	8.02	51.06
LSI-SMART	44.15	79.06	74.20	50.65	2.20	8.24	50.65
LSI-NORM	43.21	76.31	72.40	51.51	2.12	8.04	51.51
RAP	35.59	78.84	73.04	45.02	2.19	8.12	45.02

The weight μ settings for hybrid methods were MDLSA-Hybrid-NORM: $\mu = 0.25$; MLM-Hybrid: $\mu = 0.4$; MDLSA-Hybrid-BI-ACI-BCA: $\mu = 0.2$; MDLSA-Hybrid-AB-AFD-BAA: $\mu = 0.2$; MDLSA-Hybrid-SMART: $\mu = 0.2$; MF: $\mu = 0.75$; and MDLSA-Hybrid-BD-ACI-BCA: $\mu = 0.15$.

$$Recall = \frac{\text{No. of correctly retrieved documents}}{\text{No. of documents in relevant category}}. \quad (16)$$

In addition, the following measure is called "area under the precision-recall curve" (AUC) [9], [10], which is related to both above two measures

$$\text{AUC} = \sum_{i_A=2}^{n_{\max}} \frac{(P(i_A) + P(i_A - 1)) \times (R(i_A) - R(i_A - 1))}{2} \quad (17)$$

where $n_{\max}$ denotes the maximum number of retrieved documents, $P(i_A)$, and $R(i_A)$ denotes the precision and recall values with i_A documents retrieved.

2) Comparative Results: We first evaluate the retrieval performance of MDLSA and MDLSA-Hybrid based on above metrics. We empirically set the number of selected terms (or the size of vocabulary M) to 3000, i.e., $m = 3000$. We set the dimension of projected feature to 100, i.e., $d = 100$. We also included the effect study on these parameters in the next section. The numerically comparative results of different methods are summarized in Table IV, in which the results of MDLSA-Hybrid, MLM-Hybrid, and MF are based on the optimal weight μ. We also include the precision results visually shown in Fig. 3 when the retrieved documents, the most similar candidate documents from the dataset for each query, vary from 1 to 360. In Fig. 3, MDLSA-Hybrid, PCA, LSI, and VSM are based on the NORM weighting. It is observed that MDLSA-Hybrid-NORM outperforms all the other approaches with respect to AUC measure. The methods with the hybrid similarity perform better over the ones with only using either the local similarity or the global similarity. MDLSA-Hybrid, MDLSA, MLM-Hybrid, MLM-Local, and MF deliver superior results compared with the traditional techniques, PCA, LSI, RAP, PLSI, and VSM. The results suggest that LSI and PCA with appropriate preweighting and feature selection may outperform the statistical methods, i.e., PLSI and RAP, because currently these methods cannot utilize the preweighting schemes to boost their performance. In Fig. 3, it is interesting to observe that MLM-Local provides better results when a few documents are retrieved. MDLSA-related methods achieve the best results when only a single document is retrieved. In comparison with PCA, which utilizes the global semantics of documents, we listed the AUC Improvement by optimal combination of the global and local information in

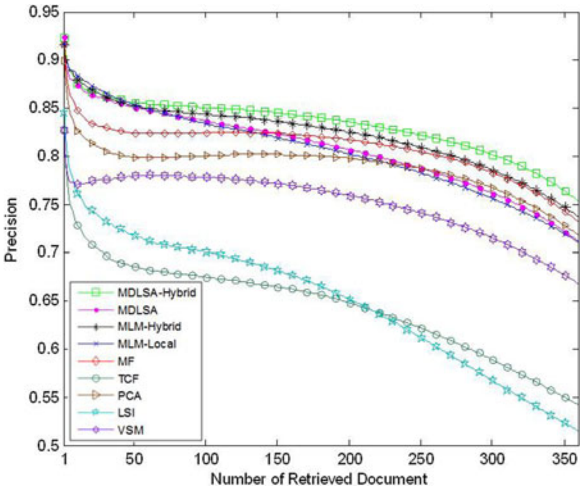


Fig. 3. Retrieval performance of different models.

TABLE V
AUC Improvement by Optimal Combination of the Global and Local Information

Weighting	PCA	MDLSA-Hybrid	Improvement
NORM	68.26	73.97	5.71%
BD-ACI-BCA	64.00	70.29	6.29%
AB-AFD-BAA	65.72	70.83	5.51%
BI-ACI-BCA	66.78	70.88	4.10%
SMART	66.11	70.78	4.67%

Table V. It is clear that combining the similarity from global and local does bring much performance efficiency. For instance, MDLSA-Hybrid achieves around 4.1% improvement of AUC value compared with MDLSA and brings around 5.7% AUC gain in contrast with PCA-NORM.

To examine the time performance of our proposed technique, we summarize the query time of different methods in Table VI. The query time of MDLSA-Hybrid, MDLSA, MF, TCF, PCA, and LSI is trivial, while MLM-Hybrid and MLM-Local require significant query time for users. As seen from Table VI, MDLSA-related methods can be up to 300 times faster than the MLM-related methods. We should notice that, in comparison with PCA and LSI, MDLSA-Hybrid requires additional time because it considers the semantics from term associations. For online applications, acceleration techniques, e.g., the concept of random indexing [41], may be employed to speed up the retrieval system.

3) Parameter Study: This section studies the effect of the parameters on the results. As shown in Table II, our studies include the number of columns selected from the projected matrix $\tilde{Z}$, i.e., the value of k associated with the matrix size of Z (see Section V-C), the weight μ involved in MDLSA-Hybrid algorithm, the vocabulary size m, and the dimension of projected features d. In this section, MDLSA-Hybrid and PCA are based on the NORM weighting.

First, we study the effect of the value of k, which is the number of columns selected from the projected matrix. As we mentioned in Section V-C, for MDLSA, in practice, setting k equal to 1 is sufficient to produce superior performance. In contrast, if $k > 1$, the matrix Z will be of dimension size $d \times k$. As a result, we have to develop an assembled metric to conduct a comparison between two documents. Let Z_p be the projected matrix from the word affinity graph associated with document p, and Z_q the projected matrix in terms of document q. The similarity between documents p and q is defined as

$$S_{\mathrm{MDLSA}}(p,q) = \frac{1}{k}\sum_{j=1}^{k} \exp\left(-1 + \frac{Z_p(\cdot,j)\cdot Z_q(\cdot,j)}{\|Z_p(\cdot,j)\|_2 \|Z_q(\cdot,j)\|_2}\right) \tag{18}$$

where $Z_p(\cdot,j)$ represents the jth column of matrix Z_p, and $Z_q(\cdot,j)$ denotes the jth column of matrix Z_q. Here, if we transform the similarity $S_{\mathrm{MDLSA}}(p,q)$ into a distance measure $D_{\mathrm{MDLSA}}(p,q) = 1 - S_{\mathrm{MDLSA}}(p,q)$, we can prove that the distance function $D_{\mathrm{MDLSA}}(p,q)$ is indeed a metric. The proof can be found in the Appendix. We plotted the AUC value against different numbers of columns in Fig. 4(a). It is obvious to see that only using the first column, i.e., $k = 1$, delivers the best performance. In the meantime, we observe that the AUC value increases gradually along with the increase of the value of k from 7. It indicates that using a larger value of k may improve the performance, but this will increase the computational cost due to the calculation of the assembled similarity $S_{\mathrm{MDLSA}}(p,q)$ between two matrices.

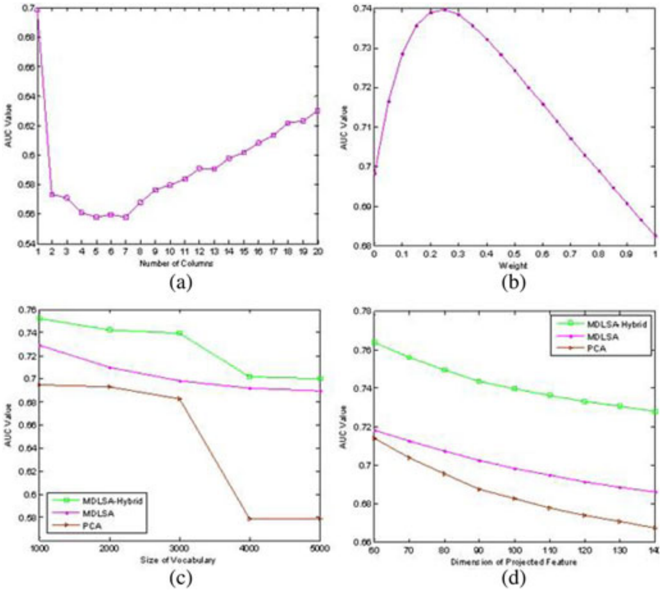


Fig. 4. AUC against (a) number of columns k, (b) weight μ, (c) vocabulary size m, and (d) dimension size of projected features d.

We then study the impact of the weight μ on the retrieval performance. Fig. 4(b) shows the AUC values produced by the MDLSA-Hybrid in terms of precision–recall curves against the weight values varying from 0 to 1 at an increment of 0.05. It is observed that there is an optimal weight to balance the importance of the global and local information in a way that the contribution of the global and local semantics is demonstrated. In this study, setting the weight μ to 0.25 for CityU1 appears to be the best option for the MDLSA-Hybrid method.

In this paper, we use the *tf-idf* weighting scheme [see (1)] to rank the importance of each term, and we select the first m terms as the vocabulary. The effect of different vocabulary sizes on the AUC performance is investigated. With the setting of $d = 100$, the results are shown in Fig. 4(c), where the vocabulary size m varies from 1000 to 5000. The results suggest that different values of m ranging from 1000 to 3000 do not have noticeable effect on the retrieval performance for MDLSA-Hybrid, MDLSA, and PCA. On the other hand, MDLSA-related methods consistently outperform the PCA for different vocabulary sizes.

At last in this section, we have also studied the effect of the projected dimension size d. With the setting of $m = 3000$, Fig. 4(d) shows the results of the AUC against the dimension of projected features that varies from 60 to 140 at an increment of 10. It is observed that the AUC values produced by MDLSA-Hybrid, MDLSA, and PCA decrease slightly with the increase of the dimension size. Thus, it indicates that the higher dimension does not necessarily bring better performance. In practice,

TABLE VI
QUERY RESPONSE PERFORMANCE OF DIFFERENT RETRIEVAL METHODS

Method	MDLSA-Hybrid	MDLSA	MLM-Hybrid	MLM-Local	MF	TCF	PCA	LSI	VSM
Query Time (s)	0.58	0.45	186.30	185.39	0.58	0.45	0.45	0.45	0.72

we can determine the optimal dimension by the use of experimental trials and prior information. It will be an interesting study on adapting this parameter for MDLSA in the future work. Again, MDLSA-related methods consistently outperform PCA for different values of d. In particular, MDLSA-Hybrid produces around 5% AUC gain for all values of d compared with PCA.

B. Document Classification

In this section, we experiment on the classification task. To evaluate the performance of our proposed methods, we use two public datasets: YahooScience and WebKB4. At present, classification, apart from retrieval, has also become important in organizing the massive amount of document data [42]–[44]. According to [26], document classification refers to automatically assigning predefined categories to free text documents. For simplicity, we have used a nearest neighbor classifier to perform the classification task based on the latent semantic features in the reduced space. It is worth pointing out that other advanced pattern classifiers, such as support vector machine [30], Bayesian [31], and neural networks [32], can be employed under our framework. Specifically, we can input the resulting feature vectors from MDLSA to these pattern classifiers and evaluate the distance between samples by setting an appropriate kernel in the classification process.

1) Evaluation Measures: To evaluate the quality of the classification, we adopted three measures that are widely used in the text classification and clustering literature [33], [34]. The first measure is the *Accuracy*, a commonly used measure, which is defined as

$$\text{Accuracy} = \frac{N_C}{n_T} \tag{19}$$

where N_C is the number of correctly classified documents, and n_T is the total number of tested documents.

The second measure is the *F-measure*, which simultaneously considers the *Precision* and *Recall* ideas from the information retrieval context, as shown in (15) and (16). In the classification context, the precision and recall of a predicted class j with respect to an actual class i are defined by

$$P_i = \frac{N_{ij}}{n_j} \tag{20}$$

and

$$R_i = \frac{N_{ij}}{n_i} \tag{21}$$

where N_{ij} is the number of members of actual class i in predicted class j, n_j is the number of members of predicted class j, and n_i is the number of members of actual class i. The F-measure of an actual class i is given by

$$F_i = \frac{2P_iR_i}{P_i + R_i}. \tag{22}$$

TABLE VII
DETAILS OF THE CLASSES IN YAHOOSCIENCE

Class	Number of Documents	Class Proportion (%)
Agriculture	159	18.47
Alternative	150	17.42
Biology	141	16.38
Chemistry	121	14.05
Earth Sciences	200	23.23
Mathematics	90	10.45

The overall F-measure for the classification result is defined as

$$F_O = \frac{\sum_i (n_i F_i)}{\sum_i n_i}. \tag{23}$$

By definition, the overall F-measure is the weighted average of the F-measure of each actual class i. The higher the overall F-measure, the better the classification, because of the higher accuracy of the predicted classes mapping to the actual classes.

The third measure is the *Entropy*, which provides a measure of "goodness" for unnested predicted classes or for the predicted classes at one level of a hierarchical classification [34]. Entropy indicates how homogeneous a predicted class is. The higher the homogeneity of a predicted class, the lower the entropy, and *vice versa*. For each predicted class j, we first calculate p_{ij}, the probability that a member of predicted class j belongs to actual class i. The entropy of each predicted class j is then computed by $E_j = -\sum_i p_{ij} \log(p_{ij})$. The total entropy of the classification result is given by

$$E_O = \sum\nolimits_j \left(\frac{n_j}{\sum_j n_j} E_j \right). \tag{24}$$

Basically, the objective of a classification task is to maximize the Accuracy and F-measure, and minimize the Entropy of the predicted classes. As such, we can accomplish high-quality classification results.

2) YahooScience: YahooScience is filed from the documents referenced the Open Directory Project, and it is publicly available.[2] This dataset has been used in [22]. The original collection of YahooScience included 907 documents in six top-level classes. For each top-level class, we first moved all the documents in its subclass to the top-level class and removed all the subclasses. We then removed all empty documents (due to the limited vocabulary size) and the documents containing only scripts. A total of 861 documents were left with YahooScience in six classes. Document size in this dataset is short. The average number of words in one document is around 900. The details of this dataset can be found in Table III. In addition, each class in YahooScience has a different number of documents. The details of the classes are listed in Table VII. The dataset was split in 25% test and 75% training data. We performed fourfold cross validation, and the results were averaged over the four folds.

[2] http://www.di.uniba.it/~malerba/software/webclass/WebClassIII.htm

TABLE VIII
CLASSIFICATION RESULTS OF DIFFERENT METHODS FOR YAHOOSCIENCE

Method	Accuracy (%)	F-measure	Entropy
MDLSA-Hybrid-NORM	92.09	0.9210	0.3256
MDLSA-Hybrid-SMART	92.09	0.9209	0.3146
MDLSA-Hybrid-AB-AFD-BAA	90.70	0.9063	0.3578
MDLSA-Hybrid-BI-ACI-BCA	90.70	0.9069	0.3514
MLM-Hybrid	90.70	0.9068	0.3882
MDLSA-Hybrid-BD-ACI-BCA	90.23	0.9025	0.3718
MLM-Local	89.77	0.8976	0.3986
MDLSA	88.84	0.8880	0.4188
MF	88.84	0.8887	0.4630
PCA-NORM	88.84	0.8887	0.4630
PCA-SMART	88.84	0.8870	0.4448
PCA-AB-AFD-BAA	87.91	0.8783	0.4615
LSI-NORM	87.44	0.8747	0.4917
LSI-BI-ACI-BCA	87.44	0.8734	0.4923
PCA-BD-ACI-BCA	86.98	0.8683	0.4936
PCA-BI-ACI-BCA	86.98	0.8689	0.4578
LSI-BD-ACI-BCA	86.98	0.8678	0.5180
LSI-AB-AFD-BAA	85.58	0.8557	0.5462
LSI-SMART	84.65	0.8458	0.5782
VSM-BI-ACI-BCA	84.65	0.8459	0.5658
VSM-AB-AFD-BAA	83.26	0.8323	0.6222
VSM-BD-ACI-BCA	81.86	0.8191	0.6697
VSM-SMART	81.86	0.8189	0.6692
VSM-NORM	81.40	0.8138	0.6829
RAP	78.60	0.7843	0.7736
PLSI	76.74	0.7657	0.8486
TCF	64.65	0.6481	1.1140
DGM	64.19	0.6505	1.0210

The weight μ settings for hybrid methods were MDLSA-Hybrid-NORM: $\mu = 0.45$; MLM-Hybrid: $\mu = 0.7$; MDLSA-Hybrid-BI-ACI-BCA: $\mu = 0.45$; MDLSA-Hybrid-AB-AFD-BAA: $\mu = 0.55$; MDLSA-Hybrid-SMART: $\mu = 0.45$; MF: $\mu = 1$; and MDLSA-Hybrid-BD-ACI-BCA: $\mu = 0.35$.

We summarized the average results of different methods in Table VIII for comparison. It is worth pointing out that DGM employs the original term affinity graphs without any dimensionality reduction and measures the between-document similarity based on the definition given in (18). We investigated the DGM because its comparison with the MDLSA will indicate the contribution of multidimensional dimensionality reduction to performance efficiency. Here, we empirically set the number of selected terms to 3000. We set the dimension of projected feature to 100. We will include the effect study on these parameters in the next section. From Table VIII, it is clear to observe that methods using the hybrid similarity deliver the best classification results over other methods. In particular, MDLSA-Hybrid-NORM achieves over 7% accuracy improvement compared with VSM, and it produces around 5% accuracy gain in contrast with LSI. Moreover, it is noted that MDLSA-Hybrid-NORM is capable of enhancing the accuracy with over 3% improvement compared with MDLSA and PCA using different preweights. It is also noted that the optimal value of weight μ is around 0.45, which indicates that the global and local information has almost equal contribution to measuring the similarity. In addition, we observe that TCF does not have contribution to accuracy gain as $\mu = 1$ for MF. It is clear to observe that MDLSA significantly outperforms the DGM method, and it produces over 20% accuracy gain in contrast with DGM. This result indicates the necessity of the dimensionality reduction to compress the term affinity graph and get rid of the impact of noise to the similarity measure. For clear comparison, we also summarized the improvement by combining the local information to PCA based on different preweights. The results are listed in Table IX. MDLSA-Hybrid brings around 3% performance gain with respect to the accuracy and F-measure, and achieves over 9% entropy improvement. Again, it indicates that combining the global and local information can accomplish performance enhancement.

We study the impact of the parameters on the classification results. MDLSA-Hybrid and PCA are based on the NORM weighting. First, we present the study of the number of columns k involved by the MDLSA algorithm. We plotted the classification accuracy against different values of k in Fig. 5(a). The results suggest that using the first column, i.e., $k = 1$, is more effective compared with the case of increasing the number of columns in the projected matrix. We then study the impact of the weight μ on the classification performance. Fig. 6(a) shows the accuracy produced by the MDLSA-Hybrid against the weight values varying from 0 to 1 at an increment of 0.05. It is observed that there may have multiple optimal weights to balance the importance of the global and local information for classification application. Moreover, we investigate the effect of different vocabulary sizes on the classification results. With the setting of $d = 100$, the results are shown in Fig. 7(a), where the vocabulary size m varies from 1000 to 5000. The results suggest that different values of m ranging from 2000 to 4000 perform better on the results for MDLSA-Hybrid and PCA. In addition, MDLSA-Hybrid consistently outperforms PCA and MDLSA for different vocabulary sizes. At last, we have also studied the effect of the projected dimension size d. With the setting of $m = 3000$, Fig. 8(a) shows the results of the accuracy against the dimension of projected features that varies from 60 to 140 at an increment of 10. It appears that setting $d = 90$ can bring the best performance for MDLSA-Hybrid. Selecting the best d is a trivial work. In general, we set d to 100 in this paper. Note that MDLSA-Hybrid can consistently outperform PCA in a significant rate for different values of d.

For completeness, we also experiment on the results of the accuracy against different proportions of training set. MDLSA-Hybrid and PCA are tested using the NORM weighting. The results are shown in Fig. 9(a). Here, we set $\mu = 0.45$ for MDLSA-Hybrid for all different proportions of training set. It is observed that MDLSA-Hybrid performs slightly better in terms of holding out 50% for training purpose. Surprisingly, PCA performs well for the case of using 90% training set. It is also suggested that, for YahooScience dataset, the global information produced by PCA has more discriminative power than the local information represented by MDLSA.

3) WebKB4: To demonstrate the performance of our proposed methods, we experiment on WebKB4, another publicly availabledataset.[3] This dataset has been used in [4] and [7]. WebKB4 is a subset of the WebKB dataset, and it contains 4199 Web pages in four categories collected from university computer science departments. We then removed all empty documents and the documents containing only scripts. A total of 4177 documents were left with WebKB4 in four classes. Average document size in this dataset is shorter. The average number of words in one document is around 290. The details of this dataset

[3] http://www.cs.cmu.edu/∼textlearning

TABLE IX
CLASSIFICATION IMPROVEMENT BY OPTIMAL COMBINATION OF THE GLOBAL AND LOCAL INFORMATION FOR YAHOOSCIENCE

	PCA			MDLSA-Hybrid			Improvement		
Weighting	Accuracy (%)	F-measure	Entropy	Accuracy (%)	F-measure	Entropy	Accuracy	F-measure	Entropy
NORM	88.84	0.8887	0.4630	92.09	0.9210	0.3256	+3.25%	+3.23%	-9.74%
BD-ACI-BCA	86.98	0.8683	0.4936	90.23	0.9025	0.3718	+3.25%	+3.42%	-12.18%
AB-AFD-BAA	87.91	0.8783	0.4615	90.70	0.9063	0.3578	+2.79%	+2.80%	-10.37%
BI-ACI-BCA	86.98	0.8689	0.4578	90.70	0.9069	0.3514	+3.72%	+3.80%	-10.64%
SMART	88.84	0.8870	0.4448	92.09	0.9209	0.3146	+3.25%	+3.39%	-13.02%

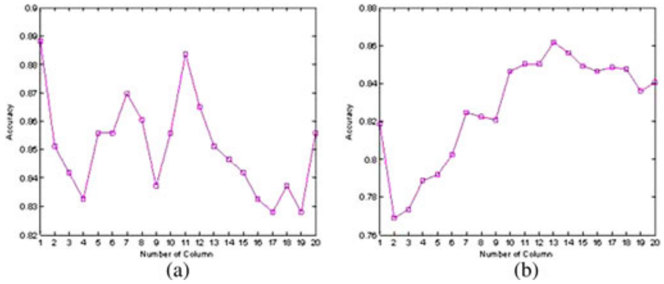


Fig. 5. Classification accuracy against different number of columns k for (a) YahooScience and (b) WebKB4.

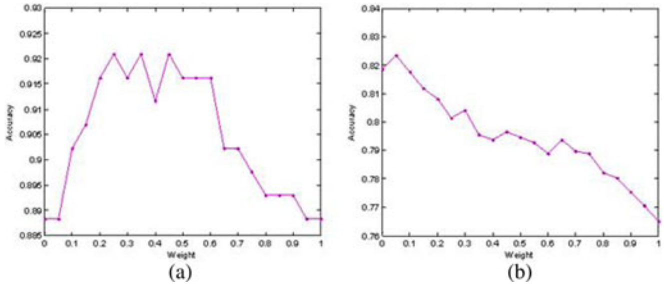


Fig. 6. Classification accuracy against different weights μ for (a) YahooScience and (b) WebKB4.

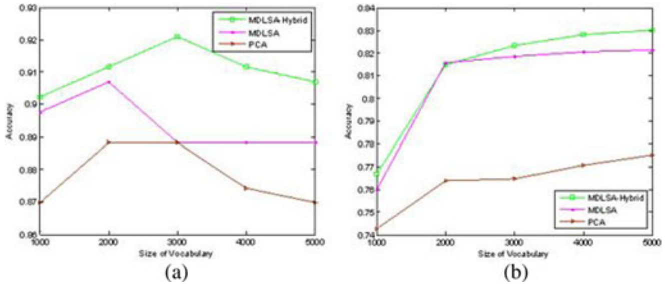


Fig. 7. Classification accuracy against different vocabulary sizes m for (a) YahooScience and (b) WebKB4.

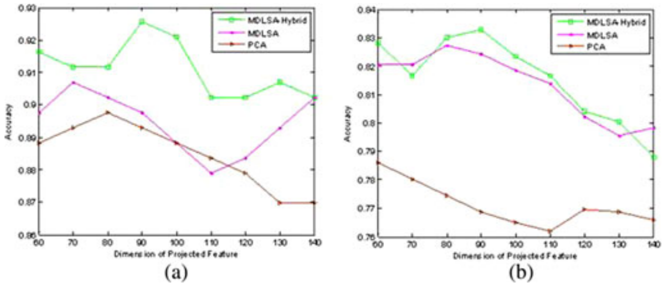


Fig. 8. Classification accuracy against different dimensions of projected features d for (a) YahooScience and (b) WebKB4.

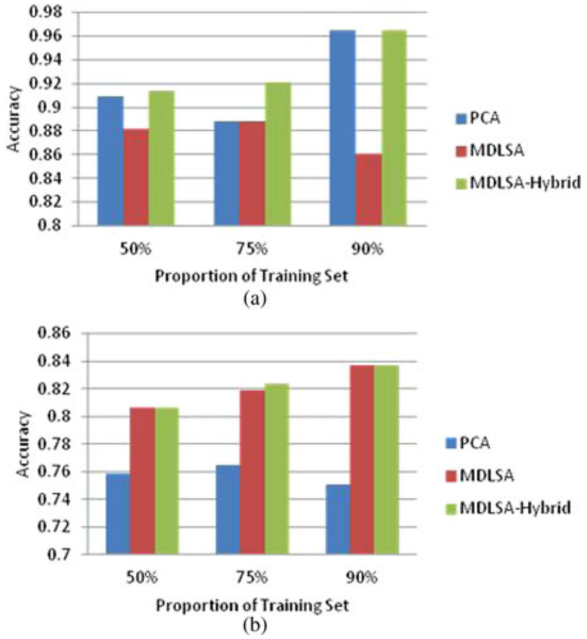


Fig. 9. Classification accuracy against different proportions of training set for (a) YahooScience and (b) WebKB4.

TABLE X
DETAILS OF THE CLASSES IN WEBKB4

Class	Number of Documents	Class Proportion (%)
Course	930	22.30
Faculty	1120	26.85
Project	503	12.06
Student	1618	38.79

can be found in Table III. Table X specifies the details of the classes. The dataset was split in 25% test and 75% training data. The results were based on fourfold cross validation.

The results of different methods are shown in Table XI for comparison. We empirically set the number of selected terms $m = 3000$ and the dimension of projected feature $d = 100$. Results show that MDLSA-Hybrid methods outperform other methods. In particular, MDLSA-Hybrid-NORM achieves over 5% accuracy improvement compared with PCA-NORM, LSI-NORM, and VSM-NORM. Setting a smaller μ for the optimal weights in both MDLSA-Hybrid-NORM and MDLSA-Hybrid-BI-ACI-BCA indicates that the local information produced by MDLSA owns more discriminative power than the global information represented by PCA. Comparative results associated with PCA and MDLSA-Hybrid in terms of different weighting schemes are listed in Table XII. It is clear that MDLSA-Hybrid outperforms PCA with respect to different evaluation measures.

TABLE XI
CLASSIFICATION RESULTS OF DIFFERENT METHODS FOR WEBKB4

Method	Accuracy (%)	F-measure	Entropy
MDLSA-Hybrid-SMART	83.30	0.8326	0.5783
MDLSA-Hybrid-BD-ACI-BCA	82.92	0.8288	0.5880
MDLSA-Hybrid-AB-AFD-BAA	82.63	0.8255	0.5919
MDLSA-Hybrid-NORM	82.34	0.8224	0.5999
MDLSA-Hybrid-BI-ACI-BCA	82.15	0.8202	0.6401
MDLSA	81.86	0.8182	0.6136
PCA-AB-AFD-BAA	81.38	0.8121	0.6277
PCA-BD-ACI-BCA	80.81	0.8071	0.6293
MF	79.94	0.7977	0.6513
PCA-SMART	79.46	0.7935	0.6571
PCA-BI-ACI-BCA	79.37	0.7934	0.6566
LSI-AB-AFD-BAA	79.37	0.7911	0.6880
MLM-Hybrid	78.02	0.7775	0.6898
LSI-BD-ACI-BCA	77.93	0.7782	0.7101
LSI-BI-ACI-BCA	77.54	0.7740	0.7220
LSI-SMART	77.06	0.7706	0.7258
PCA-NORM	76.49	0.7632	0.7220
VSM-AB-AFD-BAA	75.53	0.7513	0.7716
PLSI	75.43	0.7545	0.7507
LSI-NORM	74.47	0.7443	0.7655
VSM-BI-ACI-BCA	74.18	0.7382	0.7751
VSM-SMART	73.99	0.7366	0.7837
VSM-BD-ACI-BCA	73.80	0.7347	0.7939
RAP	71.50	0.7128	0.8561
MLM-Local	71.40	0.7088	0.8663
VSM-NORM	69.87	0.6966	0.8749
TCF	67.95	0.6762	0.9147

The weight μ settings for hybrid methods were MDLSA-Hybrid-NORM: $\mu = 0.05$; MLM-Hybrid: $\mu = 0.75$; MDLSA-Hybrid-BI-ACI-BCA: $\mu = 0.1$; MDLSA-Hybrid-AB-AFD-BAA: $\mu = 0.5$; MDLSA-Hybrid-SMART: $\mu = 0.35$; MF: $\mu = 0.65$; and MDLSA-Hybrid-BD-ACIBCA: $\mu = 0.5$.

In particular, MDLSA-Hybrid with the SMART and NORM preweights delivers significant performance improvement.

The impact of the parameters on the classification results is investigated in this section. Here, MDLSA-Hybrid and PCA employ the NORM weighting. In Fig. 5(b), we plotted the accuracy against different values of k, which is the number of columns of the projected matrix lying in the MDLSA algorithm, ranging from 1 to 20. It is interesting to observe that setting k equal to 2 from 1 degrades the performance in a significant rate, while increasing the value of k larger than 2 will gradually enhance the accuracy. It appears that using the first 12 columns, i.e., $k = 12$, reaches the optimal value, which produces around 86% classification rate, but this will require extra computational cost because of the calculation of the assembled similarity between two matrices. The effect of the weight μ on the classification performance is illustrated in Fig. 6(b), which shows the accuracy produced by the MDLSA-Hybrid against the weight values. It is observed that a small weight tends to deliver better results, and it suggests that the local information owns more discriminative power than the global information for the WebKB4 dataset. Furthermore, setting $d = 100$, we investigate the effect of different vocabulary sizes on the results shown in Fig. 7(b). The results indicate that increasing the vocabulary size may bring accuracy improvement in an insignificant rate. On the other hand, with the setting of $m = 3000$, Fig. 8(b) illustrates the accuracy results against different dimension sizes of projected features. It shows that setting d to around 90 can bring the best performance for MDLSA-Hybrid. Note that MDLSA-Hybrid and MDLSA can consistently outperform PCA in a significant rate for different values of d.

We experiment on the results of the accuracy against different proportions of training set. The results are summarized in Fig. 9(b). MDLSA-Hybrid and PCA use the NORM weighting. We set $\mu = 0.05$ for MDLSA-Hybrid for all different proportions of training set. It is observed that MDLSA-Hybrid and MDLSA, which take advantage of the local information, consistently outperforms PCA with using the global information of documents. It also indicates that, for WebKB4 dataset, the local semantics produced by MDLSA has more discriminative power than the global semantics represented by PCA.

C. Study of Multidimensional Latent Semantic Analysis at the Sentence Level

To evaluate the effects of MDLSA at a lower level, we further segmented each Web document into sentences by marking periods. This further segmentation was examined over the YahooScience dataset. The classification results were summarized in Table XIII. In comparison with Table VIII, it is observed that the MDLSA-Hybrid methods show very similar performance. In particular, MDLSA-Hybrid-SMART delivers around 1% accuracy improvement compared with the case of the segmentation at the paragraph level.

VIII. DISCUSSION

In previous works, we observe that adding local information with respect to term associations into document representation does improve the performance of various document applications [10]– [23]. However, formulating an efficient representation of this local information is still a demanding issue. Thus, this research focuses on the modeling of local information extracted from documents. Then, we develop a hybrid similarity measure that combines the global and local information together. From the experimental results, many interesting observations can be shown.

1) Using the local information from paragraphs to represent a document is able to achieve satisfied performance because this representation delivers some discriminative information, which cannot be represented by the global information from the entire document.
2) The hybrid similarity measure integrating the global and local information can further enhance the performance of document applications with setting an appropriate weighting parameter.
3) The proposed method for mining the local information can significantly improve the time performance compared with the MLM method [23].
4) The best choice of parameter k, which is the number of columns of the projected matrix lying in the MDLSA algorithm, depends on the dataset. Setting $k = 1$ usually performs well from our experimental observations.
5) The option of the optimal weight μ to balance the global and local information with respect to the hybrid similarity measure depends on the dataset. If no prior information is available, we can assign $\mu = 0.5$ to equally balance the global and local information.

TABLE XII
CLASSIFICATION IMPROVEMENT BY OPTIMAL COMBINATION OF THE GLOBAL AND LOCAL INFORMATION FOR WEBKB4

	PCA			MDLSA-Hybrid			Improvement		
Weighting	Accuracy (%)	F-measure	Entropy	Accuracy (%)	F-measure	Entropy	Accuracy	F-measure	Entropy
NORM	76.49	0.7632	0.7220	82.34	0.8224	0.5999	+5.85%	+5.92%	-12.21%
BD-ACI-BCA	80.81	0.8071	0.6293	82.92	0.8288	0.5880	+2.11%	+2.17%	-4.13%
AB-AFD-BAA	81.38	0.8121	0.6277	82.63	0.8255	0.5919	+1.25%	+1.34%	-3.58%
BI-ACI-BCA	79.37	0.7934	0.6566	82.15	0.8202	0.6401	+2.78%	+2.68%	-1.65%
SMART	79.46	0.7935	0.6571	83.30	0.8326	0.5783	+3.84%	+3.91%	-7.88%

TABLE XIII
CLASSIFICATION RESULTS OF DIFFERENT METHODS FOR YAHOOSCIENCE AT THE SENTENCE LEVEL

Method	Accuracy (%)	F-measure	Entropy
MDLSA-Hybrid-NORM	91.63	0.9159	0.3321
MDLSA-Hybrid-BD-ACI-BCA	91.16	0.9105	0.3569
MDLSA-Hybrid-AB-AFD-BAA	90.23	0.9015	0.3908
MDLSA-Hybrid-BI-ACI-BCA	91.16	0.9108	0.3445
MDLSA-Hybrid-SMART	93.02	0.9299	0.2772
MDLSA	84.19	0.8412	0.6030

The weight μ settings for hybrid methods were MDLSA-Hybrid-NORM: μ = 0.35; MDLSA-Hybrid-BD-ACI-BCA: μ = 0.35; MDLSAHybrid-AB-AFD-BAA: μ = 0.45; MDLSA-Hybrid-BI-ACI-BCA: μ = 0.4; and MDLSA-Hybrid-SMART: μ = 0.35.

6) The vocabulary size m usually produces a slight effect on the results, but given the storage space and computational burden, it can be set at a few thousands.
7) The best choice of the size of projected features d depends on the dataset, but it can be usually set at around 100.
8) The extent of our method depending on the preweighting strategy is small. In particular, the NORM weighting for LSI, PCA, VSM, and MDLSA-Hybrid delivered promising results across experiments.

From the experimentalobservations, we believe that the achievement of a number of desirable features produced by our proposed framework is based on the following reasons: 1) the word affinity graph is capable of accurately describing the term associations such that more discriminative information can be delivered; 2) MDLSA is an efficient method to compress the sparse and large-size matrix associated with the word affinity graph; and 3) the traditional document representation produced by LSI or PCA contains the global information, which derives from the features that they are independent from each other, while MDLSA includes the local information based on the word affinity graph where terms are dependent. The integration of the global and local information by a hybrid similarity measure produces a complete comparison of two documents such that the similarity of them can be evaluated effectively.

IX. CONCLUSION

This paper has presented a new document analysis method, MDLSA, which enables us to extract the local information efficiently from documents with respect to term associations. We first partition each document into paragraphs and build a term affinity graph. Each element of this graph represents the frequency of term cooccurrence in a paragraph. We then conduct a 2DPCA to achieve an optimal semantic mapping. This analysis works by finding the leading eigenvectors of the sample covariance matrix to characterize the lower dimensional semantic space. A hybrid document similarity measure is designed to further improve the performance of MDLSA. MDLSA delivers three important advantages. First, in contrast with the 2DPCA, it does not require an assembled metric to conduct matrix comparison. As a result, MDLSA is easier to make between comparisons. Second, much less time is required because MDLSA does not need the many-to-many matching compared with the MLM method. Third, MDLSA includes local semantic information of documents in comparison to the PCA and the LSI [2]. We experimented on three public datasets in terms of two tasks: retrieval and classification. The results strongly suggest that the proposed technique is accurate and computationally efficient for performing various document applications. In the future work, we plan to investigate the potential of other methods (e.g., tree structure) under our framework to represent the semantic meaning of a document instead of using a term affinity graph. It will also be interesting to investigate the performance comparison between dimensionality reduction techniques and statistical methods with respect to different vocabulary sizes and feature selection schemes.

APPENDIX

Here, we show the proof that the distance $D_{\mathrm{MDLSA}}(p,q) = 1 - S_{\mathrm{MDLSA}}(p,q)$ associated with the similarity $S_{\mathrm{MDLSA}}(p,q)$ used in Section VII-A3 is a metric. Since we can write the complete expression of $D_{\mathrm{MDLSA}}(p,q)$ by

$$\begin{aligned} D_{\mathrm{MDLSA}}(p,q) &= 1 - \frac{1}{k}\sum_{j=1}^{k}\exp\left(-1 + \frac{Z_p(\cdot,j)\cdot Z_q(\cdot,j)}{\|Z_p(\cdot,j)\|_2\|Z_q(\cdot,j)\|_2}\right) \\ &= \frac{1}{k}\sum_{j=1}^{k}\left(1 - \exp\left(-1 + \frac{Z_p(\cdot,j)\cdot Z_q(\cdot,j)}{\|Z_p(\cdot,j)\|_2\|Z_q(\cdot,j)\|_2}\right)\right) \\ &= \frac{1}{k}\sum_{j=1}^{k} D_j(p,q) \end{aligned} \tag{25}$$

it is clear that positive definiteness and symmetry hold. We now prove that the triangle inequality also holds.

Let us define

$$\rho_j(p,q) = 1 - \frac{Z_p(\cdot,j)Z_q(\cdot,j)}{\|Z_p(\cdot,j)\|_2\|Z_q(\cdot,j)\|_2} \tag{26}$$

$$\rho_j(q,r) = 1 - \frac{Z_q(\cdot,j)Z_r(\cdot,j)}{\|Z_q(\cdot,j)\|_2\|Z_r(\cdot,j)\|_2} \tag{27}$$

and

$$\rho_j(p,r) = 1 - \frac{Z_p(\cdot,j)Z_r(\cdot,j)}{\|Z_p(\cdot,j)\|_2\|Z_r(\cdot,j)\|_2}. \tag{28}$$

Given that

$$\rho_j(p,q) + \rho_j(q,r) \geq \rho_j(p,r) \tag{29}$$

we can write

$$\begin{aligned}0 &\leq (1-\exp(-\rho_j(p,q)))(1-\exp(-\rho_j(q,r)))\\ &= 1-\exp(-\rho_j(q,r)) - \exp(-\rho_j(p,q))\\ &\quad + \exp(-(\rho_j(p,q)+\rho_j(q,r)))\\ &\leq 1-\exp(-\rho_j(q,r)) - \exp(-\rho_j(p,q)) + \exp(-\rho_j(p,r))\\ &= (1-\exp(-\rho_j(p,q))) + (1-\exp(-\rho_j(q,r)))\\ &\quad - (1-\exp(-\rho_j(p,r)))\\ &= D_j(p,q) + D_j(q,r) - D_j(p,r). \end{aligned}\tag{30}$$

Hence, $D_j(p,r) \leq D_j(p,q) + D_j(q,r)$ for $j = 1, 2, \ldots, k$. As a result

$$\frac{1}{k}\sum_{j=1}^{k} D_j(p,r) \leq \frac{1}{k}\sum_{j=1}^{k} D_j(p,q) + \frac{1}{k}\sum_{j=1}^{k} D_j(q,r). \tag{31}$$

Therefore $D_{\mathrm{MDLSA}}(p,q) + D_{\mathrm{MDLSA}}(q,r) \geq D_{\mathrm{MDLSA}}(p,r)$.

Acknowledgment

The authors would like to thank reviewers for their detailed and useful comments.

References

[1] G. Salton, M. McGill, Eds. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.

[2] S. Deerwester and S. Dumais, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[3] E. Kokiopoulou and Y. Saad, "Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2143–2156, Dec. 2007.

[4] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 12, pp. 1624–1637, Dec. 2005.

[5] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. SIGIR Conf.*, 1999, pp. 50–57.

[6] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[7] N. Bouguila, "Clustering of count data using generalized Dirichlet multinomial distributions," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 4, pp. 462–474, Apr. 2008.

[8] M. Welling, M. Rosen-Zvi, and G. Hinton, "Exponential family harmoniums with an application to information retrieval," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, vol. 17, pp. 1481–1488.

[9] P. Gehler, A. Holub, and M. Welling, "The rate adapting Poisson model for information retrieval and object recognition," in *Proc. 23rd Int. Conf. Mach. Learn.*, Pittsburgh, PA, 2006, pp. 337–344.

[10] H. Zhang, T. W. S. Chow, and M. K. M. Rahman, "A new dual wing harmonium model for document retrieval," *Pattern Recognit.*, vol. 42, no. 11, pp. 2950–2960, 2009.

[11] A. Schenker, M. Last, H. Bunke, and A. Kandel, "Classification of web documents using graph matching," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 18, no. 3, pp. 475–496, 2004.

[12] M. Fuketa, S. Lee, T. Tsuji, M. Okada, and J. Aoe, "A document classification method by using field association words," *Inf. Sci.*, vol. 126, no. 1–4, pp. 57–70, 2000.

[13] C. M. Tan, Y. F. Wang, and C. D. Lee, "The use of bigrams to enhance text categorization," *Inf. Process. Manag.*, vol. 38, no. 4, pp. 529–546, 2002.

[14] M. L. Antonie and O. R. Zaiane, "Text document categorization by term association," in *Proc. IEEE Int. Conf. Data Mining*, 2002, pp. 19–26.

[15] J. Kim and M. H. Kim, "An evaluation of passage-based text categorization," *J. Intell. Inf. Syst.*, vol. 23, no. 1, pp. 47–65, 2004.

[16] X. B. Xue and Z. H. Zhou, "Distributional features for text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 3, pp. 428–442, Mar. 2009.

[17] L. A. F. Park, K. Ramamohanarao, and M. Palaniswami, "Fourier domain scoring: A novel document ranking method," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 5, pp. 529–539, May 2004.

[18] L. A. F. Park, M. Palaniswami, and K. Ramamohanarao, "A novel document ranking method using the discrete cosine transform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 1, pp. 130–135, Jan. 2005.

[19] L. A. F. Park, K. Ramamohanarao, and M. Palaniswami, "A novel document retrieval method using the discrete wavelet transform," *ACM Trans. Inf. Syst.*, vol. 23, no. 3, pp. 267–298, 2005.

[20] T. W. S. Chow and M. K. M. Rahman, "Multi-layer SOM with tree structured data for efficient document retrieval and plagiarism detection," *IEEE Trans. Neural Netw.*, vol. 20, no. 9, pp. 1385–1402, Sep. 2009.

[21] T. W. S. Chow, H. Zhang, and M. K. M. Rahman, "A new document representation using term frequency and vectorized graph connectionists with application to document retrieval," *Exp. Syst. Appl.*, vol. 36, pp. 12023–12035, 2009.

[22] H. Zhang, T. W. S. Chow, and M. K. M. Rahman, "A novel dual wing harmonium model aided by 2-D wavelet transform subbands for document data mining," *Exp. Syst. Appl.*, vol. 37, no. 6, pp. 4403–4412, 2010.

[23] H. Zhang and T. W. S. Chow, "A coarse-to-fine framework to efficiently thwart plagiarism," *Pattern Recognit.*, vol. 44, no. 2, pp. 471–487, 2011.

[24] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.

[25] J. Yang, D. Zhang, A. F. Frangi, and J.-Y. Yang, "Two-Dimensional PCA: A new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131–137, Jan. 2004.

[26] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. Int. Worksh. Mach. Learn.*, 1997, pp. 415–420.

[27] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, 1988.

[28] J. Zobel and A. Moffat, "Exploring the similarity space," *ACM SIGIR Forum*, vol. 32, no. 1, pp. 18–34, 1998.

[29] W. Zuo, D. Zhang, and K. Wang, "Bidirectional PCA with assembled matrix distance metric for image recognition," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 36, no. 4, pp. 863–872, Aug. 2006.

[30] P. J. Phillips, "Support vector machines applied to face recognition," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 1998, vol. 11, pp. 803–809.

[31] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 696–710, Jul. 1997.

[32] H. Bischof, W. Schneider, and A. J. Pinz, "Multispectral classification of Landsat-images using neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 30, no. 3, pp. 482–490, May 1992.

[33] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," Univ. Minnesota, Tech. Rep. #00-034, Aug. 2000.

[34] K. M. Hammouda and M. S. Kamel, "Efficient phrase-based document indexing for web document clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 10, pp. 1279–1296, Oct. 2004.

[35] K. Lund and C. Burgess, "Producing high-dimensional semantic spaces from lexical co-occurrence," *Behav. Res. Meth.*, vol. 28, no. 2, pp. 203–208, 1996.

[36] D. L. T. Rohde, L. M. Gonnerman, and D. C. Plaut, "An improved model of semantic similarity based on lexical co-occurrence," *Commun. ACM*, vol. 8, pp. 627–633, 2006.

[37] S. Patwardhan, "Using wordnet-based context vectors to estimate the semantic relatedness of concepts," in *Proc. Eur. Ch. Assoc. Comput. Linguistics*, 2006, pp. 1–8.

[38] M. Lapata and R. Barzilay, "Automatic evaluation of text coherence: Models and representations," in *Proc. 19th Int. Joint Conf. Artif. Intell.*, 2005, pp. 1085–1090.

[39] R. Barzilay and M. Lapata, "Modeling local coherence: An entity-based approach," in *Proc. 43rd Annu. Meet. Assoc. Comput. Linguistics*, 2005, pp. 141–148.

[40] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis," in *Proc. 20th Int. Joint Conf. Artif. Intell.*, 2007, pp. 1606–1611.

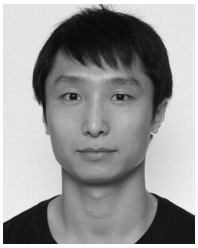
[41] P. Kanerva, J. Kristoferson, and A. Holst, "Random indexing of text samples for latent semantic analysis," in *Proc. 22nd Annu. Conf. Cognit. Sci. Soc.*, 2000, pp. 103–106.

[42] C. Silva, etc., “Distributed text classification with an ensemble kernel-based learning approach” *IEEE Trans. Syst. Man, Cybern. C, Appl. Rev.*, vol. 40, no. 3, pp. 287–297, May 2010.

[43] N. Oza, J. P. Castle, and J. Stutz, “Classification of aeronautics system health and safety documents,” *IEEE Trans. Syst. Man, Cybern. C, Appl. Rev.*, vol. 39, no. 6, pp. 670–680, Nov. 2009.

[44] N. Tsimboukakis and G. Tambouratzis, “Word-map systems for content-based document classification,” *IEEE Trans. Syst. Man, Cybern. C, Appl. Rev.*, vol. 41, no. 5, pp. 662–673, Sep. 2011.

[45] L. A. F. Pak, “Fast approximate text document clustering using compressive sampling,” *Lect. Notes Comput. Sci.*, vol. 6912, pp. 565–580, 2011.



Haijun Zhang received the B.Eng and Master’s degrees from the Northeastern University, Shenyang, China, in 2004 and 2007, respectively, and the Ph.D. degree from the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong, in 2010.

From 2010 to 2011, he was a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada. Since 2012, he has been with the Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China, where he is currently an Associate Professor of Computer Science. His research interests include multimedia data mining, machine learning, pattern recognition, evolutionary computing, and communication networks.

John K. L. Ho received the B.Sc. and M.Sc. degrees in computer, control engineering from Coventry University, West Midlands, U.K., and the Ph.D. degree from the University of East London, London, U.K.

He has many years of design experience in the field of automation when was working with GEC Electrical Projects Ltd in U.K. He is currently an Associate Professor with the Department of Mechanical and Biomedical Engineering, City University of Hong Kong, Kowloon, Hong Kong. His research interests are in the fields of data mining, control engineering, green manufacturing, enterprise automation, and product design.

Dr. Ho is the Chairman of the Control, Automation and Instrumentation Discipline Advisory Panel of the Hong Kong Institution of Engineers.

Q. M. Jonathan Wu (M’92–SM’09) received the Ph.D. degree in electrical engineering from the University of Wales, Swansea, U.K., in 1990.

He was with the National Research Council of Canada for ten years from 1995, where he later became a Senior Research Officer and a Group Leader. He is currently a Professor with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada. He has contributed to more than 250 peer-reviewed papers in computer vision, image processing, intelligent systems, robotics, and integrated microsystems. His current research interests include 3-D computer vision, active video object tracking and extraction, interactive multimedia, sensor analysis and fusion, and visual sensor networks.

Dr. Wu holds the Tier 1 Canada Research Chair in Automotive Sensors and Information Systems. He is an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, and the *International Journal of Robotics and Automation*. He has served on Technical Program Committees and International Advisory Committees for many prestigious conferences.

Yunming Ye received the Ph.D. degree in computer science from Shanghai Jiao Tong University, Shanghai, China.

He is currently a Professor with Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. His research interests include data mining, text mining, and clustering algorithms.