2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Classifying Web Documents Using Term Spectral Transforms and Multi-Dimensional Latent Semantic Representation

Haijun Zhang, Shifu Bie and Bin Luo
Department of Computer Science
Shenzhen Graduate School, Harbin Institute of Technology
Shenzhen, China
Email: aarhzhang@gmail.com; bensonku09@gmail.com; hitluobin@gmail.com

***Abstract*— This research investigates the potential of document semantic representation considering both term frequencies and term associations. In particular, we proposed a general framework of the use of term spectra to represent term spatial distributions and associations through a document. The term spectra we explored involved the use of three typical techniques: Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT), and Discrete Wavelet Transform (DWT). A term affinity graph was established to represent each document. We then employed a new document analysis method (recently developed by authors), named Multi-Dimensional Latent Semantic Analysis (MDLSA), which enables us to formulate an efficient semantic representation of a document based on the term affinity graph. Our algorithm was examined in the application of Web document classification. Experimental results demonstrate that the proposed technique not only gains much computational efficiency compared to Direct Graph Matching (DGM), but also outperforms the state-of-art algorithms such as VSM, PCA, RAP, and MLM.**

I. Introduction

BROWSING Web pages and finding information on them for business development, scientific research and entertainment have become an indispensable part of people's daily life. Web document classification/categorization is a growing demand for Website owners and ease of use for Website visitors. In general, posting information on a Web page involves two essential phases: 1) the selection of terms (or words); 2) the arrangement of terms. The first phase is very related to the basic semantics that the author wants to deliver. This semantic description can be measured by counting the term frequency (*tf*) that a term appears in the document. The second phase lies in the term associations and distributions over a document. Thus, an accurate representation of the text content must, at least, include the information from *tf* and term associations.

With respect to the *tf* features, the last two decades have witnessed the rapid development of the "Bag of Words" (BoW) models. The earliest work relying on the BoW model is the Vector Space Model (VSM) [1], which usually uses the *tf-idf* scheme for term weighting. The beauty of the VSM is the capability of reducing arbitrary length of each document to a fixed length by a term vector. Nevertheless, a lengthy vector is required because the number of words involved is usually an enormous amount. Additionally, the VSM reveals little statistical property of the semantic due to using only low level features (e.g. *tf*). To overcome these shortcomings, two popular semantic representation techniques have been proposed: 1) eigen-semantics based method by finding the solution of an eigenvalue problem; 2) statistical semantics based method using statistical inference and machine learning. The typical methods for document modeling in terms of eigen-semantics consist of Latent Semantic Indexing (LSI) [2], Principal Component Analysis (PCA), linear discriminant analysis [3], and locality preserving models [4][5]. These techniques are largely based on the use of low dimensional representations to capture the document semantics. Besides dimensionality reduction method, statistical modeling such as Probabilistic Latent Semantic Indexing (PLSI) [6], Latent Dirichlet Allocation (LDA) [7], Exponential Family Harmonium (EFH) model [8], and Rate Adapting Poisson (RAP) model [9], also have become fashionable in the last decade. Despite the great success of the aforementioned approaches, they are all in line with the nature of the BoW model, which only relies on the *tf* information. This feature extraction strategy is a rough representation of a document. As a result, it is inevitable that certain useful semantic information will be lost. For instance, two documents containing similar term frequencies may appear to be contextually different when their spatial distribution of terms is different. Thus, relying on only the *tf* information is not the most reliable way to account contextual similarity. The semantics may be very different depending on whether one considers the term interconnections and spatial distributions or not.

There have been a few research efforts on using term associations to improve the performance of document applications. For the classification of Web documents, different directed graphs are defined to represent each document [10]. Although the graph matching is successful in enhancing the classification accuracy, the process must be accomplished in polynomial time, which makes it impractical in particular for large data sets. Fuketa *et al.* [11] introduced a field understanding method by using the field association terms.Others used either bigrams [12] or term association rules [13] to enhance the classification accuracy. Another interesting study of

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

considering term associations is the spectral-based approach reported by Park *et al*. [14][15][16]. 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. Their approach does yield impressive results to enhance the text retrieval performance. Nevertheless, it is only applicable to the case of a few keywords as a query. How the term spectra contribute to a general document application, which relies on between-document similarity, still remains unclear. In the latest work, a Multi-Level Matching (MLM) strategy was designed for retrieval [17] and plagiarism detection [18]. Despite the promising performance, the major drawback of the MLM lies in the computational burden of calculating the Earth Mover's Distance(EMD)[19]. The time cost increases exponentially as the number of paragraphs increases.

In this research, we design a unified framework considering both term frequencies and term associations, and we investigate the potential of this framework to boost the performance of Web document classification. Specifically, our model starts by partitioning each Web document into paragraphs via identifying the HTML tags. The spatial distribution of a term is then characterized by the term signal. Three spectral transforms: Discrete Cosine Transform (DCT) [14], Discrete Fourier Transform (DFT) [15], and Discrete Wavelet Transform (DWT) [16], are employed to analyze each term signal at the paragraph level. We then combine the spectra of two terms to represent their association. We establish a term affinity graph (TAG) for each document, where the value of each component indicates the association of the indexed terms. To reduce the computational burden, we use the algorithm of Multi-Dimensional Latent Semantic Analysis (MDLSA), which is recently developed by authors [20], to explore a low dimensional semantic space. The MDLSA utilizes the power of Two-Dimensional Principal Component Analysis (2DPCA) [21] to achieve optimal mapping in the reduced semantic space. We also have designed a new similarity measure for between-document comparison. The proposed framework is examined in three public data sets which include HTML documents. Experimental results suggest that MDLSA with TAG delivers a slight improvement in accuracy compared to traditional methods (e.g. PCA) for the small data set with short documents, but it outperforms other methods by a relatively large amount for the data set with relatively long documents.

The remaining sections of this paper are organized as follows. The TAG construction process is illustrated in Section II. And then, we present the MDLSA algorithm in Section III. The performance of our framework is evaluated in Section IV. Finally, We conclude the paper in Section V.

II. TAG Construction Using Term Spectra

A. Term Signal

First, we introduce the common document feature extraction procedures. The preprocessing starts by separating the main text contents from Web documents, for example, HTML formatted documents. We then extract words from all the documents in a data set 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 term. 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 (the frequency of the u-th word in all documents), and the document frequency (*df*), 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 *df*, for simplicity we use the well-known *tf-idf* term-weighting measure to calculate the weight of each word

$$h_u = f_u \cdot idf, \tag{1}$$

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 data set. It is noted that this term-weighting measure can be replaced by other feature selection criterion [22]. 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 [18][23], using all the words in a data set to construct the vocabulary is not necessarily expected to deliver the improvement of performance because some words may be noisy features for some topics.

In [14][15][16], Park *et al*. proposed to map each document into a set of term vectors by grouping terms into bins. The number of bins, B, was predefined, and each document was averagely partitioned into B parts. Rather than using bins, here we use paragraphs to group terms and represent their spatial information. It is worth pointing out that the use of paragraphs, instead of the fixed number of bins in [14][15][16], is more reasonable to follow the semantic flow of a document. Given a document, if the number of paragraphs, B, is 7, (i.e. $B = 7$), a term signal is given by [1 0 0 2 0 1 0], which indicates that the term appears once in the first paragraph, twice in the fourth paragraph and once in the sixth paragraph. However, in practice we usually do not use just the term counts to calculate the document similarity. The similarity measure can be improved considerably by adding weighting to the document vectors [15][24]. A few of the promising weighting schemes are BD-ACI-BCA, AB-AFD-BAA, and BI-ACI-BCA in [25], and the Lnu.ltu (SMART) method in [26]. These techniques have been extensively examined in information retrieval [14][15][16][25][26]. For text retrieval, each of these schemes includes query weighting and document weighting. But, for classification, we need to combine the weighting in advance. The pre-weighting schemes investigated in this paper are in the form of BD-ACI-BCA:

$$w_{u,j,b} = \left(\frac{1 + \log(f_{u,j,b})}{(1-s) + sW_j/\bar{W}_j} \right) \log\left(1 + f_u^m/f_u^d\right), \tag{2}$$

AB-AFD-BAA (Okapi):

$$w_{u,j,b} = \left(\frac{f_{u,j,b}}{f_{u,j,b} + \tau_j / \bar{\tau}_j} \right) \log \left(1 + n / f_u^d \right), \quad (3)$$

BI-ACI-BCA:

$$w_{u,j,b} = \left(\frac{1 + \log(f_{u,j,b})}{(1-s) + s W_j / \bar{W}_j} \right) \left(1 - \frac{n_u}{\log_2(n)} \right), \quad (4)$$

Lnu.ltu (SMART):

$$w_{u,j,b} = \left(\frac{(1 + \log(f_{u,j,b}))/(1 + \log(\bar{f}_{u,j,b}))}{(1-s) + s \tau_j / \bar{\tau}_j} \right) \log \left(n / f_u^d \right), \quad (5)$$

where $f_{u,j,b}$ is the term frequency of the u-th word in the b-th paragraph associated with the j-th document, f_u^d is the document frequency of term u, f_u^m is the largest f_u^d for all u, W_j is the l_2 norm of the j-th document vector, $\bar{W}_j$ is the average W_j in the entire data set, τ_j and $\bar{\tau}_j$ are the number of unique terms in document j and the average unique terms, respectively, s is a slope parameter (set to 0.7 [16][25]), and n_u is a noise measure of term u [25][27].

B. Term Affinity Graph (TAG)

As we stated before, the major drawback of the traditional modeling methods such as PCA and LSI is that they lack the description of term associations and spatial distribution information over the reduced semantic space. Here we propose a new document representation that contains this description. Given a document, our goal is to build a word affinity graph such that the term associations can be illustrated. Consider a graph denoted by a matrix $G_j \in R^{m \times m}$, in which each element $g_{uv,j}$ $(u, v = 1, 2, ..., m)$ is defined by

$$g_{uv,j} = \begin{cases} c_{uv,j}, & u \neq v \\ w_{u,j}, & u = v \end{cases}, \quad (6)$$

where $w_{u,j}$ is the weighted tf associated with the term u in the j-th document, and it can be obtained by setting $B = 1$ in Eqs.(2)-(5), and $c_{uv,j}$ indicates the association of terms u and v. Note that if we do not consider term associations in a document, i.e. let $g_{uv,j} = 0$ (for $u \neq v$), the affinity graph G_j becomes a diagonal matrix with the elements corresponding to the traditional VSM. By definition, the graph G_j is a symmetric matrix. This graph contains the more semantic information of a document in a way that we can design an efficient representation including *tf* and term inter-connections in a unified framework. But the challenge is finding a way to measure the connection between two terms, i.e. the calculation of $c_{uv,j}$ in Eq.(6). On this regard, we investigate three signal processing techniques (i.e. DCF, DFT and DWT) inspired by the text retrieval system [14][15][16]. Each component $c_{uv,j}$ in a TAG is assigned by the distributional association score of term u and v. This score is measured by different discrete transforms (i.e. DCF [14], DFT [15], and DWT [16]). For clarity, Fig.1 gives us an example for describing the problem. The detailed procedures of the discrete transforms of terms signals and the combination of term spectra can be found in [14], [15] and [16], respectively.

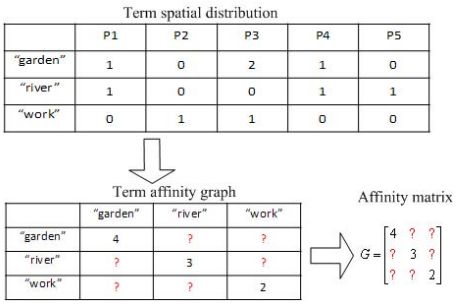

Fig. 1. An example of term spatial distribution. The top table shows the term spatial information distributed over five paragraphs. Here we assume three words, i.e. "garden", "river", and "work", are selected, and the document is partitioned into five paragraphs. The second table shows the term affinity graph transmitted from the first table. The diagonal elements represent the *tf* in this document, and the off-diagonal elements represent the term associations that we need to explore. Note that here we do not consider any weighting scheme with respect to the *df*.

C. Direct Graph Matching (DGM)

Once we have obtained the TAG for each document, where each column (or row) indicates the association scores between the indexed term and other words (including that term), we can compare two documents by calculating the similarity (or distance) between their TAGs. The similarity measure is given by

$$S_{p,q}^{DGM} = \frac{1}{m} \sum_{k=1}^{m} \exp \left(-1 + \frac{G_p(\cdot,k) \cdot G_q(\cdot,k)}{||G_p(\cdot,k)||_2 ||G_q(\cdot,k)||_2} \right), \quad (7)$$

where $G_p(\cdot,k)$ denotes the k-th column of the TAG of document p, $G_q(\cdot,k)$ denotes the k-th column of the TAG of document q, m is the vocabulary size. If $||G_p(\cdot,k)||_2 ||G_q(\cdot,k)||_2 = 0$, we set the second item, i.e. $\frac{G_p(\cdot,k) \cdot G_q(\cdot,k)}{||G_p(\cdot,k)||_2 ||G_q(\cdot,k)||_2}$, to zero. This similarity measure can be directly used to evaluate the relevance of two documents.

III. MDLSA

We can apply the way of DGM(see Section II-C) to calculate the similarity(or distance) between documents. However, the vocabulary size m is usually large. The calculation of the similarity between two documents requires m^2 operations, which increases the computational burden significantly. Besides, we note that the TAG is a sparse matrix. This graph representation contains a large quantity of noises, which spread out the original term distributional space. These noises result in the degradation of document comparison performance. Therefore, it is important to design an efficient dimensionality reduction technique, to compress the graph in a principled manner, and to model an accurate representation in a low dimensional space.

This section presents a new model, MDLSA, which considers word affinity graphs and maps them onto a low dimensional latent semantic space. First, we briefly overview the 2DPCA model that is related to the MDLSA. Second, the detailed MDLSA algorithm is presented. Third, a similarity measure is designed for between-document comparison. This part has been partially reported in our recent work [20].

A. 2DPCA

2DPCA [21] is a two-dimensional extension of the classical PCA, and it is developed in particular for face images. One promising property of the 2DPCA is the use of 2D image matrices as feature inputs, instead of using 1D stacked vectors transformed from the image matrices. If we treat the TAGs, associated with documents, as image matrices, the 2DPCA can be directly used for semantic analysis.

Given a TAG G of size $m \times m$, the goal of 2DPCA is to produce a projection $\hat{Z}$ of size $m \times d$ ($d \ll m$). In linear algebra, the projection $\hat{Z}$ can be obtained by

$$\hat{Z} = GV, \tag{8}$$

where V is a $m \times d$ linear transformation matrix. The problem comes to finding an optimal transformation V for this dimensionality reduction.

Let $\{G_1, G_2, ..., G_n\}$ be a set of training documents. By representing the TAG G_j associated with the j-th document, the graph covariance (or scatter) matrix C can be written by

$$C = \frac{1}{n}\sum_{j=1}^{n}(G_j - \bar{G})^T(G_j - \bar{G}), \tag{9}$$

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

$$J(v) = v^T C v, \tag{10}$$

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 axes, $v_1, v_2, ..., v_d$, subject to the orthogonal constraints and maximizing the criterion $J(V)$ by the form

$$\begin{aligned} &\{v_1, v_2, ..., v_d\} = \arg\max_{v} J(v), \\ &\text{subject to } \; v_\rho^T v_l = 0 (\rho \neq l, \rho, l = 1, 2, ..., d) \end{aligned}. \tag{11}$$

According to linear algebra, the optimal mapping axes, $v_1, v_2, ..., v_d$, are the orthogonal eigenvectors of C associated with the first largest d eigenvalues. If we denote these mapping axes by $V = [v_1, v_2, ..., v_d]$, the projection $\hat{Z}$ of a TAG G will be accomplished easily by the product of the resulting matrices as shown in Eq.(8). Note that this projection $\hat{Z}$ is a compact matrix, but it is still a rectangle matrix of large size, i.e. $m \times d$. Although $d \ll m$, m is the vocabulary size, which is usually large. Due to the compactness of $\hat{Z}$, it requires large storage space. On the other hand, given two documents p and q, if we compare their projections $\hat{Z}_p$ and $\hat{Z}_q$ based on the similarity measure

$$S_{p,q}^{2DPCA} = \frac{1}{d}\sum_{k=1}^{d}\exp\left(-1 + \frac{\hat{Z}_p(\cdot,k)\cdot\hat{Z}_q(\cdot,k)}{||\hat{Z}_p(\cdot,k)||_2||\hat{Z}_q(\cdot,k)||_2}\right), \tag{12}$$

the calculation of this similarity requires md operations. Moreover, the reduced semantic space may still contain a certain amount of noises resulting in a degradation of performance.

B. MDLSA

In line with the 2DPCA, we may further reduce the dimensionality of the TAGs such that the computational burden will be reduced while getting rid of noises. The proposed MDLSA model is just a typical approach of this. Given a TAG G of size $m \times m$, the objective of the MDLSA is to produce a projection Z of size $d \times d$ ($d \ll m$) resided in a low dimensional semantic space. The projection Z can be obtained by

$$Z = V^T G V. \tag{13}$$

Motivated by [28], we can regard the resulting projection Z as the result by conducting 2DPCA twice on the TAG G: one is from the row direction, and the other is from the column direction. However, we can conduct 2DPCA only once because of the symmetry of G. Thus, we use the same technique as illustrated in the last section to acquire the optimal transformation matrix V. Once V is obtained, the projection Z will be acquired easily according to Eq.(13). This projection Z has two features, i.e. compactness and small size dimension, which are very desirable for between-document comparison.

We clarify that the difference between the MDLSA and the 2DPCA is that the 2DPCA implements the projection $\hat{Z}$ of an image by multiplying the transformation matrix V only on the right side of the original image matrix (see Eq.(8)), whilst MDLSA achieves the projection Z of a document by multiplying V on the both sides of the TAG. Therefore, the MDLSA can be regarded as an extension of the 2DPCA in particular for texts.

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

Algorithm 1: *MDLSA*

Input: A training set, given the TAGs $\{G_1, G_2, ..., G_n\}$, the dimension of the reduced space d.

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

1) Input the TAGs $\{G_1, G_2, ..., G_n\}$, the dimension of the reduced space d.
2) Solve the eigenvalue problem as shown in Eq.(11), and construct the mapping V, the column vectors of which are taken from the eigenvectors associated with the d largest eigenvalues.
3) Calculate the projected graphs $Z_j = V^T G_j V$ using Eq.(13) to represent the j-th training sample.
4) Given the TAG G_t associated with a new testing document, execute Step 3), map it onto the subspace and achieve the latent semantic expression Z_t.

C. Similarity Measure

Many document applications rely on the calculation of similarity between two documents. In this paper, we have

TABLE I

DETAILS OF THE TESTED DATA SETS

Statistics	YahooScience	WebKB4	Dmoz
Class	6	4	12
Number of Documents	861	4171	4515
Maximal Number of Words in Each Document	36318	57267	61122
Average Number of Words in Each Document	913	290	959
Maximal Number of Paragraphs	427	529	872
Average Number of Paragraphs	10.96	4.17	11.67

extracted a set of features to construct a TAG G for each document. We then use the MDLSA to map these features onto the semantic space $Z \in R^{d\times d}$, which is of low dimension. For document comparison, we define the similarity measure in the form

$$S_{p,q} = \frac{1}{d}\sum_{k=1}^{d} \exp\left(-1 + \frac{Z_p(1:k,k)\cdot Z_q(1:k,k)}{||Z_p(1:k,k)||_2||Z_q(1:k,k)||_2}\right), \quad (14)$$

where $Z_p(1:k,k)$ denotes the first k elements of the k-th column in matrix Z_p for document p and $Z_q(1:k,k)$ denotes the first k elements of the k-th column in matrix Z_q for document q. By definition, we compare only the top diagonal elements of matrices Z_p and Z_q. The reason for this is that matrices Z_p and Z_q are symmetric. Note that the calculation of $S_{p,q}$ requires $d(d+1)/2$ operations. The computational burden has been significantly reduced, when comparing to the DGM method with m^2 operations, because $d \ll m$.

IV. EXPERIMENTS

As an application example, we evaluate the performance of our framework on web document classification, because it has become important in organizing the massive amount of online data. We have used a Nearest Neighbor (NN) classifier to perform this task based on the latent semantic features in the reduced space. To evaluate the quality of the classification, we adopted three measures which are widely used in the text classification and clustering literature [29]: Accuracy, F-measure and Entropy. All the experiments were performed on a PC with Intel(R) Core(TM) i7 CPU 860@ 2.80GHz and 6.00GB memory. The feature extraction programs were written in Java programming language. The classification programs were tested on MATLAB 7.5.0 (R2007b).

A. Comparative Study

Our method, MDLSA, has been examined in three public data sets: 1) YahooScience, 2) WebKB4, and 3) Dmoz. These data sets consist of HTML documents, and are firstly partitioned them into paragraphs. We have conducted extensive comparisons of the new method with several other methods including VSM [1], PCA, RAP [9], MLM [17][18] and DGM.

The VSM, regarded as a baseline method, is investigated without dimensionality reduction. The LSI and PCA are most widely used methods, and they consider only *tf* features with the same pre-weighting schemes. It is noted that the LSI is very similar to the PCA, and they both rely on the eigen-semantics. The LSI is usually used for document retrieval. Here, we only compare our method to the PCA for document classification. The RAP model is selected for comparison because it is a new statistical method and it has shown superior performance over LDA [7] and PLSI [6]. Note that the RAP model also considers only *tf* features. The MLM approach is our latest work, which has been used for document retrieval and plagiarism detection by integrating term spatial information. The DGM 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.

TABLE II

RESULTS OF DIFFERENT METHODS FOR YAHOOSCIENCE USING NN

Method	Accuracy (%)	F-measure	Entropy
MDLSA-dwt-w.3-d.100	93.02	0.9298	0.3100
PCA-w.4-d.20	91.40	0.9137	0.3833
MDLSA-dwt-w.4-d.70	90.70	0.9064	0.4055
MDLSA-dct-w.1-d.270	90.70	0.9068	0.4230
MDLSA-dwt-w.1-d.220	90.00	0.9001	0.4501
MDLSA-dwt-w.2-d.110	89.77	0.8970	0.4297
PCA-w.3-d.20	89.77	0.8971	0.4470
MLM-d.100	89.76	0.8976	0.4335
MDLSA-dft-w.4-d.260	89.53	0.8961	0.4316
MDLSA-dct-w.4-d.220	89.30	0.8921	0.4612
MDLSA-dct-w.3-d.270	89.07	0.8908	0.4748
PCA-w.1-d.30	89.07	0.8907	0.4618
MDLSA-dft-w.3-d.300	88.60	0.8865	0.4644
MDLSA-dft-w.1-d.290	87.91	0.8790	0.4974
MDLSA-dct-w.2-d.210	87.67	0.8756	0.5044
VSM-w.3	87.44	0.8748	0.5322
VSM-w.4	86.74	0.8679	0.5607
MDLSA-dft-w.2-d.280	86.51	0.8654	0.5562
PCA-w.2-d.120	86.28	0.8624	0.5402
VSM-w.1	85.81	0.8588	0.5874
VSM-w.2	84.65	0.8468	0.6167
DGM-dwt-w.3	81.16	0.8199	0.6387
DGM-dwt-w.1	79.77	0.8078	0.6685
DGM-dwt-w.4	78.37	0.7950	0.6823
DGM-dwt-w.2	76.98	0.7819	0.7148
RAP-d.230	76.28	0.7635	0.8498
DGM-dft-w.3	62.09	0.6440	1.0134
DGM-dft-w.1	61.40	0.6439	1.0028
DGM-dct-w.3	61.16	0.6369	1.0229
DGM-dft-w.4	60.93	0.6386	1.0089
DGM-dct-w.1	60.47	0.6357	1.0191
DGM-dct-w.4	60.23	0.6336	1.0183
DGM-dft-w.2	59.07	0.6223	1.0379
DGM-dct-w.2	57.67	0.6095	1.0690

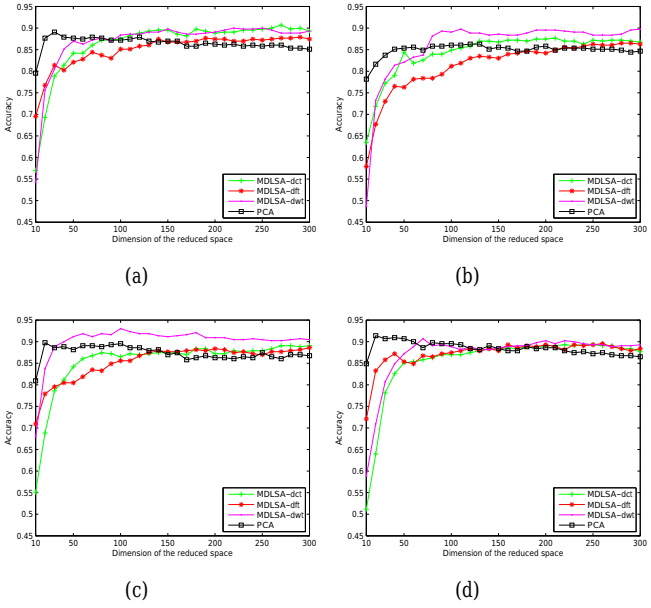

Fig. 2. Accuracy against dimension of reduced space for YahooScience Using the NN classifier with: (a) BD-ACI-BCA pre-weighting; (b) AB-AFD-BAA pre-weighting; (c) BI-ACI-BCA pre-weighting; (d) Lnu.ltu (SMART) pre-weighting

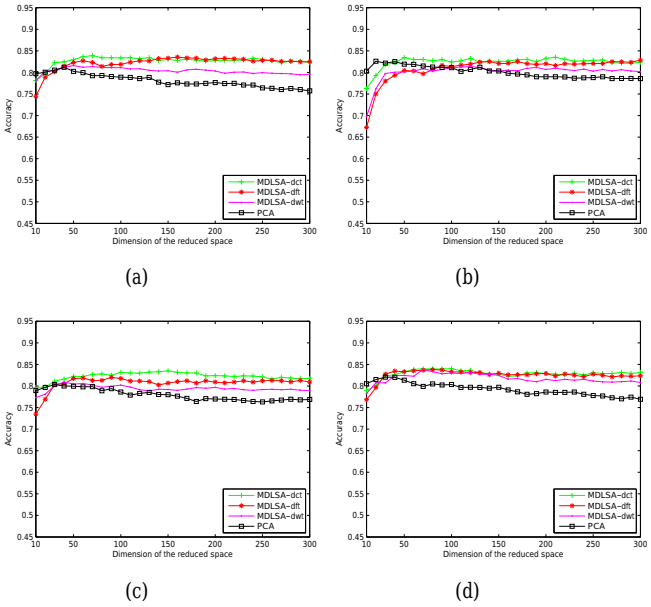

Fig. 3. Accuracy against dimension of reduced space for WebKB4 using the NN classifier with: (a) BD-ACI-BCA pre-weighting; (b) AB-AFD-BAA pre-weighting; (c) BI-ACI-BCA pre-weighting; (d) Lnu.ltu (SMART) pre-weighting

All the data sets were split into 50% testing and 50% training data, and we selected the first 3000 words as the vocabulary. For a fair comparison, we reported the results of dimensionality reduction models with their optimal dimension d in the reduced space. We also showed the effect of the dimension of reduced space on the results (see Section IV-B).

1) YahooScience: YahooScience is filed from the documents referenced the Open Directory Project, and it is publicly available[1]. The original collection of YahooScience included 907 documents in 6 top-level classes. For each top-level class, we firstly moved all the documents in its sub-class to the top-level class and removed all the sub-classes. We then removed all empty documents and the documents containing only scripts. 861 documents were left out with YahooScience in 6 classes. The average number of words in one document is around 900. The details of this data set can be found in Table I.

We summarized the average results[2] of different methods using the NN classifier in Table II, for comparison. From Table II, it is clear to observe that MDLSA significantly outperform the DGM methods, and it produces over 10% accuracy gain in contrast to DGM. This result indicates the necessity of the dimensionality reduction to compress the TAG and get rid of the impact of noise to the similarity measure. MDLSA-dwt-w.3-d.100, which means MDLSA used discrete wavelet transform, BI-ACI-BCA weighting and the optimal dimension of reduced space residing at 100, delivers the best performance over other methods. It is interesting to see that PCA-w.4-d.20 produces promising results as well. This suggests that PCA with appropriate pre-weighting may outperform statistical methods such as PLSI, LDA and RAP, because currently these methods cannot utilize the pre-weighting schemes to boost their performance. Moreover, it is noted that the optimal semantics of PCA usually resides in a lower dimension, whilst MDLSA requires a higher dimension compared to it. This is reasonable, since MDLSA considers more semantics (including term frequencies and term associations) while PCA takes only term frequencies into account. From intuition, MDLSA requires a larger dimension to allocate the semantics in order to obtain an accurate representation in the reduce space.

TABLE III
RESULTS OF DIFFERENT METHODS FOR WEBKB4 USING NN

Method	Accuracy (%)	F-measure	Entropy
MDLSA-dct-w.4-d.100	84.03	0.8388	0.5639
MDLSA-dwt-w.4-d.70	83.88	0.8379	0.5654
MDLSA-dct-w.1-d.70	83.88	0.8386	0.5654
MDLSA-dft-w.4-d.80	83.79	0.8373	0.5648
MDLSA-dft-w.1-d.160	83.55	0.8341	0.5777
MDLSA-dct-w.2-d.210	83.50	0.8330	0.5851
MDLSA-dct-w.3-d.150	83.50	0.8340	0.5801
MDLSA-dft-w.2-d.300	82.83	0.8266	0.6020
PCA-w.2-d.20	82.59	0.8258	0.5848
PCA-w.4-d.30	82.06	0.8195	0.5975
MDLSA-dft-w.3-d.90	82.01	0.8192	0.6123
MDLSA-dwt-w.1-d.50	81.58	0.8142	0.6181
MDLSA-dwt-w.2-d.120	81.34	0.8113	0.6376
PCA-w.1-d.40	81.20	0.8094	0.6287
MDLSA-dwt-w.3-d.40	80.91	0.8083	0.6360
PCA-w.3-d.20	80.34	0.6362	0.8115
VSM-w.2	75.25	0.7497	0.7829
VSM-w.3	72.90	0.7276	0.8133
RAP-d.160	72.71	0.7244	0.8378
VSM-w.4	72.61	0.7241	0.8252
VSM-w.1	71.85	0.7163	0.8431
MLM-d.100	69.69	0.6899	0.9059

2) WebKB4: To demonstrate the performance of our proposed method, we experiment on WebKB4, another publicly available data set[3]. WebKB4 is a subset of the WebKB data set, and it contains 4199 Web pages in 4 categories collected from university computer science departments. We then removed all empty documents and the documents containing only scripts. 4177 documents were left with WebKB4 in 4 classes. Average document size in this data set is short. The average number of words in one document is around 290. The details of this data set can be found in Table I.

The comparative results of different methods are shown in Table III. Results in Table III show that MDLSA-dct-w.4-d.100 delivers the best performance over other methods. MDLSA, with the use of the Lnu.ltu (SMART) pre-weighting and the DWT, performs also well. The top eight methods are all MDLSA related, and the results show that different weighting schemes have only a sight impact on the performance of MDLSA.

3) Dmoz: Dmoz is also filed from the documents referenced the Open Directory Project on health conditions and diseases, and it can be downloaded in the same Website with YahooScience. The original collection of Dmoz included

[1]http://www.di.uniba.it/∼malerba/software/webclass/WebClassIII.htm.

[2]Note for description in method names: 1) "-dct" implies use of discrete cosine transform, "-dft" implies use of discrete Fourier transform, and "-dwt" implies use of discrete wavelet transform; 2) "-w.1" implies use of BD-ACI-BCA pre-weighting, "-w.2" implies use of AB-AFD-BAA pre-weighting, "-w.3" implies use of BI-ACI-BCA pre-weighting, and "-w.4" implies use of Lnu.ltu (SMART) pre-weighting; 3) "-d.*" implies the optimal dimension of reduced space, the dimension size d varies from 10 to 300 at an increment of 10.

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

TABLE IV

RESULTS OF DIFFERENT METHODS FOR DMOZ USING NN

Method	Accuracy (%)	F-measure	Entropy
MDLSA-dwt-w.3-d.100	84.05	0.8405	0.7248
MDLSA-dwt-w.2-d.280	81.75	0.8174	0.8068
MDLSA-dwt-w.1-d.140	79.97	0.8003	0.8800
PCA-w.3-d.20	79.93	0.7994	0.8712
MDLSA-dft-w.3-d.110	79.57	0.7958	0.8727
MDLSA-dwt-w.4-d.200	79.35	0.7940	0.8930
MLM-d.100	78.20	0.7817	0.9493
MDLSA-dct-w.3-d.170	78.02	0.7800	0.9401
MDLSA-dft-w.1-d.280	77.76	0.7771	0.9550
MDLSA-dft-w.4-d.170	77.40	0.7735	0.9730
PCA-w.1-d.130	76.78	0.7676	0.9813
PCA-w.4-d.70	76.43	0.7641	0.9818
RAP-d.230	75.85	0.7595	0.9971
MDLSA-dct-w.1-d.170	75.63	0.7556	1.0340
MDLSA-dct-w.4-d.210	74.52	0.7449	1.0652
MDLSA-dft-w.2-d.200	74.44	0.7437	1.0476
VSM-w.3	74.44	0.7436	1.0683
MDLSA-dct-w.2-d.230	72.62	0.7252	1.1218
VSM-w.4	69.96	0.6986	1.2028
VSM-w.1	69.21	0.6908	1.2208
PCA-w.2-d.120	68.85	0.6875	1.2395
VSM-w.2	64.42	0.6434	1.3668

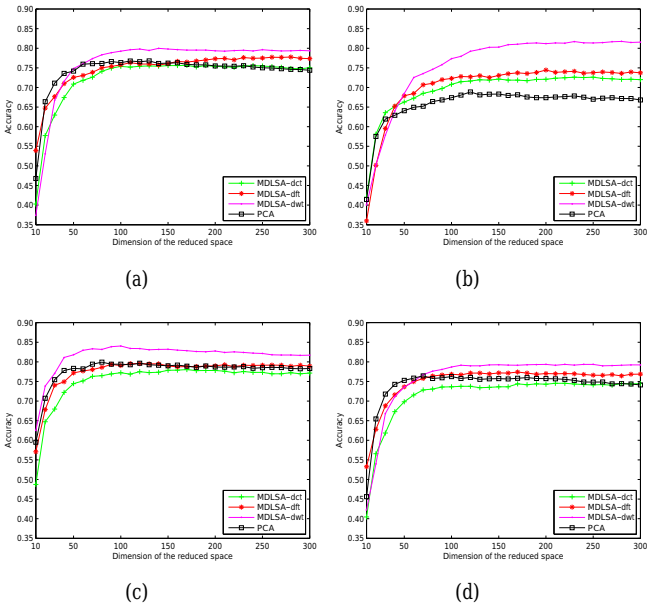

Fig. 4. Accuracy against dimension of reduced space for Dmoz using the NN classifier with: (a) BD-ACI-BCA pre-weighting; (b) AB-AFD-BAA pre-weighting; (c) BI-ACI-BCA pre-weighting; (d) Lnu.ltu (SMART) pre-weighting

5916 documents in 21 top-level classes. For each top-level class, we first moved all the documents in its sub-class to the top-level class and removed all the sub-classes. We used the largest 12 classes in our experiment. 4515 documents were left, and each class contained more than 150 files. The average number of words in one document is around 959. The details of this data set can be found in Table I.

The comparative results are listed in Table IV. With the use of the NN classifier as shown in Table IV, the best result is produced by MDLSA-dwt-w.3-d.100. It is capable of enhancing the accuracy with over 4% and the entropy with over 15% compared to traditional methods such as PCA, MLM and RAP. It is also noticed that PCA with the use of the AB-AFD-BAA pre-weights performs worse than them using other pre-weights.

TABLE V

COMPUTATIONAL OPERATIONS OF DIFFERENT METHODS FOR COMPARING TWO DOCUMENTS

	DGM	MDLSA	MLM	RAP	PCA	VSM
Operations	m^2	$d(d+1)/2$	$dN_p^3 \log(N_p)$	d	d	m

B. Dimension of Reduced Space

In order to show the impact of the dimension of reduced space on the results, we plotted the accuracy results against different dimensions of the reduced space for MDLSA and PCA. The results are visually illustrated in Figs.2-4 for different data sets. We have examined the dimensions of reduced space varying from 10 to 300 at an increment of 10. From Fig.2, we observe that the accuracy results produced by the MDLSA related methods using the NN classifier increase sharply when the dimension of reduced space varies from 10 to around 70. After reaching an optimal dimension, the result becomes very stable. This is a promising property compared with PCA, because the probability of the optimal dimension being selected will be very high if we do not have prior information for the option of the projection dimension d (see Section III-A) before conducting dimensionality reduction. For the WebKB4 data set with the NN classifier (see Fig.3), it is observed that MDLSA with the DCT consistently outperforms PCA when the dimension d varies from 50 to 300. According to observation drawn from Fig.4, MDLSA with the DWT delivers the best results compared to other methods, when increasing the dimension of reduce space from around 100 to 300.

C. Computational Time

MDLSA is a dimensionality reduction method for documents in essence. The resulting representations of MDLSA and DGM are matrices, while traditional methods such as PCA, RAP and VSM rely on the resulting vectors to represent documents. We summarize the number of operations that all the methods require for comparing two documents in Table V, where d is the dimension of the reduced space, N_p represents the number of paragraphs in a document, and m is the vocabulary size (see Section III-C). In MLM[17][18], the number of operations is based on the case that two documents have the same number of paragraphs. There is no explicit expression for the case that two documents have different numbers of paragraphs. It is observed that MDLSA requires $(d+1)/2$ times more operations than PCA and RAP, but it reduces the computational burden significantly in comparison to the DGM method that requires m^2 operations, because $d \ll m$. In fact, the time cost of MDLSA relates directly to the value of d, the dimension of the reduced space. On this concern, we experiment empirically on the YahooScience set and visually illustrate the average time cost of MDLSA and PCA for comparing two documents in Fig.5. It is clear to observe that the time cost required by MDLSA increases approximately linearly with the increase of the value of d.

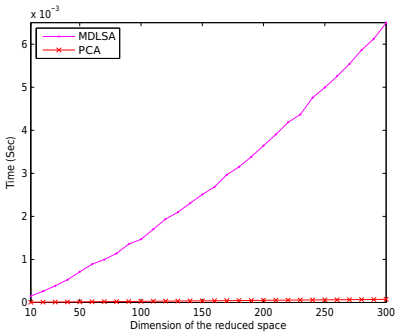

Fig. 5. Time performance against dimension of the reduced space when comparing two documents

V. Conclusion

In this paper, we introduced a new semantic analysis framework. It enabled us to formulate an accurate document representation by considering both term frequencies and term associations in a unified manner. Term signals were extracted from documents at the first place. We then investigated various spectral transforms including the DWT, the DFT and the DCT on term signals. A word affinity graph was constructed by the term spectra. We employed a new model, MDLSA[20], to project the affinity graph onto a low dimensional space. We experimented on three public data sets. The results strongly suggest that the proposed technique is accurate and computationally efficient, in particular for the data set with long documents. In the future work, it will be more interesting to apply our method to the analysis of very lengthy documents, e.g. electronic books.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grant no. 61300209, the Shenzhen Foundation Research Fund under Grant no. JCY20120613115205826 and the Shenzhen Technology Innovation Program under Grant no. CXZZ20130319100919673.

References

[1] G. Salton and M. McGill, editors. Introduction to modern information retrieval. McGraw-Hill, 1983.
[2] S. Deerwester, S. Dumais. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 1990, 41(6): 391-407.
[3] Kari Torkkola. Linear discriminant analysis in document classification. IEEE ICDM Workshop on Text Mining, 2001.
[4] E. Kokiopoulou and Yousef Saad. Orthogonal neighborhood preserving projections: a projection-based dimensionality reduction technique. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, vol. 29, no. 12, pp. 2143-2156.
[5] Deng Cai, Xiaofei He, and Jiawei Han. Document clustering using locality preserving indexing. IEEE Transactions on Knowledge and Data Engineering, 2005, vol. 17, no. 12, pp. 1624-1637.
[6] T. Hofmann. Probabilistic latent semantic indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference, 1999.
[7] D. Blei, A. Ng and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993-1022.
[8] M. Welling, M. Rosen-Zvi and G. Hinton. Exponential family harmoniums with an application to information retrieval. In Advances in Neural Information Processing Systems 17, Cambridge, MA, MIT Press, 2004, 1481-1488.
[9] P. Gehler, A. Holub and M. Welling. The rate adapting Poisson model for information retrieval and object recognition. In Proc. of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006.
[10] A. Schenker, M. Last, H. Bunke and A. Kandel. Classification of web documents using graph matching. International Journal of Pattern Recognition and Artificial Intelligence, 2004, vol. 18, no. 3, pp. 475-496.
[11] M. Fuketa, S. Lee, T. Tsuji, M. Okada and J. Aoe. A document classification method by using field association words. Information Sciences, 2000, vol. 126, no. 1-4, pp. 57-70.
[12] C. M. Tan, Y. F. Wang, and C. D. Lee. The use of bigrams to enhance text categorization. Information Processing & Management, 2002, vol. 38, no. 4, pp. 529-546.
[13] M. L. Antonie, O.R. Zaiane. Text document categorization by term association. In Proc. IEEE International Conference on Data Mining (ICDM2002), 2002, pp. 19-26.
[14] L. A. F. Park, M. Palaniswami and K. Ramamohanarao. A novel document ranking method using the discrete cosine transform. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, vol. 27, no. 1, pp. 130-135.
[15] L. A. F. Park, K. Ramamohanarao and M. Palaniswami. Fourier domain scoring: A novel document ranking method. IEEE Transactions on Knowledge and Data Engineering, 2004, vol. 16, no. 5, pp. 529-539.
[16] L. A. F. Park, K. Ramamohanarao and M. Palaniswami. A novel document retrieval method using the discrete wavelet transform. ACM Transactions on Information Systems, 2005, vol. 23, no. 3, pp. 267-298.
[17] Haijun Zhang and Tommy W. S. Chow. A Multi-level Matching Method with Hybrid Similarity for Document Retrieval. Expert Systems With Applications, 2012, vol. 29, no.3, pp. 2710-2719.
[18] Haijun Zhang and Tommy W. S. Chow. A coarse-to-fine framework to efficiently thwart plagiarism. Pattern Recognition, 2011, vol. 44, no. 2, pp. 471-487.
[19] Y. Rubner, C. Tomasi, and L.J. Guibas. The Earth Mover's Distance as a metric for image retrieval. International Journal of Computer Vision, 2000, vol. 40, no.2, pp. 99-121.
[20] Haijun Zhang, John K. L. Ho, and Jonathan Q. M. Wu. Multi-Dimensional Latent Semantic Analysis Using Term Spatial Information. IEEE Transactions on Cybernetics, 2013, vol. 43, no.6, pp. 1625-1640.
[21] J. Yang, D. Zhang, A. F. Frangi, and Jing-yu Yang. Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, vol. 26, no. 1, pp. 131-137.
[22] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. International Workshop on Machine Learning, 1997.
[23] Tommy W. S. Chow, Haijun Zhang, and M. K. M. Rahman. A new document representation using term frequency and vectorized graph connectionists with application to document retrieval. Expert Systems With Applications, 2009, 36: 12023-12035.
[24] S. T. Dumais. Improving the Retrieval of Information from External Sources. Behaviour Research Methods, Instruments & Computers, 1991, vol. 23, no. 2, pp. 229-236.
[25] J. Zobel and A. Moffat. Exploring the similarity space. ACM SIGIR Forum, 1998, vol. 32, no. 1, pp. 18-34.
[26] C. Buckley, A. Singhal, M. Mitra, and G. Salton. New Retrieval Approaches Using Smart: TREC 4. Proc. Fourth Text Retrieval Conf., Nov. 1995, pp. 25-48.
[27] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. Information Processing & Management, 1988, vol. 24, no. 5, pp. 513-523.
[28] W. Zuo, D. Zhang, and K. Wang. Bidirectional PCA with assembled matrix distance metric for image recognition. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 2006, vol. 36, no. 4, pp. 863-872.
[29] Khaled M. Hammouda and Mohamed S. Kamel. Efficient Phrase-Based Document Indexing for Web Document Clustering. IEEE Transactions on Knowledge and Data Engineering, 2004, vol. 16, no. 10, pp. 1279-1296.