ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

IJIRSET

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 International Conference on Innovations in Engineering and Technology (ICIET'14)
On 21st & 22nd March Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

Classification of Documents in E-Learning Using Multidimensional Latent Semantic Analysis

R.Archana[1], M.Ravichandran[2]

[1]Department of Computer Science and Engineering, Arunai Engineering college, Tiruvanamalai, Tamilnadu, India

[1]Department of Computer Science and Engineering, Arunai Engineering college, Tiruvanamalai, Tamilnadu, India

ABSTRACT- In this paper we consider the problem of dimensionality reduction techniques. Two techniques such as Independent Component analysis (ICA) and multidimensional latent semantic analysis (MDLSA) are proposed. A new document analysis method named multidimensional latent semantic analysis (MDLSA) which resolves the problem of in-depth document analysis, mines local information from a document efficiently with respect to term associations and spatial distributions. The MDLSA first partitions each document into paragraphs and later builds a term "affinity" graph. Each element of this graph represents the frequency of term co-occurrence in a paragraph. We then use Independent Component Analysis (ICA) which finds a linear representation of nongaussian data such that the components are statistically independent. Thus these two techniques are examined in retrieving and classifying the e-learning documents. It is also proven by experimental verifications that the proposed technique outperforms current algorithms with respect to accuracy and computational efficiency.

INDEX TERMS - Independent component analysis, Multidimensional latent semantic analysis, affinity graph.

I. INTRODUCTION

E-learning is a new education concept by using the Internet technology, it deliveries the digital content, provides a learner-orient environment for the teachers and students. The e-learning promotes the construction of life-long learning opinions and learning society. Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies information need from within large collections (usually stored on computers). Information retrieval is fast becoming the dominant form of information access, overtaking traditional database style searching. Information retrieval (IR) deals with the representation, storage, organization of, and access to information items.

The rapid development of technology has made large amount of document data available and easy access to people, which increase the growing demand of higher accuracy and speed for document retrieval. A document is retrieved based on the user input query. A user query can range from whole document to many keywords. There is another method of retrieving documents by using the document itself as input. Many document retrieval system use term frequency as feature units to build statistical models and develop natural language processing (NLP) approaches for document retrieval. The connection between terms when overlooked results in losing important semantic information of documents. There is a growing demand for semantic representation which includes the term associations and spatial distributions.

The other demand is to find low-dimensional semantic expressions of documents. More features which are extracted from documents are modeled into a lower dimensional semantic space to exploit the information present in the documents and enhance the performance of relevant data mining. The exploitation of this information for the benefit of the user depends directly on the ease and accuracy of the retrieval process. Therefore, systems that automatically organize documents in terms of their content are cost- effective. Automated classification or categorization of natural language documents is an important research topic because it promises to reduce the cost and time associated with organizing large groups of documents related to diverse topics.

Since the number of documents being produced now is more than ever before especially when we consider the Internet with its massive amount of heterogeneous documents — manual classification and categorization can be extremely difficult. Thus we have the need for accurate, automated systems that can assign categories to documents based on their content. The potential of an in-depth document information, which conveys term associations and spatial distributions

A new model is introduced 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 co-occurrence 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. I then conduct Independent Component Analysis (ICA) 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. Analysis is investigated by using term spatial information and dimensionality reduction techniques. 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 to know how to locate the low-dimensional space with the fusion of local information.

II. Related Work

Vector space model (VSM) model is one among the language models which is a statistical method of ranking of documents. This VSM model is most popular and widely used compared with others models. The frequency of each term is counted from vocabulary for a given document, where vocabulary is a list of terms/words used for feature description. This is referred to as term frequency (tf). VSM model also uses inverse document frequency (idf) which counts the number of documents where the term appears. Then the similarity between documents is computed using “cosine” distance or any other alternate distance functions. VSM model thus requires a lengthy vector, because the number of words involved is usually huge. This causes a significant increase in computational burden making the direct use of VSM model impractical for handling a large data set.

The limitations of VSM is overcame by developing latent semantic indexing (LSI), which compresses the feature vector of the VSM model into low dimension. The LSI 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. LSI is used to encode the semantic concept to some extent. Principal Component Analysis (PCA) is an alternate to LSI model, which is a classic linear technique that is able to project the high-dimensional term vectors to lower dimensional space, by finding the solution of an eigenvalue problem.

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. Independent Component Analysis is used to identify the Outer layer and Non outer layer values in the document, thus achieving accuracy. I 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.

Some of the drawbacks of 2DPCA are as follows: 2DPCA works on matrices (2D arrays) rather than on vectors (1D array). The size of the image covariance matrix of 2DPCA is much small. It needs to more coefficients for semantic representation than previous one-dimensional methods. It loses the covariance information between different local geometric structures in the semantic representation. These drawbacks lead us to move on to ICA method which is described as follows:

The problem of source separation is an inductive inference problem. There is not enough information to deduce the solution, so one must use any available information to infer the most probable solution. The aim is to process these observations in such a way that the original documents are extracted from the database system. Independent Component Analysis (ICA), a computationally efficient blind source separation technique, has been an area of interest for researchers for many practical applications in various fields of science and engineering. ICA is one of the most widely used BSS techniques for revealing hidden factors that underlie sets of random variables and measurements.

The general model for ICA is that the sources are generated through a linear basis transformation, where multiple documents can be present. Suppose we have N statistically independent documents, $si(t)$, $i = 1, ...,N$. We assume that the sources themselves cannot be directly observed and that each document, $si(t)$, is a realization of some fixed probability distribution at each time point t. Also, suppose we observe these documents using N terms, then we obtain a set of N observation documents $xi(t)$, $i = 1, ...,N$ that are mixtures of the sources. A fundamental aspect of the mixing process is that the terms must be spatially separated so that each term records a different mixture of the sources. With this spatial separation assumption in mind, we can model the mixing process with matrix multiplication as follows:

$$x(t) = As(t)$$

where A is an unknown matrix called the *mixing matrix* and x(t), s(t) are the two vectors representing the observed signals and source signals respectively. Incidentally, the justification for the description of this signal processing technique as *blind* is that we have no information on the mixing matrix, or even on the sources themselves. The objective is to recover the original documents, $si(t)$, from only the observed vector $xi(t)$. We obtain estimates for the sources by first obtaining the "unmixing matrix" W, where,

$$W = A-1.$$

This enables an estimate, ˆ $s(t)$, of the independent sources to be obtained:

$$\hat{} s(t) = Wx(t) \ (2)$$

The independent sources are mixed by the matrix A (which is unknown in this case). We seek to obtain a vector y that approximates by estimating the unmixing matrix W. If the estimate of the unmixing matrix is accurate, we obtain a good approximation of the sources. The above described ICA model is the simple model since it ignores all noise components and any time delay in the recordings.

III. System Design

In this section, we evaluate the performance of MDLSA on two document applications: retrieval and classification. We use two implementations: MDLSA and ICA.

A. Document Retrieval:

User must give their inputs as query to retrieve the documents. Here the user must give four terms as input to the system. The system will search for the term appearance in the list of documents present in the dataset. If any document is found to be relevant in the dataset, then that respective document will be retrieved. The important factor is that only the relevant and wanted data will be retrieved and the unwanted data or documents will be filtered out. The retrieved document will be dealt with the rest of the modules.

B. Computation of Weight:

The document which is retrieved in the above module is taken as input for the weight computation. The document must be partitioned into paragraphs by setting a threshold say it as 50 words. The frequency of words i.e the number of occurrence of that particular word in each paragraph must be calculated. Then we can the total number of occurrence of that particular word in a particular document. Later, Inverse Document Frequency (IDF) must be calculated by using the mathematical formulas. Once we get the values for frequency of word and IDF, we must compute the term weighting factor by using the product of frequency of words and the IDF. The frequency is represented as f_u^t and the Weight is calculated using the following equation:

$$w_u = f_u^t \cdot idf$$

C. Using MDLSA:

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 *{zi}* for training samples and *zt* 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 $\{G1,G2 , ...,Gn \}$.

3) Solve the eigenvalue problem and construct the mapping V whose column vectors are taken from the eigenvectors associated with the d largest eigen values.

4) Calculate the projected graphs ˜ $Zi = V\ T\ GiV$.

5) Select the first column of ˜ Zi to represent the ith training sample denoted as zi .

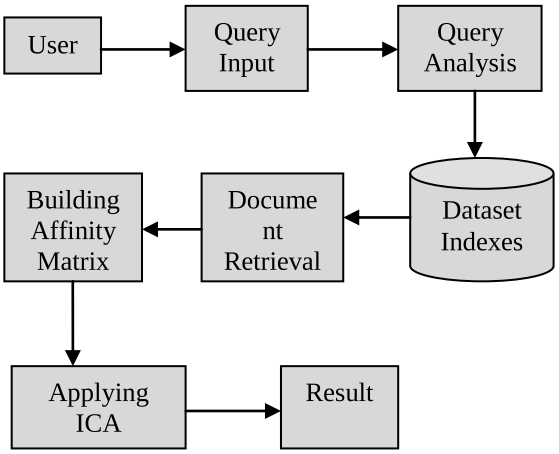
6) Given a new affinity graph Gt associated with a new testing document, repeat Steps 4 and 5, map it onto the subspace, and
achieve the latent semantic expression zt .

Thus the output of the MDLSA process is the affinity graph which has to be carried over to the next module.

D. *Applying ICA:*

Independent component analysis (ICA) is a well-known method of finding latent structure in data. ICA is a statistical method that expresses a set of multidimensional observations as a combination of unknown latent variables. The affinity graph is given as input to the ICA. ICA is the last Module. ICA is used for classification and it's used to identify the Out layer and Non Out layer.

F. Architecture Diagram



Results And Discussion

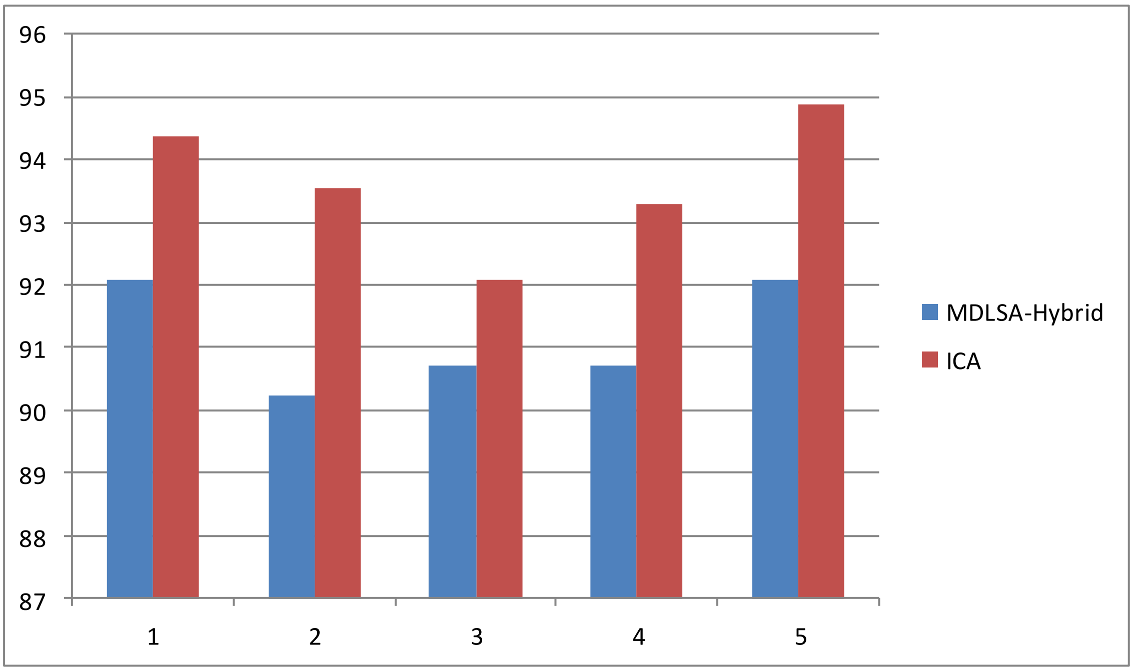
In this section, the performance of MDLSA is evaluated on two document applications: retrieval and classification. We conducted a large scale of experiments to show the retrieval performance of our proposed approach. First, we introduce the performance metrics used in this study. Second, we present the comparative results with respect to retrieval performance and query time performance. Third, we include the study on the influence of different parameters involved in the algorithms.

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

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

$$\textit{Recall} = \frac{\text{No. of correctly retrieved documents}}{\text{No. of documents in relevant category}}$$

Weighting	**MDLSA-Hybrid**			**ICA**			**Improvement**		
	Accuracy (%)	F-measure	Entropy	Accuracy (%)	F-measure	Entropy	Accuracy (%)	F-measure	Entropy
NORM	92.09	0.9210	0.3256	94.38	0.9440	0.1574	2.29	2.30	-0.1682
BD-ACI-BCA	90.23	0.9025	0.3718	93.55	0.9357	0.2514	3.32	3.32	-0.1204
AB-AFD-BAA	90.70	0.9063	0.3578	92.09	0.9210	0.1256	1.39	1.47	-0.2314
BI-ACI-BCA	90.70	0.9069	0.3514	93.28	0.9330	0.2776	2.58	2.61	-0.0738
SMART	92.09	0.9209	0.3146	94.89	0.9490	0.1387	2.80	2.81	-0.1759



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.

IV. Conclusion

A new document analysis method, MDLSA is presented, 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 co-occurrence in a paragraph. We then conduct a ICA to achieve an optimal semantic mapping. This analysis works by finding the unknown matrix known as the mixing matrix. We then obtain estimates for the sources by first obtaining the unmixing matrix. Using these matrices, ICA identifies the Out layer and Non Out layer

ontwo tasks such as retrieval and classification. The results strongly suggest that the proposed technique is accurate and computationally efficient for performing various document applications.

REFERENCES

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

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

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

[4] 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, Nov. 2009, pp. 670–680.

[5] 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, 2009, pp. 12023–12035.

[6] 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, Sep. 2009, pp. 1385–1402.

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

[8] H. Zhang, John K. L. Ho, Q. M. Jonathan Wu, and Yunming Ye, “Multidimensional Latent Semantic Analysis Using Term Spatial Information”, IEEE Trans. Cybern, vol. 43, no. 6, Dec.2013, pp. 1625-1640.

[9] 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, 2009, pp. 2950–2960.

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

[11] 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, Dec.2007, pp. 2143–2156