UDC 004.423, DOI: 10.2298/csis0902217K

Analysis of Unsupervised Dimensionality Reduction Techniques

Ch. Aswani Kumar

Networks and Information Security Division, School of Information Technology and Engineering, VIT University, Vellore-632014, India.
aswanis@gmail.com

Abstract. Domains such as text, images etc contain large amounts of redundancies and ambiguities among the attributes which result in considerable noise effects (i.e. the data is high dimension). Retrieving the data from high dimensional datasets is a big challenge. Dimensionality reduction techniques have been a successful avenue for automatically extracting the latent concepts by removing the noise and reducing the complexity in processing the high dimensional data. In this paper we conduct a systematic study on comparing the unsupervised dimensionality reduction techniques for text retrieval task. We analyze these techniques from the view of complexity, approximation error and retrieval quality with experiments on four testing document collections.

Keywords: Dimensionality reduction, Information retrieval, Latent semantic indexing, Matrix decompositions.

1. Introduction

Data such as images, text, and multimedia are high dimensional in nature. As the dimensionality of data increases query performance decreases, demand for processing power and storage space increases. This problem of high dimensionality is defined as the curse of dimensionality [21]. As a result of this curse, efficiency of data indexing structure decreases rapidly with increase in the number of dimensions. Existing indexing structures perform well in low dimensionality spaces and poorly in high dimensionality spaces. Solution for this problem is to reduce the dimensionality of the search space before indexing the data. Researchers have found that reducing the dimensionality of data results in a faster computation while maintaining reasonable retrieval accuracy [16, 20].

Information Retrieval (IR) is a domain of research that aims at providing objects satisfying the user information needs. Vector Space Model (VSM) is a standard IR model that represents documents and queries in a high dimensional term space. These spaces are susceptible to noise and have difficulty in capturing the underlying semantic structure [19]. The noisiness in the form of polysemy and synonymy coupled with high dimensionality of

vector space representation of document collections gives many challenges to text retrieval systems.

In [7] Deerwester et al., have proposed Latent Semantic Indexing (LSI), a variant of vector space IR model, which maps a high dimensional space into a low dimensional space. To approximate a source space with fewer dimensions, LSI uses matrix algebra technique termed Singular Value Decomposition (SVD). Vectors representing the documents and queries are projected in new, low dimensional space obtained by truncated SVD. But time and space complexities of SVD restrict its applicability to matrices with large size. To handle this situation researchers have explored alternate strategies for Dimensionality Reduction (DR) [20]. However, lack of empirical work comparing these techniques in a systematic manner for text retrieval task needs attention of researchers. In this research, we study and evaluate four popular DR techniques for text retrieval task. We identify the effectiveness of Singular Value Decomposition, Non-negative Matrix Factorization, Independent Component Analysis and Fuzzy K-Means algorithm. Rest of this paper is organized as follows. Section 2 presents a discussion on DR techniques. Section 3 presents the experimental details on four standard document collections. Section 4 discusses the results obtained. Section 5 provides the conclusion followed by acknowledgement and references.

2. Dimensionality Reduction

To address the curse of dimensionality, DR techniques are proposed as a data pre-processing step. This process identifies a suitable low-dimensional representation of original data. Reducing the dimensionality improves the computational efficiency and accuracy of the data analysis. Mathematically the problem of dimension reduction can be defined as: given a r-dimensional random vector $\mathbf{X}=(x_1,x_2,\ldots.x_r)^T$, the objective is to find a representation of lower dimension $\mathbf{S}=(s_1,s_2,\ldots,s_k)^T$, where $k<r$, which preserves the content of the original data, as much as possible according to some criterion. DR techniques are classified as supervised and unsupervised techniques based on the learning process. Supervised algorithms need a training set with the class label information to learn the lower dimensional representation according to some criteria and then predict the class labels on unknown text data. Linear Discriminant Analysis (LDA), Maximum Margin Criterion (MMC), Orthogonal Centroid (OC) algorithm are some of the supervised DR techniques [6, 21]. Unsupervised approaches such as SVD project the original data to a new lower dimensional space without utilizing the label information.

DR techniques functions either by transforming the existing features to a new reduced set of features or by selecting a subset of the existing features. Feature transformation techniques aim to reduce the dimensionality of data to a small number of dimensions which are linear or non-linear combinations of

the vector coordinates in the original dimensions. These techniques are believed to be successful in uncovering the latent structures in the datasets [11]. Examples of various feature transformation techniques include PCA, ICA, Projection pursuit and Factor analysis. Unsupervised feature selection techniques are much harder than the supervised techniques. A good explanation on these techniques can be found in [6]. Recently, statistical and linear algebraic projection techniques have gained popularity for DR. These techniques reduce the dimensionality by linearly transforming the original data matrix. Linear algebra based DR techniques are based on projections where the dimensionality of the matrix is reduced by multiplication or transformation of data matrix. Examples of these techniques include SVD, ICA and NMF. In the literature, research efforts have been made to implement clustering for DR [9, 10].

Few researchers have worked on comparing some of the DR techniques for text retrieval task. In [20], Vinay et al., have evaluated the performance of PCA, ICA and Random Mapping (RM) techniques. Their investigations on two testing document collections revealed that in case of text retrieval, RM is outperformed by PCA and ICA. Analysis of Moravec [16] has revealed that SVD based LSI is slow in computation but accurate when compared with other DR techniques. Moravec has analyzed the performance of LSI using SVD, Random projections and FastMap. Random projections project the document vectors into a subspace using a randomly generated matrix. FastMap is a pivot based technique on multi-dimensional scaling. However in the literature so far there is no direct comparative analysis between SVD, NMF, ICA and FKM techniques specifically for the task of text retrieval, which is the main focus of our work.

2.1. Singular Value Decomposition

A document collection of t terms and d documents is represented by a term-document matrix with t rows, d columns and with rank r. Vectors representing documents and queries are projected in new, low dimensional space obtained by truncated SVD. The SVD of a term-document matrix $\mathbf{A}$ is written as [1, 18]

$$\mathbf{A} = \mathbf{U}.\mathbf{\Sigma}.\mathbf{V}^T \quad (1)$$

If the term-document matrix $\mathbf{A}$ is t x d, then $\mathbf{U}$ is a t x r orthogonal matrix, $\mathbf{V}$ is a d x r orthogonal matrix and $\mathbf{\Sigma}$ is r x r diagonal matrix where the values on the diagonal of $\mathbf{\Sigma}$ are called the singular values. Singular values can then be sorted in decreasing order and the top k (k<r) values are selected as a means of developing a latent semantic representation of original matrix. The geometric interpretation of SVD is to consider the columns of $\mathbf{V}^T$ as defining the new axes, the rows of $\mathbf{U}$ as coordinates of the objects in the space spanned by these new axes and $\mathbf{\Sigma}$ as a scaling factor indicating the relative importance of each new axis [7]. By changing (r-k) rows of $\mathbf{\Sigma}$ to zero rows a

low rank approximation to **A** called $\mathbf{A}_k$ can be created through the truncated SVD as,

$$\mathbf{A}_k = \mathbf{U}_k . \mathbf{\Sigma}_k . \mathbf{V}_k^T \quad (2)$$

where $\mathbf{U}_k$ is the t x k term-concept matrix, $\mathbf{\Sigma}_k$ is the k x k concept-concept matrix, $\mathbf{V}_k^T$ is the k x d concept-document matrix. Only the first k columns are retained in $\mathbf{U}_k$ and k rows are retained in $\mathbf{V}_k^T$. By applying the SVD on a term-document matrix, documents will be represented in a vector space of artificial concepts. Each of the k reduced dimensions corresponds to a latent concept which helps to discriminate the documents.

2.2. Non-Negative Matrix Factorization

Non-negative Matrix Factorization (NMF) is another development in the field of DR and clustering [15]. The semantic space derived by NMF contains each axis capturing the base topic of a particular document cluster and each document is represented as an additive combination of base topics [14]. NMF is proved to be useful in approximating high dimensional data comprised of non-negative components [5]. For the data matrix **A** of size t x d with each column of t dimensional non-negative vector of original database (d vectors), NMF factorizes **A** as

$$\mathbf{A}=\mathbf{W}.\mathbf{H} \quad (3)$$

where **W** is t x k and **H** is k x d and $k \leq d$. Each column of **W** contains a basis vector and each column of **H** contains the weights needed to approximate the corresponding columns in **A** using the basis from **W**. Here the choice of k is mostly dependent upon the characteristics of the particular database within the application. In contrast to SVD, NMF does not need to be orthogonal and each document is guaranteed to take only non-negative values in all the latent semantic directions.

2.3. Independent Component Analysis

Independent Component Analysis (ICA) tries to linearly transform the original data into components that are maximally independent from each other. ICA assumes that the observed multivariate data are linear or non-linear mixtures of some unknown latent variables with unknown mixing coefficients. These latent variables are called the independent components of the data. ICA technique seeks linear projections that are as independent as possible. However, these projections are not necessarily orthogonal to each other. For DR, ICA finds k components that effectively capture variability of the original data. ICA factors a data matrix, **A** of size t x d as

$$\mathbf{A} = \mathbf{C}.\mathbf{F} \tag{4}$$

where **C** is described as the mixing matrix with *t* rows and *k* columns and **F** as the matrix of independent components with *k* rows and *d* columns.
In this study, FastICA algorithm is used to identify the latent dimensions in the data [22]. To speedup the iteration process, the observed data can be uncorrelated by a linear transformation called 'pre-whitening'. ICA is well studied by researchers in signal processing. ICA defines interestingness in terms of the directions that are statistically independent and least normally distributed. We can find applications of ICA on text data in [13].

2.4. Fuzzy K-Means Clustering

In [8] Dhillon and Modha have used centroids of clusters which are created using spherical K-means algorithm for lowering rank of the term-document matrix. The IR technique using clustering for decomposition is called Concept Indexing (CI). Concept index is a space containing linear combinations of centroids of the clusters. CI is computationally more efficient and requires less memory than LSI. Dobsa and Dalbelo-Basic [9] have proposed an improvement to CI using Fuzzy K-Means (FKM) algorithm for decomposition.
The FKM algorithm works on the assumption that there are natural tendencies of cluster structure in the data and its goal is to uncover this latent structure [9,10]. In contrast to crisp or hard clustering techniques, FKM algorithm allows the objects to partially belong to multiple clusters. FKM partitions a set of *t* dimensional vectors $\mathbf{X}=\{X_1,X_2\ldots.X_d\}$ into *k* clusters where $X_j=\{x_{j1},x_{j2},\ldots.,x_{jt}\}$ represents the j^{th} sample for $j=1...d$. Every cluster is a fuzzy set. For the j^{th} sample X_j and the i^{th} cluster center v_i, there is a membership degree u_{ij} indicating with what degree the sample X_j belongs to the cluster center v_i resulting in a fuzzy partition matrix $\mathbf{U}=(u_{ij})_{t \times k}$. The FKM algorithm is based on minimizing the heuristic global objective function $\mathbf{J}_{fuzz}$ defined as

$$\mathrm{J}_{\mathrm{fuzz}} = \sum_{i=1}^{k}\sum_{j=1}^{d} u_{ij}^{\,b} d_{ij}^{\,2} \tag{5}$$

where d_{ij} is the Euclidean distance between X_j to the cluster center v_i defined as

$$d_{ij} = \sqrt{\sum_{p=1}^{d}(v_{ip} - x_{jp})^2} \tag{6}$$

The exponent *b* in equation 5 is called as fuzzifier parameter and determines the fuzziness of the clustering. In all our experiments we consider the value of *b* to be 1.05. For higher values of *b* the clustering becomes more fuzzifier. The computational formulae of u_{ij} and v_i are

$$u_{ij} = \frac{1}{\sum_{p=1}^{k} \left(\frac{d_{ij}}{d_{pj}} \right)^{2/(b-1)}} \quad \text{where } (b \neq 1) \text{ and } v_i = \frac{\sum_{j=1}^{d} u_{ij}^{m} X_j}{\sum_{j=1}^{d} u_{ij}^{m}} \tag{7}$$

Minimization of objective function **J** is achieved by iteratively optimising u_{ij} and d_{ij}. Algorithm for FKM can be found in [10].

3. Experimental Analysis

To evaluate the above discussed DR techniques for the task of text retrieval, we have conducted experiments on four testing document collections and analyzed their performances based on standard IR metric. We have used Medline, Cranfield, CACM and CISI datasets in our experiments. These collections are widely used in IR and text mining research. Medline document collection contains a total of 1033 documents indexed by 5735 terms and 30 queries. Cranfield data collection contains 1398 documents indexed by 4563 terms and 225 queries. CACM collection consists of 3204 documents with 5763 index terms and 52 queries. CISI document collection contains 1460 documents and 76 queries indexed by 5544 terms. These document collections are pre-processed by adopting standard procedures including removal of stopwords, performing stemming, term weighting and document length normalization. In our analysis, we have applied popular TF-IDF term weighting strategy over these document collections. A good explanation on term weighting can be found in [19]. All the experiments are carried out in Matlab 6.5 environment [23]. We have evaluated the above discussed DR techniques using three parameters: computation complexity, error in dimensionality approximation and quality of the retrieval. To implement SVD and NMF, we have used functions available in Matlab and to implement ICA we have used FastICA toolbox [22].

3.1. Complexity

As the DR works in a high dimensional environment with large datasets, complexity of the reduction or clustering techniques is a major issue. Hence for a term-document matrix, the complexity of DR varies based on the reduction technique used. Though certain techniques are good at approximating the original source space, they are impractical to implement due to their high computational complexity. Truncated SVD is computationally complex. For a dense matrix **A** of size *m* x *n* the complexity of computing SVD is $O(mn^2)$ and for sparse matrix with average *c* non-zero entries per data item the complexity is O(*mnc*) [1,18]. Even with a large degree of sparsity

(c<0.01), computing truncated SVD becomes intractable with thousands of terms and documents.

There are several algorithms for computing independent components. These algorithms are iterative in nature. The most popular algorithm for computing ICA is FastICA, which has good convergence properties. The complexity of ICA is heavily dependent on the objective function and algorithm. FastICA algorithm is at least quadratic [12]. The complexity of NMF is O(mk), where m is the number of rows of the matrix and k is the number of basis vectors generated [18]. The iterative algorithm proposed in [14], updates each entry of **W** and **H** matrices on each round and converges within 100 iterations. The complexity of FKM is O($nmkT$) where k is the number of clusters and T is the number of iterations required for FKM algorithm to reach $\mathbf{J}_{fuzz}$ minimum [9].

3.2. Approximation Error

A common measure of identifying the approximation error is the Frobenius norm of the difference between the original term-document matrix and its reduced approximation. The observed error can be calculated as $\|\mathbf{A} - \mathbf{A}_k\|_F$.

Smaller the Frobenius norm, better the approximation. Table 1 gives the observed error on four document collections using each of SVD, NMF, ICA and FKM techniques. Selecting the intrinsic dimensions in data is an interesting problem in IR research. An exciting work in this direction can be found in [3]. In choosing the number of dimensions, we need the value to be large enough to fit all the real structure but small enough so that sampling error or unimportant details do not fit in the data. Selection of intrinsic dimensionality attempts to balance these two opposing effects [19]. The approximation error shown here is measured against the 100 intrinsic dimensions for Medline, CACM and CISI collection and 300 intrinsic dimensions for Cranfield collection. Approximation error decreases with increase in number of dimensions. It is evident from the table 1 that best approximation is achieved by SVD when compared with other reduction techniques.

Table 1. Dimensionality approximation error with regard to Frobenius norm

	SVD	NMF	ICA	FKM
Medline	0.0967	0.1690	0.1640	0.1950
Cranfield	0.0063	0.1730	0.1380	0.2360
CACM	0.2320	0.2920	0.2610	0.3030
CISI	0.0880	0.1450	0.1620	0.1550

3.3. Retrieval Quality

Performance in IR systems is summarized in two parameters: precision and recall. Precision is the portion of relevant documents in the set returned to the user and recall is the portion of all relevant documents in the collection that are retrieved by the system [19]. The precision figures at 11 standard recall levels are interpolated by the rule which states that the interpolated precision at the j^{th} standard recall level is the maximum known precision at any recall level between the j^{th} and the $(j+1)^{th}$ recall levels. Figure 1 presents comparison of interpolated precision results of query projections on the semantic space derived by SVD, NMF, ICA and FKM techniques at 11 standard recall levels.

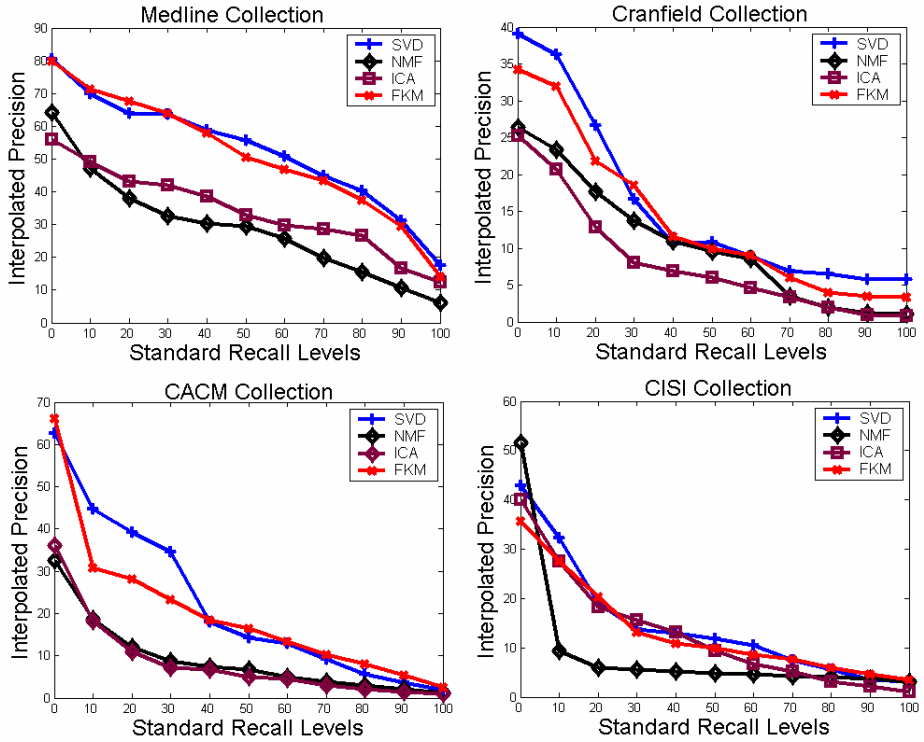

Fig. 1. Interpolated precision results for SVD, NMF, ICA and FKM based retrieval

Looking at the precision-recall curves, it is evident that semantic space derived by SVD has produced best retrieval results when compared with other techniques on all testing document collections. Semantic space derived

by FKM has presented results almost analogous to SVD on all the collections. However, other DR techniques ICA and NMF have performed poorly on all the document collections. Typically these graphs slope downward from left to right enforcing the notion that as recall increases, precision decreases. Curves close to the upper right hand corner of the graph, where recall and precision are maximized, indicate the best performance.

4. Discussion

The dimensionality curse has great influence on the effectiveness of high dimensionality indexing. An ideal DR technique should have the capability of efficiently reducing data into a lower dimensional model while preserving the properties of original data. In this work, we have analyzed the four popular DR techniques in the context of computation complexity, approximation error and retrieval quality. This study has provided some useful insights.

Summarizing the results we can observe that, projection of term-document matrix on the subspace achieved by SVD produces better approximation of the term-document matrix with regard to Frobenius norm. Although using the truncated SVD to project the term-document matrix into a lower dimensional space has the benefit of removing noise from the data, it has the drawback of being computationally expensive. Majority of the processing time in LSI is taken up with computing SVD. Given the potentially huge size of term-document matrices, recomputing the truncated SVD each time when the term-document matrix is altered can be prohibitively expensive.

Our analysis on four testing document collections with standard retrieval metric has revealed that DR computed by FKM achieves retrieval performance similar to that of LSI using SVD. Also, amount of time required by FKM is significantly smaller than that is required by SVD. NMF and ICA are computationally less expensive than SVD and their approximation error is similar or even lesser than FKM. The NMF technique is advantageous by its non-negativity constraints and simple iterative algorithm for computation. But the quality of the retrieval by NMF and ICA techniques is poor when compared with SVD and FKM techniques. Only on CISI document collection, retrieval results from ICA are similar to the retrieval results from SVD and FKM. Literature on clustering and DR shows that, in view of clustering validity functions, ICA and NMF techniques works better than SVD for clustering problems [13, 17]. However for the quality of text retrieval, these techniques are proved to be less superior when compared with SVD and FKM techniques.

5. Conclusion

The basic problem in text retrieval is the high dimensionality of the natural language. DR techniques improve the data representation by understanding

the data in terms of concepts rather than words. The objective of this paper is to provide a detailed analysis of unsupervised DR techniques for text retrieval task in terms of complexity, error in the approximation and quality of retrieval. Our results show that semantic space derived by SVD and FKM produces better retrieval results than other DR techniques. However, with less complexity FKM proves to be a better option for deriving the semantic space.

6. Acknowledgement

Author gratefully acknowledges the financial support from Dept. of Science and Technology, Govt. of India under the grant number SR/S3/EECE/25/2005. Also, author thank the anonymous reviewers for their useful suggestions.

7. References

1. Aswani Kumar, Ch., Srinivas, S.: Latent semantic indexing using eigenvalue analysis for efficient information retrieval. International Journal of Applied Mathematics and Computer Science, Vol. 16, No. 4, 551-558. (2006)
2. Aswani Kumar, Ch., Srinivas, S.: A note on the effect of term weighting on selecting intrinsic dimensionality of data, Cybernetics and Information Technologies, Vol. 9, No. 1, 5-12. (2009)
3. Aswani Kumar, Ch., Srinivas, S.: Automatic selection of intrinsic dimensionality of data, International Journal on Information Processing, Vol. 3, No. 2, 8-16. (2009)
4. Aswani Kumar, Ch., Srinivas, S.: On the performance of latent semantic indexing based information retrieval, Journal of Computing and Information Technology, Vol. 17, No. 3, 259-264. (2009)
5. Berry, M.W., Browne, M., Langville, A.N., Pauca, V.P., Plemmons, R.J. :Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics and Data Analysis. Vol. 52, No.1, 155-173. (2007)
6. Cunningham, P. :Dimension reduction. Technical Report: UCD-CSI-2007-7. (2007)
7. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R. :Indexing by latent semantic analysis, Journal of the American Society for Information Science. Vol. 41, No. 6, 391-407. (1990)
8. Dhillon, I.S., Modha, D.S.: Concept decomposition for large sparse text data using clustering. Machine Learning, Vol. 42, No. 1, 143-175. (2001)
9. Dobsa, J., Dalbelo-Basic, B.: Concept decomposition by fuzzy k-means algorithm. In Proceedings International conference on Web Intelligence. 684-688. (2003)
10. Doring, C., Lesot, M.J., Kruse, R.: Data analysis with fuzzy clustering methods. Computational statistics and data analysis. Vol. 51, No. 1, 192-214. (2006)
11. Foder, I.K.: A survey of dimension reduction techniques. Technical report URL-ID-148494, Center for applied scientific computing, Lawrence Livermore National Laboratory. (2002)
12. Hyvarinen, A., Oja, E.: Independent component analysis: Algorithms and applications. Neural Networks. Vol. 13, No. 4-5, 411-430. (2000)
13. Kolenda, T., Hansen, L.K., Sigurdsson, S.: Independent components in text. Advances in Independent Component Analysis. Springer-Verlag. (2000)

14. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems. Vol. 13, 556-562. (2001)
15. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature. Vol. 401, 788-791. (1999)
16. Moravec, P.: Testing dimensional reduction methods for text retrieval. In Proceedings of Dateso 2005 Annual international workshop on databases. 113-124. (2005)
17. Shahnaz, F., Berry, M.W., Paul Pauca, V., Plemmons, R.J.: Document clustering using non-negative matrix factorization. Information Processing and Management. Vol. 42, No.2, 373-386. (2006)
18. Skillicorn, D.B., McConnell, S.M., Soong, E.Y.: Handbook of data mining using matrix decompositions. Queen's University, Canada. (2003)
19. Srinivas, S., Aswani Kumar, Ch.: Optimizing heuristics in latent semantic indexing for effective information retrieval. Journal of Information and Knowledge Management. Vol. 5, No. 2, 97-105. (2006)
20. Vinay, V., Cox, I.J., Wood, K., Milic-Frayling, N.: A comparison of dimensionality reduction techniques for text retrieval. In Proceedings of 4th International conference on machine learning and applications.293-298. (2005)
21. Yan, J., Zhang, B., Liu, N., Yan, S., Cheng, Q., Fan, W., Yang, Q., Xi, W., Chen, Z.: Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing, IEEE Transaction on Knowledge and Data Engineering, Vol. 18, No. 3, 320-333. (2006)
22. http://www.cis.hut.fi/projects/ica/fastica/
23. http://www.mathworks.com

Ch. Aswani Kumar is Associate Professor in Networks and Information Security Division, School of Information Technology and Engineering, VIT University, Vellore, India. He obtained his Masters degree in computer science from Nagarjuna University, India and Doctorate from VIT University, India. His research interests are information retrieval, text mining and soft computing techniques. He has published 25 refereed research papers in various national, international journals and conferences. He was principal investigator to a major research project sponsored by the Department of Science and Technology, Government of India. He is a member of various professional societies including ACM, CSI, ISTE. He is editorial board member for few computer science journals and reviewer for many international journals and conferences.

Received: November 13, 2008; Accepted: November 06, 2009.