ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

Improved Correlation Preserved Indexing For Text Mining

Vinnarasi Tharania. I[1],M.Kanchana[2],V.Kavitha[3]

Professor, Department Of Information Technology, Karpaga Vinayaga College Of Engineering & Technology, China Kolambakkam, Madurantakam Taluk,Kanchipuram-603308, Tamilnadu, India[2]

Assistant Professor, Department Of Information Technology, Karpaga Vinayaga College Of Engineering & Technology, China Kolambakkam, Madurantakam Taluk,Kanchipuram-603308, Tamilnadu, India[1,3]

ABSTRACT: Data mining is an excellent domain to work where new concepts are implemented. The knowledge extraction and knowledge discovery is a major task of engineering organization. Therefore a new field of study, Knowledge Discovery in Database (KDD), and data mining is explored. And many applications are been developed. In this paper a discussion about field of data mining is also placed. Document clustering is a field where we are implementing a concept of grouping of similar objects. Some groups are formed and the documents are placed under those groups. The main motive of this paper is comparison of various algorithms and giving the best result of the clustering. A new algorithm has been proposed in this paper is ICPI (Improved correlation preserving indexing) which is performed by the correlation similarity measure space. ICPI can successfully find out the essential structures rooted in high dimensional document space. The proposed work is to provide an efficient text mining algorithm to perform mining in the document.ICPI can successfully find out essential structures rooted in high dimensional document space.

KEYWORDS: K-means, Latent Semantic Indexing (LSI), Locality Preserving Indexing (LPI), Correlation Preserving Indexing (CPI), Improved Correlation Preserving Indexing (ICPI), Knowledge Discovery in Database (KDD).

I.INTRODUCTION

More than hundred fifty years, the word "computer" started to appear in the vocabulary. Even twenty years back, a hard disk is transported only in airlines, due to huge size, for ex: a 4GB hard disk has 300Kg. Now a day, 1TB storage space is available in packet sized hard disk. The fields of application of the computers are very vast and various new applications are being searched out. The scientists applied it to prediction of weather earthquakes, storms, controlling of satellites, controlling of atomic reactions in reactors like this their application varies from small to large. The application is not limited, the use of computers are used also in designing cars, aeroplanes, ships, automobiles, bridges and tools etc. In the industry, the computers is playing a good job which applied for many tasks like word processing, record keeping, inventory controlling, patrol processing, account keeping and auditing, stock marketing and ticket reservations. Clustering is occurring as a main concept in the field of data mining. Where they are used in various applications like banking, finance etc.

II. RELATED WORK

From the year 1994 many researcher are researching in this clustering field. Till 2012 up-to-date the research is going in this field. Basically it is started with the k-means and many new researchers have found various new algorithms which improves the performance of mining. Text mining is a fundamental operation used in unverified document association, usual topic mining, and information retrieval. The well-known partitioning clustering algorithm is the K-means algorithm and its variants [2]. Latent Semantic Indexing (LSI)[4]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

is one among the most popular linear document indexing methods which constructs low dimensional representations. The main aim of LSI is to find the best subspace estimation to the original document space in the sense of minimizing the global reconstruction error. In other words, LSI[4] look forward to uncover the most representative features rather the most discriminative features for document representation.

K-MEANS

The K-mean algorithm takes the input as K and partitions a set of n Objects into K clusters so that the resulting intra-cluster similarity is high and the inter-cluster is low. The cluster similarity is measured with the help of the mean of each cluster. They are considered as the cluster's centroid or Centre of Gravity. Firstly, they Randomly select K objects each of which initially represent a cluster mean or centre. Each of the remaining objects an object is assigned to the cluster to which it is most similar based on the cluster object and the mean value. Then they compute a new mean value. This process is repeated until there is no change.

To find the mean of the cluster

$$Mf = 1/n * (x_1 f + x_2 f + \ldots\ldots\ldots + x_n f)$$

- Calculate the new means to be the centroid of the observations in the cluster.

$$O(a, b) = \sqrt{|x_{a1} - x_{b1}|^2 + |x_{a2} - x_{b2}|^2 + \ldots\ldots + |x_{ap} - x_{bp}|^2}$$

LATENT SEMANTIC INDEXING (LSI)

Latent semantic indexing is one of the most popular algorithm for document dimensionality reduction.It is fundamentally based on SVD(Singular value decomposition).suppose the rank of the term document A is r, then LSI decomposes X using SVD as follows:

$$A = U \sum V^T$$

Where $\sum = diag(a_1 \ldots\ldots, a_r)$ and $a_1 >= a_2 >= \ldots\ldots\ldots >= a_r$

are the singular value of A,$U=[a_1 \ldots\ldots, a_r]$ and a_i is called the left singular vector, and the singular vector, and $V=[v_1 \ldots\ldots\ldots, v_r]$,and v_i is called the right singular vector.LSI uses the first k vectors in U as the transformation matrix to embed the original documents into k-dimensional subspace. It can be easily checked that the column vectors of U are the eigenvectors of AA^T.

LOCALITY PRESERVING INDEXING (LPI)

The locality preserving indexing was proposed to discover the discriminant structure of the space of the document. It has been proved that it would have been more discriminant power than the LSI. However the computational complexity of the LPI is very costly due to the reason of that they involving Eigen-decomposition of two dense matrices. The LPI is not highly advised to use, since they use large data set.

The Algorithm used in Locality Preserving Indexing

- PCA Projection

We project the document set {**A***i*} into the PCA subspace by throwing away the smallest principal components. We denote the transformation matrix of PCA by W_{pca} .

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

- Constructing the Adjacency Graph

Let G denote a graph with *n* nodes. The *ith* node corresponds to the document a*i*. We put an edge between nodes *i* and *j* if a*i* and a*j* are "close", i.e. a*i* is among *k* nearest neighbors of a*i* is among *k* nearest neighbors of a*j*.

- Choosing the Weights

If node *i* and *j* are connected, put

$$S_{ab}=X_a^T X_b$$

Otherwise, put $S_{ab} = 0$. The weight matrix *S* of graph G models the local structure of the document space.

- Eigenmap:

Compute the eigenvectors and eigenvalues for the generalized eigenvector problem:

$$ALX^T\mathbf{w} = \lambda ADX^T\mathbf{w}$$

Where *D* is a diagonal matrix whose entries are column (or row, since *S* is symmetric) sums of *S*,

Dii= Σ*jS ji*. *L* = *D* −*S* is the Laplacian matrix.

CORRELATION PRESERVING INDEXING

Correlation preserving indexing (CPI), which clearly considers the multiple structure rooted in the similarities between the documents. It aims to find an optimal semantic subspace by simultaneously maximizing the correlations between the documents in the local patches and minimizing the correlations between the documents outside these patches. This is different from LSI and LPI, which are based on a dissimilarity measure (Euclidean distance), and are determined on detecting the intrinsic structure between widely separated documents rather than on detecting the intrinsic structure between nearby documents. Since the intrinsic semantic structure of the document space is often embedded in the similarities between the documents [7], CPI can effectively find the intrinsic semantic structure of the high-dimensional document space. CPI with euclidean distances performed competitively with CPI with correlation distance. Thus, similarity is an appropriate metric for measuring correlation between the documents.

$$max \sum_{i=0}^{n} \sum_{j=0}^{n} Corr(i,j)$$

$$min \sum_{i=0}^{n} \sum_{j=0}^{n} Corr(i,j)$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

III.PROPOSED METHOD

IMPROVED CORRELATION PRESERVING INDEXING

Compute Correlation between words i and j,

$max \sum_{i=0}^{n} \sum_{j=0}^{n} Corr(i,j)$=similarity

$min \sum_{i=0}^{n} \sum_{j=0}^{n} Corr(i,j)$=dissimilarity

In maximum correlation the maximum similarity is found. And with the min correlation the max dissimilarity is found. Then those values of similarity are preserved. Correlation of documents to be preserved.Correlation Preserving Index to be maintained for all documents.The proposed work, improves existing CPI method for mining similarity pattern within the documents.

There are four modules in the proposed system. They are

- Document collection and pre-processing
- Implementing the correlation function
- Identification of optimal threshold
- Ranking, Index preserving and testing

Step1: Document Collection

Collecting the document from various data set. Data set Such as IEEE Explore, ICM Digital library, Scopus etc….IEEE Explore is a scholarly research database it contains indexes, abstract and provide a full text for the article and papers on various subjects. The IEEE Explore database contains the following collection such as journal, transactions, letters and magazines. Remove authors name and keywords. Convert uppercase letters to lower case letters. Remove stop list words.

Step2: Implementation of Correlation Function

Correlation is to find the similarity.Basically +1.00 positive correlation,-1.00 negative correlation and 0 no correlation.

$$max \sum_{i=0}^{n} \sum_{j=0}^{n} Corr(i,j)$$

Corr(I,J)=$I^T J/\sqrt{I^T I}\sqrt{J^T J}$.Using this formula correlation is found for the documents pre-processed.

Step3: Identification of Optimal Threshold

From the value obtained from the correlation function a threshold value is found in this module. All the values obtained from them is stored in the form of a table. A intermediate value is obtained as the threshold value from the existing values. If the optimal threshold value is not up to the expectation then again the optimal threshold value can be changed and correlation can be done.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

Step4: Ranking, Index Preserving and Testing

The values above the threshold values are ranked in a decrement order. Then those values are stored and an index is maintained. That index value is compared with the database were the dataset is also maintained. Then the output is displayed according to the input given.

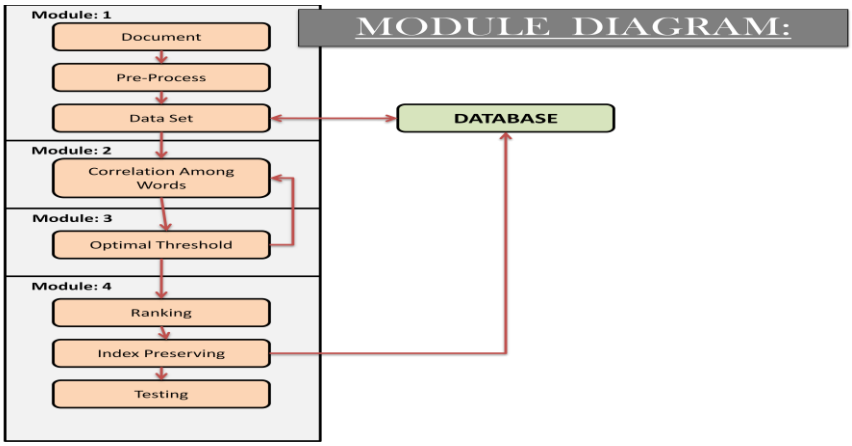

Fig.1 Architectural Structure of Improved Correlation Preserved Indexing

IV.OUTPUT ANALYSIS

The fig 4.1 explains the document collection module initial step for this project the input is collected from the IEEE EXPLORE. So the above figure shows the home page of IEEE EXPLORE.

Fig.2 - The page displays the IEEE EXPLORE

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

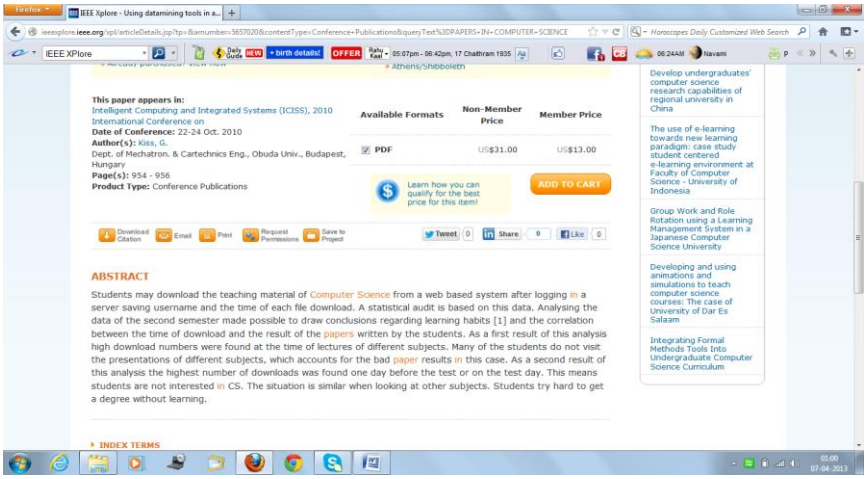

Fig.3 - Abstract of an Input Document

Here an abstract of an IEEE Paper is shown as the example with the sample abstract of Computer Science and Engineering.

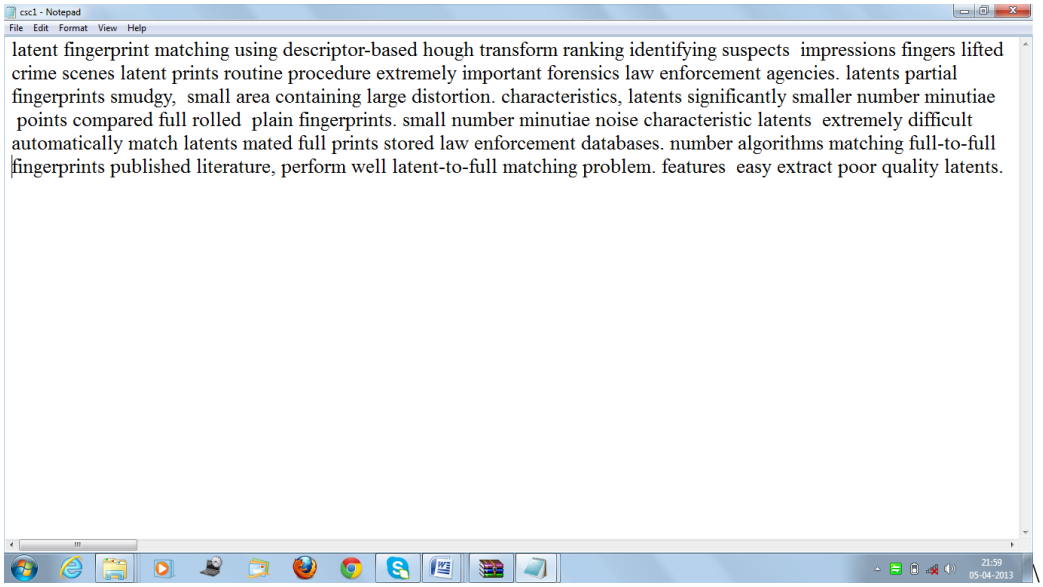

Fig.4 - Pre-processing the documents department of CSE

The above fig.4 explains about the pre-processing of the documents in the sense it removes the stop-list words and the author names from the document collected. And convert them into the lower case letters and into the text document. This is done for a document of CSE.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

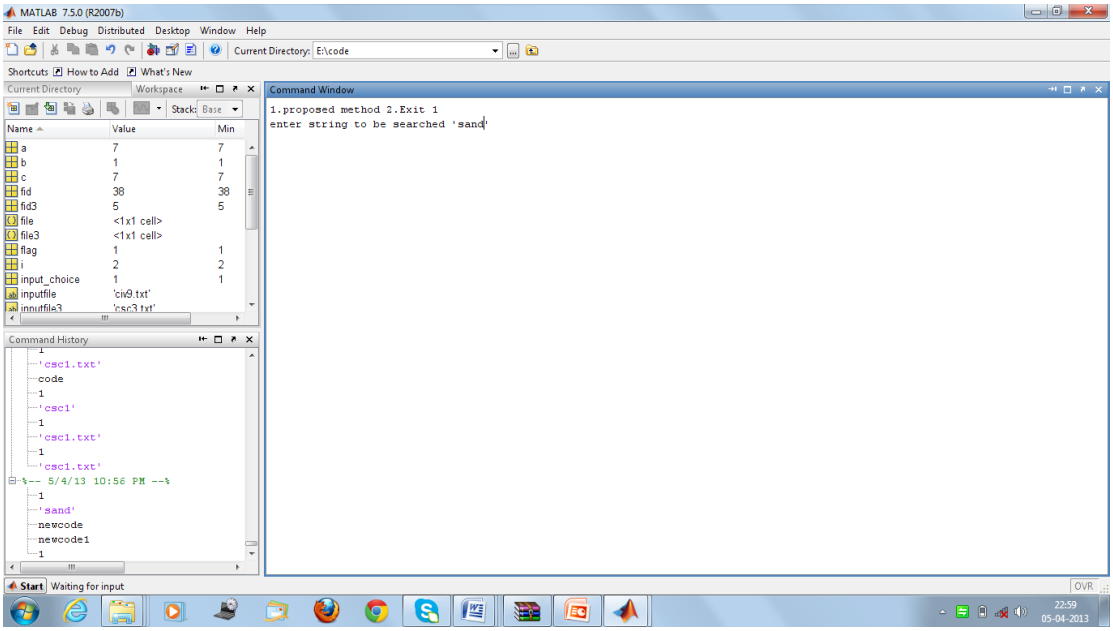

Fig.5- The string has been entered

This fig.6 is a continuous process of the previous one here in this figure the string to be searched is given.

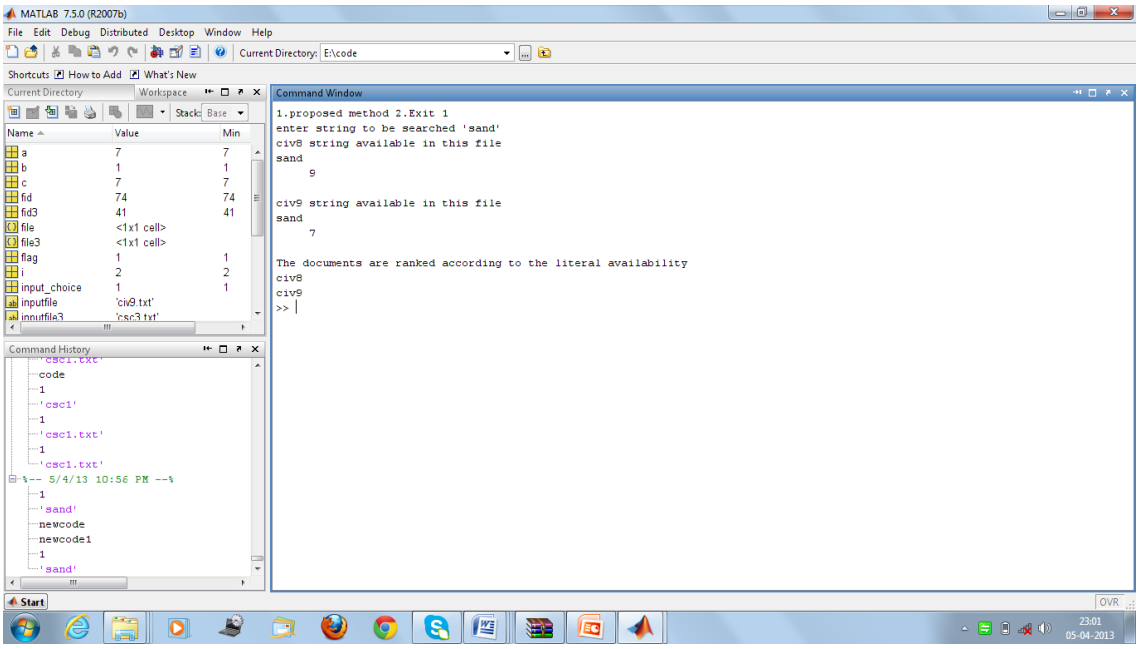

Fig.7 - Literal Count with Ranking

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

Here in the above fig.7 the count of the literal is given along with the ranking. So in this figure they explain were all the words available in which file. And according to their count they are published in an order.

V.CONCLUSION

The proposed work is implemented using few set of documents which is collected from well known data sets such as IEEE EXPLORE, ICM DIGITAL LIBRARY and SCOPUS. The emerging researches from any well established research lab in this field of research can implement the proposed work in the own data sets. The result of this ICPI can be found with help any one single document given as the input. At the same instance many new solutions have been found in this ICPI. New solution is like solution to the problem of synonymy and polysemy. Since none of the algorithms give a better solution to these problem. But even though this algorithm has been proposed still improvement of paper is going on only the survey of this is done. So the future work will be with the result of the algorithm with other algorithm will be produce soon in the future.

REFERENCES

(1) Xiaohui Cui, Thomas E. Potok," Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm".
(2) Hartigan, J. A. 1975. Clustering Algorithms.John Wiley and Sons, Inc., New York, NY.
(3) Deng Cai, Xiaofei He, and Jiawei Han, "Document Clustering Using Locality Preserving Indexing" VOL. 17, NO. 12, December 2005.
(4) S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman, "Indexing by Latent Semantic Analysis," J. Am. Soc. Information Science, vol. 41, no. 6, pp. 391-407, 1990.
(5) Xiaofei He, Deng Cai, Haifeng Liu, Wei-Ying Ma," Locality Preserving Indexing for Document Representation" *SIGIR'04*, July 25–29, 2004.
(6) Deng Cai, Xiaofei He," Orthogonal Locality Preserving Indexing" *SIGIR'05*August 15–19, 2005.
(7) D.K. Agrafiotis and H. Xu, "A Self-Organizing Principle for Learning Nonlinear Manifolds," Proc. Nat'l Academy of Sciences USA, vol. 99, no. 25, pp. 15869-15872, 2002.
(8) Taiping Zhang, Member, IEEE, Yuan Yan Tang, Fellow, IEEE, Bin Fang, Senior Member, IEEE, and Yong Xiang Document Clustering in Correlation Similarity Measure Space VOL. 24, NO. 6, June 2012.
(9)Jiawei Han and Micheline Kamber, "Data mining concepts and techniques" 2nd edition 2006.
(10)T. Hofmann, "Probabilistic Latent Semantic Indexing," Proc. Ann.Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 50-57, 1999.
(11) D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation,"Proc. Neural Information Processing Systems, pp. 601-608, 2001.
(12) D.M. Blei and J.D. Lafferty, "Correlated Topic Models," Proc.Neural Information Processing Systems, 2005.
(13) W. Li and A. McCallum, "Pachinko Allocation: Dag-Structured Mixture Models of Topic Correlations," Proc. Int'l Conf. Machine Learning (ICML), pp. 577-584, 2006.
(14) D.M. Mimno, W. Li, and A. McCallum, "Mixtures of Hierarchical Topics with Pachinko Allocation," Proc. Int'l Conf. Machine Learning (ICML), pp. 633-640, 2007.
(15) C. Zhai, A. Velivelli, and B. Yu, "A Cross-Collection Mixture Model for Comparative Text Mining," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 743-748, 2004.
(16) A. Asuncion, P. Smyth, and M. Welling, "Asynchronous Distributed Learning of Topic Models," Proc. Neural Information Processing Systems, pp. 81-88, 2008.
(17) D.J. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," Proc. Knowledge Discovery in Databases (KDD) Workshop, pp. 359-370, 1994.
(18) H. Sakoe, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-26, no.1, pp. 43-49, Feb. 1978.

BIOGRAPHY

I.VINNARASI THARANIA[1] is working as Assistant professor in Information Technology Department, Karpaga Vinayaga College of Engineering & Technology, Madurantakam Taluk , Kanchipuram-603308. She received Master of Technology (M.Tech) degree in 2013 from Vel Tech, Avadi, Chennai, India. Her research interests are Data Mining and Data Warehousing, Software Engineering etc.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

M.KANCHANA[2] received B.E degree from Madras University in 1998 and received M.E degree from Anna-University in 2006 pursing Ph.D in Anna University. Currently working as a Professor, Head of the department, Department of Information Technology. Karpaga Vinayaga College of Engineering and Technology, Chennai – 603 308, Tamil Nadu, India. Her research interests are Data Mining and Data Warehousing, Image Processing, Medical Image Processing , Networking etc.

V.KAVITHA[3] is working as Assistant professor in Information Technology Department, Karpaga Vinayaga College of Engineering & Technology, Madurantakam Taluk , Kanchipuram-603308. She received Master of Technology (M.E) degree in 2009 from Anna-University, Chennai India. Her research interests are Data Mining and Data Warehousing, Web Mining, Networking etc.