ISSN:0975-9646

Parallel Comparison of Text Document with Input Data Mining and VizSFP

Priyanka P. Palsaniya[1], D. C. Dhanwani [2]

[1]*Student ME 2nd CSE P.R.Pote (Patil) College of Engineering & Management, Amravati.*
[2] *Assistant Professor P.R.Pote(Patil) College of Engineering & Management, Amravati*

Abstract- This paper present that increasing efficiency for document processing is fundamental concept for any organization. All the documents have been processed manually but it seems very difficult if someone needs to have particular information from particular document in a less time. Therefore, parallel comparison is focusing on performance and efficient processing of multiple documents simultaneously. Design of parallel algorithm and performance measurement is the major issue. If one wants the document content to be excess as soon as possible then it require too much time. The major need of this paper is to meet the performance objectives such as time, dataset names, and size of data sets, support value and match score that tell us the whole information about the particular document and also give us the document in the re-rank manner and visualizes this result for easy analysis. There is no any Technique that can handle, manage and retrieve the information as per the user need. So, we used the combination of the clustering and mining technique to prove the result of this parallel comparison and to evaluate the accuracy i.e. stable output in the form of graph. The experimental results show that the proposed parallel comparison algorithm for each Mining, Clustering and for Comparison achieves good performance as compare to sequential.

Keywords: Data mining, Clustering, Text Mining, Parallel Computing, Fuzzy Sets, Data Visualization.

I. INTRODUCTION

Nowadays, large number of data is generated in real life applications. Data is a collection of objects(i.e. record, item) and their attributes(i.e. property, variable, field, feature or characteristic of an object).Real life data typically needs to be pre-processed (e.g. cleansed, filtered, transformed) in order to be used by the machine learning techniques in the analysis step or to get this data more easily or frequently. However, to do all this user interaction is needed which is very time consuming because much of the useful information are hidden in that enormous amount of data.

All the documents have been processed manually but it is very difficult if someone needs to have particular information from document. Text mining in document comparison using techniques, such as clustering, parallel comparison of text or word has been developed to handle the unstructured documents. Thus, document comparison using Text mining and Text clustering has become an increasingly popular and essential theme. Therefore, the basis of all this is to developed a system having a set of document and to get the list of document in which our search keyword are present then the system shall not only retrieve the files, but re-rank them in order to get the list of files most wanted by the users, and that to in a very short span of time using parallel comparison of text or word for each mining and clustering. Therefore, the techniques and algorithm get describe that involved in achieving good performance by reducing execution time. Therefore parallelization strategy get design i.e. parallel comparison, while taking into account three abstraction levels (Documents-Sentences-Words) to more precisely determine the data and to calculate Parallel fuzzy triadic similarity, called PFT-Sim, between documents, sentences and words based on parallel algorithms. And Parallel Processing Thread based on parallel processing gets used for each and every document.

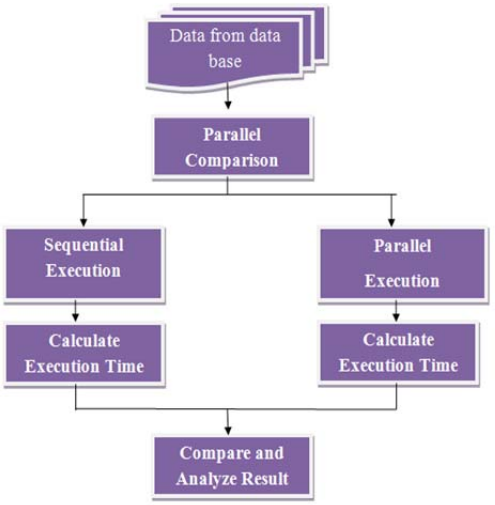

Fig. 1: Overview of Parallel Comparison.

The above diagram Overview of Parallel Comparison takes an input from a Large Database and then the user used to perform a Parallel Comparison on that data taken from a database. This Parallel Comparison get perform in two ways i.e. in Parallel and in Sequential. When user perform a Sequential Execution and Parallel Execution then time get evaluate for both the technique and finally, result get evaluate. Therefore for extracting useful information from a huge dataset we are using text mining in document comparison by using technique such as clustering.

Text mining: Text mining is a process of using tools to extract useful knowledge from a large datasets.

Text mining = information retrieval + statistics + artificial intelligence (natural language processing, machine learning / pattern recognition).

Clustering: Clustering is nothing but dividing a data objects into a group so that the object having a same property i.e. the object which are more similar are placed in one group and the object which are less similar are placed into different group [1], [2].

II. LITERATURE REVIEW

Literature survey is the most important step in software development process. Every algorithm has their own importance and behavior of the data, but on the basis of this research some technique get found that are NLP with Semantic Matching mining algorithm, Bisecting k-means clustering algorithm and PFT-SIM for comparison or VizSFP is simplest algorithm as compared to other algorithms. Generally, clustering or mining methods can be categorized as Hierarchical and Partitional (Non Hierarchical). Therefore, clustering of data can be done either by using Hierarchical and Partitonal technique. It is possible to use different types of algorithms to extract most important information from a database.

Firstly, the work is motivated by research in the field of mining and partitioning method is well technique for this. Partitioning method divide the data into different subset or group such that some criterion that evaluate the clustering quality is optimize. There are different types of Partitional algorithm like Bisecting K-mean, K-mean, Fuzzy-C-Mean, Aproiri etc. These algorithms are among the most important data mining algorithms in the research community. Second, the work is motivated by research then in the field of clustering to divide that useful data or text that we get from mining into the group or cluster. Hierarchical method can be divided into agglomerative and divisive variants. Hierarchical clustering is used to build a tree of clusters, also known as a dendrogram. Every node contains child clusters. In hierarchical clustering assign each item to a cluster such that having N items result in N clusters. Find closest pair of clusters and merge them into single cluster. So, that one cluster get reduce from the whole structure. Compute distances (similarities) between the new and old clusters. It works until all k clusters get merge.

Therefore, different approaches get used while clustering the data. In the first agglomerative approach, a bottom-up clustering method get commonly used. It works from bottom to up. This method starts with a single cluster which contains all objects, and then it splits resulting clusters until only clusters of individual objects remain. It terminate when individual cluster contain a single object. In the second divisive approach, a top-down clustering method get used which is same as agglomerative approach but works in the opposite direction to that of agglomerative technique [1].

S. Anupama Kumar and M. N. Vijayalakshm [3] illustrate that various data mining techniques like classification; clustering are apply on the student's data base. This study can be used by the learner and teaching community to increase the performance. This paper explain the many techniques of data mining according to Educational data to design a new environment Result of this paper is that education system can enhanced their performance by using data mining techniques. These papers show that every method has its own key area in which it performs accurate.

Dr. Mohd Maqsood Ali [4] presents the roll of data mining in education sector. Every university and its colleges enroll thousands of students into various courses or programs every year. Information is collected from students at the time of admissions and get store in computers. Before using the data we must understands the nature of data for predicting the behavior and performance of teacher as well as of student. The outcomes of this paper explain the application of data mining in education sector and also explain that how data mining can be applied for classifying and clustering student's characteristics.

Narendra Sharma, Aman Bajpai and Mr. Ratnesh Litoriya [5] represents the comparison between various clustering algorithm using weka tool .There are various tools in data mining which are used to analysis the data .They allow the users to analysis the data in different dimension or angles, categorize it, and summarize the relationships identified .Weka is also an data mining tool which is used for analysis the data. The main objective is to show the comparison of the different clustering algorithms of weka and to find out which algorithm will be most suitable for the users.

Comparative Study of Various Clustering Techniques in Data Mining [6]. Application to Text Mining and Bioinformatics [7]. Andreas Hotho and Alexander Maedche and Steffen Staab [8] represent the Ontology based Text-Document Clustering algorithm.

Visualization techniques can be classified as Pixel-oriented techniques, Geometric Projection techniques, Icon-based techniques, Hierarchical and Graph based techniques [9] for visualizing useful pattern. Pixel-oriented techniques will map each data value to a predefined colored pixel and present one attribute data values in separate windows. Problem with this type of technique is: if each data value is represented by one pixel, large datasets let to a problem to arrange the pixels on the screen. Geometric projection techniques present interesting patterns within the given multidimensional data sets. Icon-based technique maps each multidimensional data item to an icon. Hierarchical techniques subdivides the given dataset in to k-dimensional vector space and present the subspaces in a hierarchical fashion where as graph-based techniques present a large graph with the help of layout algorithms, query languages, and abstraction technique [10].

Jaishree Singh, Hari Ram, Dr. J.S. Sodhi [11] Represent the Improved Apriori algorithm which reduces the scanning time by cutting down unnecessary transaction records as well as reduce the redundant generation of sub-items during pruning the candidate item sets, which can form directly the set of frequent item sets and eliminate candidate having a subset that is not frequent.

B.S.Vamsi Krishna, P.Satheesh and Suneel Kumar [13] this paper represent background knowledge derived from Word Net as Ontology is applied during preprocessing of documents for Document Clustering. Document vectors constructed from WordNet Synsets is used as input for clustering. In this paper Comparative analysis is done between clustering using k-means and clustering using bi-secting k-means.

S. M. Savaresi and D. Boley [14] this paper focuses on the unsupervised clustering of a data-set. Kommineni Jenni, Sabahath Khatoon, Sehrish Aqeel [15] this paper represent that Combining multiple clustering methods is an approach to overcome the deficiency of single algorithms and further enhance their performances.

Sara Hajian and Josep Domingo-Ferrer [16] this paper represent discrimination prevention in data mining and propose new techniques applicable for direct or indirect discrimination prevention individually or both at the same time and also discuss how to clean training data sets and outsourced data sets in such a way that direct and/or indirect discriminatory decision rules are converted to legitimate (nondiscriminatory) classification rules. This paper also proposes new metrics to evaluate the utility of the proposed approaches and to compare these approaches.

Tomek Strzalkowski, Fang Lin, Jose Perez-Carballo, and Jin Wang [17] In this paper, they report on the progress of the Natural Language Information Retrieval as Natural language processing techniques may hold a tremendous potential for overcoming the inadequacies of purely quantitative methods of text information retrieval, but the empirical evidence to support such predictions has thus far been inadequate, and appropriate scale evaluations have been slow to emerge.

III. PROPOSED METHODOLOGY

In this Proposed Methodology the techniques and algorithm get describe that involved in achieving good performance by reducing the execution time and workload. Algorithms used to develop this system are NLP with semantic matching for mining, Bisecting K-means for clustering and PFT-sim for comparison. The different algorithm with their steps is:

Algorithm for Mining

Mining is the method for extraction of knowledge and for this NLP algorithm get used. NLP and its tasks are:

The steps for NLP are given below:

1. First, we have to select the one file in the form of text from database. After the selection of file then the operation such as split, tokenize, POS tag, chunk operation is performed
2. Next, splitting operation will be performed; in this splitting operation each and every sentence from the file is separated.
3. In the third step, the tokenize operation will be performed. In this step each and every word from the file is separated.
4. In the fourth step, the operation called POS tag will be performed. In this Pos tag method separate the each noun, pronoun, adjective, adverb phrases.
5. After that, the chunking operation will performed. In this action words are removed and the remaining words are compared with dataset. During comparison the common words are found.
6. Next, the actual result got in the form of support value that tells us how many words or character matches in the total document-sentence-words, and finally we analyze time required for all this processes and final document.

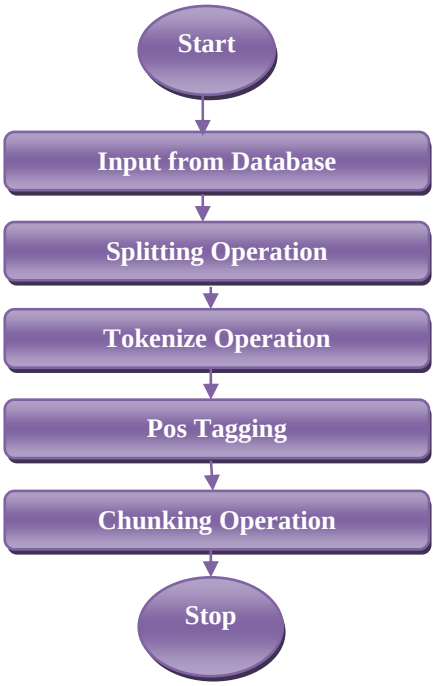

Fig. 2: Diagram for NLP Algorithm

Algorithm for Clustering

Clustering is a technique for grouping the data to get useful information and for this Bisecting K-means algorithm get used. The input for this algorithm is set of Documents that need to be clustered. Bisecting K-Means Algorithm has two main approaches are used: the first one applies the simple strategy of bisecting the greatest cluster and the second one is to split the cluster with greatest difference with respect to the centroid of the cluster. In this bisecting K-mean split the Datasets into two clusters. The Datasets having the greatest support value are place in cluster 1 and the datasets having the less or no support value are place in cluster 2.Steps for Bisecting K-means Algorithm are as follows:

1. Place inputs into the database represented by datasets that are being clustered. These inputs represent initial group centroids.
2. For each Datasets: Calculate the support value means the score of sentences or word present in that document.
   - Repeat the above step until a complete pass through all the datasets results in no data point match in the document. At this point the clusters are stable and the clustering process ends.
   - The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intra cluster distances.
3. Datasets are randomly assigned to the clusters resulting in clusters that have roughly the same number of support value.

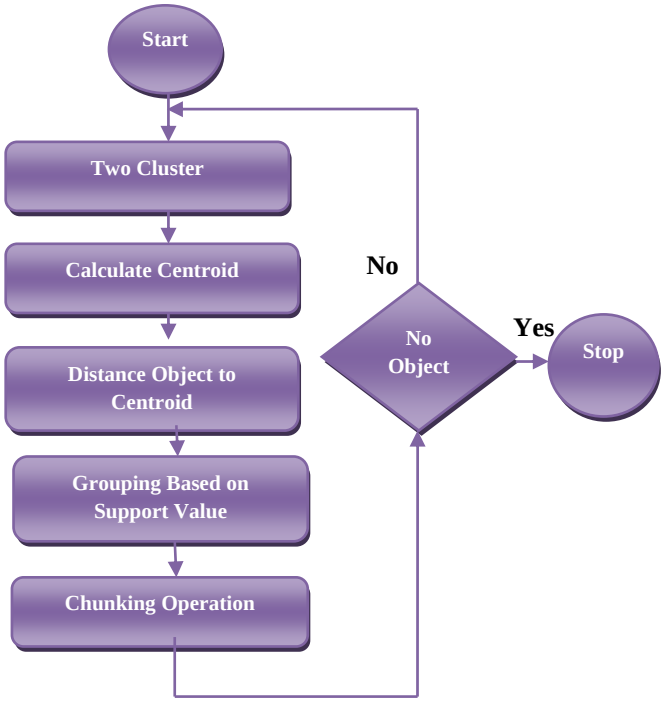

Fig. 3: Diagram for Bisecting K-mean algorithm.

Algorithm for Parallel and Sequential Comparison of Text Documents.

In this algorithm we used to compare the Document-Sentence-Word and for this we used the Comparison algorithm. The algorithm for serial and parallel comparison, Steps of this algorithm are as follows:

Algorithm 1

1. Read the document into memory, initialize a list of scores L.
2. Split the document into sentences, let the number of sentences be S.
3. Not go to each sentence and follow algorithm 2 to get the score of each sentence Ss.
4. Add all the scores of all the sentences to get the total score Ts
5. Add these scores to the list L.

Algorithm 2

1. Split the sentence into words, let the number of words be W.
2. Not go to each word and follow algorithm 3 to get the score of each word Sw.
3. Add all the scores of all the words to get the total score Tw.
4. Return to algorithm 1.

Algorithm 3

1. Let the word from document 1 be W1 and word from document 2 be W2.
2. Check if W1 is present in W2, then increment score of sentence.
3. Check if W2 is present in W1, then increment score of sentence.
4. Add the scores of all the words and return to Algorithm 2.

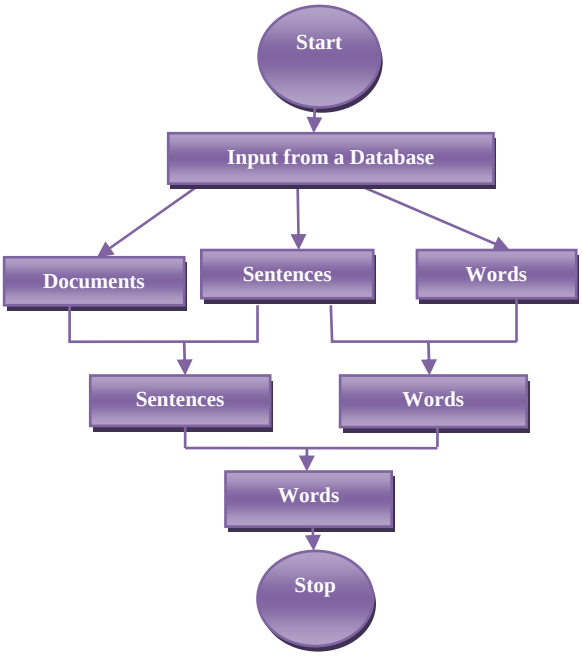

Fig. 4: Diagram for Parallel and Sequential Comparison Algorithm.

IV. RESULTS ANALYSIS

In this section, we will present the effectiveness results for all the three technique. This table 1 shows the comparison chart for mining algorithm by using dataset name, size of dataset, time as a parameter. Execution time is recorded against different datasets to analyze the speedup i.e. performance of parallel against sequential algorithm.

Table 1: Comparison Chart for Mining Algorithm using Time

Sr. No	Dataset name	Size of Dataset	Execution time of parallel mining(sec)	Execution time of sequential mining(sec)
1.	2048	2.04 MB (2,141,446 bytes)	0.001	2.475
2.	2560	2.55 MB (2,676,685 bytes)	0.001	2.609
3.	4608	4.59 MB (4,817,018 bytes)	0.001	2.187
4.	5632	5.61 MB (5,888,407 bytes)	0.001	2.045

Bar Chart shows the Performance Analysis of Mining Algorithm by using time on the x-axis and data sets on the y-axis. Fig. 5 shows that the parallel algorithm required less time as compared to sequential even if the dataset increases the performance of parallel algorithm increases as compared to sequential algorithm.

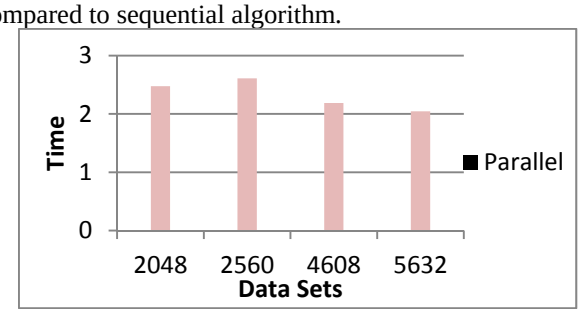

Fig. 5: Performance Analysis of Mining Algorithm using Time

This table 2 shows the comparison chart for clustering algorithm by using dataset name, size of dataset, time as a parameter. Execution time is recorded against different datasets to analyze the speedup i.e. performance of parallel against sequential algorithm.

Table 2: Comparison Chart for Clustering Algorithm using Time

Sr. No	Dataset name	Size of Dataset	Execution time of parallel clustering (sec)	Execution time of sequential clustering (sec)
1.	2048	2.04 MB (2,141,446 bytes)	0.001	0.860
2.	2560	2.55 MB (2,676,685 bytes)	0.001	0.972
3.	4608	4.59 MB (4,817,018 bytes)	0.001	2.156
4.	5632	5.61 MB (5,888,407 bytes)	0.001	1.794

Bar Chart shows the Performance Analysis of Clustering Algorithm by using time on the x-axis and data sets on the y-axis. Fig. 6 shows that the parallel algorithm required less time as compared to sequential even if the dataset increases the performance of parallel algorithm increases as compared to sequential algorithm.

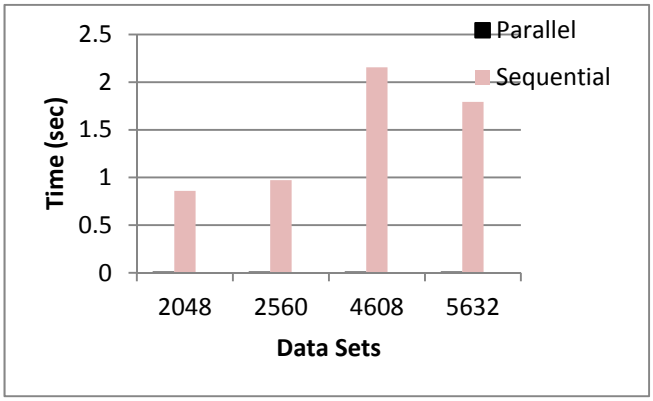

Fig. 6: Performance Analysis of Clustering Algorithm using Time

This table 3 shows the chart for Comparison algorithm by using dataset name, size of dataset, time as a parameter. Execution time is recorded against different datasets to analyze the speedup of parallel against sequential algorithm.

Table 3: Chart for Comparison Algorithm using Time

Sr. No	Dataset name	Size of Dataset	Execution time of parallel comparison (sec)	Execution time of sequential comparison (sec)
1.	2048	2.04 MB (2,141,446 bytes)	0.001	0.852
2.	2560	2.55 MB (2,676,685 bytes)	0.001	0.942
3.	4608	4.59 MB (4,817,018 bytes)	0.001	2.094
4.	5632	5.61 MB (5,888,407 bytes)	0.001	1.824

Bar Chart shows the Performance Analysis of Mining Algorithm by using time on the x-axis and data sets on the y-axis. Fig. 7 shows that the parallel algorithm required less time as compared to sequential even if the dataset increases the performance of parallel algorithm increases as compared to sequential algorithm.

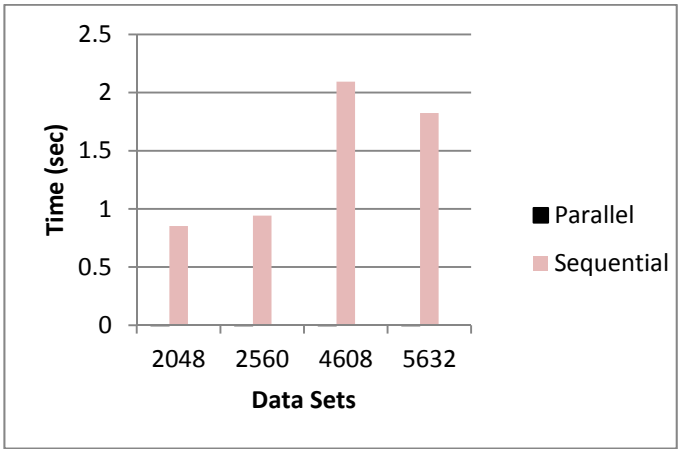

Fig. 7: Performance Analysis of Comparison Algorithm using Time

CONCLUSION

One of the major problems that the users face nowadays is getting the relevant information from the vast amount of database. In this paper, three algorithms NLP with Semantic Matching technique for mining, K-means for clustering and PFT-sim for parallel comparison are implemented on dataset taken from UCI Machine Learning Repository that overcome this problem and it enable to mine, cluster, compare and finally visualize documents information from vast amount of databases. Finally by using each algorithm, some performance parameters such as accuracy based on dataset used, efficiency based on time is calculated. It can be concluded that NLP for mining, K-means for clustering and PFT-sim for comparison technique is more accurate as compared to any other algorithm. From a proper analysis of positive points and constraints on the component, it can be safely concluded that the system is a highly efficient GUI based system. Also this system is user friendly.

REFERENCE

[1] Julian Sedding, "WordNet-based Text Document Clustering," Department of Computer Science, University of York Heslington, York YO10 5DD, United Kingdom.

[2] Preeti Baser and Dr. Jatinderkumar R. Saini, "A Comparative Analysis of Various Clustering Techniques used for Very Large Datasets," International Journal of Computer Science and Communication Networks, Vol3 (4), 271-275.

[3] S. Anupama Kumar and M. N. Vijayalakshmi, "Relevance of Data Mining Techniques in Edification Sector," International Journal of Machine Learning and Computing, Vol. 3, No. 1, February 2013, pp. 4-6.

[4] Dr. Mohd Maqsood Ali, "ROLE OF DATA MINING IN EDUCATION SECTOR," International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 2, Issue. 4, April 2013, pg. 374 – 383.

[5] Narendra Sharma, Aman Bajpai and Mr. Ratnesh Litoriya, "Comparison the various clustering algorithms of weka tools,"

International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 5, May 2012, pp. 73-75.

[6] Aastha Joshi and Rajneet Kaur, "A Review: Comparative Study of Various Clustering Techniques in Data Mining," International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013, pp. 55-57.

[7] Khalid Raza, "A NEW COSIMILARITY MEASURE: APPLICATION TO DATA MINING IN BIOINFORMATICS," International Journal of Computer Science and Engineering, Vol No2, pp. 114-118.

[8] Andreas Hotho and Alexander Maedche and Steffen Staab, "Ontology based Text-Document Clustering," Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany, pp. 1-13.

[9] Christopher L. Carmichael, Carson Kai-Sang Leung, "CloseViz: Visualizing Useful Patterns", the University of Manitoba Winnipeg, MB, Canada, ACM 978-1-4503-0216-6/10/07, July 25, 2010, pg. 17-26.

[10] Ratnesh Kumar Jain, Dr. R. S. Kasana, Dr. suresh Jain," Visualization of Mined Pattern and its Human Aspects," in the International Journal of Computer Science and Information Security, Vol. 4, No. 1& 2, 2009.

[11] Jaishree Singh, Hari Ram, Dr. J.S. Sodhi, "Improving Efficiency of Apriori Algorithm using Transaction Reduction" International Journal of Scientific and Research Publications, Volume 3, Issue 1, January 2013, ISSN 2250-3153.

[12] Suhani Nagpal, "Improved Apriori Algorithm using logarithmic decoding and pruning," International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 3, May-Jun 2012, pp. 2569-2572.

[13] B.S.Vamsi Krishna, P.Satheesh and Suneel Kumar R.," Comparative Study of K-means and Bisecting k-means Techniques in Word net Based Document Clustering," International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Vol. 1, No.6, August 2012, pp. 229-234.

[14] S. M. Savaresi and D. Boley. "On the Performance of Bisecting K-means and PDDP," Department of Computer Science and Engineering, University of Minnesota, 2001, pp. 1-14.

[15] Kommineni Jenni ,Sabahath Khatoon and Sehrish Aqeel, **"**Improve the Performance of Clustering Using Combination of Multiple Clustering Algorithms," in the International Journal of Data Mining Techniques and Applications Vol:02, December 2013, pp. 258-265.

[16] Sara Hajian and Josep Domingo-Ferrer, "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining," in IEEE transactions on knowledge and data engineering, vol. 25, no. 7, july 2013, pp. 1445-1459.

[17] Tomek Strzalkowski, Fang Lin Jose, Perez-Carballo and Jin Wang, "Evaluating natural language processing techniques in information retrieval: a tree perspective," in GE Corporate Research & Development, School of Communication, Information and Library Studies, Rutgers University, pp.1-26.