Enhance Approach for Document Clustering

Kalyani Mendhe [1], Pravin Malviya [2]
[1] MTech. CSE, Department, SBITM COE, Betul, India
[2] Professor, SBITM COE, Betul, India
Email - [1]kalyani231540@gmail.com, [2]malviyapravin2010@gmail.com

Abstract: *These days, we use new technologies in digital world to create digital documents. The examination of such documents is very difficult and essential task. Digital document clustering is nothing but an automatic organization of documents into specific clusters so that documents within a cluster have maximum similarity in comparison to documents in other clusters. Clustering technique is used to measure similarity between document set and grouping highly similar documents together. The study related to similarity measure for document clustering is not generally based on keywords but usually domain based clustering is more popular. So our main objective is to improve the accessibility and usability of text mining process for various digital documents applications. In digital document analysis time constraint is an also most important factor. So is very difficult task for examiner to do such analysis in quick period of time. That's why to do the digital document analysis within short period of time, requires particular techniques to make such complex task in a simpler way. Such special technique called digital document clustering. Here we implemented M-Clustering approach to attain enhanced document clustering for digital document analysis on the basis of interested keyword. The accuracy of clustering of documents has been improved using M- Clustering approach.*

Key Words: *Document Clustering, Digital Document Analysis, Examination, Data Mining, Text Mining*

1. INTRODUCTION:

Recently digital technology particularly in the computer world improves a lot therefore digital documents are also very important part of digital world. So, extraction of relevant data from such huge set of digital document is much more important task for that purpose we need digital document analysis.

1.1 Document Clustering

Document clustering is the process of grouping similar documents into cluster. The main advantage is to retrieve the information effectively, reduce the search time and space, to identify the outliers, to handle the high dimensionality of data and to provide the summary for similar documents. It provides the efficient way of representing and visualizing the documents in which it provides better navigation. The Similarity measure used to find similarity between documents, document representation, and algorithm or technique used to cluster the documents plays major role in document clustering. Document Clustering has been used in variety of application such as recommended system, search optimization, Duplicate content detection, Document Summarization and document Investigation.

Clustering is a division of data into groups of similar objects. Each group, called cluster, it consists of objects that are similar between themselves and dissimilar to objects of other groups. In other words, the goal of a good document clustering scheme is to minimize Intra-cluster distances between documents, while maximizing inter-cluster distances (using an appropriate distance measure between documents). A distance measure (similarity measure) thus lies at the heart of document clustering. Clustering is the most common form of unsupervised learning and this is the major difference between clustering and classification. No super-vision means that there is no human expert who has assigned documents to classes.

In clustering, it is the distribution and makeup of the data that will determine cluster membership. Clustering is sometimes speciously referred to as automatic classification; however, this is inaccurate, since the clusters found are not known prior to processing whereas in case of classification the classes are pre-defined. In clustering, it is the distribution and the nature of data that will determine cluster membership, in opposition to the classification where the classifier learns the association between objects and classes from a so called training set, i.e. a set of data correctly labelled by hand, and then replicates the learnt behaviour on unlabeled data.

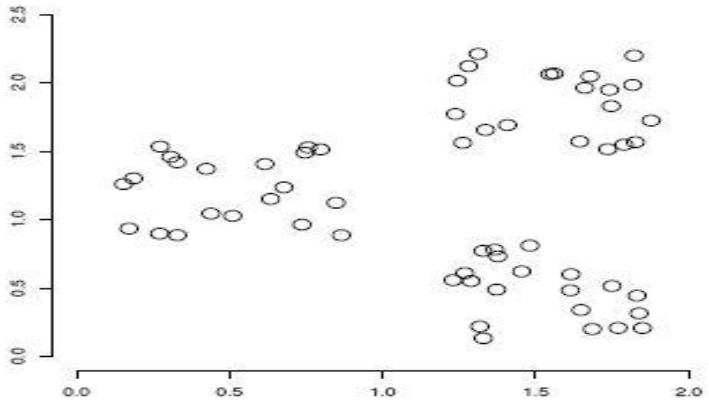

Figure 1.1: An Example of a Data Set with a Clear Cluster Structure

1.1.1 Challenges in Document Clustering

Document clustering is being studied from many years but still it is far from a trivial and solved problem. The challenges in document clustering are:

- Selecting most appropriate features of the documents that should be used for clustering.
- Selecting correct similarity measure between documents.
- Selecting significant clustering method utilizing the above similarity measure.
- Implementing most feasible clustering algorithm in an efficient way in terms of memory usages and CPU resources.
- Finding ways of assessing the quality of the performed clustering.

Furthermore, with medium to large document collections (12,000+ documents), the number of term-document relations is fairly high (millions+), and the computational complexity of the algorithm applied is thus a central factor in whether it is feasible for real-life applications.

We have produced result for sample dataset using two algorithms. The first algorithms are existing algorithms K-mean second one hybrid algorithm that is M-Clustering proposed by us to be better. The results for the M-Clustering algorithm have been generated to compare them with the existing algorithms.

2. METHODOLOGIES AND IMPLEMENTATION

The main objective will be firstly to collect information for dataset. Afterwards remove stop words and the unique words along with count from those datasets will be our next basic objective. Once the search keywords are input we will then perform the clustering using the M-Clustering algorithm.

2.1 Proposed M-Clustering Algorithm

Let's observe the special requirements for good document clustering algorithm:

The document model should better conserve the relationship between words like synonyms in the documents since there are different words of same meaning. Relate a meaningful label to each final cluster is necessary. The high dimensionality of text documents must be reducing. So to achieve this feature in our implemented system we enhance approach to improve document clustering in document analysis. For that we were implementing hybrid approach to accomplish this proposed approach.

We implementing new text clustering algorithm such as M-Clustering algorithm which will gives us the better clustering result .The main idea of M-Clustering algorithm is to use the relative attribute frequencies of the clusters mode in the dissimilarity measures in the M-mode objective function .It has been shown that M-Clustering algorithm is very efficient. Due to the modification proposed in forming representatives for clusters of categorical objects, the dissimilarity between a categorical object and the representative of a cluster is defined based on simple matching as follows.

2.1.1 Steps of M-Clustering Algorithm

- First initialization of logical M-partition of input dataset.
- Find centriod M, one for each cluster then
- For each di, calculate the dissimilarities d (di, pl), l = 1, Reassign di to cluster new specific cluster such that the dissimilarity between di and pl is very less.
- Update both pl and new cluster involve in (say p7)
- Repeat Step (iii) if convergence criteria are not meet. Otherwise stop

3. EXPERIMENTAL EVALUATION:

The snapshots of core modules of proposed framework are explained below. Figure 3.1 illustrates user interface of the proposed system. The examiner has to enter the folder path in which he/she has to search for evidence. In Figure 5.1 the Sample Dataset' folder is selected for testing the system by using browse button after clicking on browse button path will be selected.

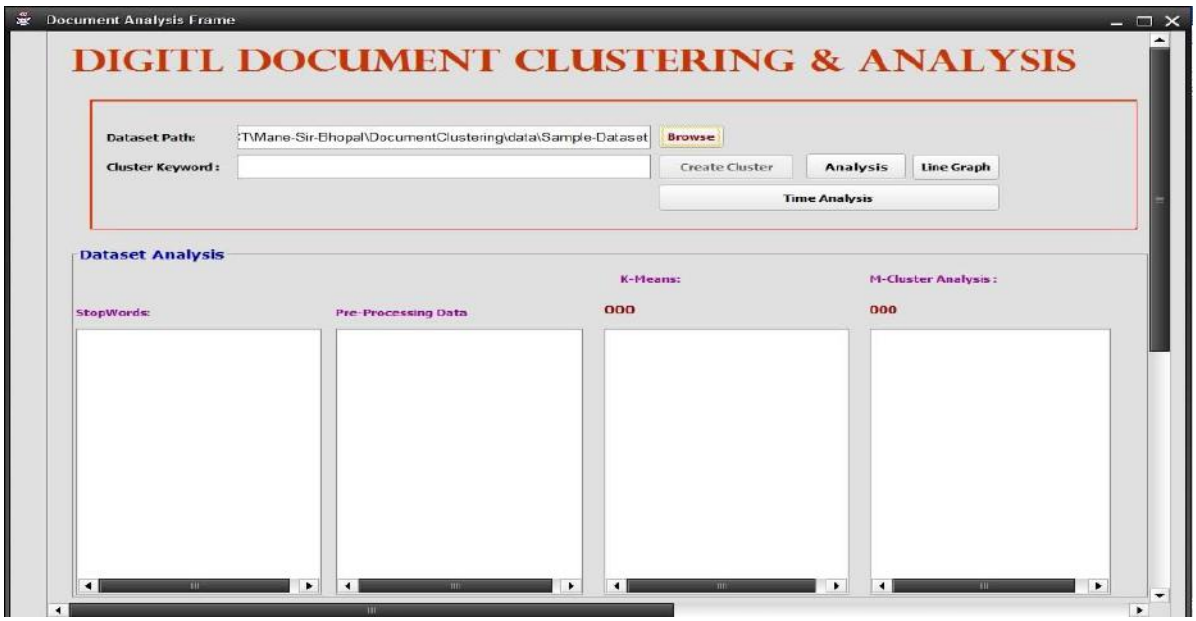

Figure 3.1: GUI of Proposed System

So, after selecting path user gives input query keyword for searching user relevant documents or clustering files according to their requirements using clicked on create cluster button .On this GUI there are three buttons analysis Line graph and Time Analysis we will see working of that in next figure.

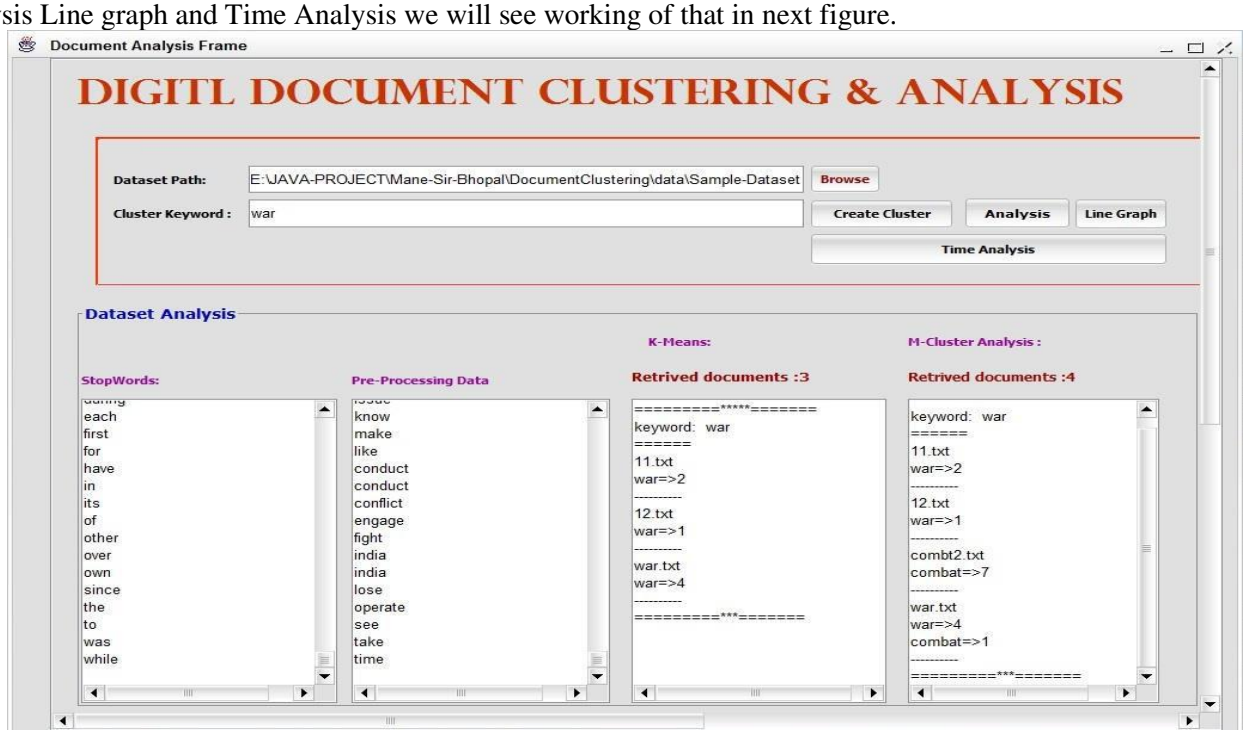

Figure 3.2: Result after Pre-processing and Clustering

Above figure 3.2 shows pre-processing and clustering results after applying pre-processing and clustering results. We performed pre-processing on the documents to reduce the high dimensionality of the documents while keeping the necessary information to do proper text mining. Here pre-processing is done in three steps, namely stop-words removal, stemming and indexing.

In stop-words removal, common words such as pronouns (he, she, it), conjunctions (and, or, but) and prepositions (of, a, an, the, where) are removed as they do not convey any meaning and have no effect on the significant information. We show result of stop words removal process on the left side of GUI it shows list of stop words which are removed from documents.

Stemming reduces the words to their root form. For example the words 'thinker', 'thinking', 'thinks' are all stemmed to 'think'. In indexing a set of distinct terms acquired after stemming process is chosen and the weight of these terms is computed for every document. After that we shows pre-processed data results on GUI right side of stop words list from this data all stop words are remove and stemming also performed on that we get noise free data because of that searching speed of process will increase.

After that we show results of K-mean, and M-Clustering algorithm results in above GUI. Above that result we shows how many documents are retrieved after applied it. And study on that which clustering algorithm gives best result we will see this analysis in experimental results.

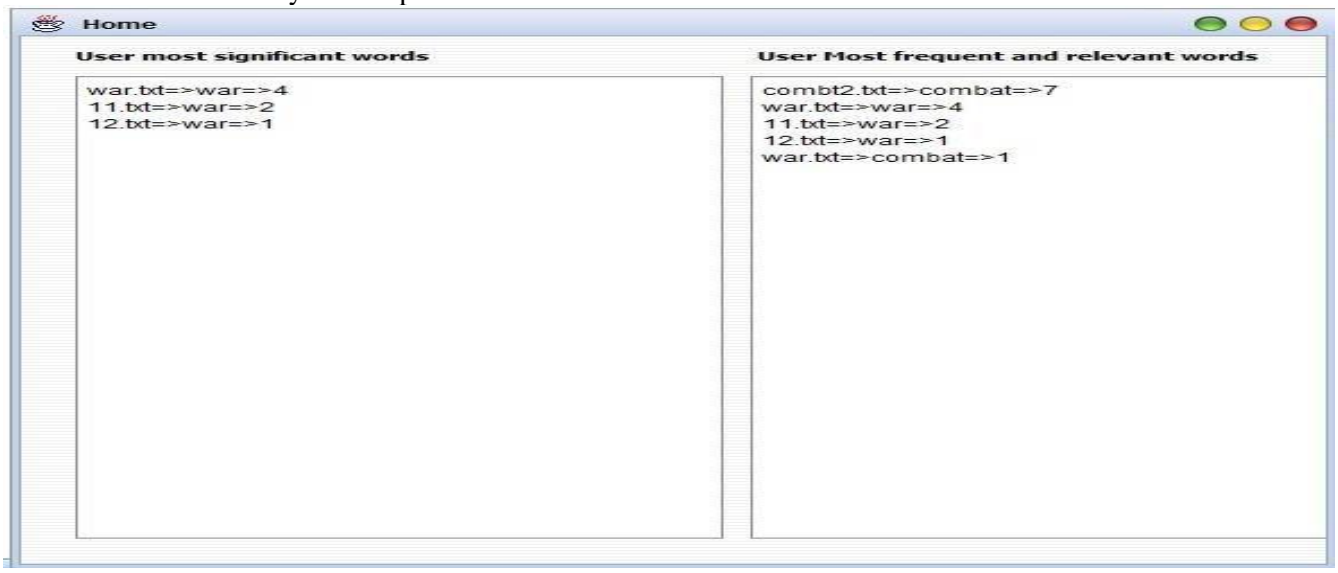

Figure 3.3: User Frequent Keyword Suggest to Examiner

Figure 3.3 shows final result of M-Clustering hybrid algorithm which is significant to the examiner .After pre-processing and clustering of documents, the system calculates the frequency of occurrence of each keyword in each pre-processed documents.

The top frequent keywords that are commonly appearing in the dataset are identified. We shows files which contains maximum number of war base word query it appears first which is important to the finding relevant documents. we shows two results in above figure such as examiner most significant word in that only base word search result depicted and another one is most frequent and relevant words means synonyms of base words also shown over here.

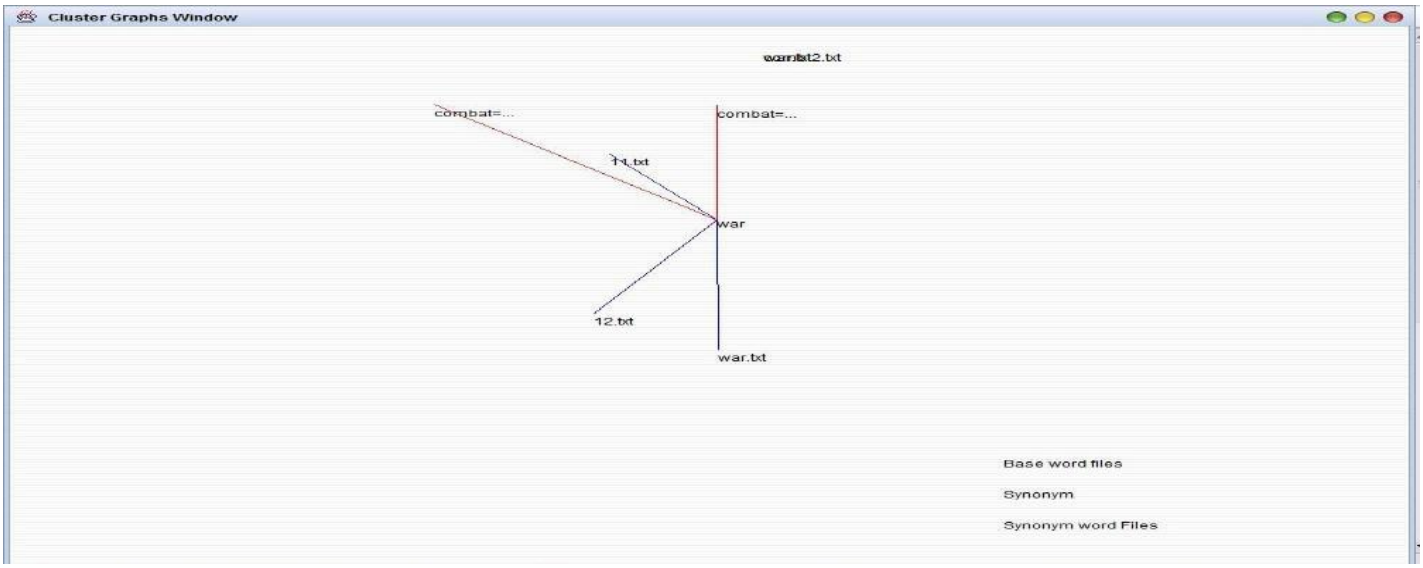

Figure 3.4: Clustering Graph

Figure 3.4 shows graphical view of final M-Clustering result in that we shows graph which represents clusters object after applying clustering algorithm. War base word is considered as centroid of cluster and all files represent nearest neighbours with minimum distance from it and we also shows synonym word file i.e. relative match of base word in files. Blue line represents base word files and red line represent synonym and green line represents synonym word files.

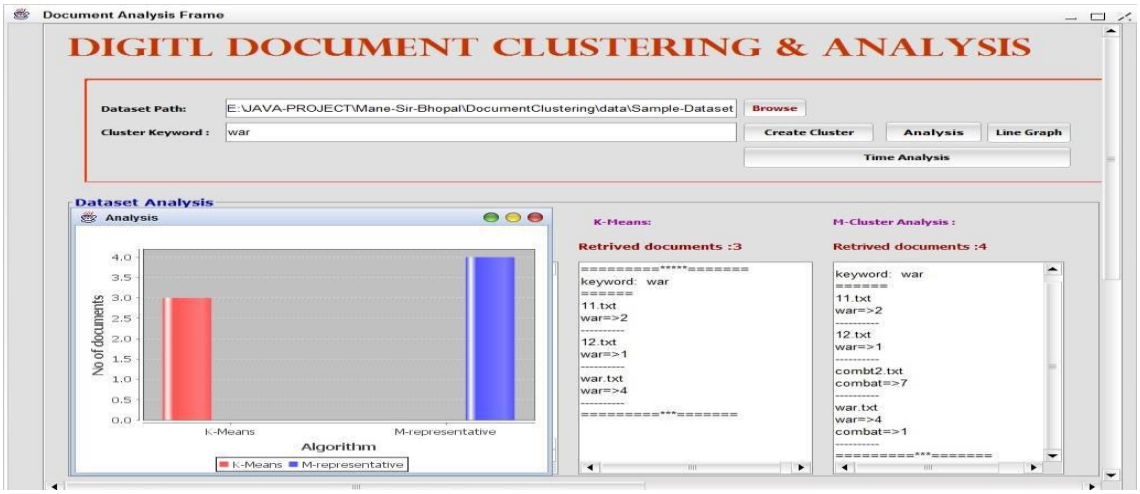

Figure3.5: Number of Retrieved Documents

Figure 3.5 shows bar chart of no of retrieved documents after applying two clustering algorithms and from that analysis we depicted that which clustering algorithm is efficient. In this analysis red bar shows K-mean algorithm result it retrieved 03 documents, blue bar shows M-Clustering algorithm result it retrieved 04 documents which are related to war in whole dataset which contain 47 files if we overlook result the M-Clustering we gives best result than K-mean algorithms.

The implemented work is can be improved document clustering quality by reducing the noise in the data by pre-processing the structure of data representation and also by applying new clustering techniques and also applying some existing algorithm. So, on the basis of that result we show how our new clustering method is better than that existing clustering method. We first study the use of pre-processing of data representation and then applying the standard clustering methods.

We then detect the effectiveness of a new clustering algorithm in which the noise was reduced by first clustering the features of the data and then clustering the data on the basis of their feature's clusters and by applying that algorithm we will get better clusters to improve the quality of document clustering for digital document analysis.

3.1 Dataset

We have used sample dataset for performing clustering result .In which we collected data related to war investigation in which we have taken 100 documents which contain some .txt files,.pdf files and .docx files in that dataset there are files which are related to murder, laws, corruption, crime etc.

3.2 Accuracy

Accuracy of the system is tested base keywords and synonyms. The proposed system is tested to search for different keywords and the precision, recall and f-measure are calculated for individual results obtained. To test the accuracy of the system experiments are conducted on demo dataset consisting of 100 documents from different categories.

3.3 Scalability

Scalability is ability of system in that we calculate how much time system takes for Clustering and preprocessing on different datasets. We calculate scalability of our performing system which shown in table.

4. CONCLUSION:

The implemented approach has applicable to clustering of text digital document only here scalability & time limit is also very important factor that arises when performing clustering on large dataset. So we conclude that it is barely possible to get a best general algorithm, which can be work in clustering of all types of datasets.

Thus we tried to implement enhance digital text clustering algorithms which can work well in categorical or numerical datasets. The implemented M-Clustering algorithm, suits the set of documents in which the required classes are related to each other and we require a strong basis for each cluster. Thus, this algorithm can be very effective in search engine like application**.**

REFERENCES:

1. M. R. Clint, M. Reith, C. Carr, and G. Gunsch, an Examination of Digital Forensic Models, 2003.
2. A. Kao and S. R. Poteet, "Natural Language processing and Text mining", Springer Verlag London Limited, 2007.
3. Y. Zhao, G. Karypis, and U. M. Fayyad, "Hierarchical clustering algorithms for document datasets", Data Mining Knowledge Discovery, vol.10, 2005.
4. Aggarwal, C. C. Charu, and C. X. Zhai, Eds., "Chapter 4: A Survey of Text Clustering Algorithms", Mining Text Data, New Springer, York, 2012.

5. D.Napoleon and P.Ganga Lakshmi, "An Enhanced K-means Algorithm to Improve the Efficiency Using Normal Distribution Data Points", International Journal on Computer Science and Engineering (IJCSE), vol. 02, issue 07, 2010.
6. G.Gandhi and R. Srivastava, "Analysis and implementation of modified K-medoids algorithm to increase scalability and efficiency for large dataset", International Journal of Research in Engineering and Technology (IJRET), Vol.03 Issue-06, Jun-2014.
7. K. Murugesan and J. Zhang, "Hybrid Bisect K-Means Clustering Algorithm", Department of Computer Science, University of Kentucky Lexington, USA.
8. R. Mundhe, A.Maind and R.Talmale, " Information Retrieval Using Document Clustering for Forensic Analysis" International Journal of Recent Advances in Engineering & Technology (IJRAET), Vol.2, Issue -5, 2014.
9. B. K. L. Fei, J. H. P. Eloff, H. S. Venter, and M. S. Oliver, "Exploring forensic data with self-organizing maps", Proceedings IFIP Int. Conf. Digital Forensics, 2005.
10. W.Liao, Y.Liu and A. Choudhary, "A Grid-based Clustering Algorithm using Adaptive Mesh Refinement", Appears in the 7th Workshop on Mining Scientific and Engineering Datasets 2004.
11. K. Stoffel, P. Cotofrei, and D. Han, "Fuzzy methods for forensic data analysis", IEEE International Conference Soft Computing and Pattern Recognition, 2010.
12. K. Nagarajan and Dr. M. Prabakaran ,"A Relational Graph Based Approach using MultiAttribute Closure Measure for Categorical Data Clustering", The International Journal Of Engineering And Science (IJES) ,Vol. 3, 2014.
13. H. Chen, W. Chung, Y. Qin, M.Chau, J.Xu, G.Wang, R. Zheng, and H. Atabakhsh. "Crime data mining: an overview and case studies", Proceedings of the 2003 annual national conference on Digital government research ,Digital Government Research Center, 2003 pages 1–5.
14. G. Forman, K. Eshghi, and S.Chiocchetti, "Finding similar files in large document repositories", KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, ACM, New York, NY, USA, 2005.
15. A.B.Schatz and G.Mohay, "A correlation method for establishing provenance of timestamp in digital evidence", Digital Investigation, volume 3, supplement1, 6th Annual Digital Forensic Research Workshop, 2006, pp. 98–107.
16. T. Abraham, "Event sequence mining to develop profiles for computer forensic investigation purposes", ACSW Frontiers '06: Proceedings of the 2006 Australasian workshops on Grid computing and e-research, Australian Computer Society, Australia, 2006, pp. 145–153.
17. J.G.Clark and N.L.Beebe, "Digital forensics text string searching: Improving information retrieval effectiveness by thematically clustering search results", In Digital Investigation, vol.4, 6th Annual Digital Forensic Research Workshop, 2007, pp. 49–54.
18. G. Thilagavathi and J. Anitha, "Document Clustering in Forensic Investigation by Hybrid Approach", International Journal of Computer Applications, vol. 91, April 2014.
19. L.F.D.C Nassif and E.R. Hruschka, "Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection", IEEE Transactions on Information Forensics and Security, vol.8, issue 1, January 2013.