ISSN:0975-9646

A New Approach for Improving Computer Inspections by Using Fuzzy Methods for Forensic Data Analysis

P. Jyothi[1], Dr S. Murali Krishna[2]

[1](M.Tech) Department of Computer Science & Engineering,
Madanapalle Institute of Technology &Sciences, Madanapalle, Andhra Pradesh, India

[2]Professor and Head of the Department, Department of Computer Science & Engineering,
Madanapalle Institute of Technology & Sciences, Madanapalle, Andhra Pradesh, India

***Abstract*- Now a day's digital world data in computers has great significance and this data is extremely critical in perspective for upcoming position and learn irrespective of different fields. Therefore we assessment of such data is vital and imperative task. Computer forensic analysis a lot of data there in the digital campaign is study to extract data and computers consist of hundreds of thousands of files which surround shapeless text or data here clustering algorithms is of plays a great interest. Clustering helps to develop analysis of documents under deliberation. This document clustering analysis is extremely useful to analyze the data from seized devices like computers, laptops, hard disks and tablets etc. There are six algorithms used for clustering of documents like K-means, K-medoids, single link, complete link, Average Link and CSPA. These six algorithms are used to cluster the digital documents. Existing algorithms are operated on single document at a time. In the proposed approach of these working algorithm applied on multiple documents at single time. Now we use clustering algorithm named as Hierarchical Agglomerative Clustering algorithm which gives better result compared with existing techniques. This algorithmic process will estimate the number of clusters in input documents and it takes less time for analyze the clusters in forensic analysis; they also find similarity between all documents in clusters.**

***Keywords*- Clustering, Forensic Analysis, Clustering Algorithms, Hierarchical Agglomerative Clustering Algorithm.**

I. INTRODUCTION

In computer forensic process is impacted by large amount of data. This has roughly distinct as restraints that merge element of law and computer sciences to gather and examine information from computer systems. In our study there are hundreds of files are there in instructed format. For this analysis they have some methods like machine learning and data mining are of great importance. Clustering algorithms are usually needed to grouping data in files, where there is practically no prior knowledge about the information [1] [13]. From a more specialized perspective, our datasets comprise of unlabeled objects. In addition, actually expecting that named datasets could be accessible from previous analysis, there is very nearly no hope that the same classes would be still legitimate for the upcoming information, got from different computers also related to different examinations. More definitely, it is likely that the new information would come from different locations. In this way, the utilization of clustering algorithms, which are fit for discovering latent patterns from content documents found in seized computers, can improve the analysis performed by the expert examiner. The methods of rational clustering algorithm objects within a substantial group are more like one another than they are two objects belongs to alternative group [1]. Along those data partition has been actuated from data. The export examiner may concentrate on interesting on delegated documents from the obtained set of groups by performing this task of examination of all documents. In a more functional and sensible situations, domain experts are rare and have limited time accessible for performing examinations. Therefore it is sensible to expect that finding a significant document. The examiner could prioritized the investigation of different documents belongs to the cluster interest.

Clustering algorithm has been mulled over for a considerable length of time and the literature on the subject is huge. Therefore, we decided to demonstrate the capability of the proposed methodology, namely: the partition algorithms K-Means, K-Medoids, the hierarchical single link, complete link, average link and the cluster ensemble algorithm known as CSPA[3].It is well known that the number of clusters demonstrating parameter of many algorithms and it is generally having an earlier knowledge. However the number of clusters has not been examined in the computer forensics. Really we could not even spot one work that is sensibly close in its application area and that reports the utilization of number of algorithms capable finding the number of clusters[3].

II. REVIEW OF RELATED RESEARCH

In our software development process research is the most important one. In this is based on the time factors, economy and company strength we can determine the developing process. Once the programmer start the work based on experts suggestions and gather related information to different websites based on their work. Before building the system each and developer can maintain the above requirements report.

B. Fei*et al*. [3] have discusses the application of a self-organizing map (SOM) is to support decision making by computer forensic investigators and assist them in conducting data analysis in a more efficient manner and also SOM produces patterns similarity in data sets. Author explores great ability to interpret, explore data generated by computer forensic tools.

Alexander Strehl et al. [2] has introduces three effective and efficient techniques to obtaining high quality combiners. In first combiner induce Partitioning and re-clustering of objects is based on similar measure, second is based on hyper-graph and third is based on collapse group of clusters into meta-clusters which participate to find individual object to the combined clustering. By using the three approaches to provide the low computational costs and feasible to use a supra-consensus function against the objective function and provide the best results.

L. F. Nassif et al. [5] have present an approach that applies document clustering algorithms to forensic analysis of computers seized in police examinations. Author represents experimentation with six well-known clustering calculations (K-Means, K-medoids, Single Link, Complete Link, Average Link, and CSPA) applied to five certifiable datasets acquired from computers seized in true examinations. Investigations have been performed with different combination of parameters, resulting in 16 different instantiations of algorithms. Moreover, two relative legitimacy records were utilized to consequently appraise the quantity of clusters. In the event that suitably introduced, partition algorithms (K-means and K-medoids) can likewise respect great results.

Ying Zhao et al. [8] have use high quality clustering algorithms play an essential part in giving intuitive navigation and browsing mechanism by sorting large amount of data into a little number of meaningful clusters. Specifically clustering algorithm, hierarchical clustering that assemble meaningful hierarchies out of large amount of accumulations. This all concentrates on document clustering algorithm that manufactures such various leveled solutions.

A. Hierarchical Agglomerative Algorithm

Hierarchical clustering algorithms are top-down and bottom up. Bottom-up algorithms treat each one record as a single cluster at the beginning and after that progressively merge (or agglomerate) sets of clusters until all groups have been merged into a single group that contains all documents. Bottom-up hierarchical clustering is in this way called hierarchical agglomerative grouping or HAC. Top-down clustering requires a technique for splitting a group. It continues by splitting clusters recursively until individual documents are arrived. In the Hierarchical Agglomerative Algorithm assumes each one document as a solitary of cluster at beginning and after that progressively agglomerative pair of clusters into single group of clusters have been agglomerate into single group that contain similar type of documents.

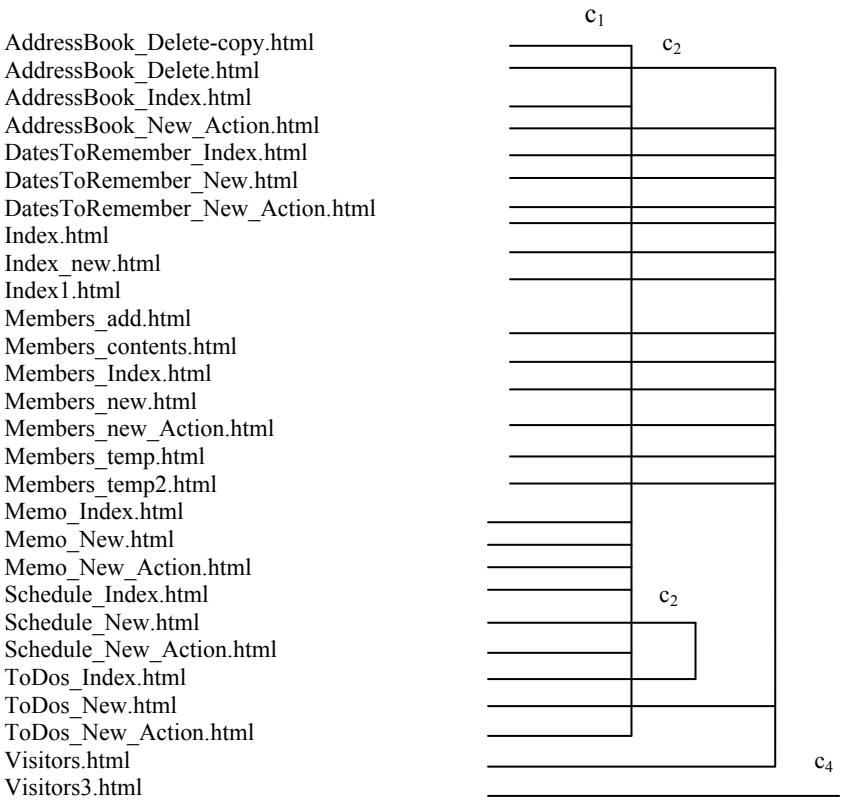

Fig. 1 A diagram shows the clustering of 28 documents.

a) Cluster Formation:

By passing up from the bottom to top process of clusters the dendrograms used to reproduce the historical environment of consolidations that brought about the represented clustering. For example, we see that the two documents permitted AddressBook_Delete-copy.html and AddressBook_New_Action.html were combining in Figure (1) and that the last combine added Visitors.html to a cluster comprising of the other 27 documents.

b) ***Finding Similarity:***

The Hierarchical Agglomerative clustering is usually defined as a dendrograms as illustrate in Fig (1). Each combination is signified by a horizontal line. That line represents similarity between two documents, where documents are viewed as single clusters. We call this similarity the combination similarity of the combined documents. For example, the combination similarity of the documents Schedule_New.html and ToDos_Index.html consisting of Figure (1) is ≈ 0.19 . This defines the cosine similarity of clusters as 1.0.

III. FRAMEWORK REQUIREMENTS

For finding the meaningful data from the dataset, researchers have used data mining techniques, in which clustering is one of the popular techniques. Let DS will taking as our dataset represented as DS={d_1, d_2, … d_n}; $1 \leq I \leq n$, where n is the number documents in a dataset DS. In our propose system basically there are three important steps which are as follows

1) Preprocessing
2) Cluster Formulation
3) Forensic analysis

A. Preprocessing: In preprocessing step there are three steps such as a) fetch a file contents, b) Stemming, c) Stop word Removal. These 3 steps are used to remove the noise and inconsistent data. In first step fetch the dataset and perform the second operation with the help of porter stemming. In this stemming is based on the idea that the suffixes in the English language are mostly made up of a combination of smaller and simpler suffixes. If the words end with ed, ing, lyetc that words are removed. This step is a linear step stemmer[16]. In this last step is remove the stop words with the help of Stop token filter.[17] Stop words in a document like to, I, has, the, be, or etc. stop words are the foremost frequent words with in the English language. Stop words blot your index while not providing any additional worth. At that point, we received a customary statistical methodology for text mining, in which documents are meant in a vector space model. In this each one model, each one document is denoted by vector containing the frequencies of events of words. To process the distance between reports, two measures have been utilized specifically: cosine-based separation and hierarchical agglomerative clustering. After these steps our data will be relevant.

B. Cluster Formulation: This session exhibits the mining of datasets from the preprocessed dataset. For each document the similarity of the concentrated words from the preprocessed step is processed and the top comparability documents are clustered first this. This session depicts the mining of successive item sets from the preprocessed content documents. For each document the recurrence of the concentrated words from the preprocessing step are registered and the top continuous words from each are taking out. From the set of top frequent words, the binary database is framed by getting the unique words.

Hierarchical Agglomerative Clustering Algorithm:Hierarchical agglomerative algorithms treat every one document as a singleton cluster toward the starting and thereafter dynamically consolidation set of clusters until all clusters have been melded into a single cluster that contains all documents.

Input: List of Documents D=d_1,d_2… d_n

Output: Clusters resultC= {c_1, c_2…c_n }

1. For i=1 to n do
2. For the given list of documents each document is treated as a specified
3. Finding parsers //those are theunique words in documents
4. Suppress non-dictionary words
5. Get unique edges in this documents
6. Initialize clusters

a. For n←1 to N

b. Applying clustering to the items

Constructing histogram //for analyzing clusters

h_{min} should be 1.0; h_{max} should be 0.0

For T to 1←n-2

For J to 1←n-1

t_{sim}=sim(doc[t], doc[j]) { If(h_{min}>t_{sim}) h_{min}=t_{sim}; If (h_{max}<t_{sim}) h_{max}=t_{sim}; }

7. Finding analogousity.

c) **Forensic Analysis:**

This will be the last step in proposed method. Here the algorithm process initially provides a topological arrangement between neurons at convergence of documents. Here we can analyze the number of clusters from our selected dataset. At final step this process will calculate the similarity between formed documents with less time compared to other algorithms.

IV. IMPLEMENTATION

The implementation process of clusters can done through number of steps that process will needed for the purpose of good cluster similarity between clusters. The follower can do these steps very care full. In this process first collect the documents from local systems then perform preprocessing- In preprocessing step there are three steps such as a) fetch a file contents, b) Stemming, c) Stop word Removal. These 3 steps are used to remove the noise and inconsistent data. In first step remove the stop word prepositions, pronouns, irrelevant documents data(a, an ,the etc)[17] and later on to do stemming on that file which will be removing Portuguese words(ing and edetc)[16] from the upcoming data. At that point, we received a customary statistical methodology for text mining, in which documents are

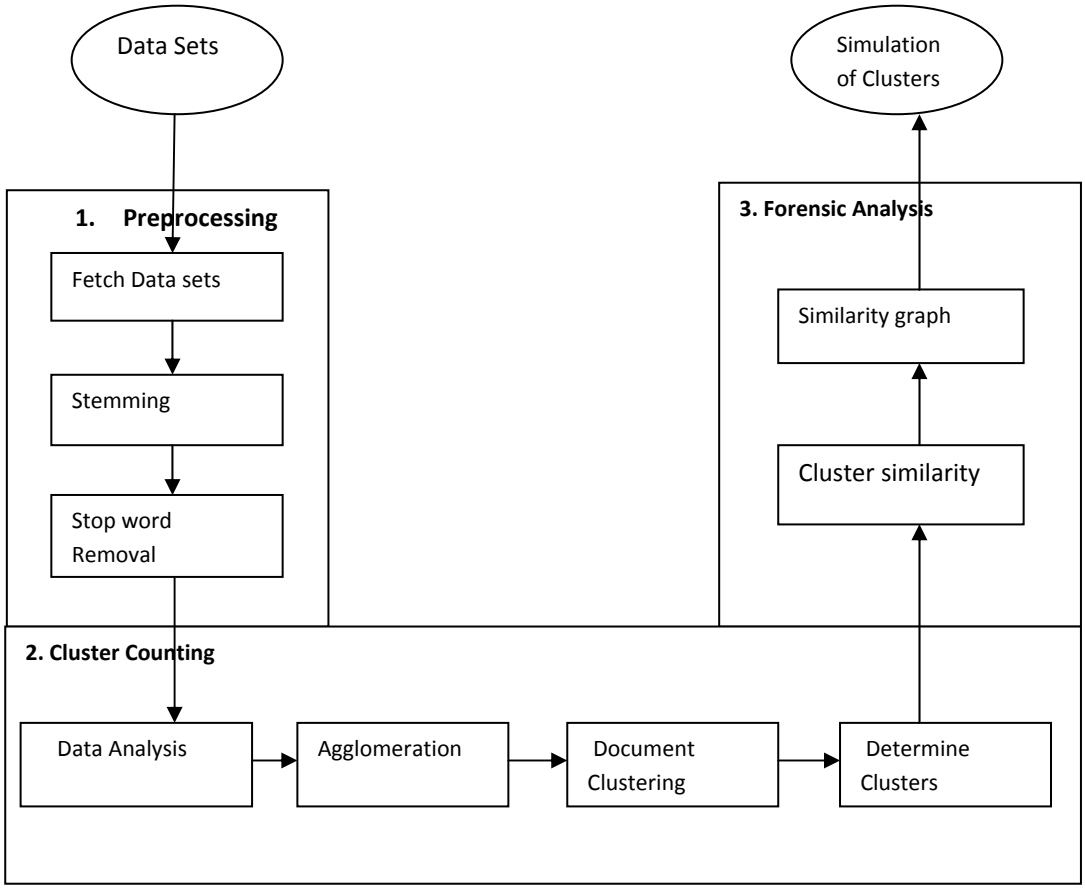

Fig. 2 Architecture implementation of clustering process.

meant in a vector space model.[19] In this model, each one document is denoted by vector containing the frequencies of events of words. To process the distance between reports, two measures have been utilized specifically: cosine-based separation and hierarchical agglomerative clustering. After these steps our data will be relevant.

The next step of this process will take clusters counting. Here, data will collect from the previous step. Then analyze data for estimating relevant clusters, by using agglomerative algorithm. In this algorithm data is analyzes from multiple documents, and then divide similar documents and dissimilar documents. Based on the priority of data high priority documents are saved under one cluster and comparison of first cluster less similarity documents are stored under next cluster based on the algorithmic perspective. Then, the last step algorithm finds the similarity of all cluster consisting documents. In this comparison include clusters containing each and every document is compared and finds similarity between them. This algorithm plays a important role in this process compared to other algorithms.

V. EXPERIMENTAL RESULTS

In this proposed approach experimentation developed by java (JDK 2.0). In the process of running and executing the main file. After executing main file dataset containing documents are loaded that are shown in figure (3). Here we are taking 26 documents [0-25] these all documents are under 8 different areas. Like Address Book [0-2], Members [8-14] etc those are shown in figure (4). These 8 different partitions are clustered into 4 groups named as C_1, C_2, C_3, C_4 . In C_1 under documents are [0, 1, 7, 8, 9, 11, 15], C_2 under documents are [2, 3, 4, 5, 6, 13, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24], C_3 under documents are [10, 12] and C_4 under documents are [25] shown in figure (5) and finding similarity between all documents shown in figure (6).

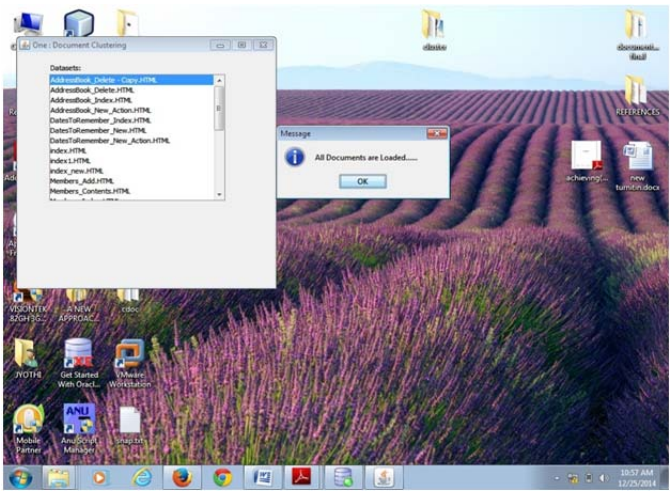

Fig. 3 Loading documents.

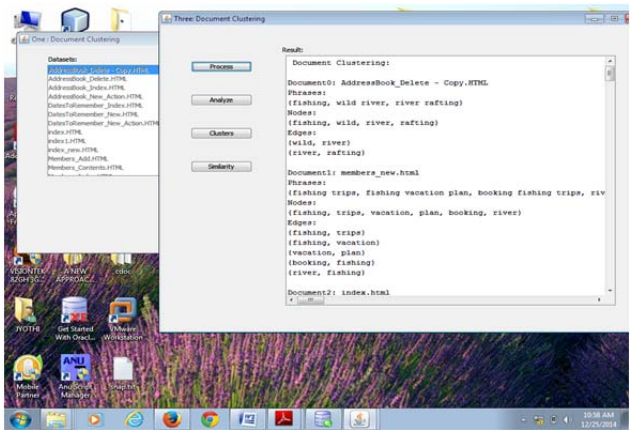

Fig. 4 Processing input documents

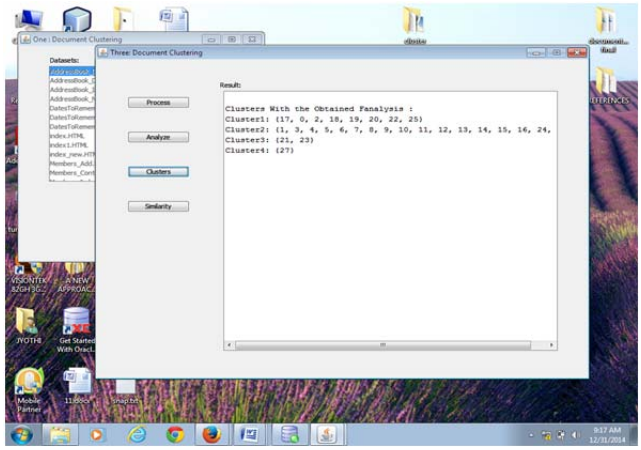

Fig. 5 Cluster Analysis.

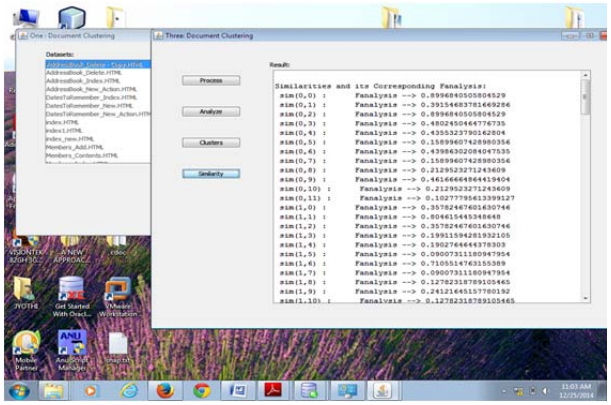

Fig. 6 Findingsimilarity between documents.

VI. CONCLUSION

We use an approach for clustering documents which can become an ideal application forensic analysis of computers. There are several practical results based on our work which are extremely useful. In our work, the algorithm known as Hierarchical Agglomerative Clustering algorithm that yields the best results. In spite of this algorithm we find the number of clusters in our input documents and finding the similarity between the documents.

REFERENCES

[1] B. S. Everitt, S. Landau, and M. Leese, Cluster Analysis. London,U.K.: Arnold, 2001.

[2]. S. Haykin, Neural Networks: A Comprehensive Foundation. EnglewoodCliffs, NJ: Prentice-Hall, 1998.

[3]. A. Strehl and J. Ghosh, "Cluster ensembles: A knowledge reuse framework for combining multiple partitions," J. Mach. Learning Res., vol.3, pp. 583–617, 2002.

[4]. B.K.LFei, J. H. P. Eloff, H. S. Venter, andM. S. Oliver, "Exploring forensic data with self-organizing maps," in Proc. IFIP Int. Conf. DigitalForensics, 2005, pp. 113–123

[5]. K. Stoffel, P. Cotofrei, and D. Han, "Fuzzy methods for forensic data analysis," in Proc. IEEE Int. Conf. Soft Computing and Pattern Recognition, 2010, pp. 23–28.

[6]. L. F. Nassif and E. R. Hruschka, "Document clustering for forensic computing: An approach for improving computer inspection," in Proc.Tenth Int. Conf. Machine Learning and Applications (ICMLA), 2011.

[7]. S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, andR. Zunino, "Text clustering for digital forensics analysis," Computat.Intell. Security Inf. Syst., vol. 63, pp. 29–36, 2009.

[8]. Y. Zhao, G. Karypis, and U. M. Fayyad, "Hierarchical clustering algorithms for document datasets," Data Min. Knowl. Discov., vol. 10, no. 2, pp. 141–168, 2005.

[9] Y. Zhao and G. Karypis, "Evaluation of hierarchical clustering algorithms for document datasets," in Proc. CIKM, 2002, pp. 515–524.

[10] Y. Loewenstein, E. Portugaly, M. Fromer, and M. Linial, "Effcient algorithms for exact hierarchical clustering of huge datasets: Tackling the entire protein space," Bioinformatics, vol. 24, no. 13, pp. i41–i49, 2008.

[11] B. Mirkin, Clustering for Data Mining: A Data Recovery Approach.London, U.K.: Chapman & Hall, 2005.

[12] L. Hubert and P. Arabie, "Comparing partitions," J. Classification, vol.2, pp. 193–218, 1985.

[13] C. M. Bishop, Pattern Recognition and Machine Learning. NewYork: Springer-Verlag, 2006.

[14] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[15] K. Kishida, "High-speed rough clustering for very large document collections,"J. Amer. Soc. Inf. Sci., vol. 61, pp. 1092–1104, 2010, doi:10.1002/asi.2131.

[16]. Daniel Waegel. "The porter Stemmer".CISC889/Fall 2011.

[17]. Swatantrakumarsahu and NeerajSahu, G.S,Thakur, "Classification of Document Clustering Approach". Volume 2, issue 5, may 2012.