ijecs open access www.ijecs.in
*International Journal Of Engineering And Computer Science ISSN: 2319-7242*
*Volume 4 Issue 9 Sep 2015, Page No. 14235-14238*

# Document Clustering and Automatic Labeling for Forensic Analysis Using High Performance Clustering Algorithm

*Asmita V. Mane,[1], Prof. Gitanjali Shinde[2]*

[1]Smt.Kashibai Navale college of Engineering, Savitribai Phule Pune University,
Pune
*asmitamane43@gmail.com*

[2]Smt.Kahibai Navale College of Engineering, Savitribai Phule Pune University,
*Pune*
*Gr83gita@email.com*

**Abstract: Forensic document the strategy for examination of different unlawful acts by computer based techniques is called as computerized scientific investigation . A huge number of records are generally inspected in computer forensic investigation. The greater part of the information in that advanced records are fundamentally unstructured. To handle or investigation of this expansive, unstructured information by inspectors is hard to be performed and drawn out. Many calculations or algorithms are applied for document clustering can encourage the revelation of new and helpful learning from the archives under examination. Clustering algorithm are key piece of this work in which number of unstructured records are given as an input and the yield is organized document position. We exhibit a methodology that applies clustering algorithm of documents to measurable examination of seized computers in examinations. One particular calculation is not a Cluster examination itself but rather the general undertaking to be explained. Different algorithms are utilized that essentially vary as a part of their idea of what they constitutes a group. We characterize the proposed methodology with K-Means algorithm and hierarchical algorithm. Likewise we can create number of clusters dynamically and cluster labeling . The execution of computer for examining a few documents is enhanced in our test approach.**

**Keywords: T**ext: Clustering, Forensic Analysis, Unstructured Documents, Clustering Algorithm.

## 1. Introduction

In Forensic Investigation handle in which the advanced gadget like computers are utilized to break down the computerized proof those are truths which are under the investigation. The computerized evidence are the advanced information which underpins the occurrence theory. Volume of information of advanced world expanded from 161 hex bytes to 988 hex bytes around 18 times the measure of data present in every one of the books ever composed—and it keeps on growing exponentially. Examination of this reports is troublesome errand of the computerized legal the investigation of different records having a place examination process. This substantial measure of information straightforwardly affect on computer criminology. It normally more unpredictable to analysing in light of the fact that it is unstructured. It is more mind boggling to inspecting a huge number of records if the quantity of archives are huge. The investigation of extensive measure of information surpasses the master's capacity of examination and understanding of information. In this way, automated document analysis, similar to those broadly utilized for machine learning and information mining, are of principal significance.

Clustering algorithms are utilized for exploratory information examination, where there is almost no earlier learning about information. Clustering algorithms are utilized as a part of procedure of advanced measurable investigation. Clustering algorithms are utilized as a part of procedure of advanced legal examination. These techniques are utilized to change over unstructured archive to organized report. This is definitely the case in a few utilization of Computer Forensics, more specialized perspective, our data sets comprise of unlabelled articles the classes or classifications of archives that can be found are obscure priory. Additionally, accepting that named datasets could be accessible in past investigations. From this, the utilization of clustering algorithms, which is equipped for discovering inert examples from content reports found in seized computers, can upgrade the investigation performed by the master inspector.

The grouping calculations is that protests inside of a substantial bunch are more like one another than articles fitting in with an alternate group. Along these lines, once the information allotment has been made from information, the master analyst may at first concentrate on surveying agent records from the acquired set of clusters. After this preparatory investigation, analyst might inevitably choose to investigate different records from every cluster. By doing this, one can keep away from the hard undertaking of looking at all archives in any case, regardless of the possibility that so craved, it still should be possible. As a general rule space specialists are rare and have restricted time for performing examinations along these lines, it's sensible to accept that, subsequent to discovering a pertinent record, the inspector could organize to the group of premium, in light of the fact that it is likely that these are likewise important to the examination. Such approach of document, clustering, can in reality to enhance the investigation of seized computers;

DOI: 10.18535/ijecs/v4i9.29

## 2. Related Work

In this section we review work done on forensic analysis in the literature. The computerized world is growing quickly. An overview and estimate of overall data development is worth thought. Information storage, systems administration and security is a major test for always expanding advanced information. Data security and protection assurance will turn into a meeting room concern as associations what's more, their clients turn out to be progressively entwined continuously. This will require the execution of new security advancements addition to new training, policies, and procedures.

The essential idea of bunch, parameters needed for grouping, different bunching strategies are decisively expounded in [3]. In the book[3], the writer has extremely very much talked about the fundamental idea of clustering, recognition of clustering graphically. A brief learning about vicinity routines alongside various levelled, Partitional and probabilistic systems is likewise clarified in subtle element. The creator has additionally given rules to genuine execution.

Cluster Ensembles - The Knowledge Reuse Framework for Combining Multiple Partitions.[5]. This paper display the issue of joining numerous dividing of set of articles into a solitary solidified clustering without getting to the elements or calculations that stop mined these partitioning's. Presented the cluster ensemble issue and to take care of this issue gave a three viable and productive calculations. It characterize a common data based target work that empowers to naturally choose the best arrangement from a few calculations and to manufacture asupra-agreement work also

Text Clustering for Digital Forensics Analysis.[6] Present an effective digital text analysis strategy, based on clustering based text mining techniques, is introduced for investigational purposes. It gives an overview on the possibilities offered by textual clustering when applied to Digital Forensics analysis.

Term-Weighting Approaches In Automatic Text Retrieval[7]. This article shows the insights gained in automatic term weighting and provides single term indexing with which other more elaborate content analysis procedures can be compares.[7].

Fuzzy methods for forensic data analysis [8]: In this paper depict a strategy and a programmed method for construing exact and effectively justifiable master framework like standards from forensic information. This approach is in view of the fuzzy set hypothesis.

Exploring forensic data with self-organizing maps: SOM examines the use of a self-organizing map guide (SOM), an unsupervised learning neural system demonstrate in a more effective way. The paper investigates that how a SOM can be utilized as a premise for further examination furthermore, it shows how SOM perception can give investigator more prominent capacities to translate and investigate information created by computer criminological tools.

## 3. Proposed Scheme

Figure 1 shows the system architecture of proposed system. In previous system six well known algorithm used for clustering. Clustering by this algorithm efficient but difficulty when examiner searches for particular file. There is no any label to a clusters so didn’t get a which cluster contain which file. So in proposed system used labeling method for clustering. In this approach use a K-means with vector space model and weighted hierarchical algorithm for clustering database seized in police investigation. After clustering system gives labeled to cluster according to high term frequency (word count). In proposed system basically there are four important steps which are as follows:

1). Preprocessing,
2). Estimation of Clusters,
3). Cluster Merge,
4). Labeling.

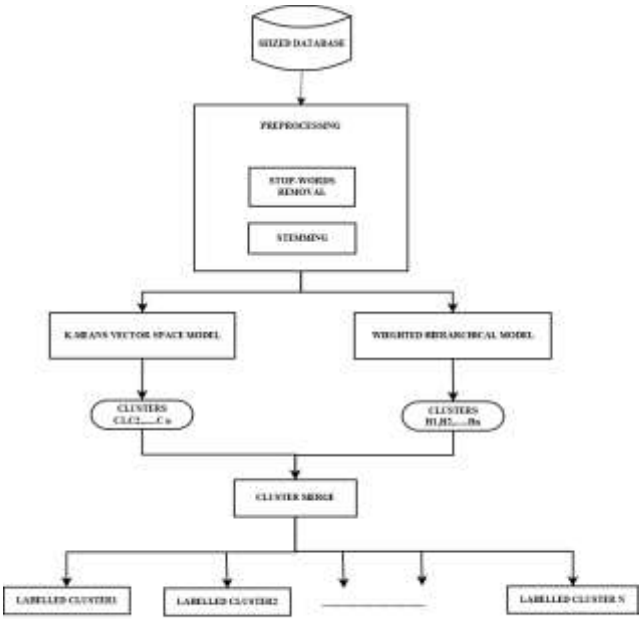

**Figure 1**: System Architecture.

**Preprocessing:** Preprocessing is performed get clean plain text document**.** After converting particular formatted file in plain text it gives to get preprocessed. In preprocessing there are two steps as follows

1)Stopwords Removal : All stopwords like is, are, the, it etc. are removed from documents using stop word library maintained in system.

2)Stemming: Stemming is a process of getting base or root word after removing ing, ed etc. from word. Snowball Stemming algorithm is used.

**Estimation Of Cluster:** There number of algorithms are available for clustering. That clustering algorithms give efficient output. In proposed system choose k-means clustering algorithm with vector space model (VSM) and weighted hierarchical clustering algorithm for large database. In k-means we use a VSM to get accurate centroid. VSM take product of TF and IDF as a coordinate.

**Cluster Merge:** Sometimes clusters from above algorithms are not accurate or getting same type of cluster from both algorithm. So for accuracy we merge a two same type of clusters together. Because of this get a one accurate cluster. System use cosine distance between clusters getting from k-means and hierarchical clustering.

**Labeling:** After getting accurate clusters labeling performed. For labeling term frequency used to give name for particular cluster. From all this stuff examiner get accurate cluster with label.

In this way examiner get great ability to examine documents efficiently and effectively.  There is less time consumed for forensic analysis. Fig. 2 shows flow of system.

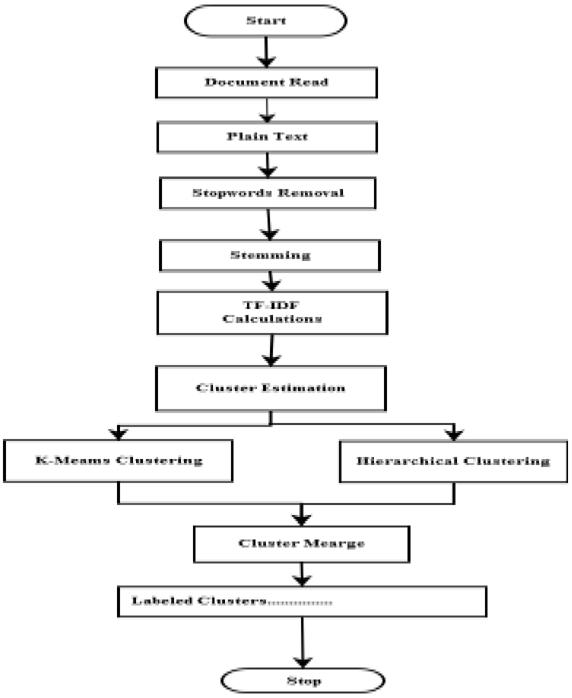

**Figure 2 :** System Flowchart.

## 4. Result

Fig 2 shows the final result of proposed system. In figure shows the how many clusters to be formed by system for given database.

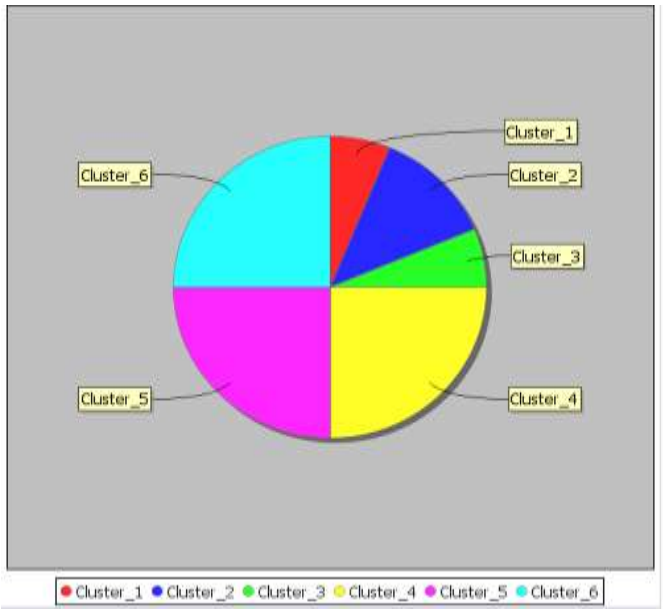

Fig 2: Clusters Formed by System.

In fig 3 shows the time consumption analysis by k-means , hierarchical and proposed  system.

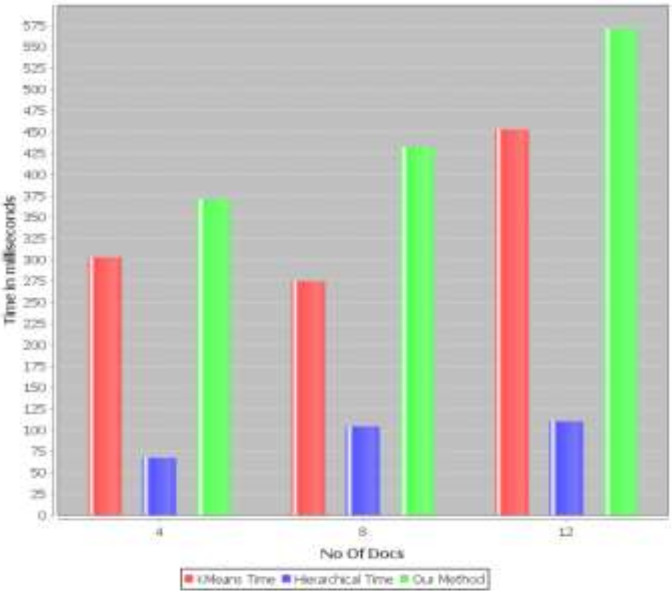

Fig 3: Time Consumption Analysis.

## 5. Conclusion

The result of the data recovery utilizing archive clustering for forensic investigation in this proposition is the quantity of labeled cluster, which gives the better representation as edge which demonstrates the most applicable information show in the specific cluster. The task of assignment of labels to clusters empowers the forensic inspector to distinguish the substance of every cluster all the more rapidly—in the long run even before analyzing their substance. K-mean calculation is successful; it gives good output and does not require the client to determine numerous parameter.  Weighted algorithm help in large database where k-means less efficient. Both algorithm helps in get accurate clustering. The automatic labeling methodology gives the quick and proficient investigation, decrease manual work, as framework examinations all things the proofs assembled will be more exact, additionally build the execution of criminological investigation with rate up the computer inspection process.

## References

[1] Luís Filipe da Cruz Nassif and Eduardo Raul Hruschka, “Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection” , IEEE Transactions On Information Forensics And Security, Vol. 8, No. 1, January 2013Shamir, Adi. "How to share a secret." Communications of the ACM  (1979) VOL:22, NO:11

[2] J. F. Gantz, D. Reinsel, C. Chute, W. Schlichting, J. McArthur,  S. Minton, I. Xheneti, A. Toncheva, and A. Manfrediz, “The expanding digital universe: A forecast of worldwide information growth through 2010,” *Inf. Data*, vol. 1, pp. 1–21, 2007.

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

[4] L. Kaufman and P. Rousseeuw, Finding Groups in Gata: An Introduction to Cluster Analysis. Hoboken, NJ: Wiley-Interscience, 1990.U.K.: Arnold, 2001.

[5] 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.

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

[7] G. Salton and C. Buckley, “Term weighting approaches in automatic text retrieval,” Inf. Process. Manage., vol. 24, no. 5, pp. 513–523, 1988.

[8] 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.

[9] B. K. L. Fei, J. H. P. Eloff, H. S. Venter, andM. S. Oliver, “Exploring forensic data with self-organizing maps,” in *Proc. IFIP Int. Conf. Digital Forensics*, 2005, pp. 113–123