Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Elevating Document Clustering For Forensic Analysis Investigation System

Vilas V Pichad*, Asst. Prof. Sachin N Deshmukh
Department of CS & IT, DR. BAMU,
Aurangabad, India

Abstract— Current digital word technologies information in computer world, there is extremely large increase in crime like money laundering, drug trafficking, unauthorized access, ethical hacking, fraud detection in different domain etc. So investigation of such cases deserves a much more important, in computer devices plays a very important role. In this document clustering algorithms for digital forensic analysis of computers seized devices play important role in real word investigation case conducted by police investigations. Document clustering for forensic analysis is used to study the source and content of various messages as evidence, identifying the actual criminal with the help of related evidence, etc. to collect credible evidence to bring criminals to justice.
This paper presents the results of an experimental study of some common document clustering techniques. In particular, we compare the approaches using document clustering algorithm. In the Digital forensic analysis of investigation, there are total three well-known document clustering algorithms are used. It has been implemented to be used with k-means, Average link and complete link algorithm. This is applied datasets for five real-world investigation cases conducted by the Brazilian Federal Police Department. Partational k-means algorithm provides good result but when we change the centroid at every time result will be different. Average link and complete link provides the better results. This result different existing text clustering and Document clustering multithreading methods is used with computer forensic analysis.

Keywords— Digital forensic analysis, partitional clustering algorithm, hierarchical clustering algorithm, Document clustering.

I. INTRODUCTION

There is explored and utilize the huge amount of text documents is a major question in the areas of information retrieval and text mining. Document clustering (referred to as text clustering) is one of the most important text mining methods are used, to help the organizing a large amount of documents into clusters. Today there is fast increase in crime relating cases to Internet and Computers device has caused a growing need for computer forensics. Computer forensic is analysing huge number of files from computer seized devices, which is computer forensic tools, can be exist in the form of computer software. These tools have been developed to help computer forensic investigators in a computer investigation. But in computer forensic process all the information and files are stored and received in digital form. This digital information stored in computer seized devices has an important factor from an investigative perspective [1]. Which treated as evidence in the court of law to prove what occurred based on such evidences. So collection of evidences from digital devices is also key task of forensic analyst. When the forensic examiner collects all such evidences then his job is to establish based on the collected evidences. But such computer seized devices contains huge set of files and documents, so it is not easy to do the analysis of each and every files individually.

For performing examinations experts in particular domain have limited time, so to finding a specific document, the examiner could prioritize all the documents to the analysis. Such an approach, can indeed improve the forensic analysis of seized computers. At the same time court of law requires quick result of such cases to improve the speed of forensic analysis process has a more important. The clustering algorithms play important role in forensic analysis of digital documents encloses, complex and unstructured data, to improve such forensic analysis process requires fast text clustering and document clustering techniques. The process of analysing large volumes of data may consume a very large amount of storage in small time in the storage media. The police investigations forensic data of clustering algorithms are typically used for examining data analysis, where there is datasets consist of unlabelled objects the classes or categories of documents that can be found are a prior unknown, this is exactly the case in several applications of computer forensics [2]. For, the implementation of project is that the target of effective document clustering can be improved by dropping the noise in the data by pre-processing and the structure of data representation and also by applying some clustering techniques. We first study the effectiveness of pre-processing technique to remove the irrelevant data from the document, for the purpose used stop word removal and steaming technique. Then applying the partition and hierarchical clustering methods provides the best result. The partition k-means [3] clustering algorithm provides good results but when we change centroid every time result gets change. Hierarchical clustering algorithm average link and complete link [4] provide the best result. All the results are shown by graphical representation.

II. RELATED WORK

In past few years many Document clustering for forensic analysis field has been proposed lots of algorithm for clustering the documents. The algorithm for clustering the data is called by most of the study they used unsupervised learning of Gaussian Mixture Model [5] used to requires an a-priori selection of model order, name is the number of components to be incorporated into the model, Expectation-maximization (EM) [6], Self-Organizing Maps (SOM) based algorithms were used for clustering files with the aim of making the decision-making process performed easy task to the examiners arrange files more quickly [7]. Cluster the result form key word searches this kind of algorithm has been used in [8] order to increase information retrieval efficiency. There are unsupervised feature selection methods such as Term Variance (TV) dimension reduction technique also used document clustering and text clustering [9]. E-mail for forensic analysis explains classification and clustering algorithm presented in [10].Maximum of the e-mails is grouped by using structural, syntactic lexical and domain specific features explain in [11]. Recently Fuzzy C-means (FCM) and k-means [3] method for mining association rule from clustering method of forensic data analysis based on the fuzzy membership of each data point to each of the clusters of data formed. The objective of the fuzzy c-means algorithm is to minimize the sum of the weighted squared distances between the data points and the cluster centres [12].

Recently, Nassif and Hruschka[1] proposed an approaches of document clustering algorithm for forensic analysis in the computer seized devices also overcame the limitation of document clustering algorithm, also explain better cluster analysis technique[13] for the purpose estimating the number of cluster automatically they use the relative validity index criteria [14] which overcomes the limitations of previous techniques. They have tested the result of five dataset, which has real word investigation cases conducted by Brazil police department. For testing they have used six well-known clustering algorithm partition k-means, k-medoids algorithms, Cluster similarity partition algorithm CSPA [15] and Hierarchical single link, complete link and Average link algorithms.

III. RESEARCH METHODOLOGY

A. System Architecture

1) Pre-processing Step:

In Pre-Processing step there are three steps:

These 3 steps are used to remove the noise and inconsistent data. In first step select the dataset then remove stop word and perform the third 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.

a) Document Selection:

Documents selection are used to collect large text document from the document folder or various source for the purpose of clustering, including the processes crawling, filtering, indexing etc. which are used to store and retrieve in to the database. Document file must be in '.txt' or '.xlsx' format for pre-processing and further analysis. The basic idea is to check the file content of our input file will be browse by our software for further processing.

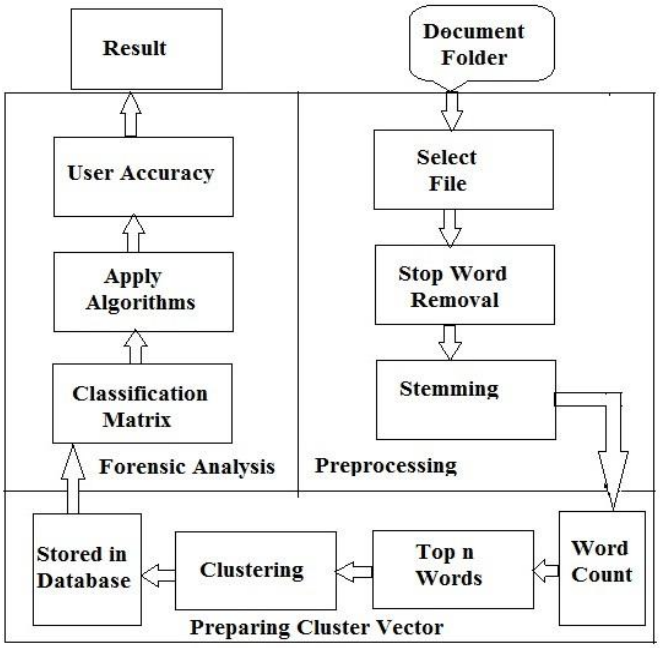

Fig1: System Architecture

b) StopWord Removal

The document file contains lot of input contains as stop words such as nouns, pronouns, adjectives etc. these are not affect the real meaning of the document. The removal of stop words is the most common term filtering technique used. For browsing the document/dataset, it accesses the stop word list from the database and removes stop word from the dataset [16].

c) Stemming:

In the pre-processing stemming is defined as reducing the words to their base form by removing suffix like -ing ,-ed etc. the words like 'plying', 'plays', played, 'player' all are converted to stem 'play', By Using the Porter Stemming Algorithm. These are the necessary step to filtering the data in the text clustering [17].

2) Preparing Cluster Vector

After pre-processing is done, for preparing cluster vector there is need to count the remaining word from the dataset. For preparing the cluster vector one will need to find top 100 words from the file on which pre-processing step is already done. The words having the largest difference between them, we prepare such task in the implementation. Then the object should have been assigned to its current cluster or to a neighbouring one [18].

Term Frequency:

In the multiple documents there are large numbers of individual occurrences of the words. So, Term Frequency (TF) [9], also called as Term Variance (TV) can be calculated. For each document weight of the words should be computed based on the Term Frequency (TF). It is a dimensionality reduction technique used to increase the efficiency and effectiveness of clustering algorithms. The results can be saved for containing the correlations between the documents based on the content. For preparing the cluster vector one will need to find top n words from the file on which pre-processing step is already done.

Similarity Calculation:

We should identify the similarities between the various documents based on the contents. In the dataset each document is compared with remaining other documents sequentially. The distance between the paragraphs can be computed from the similarity between the name or words in the document by using with the cosine base distance and the dissimilarities between the names or words in the documents using with the cosine and Euclidian distance. It is very advantage of a number of properties involving the composite and centroid vectors of a set of documents. The dissimilarities of the documents can be calculated based on the correlations between the documents available for clustering.

Cluster Formation:

Cluster formation is based on the variances multiple clusters are formed in the document by performing similarity calculation. At the end of resulting data clusters can be formed as single cluster. Then, the clustering process is repeated over and over again until a partition without singletons is found. Finally all the process, all singletons are processed into the resulting data partition (for evaluation purposes) as single clusters. Finally clusters are stored in the database for further processing.

3) Forensic Analysis

This will be the last step of proposed method. One can say that for the forensic data analysis classification matrix need to be made with the help of term frequency. It involves the construction of classification matrix based on frequency of occurrence of word. Applying k-means, complete link and Average link algorithm for the particular dataset to calculate the result. At last one can find accuracy of his work.

B. Clustering Algorithm

1) k-means Algorithm:

k-means is one of the simplest unsupervised learning clustering algorithms that solve the well-known clustering problem. The main basic idea is to find k centroids, one for each cluster. Randomly, choose centroids it should be placed in an anywhere with different location, it gives different result at every time change the location of centroid. So, the better choice is to place them as much as possible far away from each other. Another step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed. Again at this point we need to re-calculate k new centroids as centers of the clusters resulting from the earlier step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. This algorithm produces a separation of the objects into groups of cluster but every time we change centroid location result become different.

2) Complete Linkage:

Complete link algorithm is based on hierarchical clustering. All the clusters are sequentially combined into larger clusters, till all elements end up being in the same cluster. At every step, the two clusters separated by the shortest distance are combined. Complete link is the link between two clusters covers all element pairs, and the distance between clusters equals the distance between those two elements (one in each cluster) that are farthest away from each other. The method is also known as farthest neighbours clustering which is proposed by Johnson.

The following complete link algorithm is an agglomerative scheme that erases rows and columns in a proximity matrix as old clusters are merged into new ones. The N*N proximity matrix D contains all distances d (i, j). The clustering are assigned sequence numbers 0, 1... (n − 1) and L(k) is the level of the kth clustering. A cluster with sequence number m is denoted (m) and the proximity between clusters (r) and (s) is denoted d [(r), (s)]. [3][19]

The algorithm is composed of the following steps:

1. Start with the disjoint clustering having level L (0) = 0 and sequence number m = 0.

2. Find the most similar pair of clusters in the current clustering, say pair (r), (s), according to d [(r), (s)] = max d [(i), (j)] where the maximum is over all pairs of clusters in the current clustering.

3. Increment the sequence number: m = m + 1. Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L (m) = d [(r), (s)]

4. Update the proximity matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The proximity between the new cluster, denoted (r, s) and old cluster (k) is defined as d [(k), (r, s)] = max d [(k), (r)], d [(k), (s)].

5. If all objects are in one cluster, stop. Else, go to step 2.

3) Average Linkage:

Average link algorithm is based on hierarchical clustering. Average method is combine clusters with small variance and produce large number of cluster with the same variance. The distance between two clusters is defined by the average of the distances pairing two objects each from a cluster.

IV. PERFORMANCE ANALYSIS

Proposed "Elevating Document Clustering for Forensic Analysis Investigation System" This simplifies the effect of document clustering and thus successfully uses the variation of the dominant principal to identify the presence of rare but abnormal data. Clustering algorithm has been spiced over for a considerable length of time and the literature on the subject is huge.

The Result was produced depending upon content in the dataset; we use dataset investigation cases conducted by Brazil federal police department. For the purpose of applying algorithm, we select cluster at randomly in to the vector data. So, there is variation between the results at every iteration. In our experiments, we implement all methods in java using Eclipse. Run the program on server we install Apache Tomcat version 7, it is an open source software implementation of the Java Servlet, Java Server Pages, Java Expression Language and Java Web Socket technologies, developed under the java community.

Before each experiment, we do selection of the document then doing pre-processing for preparing cluster vector. The main partational and hierarchical document clustering algorithm is used clustering result.

Following figure 2 show basic view of our project:

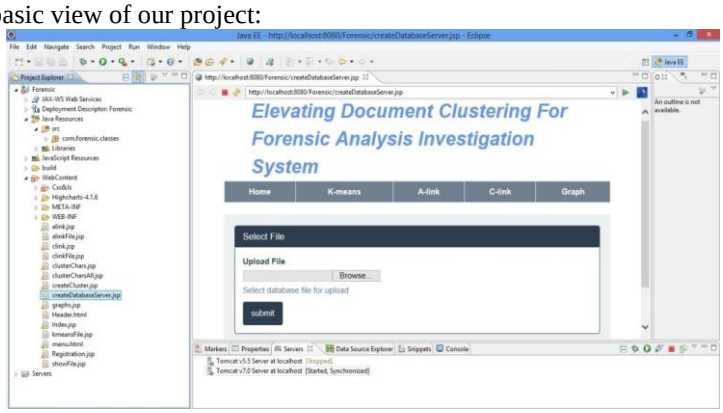

Fig 2: View of project

Generally speaking, the result of clustering can be represented in accuracy as shown in table1. Brazil police department consist of five dataset having thousands of documents. The name of Dataset is A, B, C, D and E used for the experimental setup, applying algorithm on each dataset to find better accuracy. Three performance measures k-means, complete link, average are shown by following table1.

Table I: Accuracy for partition and hierarchical clustering algorithm with the dataset

Dataset/Algorithm	K-Means	Average Link	Complete Link
A	0.75156	0.82010	0.88416
B	0.79094	0.82319	0.87827
C	0.79406	0.84415	0.85684
D	0.76191	0.83168	0.89683
E	0.75700	0.84869	0.87102

Above table 1, we observe that our proposed hierarchical clustering algorithm achieved better or comparable results than partition algorithm with better accuracy, while our system is the most computationally efficient one among the method considered. Above output showing k-means result is less than average link and complete link algorithm results.so average link and complete link provide the best result.

Following figure 3 shows results for apply algorithm on data set E.

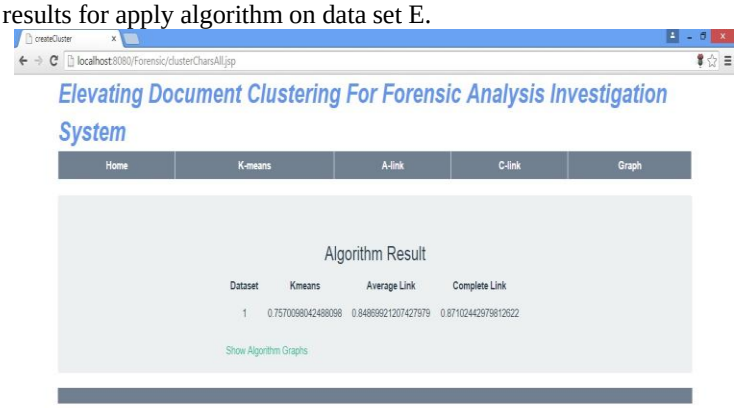

Fig 3. Output of partition and hierarchical clustering algorithm using E dataset

Above output showing k-means result is less than average link and complete link algorithm resuls.so average link and complete link provide the best result.

Three performance measures k-means, complete link, average link are shown by graph.

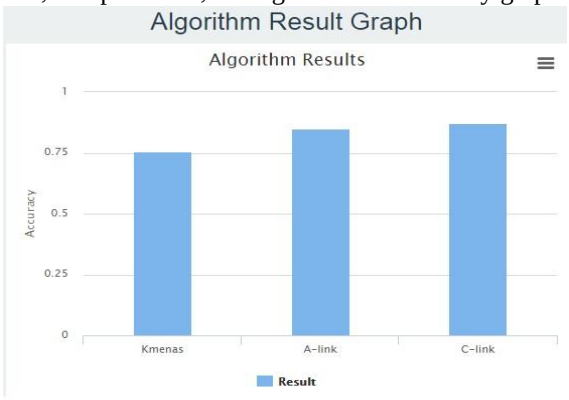

Fig 4: Clustering Algorithm Accuracy graph on data Set E

In the graph k-means algorithm shows 0.75 accuracy, Average link algorithm shows 0.84 accuracy and complete link algorithm provides 0.87 accuracy.

V. CONCLUSIONS

We introduce a new way clustering algorithms for computer forensic analysis; the scope of clustering data is very difficult step. Experimental result show that our formulation can perform very well on large data to be clustered in compute forensic to overcome this problem, we presented an approach that applies document or text clustering methods to forensic analysis. We have also tested our approach on computers seized in real world investigations cases conducted by Brazil police Department. From the experimental results, we demonstrate that the combining the algorithms can improve the efficiency and effectiveness of the clustering algorithm. We provide effective system to increase the speed of clustering. Furthermore, our method does not need to keep only particular dataset, it will applicable entire text dataset for the testing we used various dataset. Therefore, our approach is able to achieve the satisfactory results and overcome the limitation of existing system.

In our work, we are trying to implement the k-means algorithms and the algorithm known as Average Link and Complete Link yielded the best results. These algorithms are suitable for our work domain because it provides a summary view of documents which are being inspected. All the large number of case or documents is collected and providing corresponding output. Thus our technique is useful for large scale text data

.

ACKNOWLEDGMENT

I express my sincere gratitude towards my guide Dr. Sachin N. Deshmukh for his constant help, encouragement and inspiration throughout the project work. Without his invaluable guidance, this work would never have been a successful one. I would also like to thank all the faculty members and staff of Computer Science and IT department for making this journey successful. I would also like to thank Brazil police department to provide dataset on my research work.

REFERENCES

[1] U.S. Department of Justice, Electronic Crime Scene Investigation: A Guide for First Responders, I Edition, NCJ 219941, 2008, http://www.ncjrs.gov/pdffiles1/nij/219941.pdf.

[2] Filipe daCruz, 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 2013.

[3] A. K. Jain, R. C. Dupes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[4] R. Xu, D. C.Wunsch, II, Clustering. Hoboken, NJ: Wiley/IEEE Press, 2009.
[5] Jensen, J.H., Ellis, D., Christensen, M.G., Jensen, and S.H.: Evaluation distance measures between Gaussian mixture models of mfccs. Proc. Int. Conf. on Music Info. Retrieval ISMIR-07 Vienna, Austria pp. 107–108 (October, 2007).
[6] C. M. Bishop, Pattern Recognition and Machine Learning. New York: Springer-Verlag, 2006.
[7] B. K. L. Fei, J. H. P. Eloff, H. S. Venter, and. S. Oliver, "Exploring forensic data with self-organizing maps," in Proc.IFIP Int. Conf. Digital Forensics, 2005, pp. 113–123.
[8] N. L. Beebe and J. G. Clark, "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results," Digital Investigation, Elsevier, vol. 4, no. 1, pp. 49–54, 2007.
[9] L. Liu, J. Kang, J. Yu, and Z. Wang, "A comparative study on unsupervised feature selection methods for text clustering," in Proc. IEEE Int. Conf. Natural Language Processing and Knowledge Engineering, 2005, pp. 597–601.
[10] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem, "Towards an integrated e-mail forensic analysis framework, "Digital Investigation, Elsevier, vol. 5, no. 3–4, pp. 124–137, 2009.
[11] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi, "Mining write prints from anonymous e-mails for forensic investigation," Digital Investigation, Elsevier, vol. 7, no. 1– 2, pp. 56–64, 2010.
[12] 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.
[13] B. S. Everitt, S. Landau, and M. Leese, Cluster Analysis. London, U.K.: Arnold, 2001.
[14] L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka, "Relative clustering validity criteria: A comparative overview," Statist. Anal. Data Mining, vol. 3, pp. 209–235, 2010.
[15] 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.
[16] R. Mihalcea and P. Tarau.TextRank: "Bringing order into texts". In Proceedings of EMNLP 2004. pp. 404–411. https://github.com/ceteri/textrank.
[17] C.J. van Rijsbergen, S.E. Robertson and M.F. Porter, 1980. New models in probabilistic information retrieval. London: British Library. (British Library Research and Development Report, no. 5587).
[18] L. Kaufman and P. Rousseeuw, Finding Groups in Gata: An Introduction to Cluster Analysis. Hoboken, NJ: Wiley-Interscience, 1990.
[19] D. Defays (1977). "An efficient algorithm for a complete link method.

