International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Document Clustering Approach for Forensic Analysis: A Survey

Prachi K. Khairkar[1], D. A. Phalke[2]

[1, 2]Savitribai Phule Pune University, D Y Patil College of Engineering, Akurdi, Pune, India (411044)

Abstract: ***Digital world is consist of a huge volume of data. In forensic analysis, large amount of data is examined. So it's a difficult task for computer examiner to do such analysis in quick period of time. That's why to do the forensic analysis of documents within short period of time requires special approach such as Document Clustering. This paper reviews different existing Document clustering methods for example, K-means, K medoids, Single Link, Complete Link and Average Link in accordance with computer forensic analysis. In this paper we also give comparative study of different computer forensic analysis techniques and we propose enhance clustering algorithm which will improve accuracy of clustering to finding relevant documents from huge amount of data and speeds up computer examination process.***

Keywords: Document clustering, Forensic analysis, Text clustering, Clustering algorithms, Outlier Detection.

1. Introduction

In Computer Forensic analysis thousands of files are usually examined that allowing the evidence on suspected computer by analyzing the communication logs and the data on the computer storage device. Everyone faces the problem of handling large amount of data. Daily, thousands of files can be investigated per computer. The process of analyzing large volumes of data may consume a lot of time.

1.1 Forensic analysis

In general, Digital forensics is the application of investigation and analysis technique to collect and defend evidence from a particular computing device in a way that is proper for presentation in a court of act. Forensic analysis deals with the reorganization, collection, preservation, examination, analysis, extraction as well as documentation of digital evidences. *Computer Forensics*, which can be broadly defined as the discipline that combines elements of law and computer science to collect and analyze data from computer systems in a way that is admissible as evidence in a court of law. Data in those files consists of unstructured text or data whose data analysis by forensic examiner, which is difficult to be performed and requires lots of time to get clue for further investigation. The clustering algorithm is the data having more similar characteristic of information within a cluster [1].

The method of analyzing the various crimes using the computer based methods is called as digital forensic analysis (DFA). Digital forensics is a part of forensic science encompassing and has become an important tool in the identification of computer-based and computer-assisted crime. So the key factor to improve such forensic analysis process requires text clustering and document clustering techniques. The text clustering and document clustering simplifies the job of forensic examiner in forensic investigation.

1.2 Text clustering

Generally, most of computer seized devices consist of textual data which is to be processed by examiners, but this textual information resides in unstructured format which makes difficult job. Text mining provides an effective, automatic platform to support the analysis of digital textual evidences, which is the key point for forensic analysis process [3].Text clustering involves following steps:

1. Collection of documents
2. Pre-processing
   a) Tokenization
   b) Stop Words Removal
   c) Stemming
3. Text Clustering:

Based on preprocessing of data, text clustering is performed on preprocessed data. Text clustering produces sets of clusters as an output.

Forensic Analysis Process:

In forensic analysis process, the results of text clustering are used for collection of relevant files and documents according to reported case.

1.3 Document clustering

Computer forensic analysis involves the examination of the large volume of files. Among all of that files those file which are relevant to the forensic examiner interest need to be find quickly. Document clustering is the process of grouping similar documents into cluster which benefit is to retrieve the information effectively, reduce the search time and space, to remove outliers, to handle the high dimensionality of data and to provide the summary for similar documents. These document clustering provides different set of clusters among which forensic examiner analyze only relevant documents related to investigation of reported case. It helps to improve speed of the forensic analysis process. It will also help for forensic examiner to analyze the files and documents by only

analyzing representative of the clusters. The main purpose of using algorithms for clustering documents is to facilitate the discovery of new and useful knowledge from the documents under analysis.

Document clustering is considered as a centralized process has been in use in a number of different areas of text mining and information retrieval. Clustering is considered to be the most important unsupervised learning problem which deals with finding a structure in a collection of unlabeled data. A cluster is defined as a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. The main requirements that a clustering algorithm should satisfy are:

- Scalability,
- Handling many types of attributes,
- Identifying clusters with absolute shape,
- Determining input parameters,
- Capability to deal with noise and outliers,
- Unfitness to order of input records,
- High dimensionality,
- Interpretability and usability.

2. Literature Survey

The paper [1] has presented an approach that applies document clustering methods to forensic analysis of computers seized in police investigations. Reddy, BG Obula [2] in their paper gives comparative statement about different clustering algorithms by taking constraints i.e Data type, Cluster Shape, Complexity, Data Set, Measure, Advantages and Disadvantages. The paper [4] has introduced the different aspects of text mining and document clustering which are used for the analysis of forensic. In paper [5] author has used the clustering technique based on labeled clustering , for finding the correct disease of the patient and this clustering is done as soon as the update is made in the database it will provide us the current status of the patient and the treatment they are supposed to undergo. Zhao, Ying and Karypis, George in [6] this paper focuses to evaluate different hierarchical clustering algorithms and author has compared various partitional and agglomerative approaches. Karypis Steinbach, and Michael in their paper [7] presented the results of an experimental study of some common document clustering techniques: agglomerative hierarchical clustering and K-means. From survey on computer forensic analysis it can be concluded that clustering on data is not an easy step as there is huge data available. Hence, paper [10] presented an approach that applies document clustering methods to forensic analysis of computers seized in police work. Multithreading technique is used for document clustering for forensic data which will be useful for police investigations. In [13] this paper an effective digital text analysis strategy, believing on clustering based text mining techniques, is used for investigational purposes. The paper [14] is the survey of various clustering techniques. These techniques can be divided into several categories: Partitional algorithms, Density based, Hierarchical algorithms, and comparison of various clustering algorithms is surveyed and shows how Hierarchical Clustering can be better than other techniques.

3. Clustering Algorithms and Pre-processing

3.1 Collection of data

Document Clustering aims to automatically group related documents into clusters and also one of the tasks in machine learning and artificial intelligence and has received much attention in recent years. Clustering is one of the techniques of data mining extracting knowledge from large amount of information. Collection of data involves the processes like obtain the files and documents from the computer seized devices. The collection of such files and documents involves special techniques.

3.2 Pre-Processing Steps

It is done to represent the data in a form that can be used for clustering. There are many types of representing the documents like, graphical model, vector-Model, etc. Many measures are also used for weighing the documents and their similarities.

3.2.1 Stemming

Stemming is the process of reducing words into their base form and stem. For example, the words "connected", "connection", "connections" are all reduced to the stem "connect."

3.2.2 Stopword Removal

A term, which is not thought to convey any meaning as a dimension in the vector space (i.e. not context) is known as stopword. A typical method to remove stopwords is by compare each term with a compilation of known stopwords. This can be done by removing terms with low document frequencies and applying a part of speech tagger and then rejected all stop words such as nouns, verbs, pronouns, adjectives etc.

3.2.3 Term Frequiency

Reduction technique known as Term Variance (TV) is also used to increase efficiency of clustering algorithms. As clusters are formed, which containing documents, term variance are used to estimate top n words which have greatest occurrences over documents within clusters.

3.2.4 Similarity Computation

Also it is important to find out distances between two documents when they are resides in different clusters and for finding out distances between them, cosine-based distance and Levenshtein -based distance [11].

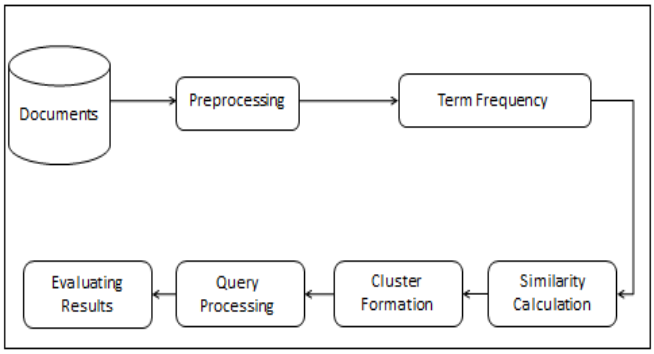

Figure 1: Data flow diagram

3.3 Estimating the Number of Clusters From Data

In order to estimate the number of clusters, a mostly used approach called as silhouette consists of getting a set of data partitions with different numbers of clusters and then selecting that particular partition that provides the best result according to a specific quality criterion (e.g., a relative validity index). Such a set of partitions may result directly from a hierarchical clustering dendogram or, alternatively, from multiple runs of a partitional algorithm (e.g., K-means) starting from different numbers and initial positions of the cluster prototypes.

Let us consider an object belonging to cluster A. a(i) denotes the average dissimilarity of i to all other objects of A. Let us consider cluster C. The average dissimilarity of i to all objects of cluster C will be called C(i). After computing d (i,C) for all clusters C ≠ A, the one which is smallest is selected, b(i) = min d(i,C), C ≠ A. This value represents the dissimilarity of i to its neighbour cluster, and the silhouette for a given object, s(i) is as below:

$$s(i) = \frac{b(i) - a(i)}{\max\{ a(i), b(i) \}}$$

It can be verified that $-1 \leq s(i) \leq 1$ [1].

3.4 Clustering Algorithms

There are various algorithms which can used in clustering like K-means, K-medoids, Single Link, Complete Link and Average Link. One of the simplest unsupervised algorithms is the K-means. In single link clustering, two groups have been merged such that their closest pair of documents have the highest similarity compared to any other pair of groups. In complete linkage many elements in the clusters are distant to each other. It produces more Compact clusters and most useful hierarchies than any other clustering. In average linkage clustering pairing of clusters takes place with the highest cohesion. The main key point of K-medoids is to determine optimal value from the original set of values.

3.4.1 Partitioning Algorithms

Partitioning methods are divided into two methods, as centroid and medoids algorithms. In centroid algorithms, each cluster is represented by using the gravity centre of the instances. The medoid algorithms establish each cluster by means of the instances closest to gravity centre.

K-means:

K-Means is one of the simplest unsupervised learning methods among all partitioning based clustering methods. If *k* is the number of desired clusters then it classifies a given set of *n* data objects in *k* clusters. A centroid is defined for each cluster. Data objects having centroid nearest (or most similar) to that data object are placed in a cluster. After processing all data objects, calculating centroids, are recalculated, and the entire process is repeated until no change. Based on the newly calculated centroids, all data objects are considered to the clusters. In each iterations centroids change their location as centroids move in each iteration. This process is continued until no change in the position of centroid. This results in *k* cluster representing a set of *n* data objects. An algorithm for k-means method is given below.

Algorithm :
Input : '*k*', the number of clusters to be partitioned, '*n*', the number of objects.
Output: A set of '*k*' clusters based on given similarity function.
Steps: **i)** Arbitrarily choose '*k*' objects as the initial cluster centers;
ii) Repeat,
a. (Re)assign each object to the cluster to which the object is the most similar; based on the given similarity function;
b. Update the centroid (cluster means), by calculating the mean value of the objects for each cluster;
iii) Until no change. [8]

K-medoid:

The k-means method uses centroid to represent the cluster and it is sensitive to outliers. That is, a data object with an extremely large value may disrupt the distribution of data. This problem can be overcome by using medoids to represent the cluster rather than centroid. A medoid is the center data object in a cluster.

Here, k data objects are selected randomly as medoids to represent k cluster and remaining all data objects are placed in a cluster having medoid nearest (or most similar) to that data object. New medoid is determined after processing all data objects, which gives cluster in a better way and the entire process is repeated. Again all data objects are clustered to the clusters based on the new medoids. In each iteration, medoids are changing their location step by step. This process is continued until no there is no move. As a result, *k* clusters are found representing a set of n data objects.

Algorithm :
Input : '*k*', the number of clusters to be partitioned, 'n', the number of objects.
Output: A set of '*k*' clusters that minimizes the sum of the dissimilarities of all the objects to their nearest medoid.
Steps:
i) Arbitrarily choose '*k*' objects as the initial medoids;
ii) Repeat,
a. Assign each remaining object to the cluster with the nearest medoid;
b. Randomly select a non-medoid object;

c. Compute the total cost of swapping old medoid object with newly selected nonmedoid object.
d. If the total cost of swapping is below zero, then perform that swap operation to form the new set of k- medoids.
iii) Until no change.

3.4.2 Expectation Maximization

The EM algorithm is well established clustering algorithms, in statistic community. The EM is model-based clustering algorithm that assumes the data were generated by a model and then tries to recover the original model from the data. This model then defines clusters and the cluster membership of data. The EM algorithm is a generalization of K-Means algorithm in which the set of K centroids as the model that generate the data. It recurs between an expectation step, corresponding to redesigning, and a maximization step, corresponding to re-calculation of the parameters of the model. EM is chosen to cluster for the following reasons among others:

- It has a strong statistical basis.
- It is linear in database size.
- It is robust to noisy data.
- It is able to accept any number of clusters as input.
- It can handle high dimensionality.
- It merges fast if a good initialization is given

3.4.3 Fuzzy C-Means (FCM) algorithm

Traditional clustering approaches generate partitions; each pattern belongs to one cluster. The clusters in a hard clustering are disjoint. Fuzzy clustering extends this presentation to associate each pattern with every cluster using a membership function. Larger membership values specify higher confidence in the assignment of the pattern to the cluster. Widely used one algorithm is the Fuzzy C-Means (FCM) algorithm, which is based on k-means. The Fuzzy C-Means Clustering (FCM) is an unsupervised goal oriented clustering algorithm which attempts to find the most characteristic point in each cluster, this can be considered as the center of the cluster, then using membership functions for each instance in the clusters. It is efficient, straightforward, and can be implemented easily, but it is sensitive to initialization (due to random selection of initial center points) and so is easily trapped in local optima [4].

3.4.4 Hierarchical Clustering

Hierarchical clustering approaches attempt to create a hierarchical decomposition of the given document collection thus achieving a hierarchical structure. Hierarchical methods are usually classified into Agglomerative and Divisive methods depending on how the hierarchy is constructed. Agglomerative methods start with an initial clustering of the term space, where all documents are considered representing a separate cluster. The closest clusters using a given inter-cluster similarity measure are then merged continuously until only 1 cluster or a predefined number of clusters remain [6].

Steps in simple Agglomerative Clustering Algorithm:
1. Compute the similarity between all pairs of clusters i.e. calculate the similarity between the ith and jth clusters, which is represented by ijth entry of a similarity matrix.
2. Merge the most similar (closest) two clusters.
3. Update the similarity matrix to reflect the pairwise similarity between the new cluster and the original clusters.
4. Repeat steps 2 and 3 until only a single cluster remains.

Divisive clustering algorithms start with a single cluster containing all documents. It then continuously divides clusters until all documents are contained in their own cluster or a predefined number of clusters are found [6]. Agglomerative algorithms are usually classified according to the inter-cluster similarity measure between them. The most popular of these are single-link, complete-link and group average. Single link method where the distance between clusters is the minimum distance between any pair of elements drawn from these clusters, in the complete link it is the maximum distance and in the average link it is correspondingly an average distance. The clusters produced by the single link algorithm are formed at low dissimilarities in the dissimilarity dendrogram. On the other hand, the clusters produced by the complete link algorithm are formed at high dissimilarities in the dissimilarity dendrogram.

3.5 Outlier Detection

Silhouette is an approach to remove outliers. Fundamentally, if the best partition chosen by the silhouette has singletons (i.e., clusters with single object), these are removed. Then, the clustering process is repeated recursively until a partition without singletons is found. Finally, all singletons are incorporated into the resulting data partition (for evaluation purposes) as single clusters [10].

4. Comparative Statement of Various Clustering Techniques

Table 1: Clustering Techniques Comparison

Clustering Technique	*Complexity*	*Measure*	*Advantages*
K-means	O(nki)	Mean	K-means is relatively scalable and efficient in processing large data sets.
K-medoid	O(tkn)	Medoid	1. K-Medoids method is more robust than k-Means in the presence of noise and outliers. 2. K-Medoids algorithm seems to perform better for large data sets.
Single link	O(n^2)	Similarity Measure	1. Theoretical properties, efficient implementations, widely used. 2. No cluster centroid or representative required, so no need arises to recalculate the similarity matrix.
Average link	O(n^2 log n)	Similarity Measure	It is a structure intermediate between the loosely bound single link clusters and tightly bound complete link clusters.
Complete link	O(n^2 log n)	Similarity Measure	It gives good results as compared to single and average link.
Fuzzy C-means	O(ndc^2i)	Similarity Measure	Gives best result for overlapped data set and comparatively better then k-means algorithm.

5. Summary

Digital Forensic Investigation is the branch of scientific forensic process for investigation of material found in digital devices related to computer crimes. Process is to analyze the documents present on computer. Due to increasing number of documents and larger size of storage devices makes very difficult to analyze the documents on computer. The main advantage of document clustering here is an unsupervised i.e. number of clusters cannot be predetermined or given by users. Time taken to perform the clustering takes less time based on subjects declared. In this paper, different clustering techniques are compared based on parameters such as similarity measures, complexities and their advantages.

6. Acknowledgment

The authors would like to thank the publishers and researchers for making their resources available. We also thank the college authority for providing the required infrastructure and support. Finally we would like to extend our heartfelt gratitude to friends and family members.

References

[1] L. F. da Cruz Nassif and E. R. Hruschka, “Document clustering for forensic analysis: An approach for improving computer inspection," Information Forensics and Security, IEEE Transactions on, vol. 8, no. 1, pp. 46-54, 2013.

[2] Reddy, BG Obula, et al. "Literature Survey On Clustering Techniques." *IOSR Journal of Computer Engineering (IOSRJCE) ISSN*: 2278-0661.

[3] Umale, Bhagyashree, and M. Nilav. "Survey on Document Clustering Approach for Forensics Analysis."

[4] Stoffel, Kilian, Paul Cotofrei, and Dong Han. "Fuzzy methods for forensic data analysis." *SoCPaR*. 2010.

[5] ISRAIL, K., and CC KALYAN SRINIVAS. "Improving Computer Inspection by Using Forensic Cluster Analysis to Develop the Document." (2014).

[6] Zhao, Ying, and George Karypis. "Evaluation of hierarchical clustering algorithms for document datasets." *Proceedings of the eleventh international conference on Information and knowledge management*. ACM, 2002.

[7] Steinbach, Michael, George Karypis, and Vipin Kumar. "A comparison of document clustering techniques." *KDD workshop on text mining*. Vol. 400. No. 1. 2000.

[8] Patil, A. J., C. S. Patil, R. R. Karhe, and M. A. Aher. "Comparative Study of Different Clustering Algorithms."

[9] Beebe, Nicole Lang, and Jan Guynes Clark. "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results." *Digital investigation* 4 (2007): 49-54.

[10] Abhonkar, Prashant D., and Preeti Sharma. “Evaluating Forensic Investigation System for File Clustering.”

[11] Vidhya, B., and R. Priya Vaijayanthi. "Enhancing Digital Forensic Analysis through Document Clustering."

[12] Iqbal, Farkhund, Hamad Binsalleeh, Benjamin Fung, and Mourad Debbabi. "Mining write prints from anonymous e-mails for forensic investigation." *digital investigation* 7, no. 1 (2010): 56-64.

[13] Decherchi, Sergio, et al. "Text clustering for digital forensics analysis." *Computational Intelligence in Security for Information Systems*. Springer Berlin Heidelberg, 2009. 29-36.

[14] Popat, Shraddha K., and M. Emmanuel. "Review and Comparative Study of Clustering Techniques."