IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A SURVEY ON DOCUMENT CLUSTERING APPROACH FOR COMPUTER FORENSIC ANALYSIS

Monika Raghuvanshi*, Rahul Patel
Acropolise Institute Of Technology and Research Indore Madhya Pradesh

DOI: 10.5281/zenodo.60125

ABSTRACT

In a forensic analysis, large numbers of files are examined. Much of the information comprises of in unstructured format, so it's quite difficult task for computer forensic to perform such analysis. That's why to do the forensic analysis of document within a limited period of time require a special approach such as document clustering. This paper review different document clustering algorithms methodologies for example K-mean, K-medoid, single link, complete link, average link in accorandance with computer forensic analysis.

KEYWORDS: Forensic analysis, Document clustering, Clustering algorithm, Unstructured.

INTRODUCTION

Due to advancement in technology, it is estimated that the digital data density increased 18 times in latest 5 to 6 year[1]. This increased data has direct impact in computer forensic. in general computer forensic is the application of investigation and analysis techniques in which evidence are collected from a particular computing device in a manner that is proper for presentation in a court and according to the law. Document clustering has shown to be very useful for computer forensic analysis.

Computer Forensic Analysis

Computer forensic analysis is a branch of forensic science encompassing the investigation of material found in digital device in a way that is proper for presentation in a court and according to the law. Document analysis in a computer device is a key task of the computer forensic investigation process. But this task may be daunting due to large no of document usually stored on a hard disk. The clustering algorithm are used in the process of computer forensic analysis .these methods are basically used to covert unstructured documents to structured documents for further investigation.

Document Clustering

Document clustering provides an effective, automatic platform to support the analysis of digital textual evidence, which is the key point for forensic analysis process. The process of grouping a set of physical or abstract object into class of similar object is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the object in other cluster.

LITERATURE SURVEY

There are only a few studies reporting the use of clustering algorithms in the Computer Forensics field. Essentially, most of the studies describe the use of classic algorithms for clustering data—e.g., Expectation-Maximization (EM) for unsupervised learning of Gaussian Mixture Models, K-means, Fuzzy C-means (FCM), and Self-Organizing Maps (SOM). These algorithms have well-known properties and are widely used in practice. For instance, K-means and FCM can be seen as particular cases of EM [2]. Algorithms like SOM [3], in their turn, generally have inductive biases similar to K-means, but are usually less computationally efficient.

In [4], SOM-based algorithms were used for clustering files with the aim of making the decision-making process performed by the examiners more efficient. The files were clustered by taking into account their creation dates/times and their extensions. This kind of algorithm has also been used in [5] in order to cluster the results from keyword searches.

An integrated environment for mining e-mails for forensic analysis, using classification and clustering algorithms, was presented in [6]. In a related application domain, e-mails are grouped by using lexical, syntactic, structural, and domain-specific features [7]. Three clustering algorithms (K-means, Bisecting K-means and EM) were used. The problem of clustering e-mails for forensic analysis was also addressed in [8], where a Kernel-based variant of K-means was applied. The obtained results were analyzed subjectively, and the authors concluded that they are interesting and useful from an investigation perspective. More recently [9], a FCM-based method for mining association rules from forensic data was described.

The literature on Computer Forensics only reports the use of algorithms that assume that the number of clusters is known and fixed *a* priori by the user. Aimed at relaxing this assumption, which is often unrealistic in practical applications, a common approach in other domains involves estimating the number of clusters from data. Essentially, one induces different data partitions (with different numbers of clusters) and then assesses them with a relative validity index in order to estimate the best value for the number of clusters [2], [3], [10]. This work makes use of such methods, thus potentially facilitating the work of the expert examiner—who in practice would hardly know the number of clusters apriori.

DOCUMENT CLUSTERING PROCESS

Clustering is the most common form of unsupervised learning which deals with finding a structure in a collection of unlabeled data. .Clustering of document is an automatic grouping of text document within a cluster have a high resemblance in comparison to one another ,but are different from document in other clusters. It is important to emphasize that getting from a collection of document to a clustering of the collection is not merely a single process , but is more a process in multiple stage.

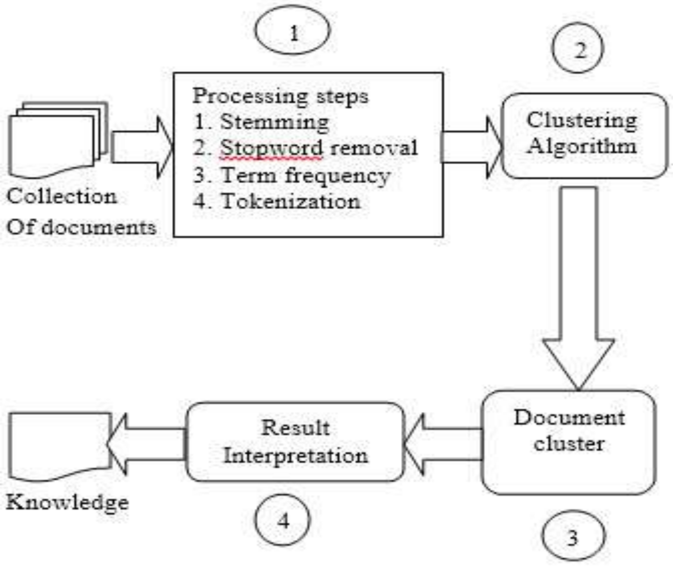

Figure1: Document clustering process

Collection of data:

Methods like crawling, indexing and filtering etc which are used to collect the documents that needs to be clustered.

Preprocessing Steps:

Preprocessing is done to represent the data in a form that can be used for clustering

Stemming :

Stemming is a technique for the reduction of words into their steam or base form many words e.g. agreed, agreeing, disagrees, agreement, and disagreement belong to agree.

Stop word Removal

Prepositions, articles, and pronouns etc are the most common words in any text document does not provide meaning of the document. These words are eliminated. These words are not necessary for text mining application.

Term Frequency

The simplest possible method for feature selection in document clustering is document frequency that is used to filter out irrelevant feature. In other word, words which are too frequent in the corpus can be removed because they are

Tokenization

Splits sentences into separates tokens, the main use of tokenization is to identifying meaningful keyword

Clustering Algorithm

The clustering algorithm is used in the process of digital forensic analysis. These methods are basically used to convert unstructured document to structured document for further investigation. In this work we used a different clustering algorithm as follows.

K-Means

K-means is the most important flat clustering algorithm. The objective function of K-means is to minimize the average squared distance of objects from their cluster centers, where a cluster center is defined as the mean or centroid μ of the objects in a cluster C:

The ideal cluster in K-means is a sphere with the centroid as its center of gravity. Ideally, the clusters should not overlap. A measure of how well the centroids represent the members of their clusters is the Residual Sum of Squares (RSS), the squared distance of each vector from its centroid summed over all vectors

K-means can start with selecting as initial clusters centers K randomly chosen objects, namely the seeds. It then moves the cluster centers around in space in order to minimize RSS. This is done iteratively by repeating two steps until a stopping criterion is met

1. Reassigning objects to the cluster with closest centroid
2. Recomputing each centroid based on the current members of its cluster.

We can use one of the following termination conditions as stopping criterion

- A fixed number of iterations I has been completed.
- Centroids μ do not change between iterations.
- Terminate when RSS falls below a pre-estabilished threshold.

K-Medoids

The k-mean alogorithm is sensitive to outlier because an object with an extremely large value may substantially distort the distribution of data. This problem can be over come 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 medoids nearest to that objects. 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 cluster 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-cluster are found representing a set of n data objects.

Expectation Maximization

The EM algorithm fall within a subcategory of the flat clustering algorithms, called Model-based clustering. The model-based clustering assumes that 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 alternates between an expectation step, corresponding to reassignment, and a maximization step, corresponding to recomputation of the parameters of the model.

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. Simple Agglomerative Clustering Algorithm:

1. Compute the similarity between all pairs of clusters i.e. calculate a similarity matrix whose ij entry gives the similarity between the I and j clusters.
2. Merge the most similar (closest) two clusters.
3. Update the similarity matrix to reflect the pair wise 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. Agglomerative algorithms are usually classified according to the inter-cluster similarity measure they use. The most popular of these are single-link, complete-link and group average. In the single link method, the distance between clusters is the minimum distance between any pair of elements drawn from these clusters (one from each), in the complete link it is the maximum distance and in the average link it is correspondingly an average distance

A hierarchical clustering algorithm creates a hierarchical decomposition of the given set of data objects. Depending on the decomposition approach, hierarchical algorithms are classified as agglomerative (merging) or divisive (splitting). The agglomerative approach starts with each data point in a separate cluster or with a certain large number of clusters. Each step of this approach merges the two clusters that are the most similar. Thus after each step, the total number of clusters decreases. This is repeated until the desired number of clusters is obtained or only one cluster remains. By contrast, the divisive approach starts with all data objects in the same cluster. In each step, one cluster is split into smaller clusters, until a termination condition holds. Agglomerative algorithms are more widely used in practice. Thus the similarities between clusters are more researched.

CONCLUSION

In this paper, different clustering techniques for computer forensic analysis are presented .the aim of these paper is to present different methodology of computer forensic analysis with phases included into it. There I huge amount of data to be cluster in computer forensic so to overcome this problem, this survey paper represent an different approach for document clustering method for forensic analysis of computers seized in police investigation

REFERENCE

[1] 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.

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

[3] S. Haykin, Neural Networks: A Comprehensive Foundation.Englewood Cliffs, NJ: Prentice-Hall, 1998.

[4] 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

[5] 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,

[6] 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.

[7] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi, "Mining writeprints from anonymous e-mails for forensic investigation," *Digital Investigation, Elsevier*, vol. 7, no. 1–2, pp. 56–64, 2010.

[8] 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.

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

[10] 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.