Enhanced Document Clustering using K-Means with Support Vector Machine (SVM) Approach

Prachi K. Khairkar[1]
Savitribai Phule Pune University
D Y Patil College of Engineering,
Akurdi, Pune. India
prachik4590@gmail.com

Mrs. D. A. Phalke[2]
Savitribai Phule Pune University
D Y Patil College of Engineering,
Akurdi, Pune. India
a_dhanashree@yahoo.com

Abstract—Today's digital world consists of a large amount of data. Volume of data in the digital world is increasing continuously. Dealing with such important, complex and unstructured data is important. These files consist of data in unstructured text, whose analysis by computer examiners is difficult to be performed. In forensic analysis, experts have to spend a lot of time as well as efforts, to identify criminals and related evidence for investigation. However crime investigation process needs to be faster and efficient. As large amount of information is collected during crime investigation, data mining is an approach which can be useful in this perspective. Data mining is a process that extracts useful information from large amount of crime data so that possible suspects of the crime can be identified efficiently. Algorithms for clustering documents can provide the learning of knowledge from the documents under analysis. This can be done by applying different clustering algorithms to different datasets. Clustering algorithms indeed tends to induce clusters formed by either relevant or irrelevant documents, further extending work by using Clustering Technique Cascaded with Support Vector Machine, thus contributing to enhance the experts job and investigation process can be speed up.

Keywords- *Document Clustering; Clustering Analysis; K-means; Support Vector Machine*

*****

I. INTRODUCTION

Digital world consist of information in computers and this information is very essential for future references and studies irrespective of various fields. There is exponential growth of the data in digital world. So, clustering algorithms play important role in forensic analysis of digital documents. Digital world contains very important, complex and unstructured data.

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.

A. 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.

B. 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.

4. 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.

C. *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.

II. RELATED WORK

A. *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. Partitioning methods are as k-means and k-medoid.

B. *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.

C. *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 recalculation of the parameters of the model.

L. F. da Cruz Nassif and E. R. Hruschka [1] have 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. Stoffel, Kilian, Paul Cotofrei, and Dong Han [3] have introduced the different aspects of text mining and document clustering which are used for the analysis of forensic. Author [4] 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 [5] 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 [6] 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, Prashant D [7] has 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.

Decherchi, Sergio, et al [8] in their paper have used an effective digital text analysis strategy, believing on clustering based text mining techniques, for investigational purposes. The author M. Emmanuel [9] has explained 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.

III. PROPOSED WORK

Suppose an investigator of a digital forensic case is interested in clustering a collection of related documents. The There are various algorithms for this process such as k-means, kmedoids, hierarchical, expectation maximization, etc. datasets are made up of unlabeled categories or classes of documents which were initially identified as unknown. In such cases even if we consider the thing that the availability of labelled dataset is possible through the past analysis, but there is no possibility that same classes or groups available in input dataset or for next incoming raw dataset which is collected from different digital devices as well as related to various processes of investigations. The new data sample can come from the different types of sources. Therefore to provide the efficient solution to process such input datasets in forensic analysis, the clustering algorithms are used [1].

Proposed system tries to improve the performance and quality of the output generated by the clustering technique by cascading it with Support Vector Machine (SVM). From the large volume data document, it is important to retrieve information needed by using relevant document. Text summarization is a process of extracting content from a document and generating summary of that document thus presenting important content to user in a relatively condensed form [12]

 ISSN: 2321-8169
4112 - 4116

A. System Architecture

The system architecture consists of following steps:

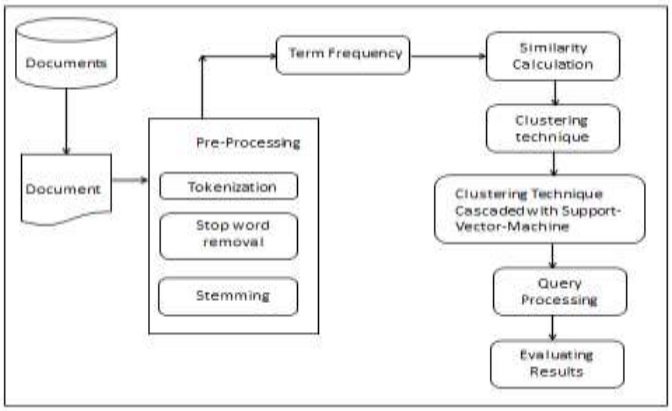

Figure 1. System Architecture

- **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.

- **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.

- **Tokenization:** In this phase sentences are divided into streams of individual tokens that are differentiated by spaces.

- **Stopword Removal:** A term, which is not thought to convey any meaning as a dimension in the vector space 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 [1,2].

- **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."

- **Term Frequency:** 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.

- **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].

- **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 $C \neq A$, the one which is smallest is selected, b(i) = min d(i,C), $C \neq A$. This value represents the dissimilarity of i to its neighbour cluster, and the silhouette for a given object, s(i) is as below [1]:

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

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

- **Document Clustering:** 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. It is important to emphasize that getting from a collection of documents to a clustering of the collection, which is not only consist of a single operation, but also is more a process in multiple stages which includes more traditional information retrieval operations such as crawling, indexing, weighting, filtering etc.

- **Outlier Detection:** To remove outliers, a simple approach that makes recursive use of the silhouette is used. Fundamentally, if the best partition chosen by the silhouette has singletons (i.e., clusters formed by a single object only), these are removed. Then, the clustering process is repeated over and over again-until a partition without singletons. At the process end, all singletons are incorporated into the resulting data partition as single clusters [10].

- **Clustering Technique Cascaded with Support-Vector-Machine:** For this phase, an algorithm will be generated containing clustering technique cascaded with the machine learning technique i.e Support-Vector-Machine (SVM). This helps in retrieving documents have most of the redundant information [16].

 ISSN: 2321-8169
4112 - 4116

SVM is a supervised machine learning algorithm used for data classification and estimating the relationship between variables. It's a supervised algorithm because there's an initial training phase involved where you feed the algorithm data that has already been classified (labeled) [12].

Suppose a big set of features to ensure that the two classes are linearly separable. The best separating line to be used is obtained by using support vector machine which maximizes the distance between the hyperplane and the "difficult points" also called as "Support vectors" close to decision boundary.

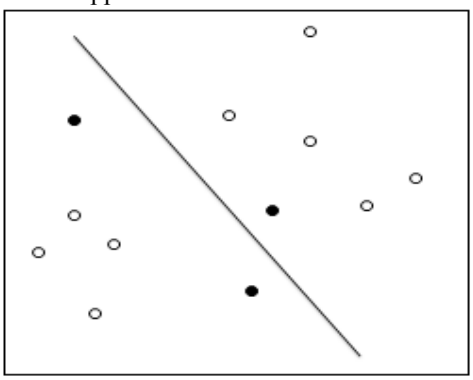

Figure 2. Support Vectors are indicated by Dark Circles

IV. ALGORITHM

SVM with 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 [2,12].

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]
iv) Finding the closest pair of data points '*n*'.
v) Adding a Point to the Support Vector Set SV
vi) Repeat till all such points are pruned.

V. IMPLEMENTATION AND RESULTS

In this section we represent the input, result of practical work done. There are various algorithms which can be used in clustering like K-means, K-medoids, hierarchical, expectation maximization algorithms. One of the simplest unsupervised algorithms is the K-means. For this implementation, the dataset used is a set of documents collected from computer like bank statements, bills, financial documents, web logs etc. Input dataset includes different documents in different file formats such as document file, image file, etc. Subsequently, those documents were converted into plain text format and preprocessed further as described in section III.

We have used windows XP/7 operating system, Java programming language and Netbeans tool. Result consists of documents clustered on the basis of their similarity. Different clusters are formed using clustering algorithms which help to search and evaluate the query. Dataset contains varying amount of documents, attributes, singletons, number of documents in each clusters. For the datasets, the best data partitions are formed by clusters containing either relevant or irrelevant documents.

TABLE I. EXPERIMENTAL RESULTS

Sr No.	Clustering Algorithm	Results (Accuracy)
1	K-means	15 ms
2	K-means with Support Vector Machine (SVM)	10 ms

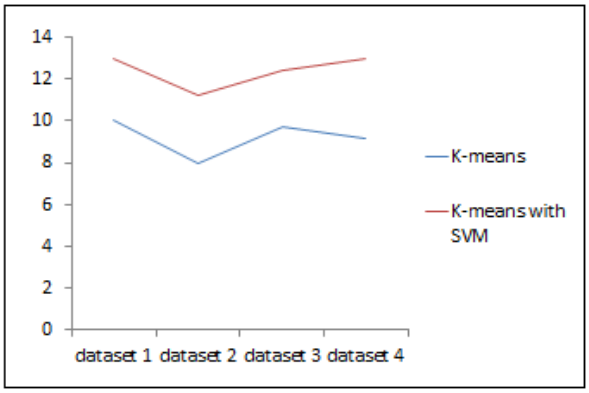

Figure 3. Graphical Representation

VI. CONCLUSION

Digital Forensic Investigation is the category 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. In this paper, different clustering techniques are applied to available dataset and compared based on parameters such as similarity measures, complexities and their advantages. Hence, clustering technique cascaded with support vector machine to improve performance accuracy and quality of system. Future work includes overlapping partitions such fuzzy C-means, automatic approach for cluster labeling to identify the semantic content of each cluster more quickly.

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] Umale, Bhagyashree, and M. Nilav, "Survey on Document Clustering Approach for Forensics Analysis," IJCSIT) International Journal of Computer Science and Information Technologies 5.3 (2014).

[3] Stoffel, Kilian, Paul Cotofrei, and Dong Han, "Fuzzy methods for forensic data analysis," SoCPaR, , pp. 23-28, 2010.

[4] Israil, K., and CC Kalyan Srinivas, "Improving Computer Inspection by Using Forensic Cluster Analysis to Develop the Document," (2014).

[5] 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.

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

[7] Abhonkar, Prashant D., and Preeti Sharma, "Elevating Forensic Investigation System for File Clustering,"

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

[9] Popat, Shraddha K., and M. Emmanuel "Review and Comparative Study of Clustering Techniques," International Journal of Computer Science and Information Technologies 5, no. 1 (2014): 805-812.

[10] Vidhya, B., and R. Priya Vaijayanthi, "Enhancing Digital Forensic Analysis through Document Clustering," International Journal of Advanced Research in Computer Engineering and Technology (IJARCET) Volume 3 (2014).

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

[12] M. S. Patil , M. S. Bewoor, S. H. Patil, "A Hybrid Approach for Extractive Document Summarization Using Machine Learning and Clustering Technique ," International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 1584-1586, 2014.

[13] 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.

[14] Iqbal, Farkhund, Hamad Binsalleeh, Benjamin Fung, and Mourad Debbabi, "Mining writeprints from anonymous e-mails for forensic investigation," digital investigation 7, no. 1 : 56-64, 2010.

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

[16] S. K. G. Madan Kumar, "File clustering using forensic analysis system ," International Journal of Computer Science and Mobile Computing, volume 3, no. 7, page numbers: 948-954, July 2014.