Information Retrieval Using Document Clustering for Forensic Analysis

Raksha K. Mundhe, Ankush Maind, R.B. Talmale
IVth SEM, M.Tech – CSE, Department of CSE, H.O.D. of CSE Department
Tulasiramji Gaikwad-Patil College of Engineering and Technology, Nagpur, India
Email: raksha.k.mundhe@gmail.com, ankushmaind@gmail.com, roshanikambe@rediffmail.com

Abstract - In computer forensic analysis, retrieved data is in unstructured text, whose analysis by computer examiners is difficult to be performed. In proposed approach the forensic analysis is done very systematically i.e. retrieved data is in unstructured format get particular structure by using high quality well known algorithm and automatic cluster labelling method. Two relative validity indexes were used to automatically estimate the number of clusters with automatic labelling to it; which makes it very easy to retrieve most relevant information for forensic analysis.

Index terms: forensic analysis, clustering, automatic labelling method, indexing.

I. INTRODUCTION

It usually involves examining hundreds of thousands of files per computer. The search engines commonly build a very large centralized database to index a portion of the Internet and help to reduce the information overload problem by allowing a user to do a centralized search. However, they also bring up another problem: too many web pages are returned for a single query. To find out which documents are useful, users have to sift through hundreds of pages to find out that only a few of them are relevant. One way to tackle this problem is to cluster the search result documents based on their semantic similarity so that users can scan a few coherent groups instead of many individual documents. This activity exceeds the expert’s ability of analysis and interpretation of data. Therefore in proposed approach, methods and clustering used keep paramount importance. Explosion of research aimed at facilitating retrieval and visualization capabilities of data clustering structures. Clustering algorithms are used for data sets appearing in statistics, computer science, and machine learning; the travelling salesman problem is used to find out the shortest possible distance between clusters [6].

The rationale behind clustering algorithms is that objects within a valid cluster are more similar to each other than they are to objects belonging to a different cluster. Such an approach, based on document clustering, can indeed improve the analysis of seized computers. Therefore, I decided to choose a set of six representative algorithms in order to show the potential of the proposed approach, namely: the partitional K-means and K-medoids, the hierarchical Single/Complete/Average Link, and the cluster ensemble algorithm known as CSP and automatic labelling cluster method.

This paper is organized in the following way. in section II some earlier related work is explained. In section III, information retrieval using document clustering. In section IV, the conclusion.

II. RELATED WORK

Information Retrieval:

Information Retrieval (IR) is a large and growing field, Search engines like Google is used by many people to retrieve data or to return thousands of related web pages. Clustering can be used to group search results into small number of clusters, each of which captures a particular aspect of the query. For instance, a query of “Library” might return List of books, author, and publication. Each category (cluster) is broken into subcategories (sub-clusters), producing hierarchical structure that further assists a user’s explanation of the query results.

Six well-known algorithms two relative validity indexes were used to automatically estimate the number of clusters. Dendrograms provide summarized view of the document being inspected also provide very informative descriptions and visualization capabilities of data clustering structures. Clustering algorithms are used for data sets appearing in statistics, computer science, and machine learning; the travelling salesman problem is used to find out the shortest possible distance between clusters [6].

Vector Space Model is used for Information Retrieval. In this document is represented by vector containing the frequencies of occurrences of words, which are defined as delimited alphabetic strings, whose number of characters are between 4 and 25. TV selects a number of
attributes (near about 100 words) that have greatest variance over the documents. Dimensionality reduction technique increases both effectiveness and efficiency of clustering algorithm. In order to compute distance between documents two measures are used.

1) Cosine based distance.
2) Levenshtein based distance.

By using the cosine function as the measure of similarity between documents we can take advantage of a number of properties involving the composite and centroid vectors of a set of documents. In particular, if \( S_i \) and \( S_j \) are two sets of unit-length documents containing \( n_i \) and \( n_j \) documents respectively, and \( D_i \), \( D_j \) and \( C_i \), \( C_j \) are their corresponding composite and centroid vectors then the following is true:

1. The sum of the pair-wise similarities between the documents in \( S_i \) and the document in \( S_j \) is equal to
2. The sum of the pair-wise similarities between the documents in \( S_i \) is equal to \( D_i \).

Note that this equation includes the pair wise similarities involving the same pairs of vectors.

Three effective and efficient combiners are used to solve cluster ensemble problem as an optimization problem based on a hyper-graph model which results cluster ensembles can

1) Improve quality and robustness, and
2) Enable distributed clustering.

An automatic procedure and methodology are described for inferring accurate and easily understandable expert-system-like rules from forensic data. For this FCM-based mining association rule based on the fuzzy set theory is used to generate most appropriate result [3]. Partitional algorithms always lead to better clustering solutions than agglomerative algorithms, which suggests that partitional clustering algorithms are well-suited for clustering large document datasets due to not only their relatively low computational requirements, but also comparable or even better clustering performance.

The consistency of clustering solutions at different levels of granularity allows flat partitions of different granularity to be extracted during data analysis, making them ideal for interactive exploration and visualization. There is the common belief that in terms of clustering quality, partitional algorithms are actually inferior and less effective than their agglomerative counterparts[2]. For partitional clustering algorithms, we used six functions [5] that have been shown to produce high-quality partitional clustering solutions. The vector-space model is to represent each document. In this model, each document \( d \) is considered to be a vector in the term-space. In particular, we employed the \( t f = idf \) term weighting model, in which each document can be represented as

\[
(tf_1 \log(n/df_1), tf_2 \log(n/df_2), \ldots, tf_m \log(n/df_m))
\]

Where, \( tf_i \) is the frequency of the \( i \)th term in the document and \( df_i \) is the number of documents that contain the \( i \)th term.

A self-organizing map (SOM) [3] assists to computer forensic investigators conducting data analysis in a more efficient manner. A SOM is used to search for patterns in data sets and produce visual displays of the similarities in the data. SOM provide greater abilities to interpret and explore data generated by computer forensic tools. The data set obtained directly from hard drive can be critical to an investigation. Patterns in the dataset could help forensic investigators to locate information and guide them to the next step in their search. The technique is used to create graphical representations of large datasets that offer investigators a fresh perspective from which to study the data.

Computer forensics deals with the presentation, identifications, extraction and documentation of digital evidence [6]. Computer forensic tools have advanced from using command line environments to providing sophisticated graphical user interfaces that significantly enhance investigative activities. One useful feature is the presentation of the files in a spreadsheet-style format. The ability allows investigators to view all the files on a particular storage medium as well as information regarding each file. The details include file name, file creation date and time, logical size etc.

In fuzzy set theory [4], the fuzzy methods improve the effectiveness and the quality of the data analysis phase for crime investigation. The paper proposes a framework for applying fuzzy tools in digital investigation. The main goal is the extraction of expert-system-like rule sets based on fuzzy sets that can be presented to the experts in order to support them in their daily activities. This framework is conceived to be a potential starting point to a future standard framework for guiding the use of computational intelligence techniques in gathering digital evidence admissible in a court of law. Fuzzy clustering is used in [3] to detect the explanation of criminal activities for crime hot-spot areas and their spatial trends. Compared with two hard-clustering approaches (median and k-means clustering problem), the empirical results suggest that a fuzzy clustering approach is better equipped to handle crime spatial outliers. A two stage fuzzy decision classifier, using reference fuzzy set information, is used to create a text-independent Automatic Speaker Identification.

Finally, a framework of intelligent decision-support model based on a fuzzy self-organizing map (FSOM) network to detect and analyze crime trend patterns from temporal crime activity data is proposed in. The resultant model can support police managers in assessing more appropriate law enforcement strategies, as well as improving the use of police duty deployment for crime prevention. Fuzzy logic (together with neurocomputing and genetic algorithms) is one of the techniques of soft computing, i.e. computational methods tolerant to sub optimality, impreciseness
(vagueness) and partial truth and giving quick, simple and sufficiently good solutions.

III INFORMATION RETRIEVAL USING DOCUMENT CLUSTERING

Information retrieval (IR) is a technique of finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). The term “unstructured” refers to data which does not have clear. The field of information retrieval also covers supporting users in browsing or filtering document collections or further processing a set of retrieved documents. Given a set of documents, clustering is the task of coming up with a good grouping of the documents based on their extensions. There are some methods and algorithms used for document clustering explained below

Scatter/Gather cluster based approach:

In Scatter/Gather approach most useful algorithm is Hierarchical Algorithm. Scatter/Gather clustering algorithms can be used to compute a hierarchical clustering solution using a repeated cluster bisectioning approach [19, 20]. In this approach, all the documents are initially partitioned into two clusters. Then, one of these clusters containing more than one document is selected and is further bisected. This process continues n − 1 times, leading to n leaf clusters, each containing a single document. It is easy to see that this approach builds the hierarchical agglomerative tree from top (i.e., single all-inclusive cluster) to bottom (each document is in its own cluster). In the rest of this section we describe the various aspects of the partitional clustering algorithm that we used in our study. Hierarchical Algorithm divides into two parts first one is agglomerative and second one is divisive Clustering. In Agglomerative method merging of the most similar pairs of data points is done until one big cluster left. This is called a bottom-up approach. In divisive method splitting of large data is done. It is top- down approach. The concept is explained below diagrammatically.

Comparative study of K-means and Hierarchical clustering:

i. Computation Time:
   - Hierarchical clustering: O( m n^2 log(n) )
   - K-means clustering: O( k t m n )

Where t: number of iterations, n: number of objects, m-dimensional vectors,k: number of clusters.

ii. Memory Requirements:
   - Hierarchical clustering: O( mn + n^2 )
   - K-means clustering: O( mn + kn )

Hybrid algorithm:

Hybrid hierarchical clustering algorithm will be the integration of density based clustering and hierarchical cluster.

Density based clustering:

Clustering based on density (local cluster criterion), such as density-connected points. Each cluster has a considerable higher density of points than outside of the cluster. DBSCAN requires two parameters: ε (eps: Maximum radius of the neighborhood) and the minimum number of points required to form a cluster (minPts). It starts with an arbitrary starting point p that has not been visited from the group of point D. This point’s ε-neighbourhood is retrieved, and if it contains sufficiently many points less than or equal to MinPts it is called a core point and a cluster is started. Otherwise, the point is labelled as noise. Note that p might later be found in a sufficiently sized ε -environment of a different point and hence be made part of a cluster. If in the range of p's ε radius the number of the elements is less than MinPts, we can call p as the boundary, p is marked as noise node temporarily. Then, DBSCAN will dispose the next document in set D. As the first and the last step is the same as the threshold clustering method, so the two steps are ignored here. The main workflow of DBSCAN clustering is shown as follows:

Step1: Scan the point p in the set D one by one. Judge whether it has been clustered in a cluster. If so, skip this document, otherwise turn to Step 2. If the scan of all the digests in the set D is completed, then turn to Step 3.

Step2: Get the number of neighbours of p within the range of ε. This step is done by calculating the distance between p and all other document. The calculation of distance between two mails as shown in previous section include the average of the smallest three distances between all of the two document. If the number isn’t less than MinPts, set the document p as the core document, then scan each of the neighbours of p and turn to Step 1 for recursive queries. Finally, all elements from recursive clustering are marked as a new cluster, and then turn to Step 1 to dispose the next document of set D.

Step3: Scan all the documents in set D, if a document isn’t in a cluster, it should be marked as a noise

---

Figure 1: Hierarchical concept
Major features of DBSCAN algorithm are
i. Discover clusters of arbitrary shape
ii. Handle noise
iii. One scan

Hierarchical algorithm is same as explained in Scatter/Gather based approach.

Automatic cluster labeling method:
The search engines commonly build a very large centralized database to index a portion of Internet and help to reduce the information overload problem by allowing a user to do a centralized search. However, they also bring up another problem: too many web pages are returned for a single query. To find out which documents are useful, users have to sift through hundreds of pages to find out that only a few of them are relevant. In the proposed approach, one way to tackle this problem is to cluster the search result documents based on their extension similarity so that users can scan a few coherent groups instead of many individual documents. The first method uses a $x^2$ test of significance to detect different word usage across categories in the hierarchy which is well suited for testing dependencies when count data is available. The second method selects words which both occur frequently in a cluster and effectively discriminate the given cluster from the other clusters. Indexing and the technique used for automatic labelling provide the fast and efficient related data retrieval.

IV. CONCLUSION
Many data mining techniques have been proposed in the last decade. The outcome of the information retrieval using document clustering for forensic analysis is the number of labeled cluster, which provides the better visualization in the form of frame which shows the most relevant data present in the particular cluster. K-mean algorithm is effective; it minimizes the squared-error criteria computationally efficient and does not require the user to specify many parameters [6]. K-medoid could be more robust to noise and outliers as compared to k-means because it minimizes a sum of general pair wise dissimilarities instead of a sum of squared Euclidean distances.

The hybrid algorithm chooses the most accurate label to the cluster. And provides
1. Fast and efficient analysis.
2. As system analyzes all the evidences gathered will be more accurate.
3. It has the potential to increase the performance of forensic analysis with speed up the computer inspection process.

In this paper the high quality clustering algorithm, and methods used which will provide the automatic labelling to the cluster, and will provide the indexing to text, doc, and pdf file.

VI. REFERENCES


