Bharathi K.S *et al*, Journal of Computer - JoC,
Available Online at: www.journal.computer
Vol.1 Issue. 2, July- 2016, pg. 1-8 **ISSN: 2518-6205 (Online)**

Automatic Labeling of Text Document Clusters using Singular Value Decomposition

Bharathi K.S[1], Dr. Asha T.[2]
[1]Department of Computer Science, Bangalore Institute of Technology, Bangalore, Karnataka, India
[2]Professor, Department of Computer Science, Bangalore Institute of Technology, Bangalore, Karnataka, India
[1] bharathikspatla@gmail.com; [2] asha.masthi@gmail.com

Abstract— Analysis of text documents is difficult due to unstructured information it contains. Clustering of these documents helps to improve analysis under consideration. Most widely used text mining methods such as partitional algorithm k-means and hierarchical clustering methods based on linkage criterion such as single link, average link and complete link are used in this paper. The clusters are then labeled by using singular value decomposition method in a mathematical way. The labeling of the clusters makes the analyst job easier by quick capture of the cluster summary on the screen. Relative validity index is used to determine the efficiency of clustering process. It is used for estimation of number of clusters at which the process is efficient. Cluster analysis is very useful for forensic domain wherein crime investigations are performed to analyze the information from seized digital devices.

Keywords – Clustering, forensic domain, text mining.

I. INTRODUCTION

The volume of data in the digital world continues to grow exponentially every year. This large amount of data has a direct impact on computer forensics. This application domain involves examining hundreds of thousands of files per computer. It will be tedious for an expert examiner to perform manual analysis and interpretation of data. Hence, methods for automated data analysis, like clustering techniques are of huge importance. Cluster analysis is a works directly on the data, thus belonging to the class of unsupervised learning. Clustering algorithms are most widely used in data mining for exploring and analyzing data, where users will not have prior knowledge about the data. These methods ensure that the objects within a cluster are more alike than the objects in a different cluster.

Categories of clustering algorithms are determined by the type of model used. There are approximately 100 algorithms published on clustering. An overview of algorithms can be found in [1]. Most suitable clustering algorithm for a given problem needs to be chosen frequently on an experimental basis. Observation from studies is that, there is no objectively correct clustering algorithm; it depends on perception of people. In addition to clustering the documents, labeling helps the examiner's job much easier, by browsing the clusters based on the label. By this process one can avoid the tedious task of examining every document individually.

In this paper four representative algorithms are chosen in order to illustrate the potential of the proposed approach. Those algorithms are the partitional k-means, the hierarchical single link, complete link, average link algorithms. Clusters formed are then summarized to provide a meaningful label for it.

Clusters are illustrated in the form of central vector in partitional clustering that may not be necessary to be member of the given dataset. In this type of clustering if the number of clusters is fixed by the user say 'k' then

Bharathi K.S *et al*, Journal of Computer - JoC,
Available Online at: www.journal.computer
Vol.1 Issue. 2, July- 2016, pg. 1-8 **ISSN: 2518-6205 (Online)**

it is called as k-means clustering. It is also called as Lloyd's algorithm. Here the cluster centroids are calculated and objects are assigned to the nearest cluster center.

Based on the distance between objects, the hierarchical clustering connects objects to form clusters. At a larger extent the cluster can be described here as the maximum distance required connecting the parts of it. Different clusters will form at different distances, which are represented using a tree diagram called as dendrogram, thus naming hierarchy as a true meaning. In a dendrogram, y-axis indicates the distance at which the clusters merge, whereas the x-axis indicates the objects that are clustered. A hierarchy of clusters is formed that merge the clusters with each other at certain distances.

II. LITERATURE SURVEY

Very few studies have been present describing the use of clustering algorithms in Computer Forensics field. Many among them describe the use of classic algorithms for clustering of data. B.Fei and J.Eloff [2] have discussed on the application of self-organizing map(SOM) that will support the decision making process in forensic analysis along with demonstration of SOM visualization for the interpretation of data. Several applications of SOM are identifying associations in data, classification of data, clustering of data, and discovering latent patterns in data useful in forecasting.

R. Hadjidj and M. Debbabi [3] have presented e-mail forensic analysis software tool, developed by amalgamating existing state-of-the-art statistical and machine learning techniques accompanied with social networking techniques. In the proposed structure authors engulf two proposed authorship attribution methods.

F. Iqbal and H. Binsalleeh [4] have formally defined the problem of authorship attribution and formulated a new notion of write-print based on the concept of frequent patterns. The problem has been refined into three sub problems: As a first step, need to identify the write-print of each suspect. Next step is to determine who the author of the malicious e-mail. Finally need to extract evidence for supporting the conclusion on authorship.

S. Decherchi and S. Tacconi [5] have proposed a methodology with effective digital text analysis strategy, which relies on clustering based text mining techniques. Publicly available Enron datasets are used for the experimentation purposes. Investigational activity is performed on seized digital devices, wherein they can provide valuable information and evidences about facts and/or individuals.

The authors have designed the methodology in two phases that aims at generating a collection of raw text file from information stored in digital devices. As the first step design for bit- stream acquisition and early analysis is carried out; and the second step consists textual information extraction from relevant files previously found.

III. SYSTEM OVERVIEW

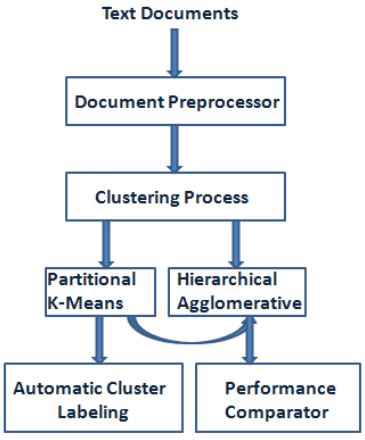

Fig. 1. System architecture

Bharathi K.S *et al*, Journal of Computer - JoC,
Available Online at: www.journal.computer
Vol.1 Issue. 2, July- 2016, pg. 1-8 **ISSN: 2518-6205 (Online)**

This section gives the high-level overview of the automatic cluster labeling system. Fig. 1 shows the architecture of the proposed method.

The document preprocessing is carried out for the input text documents. As the first step, stop words are removed. Few of those stop words are pronouns, prepositions and document metadata which is immaterial have been removed. Next step is to represent the documents in a vector space model in a traditional statistical approach for text mining. Here, each document is represented by a vector comprised of the frequencies of occurrences of words. Each word is defined as delimited alphabetic strings, whose number of characters is between 4 and 25.

Estimate of the number of clusters for which the clustering is said to be efficient. Most common approach consists of getting a set of data partitions with different numbers of clusters. And as a next step, run the algorithm for the different partitions. Finally choose the partition that gives the best result according to a quality criterion a relative validity index. The silhouette method is used here as the relative validity index. The set of partitions would result directly from a hierarchical clustering dendrogram. And in the case of partitional algorithm (e.g., k-means) it is obtained from multiple runs starting from different numbers and initial positions of the cluster prototypes.

The clusters formed from k-means are labeled automatically to give each cluster a meaningful summary. Content of the documents belonging to each cluster is combined together in a text document and summarization method if applied on them. Singular value decomposition method is used here for summarizing the cluster with the best sentence describing the whole cluster, thus forming a meaningful label.

The performance of the algorithms is compared with respect to the silhouette method of relative validity index as well as adjusted random index.

A. K-Means Clustering

K-Means is the simplest clustering algorithm which works based on the partitions. It is an unsupervised learning method in text mining domain. The 'K' in the very name of algorithm is the number of clusters the user wants to create for the set of input documents.

The procedure followed in the algorithm is explained in form of steps are as follows:

Input: The number of clusters 'K' and Term Variance
Matrix generated from the preprocessing.

Output: A set of K-clusters.

1: Load the term matrix obtained from pre-processing
2: Convert the term matrix into vector of double points from the strings
3: Read the K value
4: Randomly select cluster centroids from the term matrix based on K-value
5: **for** each document vector compute the distance from to each of the centroids defined
6: **end for**
7: Group the document vector with the respective clusters based on to which cluster centroid point it is nearest
8: **for** each cluster formed in first iteration
Calculate the new centroid by taking average of all the document vectors belonging to that particular cluster
end for
9: Repeat the steps 5 through 7 until the centroid point does not change, which indicates centroid points are stable
10: Return K clusters

Bharathi K.S *et al*, Journal of Computer - JoC,
Available Online at: www.journal.computer
Vol.1 Issue. 2, July- 2016, pg. 1-8 **ISSN: 2518-6205 (Online)**

Euclidean distance formula is used here for the distance calculation between a point **X** (**X**1**, X**2**,** etc.) and a point **Y** (**Y**1**, Y**2**,** etc.) given as:

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

B. Estimating the number of clusters

Consistency of the clustering is directly proportional to the number of clusters and measured in variety of ways. This paper uses Silhouette method for validation and interpretation of consistency within the clusters of data. This is a widely used relative validity index approach to estimate the number of clusters. It provides a representation in graphical way to show how well each object lies within its own cluster. Thus cohesiveness of an object is measured i.e. similarity of an object to its own cluster compared to other cluster.

The indication of high value means the object is well matched to its own cluster and poorly matched to neighboring clusters. Obtaining high value to most of the clusters indicate that the clustering configuration is good. Instead, if low or negative value is obtained for most of the objects then the clustering is said to be inappropriate. And it might have too many or too less clusters.

Consider an object *i* belonging to cluster A. Let *a(i)* be the average dissimilarity of *i* to all other objects of A. Taking into account another cluster C, d(i, C) is the average dissimilarity of i to all other objects of C. When the d (i, C) for all the clusters C $\neq$ A is computed, the smallest one is selected, i.e., b (i) = min d (i, C), C $\neq$ A. This value stands for the dissimilarity of i to its neighboring clusters. The formula for Silhouette calculation s(i) for a given object i, is given by:

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

For K-Means algorithm the *s(i)* is calculated from multiple runs of the algorithm starting from different initial positions of cluster prototypes and different numbers. i.e., different values for K are given in multiple runs. Whereas, in case of hierarchical clustering method, the best partition results directly from the dendograms (tree diagram to illustrate the arrangement of clusters).

C. Hierarchical Agglomerative Clustering

Hierarchical cluster analysis is a method which seeks to build hierarchy of clusters. The resultant hierarchy is in the form of a tree diagram called as dendrogram. Partitions are formed in series, starting from the each object into an individual cluster to until all objects contained into a single cluster.
Approaches adopted in hierarchical clustering are of two types:

- **Agglomerative:** Follows 'bottom up' approach, starts with each object is contained in its own cluster and on the similarity basis pair of clusters merged as one moves up the hierarchy.
- **Divisive:** Follows 'top down' approach, starts with all objects are contained in a single cluster and recursive splits are performed as one move down the hierarchy.

In this paper, the agglomerative cluster analysis is used which proceeds by series of fusions of the '*n*' objects into partitions like P_n, P_{n-1}, ... , P_1. The initial partition P_n consists of n single object clusters. The last partition P_1, consists of single group consisting of all n objects. At each stage of merge two closest clusters are joined together. The closeness of the clusters is measured by the similarity metric. Shape of the clusters is influenced by the type of metric chosen, because the objects seem close to one another in one metric and farther away according to another. As an example, the distance between the point(1,0) and the origin(0,0) is always 1

Bharathi K.S *et al*, Journal of Computer - JoC,
Available Online at: www.journal.computer
Vol.1 Issue. 2, July- 2016, pg. 1-8 **ISSN: 2518-6205 (Online)**

according to usual norms, but the distance between the point(1,1) and the origin(0,0) differs based on distance metric i.e. 2 under Manhattan distance, 1 under maximum distance, or $\sqrt{2}$ under Euclidean distance. Euclidean distance measure is used in this paper. Formula for distance between 2 objects *a* and *b* is:

$$\|a - b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$$

The algorithm is described as follows:

Input: N text documents to be clustered, an NxN distance (similarity) matrix.

Output: Dendrogram displaying the resultant cluster assignment for the documents.

Algorithm:

1: Start by assigning each document to its own cluster, so that if there are N documents then N clusters are formed each containing just one document.

2: Locate the closest pair of clusters and merge them into a single cluster, the number of clusters is reduced by one now. (Euclidean distance measure used for the similarity check)

3: Compute the distance between the newly formed clusters from merge and each of the old clusters.

4: Repeat steps 2 and 3 until a single cluster of size N is formed containing all the documents.

5: Return the dendrogram of clusters.

In the algorithm mentioned above, step 3 can be done in different ways, which is what distinguishes the below clustering methods in agglomeration.

1. Single linkage clustering: This is also called a minimum linkage method. Here the distance between one cluster and another is considered to be equal to the shortest distance from any object of one cluster to any object of another cluster.

2. Average linkage clustering: This method is also called as mean linkage clustering. Here distance between one cluster to another is considered to be equal to the average distance from any object of one cluster to any object of another cluster.

3. Complete linkage clustering: This is also termed as maximum linkage clustering. Here the distance between one cluster and another is considered to be equal to the longest distance from any object of one cluster to any object of another cluster.

The result of all these methods is dendrogram, the branching diagram representing the relationships of similarities among the group of objects. Fig 2 shown below depicts a dendrogram for the text documents input.

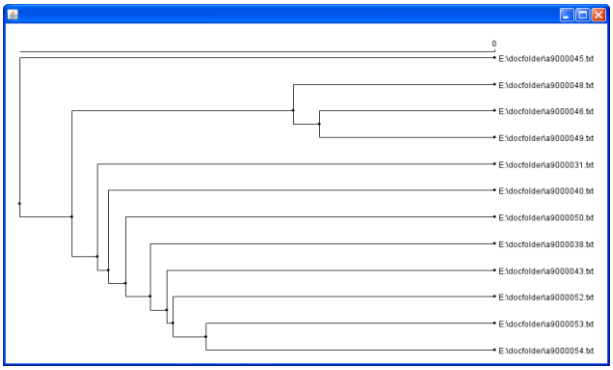

Fig.2. Dendrogram for the document clusters

Bharathi K.S *et al*, Journal of Computer - JoC,
Available Online at: www.journal.computer
Vol.1 Issue. 2, July- 2016, pg. 1-8 **ISSN: 2518-6205 (Online)**

Here the rows of x-axis represent the objects to be clustered, and the distance on y-axis refers to the distance measure between the objects. Each branch is called as a *clade* and the terminal end of each clade is a *leaf*.

Arrangement of the clades illustrates which leaves are most similar to each other. The measure of how similar or different the cluster is is revealed from the height of the branch points. Suppose greater the height, the greater the difference. A dendrogram can be used to represent the relationships between any kinds of objects as long as we can measure their similarity to each other.

D. Cluster labeling

Cluster labeling is the process of deciding human readable, descriptive labels for the clusters produced by the document clustering algorithm. Generally, standard clustering algorithms do not produce any labels to the clusters. Instead, the analyst has to browse each cluster i.e. the documents it contains to get the information. If the clusters are labeled then based on the label analyst can access whichever topic he is interested in.

In this paper Singular Value Decomposition (SVD) method is used to label the clusters. The formula is:

$$A = U\Sigma V^T$$

A: Input matrix (m x n)

U: Words x Extracted Concepts (m x n)

Σ: Scaling values, diagonal descending matrix (n x n)

V: Sentences x Extracted Concepts (n x n)

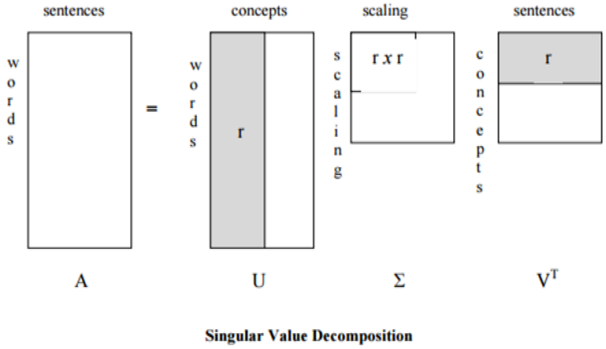

Fig.3. Decomposition of matrix in SVD

Inputs to the process are **c**lustered text documents. The detailed steps in the SVD process are as shown below:

- **Step 1:** Club the contents of each cluster into an individual text file.
- **Step 2:** Pre-process the each text file formed from step 1, to separate each sentence.
- **Step 3:** Create a Term Sentence matrix M.
- **Step 4:** Create a SVD matrix from M, to decompose it into U, Σ and V.
- **Step 5**: Select top K sentences.
- **Step 6**: Return the cluster label.

In brief the above process is explained this way. The clusters formed from the document clustering algorithm are taken as the input. Then each cluster contents are copied into a text file for further processing.

Bharathi K.S *et al*, Journal of Computer - JoC,
Available Online at: www.journal.computer
Vol.1 Issue. 2, July- 2016, pg. 1-8 **ISSN: 2518-6205 (Online)**

So, now in each text file sentences are present which belong to the cluster. It is required to separate these sentences. The pre-processor does this job of separation of each sentence and to create a term sentence matrix.

Term sentence matrix is now decomposed into three different matrices which are U, Σ and V as per fig. 3. The matrix V is transposed to get V^T. From this transposed matrix choose the best sentence which describes the cluster content without much distortion.

E. Performance comparison

The performance k-means is measured with a plot of varied k values and their silhouette measure for each of the k value. The clustering is said best when the data partition has the maximum silhouette value as an average. Fig 4 illustrates the graph plot based on silhouette measure.

Between the k-means and hierarchical clustering, to choose the best one adjusted random index (ARI) method is used. Considered the partition P which is obtained by running the hierarchical linkage algorithm, and compared it with the reference partition R. The greater value suggests better is the partition. Fig 5 illustrates the graph plot based on ARI measure.

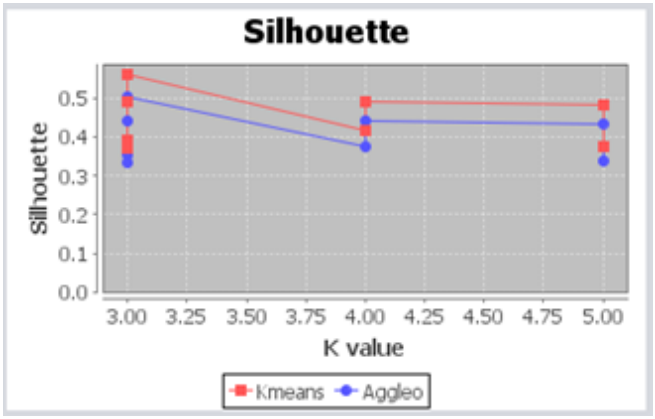

Fig. 4. Performance plot using silhouette values

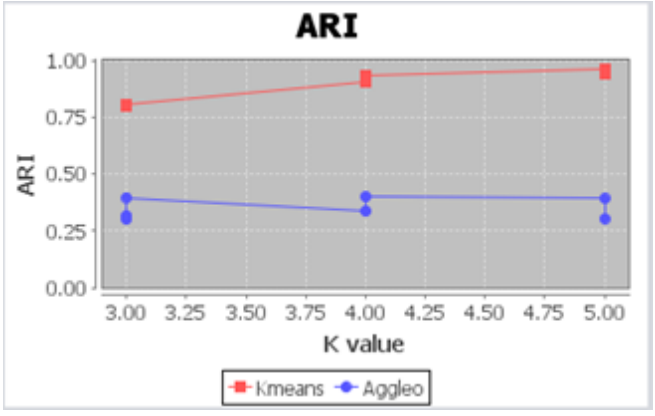

Fig. 5. Performance plot using ARI values

IV. CONCLUSION

Hierarchical clustering algorithms known as single, average and complete link are presented in this paper which yields the best results. The dendrograms produced by these algorithms are proven to be very useful in text analysis. They provide a graphical view of how documents are better summarized for analysis. K-means does offer good results on proper initialization of the cluster count k. Silhouette method of relative validity index gives the best measure of k value. The computational efficiency is thus measured for the k-means algorithm. Assignment of labels to clusters enables the analyst to identify the content of each cluster very quickly, thus avoiding the time spent in searching the individual documents and their contents. For the smaller datasets hierarchical clustering is more efficient where as for the larger datasets k-means is computationally efficient.

Labeling the hierarchical clusters is worth of investigation further. Because along with the graphical representation, can also view the actual content summary which is very useful for analysis. As the future work, one can perform clustering of the documents that contain images.

REFERENCES

[1] Https://en.wikipedia.org/wiki/List_of_algorithms#Statistics.

[2] B. K. L. Fei, J. H. P. Eloff, H. S. Venter, and M. S. Oliver, "Exploring forensic data with self organizing maps," in Proc. IFIP Int. Conf. Digital Forensics, 2005, pp. 113–123.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

