ISSN 2395-1621

Document Clustering using k means Algorithms

[#1]Pradeep T. Bhutare, [#2]Chaitanya P. Shewale, [#3]Ankush B. Bharekar,
[#4]Akshay S. Pandav

[1]pradeepbhutare@gmail.com
[2]chaitanyashewale99@gmail.com
[3]bharekarankush91@gmail.com
[4]akshay.pandav@gmail.com

[#1234]Dept of Computer Engineering
NBN Sinhgad School of Engineering
Pune, India

ABSTRACT

In computer forensic analysis, hundreds of thousands of files are saving into the digital formats; these files are usually into unstructured textdata format. The analysis of that unstructured data is very difficult to examine by computer examiners.In this paper, we build an integrating system for extracting an automated document clustering method for grouping whole data into cluster format. We present an approach that applies document clustering algorithms to forensic analysis of computers seized in police investigations.In data mining, document Clustering is based on K-means and K-medoid algorithm for clustering the documents by using centroid data point. In the context, we provide security to our database server.

Keywords— **Clustering, text mining.**

ARTICLE INFO

Article History

Received :28th September 2015

Received in revised form :

6th October 2015

Accepted : 8thOctober , 2015

Published online :

12th October 2015

I. INTRODUCTION

The Data size is increasing day by day. This large amount of data is directly or indirectly impact on computer forensics. To analyze data manually required very large time and it is tedious work. Manual data analysis is done by human there is chance of getting errors in analysis. So we proposed automated data analysis. To solve these problem by using K-means and K-medoid algorithms (clustering analysis techniques). The main purpose of clustering is to locate information and in the present day context, to locate most relevant electronic resources. Clustering is a method in which we make cluster of data objects that are somehow similar in characteristics. Our dataset may be labeled or unlabelled. Labeled dataset analysis is easy as compared to unlabeled dataset analysis. The concept behind clustering algorithms is that data objects within a valid cluster are very similar to each other than they are to objects belonging to a different cluster. Thus, once a data partition has been induced from data, the examiner might initially focus on reviewing representative documents from the obtained set of clusters. Then, after this preliminary analysis, she/he may eventually decide to scrutinize other documents from each cluster. By doing so, one can avoid the hard task of examining all the documents but, even if so desired, it still could be done.

Fig. Forensic Analysis

From the figure forensic analysis do collect evidence. Look in the logs, repairing the system, tracking the hacker, keystroke loggers, determining the spy. Solution of this problem is allowable, authentic, accurate and complete.

Forensic acquisition puts relevant data into the preliminary phase. It is the selective storage. It involves two steps viz.

1. Textual information extraction.
2. Textual data analysis

Digital investigation important for textual evidence. Examples of investigations are e-mails, instant messaging, word processing documents n/w activity logs. In physical level every byte search at the digital evidence. Second identifies the specific text string. It moves to the next investigation. Text string search have Information Retrieval (IR) overhead, and make noise. Small device have a capacity of 80 GB, these problems solved two solution. First one have decrease the number of irrelevant search hits. Second one has present the search hits a manner which enables the investigator to find the relevant hits more quickly. Indexing algorithms and ranking algorithms combines fail in the first solution. At the second solution it works. Main function is improving the (IR) information retrieval.

II. RELATED WORK

Document clustering is a technique in which, the information that is coherent is actually stored together . For maximizing the efficiency of find and the retrieval in database , the number of disk accesses is to be minimized. In clustering, since the objects of same properties are placed in single class of objects, a single access to the disk can retrieve the entire class of objects. If the clustering takes place in some predefined algorithmic space, we may create population into subsets with unique characteristic, and then decrease the problem space by acting on only a cluster head from each cluster. Clustering is eventually a process of decreasing a mountain of data to formatted groups. For relating and computational simplification, these groups may consist of "same" items. There are two approaches to document clustering, unfussy in information retrieval; they are known as terms and item clustering. Terms clustering is a method, which groups duplicate terms, and this grouping minimizes, noise and gaining frequency of assignment.

There are only a very few studies reporting the use of grouping or clustering algorithms in the *Computer Forensics* domain. Necessary, most of the studies explain the use of classic algorithms for clustering data. e.g., Expectation-Maximization (EM) for improper learning of Gaussian Mixture Models , K-means. These algorithms have known properties and are mostly used in practice. For instance K-means, K-medoids, Single Link, Complete Link and Average Link, can be seen as particular cases of Expectation maximization . The literature on *Computer Forensics* only reports the use of algorithms that consider that the number of clusters is well known and fixed *a priori* by the user. Goal at relaxing this consideration, which is often improper in real life applications, a common approach in other field involves estimating the number of clusters from data. Necessarily, one induces different data partitions and then assesses them with a relative validity index in order to estimate the better cost for the number of clusters. In this section, we discuss related work on document clustering and clustering algorithm.

A.DOCUMENT CLUSTERING

Extraction and fast information retrieval or filtering. Related to data clustering. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories. Document clustering involves descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document cluster is generally considered to be a centralized process. Example of document clustering is web document clustering. Application of document clustering can be categorized to two types that are online and offline.

B.PRE-PROCESSING STEPS

Stop words doing before clustering algorithm. It defines remove of prepositions, pronouns, articles and irrelevant document, Meta data. It enables snowball steaming. Text mining using traditional satisfies approach. Identifies vector space model. In this model [4] have effectiveness, efficiency, clustering algorithm. Transformation vector selects a number of attributes that have been used namely, cosine-based distance and leven steins-based distance.

C.CLUSTERING ALGORITHM

Machine learning data mining fields using Cluster Ensemble Based Algorithm (CSPA)Mediods have centroids. This property makes it particularly interesting for applications in which1) centroids cannot be computed, and2) distances between pairs of objects are available-MEANS AND k-mediods are sensitive to initialization considering partitioned algorithms. Every partition represented by the dendogram subsequently choosing best results. CSPA algorithm essentially finds a consensus clustering from a cluster ensemble formed by a set of different data partitions. After applying clustering algorithms to the data a similarity matrix computed. Each element of this matrix represents pair-wise similarities between objects. The similarity between two objects is simply the fraction of the clustering solutions in which those two objects lie in the same cluster.

III. IMPLEMENTATIONS K-MEANS ALGORITHM

K-means is one of the simple and easy unsupervised learning algorithms that partition feature vectors into k clusters so that the within group sum of squares is minimized. K-means clustering is a method of vector quantization originally from signal processing that is popular for cluster analysis in data Mining. From the figure K-Means follows a simple way to classify a given data set and looks like

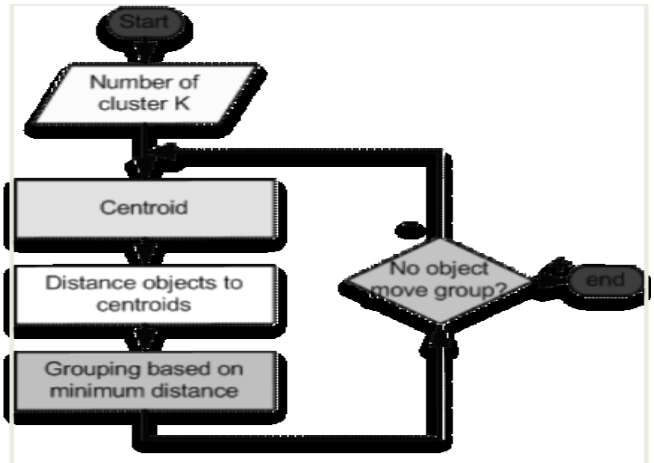

Fig.2. K-Means process

STEPS

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, *'c_i'* represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

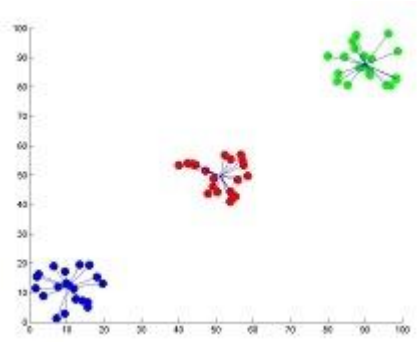

Fig 3. Showing the result of k-means

IV. EQUATIONS

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled. A colon is inserted before an equation is presented, but there is no punctuation following the equation. All equations are numbered and referred to in the text solely by a number enclosed in a round bracket (i.e., (3) reads as "equation 3"). Ensure that any miscellaneous numbering system you use in your paper cannot be confused with a reference [4] or an equation (3) designation.

V. FIGURE AND TABLE

To ensure a high-quality product, diagrams and lettering MUST be either computer-drafted or drawn using India ink.

Figure captions appear below the figure, are flush left, and are in lower case letters. When referring to a figure in the body of the text, the abbreviation "Fig." is used. Figures should be numbered in the order they appear in the text.

Table captions appear cantered above the table in upper and lower case letters. When referring to a table in the text, no abbreviation is used and "Table" is capitalized.

VI. CONCLUSION

A conclusion section must be included and should indicate clearly the advantages, limitations, and possible applications of the paper. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

ACKNOWLEDGEMENT

An acknowledgement section may be presented after the conclusion, if desired

REFRENCES

[1] Navin Kumar Tyagi1, A.K. Solanki2& Sanjay Tyagi3," An algorithmic approach to data processing in web usage mining ",International journal of information technology and knowledge management.

[2] Srinivasa K G * , Venugopal K R 1 and L M Patnaik 2, "Feature Extraction using Fuzzy C - Means Clustering for Data Mining Systems". IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.3A, March 2006 Alexey Tsymbal 1,3 , Seppo Puuronen 1 , Mykola Pechenizkiy 2 , Matthias Baumgarten 3 , David Patterson 3, "Eigenvector-based Feature Extraction for Classification". FLAIRS-2002.

[3] Yu Tao, Vallipuram Muthukkumarasamy, Brijesh Verma and Michael Blumenstein, "A Texture Feature Extraction Technique Using 2D–DFT and Hamming Distance". Compuatational intelligence and multimedia appliactions ,2003.

[4] Fabian M¨orchen ∗, "Time series feature extraction for data mining using DWT and DFT (November 5, 2003)".

[5] Ying Zhao and George KarypisC, "Comparison of Agglomerative and Partitional Document Clustering Algorithms ".

[6] K.Sasirekha,P.Baby,"Agglomerative Hierarchical Clustering Algorithm- A Review". Proceedings of the eleventh international conference on Information and knowledge management 2002

[7] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, "An Efficient kMeans Clustering Algorithm: Analysis and Implementation". IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002.

[8] B. S. Everitt, S. Landau, and M. Leese, Cluster Analysis. London, U.K.: Arnold, 2001.

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

[10] Osama Abu Samu computer science department yarmouk university,"comparison between data clustering algorithm", may 2 2007.

[11] paresh chandra barman,Md. Sipon Miah, Bikash Chndra Singth,"Feature extraction clustering in text miningusing NMF basis probability", Ulab journal of science and engineering ,november 2,2011.

