ISSN 2395-1621

DOCUMENT CLUSTERING BY K-MEANS: CRIMINAL DATABASE

[#1]Prof. Rahul Samant, [#2]Naman Mathur, [#3]Prateek Sonaje, [#4]Omkar Thatte, [#5]Sukumar Zurange

rahul.samant@sinhgad.edu
namanmathur16@gmail.com
pksonje44@gmail.com
omkart10@gmail.com
sukumarzurange@gmail.com

Information and Technology Engineering,
NBN Sinhgad School of Engineering
Ambegaon (bk), Pune- 411041.

ABSTRACT

In computer forensic analysis, hundreds of files are usually examined. Much of the data in those files consists of unstructured text, whose analysis by computer examiners is difficult to be performed. In this context, automated methods of analysis are of great interest. In particular, algorithms for clustering documents can facilitate the discovery of new and useful knowledge from the documents under analysis. We present an approach that applies document clustering algorithms to forensic analysis of computers seized in police investigations. We illustrate the proposed approach by carrying out extensive experimentation with K-means algorithm. Experiments have been performed with different combination of parameters.

Keywords— K-means, automated methods, clustering

ARTICLE INFO

Article History

Received: 28th May 2018

Received in revised form :

28th May 2018

Accepted: 31st May 2018

Published online :

3rd June 2018

I. INTRODUCTION

In our particular application domain, it usually involves examining hundreds of thousands of files per computer. This activity exceeds the expert's ability of analysis and interpretation of data. Therefore, methods for automated data analysis. Like those widely used for machine learning and data mining are of paramount importance. In particular, algorithms for pattern recognition from the information present in text documents are promising as it will hopefully become evident later in the paper. The concept of clustering has been around for a long time. It has several applications, particularly in the context of information retrieval and in organizing web resources.

There is an exponential increase in the number of digital data and text documents. Thus, it is very difficult to organize these large collection of text documents in an effective way and locating interesting information or patterns [1] has become a vital task. To speed up the searching process for similar documents for finding required documents, document clustering is a method widely used. Clustering organize a large set of documents into a number of similar clusters [2]. So, the documents in the same cluster are more similar to one another than to documents in another cluster.

K-Means algorithm is used by our document clustering technique which is explained in **detail in section II.** Pre-processing is done to convert the words to their base form, to remove stop words, duplicate words before applying vector space model to the text documents.

Clustering algorithm identifies [4] the accurate data from the analysis of little knowledge or no prior knowledge data. Computer forensics have unlabelled [4] objects. In previous analysis have labelled object design or supervised learning setting. Preliminary analysis defines data partition from the data and expert examiner only focus on reviewing representative documents from the obtained set of clusters. In preliminary process avoid the hard work of examiners. After finding relevant document the examiner could pass the analysis of the other document to investigation. Text clustering [1] in digital evidence defines information and data of investigate value. That are stored in digital device or transmitted in digital device. This type of seized device established by digital forensic analysts. It deals with massive amount of data and increasing capacity of data. Investigate activity have two aspects is acquisition and retrieval information extracted from digital device. Forensic acquisition puts most relevant data into the preliminary phase. It is the selective storage. It involves two steps. That is textual information extraction have digital device text files and early analysis (bit-stream acquisition) and textual

data analysis via clustering based text mining tool identifying, tracking, extracting and classifying discovering. Text clustering for forensic analysis based on dynamic adaptive clustering model. Digital investigation important for textual [3] evidence. Examples of investigations are e-mails, internet browsing history, 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 devices have a capacity of 80gb. These problems are with solved two solutions. First one includes decrease in 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. Fuzzy Methods [2] defines crime data analysis and utilization important for intelligence. Intelligence based approach for law enforcement. Must it necessity. analysis related to type of intelligence. Forensic intelligence defines the accurate, timely and useful products of logically processing forensic case data. Results of forensic intelligence have discipline specific activities. Information technology used to produce the information sets and digital evidence the methods from Artificial intelligence. Artificial intelligence defines the science and engineering of making intelligent machine. Computational intelligence includes a number of computational methods as neural networks, fuzzy systems. Fuzzy methods improve the quality of data analysis phase. Fuzzy tools apply digital investigation. Forensic analysis evidence has computational intelligence methods and techniques and assigning analysis. Evolutionary algorithm and genetic algorithm solve the problem of missing persons. Writer identification solves the problem of hand writing analysis. Fuzzy methods important a role and learning complex data structures and patterns classifying them to make intelligent decisions. Comparing k-means and k-medidos it works best.

II. LITERATURE SURVEY

K- MEANS

k-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centres, one for each cluster. These centres should be placed in a cunning way becauseof different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centre. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycentre of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centre. A loop has been generated. As a result of this loop we may notice that the k centres change their location step by step until no more changes are done or in other words centres do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\|x_i - v_j\| \right)^2$$

where,

'||xi - vj||' is the Euclidean distance between xi and vj.

'ci' is the number of data points in ith cluster.

'c' is the number of cluster centres.

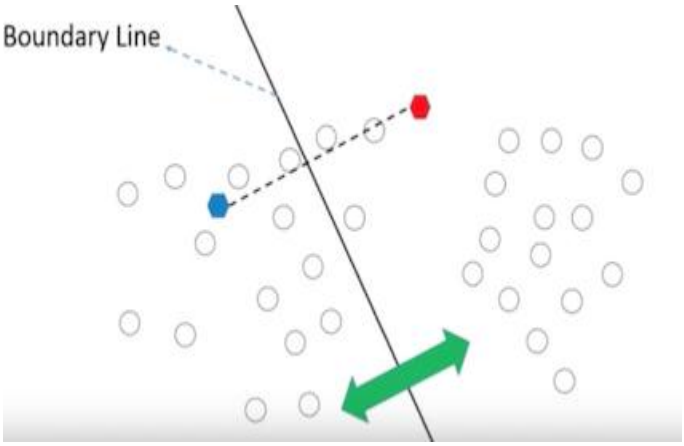
As, you can see, k-means algorithm is composed of 3 steps:

Step 1: Initialization

The first thing k-means does, is **randomly** choose K examples (data points) from the dataset (the 4 green points) as initial centroids and that's simply because it does not know yet where the center of each cluster is. (a centroid is the center of a cluster).

Step 2: Cluster Assignment

Then, all the data points that are the closest (similar) to a centroid will create a cluster. If we're using the Euclidean distance between data points and every centroid, a straight line is drawn between two centroids, then a perpendicular bisector (boundary line) divides this line into two clusters.



Step 3: Move the centroid

Now, we have new clusters, that need centers. A centroid's new value is going to be the mean of all the examples in a cluster.

We'll keep repeating step 2 and 3 until the centroids stop moving, in other words, K-means algorithm is converged.

K-means clustering comes under partitioning clustering algorithm. It partitions given data into K clusters. Several other clustering algorithms are proposed for dealing with

document clustering task including Novel algorithm for automatic clustering suggested how clustering is done automatically, Improved partitioning K-means algorithm presented new method for initializing centroids. Ontology based k-means algorithm presented how ontological domains are used in clustering documents.

STOP WORDS

When working with text mining applications, we often hear of the term “stop words” or "stop word list" or even "stop list". Stop words are basically a set of commonly used words in any language, not just English. Stop words are critical to many applications because if we remove the words that are very commonly used in a given language, we can focus on the important words instead. For example, in the context of a search engine, if we search query is “how to develop JAVA applications”, If the search engine tries to find web pages that contained the terms “how”, “to” “develop”, “JAVA”, “applications” the search engine is going to find a lot more pages that contain the terms “how”, “to” than pages that contain information about developing JAVA applications because the terms “how” and “to” are so commonly used in the English language. So, if we disregard these two terms, the search engine can actually focus on retrieving pages that contain the keywords: “develop” “JAVA” “applications” – which would more closely bring up pages that are really of interest.

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.

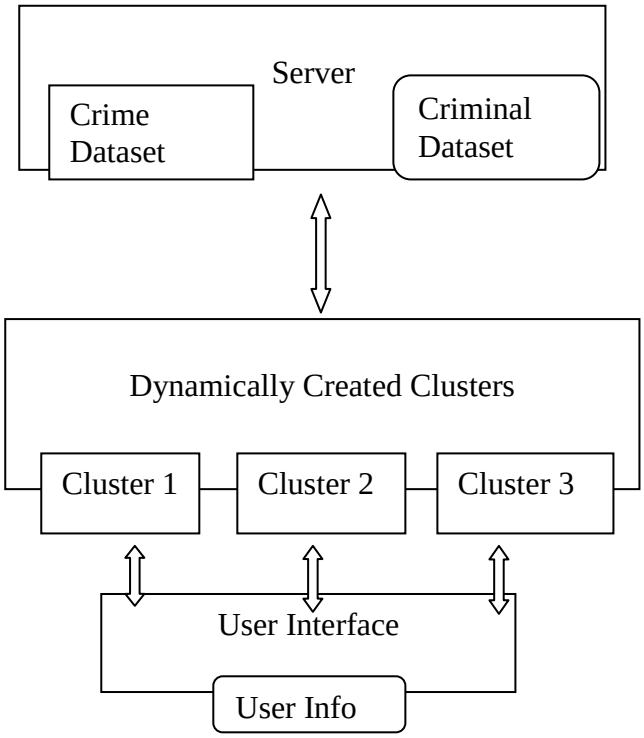
CLUSTERING ALGORITHM

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 dendrogram 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 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. SYSTEM ANALYSIS PROPOSED ARCHITECTURE

Traditional K-means algorithm has problems like K-means algorithm works well for a few documents. But when number

of documents is increased, it could not cluster the documents. It is not automatic. The number of clusters needs to be specified in advance. To resolve these issues, we started with improved document clustering using k-means algorithm and proceeded with and algorithm explained in section **II**. The main features of proposed algorithm are capable of handling large documents, since it is partition-based algorithm and it is also automatic. Proposed algorithm considers feature vectors for automatically clustering the documents present in a corpus collection.



IV. CONCLUSION

In this paper, collection of text document is done and then the information is extracted in that document in brief formats. It reduces the work of data examiners. It helps police departments.

Clustering has a number of applications in every field of life. We are applying this technique whether knowingly or unknowingly in a day-to-day life. It is the first step in data mining analysis. The k-means algorithms achieved good results when properly initialised. It identifies groups of related records the can be used as a starting point for exploring further relationships. In addition, some of our results suggest that using the file names along with the document content information may be useful for cluster ensemble algorithms.

More importantly, we observed that clustering algorithms indeed tend to induce clusters formed by either relevant or irrelevant documents, thus contributing to enhance the expert examiner’s job.

V. ACKNOWLEDGEMENT

We thank Prof. Rahul Samant Sir for his invaluable insights and guidance for this project. Also, we are grateful to MD Technologies for their support.

REFRENCES

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

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

[4] L. Kaufman and P. Rousseeuw, Finding Groups in Gata: An Introduction to Cluster Analysis. Hoboken, NJ: Wiley-Interscience, 1990.

[5] Ranjana Agrawal ,Madhura Phatak, "A Novel Algorithm for Automatic Document Clustering," 3rd IEEE International Advance Computing Conference (IACC) ,pages 877 - 882,IEEE, 2013.

[6] Zonghu Wang, Zhijing Liu, Donghui Chen, Kai Tang,"A New Partitioning Based Algorithm For Document Clustering",Eighth International Conference on Fuzzy Systems and Knowledge Discovery,pages 1741 - 1745 IEEE,20 II.

[7] S.C. Punitha, R. Jayasree andDr. M. Punithavalli, "Partition Document Clustering using Ontology Approach", Multimedia and Expo, 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Jan. 04 06,pages 1-5, 2013.

[8] Lus Filipe da Cruz Nassif and Eduardo Raul Hruschka, "Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection," IEEE transactions on information forensics and security, Vol. 8, NO. I ,pages 46 - 54 Jan 2013.

