Overview of K-means and Expectation Maximization Algorithm for Document Clustering

Bhagyashree Umale
Research Scholar,
Dr D Y Patil School of Engg & Technology, Pune.

Nilav M.
Assistant Professor, Guide,
Dr D Y Patil School of Engg & Technology, Pune.

ABSTRACT
Advances in data collection and storage capabilities during the past decades have led to an information overload in most sciences. Computer forensics is a new and fast growing field that involves carefully collecting and examining electronic evidence that not only assesses the damage to a computer as a result of an electronic attack, but also to recover lost information from such a system to prosecute a criminal.

Nowadays the digital content involved in a crime is nowhere simple to read & infer. Its increasingly a labyrinth of data/files/folders, which needs to be analyzed, to get ahead into investigation & solving the crime cases worldwide. In light of this, the computer based document clustering, for the forensics analysis of digital content/data, is a very important tool/program. It reduces the much of manual effort & redundancy, & makes the resolution of crimes cases faster.

The process of clustering is based on processing of multiple text files simultaneously. These text files may comprise very huge raw/text data, which needs to be converted into structured form in order to do further processing of crime analysis. Huge volumes of data need

to be analyzed & this process may be slow if commercial and open source forensic tools are used. In early days, forensics was largely performed by computer proffesionals who worked with law enforcement on an ad-hoc, case-by-case basis. There are many algorithms suggested by various experts for the data analysis. A study of investigation work over the different document clustering methods for forensic analysis is used for this survey. In this paper, we are aiming to explain partitional algorithms namely – kmeans and its variant i.e., Expectation Maximization Algorithm.

General Terms
Document clustering, partitional, hierarchical

Keywords
Computer Forensics Analysis, Expectation-Maximization, k-means.

1. INTRODUCTION
Documents analysis process in computer device is key task of the digital forensic investigation process & this process becomes more complex, if the number of documents available to process very large. The complexity increases further, if the digital device (under investigation) has a large storage. There aresome methods and tools already presented by various researchers for the analysis of multiple documents. These existing methods of DFI propose a multi- level search approach, for giving the accurate results and producing digital evidence that is related to the current investigation task. The inherent drawback of these methods is, no provision for crime investigator/end user to search the documents relevant to the specific subject in which end user is interested, or to group the document set based on a given subject.

The DFI system first takes the input as raw/text files related to crime data which is in unstructured format. This data is further required to be converted into structured form using the text mining methods. There are many clustering algorithms presented previously those are especially tailored to be used for the analysis of forensic. Such clustering methods are basically used for the data analysis purpose in which there is very less or no prior information about the input data. All computer forensics applications produce end results with same attribute/lacunae. While technically speaking, 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 availability of labeled dataset is possible through the past analysis, but there is no certainty that same classes or groups available in input dataset or for next incoming raw dataset which is being collected from different digital devices as well as related to various processes of investigations. The inbound data sample can come from the different types of sources. Therefore to provide an efficient solution, for processing, such heterogeneous input datasets in forensic analysis, the clustering algorithms are used. Such clustering methods are able to find out the latent patterns from the text documents those are available from seized computers. Clustering algorithms improves the process of analysis which is performed by end users. The methodology behind such clustering algorithms is that objects inside the valid cluster are more likely to same with each other as compared to objects belonging to a various other clusters [1]. Hence once the data partition has been induced from the data, the investigator/end user might initially focus on checking similar documents from the obtained set of clusters. After this preliminary analysis, the team may eventually decide to scrutinize other documents from each cluster. Thus with this, we can improvise the difficult task of analyzing the documents individually & at the same time manual scrutiny is available, if it is required in some complicated criminal cases.

We have studied, the recent investigation based work done over different clustering algorithms such as k-means clustering, K-medoids, Single Link, Complete Link, Average Link, and CSPA) with different digital forensic datasets in [1]. In [1], author presented the methodology which the document clustering algorithms were used for the forensic analysis of digital data/evidence in the criminal cases being investigated by police. A number of different practical results were reported as well as discussed with different datasets of forensic computing. However as per the author's statements, it still requires more investigations and analysis. In this paper we are analyzing partitional algorithms, for forensic analysis of digital data/evidence.

From all the studies we may conclude that, typical clustering methods are: partitioning methods, hierarchical methods, density-based methods, grid-based methods and model-based methods. Our research is focused on model-based clustering.

In next section 2 we have presented the work related to the various methods, which are used in clustering algorithms. In section 3, the approach for clustering is depicted. Finally conclusion and future work is predicted in section 4. Section 5 refers to acknowledgement.

2. RELATED WORK

In this section we try to briefly walkthrough the different methods of document clustering in digital forensics

- Computer Forensics field uses very selective clustering algorithms. Most of the studies describe the use of classic algorithms for clustering data—e.g., Expectation-Maximization (EM) for unsupervised learning of Gaussian Mixture Models, K-means, Fuzzy C-means (FCM), and Self-Organizing Maps (SOM). These algorithms have well-known properties and are widely used in practice.
- In [1], document clustering algorithms are implemented using various datasets. Two relative validity indices namely silhouette and simplified silhouette are used for estimating the number of clusters from data. Reference partition is used for evaluating data clustering algorithms. Limitations are also explained which throw light on the fact that the success of any clustering algorithm depends upon the input data. In these days, all the data required is available in digital form. A survey and forecast of worldwide information growth is worth consideration. Data storage, networking and security are the important aspects which play a major role in crime investigation [3].
- In [9], Self Oriented Maps based algorithms were used. This helps the examiners to perform clustering more efficiently. The files were clustered by taking into account their creation dates/times and their extensions. This kind of algorithm has also been used in [10] in order to cluster the results from keyword searches. The underlying assumption is that the clustered results can increase the information retrieval efficiency, because it would not be necessary to review all the documents found by the user anymore.
- The partional K-means and K-medioids are discussed in [5] and [6]. Details about convergence of algorithm are also discussed in the same. Cluster ensemble problem and the algorithms to solve that problem is discussed in [7]. In [8] the author stresses that, Clustering is a useful exploratory technique for gene-expression data. According to it, evolutionary algorithms automatically estimate the right number of clusters. Relative cluster validity criterion is discussed in [9]. External cluster validity criteria, such as rand index, adjusted rand index and jaccard coefficient are explained in detail. The already existing studies mention that the number of clusters is known and fixed a priori by the user. This assumption of entering number of clusters by the user is unrealistic in real world applications. Hence, a common way is to find out the number of clusters from the given data. Therefore, different data partitions (with different numbers of clusters) can be considered and then assesses them with a relative validity index in order to estimate the best value for the number of clusters [4], [5], [11].
- M. Laszlo and S. Mukherjee [12] propose the usage of Hyper-Quad trees (HQ) as the initialization algorithm to obtain the initial cluster centers/centroids which serve as input to various clustering algorithms such as K-Means, EM. Related work in this domain can be pursued to achieve increased efficiency in the computation of centroids derived from the initialization algorithm.
- J. Han and M. Kamber [13] provide a detailed description of the widespread concepts of data mining and the tools required to manipulate data. Fault prediction using quad tree and Expectation Maximization clustering algorithm, limits the research in this book to the section of "Cluster Analysis". The cluster analysis section in this book describes different types of clustering methods. In [14], a detail chapter of mixture models and EM introduces the concepts related to Expectation Maximization Algorithm.
- M. Steinbach, G. Karypis, and V. Kumar [15] discuss about comparison of document clustering techniques.

3. Approach used by CLUSTERING Algorithms

3.1 3.1 Types of Clustering

Document clustering is a completely unsupervised task with the goal of discovering groups of similar documents in a collection without a-priori knowledge. There are two typical categories of clustering algorithms, the partitional and the hierarchical. K- means and the single/complete/average link clustering are the representatives of these two categories, respectively. There are many comparisons between K-means and hierarchical clustering. But our consideration is speed, since we are going to apply clustering algorithms on big social network data, which is always of GB or TB size.

The hierarchical clustering is extremely computational expansive as the size of data increases, since it needs to compute the D_D similarity matrix, and merges small clusters each time using certain link functions. In contrast, K-means is much faster. It is an iterative algorithm, which updates the cluster centroids (with normalization) each iteration and re-allocates each document to its nearest centroid. A comparison of K-means and hierarchical clustering algorithms can be found in [15].

3.2 General Approach

Various algorithms for clustering using k- means, k-medioids, hierarchical clustering can be used. These documents need to be preprocessed to remove the unwanted information which is not useful for investigation.

In general, the documents will be stored in files maybe in the same directory or different directories, where we take set of documents as input to system then apply preprocessing methods which include the following steps :

1. Stop word removal

2. Stemming

3. Vector space model

Preprocessing includes stop word removal and stemming.

The stopwords, which are the most frequently used words, are collected in a separate text file. Preprocessing is performed and using cosine distance formula the required matrix is generated which depends upon the number of input files.To perform cluster analysis, the documents need to be represented in vector form. The vector space model is used for the same. Calculating clusters depends upon the distance between similar words. Cosine based formula or any distance

formula is used for this purpose. Also, Levenshtein distance is calculated to find the distance between two documents. Clusters are estimated using various methods such as K-means, K- medioids, hierarchical- single, complete and average link, CSPA and Expected Maximization algorithm. Silhouette is used as a relative validity criterion. External validity criterion may be using Random Index Analysis or adjusted random index or Jaccard coefficient.

3.3 Algorithms

3.3.1 K-means Algorithm

In the k-Means algorithm, the labeling function is computed by comparing the distances of a data point xi from the vectors which represent the clusters (the centroids cj). The centroids are the model parameters which are estimate by using iterative steps.

According to [13], the k-means algorithm defines the centroid of a cluster as the mean value of the points within the cluster.

Algorithm:

The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input: c: the number of clusters

D: a dataset containing n objects

Output: A set of k clusters

Method:

i. Arbitrarily chose k objects from D as the initial cluster centers;

ii. Repeat

iii. (re)assign each object to the cluster to which the object is most similar, based on the mean value of theobjects in the cluster;

iv. update the cluster means, that is calculate the mean value of the objects for each cluster;

v. until no change

The time complexity of the k-means algorithm is O(nkt) , where n is the total number of objects, k is the number of clusters, and t is the number of iterations.

Therefore, the method is relatively scalable and efficient in processing large datasets.

3.3.2 Expectation Maximization Algorithm

Expectation Maximization is a type of model based clustering method. It attempts to optimize the fit between thegiven data and some mathematical model. Such methods are often based on the assumption that the data are generated by a mixture of underlying probability distributions. The EMalgorithm is an extension of the K-Means algorithm. It is iterative in nature and finds maximum likelihood solutions. With reference to [13], Expectation Maximization consists of two steps:

The expectation step assigns objects to clusters according to the current fuzzy clustering or parameters of probabilistic clusters.

The maximization step finds the new clustering or parameters that maximize the expected likelihood in probabilistic model-based clustering.

Algorithm

Input: c: the number of clusters

D: a dataset containing n objects

Output: A set of k clusters

Method:

1) First find initial centers/centroids which will be the initial input.

2) Compute distance between each data point and each centroid using cosine distance formula or any other distance formula.

3) Assign weights for each combination of data point and cluster based on the probability of membership of a data point to a particular cluster.

4) Repeat

i) (Re) assign each data point to the cluster with which it has highest weight i.e., highest probability.

ii) If a data point belongs to more than one cluster with the same probability, then (re)assign the data point to the cluster based on minimum distance.

iii) Update the cluster means for every iteration until clustering converges.

EM has a strong statistical basis, it is linear in database size, it is robust to noisy data, it can accept the desired number of clusters as input, it provides a cluster membership probability per point, it can handle high dimensionality and it converges fast given a good initialization.

EM offers many advantages besides having a strong statistical basis and being efficient. One of those advantages is that EM is robust to noisy data and missing information. In fact, EM is intended for incomplete data.

The complexity of EM depends upon the number of iterations and time to compute E and M steps.

4. CONCLUSION AND FUTURE WORK

Due to availability of high speed net connections and newer portable devices, forensic analysis is becoming a complicated process. Existing digital forensic tools for analyzing a set of documents provide multiple levels of search techniques to answer questions and generate digital evidence related to the investigation. However, these techniques stop short of allowing the investigator to search for documents that belong to a certain subject he is interested in, or to group the documents.

Most importantly, it is observed that clustering algorithms find out similar words and collect them in a single cluster which helps the forensic examiner for detection. Furthermore, our studies of the proposed approach in real world applications show that it has the capacity to fasten the computer inspection process. In this paper two clustering methods are discussed**.**

The future work may include modifying the existing EM algorithm by combining the Quad Tree approach and the EM algorithm which gives a clustering method that not only fits the data better in the clusters but also tries to make them compact and more meaningful [2].

5. ACKNOWLEDGEMENT

We thank Luís Filipe da Cruz Nassif and Eduardo Raul Hruschka, for providing their valuable research work .

6. REFERENCES

[1] Luís 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. 1, JANUARY 2013.

[2] Meenakshi PC, Meenu S, Mithra M, Leela Rani P, "Fault Prediction using Quad Tree and Expectation Maximization Algorithm", International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 2– No.4, May 2012

[3] J. F. Gantz, D. Reinsel, C. Chute, W. Schlichting, J. McArthur, S. Minton, I. Xheneti, A. Toncheva, and A. Manfrediz, "The expanding digital universce: A forecast of worldwide information growth through 2010," I*nf. Data*, vol. 1, pp. 1–21, 2007.

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

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

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

[7] A. Strehl and J. Ghosh, "Cluster ensembles: A knowledge reuse framework for combining multiple partitions," *J. Mach. Learning Res.*, vol. 3, pp. 583–617, 2002.

[8] E. R. Hruschka, R. J. G. B. Campello, and L. N. de Castro, "Evolving clusters in gene-expression data," *Inf.Sci.*, vol. 176, pp. 1898–1927, 2006.

[9] L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka, "Relative clustering validity criteria: A comparative overview," *Statist. Anal. Data Mining*, vol. 3, pp. 209–235, 2010.

[10] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988.

[11] B. D. Carrier and E. H. Spafford., " An event-based digital forensic investigation framework". In *Proceedings of the 4th Digital Forensic Research Workshop*, 2004.

[12] M. Laszlo and S. Mukherjee, "A Genetic Algorithm Using Hyper-Quad trees for Low-Dimensional K-Means Clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no4, pp. 533-543, 2006.

[13] Han, Kamber, Pei, "Data Mining : Concepts and Techniques", MK Third Edition

[14] C M Bishop, "Pattern Recognition and Machine Learning" NewYork Springer-Verlag 2006

[15] M. Steinbach, G. Karypis, and V. Kumar. "A comparison of document clustering techniques". Technical Report 00-034, University of Minnesota, 2000