International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.5., 2015
(Sept.-Oct.)

RESEARCH ARTICLE

ISSN: 2321-7758

IMPLEMENTATION OF DOCUMENT CLUSTERING FOR FORENSIC ANALYSIS WITH BISECTING K-MEANS ALGORITHM

FATHIMA SHAIK[1],Dr.A.KRISHNA MOHAN[2]

[1]M.Tech (IT), Department of CSE, UNIVERSITY CAMPUS, JNTUK KAKINADA
ANDHRA PRADESH, INDIA

[2]Professor & HOD , Department Of CSE
UNIVERSITY CAMPUS, JNTUK KAKINADA,ANDHRA PRADESH, INDIA

FATHIMA SHAIK

Dr.A.KRISHNA MOHAN

ABSTRACT

In Forensic Analysis thousands of files are generally examine. Data in those files consists of shapeless content analyze it by examiners is especially complicated. Algorithms for clustering credentials can assist the innovation of new and practical information from the credentials below investigation. Cluster study itself is not one particular algorithm but the common assignment to be solve. It can be achieve by different algorithms that vary extensively in their view of what constitute a cluster. Here we offer an approach that apply to clustering of Documents detained in police investigation. We define the future approach with Bisecting K-Means algorithm. Our experimentation show that the concert of computer for inspect numerous files is enhanced. In conclusion we also present and talk about numerous useful outcome that can be useful for researchers and practitioners of forensic computing.

Keyword: clustering, forensic computing, k-means, bisecting k-means, data mining.

I. INTRODUCTION

The quantity of data in the digital world improved from 161 hexabytes in 2006 to 988 hexabytes in 2010.this huge quantity of data has a straight contact in computer Forensics. In our picky function area it habitually involve groping hundreds of thousands of files per computer. This movement exceed the expert's capability of investigation and understanding of information. for that reason method for automatic information scrutiny. similar to those broadly use for machine learning and data mining are of supreme significance. In picky, algorithms for pattern recognition from the information here in text documents are hopeful as it will confidently become evident shortly in the paper. The theory of cluster has been approximately for a lengthy time. It has numerous application, mainly in the perspective of information retrieval and in organize web resources. The major use of cluster is to place information and in the near daytime

perspective to place for the most part significant electronic property. The study in cluster ultimately lead to repeated indexing to key as well as to get back electronic proceedings. Clustering is a technique in which we build cluster of substance that are in some way comparable in individuality. The final endeavor of the cluster is to afford a combination of comparable report. Clustering is frequently bemused with arrangement, but there is a little variation between the two. Clustering algorithms are usually use for investigative data study. This is correctly the folder in numerous application of Computer Forensics, as well as the one address in our effort. beginning a more practical perspective, our datasets consist of unlabeled things the module or category of credentials that can be originate are a priori indefinite. in addition, still assume that labelled datasets could be accessible as of earlier study. In this situation, the utilize of cluster algorithms, which are able of discovery hidden pattern from text documents initiate in detained computers, can develop the study perform by the professional assessor. address in our effort. From a additional technological point of view our datasets consist of unlabeled objects the module or category of credentials that can be establish are a priori indefinite. furthermore, still assume that labelled datasets could be obtainable from earlier study. In this situation, the utilize of cluster algorithms, which are able of discovery hidden pattern from text documents set up in held computers, can develop the study perform by the specialist assessor. Clustering algorithms have been considered for decades, and the prose on the issue is vast.

II. RELATED WORK

In database management, data clustering is a technique in which, the information that is logically parallel is physically store as one. In classify to raise the effectiveness of investigate and the recovery in database management, the amount of diskette access is to be minimize In clustering, from the time when the items of related property are located in one group of items, a particular entrée to the diskette can get back the whole class. If the clustering take position in a few theoretical algorithmic gap, we may cluster a residents into subsets with related feature, and then decrease the difficulty gap by performing on no more than a delegate from every split. Clustering is finally a method of dropping a mass of information to controllable loads. For cognitive and computational overview, these many may consist of "related" items. There are two approach to document clustering, predominantly in information retrieval; they are known as term and item clustering. Term clustering is a technique, which group unneeded terms, and this group reduce, blast and enhance occurrence of task There are only a small number of reading experience the use of cluster algorithms in the Computer Forensics field. basically, the majority of the study illustrate the utilize of typical algorithms for clustering data—e.g., Expectation-Maximization (EM) for unofficial knowledge of Gaussian Mixture Models, Kmeans, Fuzzy C-means (FCM), and Self-Organizing Maps (SOM). These algorithms have distinguished property and are extensively use in perform. For example, K-means and FCM can be seen as meticulous cases of EM. The literature on Computer Forensics only intelligence the utilize of algorithms that imagine that the number of clusters is known and set a priori by the consumer. expected at peaceful this theory, which is often impractical in useful application, a familiar approach in other domain involve estimate the number of clusters from data. basically, one induce dissimilar statistics partition (with different numbers of clusters) and then assess them with a comparative strength file in order to approximation the most excellent price for the amount of clusters.

III. CLUSTERING ALGORITHMS AND PREPROCESSING STEPS

Clustering is a distribution of records into group of related items. instead of the information by smaller amount cluster essentially lose definite well detail, but achieve generalization. It model information by its clusters. Data model place cluster in a past perception embedded in arithmetic, data, and arithmetical examination. From a device learn point of view cluster communicate to secreted pattern, the explore for cluster is invalid learn, and the resultant structure characterize a records theory. From a realistic point of view cluster acting an exceptional function in data mining application such as systematic information study, in order retrieval and text mining. Clustering is the theme of lively

investigate in numerous field such as data, prototype detection, and machine learning. This study focus on cluster in data mining.

Data mining add to cluster the complication of especially huge datasets with extremely a lot of attribute of dissimilar type. This impose exclusive computational necessities on appropriate clustering algorithms. In this paper we use Bisecting K-Means Algorithm.

B. Preprocessing

earlier than operation clustering algorithms on text datasets, we perform a number of preprocessing steps. In exacting, stop words (prepositions, pronouns, articles, and irrelevant document metadata) have been detached. Also, the Snow balls stemming algorithm for Portuguese words has been used. Then, we adopt a conventional numerical advance for text mining, in which documents are represent in a vector space model. In this representation, each document is represent by a vector contain the frequencies of occurrences of words, which are define as delimited alphabetic strings, whose number of typescript is between 4 and 25. We also use a dimensionality reduction method known as Term Variance (TV) that can enhance both the usefulness and effectiveness of clustering algorithms. TV selects a number of attributes (in our case 100 words) that have the maximum variances over the documents. In order to calculate distances between documents, two events have been used, namely: cosine based distance and Levenshtein-based distance. The later has been used to calculate distances between file (document) names only.

C. Calculating the number of Clusters

In order to guess the amount of clusters, a broadly utilize move toward consists of receiving a set of data partitions with dissimilar records of clusters and then select that exacting divider that provides the best result according to a definite value measure (e.g., a relative validity index). Such a set of partitions may effect straight from a hierarchical clustering dendogram or, on the other hand, from numerous runs of a partitional algorithm (e.g., K-means) starting from different numbers and initial positions of the cluster prototypes.

D. Clustering Techniques

Generally clustering methods are typically divided into hierarchical and partitioning. Hierarchical clustering is more subdivided into agglomerative and divisive.

Agglomerative: Start with the points as individual clusters and, at each step, merge the most similar or closest pair of clusters. This requires a definition of cluster similarity or distance.

Divisive: Start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain.

In this case, we need to make a decision, at every step, which cluster to divide and how to achieve the divide. Hierarchical technique generate a nested series of partitions, with a particular every complete cluster at the top and singleton clusters of individual points at the base. every middle point can be viewed as combining two clusters from the next minor level. The outcome of a hierarchical clustering algorithm can be graphically displayed as tree, called a dendrogram. This tree graphically displays the integration procedure and the intermediate clusters. For document clustering, this dendrogram provides a catalog, or hierarchical index.

E. Removing Outliers

We evaluate a easy advance to remove outliers. This move toward makes recursive use of the profile. essentially, if the finest division selected by the profile has singletons (i.e., clusters formed by a single object only), are removed. Then, the clustering procedure is continual more and more again—awaiting a divider without singletons is create. At the end of the process, all singletons are included into the resultant information divider (for evaluation purposes) as single clusters.

IV. EXISTING SYSTEM:

Clustering algorithms have been consider for few years , and the literature on the topic is vast. Therefore, we confident to wish a set of (six) delegated algorithms in order to illustrate the possible 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 CSPA. These algorithms were run with dissimilar grouping of their parameters, ensuing in sixteen dissimilar algorithmic

instantiations. Thus, as a role of our work, we evaluate their comparative performances on the studied application domain—using five real-world examination cases conducted by the Brazilian Federal Police Department. In order to make the proportional analysis of the algorithms more practical, two comparative validity indexes have been used to approximation the number of clusters mechanically from data The basic K-means clustering technique is presented below. We elaborate on various issues in the subsequent sections.

Basic K-means Algorithm for finding *K*-clusters.

1. Select *K* points as the initial centroids.
2.Allot all points to the closest centroid.
3. Re compute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroids don't change.

DISADVANTAGES OF EXISTING SYSTEM:

The literature on *Computer Forensics* only reports the use of algorithms that guess that the number of clusters is known and fixed *a priori* by the user. Aimed at relaxing this supposition, which is often impracticable in practical applications, a common approach in other areas involves estimates the number of clusters from data.

V.PROPOSED SYSTEM:

For what follows we will use a bisecting K-means algorithm as our primary clustering algorithm.

This algorithm starts with a single cluster of all the documents and works in the following way:

Basic "Bisecting K-means" Algorithm for finding *K* clusters.

1. Pick a cluster to split.
2. Find 2 sub-clusters using the basic K-means algorithm. (Bisecting step)
3. Repeat step 2, the bisecting step, for ITER times and take the split that produces the clustering with the highest overall similarity.
4. Repeat steps 1, 2 and 3 until the desired number of clusters is reached.

ADVANTAGES OF PROPOSED SYSTEM:

Most importantly, we observed that clustering algorithms in fact lean to make clusters shaped by either significant or unrelated papers, thus causal to improve the proficient examiner's job. moreover our assessment of the proposed approach in applications illustrate that it has the possible to speed up the computer inspection process.

VI.RESULTS AND ANALYSIS:

6.1 Entropy:

We use entropy as a determine of excellence of the clusters. Let CS be a clustering decision For each cluster, the class distribution of the statistics is intended first, i.e., for cluster j we compute p_{ij}, the"probability" that a member of cluster j belongs to class i. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula

$E_j = -\sum p_{ij} \log(p_{ij})$

where the sum is taken over all classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each

cluster:

$E_{CS}=\sum_{j=1 \text{ to } m} n_j * E_j / n$

where nj is the size of cluster j, m is the number of clusters, and n is the total number of data points.

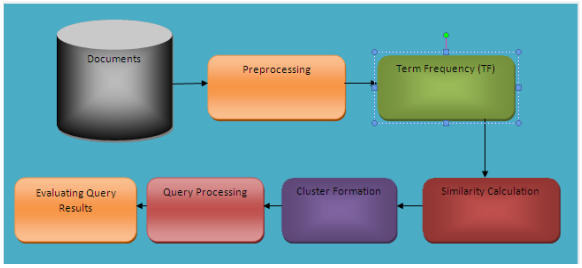
1	comparision of k-means and bisecting k-means with entropy				
2					
3	DATASET	K	BISECTING K-MEANS	BISECTING K-MEANS WITH REFINEMENT	REGULAR K-MEANS
4	animal.txt	16	1.3305	1.1811	1.3839
5	badminton.txt	16	1.6315	1.7111	1.6896
6	basketball.txt	16	1.5494	1.5601	1.8557
7	carrom.txt	16	0.4713	0.4722	0.5228
8	chemical.txt	16	0.6909	0.6927	0.7426
9	chemicalreaction	16	1.3708	1.4053	1.3198
10	chess.txt	16	0.957	0.9511	1.071
11	cricket.txt	16	0.9799	0.9445	0.9673

Cosine similarity:

The similarity between two documents must be measured in some way if a clustering algorithm is to be used. There are a number of possible measures for computing the similarity between documents, but the most common one is the cosine measure, which is defined as *cosine*(*d1, d2*) = (*d1* . *d2*) / ||*d1*|| ||*d2*|| ,

where indicates the vector dot product and ||*d*|| is the length of vector *d*.

BLOCK DIAGRAM:



ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise actions and events with carry for variety, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to illustrate the business and prepared step-by-step workflows of components in a scheme. An activity diagram shows the overall flow of control.

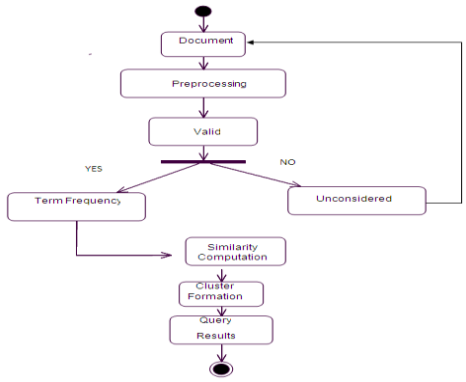


Figure: Flow chart for clustering technique in data mining

VII. CONCLUSION

Clustering has a number of functions in every pasture of life. We are relating this technique whether consciously or unconsciously in day-to-day life. One has to cluster a group of object on the base of relationship either knowingly or automatically. Clustering is repeatedly one of the first steps in data mining examination. The Bisecting K-means algorithm also achieved good results when appropriately initialized. consider the approach for estimating the number of clusters, the relative validity measure known as silhouette has shown to simplified version. It identify group of correlated records that can be use as a starting point for explore additional relations. In adding, some of our consequences propose that using the file names along with the document content information may be useful for cluster ensemble algorithms. mainly significantly, we practical that clustering algorithms indeed tend to make clusters formed by either applicable or unrelated documents, thus underlying to improve the expert examiner's job. in addition, our estimation of the future progress in five real-world applications show that it has the possible to speed up the computer examination process.

VIII. REFERENCES

[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]. R. Xu and D. C.Wunsch, II, *Clustering*. Hoboken, NJ: Wiley/IEEE Press, 2009.

[6]. 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.

[7]. 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.

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

[9]. N. L. Beebe and J. G. Clark, "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results," *Digital Investigation, Elsevier*, vol. 4, no. 1, pp. 49–54, 2007.

[10]. 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.

ABOUT AUTHORS:

FATHIMA SHAIK ,Bachelor Degree in Computer Science And Engineering from MLWEC , Guntur , in 2013 , and now pursuing M.Tech degree in Information Technology from University Campus , JNTU KAKINADA ,Andhra Pradesh . Her research interests include network security and Data mining .

Dr. A.Krishna Mohan ,currently working as a HOD and associate prof. in cse dept.UCEK,JNTUK ,East Godawari, A.P. Ph.D in Computer Science and Engineering, from J.N.T. University Kakinada.He has Few Decades Of Experience In Teaching.