International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Enhanced Document Clustering for Forensic Analysis

Rahul D. Kopulwar[1], Fazeel Irshad Zama[2]

[1]WCEM,Nagpur. R.T.M.N.U., Maharashtra, India

[2] Professor, WCEM,Nagpur, R.T.M.N.U., Maharashtra, India.

Abstract: ***Today, a crime can be in any form but the crime in digital form is increasing now a days. Therefore, the branch of computer science called digital forensic science has increasing importance. The forensic analysts analyze the computer data for particular proof against any crime. But, computers having huge amount of data files really creates havoc to analyze it. Therefore, it needs a good clustering techniques that reduces the efforts of forensic analysts. Already, there are many clustering techniques for the analysis purpose like (k-means, k-medoid, single link, complete link, etc.) and found some good results. But, we used a new approach for the same purpose by applying k-representative as a key algorithm for clustering. We applied k-representative on the datasets collected from different sources and found really a good result as compared with the others. Also we focused on good preprocessing techniques like stemmer and porter algorithms. We experimented our techniques with different types of document formats and it works better with all formats. Finally, we shows results with comparative techniques with the help of comparative graphs.***

Keywords: Data mining, preprocessing, clustering, k-representative**.**

1. Introduction

In the world of technology, computers are of great importance. And advancement in computer architecture result in high processing power as well as huge storage space in computers which can contain very huge amount of data in it. But this can really creates havoc if someone wants to search a specific file in it. For example: if forensic analyst wants to search for a particular file as a proof against any crime then he might have to scan complete computer. It takes a lot of time and it can not a certain accuracy of results.

Computer forensic department is a branch of computer science which mainly focuses on crime related to digital world. Cyber crime is the modern type of crime generally performed by computer technical persons. Therefore, many evidences might be hidden in those computer by which the crime had happened or the computer found at crime scene. Many times, it takes much time to scan all the computer data and search for the desired files which can present in court as a proof against the criminal. Thus, an expert forensic analysts generally scans the computers manually and try to collect evidences. But, it could takes a lot of efforts and also a long time. So, many time it could be an advantage for the criminal as an evidence can not be present on time. Also, the accuracy of analysis is mostly depends on the knowledge and experience of the examiner. Thus, in order to overcome this problem the concept of document clustering can be very useful. The clustering algorithms can be very useful where no knowledge about the data in related document are known a priori[2],[3]. Thus, clustering helps a lot to partition data into group of related documents. There are various clustering approaches with well known algorithms like k-means, k-medoid, single link, complete link, etc. K-means like algorithms works on relatively validity index to estimate the cluster numbers automatically[1]. In our proposed system, we mainly focuses on preprocessing steps like removal of stop words and to stem the words which can help to create the data which can be well organized and can be treat words like stems as keywords for the clustering purpose. We used the K-representative algorithm as our key clustering algorithm. We apply this clustering algorithm on the dataset which we considered as FIR collected from various internet sources. The data found on internet was in unstructured and unorganized form as well. Thus, good preprocessing techniques can help to clean the data and make that data to be effectively used in clustering process. Finally, we found that K-representative gives better result than K-means and K-medoid and other clustering algorithms. We analyze our result in the form of processing time and the size of clusters. Our proposed method forms the maximum clusters where related documents found. We presented our results with the help of graph for better summarization and visual presentation purpose.

2. Background work

Luís Filipe da Cruz Nassif. proposed various techniques of forensic clustering on computers seized on the crime scene. They performed their experiment on real time dataset collected from Brazil police department. They applied six clustering algorithm i.e. K-mean, K-medoids, single link, complete link, Average link and CSPA etc. They calculate the ARI for algorithm and found that Average link for 100 terms was stable and accurate[1]. B. Vidhya et al. proposed text and document clustering technique. They proposed their methods with the help of K-means and ant colony algorithm. They found that K-means was one of the simplest algorithm[4]. Charu et al. proposed the system of document clustering with the help of side information. The side information could be anything like log files, reference path, etc. They proposed the COATES and COLT algorithm applied on the dataset which already clustered by another clustering algorithms like K-means, etc. K. Nagarajan et al. proposed a system which provided a graph based approach

which represents the relation between the data points and clusters.

3. Preprocessing and Clustering Algorithm

3.1 Removal of Stop word

Stop words are the words which we used more frequently in Natural Languages. They are frequently used common words in any natural languages like preposition, pronoun, conjunction, etc. for example : am, is, was, where, etc. Such frequently appearing words are not very much important in mining usage. So, it can be skip to reduce the document size for mining and for data cleaning usage as well [5]. But, one problem mainly found with the removal of stop words is that removal of stop word is domain (language) specific and thus the stop word in one domain may not be the stop word in other[6].

3.2 Stems

Today, in information retrieval system and NLP, the word stem is of great importance as it facilitates the indexing of documents. Usually, the prefixes and suffixes are remove to form the stem. By using the stemming technique we can increase the number of retrieved documents[7]. That means we can ultimately increases the recall rate without any affect on the fetched precision. But, there is also a problem of stemming errors for example : over stemming and under stemming which may affect the recall and precision rates[7].

3.3 Indexing

The process of expressing the main subject or the theme of a text in document is called indexing. Text headings are often taken as indexes. There are primarily two categories of indexing:

1) Classification
2) Co-ordinate

With classification indexing, or classifying, the texts are included an appropriate class (one or several) depending on their content. All texts with basically the same semantic content are brought together. The index number of this class is assigned to each text within it and the number is then serves as it's search specification.

In coordinate indexing, the basic semantic content of the text is expressed by a list of significant words selected either from the text itself or it's headings or from a special normative dictionary. In the first instance, such lexical units are termed key words, and in the second descriptors. Each key word or descriptors designates a class that potentially includes all the texts that have the word in the basic semantic content.

3.4 Clustering

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. Clustering helps users to understand the natural grouping or structure in a data set. Clustering is unsupervised classification that means no predefined classes. It used either as a stand-alone tool to get insight into data distribution or as a preprocessing step for other algorithms.

A good clustering method will produce high quality clusters in which the intra-class (that is, intra-cluster) similarity is high. And the inter-class similarity is low. The quality of a clustering result also depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns. However, objective evaluation is problematic: usually done by human / expert inspection [10].

4. Proposed Work

We proposed the techniques for Document Clustering to facilitate the forensic analysts to do their work efficiently. The stepwise description of the proposed techniques are as follows:

4.1 Data Collection

We collect our dataset used for the proposed system from various sources. It is a data to be considered as real time police investigation reports. The documents we collected is in various formats like doc, docx, pdf, etc. Also, it is not necessary to use only pre maintained dataset rather we can use any dataset on runtime. For example : the dataset from external devices like pen drives and other.

4.2 Preprocessing

Preprocessing of text documents is necessary to clean data and to provide algorithms only the required data. The preprocessing techniques used in our system is described below:

4.2.1 Removal of Stop Words

We maintained a stop word dictionary having all possible stop words. We scan our documents to find such stop words and remove it as well as we maintained the separate removed stop word list to keep the record for number of stop words found in particular document.

4.2.2 Stemming

After stop word removal, we performed stemming of words. We maintained indexed stems. For first index position we kept the original stem, then we scan the document to make the stems. For example : bail / bailed / bailing. So, if we found any word like bailed or bailing then we replace these words as bail.

4.2.3 Synonyms

For better results, we maintained a synonym dictionary. If we don't get accurate word matching then these synonyms could help us to create the related clusters. For example: bail, warranty, surety, bond, guarantee, warrant. Our system finds any of word and consider it as similar word so that it place these words in same category.

We put a text field to search any query by forensic analysts. There is no need to scan and manually check the cluster of interest. Instead, one can search for the interested clusters by entering any keyword or the query. We maintained the indexing of keywords and the files in which the keywords can be found. We retrieve all these files and then the above preprocessing steps are applied on these files. Thus, we get the keywords found in all files. We then input these result to three different algorithm i.e. K-means, K-medoid and K-representative. We find Jaccard coefficient as given below to calculate the similarity distance between two keywords. Thus, formed the clusters having similar group of words.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Our key algorithm for our system is K-representative. We found that K-representative is really gives good results. The pseudo code for K-representative given as follows:

Input: L1, L2, k, U

Output: Answer set ANS

Methods:

1. Initialize ANS = ∅.
2. **while** (|ANS| < k) **do**
3. **For** each O∈U – ANS
   - 3.1 Compute Dist(O, L_1), and Dist(O, L_2) .
   - 3.2 Compute Dist(O, ANS).
   - 3.3 Compute Rep(O,ANS).
4. Find the object P whose Rep(P,ANS) is maximal.
5. ANS = ANS ∪ {P}.
6. **end while**
7. Return ANS.

The input of the algorithm are: Positive set L1 and Negative set L2 and the Unlabelled dataset U. The ANS returns the K number of clusters. The overall system architecture is as shown below in figure 1:

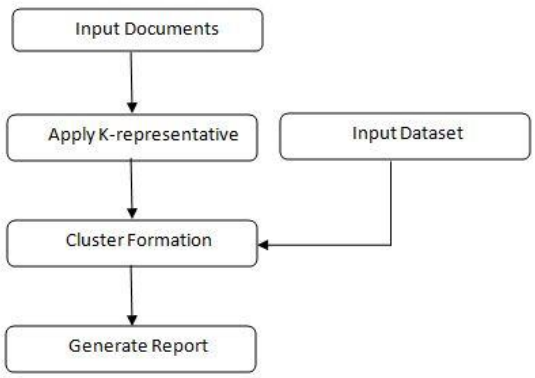

Figure 1

5. Experimental Results

We found that the processing time takes for the preprocessing as well as to form clusters is better. We have test our proposed system for 100 different documents to calculate the processing time and the results found are given in table-1 :

Table 1

Number of documents (Samples)	Time(second)		
	Preprocessing	Clustering	Total
10	11.75	0.515	12.264
25	20.530	0.390	20.92
50	32.136	0.624	32.76
75	44.834	1.123	45.947
100	55.646	1.5	57.146

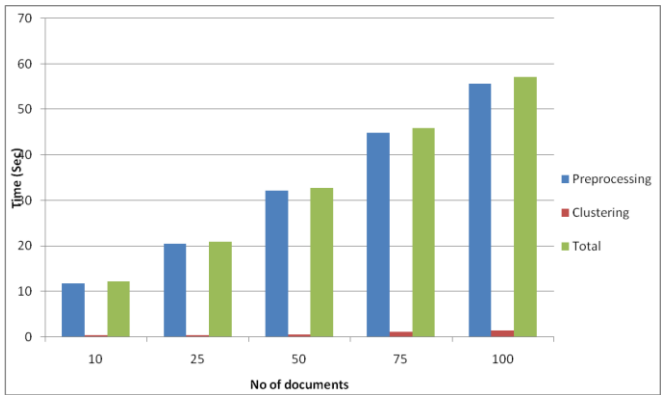

Figure 2

We tested our system with three clustering algorithms as K-means, K-medoid and K-representative. We input five different keywords to be search in documents like crime, corruption, law, legal, etc. And it is found that K-representative algorithm retrieved more number of relevent documents as compared to the other two. The k-means algorithm retrieved 26 relevent documents and k-medoid retrieved 28 whereas k-representative algorithm retrieved 38 relevent documents. The results are shown in table-2 as below:

Table 2

K- Representative Result					
Keywords/Parameter	crime	local	corruption	law	legal
Total no of relevant result in system	45	8	10	38	15
No of retrieved records	40	6	8	35	10
No of relevant records	38	5	5	32	8
No of relevant record not retrieved	5	2	2	6	5
No of irrelevant record retrieved	2	1	3	3	2
Precision	0.95	0.83	0.71	0.91	0.80
Recall	0.84	0.71	0.62	0.84	0.61

Finally, we calculate the precision and recall values to analyze the result for our system. The precision and recall results table-3 is given as:

Table 3

	Average precision	Average Recall	Average Accuracy
K-Mean	0.784	0.462	0.623
K-Medoid	0.74	0.62	0.670
K-representative	0.84	0.724	0.782

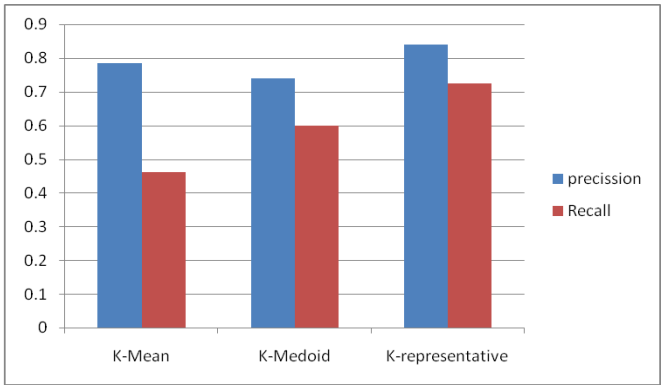

Table 4

6. Conclusion

We presented an approach of clustering for the analysis of documents found in computers seized in crime site. The data in such computers generally found to be in unstruactured form and it needs to be convert in properly structured form to efficiently analysed by the forensic experts. Thus, in order to facilitate such requirements we proposed the enhanced approach for clustering. We experiment with enhanced preprocessing steps and used K-representative as the key algorithm for clustering. We found that K-representative gives good result as compared other algorithms used for this experiments. It shows better computational speed and the amount of relevent retrive data. We can also conclude that the result accuracy is mostly depends on the preprocessing of data which we used to clean the data and that data to be given input to clustering algorithms. But, the accuracy of present preprocessing techniques not found to be that accuarate. Hence, one can need to enhance the technique to improve the clustering accuracy.

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] 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] B. Vidhya and R. Priya Vaijayanthi, " Enhancing Digital Forensic Analysis through Document Clustering", International Journal of Innovative Research in Computer and Communication Engineering, Vol.2, Issue 1, March 2014.
[5] C.Ramasubramanian and R.Ramya," Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 201.
[6] Eduard Dragut, Fang Fang, Prasad Sistla, Clement Yu,"Stop Word and Related Problems in Web Interface Integration", VLDB '09, August 24-28, 2009, Lyon, France Copyright 2009 VLDB Endowment, ACM 000-0-00000-000-0/00/00.
[7] Ms. Anjali Ganesh Jivani," A Comparative Study of Stemming Algorithms", Int. J. Comp. Tech. Appl., Vol 2 (6), 1930-1938.
[8] Pascal Soucy and Guy W. Mineau," Beyond TFIDF Weighting for Text Categorization in the Vector Space Model"
[9] "TF/IDF ranking, solution"
[10] JERZY STEFANOWSKI, "Data Mining - Clustering", Institute of Computing Sciences Poznan University of Technology Poznan, Poland Lecture 7 SE Master Course 2008/2009.
[11] Henry Lin," Clustering", 15-381 Artificial Intelligence.
[12] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn and Supachanun Wanapu," Using of Jaccard Coefficient for Keywords Similarity", Proceedings of the International MultiConference of Engineers and Computer Scientists 2013 Vol I,IMECS 2013, March 13 - 15, 2013, Hong Kong

Author Profile

Rahul Diwakar Kopulwar, M.Tech student of WCEM, Nagpur. Maharashtra, India. His area of interest is Computer Security and Data Mining .

Prof. Fazeel Irshad Zama, received M. Tech. degree from R.T.M.N.U., Maharashtra, India. He is working as faculty of Computer Science at WCEM, Nagpur, Maharashtra, India. His area of interest is Data Mining.