ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 4, September 2014

An Approach to Improve Computer Forensic Analysis via Document Clustering Algorithms

J. Shankar Babu[1], K.Sumathi[2]

Associate Professor, Dept. of C.S.E, S.V. Engineering College for Women, Tirupati, Chittoor, Andhra Pradesh, India[1]

M.Tech Student, Dept. of C.S.E, S. V. Engineering College for Women, Tirupati, Chittoor, Andhra Pradesh, India[2]

ABSTRACT: In computer Forensic analysis thousands of files are usually examined. The computer examiners feel much difficult to analyze data in those files, since it consists of only unstructured text or information. The majority of the tools available on the market have the ability to permit investigators to analyze the information or data that was gathered from a computer Systems. In this context, forensic analysis plays a major role by examining suspected documents seized in police investigations. We proposed an approach clustering algorithms to estimate the number of clusters formed while analyzing the document. New and useful knowledge is discovered while clustering the documents by our algorithm. K-means, K-medoids, Single Link, Complete Link, Average Link, and Cluster-based Similarity Partitioning Algorithm (CSPA) aredifferent efficient algorithms used for clustering documents. To find the number of clusters formed, we use two relative validity indexes in our approach. As a final point, we present several practical results of forensic computing.

KEYWORDS: CSPA, Forensic computing, Preprocessing, Term Frequency.

I. INTRODUCTION

It is estimated that the digital data density is increasing exponentially in recent 5 to 6 years. In Computer Forensic analysis thousands of files are usually examined that allowing the evidence on suspected computer by analyzing the communication and the data on the computer storage device.

It takes a lot of time to analyze all the documents in the seized computers. Clustering algorithms play important role in forensic analysis of digital documents since it contains very important, complex and unstructured data. So we present a novel approach clustering algorithm Cluster-based Similarity Partitioning Algorithm (CSPA).Since clustering algorithms don't have any prior knowledge about data in the document, a great importance is given to the pattern recognition techniques as well as data mining techniques.

Since clustering groups the similar data in the documents which helps to perform searching and finding efficiently. The same concept is carried out here;So that particular cluster $c_i$consist of documents D_i contains some sort of similar content. Now experts can focus on particular cluster for any document rather than analyze all documents. Also there are in all 6 clustering algorithms namely K-means, K-medoids, Single link, complete link, Average Link, CSPA. With different parameters, when algorithms were executed gives sixteen new instantiations. Silhouette and its simplified version are two relative indexes which are used to finding out number of clusters that will be made after algorithm implementation.

II . CLUSTERING ALGORITHM AND PREPROCESSING

Preprocessing is the important step to perform before running the clustering algorithms. Preprocessing stepconsists of 1) removal of stopwords such as pronouns, nouns, adjectives etc., which will not affect the meaning of the document and 2) stemming algorithm for Portuguese words. The main statistical approach for text mining is adopted after preprocessing. According to their frequencies of occurrence, each document is represented as vector model format consisting of words.

For each document the weight of the words should be computed based on the Term Frequency(TF) which is a reduction technique used to increase the efficiency of clustering algorithms. Term Frequency (TF) is the number

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

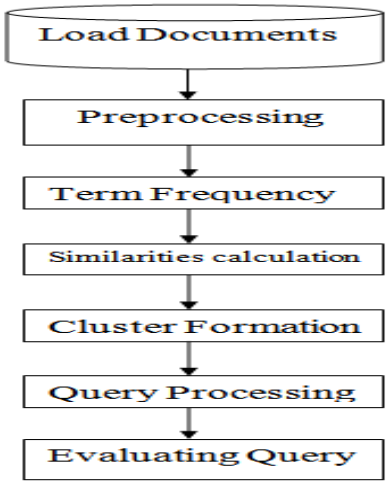
Vol.2, Special Issue 4, September 2014

Cluster	Information
C1	Credit card theist in AP
C2	Murders
C3	ATM Misuse
C4	1 Loan Agreements 2 Bank Accounts
C5	6 LIC Policies 2 Check receipt
C6	5 Investment Club Status
C7	5 Grocery List

Table1: Information Found in Clusters

of occurrences of the word in the document .The distance between the paragraphs can be computed from the dissimilarities between the names in the documents using with the Levenshtein distance. The dissimilarities of the documents can be calculated based on the correlations between the documents available for clustering. Finally, the resulting data clusters can be formed as single cluster. Suppose there is object x in cluster A. And dissimilarity of x with other objects in same cluster that is A is a (x). Now consider cluster C and average dissimilarity of object x to cluster C is d(x, C). As we have to find out dissimilarities within neighbor clusters following technique is used. After computing the dissimilarity d(x, C) with all clusters except A, smallest one is selected that is b(x)=min d(x,C).
Value of dissimilarity neighbors is finding out by formula
$S(x) = b(x)-a(x)/ \max \{a(x), b(x)\}$
Value of S(x) is verified in between -1 to 1. If value of S(x) is higher, object x belongs to particular cluster but if S(x) is zero, then it is not clear that whether object belongs to current cluster or adjacent one. S(x) is carried out over i=1, 2,…,n where n is number of objects and then average is computed. And best clustering is maximum S(x).Hence to finding out effective S(x), i.e. called simplified Silhouette, one can compute only the distances among the objects and the centroids of the clusters. So a(x) is dissimilarity correspond to or simply belongs to cluster (A) centroid. Now it is very easy to get only one distance rather than finding out whole dissimilarity between all objects within cluster. Also rather than finding out d (x C). C not equal to A, we will find only distance with centroid.

III DATA FLOW DIAGRAM FOR FORENSIC ANALYSIS OF DOCUMENTS



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 4, September 2014

Load the documents:

Initially,we should load a folder which consists number of documents for clustering. For an instance, in our approach the folder consists of six documents, where as our software takes only one file at a time. The document isloaded by our software for further processing. Then it will undergo the following process in preprocessing.

Preprocessing:

In preprocessing we remove all unwanted words in the documents.Preprocessing step consists of 1) removal of stopwords and 2) stemming.

Removal of Stopwords:

The loaded document as input contains lot of stopwords such as pronouns, nouns, adjectives etc., which will not affect the meaning of the document.The removal of stopwords is the most common term filtering technique used. There are standard stopword lists available but in most of the applications these are modified depending on the quality of the dataset. Some other term filtering methods are:

• Removal of terms with low document frequencies. This is done to improve the speed and memory consumption of the application.

• Numbers do not play much importance in the similarities of the documents except dates and postal codes. Thus these can also be removed.

For an instance,assume our loaded input document contains the sentence as

He is preparing for exam.

In the above sentence, it contains words like He, is,for are stopwords which are not needed for further Preprocessing step i.e., stemming. So we will remove those words from our original sentence and we just pass words like exam, preparing to further step.

Stemming:

Using Stemming algorithm we bring down the word to its original base form. Consider the sameexample of sentence. He is preparing for exam.In this sentence, after removing of stopwords, we get words preparing,exam for stemming.In stemming, we will bring the word ‘preparing’to its base form as to ‘prepare’.

For this, we usea stemming algorithm called Port stemmer.

Term Frequency:

After preprocessing, we will find out the number of individual occurrences of the words on the multiple documents.Based on the Term Frequency,Inverse document frequency(IDF) can be calculated. For each document the weight of the words should be computed based on the Term Frequency (TF) which is a reduction technique used to increase the efficiency of clustering algorithms. The results can be saved for entertaining the correlations between the documents based on the content.

Similarity Calculation:

In similarity calculation,we should identify the similarities between the documents based on the contents.Each document is compared with remaining other documentssequentially.The distance between the paragraphs can be computed from the dissimilarities between the names or words in the documents using with the Levenshtein distance. The dissimilarities of the documents can be calculated based on the correlations between the documents available for clustering.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 4, September 2014

Cluster Formation:

Based on the variances multiple clusters are formed as Noun cluster, pronoun cluster and adjective cluster.In noun cluster, all the nouns in the remaining documents are also grouped .In the same way all the pronouns and adjectives in the remaining documents are grouped to form pronoun cluster and adjective cluster respectively.
Finally, the resulting data clusters can be formed as single cluster.

Query processing:

The examiner passes the queries, from that he can retrieve the content from the cluster.
For an instance, the examiner has given the query as 'Bangalore'.
Here, after searching it is identified that Bangalore is noun and the documents containing noun Bangalore is displayed.

Evaluating query result:

Finally, we evaluate the query result using the Noun based summarization. Now the content can be retrieved from the multiple documents.

IV. LIMITATIONS

In our approach**,**our clustering algorithms has given good results and raised an issue called scalability.To deal with this issue several sampling techniques and other algorithms like partitional k-means are used.In some cases we use the combination of partitional and hierarchical clustering algorithms to avoid computational difficulties in our work.We adopted, a simplified version of silhouetteswhich includes computation of distance between objects and cluster centroids for estimating the no.of clusters. It takes computational cost of O (k.N.D), where k is the number of clusters, N is number of objects in dataset and D is number of attributes.
In practice, it is not compulsory to use the particular clustering algorithms or particular scalable methods .we can use any of them based upon the amount of the data. Since we don't have any time constraint and computational cost requirements both partitional and hierarchical will give good results.As we know that the examiners in forensic department to analyze their data will take month's time to the final conclusion.

V. CONCLUSION

In this paper, we presented an approach to find the accurate information in computers seized by police department. In past days, forensic analysis by data examiners takes long time and difficult. We used clustering algorithms to estimate the number of clusters formed while analyzing the documents. New and useful knowledge is discovered while clustering the documents by our approach. We also discussed how the data clustering in forensic analysis is performed and how the accurate results are obtained.

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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 4, September 2014

[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.
[11] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi, "Mining writeprints from anonymous e-mails for forensic investigation," *Digital Investigation, Elsevier*, vol. 7, no. 1–2, pp. 56–64, 2010.
[12] S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, and R. Zunino, "Text clustering for digital forensics analysis," *Computat. Intell. Security Inf. Syst.*, vol. 63, pp. 29–36, 2009.
[13] K. Stoffel, P. Cotofrei, and D. Han, "Fuzzy methods for forensic data analysis," in *Proc. IEEE Int. Conf. Soft Computing and Pattern Recognition*, 2010, pp. 23–28.
[14] 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.
[15] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no.5,pp.513–523,1988.

BIOGRAPHY

J. Shankar Babu is an Associate Professor in the Department of Computer Science and Engineering, S.V. Engineering College for Women, Tirupati, Andhra Pradesh, India. He has Published one paper.His areas of Interest are Image Processing and Data WareHousing and Data Mining.

K.Sumathi is a student in Master of Technology in the Department of Computer Science and Engineering, S. V Engineering College for Women, Tirupati, Andhra Pradesh, India. She received Bachelor of Technology (B. Tech) in Computer Science and Engineering in the year 2012 from VITS, Proddatur,Kadapa, Andhra Pradesh, India.