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. 4, Issue 5, May 2016

Document Clustering for Forensic Analysis

Smita Avinash Saravade, Nita Arun Gawali, Sharwari Shankar Dadali

BE, Dept. of Computer, P.E.S .MCOE, Savitribai Phule Pune University, Pune, India

BE, Dept. of Computer, P.E.S .MCOE, Savitribai Phule Pune University, Pune, India

BE, Dept. of Computer, P.E.S .MCOE, Savitribai Phule Pune University Pune, India

ABSTRACT: In computer forensic analysis, thousands of files are usually examined. These files consist of unstructured text, whose analysis by computer examiners is difficult to be performed. In particular, algorithm for clustering documents can facilitate the discovery of new and useful knowledge from the document under analysis. We propose the system that applies document clustering algorithm to forensic analysis of computers sized in police investigation. In this proposed system we used k-means algorithm.

KEYWORDS: Clustering, Forensic computing, Text mining.

I. INTRODUCTION

There is large amount of data that has a direct impact in Computer Forensics, which can be broadly defined as the discipline that combines elements of law and computer science to collect and analyze data from computer systems in a way that is admissible as evidence in a court of law. It usually involves examining thousands of files per computer. This activity exceeds the expert's ability of analysis and interpretation of data.
Clustering algorithms are typically used for analysis of very large data set, where there is no prior or little knowledge about the data. This is precisely the case in several applications of Computer Forensics, including the one used in our work. Our datasets consist of unlabeled objects—the classes or categories of documents that can be found are a priori unknown. Even assuming that labeled datasets could be available from previous analyses which can help us in clustering. Clustering algorithms are used which are capable of finding patterns from computers for enhancing our analysis.
Clustering is nothing but grouping of similar data objects in one cluster that is objects in one cluster are similar to each other than the objects in another cluster. Domain experts have limited time for performing examinations or doing analysis of data. Clustering prioritize the relevant documents required for analysis. Clustering helps to speed up the computer forensic process.

II. RELATED WORK

The project gets divided into the following four different modules. They are: -

(a) Pre-processing: Before running clustering algorithms on text datasets, we will be performing some pre-processing steps. In particular, prepositions, pronouns, articles, and irrelevant document metadata have been removed.

(b) Text Mining: We will be adopting a traditional statistical approach for text mining, in which documents are represented in a vector space model In this module, each document is represented by a vector containing the frequencies of occurrences of words, We will be also using a dimensionality reduction technique known as Term Variance (TV) that can increase both the effectiveness and efficiency of clustering algorithms. TV selects a number of attributes that have the greatest variances over the documents. Order to compute distances between documents cosine-based distance and Leven-shtein based distance are used.

(c) Estimating number of clusters: Inorder to estimate the number of clusters, a widely used approach consists of getting a set of data partitions with different numbers of clusters and then selecting that particular partition that

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. 4, Issue 5, May 2016

provides the best result according to a specific quality criterion. For the moment, let us assume that a set of data partitions with different numbers of clusters is available, from which we want to choose the best one according to some relative validity criterion. Note that, by choosing such a data partition, we are performing model selection and, as an intrinsic part of this process, we are also estimating the number of cluster.

(d) Clustering Technique: In this proposed system we will be using k-mean algorithm for performing document clustering.

(e) Removing Outliers: We assess a simple approach to remove outliers. This approach makes recursive use of the silhouette. Fundamentally, if the best partition chosen by the silhouette has singletons (i.e., clusters formed by a single object only), these are removed. Then, the clustering process is repeated over and over again until a partition without singletons is found. At the end of the process, all singletons are incorporated into the resulting data partition (for evaluation purposes) as single clusters.

III. EXISTING SYSTEM

In existing system we were working on labelled datasets which is called as supervised dataset ,where there is little or prior knowledge about the data. But from technical point of view, our datasets consists of unlabelled objects-the classes or categories of document that can be found are a priori unknown. We can assume that labelled datasets could be available from previous analyses, there is almost no hope that the same classes would be still valid for upcoming data, obtained from other computers and associated to different investigation processes. Clustering algorithms are usedwhich are capable of finding latent patterns from text documents found in seized computers, can enhance the analysis performed by the expert examiner. The rationale behind clustering algorithms is that objects within a valid cluster are more similar to each other than they are to objects belonging to a different cluster. Thus, once a data partition has been induced from data, the expert examiner might initially focus on reviewing representative documents from the obtained set of clusters. Then, after this preliminary analysis, (s) he may eventually decide to scrutinize other documents from each cluster. By doing so, one can avoid the hard task of examining all the documents (individually) but, even if so desired, it still could be done.

DISADVANTAGES OF EXISTING SYSTEM:
The literature on Computer Forensicsonly reports the use of algorithms that assume that the number of clusters is known and fixed *a* prioriby the user. Aimed at relaxing this assumption, which is often unrealistic in practical applications, a common approach in other domains involves estimating the number of clusters from data.

IV. PROPOSED SYSTEM

In proposed system, we can work upon unsupervised datasets, which is type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.We are using clustering algorithms which indeed tend to induce clusters formed by either relevant or irrelevant documents, thus contributing to enhance the expert examiner's job. Furthermore, our evaluation of the proposed approach in applications shows that it has the potential to speed up the computer inspection process.

V.BLOCK DIAGRAM OF SYSTEM

Architecture diagram consist of 4 modules

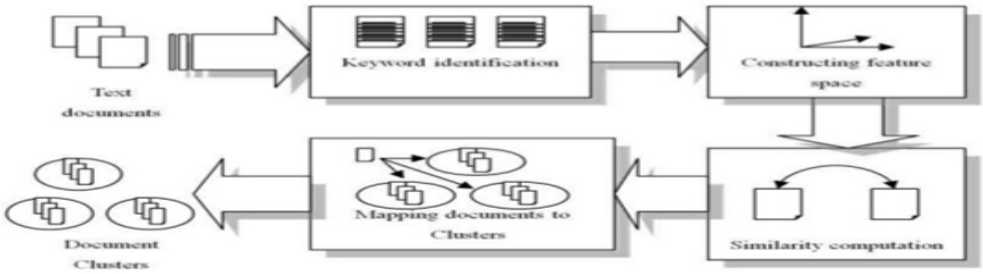
1. Keyword identification
2. Constructing feature space
3. Similarity computation
4. Mapping documents to the

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. 4, Issue 5, May 2016



VI.ALGORITHM

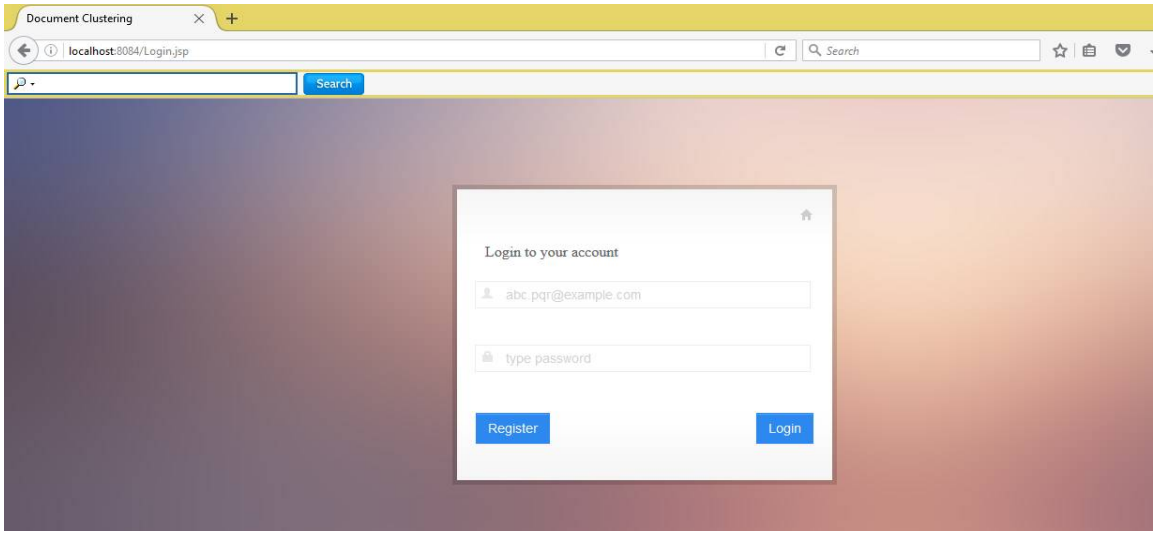
CLUSTERING ALGORITHM

Step 1 : Randomly selects numbers of cluster k
Step 2: Repeat
Step 3: Allocate or reallocate the number of objects to the cluster with minimum distance
Step 4: Calculate the mean
Step 5: Until the number of objects in the cluster remains same.

K-means algorithm is sensitive to the initialization of number of clusters.

VII. RESULT

1.User Login : User can login using registered email I'd and password.



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

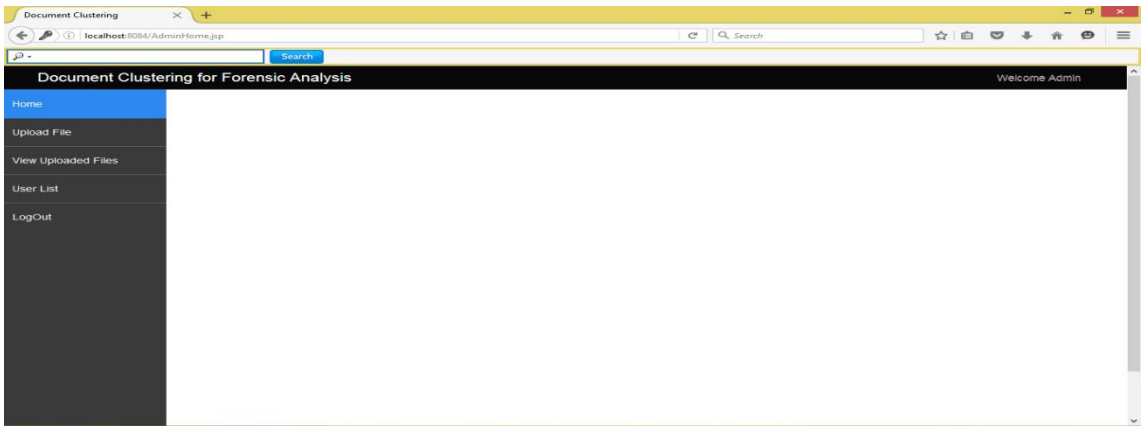
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

2.**Admin Login** : Admin has authority to

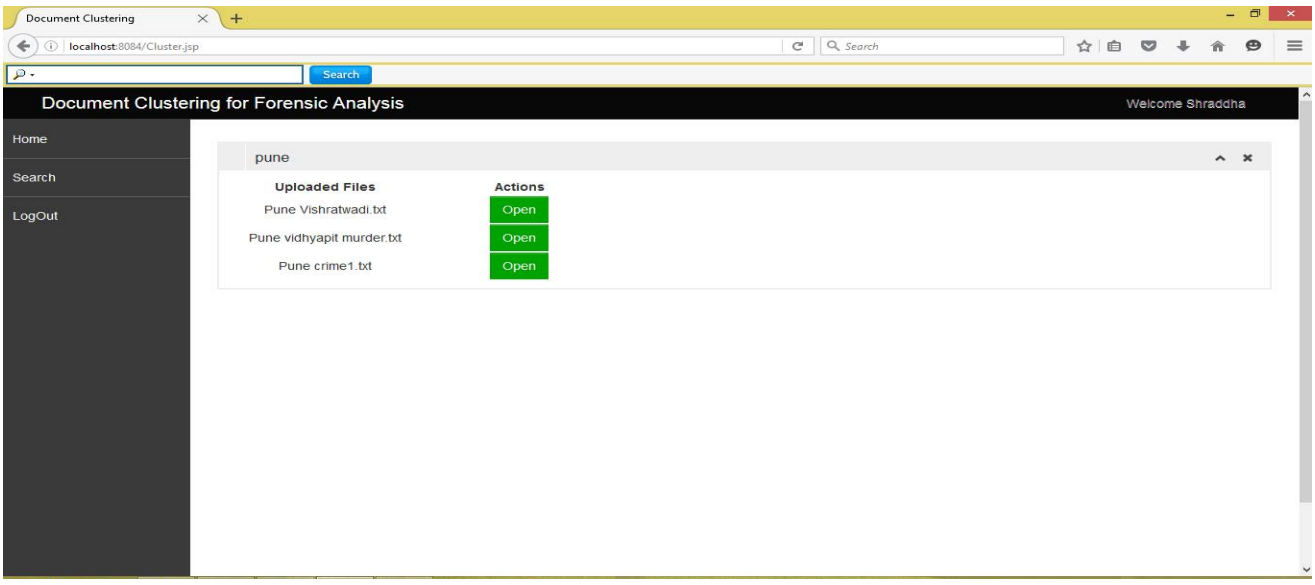
- Delete Files
- Upload Files
- Search Files
- Add User
- Delete User



3.Preprocessing : It includes

- Removing Stopwords.
- Potter Steming.
- Calculate Frequency of Keywords.

4.Clustering: Clustering is done based on related documents to the users query



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. 4, Issue 5, May 2016

VIII. CONCLUSION

We present an approach that applies document clustering method to forensic analysis of computer seized in police investigations. There are several practical results based on our work which are extremely useful for the experts working in forensic computing department.

IX.FUTURE WORK

1. Investigating automatic approaches for cluster labeling.
2. Different latest clustering algorithms to obtain best result.
3. Clustering of other unstructured text.

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. 121, 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. 583617, 2002.