ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 11, May 2015

Improving Computer Inspection: Through Document Clustering and Automatic Cluster Summarization

Adagale Sushadevi Shamrao, Amrit Priyadarshi, Shubhangi Sagar Vairagar

Abstract*— Document clustering or unsupervised document classification is an automated process of grouping documents with similar content. Document clustering is an important task in many Information Retrieval systems. Also document clustering Algorithms can help in discovery of new and useful knowledge or novel class from the documents under analysis. This knowledge or novel class is very important issue while handling forensic analysis. Digital Forensic Investigation is the branch of scientific forensic process for investigation of material found in digital devices related to computer crimes. In computer forensics, hundreds of thousands of files per computer are examined. Hence methods for automated data analysis, such as clustering are required. Labeling large datasets with clusters bridges the effective cluster analysis to the large dataset. Labeling irregular shaped clusters, distinguishing outliers and extending cluster boundary are the main problems in this stage. We address these problems and propose a cluster labeling algorithm which is very intuitive and easy to use.

Index Terms—**Clustering, forensic computing, text mining.**

I. INTRODUCTION

The latest edition of the annual Internet Trends report finds continued robust online growth. This survey shows that in near future digital information system may grows t 7-8 zettabytes of data up to year 2015. This survey itself shows that digital document handling is very important but complex task exist. This large amount of data 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. In our particular application domain, it usually involves examining hundreds of thousands of files per computer. This activity exceeds the expert's ability of analysis and interpretation of data. Therefore, methods for automated data analysis, like those widely used for machine learning and data mining, are of paramount importance. In particular, algorithms for pattern recognition from the information present in text documents are promising, as it will hopefully become evident later in the paper. From a more technical viewpoint, our datasets consist of unlabeled objects—the classes or categories of documents that can be found are a priori unknown. Moreover, even assuming that labeled datasets could be available from previous analyses, there is almost no hope that the same classes (possibly learned earlier by a classifier in a supervised learning setting) would be still valid for the upcoming data, obtained from other computers and associated to different investigation processes. More precisely, it is likely that the new data sample would come from a different population. In this context, the use of clustering algorithms, which are capable of finding latent patterns from text documents found in seized computers, can enhance the analysis performed by the expert examiner [1]. In this work, however, the clustering and labeling tasks are separated into two independent processes. First, a cluster partition of the data set is produced by a fully unsupervised clustering algorithm. Then, given a small set of labels (also referred to as prototype of labeled seed), a cost matrix is computed based on the distribution of labels throughout the clusters. The cluster labeling objective is then formulated as an assignment problem that is solved using the Hungarian algorithm [2]. Thereby, an optimum cluster labeling given the labeled seeds is ensured.

II. LITERATURE SURVEY

The digital world is expanding rapidly. A survey and forecast of worldwide information growth is worth consideration. The concept of relative validity index to estimate the best value for the number of clusters is used. Illustration of partitional algorithms like the K-means, K-medioids, the hierarchical, Single /Complete /Average Link and the cluster ensemble algorithms known as CSPA a traditional statistical approach for text mining, in which documents are represented in vector space model is adopted. The property of medioids makes it useful for applications in which (i) Centroids cannot be computed and (ii) distances between pair of objects are available, as for computing dissimilarities between names of documents with the Levenshtein distance. **Amparo Albalate, Aparna Suchindranath, David Suendermann, Wolfgang Minker "A semi-supervised cluster-and-label approach for utterance classification"** in propose a semi-supervised cluster-and-label algorithm for utterance classification. The approach assumes that the underlying class distribution is roughly captured through fully unsupervised clustering. Then, a minimum number of labeled examples is used to automatically label the extracted clusters so that the initial label set is augmented to the whole clustered data.

Advantages:

1. A semi-supervised cluster-and-label algorithm for utterance classification.
2. An unsupervised clustering algorithm is used to obtain a cluster partition of the utterance training set.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 11, May 2015

Disadvantages:
1. Its not suitable for forensic dataset.

2]Vidhya B, Priya Vaijayanthi R " Enhancing Digital Forensic Analysis through Document Clustering " January 2014

Digital forensic is the process of uncovering and interpreting process of uncovering and interpreting electronic data for use in a court of law. The goal of the process is to preserve any evidence in its most original form while performing a structured investigation by collecting identifying and validating the digital information for the purpose of reconstructing past events. Digital forensics deals with the analysis of artifacts on all types of digital devices.
The role of digital forensics is to facilitate the investigation of criminal activities that involve digital devices, to preserve, gather, analyze and provide scientific and technical evidence, and to prepare the documentation for law enforcement authorities. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories. Document clustering involves descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document cluster is generally considered to be a centralized process.

Advantage:
1. Digital forensics deals with the analysis of artifacts on all types of digital devices.

Disadvantage:
1. Depend on single clustering algorithms and not handles annotation.

3]K.Pallavi, S.NagarjunaReddy, Dr.S.Sai Satyanarayana Reddy "Clustring of Document in Forensic analysis for imporving computer inspection"July 2014

In Forensic Analysis thousands of files are usually examined. Data in those files consists of unstructured text analyzing it by examiners is very difficult. Algorithms for clustering documents can facilitate the discovery of new and useful knowledge from the documents under analysis. Cluster analysis itself is not one specific algorithm but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster. Here we propose an approach that applies to clustering of Documents seized in police investigations. We define the proposed approach with K-Means algorithm.

Advantage:
1. Efficient algorithm which work with very normal datasets.

Disadvantage:
1. Works only with K-Means algorithm

4]G. Madan Kumar, Sunil Kumar. V "File Clustering using Forensic Analysis System " July 2014

Clustering is the unverified organization of designs that is data items, remarks, or feature vectors into groups (clusters). To find a noble clarification for this automated method of analysis are of great interest. In particular, algorithms such as K-means, K-medics, Single Link, Complete Link and Average Link can simplify the detection of new and valuable information from the documents under investigation. present a tactic that applies text clustering algorithms to forensic examination of computers seized in police investigations using multithreading technique for data clustering. Our experiments show that the Average Link and Complete Link algorithms provide the best results for our application domain. If suit-ably initialized, partition algorithms (K-means and K-medoids) can also yield to very good results.

Advantage:
1. Clustering algorithms to forensic examination of computers seized in police investigations using multithreading technique for data clustering.

Disadvantage:
1. Huge computation required and not handles annotation.

5]Lus Filipe da Cruz Nassif and Eduardo Raul Hruschka, IEEE" Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection" IEEE transactions on knowledge and data engineering, JANUARY 2013

In computer forensic analysis, hundreds of thousands of files are usually examined. Much of the data in those les consists of unstructured text, whose analysis by computer examiners is difficult to be performed. In this context, automated methods of analysis are of great interest. In particular, algorithms for clustering documents can facilitate the discovery of new and useful knowledge from the documents under analysis.
Advantages:
1. Document clustering methods to forensic analysis of computers seized in police investigations..

Disadvantages:
[1] It creates number of cluster. But to do analysis it not provides cluster labels which make analysis easy.

III. ARCHITECTURE

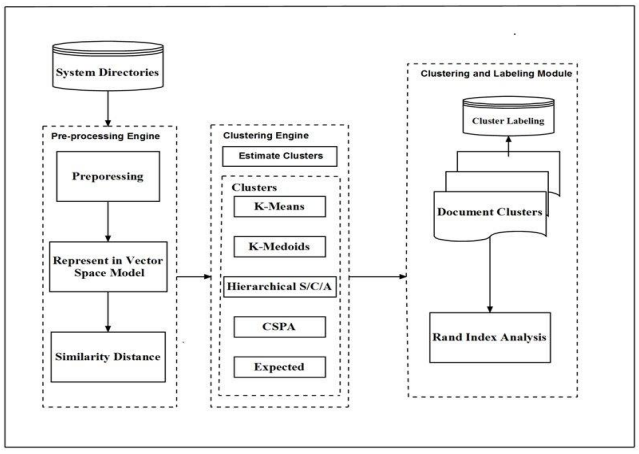


Fig 1: system architecture

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 11, May 2015

It will also include the study of partitioned algorithms such as K-means & K-medioids; hierarchical algorithms – single, complete, average link and cluster ensemble based algorithms known as CSPA. The new proposed main work is to develop a novel hierarchal algorithm for document clustering in Computer Forensics, which provides labels for each clusters so that, it is very easy to do analysis of data for every cluster. Major functional components are:-

I. **System Data:**
This module is responsible for reading, listing all relevant documents. This module is act as reading input. Documents are stored in various format e.g. unstructured or semi-structured format. All relevant documents are listed out.

II. **Pre-processing**
In this process we remove unwanted data or manipulate data which helps to cluster algorithm work efficiently. This process includes stop word removing, stemming, pruning etc.

III. **Represent data in vector**
Vector space model or term vector model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers.

IV. **Similarity distance**
This takes input as data vector and using similarity algorithm we get weight matrix for data vector or documents.

V. **Estimate clusters**
A. K-means. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

B. CSPA – Using METIS algorithm to partitions the data in Similarity graph to obtain desired number of clusters.

C. K-medioids - The *k*-medoids algorithm is a clustering algorithm related to the k-means algorithm and the medoid shift algorithm.

D. EM - an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step.

VI. **Labeling** – here we get input as various cluster and along with its members. We presented a semi-supervised cluster-and-label approach to classification of utterances has been presented. Then, the take output cluster partition to this algorithm as well as a small set of labeled prototypes (also referred to as labeled seeds) are used to determine the optimum cluster labeling related to the labeled seed. We formulated the cluster labeling problem as an assignment optimization problem whose solution is obtained by means of the Hungarian algorithm.

VII. **Project Analysis** –
1. Label accuracy
2. Rand index

IV. IMPLEMENTATION DETAILS - DOCUMENTS CLUSTERING

A. Estimating number of clusters

In order 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 provides the best result according to a specific quality 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 clusters. A widely used relative validity index is the so-called silhouette [6]

B. Clustering Algorithms

The clustering algorithms adopted in our study—the partitional K-means [7] and K-medoids [8], the hierarchical Single/Complete/Average Link [9], and the cluster ensemble based algorithm known as CSPA [10]—are popular in the machine learning and data mining fields, and therefore they have been used in our study.

C. Cluster labeling

Given the training data, XT=X(t)T S the set Y(t)T of labels associated with the portion X(t)T of the training set, the set K of labels for the k existing classes1, and a cluster partition C of XT into disjoint clusters, the optimum cluster labeling problem is to find a objective mapping function
L: C K, K= 1, 2, 3, .., k
That assigns each cluster in C to a class label in K, while mini minimizing the total labeling cost. This cost is defined in terms of the labeled seed (X(t)T, Y(t)T) and the set of clusters C. Consider the following matrix of overlapping products N:

N = ni1 . . . ni2 . . . . . nik

n21 . . . n22 . . . . . n2k

nk1 . . . nk2 . . . . . nkk

With constituents nij, denoting the number of labeled patterns from X(t)T with class label y=I that fall into cluster Cj. The labeling objective is to minimize the global cost of the cluster labeling denoted by L:
Total cost (L) =∑ wi.Cost(L(Ci)) …..... (1)
Where W= (w1, , wk) is a vector of weights for the different clusters. For example, it may be used if cluster sizes show significance differences among the clusters. In this approach, the weights are assumed to be equal for all clusters, so that wi = 1,
The individual cost of labeling the cluster Ci, with class L(Ci) is defined as the number of samples from class L(Ci)(in the labeled seed) that fall outside the cluster Ci i.e. :

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 11, May 2015

Cost (L (Ci)) = ∑ ηL(Ci)k.....(2)
Applying Equation 2 to the total cost definition of Equation 1 yields:
Total Cost (L) = ∑ci ε c∑c ≠ cj(Ci)k..... (3)
In this applied the Hungarian algorithm to achieve the optimum cluster labeling in Equation 3. It requires the definition of the cost matrix C[kxk] whose rows denote the clusters and the columns refer to class labels in K. The elements Cij denote the individual costs of assigning the cluster Ci to class label j, i.e. Cij = Cost (L (Ci) = j). The reader is referred to [6] for further details about the assignment problem and the Hungarian algorithm

V. RESULT

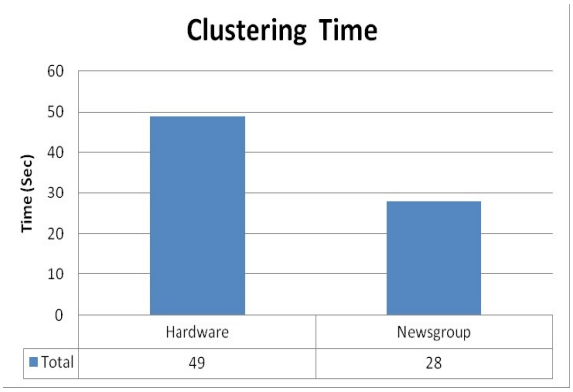


Fig.2: clustering time compare

We tested this project on standard Hardware dataset and Newsgroup dataset. Here we showed how much time it takes to cluster all documents. It shows if dataset size increases clustering time increases.

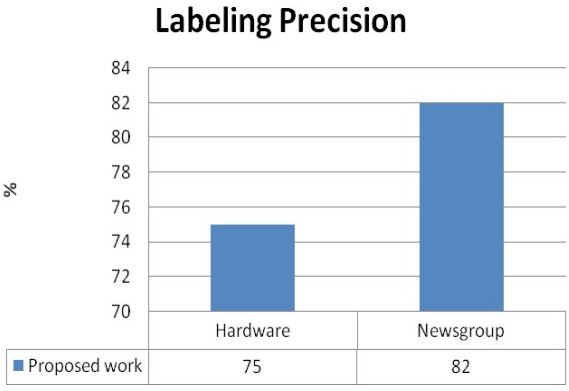
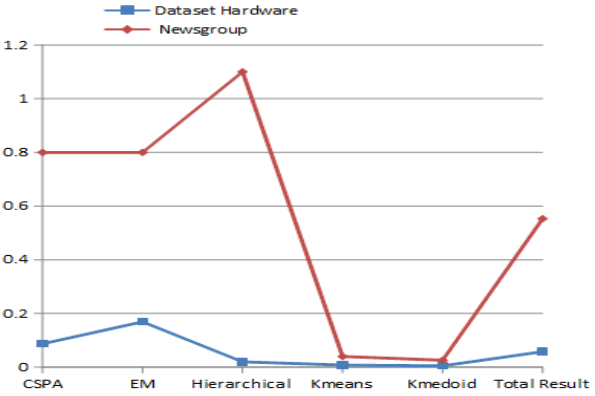


Fig 3::Labeling Precision



We also calculated how our labeling algorithm works. We used precision as parameter to calculate or accuracy of clustering labeling. Graph shows that our clustering algorithm results are very good.

We tested this project on standard Hardware dataset and Newsgroup dataset. Above graph shows the variations of different dataset hardware vs different algorithms. This graph highlighted with blue colour. Also the graph shows the variations of newsgroup hardware vs different algorithms. This graph highlighted with red color.

VI. CONCLUSION

We presented an approach that applies document clustering methods to forensic analysis of computers seized in police investigations. Also here we presented efficient clustering labeling algorithm which helps forensic analyst to handle cluster efficiently after getting the cluster labels. Here we showed how cluster labeling work and gives better labeling. By using these labels it's very easy to analyst to handle cluster.

VII. ACKNOWLEDGMENT

Our heartfelt thanks go to DGOI Faculty Of Engineering providing a strong platform to develop our skills and capabilities. I would like to thank to our guide respected teachers for their continuous support and incentive for us. Last but not least, I would like to thanks to all those who directly or indirectly help us in presenting the paper.

REFERENCES

[1] Luís Filipe da Cruz Nassif and Eduardo Raul Hruschka Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection – 2014

[2] Amparo Albalate1, Aparna Suchindranath1, David Suendermann2, Wolfgang Minker1 "A semi-supervised cluster-and-label approach for utterance classification"

[3] K.Pallavi1, S.NagarjunaReddy2, Dr.S.Sai Satyanarayana Reddy "CLUSTERING OF DOCUMENTS IN FORENSIC ANALYSIS FOR IMPROVING COMPUTER INSPECTION 2014"

[4] Forensic Analysis of the Tor Browser Bundle on OS X, Linux, and Windows Runa A. Sandvik runa@torproject.org Tor Tech Report 2013-06-001 June 28, 2013

[5] Jianchu KANG, Jing YU, Zhongliang WANG"A Comparative Study on Unsupervised Feature Selection Methods for Text Clustering Luying LIU"

[6] Gharib1,2, Mohammed M. Fouad3 , Abdulfattah Mashat1 ,Ibrahim Bidawi1"Self Organizing Map -based Document Clustering Using WordNet" Ontologies 2012 Tarek F.

[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, and M. S. Oliver, "Exploring forensic data with self-organizing maps," in Proc. IFIP Int. Conf. Digital Forensics, 2005, pp. 113–123.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 11, May 2015

[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.

AUTHOR'S PROFILE

Miss. Adagale Sushadevi Shamrao Student, Computer Engineering DGOI, COE, Swami-Chincholi, Daund, Pune, India.Email:susha0810@gmail.com

Prof. Amrit Priyadarshi, Assistant Professor Computer Engineering DGOI, COE, Swami-Chincholi, Daund, Pune, India. Email:amritpriyadarshi@gmail.co

Ms. Shubhangi Sagar Vairagar. Computer Engineerin Pune, India Email vairagarss@gmail.com