File Clustering for Similar Dataset an Advance for Improving Computer Scrutiny

[1] M.N.BHASKAR, [2] K. CHANDANA

[1] Assistant Professor, Dept of CSE, Shri Shirdi Sai Institute of Science & Engineering, Affiliated to JNTUA, AP, India.
[2] M.Tech Dept of CSE, Shri Shirdi Sai Institute of Science & Engineering, Affiliated to JNTUA, AP, India.

***Abstract*— The evolution of the insecurity problematics, the arrival of new threats (terrorism, cybercrime, etc.) and the development of new technologies are factors which dramatically increased the importance of intelligence in the process of management, analysis, and utilization of the growing volumes of available crime data. The design of specific intelligence processes and computational systems for crime analysis is related to the ”type” of intelligence that is considered. Current digital forensic text string search tools fail to group and/or order search hits in a manner that appreciably improves the investigator’s ability to get to the relevant hits first. Text string search results are extremely noisy, which results in inordinately high levels of information retrieval (IR) overhead and information overload.**

Algorithms for clustering documents can facilitate the discovery of new and useful knowledge from the documents under analysis. In this paper, we propose the approach of applying document clustering algorithms to forensic analysis of computers seized in police investigations.

***Keywords:* Forensic analysis, Document clustering, computer forensics, forensic computing, text mining**

I. INTRODUCTION

Digital forensic tools are not being developed fast enough to keep pace with the variety of forensic targets. The explosion of growth that technology and in particular the computing world, has resulted in highly sophisticated equipment. This has in essence intensified the criminals’ potential to perform criminal activity. Forensic investigators follow a generalised methodology when conducting an investigation to ensure credibility and integrity of the digital devices. There are a number of frameworks and methodologies that cover the digital forensic investigation differently. Computer Forensics combines elements of law and computer science to collect and analyze the data. Examining hundreds and thousands of data has a direct impact in the field of computer forensics. Hence, the methods used for automated data analysis are of significance in the data analysis.

Text clustering techniques[3] are applied pre-retrieval and/or post-retrieval. thematic clustering of text string search hits will lead to separation between investigatively relevant and investigatively irrelevant hits. The computational expense of model-based clustering approaches varies between approaches, but is often higher order with respect to input size.

In this paper, we propose a comparative analysis of the six-well known algorithms K-means, K-medoids, Single Link, Complete Link, Average Link, and CSPA applied to five real-world datasets obtained from computers seized in real-world investigations. 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. In a practical scenario, domain experts (e.g., forensic examiners) are scarce and have limited time available for performing examinations. Thus, it is reasonable to assume that, after finding a relevant document, the examiner could prioritize the analysis of other documents belonging to the cluster of interest, because it is likely that these are also relevant to the investigation.

II. EXISTING SYSTEM

In practical, the widely used existing clustering algorithms in the field of computer forensics are K-Means, Expectation-Maximization (EM) for unsupervised learning of Gaussian Mixture Models, Fuzzy C-means (FCM), and Self-Organizing Maps (SOM).

Fuzzy clustering[2] is used to detect the explanation of criminal activities for crime hot-spot areas and their spatial trends. Compared with two hard-clustering approaches (median and k-means clustering problem), the empirical results suggest that a fuzzy clustering approach is better equipped to handle crime spatial outliers. The Fuzzy C-Means Clustering (FCM) is an unsupervised goal oriented clustering algorithm. SOM-based algorithms used for clustering files aims at facilitating an efficient decision-making process by the examiners. The files were clustered by taking into account their creation dates/times and their extensions.

E-mail forensics use an integrated environment of classification and clustering algorithms. Practically, e-mails are grouped by lexical, syntactic, structural and domain-specific features. In the field of Computer forensics, the number of clusters should be fixed and are known in prior to the user which is not feasible all the time for all kinds of investigations.

III. PROPOSED SYSTEM

Considering the partitional algorithms, it is widely known that both K-means and K-medoids[2] are sensitive to initialization and usually converge to solutions that represent local minima. In this work, we propose a nonrandom initialization in which distant objects from each other are chosen as starting prototypes. the quality of every partition represented by the dendrogram, subsequently choosing the one that provides the best results. 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. set of partitions may result directly from a hierarchical clustering dendrogram or, alternatively, from multiple runs of a partitional algorithm (e.g., K-means) starting from different numbers and initial positions of the cluster prototypes. The relative validity index is the so-called *silhouette* , which has also been adopted as a component of the clustering algorithms[5].

Let us consider an object i belonging to cluster $\mathbf{A}$. The average dissimilarity of i to all other objects of $\mathbf{A}$ is denoted by $a(i)$. Now let us take into account cluster $\mathbf{C}$. The average dissimilarity of i to all objects of $\mathbf{C}$ will be called $d(i, \mathbf{C})$. After computing $d(i, \mathbf{C})$ for all clusters $\mathbf{C} \neq \mathbf{A}$, the smallest one is selected, i.e., $b(i) = \min d(i, \mathbf{C})$. $\mathbf{C} \neq \mathbf{A}$. This value represents the dissimilarity of i to its neighbor cluster, and the silhouette for a give object, $s(i)$, is given by:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

The average silhouette just addressed depends on the computation of all distances among all objects. The CSPA algorithm [6] essentially finds a consensus clustering from a cluster ensemble formed by a set of different data partitions. A similarity matrix is constructed in which each element of this matrix represents pair-wise similarities between objects. The similarity between two objects is simply the fraction of the clustering solutions in which those two objects lie in the same cluster. For the hierarchical algorithms (Single/Complete/Average Link), the best partition elected according to the relative validity index is taken as the result of the clustering process.

For each partitional algorithm (K-means/medoids), we execute it repeatedly for an increasing number of clusters. For each value of , a number of partitions achieved from different initializations are assessed in order to choose the best value of and its corresponding data partition. 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 sdata partition (for evaluation purposes) as single clusters.

In particular, any kind of content that is digitally compliant can be subject to investigation. In the datasets assessed in our study, for instance, there are textual documents written in different languages. documents have been originally created in different file formats, and some of them have been corrupted or are actually incomplete in the sense that they have been (partially) recovered from deleted data. The obtained data partitions were evaluated by taking into account that we have a *reference partition* for every dataset.

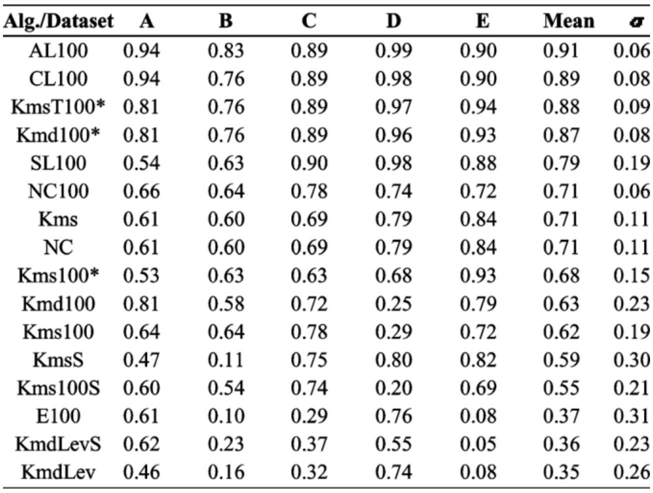
TABLE II
DATASET CHARACTERISTICS[1]

Dataset	*N*	*K*	*D*	*S*	# Largest cluster
A	37	23	1744	12	3
B	111	49	7894	28	12
C	68	40	2699	24	8
D	74	38	5095	26	17
E	131	51	4861	31	44

where we have documents(N), groups(K), attributes(D) , singleton(S), number of documents per group(#).

Considering the algorithms that recursively apply the Silhouette for removing singletons (KmsS and Kms100S), Table II shows that their results are relatively worse when compared tothe related versions that do not remove singletons (Kms and Kms100).

Table II ADJUSTED RAND INDEX (ARI) RESULTS



Alg./Dataset	A	B	C	D	E	Mean	σ
AL100	0.94	0.83	0.89	0.99	0.90	0.91	0.06
CL100	0.94	0.76	0.89	0.98	0.90	0.89	0.08
KmsT100*	0.81	0.76	0.89	0.97	0.94	0.88	0.09
Kmd100*	0.81	0.76	0.89	0.96	0.93	0.87	0.08
SL100	0.54	0.63	0.90	0.98	0.88	0.79	0.19
NC100	0.66	0.64	0.78	0.74	0.72	0.71	0.06
Kms	0.61	0.60	0.69	0.79	0.84	0.71	0.11
NC	0.61	0.60	0.69	0.79	0.84	0.71	0.11
Kms100*	0.53	0.63	0.63	0.68	0.93	0.68	0.15
Kmd100	0.81	0.58	0.72	0.25	0.79	0.63	0.23
Kms100	0.64	0.64	0.78	0.29	0.72	0.62	0.19
KmsS	0.47	0.11	0.75	0.80	0.82	0.59	0.30
Kms100S	0.60	0.54	0.74	0.20	0.69	0.55	0.21
E100	0.61	0.10	0.29	0.76	0.08	0.37	0.31
KmdLevS	0.62	0.23	0.37	0.55	0.05	0.36	0.23
KmdLev	0.46	0.16	0.32	0.74	0.08	0.35	0.26

Both the silhouette and its simplified version estimate the number of clusters by taking into account two concepts: cluster compactness (average intracluster dissimilarity) and cluster separability (inter-cluster dissimilarity). Both arematerialized by computing average distances. Also, the average distance from a given object to all the objects of a cluster tends to be greater than the distance of that object to the cluster's centroid. As far as the adopted dimensionality reduction technique is concerned—Term Variance (TV) we observed that the selection of the 100 attributes (words) that have the greatest variance over the documents provided best results than using all the attributes in three out of five datasets.

IV. CONCLUSION

The computational cost of estimating the number of clusters, depends on the computation of all distances between objects, leading to an estimated computational cost of $O(N^2.D)$, where N is the number of objects in the dataset and D is the number of attributes, respectively. The simplified silhouette is based on the computation of distances between objects and cluster centroids, thus making it possible to reduce the computational cost from $O(N^2.D)$ to $O(k.N.D)$, where , the number of clusters, is usually significantly less than . The partitional K-means and K-medoids algorithms also achieved good results when properly initialized. Considering the approaches for estimating the number of clusters, the relative validity criterion known as *silhouette* has shown to be more accurate than its basic version.

The future work aims at investigating automatic approaches for cluster labeling. The assignment of labels to clusters may enable the expert examiner to identify the semantic content of each cluster more quickly—eventually even before examining their contents.

.

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] AK. 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.

[3] JN. 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.

[4] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, pp. 193–218, 1985.

[5] B. Mirkin, *Clustering for Data Mining: A Data Recovery Approach*. London, U.K.: Chapman & Hall, 2005.