ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 6, Issue 5, November 2016

Document clustering -- an approach for improving computer inspection

K.Navya[1], D. Kamal Kumari [2], S. Ram Prasad Reddy [3],

[1]M.Tech, CSE, Vignan's Institute of Engineering for Women, Visakhapatnam, AP.
[2] Assistant Professor, CSE, Vignan's Institute of Engineering for Women, Visakhapatnam, AP.
[3] *Professo*r, CSE, Vignan's Institute of Engineering for Women, Visakhapatnam, AP.

Abstract*— Clustering is approach that organizes a large quantity of unstructured data documents into small number of meaningful and coherence clusters. present an approach the applied document clustering algorithms to forensic analysis of computers seize in police investigations. The proposed approach is carried out with extensively experiment on six well-known clustering algorithms (K-means, K-medoids, Single Link, Complete Link, Average Link, and CSPA) applies to five real-world datasets obtained from computers seize in real-world investigations. An experiment has been performed with different combinations of parameters, resulting in 16 different instantiations of algorithms. In addition, two relative validity indexes are used to automatically estimate the number of clusters. Experiments show that the Average Link and Complete Link algorithms provide the best results for our application domain. If suitably initialized, partitional algorithms (K-means and K-medoids) can also produce to very good results. Finally, we also present and discuss several practical results that can be useful for researchers and practitioners of forensic computing.

Index Terms— Clustering; forensic computing; text mining; k-means clustering; term frequency .

I. INTRODUCTION

Clustering is a divided of data into groups of similar objects. Each group, called as cluster. In clusters consists of objects that are similar between themselves and dissimilar to objects of other groups. In other words, the goal of good document clustering is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances (using to appropriate distance measure between documents). A distance measure (are dually, similarity measure) thus lies at the heart of document clustering.

Clustering is the mostly common form of unsupervised learning is the major difference between clustering and classification. No supervised means that there is no human expert who has assign documents to classes. In clustering, it is the distribution and make of the data that will determined cluster membership. Clustering is sometimes erroneously referred to as automatic classification; however, this is the inaccurate, since the clusters find are not known prior to processing where in case of classification and the classes are pre-defined. In clustering, it is the distribution and the nature of data that will determine cluster membership, in opposition to classification where the classifier learn the association between the objects and classes from a so called training set, i.e. a set of data correctly labelled by hand, and then replicate the learnt behaviour on unlabelled data. By clustering the articles we could reduce domain of search for recommendations as most of the users had interest in the news corresponding number of clusters. This improved our time efficiency to a great extent.

Clustering algorithms has been studied for decades, and the literature on the subject is huge. Therefore, we are decided to choose set of (six) representative algorithms in order to show the potential of the proposed approaches, name as: the partitional K-means and K-medoids, the hierarchical Single/Complete/Average Link, and the CSPA .These algorithms were run with different combinations of their parameters; result in sixteen different algorithmic instantiations.

It is important to emphasize that getting from a collection of documents to a clustering of the collection, is not only a single operation, but is more a process in multiple stages and include more traditional information retrieval operations that are crawling, indexing, weighting, filtering etc. Some of other processes are central to the quality and performance of most of clustering algorithms, and it is thus necessary to consider these stages together with a given clustering algorithm to arm its true potential. We will give the brief overview of the clustering process.

II. METHODOLOGY

A. Collection of data

Document clustering is the process of grouping similar documents into cluster which benefit to the retrieve information effectively, reducing the search time and space.Aim is automatically group related documents into clusters and also one of the tasks in machine learning and artificial intelligence and have received much attention in recent years. Clustering is one of the techniques of data mining extract the knowledge from large amount of information. Collections of files involve the processes like obtain the data and documents from the computer seized devices. The collection of data files and documents involves the special techniques.

B. Pre-Processing Steps

It is done to represent the data in a form that can use for clustering. There are many types of representing the documents

DOI:10.17605/OSF.IO/D3FR6

Example vector-Model, etc. Many measures are also used for weighing the documents and their similarities between them.

Vector space model:

Document and queries are both vectors.

$\vec{d} = (wi1, wi2, \ldots..wit)$

Each wi1 is a weighting for term j in document 1.

“Bag of words representation”.

Similarity of document vector to query vector=cosine of angle between them.

1. *Stemming*

Stemming is process of changing the words in the document to obtain a root word with certain rules. Stemming is used to maximize the information retrieval in a document.

A stemming algorithm has three main objectives. The first is grouping of words according to their topics. Many words from the same root derivation and derivation generate though additional affix (prefix, infix, and suffix).The second goal of a stemming algorithm is related to the process of finding information that has the same root, and those the grouping term by the root word makes it easy to index the documents. The third goal is the incorporate of variety of the same root to reduce the words that ate taken into account in the process of collection data there by reducing the space required to store the structure used by the information retrial system.

2. *Stop word Removal*

A term, which is not thought to convey any meaning as a dimension in the vector space (i.e. not context) is known as stop word. A typical method to remove stop words is by compare each term with a compilation of known stop words. This can be done by removing terms with low document frequencies and applying a part of speech tagger and then rejected all stop words such as nouns, verbs, pronouns, adjectives etc.

3. *Term Frequency*

Reduction technique known as Term Variance (TV) is also used to increase efficiency of clustering algorithms. As clusters are formed, which containing documents, term variance are used to estimating a top n word which has greatest occurrence over documents within clusters?

Term frequency: the number of times the term occurrence in the document.

Tf(t,d)

Inverse term frequency: total number of documents containing the term.

Idf(t,D)

$\{d \in D : t \in d\}$

Tf-idf

Tfidf(t,d,D)=tf(t,d)*idf(t,D)

4. *Similarity Computation*

It is important to find out distances between two documents when they are resides in different clusters and for finding out distances between them, cosine-based distance.

C. *Estimating the Number of Clusters from Data*

In order to estimating the number of clusters, a most of clusters used approach called as silhouette consists of getting a set of data partitions with different numbers of clusters and the selecting that particular partition that are provide the best result according to a specific quality criteria (e.g., a relative validity index). A set of partitions may result directly from a hierarchical clustering genogram or, alternatively, from multiple runs of a partitional algorithm (e.g., K-means) starts from different numbers and starting positions of the cluster prototypes. Let us consider an object belonging to cluster A. a(i) denotes the average dissimilarity of i to all other objects of A. Let us consider cluster C. The average dissimilarity calculates to i to all objects of cluster C will be called C(i). After compute the d (i,C) for all clusters C ≠ A, the one which is smallest one is selected, b(i) = min d(i,C), C ≠ A. This value represent the dissimilarity distance of i to its neighbour cluster, and the silhouette for a given object, s(i) is as below:

It can verified that $s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$

≤ s (i) ≤ 1 [1].

D. *Clustering Algorithms*

There are various algorithms which can use in clustering like K-means, K-medoids, Single Link, Complete Link and Average Link. One of simplest un supervision algorithms is the K-means. In single link clustering algorithm, two groups have been merged and their closest pair of documents has the highest similarity compressions to any other pair of groups. In complete linkage many elements in the clusters are different to each other. It produces more Comply clusters and most useful hierarchies than any other clustering.

1. *K-means*

K-Means algorithm is one of the simplest unsupervised learning methods among all partitioning based clustering methods. If k is the number of to desire clusters then it classifies the given set of n data objects in k clusters. Centroid is defined for each cluster. All Data objects have centroid nearest (or most similar) to that data object are placed in the cluster. After the processing all data objects, calculating centroids, and recalculated, and the entire process is repeats until no change. Based on the new calculated centroids, all data objects are considered to the clusters. In each iteration centroids change their location and as centroids move into the each iteration. This process is continuing until no change in the position of centroid. This results in k cluster representing a set of n data objects.

DOI:10.17605/OSF.IO/D3FR6

2. *Hierarchical Clustering*

Hierarchical clustering approaches attempt to create a hierarchical decomposition of the given document collection thus achieving a hierarchical structure. Hierarchical methods are classified into Agglomerative and Divisive methods depends how the hierarchy is constructed. Agglomerative method start with an initial clustering of the term space, where all documents are to be considered represents a separate cluster. The closest clusters using an inter-cluster similarity measure are then merged continuous until only 1 cluster or a predefined number of clusters remain.

E. *Removing Outliers*

We assert a simple approach to remove outliers. These approaches make recursive use of the silhouette. Fundamentally, if the best partition chose by the silhouette have singletons (i.e., clusters formed by a single object), these are removed. Then, the cluster processing is repeated over and over again—until partitions without singletons is found. At the end of the process, all singletons are incorporate into the resulting the data partition as single clusters.

III. RESULTS AND DISCUSSION

A. processing techniques

The Dataflow Diagram [figure 1] is also called as bubble chart. It is a simple graph formalism that can be used to represent a system in terms of input data to the system, various processing carries on this data, and the output data is generated by this system.

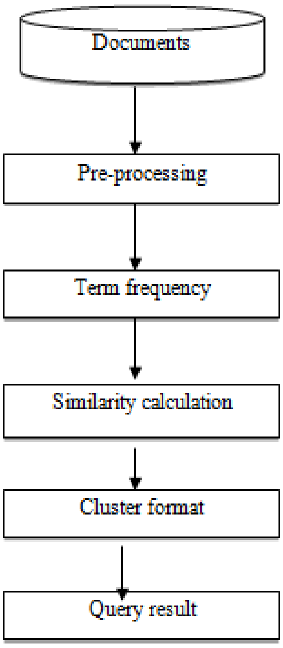


Fig1: flow chart

B. K-means algorithm: An algorithm for k-means method is given below.

Input: ‘*k*’, the number of clusters to partitioned, ‘*n*’, the number of objects.

Output: A set of ‘*k*’ clusters based on given similarity function.

Steps:

```
i) Arbitrary choose ‘k’ objects as the initial cluster
      centers;
ii) Repeat,
a. Assign each object to the cluster to which the object is
      the most similar; based on the similarity function;
b. Update the centroid (cluster means), by calculate the
      mean value of the objects for each cluster;
iii) Until no change.   [Figure2]
```

1 Basic Euclidean distance: Euclidean is one of the distance measures used on K Means algorithm. Euclidean distance between an observation and initial cluster 1 and 2 centroids is calculated. Based on Euclidean distance measures each observation is assigned to one of the clusters - based on minimum distance.

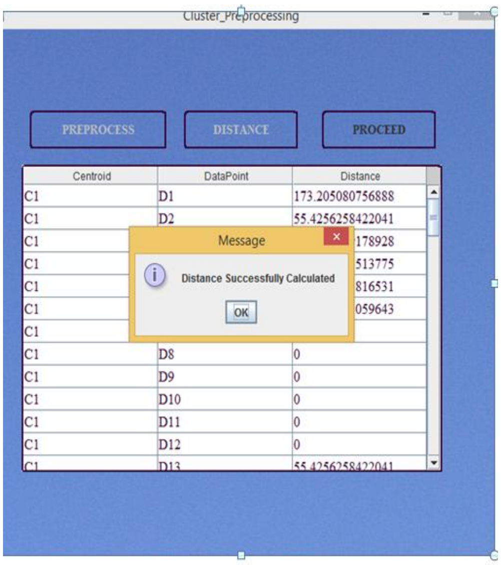
$$ED = \sqrt{(xh - h1)^2 + (xw - w1)^2}$$

Xh=observation value of variable height.

H1=centroid value of cluster 1 for variable height.

Xw=observation value of variable weight.

W1=centroid value of cluster 1 for variable weight.



Centroid	DataPoint	Distance
C1	D1	173.20508075688
C1	D2	55.4256258422041
C1		178928
C1		513775
C1		816531
C1		059643
C1		
C1	D8	0
C1	D9	0
C1	D10	0
C1	D11	0
C1	D12	0
C1	D13	55.4256258422041

Distance Successfully Calculated

Fig3: distance measure

C. Hierarchical clustering algorithm

This way we go on grouping the data until one cluster is formed. Now on the basis of dendogram graph we can calculate how many numbers of clusters should be actually present.

Step1: Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item

Step2: Let the distances (similarities) between the clusters the same as the similarity distances between the items they contain.

Step3: Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.

Step4: Compute distances (similarities) between the new cluster and each of the old clusters.

Step5: Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

1. Cosine Similarity Algorithm

Document clustering is particular useful in many applications such as automatic categorization of documents, grouping of search engine results, building taxonomy of documents, and others. Hierarchical Clustering method provides a better improvement in achieves the result.in paper two key parts of successful Hierarchical document clustering. The first part is a document index model, which allows for the incremental construction of the index of the document set with an emphasize on efficiency, rather than rely on single-term indexes only. It provides efficient phrase matching that is used to judge the similarities between documents. This model is flexible in that it could revert to a compact representation of the vector space model if we are not choose to index phrases. The second part is an incremental document clustering based on maximizing the tightness of clusters by carefully watching the pair-wise document similarity distribution. The combination of components creates an underlying model for robust and accurate document similarity calculation that leads to very much improved results in Web document clustering over traditional methods.

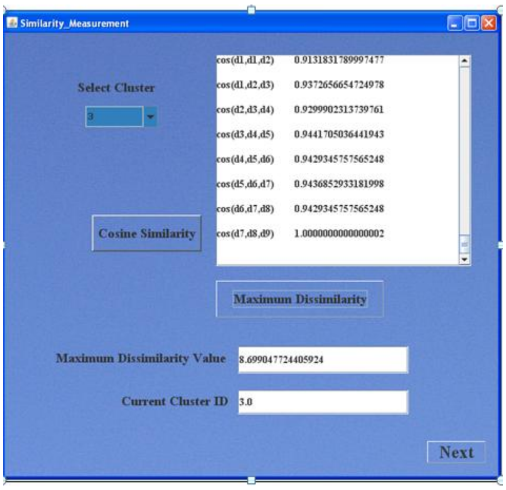
Given the set of N items to be clustered, and an N*N similarity distance matrix, the basic process of hierarchical clustering is this:

STEP 1 - Start by assigning each item to a cluster, so if you have N items, you now have N clusters, each containing just one item. Let the similarly distances between the clusters the same as the distances (similarities) between the items they contain.

STEP 2 - Find the closest mean of similar pair of clusters and merge them into a single cluster, so that now you have one cluster less with the help of tf - idf.

STEP 3 - Compute similarity distances between the new cluster and each of the old clusters.

STEP 4 - Repeat steps 2 and 3 until all items are clustered into a single cluster of size



```
cos(d1,d1,d2)    0.9131831789997477
cos(d1,d2,d3)    0.9372656654724978
cos(d2,d3,d4)    0.9299902313739761
cos(d3,d4,d5)    0.9441705036441943
cos(d4,d5,d6)    0.9429345757565248
cos(d5,d6,d7)    0.9436852933181998
cos(d6,d7,d8)    0.9429345757565248
cos(d7,d8,d9)    1.0000000000000002
```

Maximum Dissimilarity Value 8.699047724405924

Current Cluster ID 3.0

Fig5: cosine similarity

D. Comparative Statement of Clustering Techniques

Table 1: Clustering Techniques Comparison

Clustering Technique	Clustering Technique	Measure	Advantages
K-means	O(nki)	Mean	K-means is relatively scalable and efficient in processing large data sets
Single link	O(n^2)	Similarity Measure	1. Theoretical properties, efficient implementations, widely used. 2. No cluster centroid or representative required. no need arises to recalculate the similarity matrix
Average link	O(n^2 log n)	Similarity Measure	It is a structure intermediate between the loosely bound single link clusters and tightly bound complete link clusters.
Complete link	O(n^2 log n)	Similarity Measure	It gives good results as compared to single and average link.

An approach to that applies document clustering methods to forensic analysis of computers. This approach is can be very useful for researchers of organization relevant to working with data documents. More specifically, in the experiments the hierarchical algorithms known as Average

Link and Complete Link presented the best results. The figure main indicates the Clustering Accuracy. In the figure8 clustering accuracy of the techniques is representing in the form of a graph.

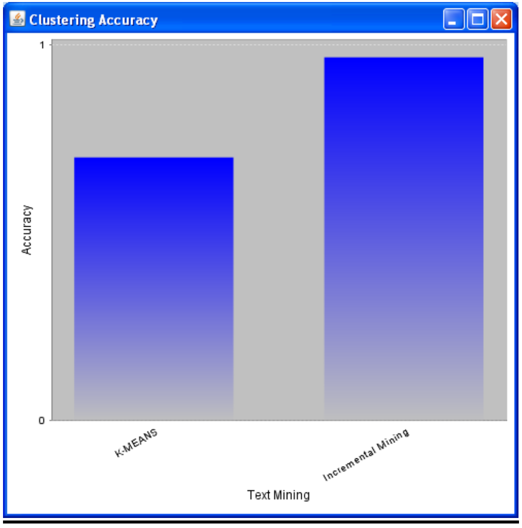


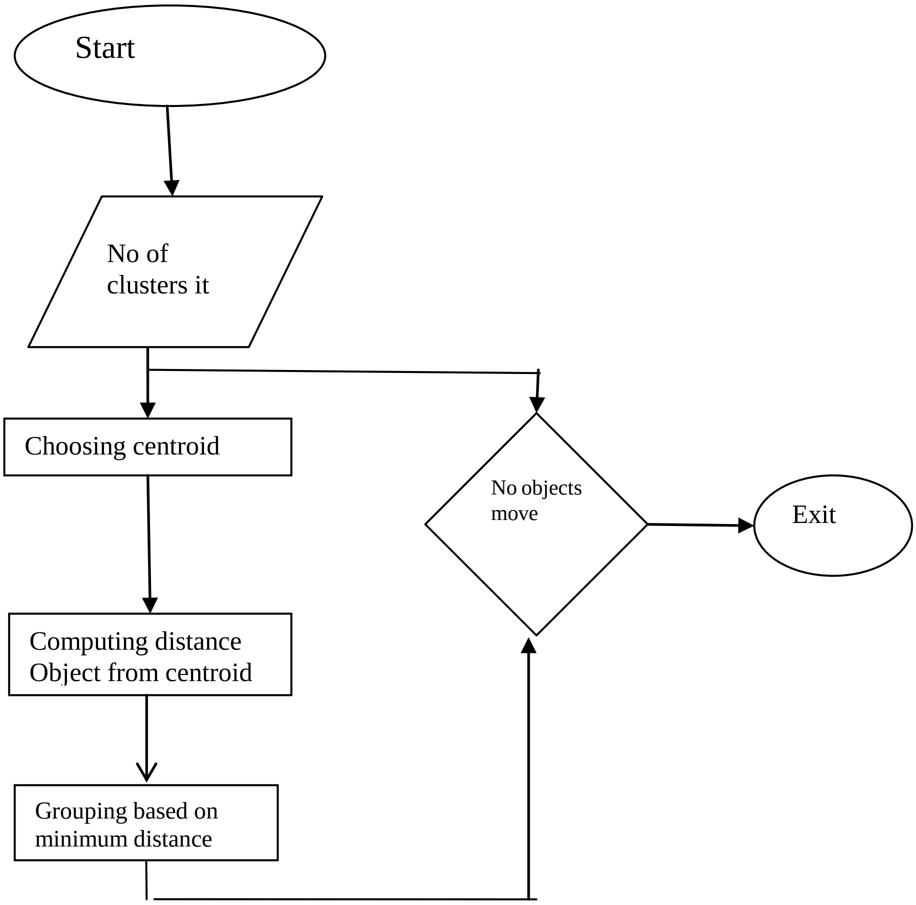
Fig 6: clustering accuracy

IV. CONCLUSION

We present an approach that applies document clustering methods to document analysis of computer document inspection. Also reported and discussed several practical results that can be very useful for researchers of document computing. It is specifically, in our experiments the hierarchical algorithms known as Average Link and Complete Link presented gives the best results. In despite their usually high computational costs, we have shown that they are particularly suitable for the studied the application domain because the dendrograms. As already observed in other application domains, dendrograms provide very useful information descriptions and visualization capabilities of data clustering structures.

The partitional K-means and K-medoids algorithms also give the good results when properly initialized. Considering the approaches for estimating the number of clusters, the relative validity criterion known as silhouette has shown to simplified version. In addition, some of our results suggest that using the data file names along with the document content information may be useful for clustering algorithms. Most importantly, we observed that clustering algorithms 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 five real-world applications shows that it has the potential to speed up the computer inspection process. Aimed at further leveraging the use of data clustering algorithms in similar applications, a promising for future work involve to investigate automatic approaches for cluster labelling.

Fig 2: k means algorithm flow chart



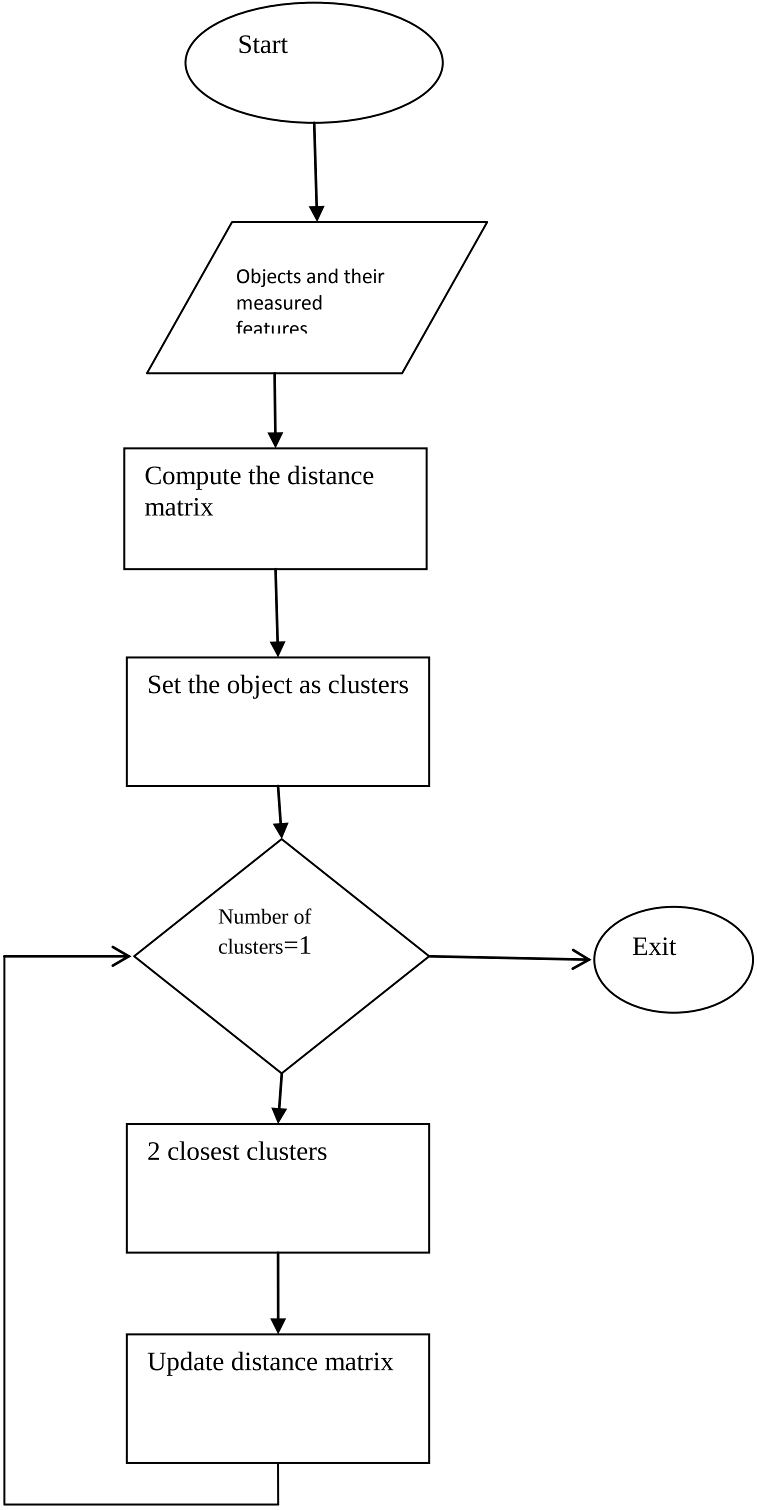

Start
Objects and their measured features
Compute the distance matrix
Set the object as clusters
Number of clusters=1
Exit
2 closest clusters
Update distance matrix

Fig 4: hierarchal cluster algorithm flow chart

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] 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.

[3] 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.

[4] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[5] 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.

[6] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," Inf. Process. Manage. vol. 24, no.5, pp. 513–523, 1988.

[7] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," Soviet Physics Doklady, vol. 10, pp. 707–710, 1966.

[8] Y. Zhao,G. Karypis, and U. M. Fayyad, "Hierarchical clustering algorithms for document datasets," Data Min. Knowl. Discov., vol. 10, no. 2, pp. 141–168, 2005.

[9] M. Widjaja and S.Hansun, "implementation of porter's modified stemming algorithm in an Indonesian word error detection plugin application", International Journal of Technology: 139-150, 2015.

[10] T. Kanungo, D. M. Mount, and N. S. Netanyahu ,"An efficient k-means clustering algorithm: analysis and implementation", IEEE transaction pattern anly.mach.,vol 24,no7,2002.

[11] Adam Coates and Andrew Y. Ng," Learning feature representations with k-means" Originally published in: Tricks of the Trade, 2nd edn, Springer LNCS 7700, 2012.

[12] C Murugananthi and D Ramyachitra, "Performance evaluation of partition and hierarchical clustering algorithms for protein sequence" International Journal of Computational Intelligence and Informatics, Vol. 3: No. 4, January - March 2014.

[13] M-J. Lesot* and M. Rifqi* ,"Similarity measure for binary and numerical data's survey", Int. J. Knowledge Engineering and Soft Data Paradigms, Vol. 1, No. 1, 2009.

[14] Sergio Decherchi, Simone Tacconi, "Text clustering for digital forensics analaysis" Dept. Biophy. and Elect. Engin., University of Genoa,16145 Genoa, Italy.

[15] Charu C. Aggarwal and Cheng Xiang Zhai ,"A survey of text clustering algorithms",

[16] David Dubin,"The most influential paper gerard saltonnever wrote", Daniel Street, Urbana, I61820-6211LIBRARY TRENDS, Vol. 52, No. 4, Spring 2004.

DOI:10.17605/OSF.IO/D3FR6