www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 2 Feb. 2017, Page No. 20285-20288
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i2.23

An Improved Document Clustering with Multiview Point Similarity /Dissimilarity measures

Monika Raghuvanshi[1], Rahul Patel[2]
[1]Acropolise Institute Of Technology and Research
Indore Madhya Pradesh
Erraghuvanshi17@email.com
[2] Acropolise Institute Of Technology and Research
Indore Madhya Pradesh
rahulpatelcs@gmail.com

ABSTRACT*: Clustering is a technique of an unsupervised learning aimed at grouping a set of objects into a clusters, each cluster consist of objects that are similar to one another within the same clusters and are dissimilar to objects belonging to other cluster. The similarity between a pair of objects can be defined either explicitly or implicitly All clustering methods have to assume some cluster relationship among the data objects that they are applied on. Similarity between a pair of objects can be defined either explicitly or implicitly. In this we introduce a novel multiviewpoint-based similarity measure and two related clustering methods.*
The major difference between a traditional dissimilarity/similarity measure and ours is that the former uses only a single viewpoint, which is the origin, while the latter utilizes many different viewpoints, which are objects, assumed to not be in the same cluster with the two objects being measured. Using multiple viewpoints, more informative assessment of similarity could be achieved. Theoretical analysis and empirical study are conducted to support this claim. Two criterion functions for document clustering are proposed based on this new measure. We compare them with well-known k-mean clustering algorithms that uses a Euclidean distance measures on various document collections to verify the advantages of our proposal

KEYWORDS: *Document clustering, Clustering algorithm, Unstructured, similarity measures, multiview point similarity.*

I INTRODUCTION

Data clustering which is the unsupervised classification is one of the important techniques of data mining, where similar data objects are groups into clusters so that data in each clusters share some common characteristics according to defined distance measures. Document clustering is the process of organizing a collection of text documents into clusters based on some similarity measures. Document within a same clusters are more similar to each other than those document belong to a different clusters.

The similarity between a pair of objects can be defined either explicitly or implicitly All clustering methods have to assume some cluster relationship among the data objects that they are applied on. Similarity between a pair of objects can be defined either explicitly or implicitly. In this we introduce a novel multiviewpoint-based similarity measure and two related clustering methods.

The major difference between a traditional dissimilarity/similarity measure and ours is that the former uses only a single viewpoint, which is the origin, while the latter utilizes many different viewpoints, which are objects, assumed to not be in the same cluster with the two objects being measured. Using multiple viewpoints, more informative assessment of similarity could be achieved. Theoretical analysis and empirical study are conducted to support this claim. Two criterion functions for document clustering are proposed based on this new measure

1.1 SIMILARITY MEASURES

Accurate clustering requires a precise definition of the closeness between a pair of objects, in terms of either the pair wise similarity or distance. A variety of similarity or distance measures have been proposed and widely applied, such as cosine similarity, Jaccard coefficient, Euclidean distance and Pearson Correlation Coefficient.

1.2 DOCUMENT CLUSTERING

Document clustering provides an effective, automatic platform to support the analysis of digital textual evidence, which is the key point for forensic analysis process. The process of grouping a set of physical or abstract object into class of similar object is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the object in other cluster.

II LITERATURE SURVEY

There are only a few studies reporting the use of clustering algorithms in the Computer Forensics field. Essentially, most of the studies describe the use of classic algorithms for clustering data—e.g., Expectation-Maximization (EM) for unsupervised learning of Gaussian Mixture Models, K-means, Fuzzy C-means (FCM), and Self-Organizing Maps (SOM). These algorithms have well-known properties and are widely used in practice. For instance, K-means and FCM can be seen as particular cases of EM [2]. Algorithms like SOM [3], in their turn, generally have inductive biases similar to K-means, but are usually less computationally efficient.

In [4], SOM-based algorithms were used for clustering files with the aim of making the decision-making process performed by the examiners more efficient. The files were clustered by taking into account their creation dates/times and their extensions. This kind of algorithm has also been used in [5] in order to cluster the results from keyword searches.

An integrated environment for mining e-mails for forensic analysis, using classification and clustering algorithms, was presented in [6]. In a related application domain, e-mails are grouped by using lexical, syntactic, structural, and domain-specific features [7]. Three clustering algorithms (K-means,

Bisecting K-means and EM) were used. The problem of clustering e-mails for forensic analysis was also addressed in [8], where a Kernel-based variant of K-means was applied. The obtained results were analyzed subjectively, and the authors concluded that they are interesting and useful from an investigation perspective. More recently [9], a FCM-based method for mining association rules from forensic data was described.

The literature on Computer Forensics only reports the use of algorithms that assume that the number of clusters is known and fixed *a* priori by 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. Essentially, one induces different data partitions (with different numbers of clusters) and then assesses them with a relative validity index in order to estimate the best value for the number of clusters [2], [3], [10]. This work makes use of such methods, thus potentially facilitating the work of the expert examiner—who in practice would hardly know the number of clusters apriori.

2 DOCUMENT CLUSTERING PROCESS

Clustering is the most common form of unsupervised learning which deals with finding a structure in a collection of unlabeled data. .Clustering of document is an automatic grouping of text document within a cluster have a high resemblance in comparison to one another ,but are different from document in other clusters. It is important to emphasize that getting from a collection of document to a clustering of the collection is not merely a single process , but is more a process in multiple stage

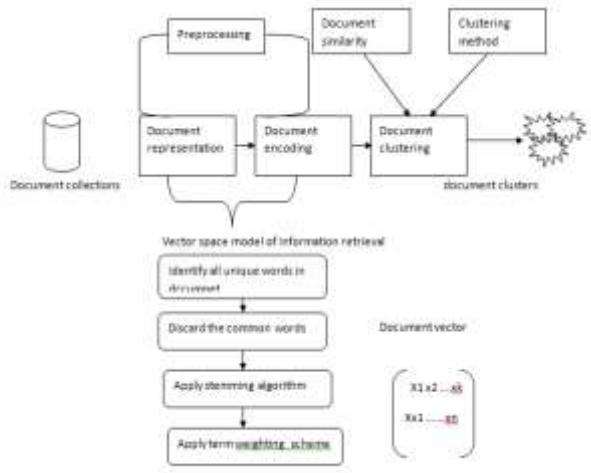


Figure1: Document clustering process

2.1 COLLECTION OF DATA:

Methods like crawling, indexing and filtering etc which are used to collect the documents that needs to be clustered.

2.2 Preprocessing Steps:

Preprocessing is done to represent the data in a form that can be used for clustering

2.2.1 Stemming :

Stemming is a technique for the reduction of words into their steam or base form many words e.g. agreed, agreeing, disagrees, agreement, and disagreement belong to agree.

2.2.2 Stop word Removal

Prepositions, articles, and pronouns etc are the most common words in any text document does not provide meaning of the document. These words are eliminated. These words are not necessary for text mining application.

2.2.3 Term Frequency

The simplest possible method for feature selection in document clustering is document frequency that is used to filter out irrelevant feature. In other word, words which are too frequent in the corpus can be removed because they are

2.2.4 Tokenization

Splits sentences into separates tokens, the main use of tokenization is to identifying meaningful keyword

2.3 CLUSTERING ALGORITHM

The clustering algorithm is used in the process of digital forensic analysis. These methods are basically used to convert unstructured document to structured document for further investigation. In this work we used a different clustering algorithm as follows.

2.3.1 *K-Means*

K-means is the most important flat clustering algorithm. The objective function of K-means is to minimize the average squared distance of objects from their cluster centers, where a cluster center is defined as the mean or centroid μ of the objects in a cluster C:

The ideal cluster in K-means is a sphere with the centroid as its center of gravity. Ideally, the clusters should not overlap. A measure of how well the centroids represent the members of their clusters is the Residual Sum of Squares (RSS), the squared distance of each vector from its centroid summed over all vectors

K-means can start with selecting as initial clusters centers K randomly chosen objects, namely the seeds. It then moves the cluster centers around in space in order to minimize RSS. This is done iteratively by repeating two steps until a stopping criterion is met

1. Reassigning objects to the cluster with closest centroid
2. Recomputing each centroid based on the current members of its cluster.

We can use one of the following termination conditions as stopping criterion

- A fixed number of iterations I has been completed.
- Centroids μ do not change between iterations.
- Terminate when RSS falls below a pre-estabilished threshold.

2.3.2 Modified K-mean Alogrithm

Improved k-mean method is used.

Steps of improved method:

INPUT: D = {do1, do2, Do3,..., doi,..., don }
set of documents di = { x1, x2, x3,..., xi,..., xm }
k - Number of desired clusters.
OUTPUT: k clusters sets
Step 1. Select k=2 initial cluster centers Ci randomly from data Xi.

Repeat following steps for every cluster center.

Step 2. Find Euclidean distance of each data objects Xi from cluster centers and assign objects to cluster with minimum distance.

Step 3. Find Min_dist and Max_dist distance along with corresponding nearest object min_obj and farthest object max_obj.

Step 4. Calculate two sets of objects NPT and MPT contain densely connected objects to min_obj and max_obj within distance:

$$avg\_dist = (Min\_dist + Max\_dist)/3$$

Step 5. Selecting K

i)$NPTi \cap MPTi = \Phi$

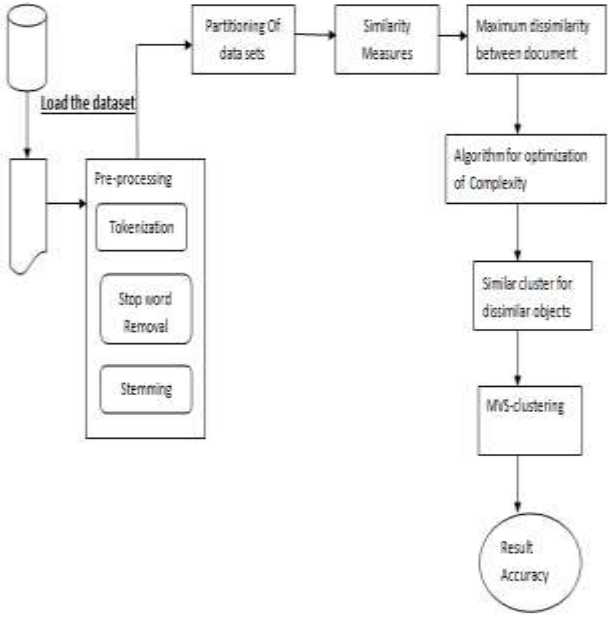
ii) $NPTi \cap NPTj = \Phi$ and $MPTi \cap MPTj = \Phi$

If (i) valid then split Ci and if both (i) and (ii) valid split both center and assign new center as min_obj and max_obj of corresponding cluster. If either condition is valid then goto step 2.

Step 6. Find mean for every cluster.

Step 7. If no change in cluster centers then exit.

The above Modified k-means algorithm has additional steps in traditional k-means algorithm for better cluster center selection. We use Euclidean distance for assigning object to proper cluster by using these calculated distances and we find nearest min_obj and farthest max_obj objects from cluster center and record its minimum Min_dist and maximum Max_dist distance values. For selecting better cluster centers we use two sets of densely connected objects. The NPT set contain objects within avg_dist from min_obj and MPT set contain objects within avg_dist distnace from max_obj.



III RESULT

It shows result of the test performed over different dataset basically we check purity of two method one is the K-mean and other is incremental clustering technique.

A Evaluation Measurements

The quality of resulting clusters was evaluated using two measures –purity and entropy, which are widely used to evaluate the performance of clustering.

a) Purity:

The purity indicates the coherence of cluster that is the degree to which a single category represents a cluster contains documents belonging to that category. More precisely purity can be defined as the classification rate that all sample of the cluster are predicated to be member of the actual dominates class for the cluster. For ideal cluster .i.e. where purity value is which only contain documents from a single category. In general, the quality of the cluster is depends on the purity value , means higher the purity value the better the quality of cluster is. In clustering purity is an important measure to evaluate the overall quality of clustering solution, higher the purity better the quality of cluster.

b) E**ntropy:**

The disorder of objects within the cluster is often express in term of entropy with theoretic-informatics terms. for each cluster , the distribution of data is calculated ,i.e. the probability that a member of cluster k belong to category i pik;

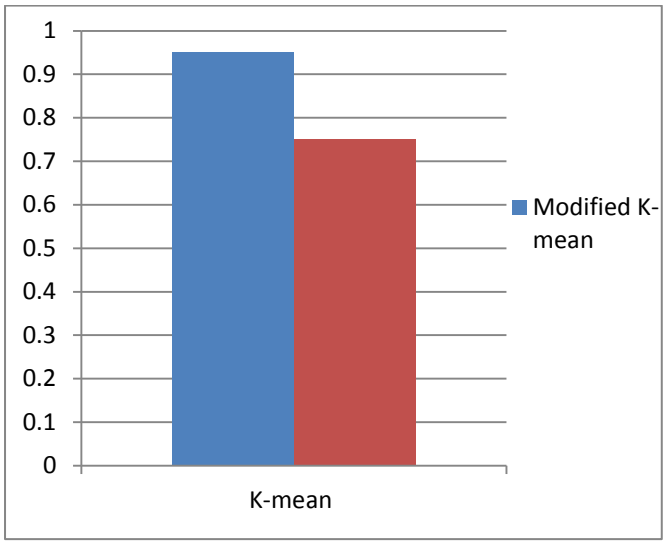
the entropy of cluster k is calculated as as ;

$$E_K = -\sum_i P_{ik} \log(P_{ik})$$

The total entropy of all clusters is calculated as the sum of entropy of all clusters;

$$E_{en} = \frac{\sum_{k=1}^{n} m_k E_k}{n}$$

where n is the total no of clusters $m_k$is the size of kth clusters and m is the total no of documents.



IV CONCLUSION

Document clustering play an important role in selecting documents among thousands documents. In this research work we have studied the updating to k-mean clustering. K-mean clustering algorithm is very popular and simple algorithm, but it has some drawback. There have been various algorithm which aims to improve the k-mean algorithm and work around the limitations of the k-mean algorithm. The performance of k-mean algorithm depends on initial cluster centre selection. The challenging issue in k-mean is selection of appropriate value of k and cluster centre. The proposed research work can choose better value of k by selecting high dense objects as cluster centre and by splitting. so that they can provide a efficient clustering for k mean algorithm.

The simplicity of k-men algorithm makes it choice

for many clustering applications. However k-mean algorithm for document clustering suffers the problem of initializations, dead point problem, and the predetermined number of cluster k. we introduced a novel method for initializations that well to find appropriate initial centre for k mean. The experiments results shows the advantage of proposed algorithm which are high clustering accuracy and purity of clustering.

V FUTURE SCOPE

The algorithms proposed in this thesis are at their rudimentary stage and there are many possible improvements that can be implemented, aimed at further leveraging the use of data clustering algorithms in similar applications, a promising for future work involves 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. In future, the experiment can be conducted for volume of large datasets. New clustering algorithm approach can also be explored to further improve the effectiveness of clustering.

VI REFERENCE

[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] C.M.Bishop,Pattern Recognition and Machine Learning. NewYork: Springer-Verlag, 2006.

[3] S. Haykin, Neural Networks: A Comprehensive Foundation.Englewood Cliffs, NJ: Prentice-Hall, 1998.

[4] 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

[5] 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,

[6] 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.

[7] 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.

[8] 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.

[9] K. Stoffel, P. Cotofrei, and D. Han, "Fuzzy methods for forensic dataanalysis," in *Proc. IEEE Int. Conf. Soft Computing and Pattern Recognition*, 2010, pp. 23–28.

[10] 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.