An Enhanced and Improved Methodology for performing Document Clustering

[1]Aayushi Rathore, [2]Prof. Balwant Prajapat

***Abstract*: Keeping similar documents together is crucial for the purpose of document organization, summarization, topic extraction and information retrieval in an efficient way. Initially, clustering is applied for enhancing the information retrieval techniques. Of late, clustering techniques have been applied in the areas which involve browsing the gathered data or in categorizing the outcome provided by the search engines for the reply to the query raised by the users. In this paper, we are providing a methodology for more accurate document clustering. The experimental results have shown that the clustering accuracy cum purity of the proposed technique is better than that of the present clustering technique.**

***Keywords*: Document clustering, preprocessing, improved clustering method, term frequency, Euclidian distance**

1. Introduction:

Due to advancement in technology, it is estimated that the digital data density increased 18 times in latest 5 to 6 year[1,2,4]. This increased data has direct impact in computer forensic. in general computer forensic is the application of investigation and analysis techniques in which evidence are collected from a particular computing device in a manner that is proper for presentation in a court and according to the law. Document clustering has shown to be very useful for computer forensic analysis.

Computer forensic [3,5] analysis is a branch of forensic science encompassing the investigation of material found in digital device in a way that is proper for presentation in a court and according to the law. Document analysis in a computer device is a key task of the computer forensic investigation process. But this task may be daunting due to large no of document usually stored on a hard disk. The clustering algorithm are used in the process of computer forensic analysis .these methods are basically used to covert unstructured documents to structured documents for further investigation.

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.

2. Methodology for Document Clustering:

Clustering [8,9,10] 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.

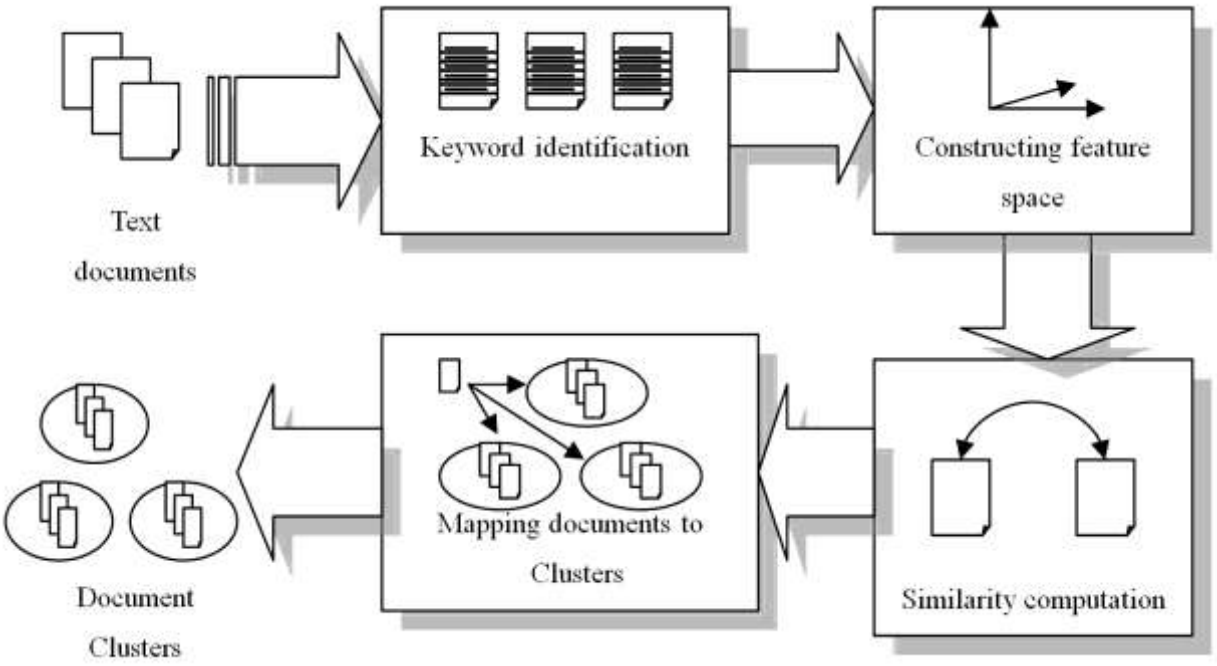

Figure1: Document clustering process

- **Collection of Data :**

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

- **Preprocessing Steps:**

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

- **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.

- **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.

- **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

- **Tokenization**

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

- **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.

- **K-Means**

K means is used in the most of the present frameworks.

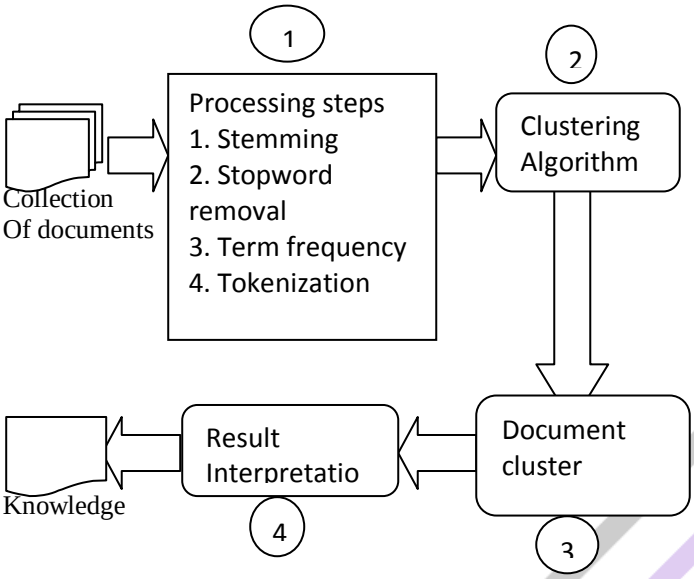

Figure2: Document clustering process

3. Steps of proposed and improved Clustering Technique

Output: D = {d1, d2, d3,..., di,..., dn } //set of documents di = { x1, x2, x3,..., xi,..., xm } k // Number of desired clusters.

Input: A set of k clusters.

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 ∩ MPTi=Φ ii) NPTi ∩NPTj=Φ and MPTi ∩MPTj=Φ 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.

4. Conclusion:

Most importantly, we observed that clustering algorithms 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 five real-world applications show that it has the potential to speed up the computer inspection process. As clustering plays a very vital role in various applications, many researches are still being done. The upcoming innovations are mainly due to the properties and the characteristics of existing methods. This paper presents an introduction to the present document clustering concept along with the methods used for document clustering. An updated clustering technique is also discussed in detail along with the model.

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. 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.