A REVIEW OF NOVEL AND EFFECTIVE APPROACH FOR CLUSTERING DOCUMENTS

Yashashvi Hirwale[1], Nisha Bhalse[2]
[1]Student, [2]Assistant Professor
Department Of Computer Science & Engineering
IES-IPS Academy, Indore
[1]yashashvi.hirwale1@gmail.com
[2]nishabhalse@gmail.com

Abstract— **In the computerized inspection of text documents usually substantial amount of files are explored every day. Much of the data contained in those files consists of disorganized text, whose examination is difficult to be achieved by computer research workers. In this situation, computerized methods of search are of great significance to the analyst. In particular, algorithms for clustering of documents can be a start to the expedition of new and beneficial knowledge from the records under search. Here we will represent an approach that is useful for recommendations of readers of a news portal. We will illustrate the suggested approach by performing extensive tests on datasets with the renowned k-means clustering algorithm. Experiments will be performed on the set of data which is ready for use and distinct specifications of the database are analyzed for achieving the defined objective. In addition, two relative validity indexes were used to naturally find out the number of clusters. The objective of this work is to increase the efficiency and preciseness of existing clustering algorithm. At last, we will also represent several drawbacks of some classical clustering techniques that can be valuable for researchers and masters of text mining.**

Index terms- **Document clustering,, k-means algorithm, text mining.**

I. INTRODUCTION

Due to the increasing requirement and frequent use of storage houses at all places it is becoming very problematic to organize and take care of the data contained in these repositories effortlessly and adequately for the desire of achieving departmental benefits. It is predicted that the volume of electronic data has increased from 165 hexabytes in 2007 to 988 hexabytes in 2010 about 18 times the amount of information present in the records and it still continues to increase exponentially. This large amount of data has a direct impact on *Computer Analysis*, which can be broadly defined as the discipline of computer science to collect and analyze data from computer systems in a way that can be helpful for the domain experts to perform the examination in limited amount of time. The disciplines in which data is directly used are influenced the most. In our research area of expertise, analysis process involves examining huge data files per day by the researchers. This activity surpasses the master's ability of analysis and interpretation of data. Therefore, computerized methods are of great concern like the one used for learning of machines and extraction of data. In some specific applications, algorithms for pattern recognition from the facts present in text files are promising, as it will hopefully become apparent later in the paper.

Clustering algorithms are usually used for exploratory analysis of data. Clustering is usually performed on the data where there is no or prior knowledge about the type of data .This is exactly the case found in various applications of Computer analysis, including the one addressed in our work. From a technological view, datasets contains unidentified objects and the type of documents that are found are not previously known. Furthermore, assuming that recognized data can be ready for use from previous analyzed process, there is nearly no hope that the same class or type (as previously identified by a classifier) would be still right for the forthcoming data, acquired from computers and related to different thorough check processes. More exactly, it is likely to happen that the fresh data sample would arrive from a unlike sources. In this circumstance, the employment of clustering algorithms, which are capable of finding hidden motifs from text documents found in computers, can embellish the examination performed by the knowledgeable tester.

The logic for belief trailing clustering techniques is that objects within the same cluster are very much alike to one another than they are to objects belonging to a dissimilar cluster. The expert examiner may first closely think on criticizing the ideal documents from the obtained set of clusters once the data partition has been induced from the data. After this introductory analysis, we may ultimately determine to closely analysis other documents from other clusters. One can stay away from the tough task of inspecting all the documents by doing the introductory analysis (individually but, even if it is required, it can still be done.)

In a more feasible and functional scenario, domain experts are limited and have defined amount of time for performing examinations. Thus, it is feasible to consider that, after finding an appropriate document the investigator could prioritize the analysis of other documents belonging to the cluster of interest.

The evaluation of the number of clusters automatically has not been investigated in the *Computer Analysis* literature and it is also a critical and detracting specification in many algorithms which is usually *a priori* unknown. Practically , we could not report one work that is capable of estimating the number of clusters with the use of algorithms and that is reasonably close in its application domain .Perhaps even more surprising is the lack of studies on hierarchical clustering algorithms, which date back to the sixties. Our study considers such classical algorithms, as well as recent developments in clustering.

The remainder of this paper is organized as follows. Section II presents the system model. Section III briefly addresses the adopted clustering algorithms and the work that has been performed previously by the expert's which has given us motivation to accomplish our research on this topic. Section IV presents the proposed methodology and section V reports our experimental results. At last Section VI concludes the paper and Section VII addresses the future work.

II. SYSTEM MODEL

The objective of the proposed model is to develop a more powerful and adoptive clustering algorithm in comparison to the classical K-means method. The clustering accuracy of the new method will be increased as compared to the existing method. The proposed system can be explained as follows:

CLUSTERING AND PREPROCESSING STEPS:

A. Preprocessing Steps:

In order to remove the stop words Preprocessing steps are employed. Before employing the clustering algorithm on set of data preprocessing steps are performed. For mining the text the snowball stemming algorithm and a classical statistical technique is adopted which will represent documents in vector space model and each document is represented by a vector containing the frequencies of occurrences of words. To enhance the performance and proficiency of clustering algorithms Term Variance (TV-a dimensionality reduction technique) will also be used. The number of attributes that have the greatest variances over the documents is selected by TV. Two measures cosine-based distance and Levenshtein-based distance are adopted for computing the distance between documents. Levenshtein-based distance calculates distances between file names only.

B. Estimation Of Number Of Clusters:

A conventional approach for estimating the number of clusters consists of selection of the best result from several results according to an explicit quality criterion (e.g., a relative validity index). For selecting the best result a set of data partitions from different number of clusters are collected and then an appropriate partition that yields the best result are evaluated. The data partitions are resultant of the hierarchical clustering dendrogram or from numerous runs of a K-means algorithm starting with different numbers and initial positions of the cluster prototypes. By choosing the set of data partitions on the basis of some relative validity criterion we are performing model selection which is an important step of this process and the number of clusters is also estimated. A conventional relative validity index called *silhouette,* has been adopted as a component of the algorithm. The average of the silhouette depends on the computation of all distances among all objects. A more computationally efficient criterion, called *simplified*

Silhouette, can also be used and one has to compute only the distances among the objects and the centroids of the clusters.

C. Clustering Technique Adopted:

The k –means clustering algorithm has been adopted in our study. Mainly popular in the machine learning and data mining fields, and therefore it is adopted in our study. K-Means is an unsupervised learning algorithm and one of the classical techniques adopted for text mining. K-means is also a renowned clustering algorithm that is employed in many applications to solve the clustering problem. The procedure followed to determine the number of clusters is simple. The algorithm works on the fact that each point in a cluster should be near to the center of that cluster. Algorithmic steps for K-Means is shown as below

Let $X = \{x_1,x_2,x_3,\ldots\ldots..,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots.,v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1 / c_i) \sum_{j=1}^{c_i} x_i$$

where, *'c_i'* represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

D. Removal of Outliers:

The recursive use of silhouette is one of the simple and effective approaches that lead to the removal of outliers. Basically, if the best partition chosen by the silhouette has singletons (i.e. clusters composed by a single object only), these are eliminated. Then, the clustering process is repeated-until a partition without singletons is found. At the end of the process, all singletons are integrated into the resulting data partition (for evaluation purposes) as single clusters. The proposed system model can be explained diagrammatically as:

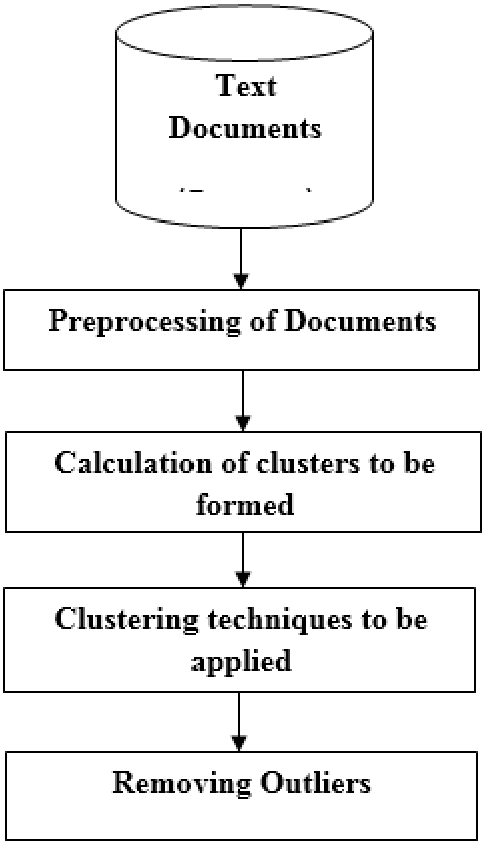
Text Documents
Preprocessing of Documents
Calculation of clusters to be formed
Clustering techniques to be applied
Removing Outliers

Figure 1: diagram of the proposed system.

III. PREVIOUS WORK

A lot of work has been done previously in the field of computer analysis. In order to get the desired objective the use of clustering algorithms are done. Most of the studies illustrate the application of classical clustering algorithms for clustering documents e.g., K-means, Fuzzy C-means (FCM). Expectation-Maximization (EM) for unsupervised learning of Gaussian Mixture Models.

Agrawal *et al.* [1] has depicted about data mining applications and different methods of clustering documents. The objective of their work is to identify the clustering ability of the algorithms for identifying the clusters embedded in subspaces. These subspaces consist of high dimensional data and scalability. HPSO (Hybrid Particle Swarm Optimization) is a new and innovative clustering technique addressed in [2] which combines features of partitional and hierarchical clustering techniques and proved to be very efficient and powerful for performing hierarchical clustering. It employs the swarm intelligence of ants in a decentralized environment. C.Aggarwal *et al* [4] has described in his paper that ambiguous and unlike data segments are found in most of the documents but the results of the existing approaches in terms of quality of clustering are not satisfactory. In [5] an approach that automatically extracts data from large data-intensive web sites is presented. To extract valuable data a number of websites are examined and a scheme is inferenced which describes it as a directed graph with nodes. It describes classes of structurally similar pages and arcs representing links between these pages. After locating the classes of interest, a library of wrappers can be created, one per class with the help of an external wrapper generator. In [7] a technique for clustering ambiguous data streams has been described. A frequency histogram is used to trace the characteristic of categorical statistic. This technique initially, creates ‘n’ clusters by applying K-prototype algorithm. The new approach proves to be more useful than U-Micro in respect to clustering quality. L.Taoying *et al.* [8] had made use of cluster ensemble to develop an incremental clustering technique for categorical data. They reduced the use of unnecessary attributes and emphasized on the use of true values to form cluster membership. In [9] the global K-means clustering technique for creating clusters of the documents has been proposed. It creates cluster by making the use of K-dimensional tree approach. In [12] modified K-means algorithm has been proposed that solves the empty cluster problem. This modified K-means algorithm had produced effective result and experiments had proved this method better than the traditional clustering techniques. In [14] software mining mission with an integration of text mining and link study technique is presented. This technique is concerned with the inter links between instances. LATINO is an ontology-learning framework developed by Grcar *et al.* [6]. It is an open source platform, offering text mining, link analysis, machine learning, etc. S.Khan and A.Ahmad [17] in their work has proposed an iterative clustering technique. This technique is feasible for clustering techniques for continuous data. S.Lee and X.Zeng [18] suggested clustering-based method to identify the fuzzy system. It tried to present a modular approach, based on hybrid clustering technique.

IV. PROPOSED METHODOLOGY

One of the challenging issues is to perform clustering of similar documents efficiently as the amount of documents and the need of organizing these documents from the huge and multiple store houses has great demand and need in IT field. Efficiency and similarity are the most important and essential feature of document clustering algorithms. One of the essential part of document clustering is the measurement of similarity between two patterns. In news articles we first tried to use the synonyms of words on the place of same word to calculate the similarity between documents but it’s not the synonymy but the co-occurrence of words which plays importance role for calculating the similarity. For example the words like ‘Narendra Modi’ and ‘India’ are not synonyms but the news articles containing these words will belong to same clusters. This gave us the motivation to first cluster the words based on their co-occurrence in the given dataset and then use this cluster to cluster the documents.

We are developing an application for recommendations of readers of online news of news portal. By clustering the articles we could reduce our domain of search as most of the users had interest in the news related to their own topic of interest. This will improve time efficiency to a great extent and we can easily identify the articles related to same topics from different sources. The aim of this work is to improve the effectiveness and efficiency of document clustering by reducing the noise in data by pre-processing the data and identifying the drawbacks of k-means clustering algorithm to get their solutions. We will first study the effectiveness of pre-processing step and then apply proposed clustering algorithm. We will then detect the effectiveness of a new clustering algorithm in which the noise is reduced.

V. EXPERIMENTAL RESULT

A. DATASET

Sets of documents that appear in computer forensic analysis applications are quite diversified. The experiments are performed on dataset that are freely available. The data partitions were evaluated by taking into account that we have a *reference partition* (ground truth) for every dataset. Such reference partitions have been provided by an expert examiner. The datasets contain varying amounts of documents (N), groups (K), attributes (D), singletons(S), and Number of documents per group (#).

The experiment can be performed by making the use of datasets described below:

1. 20 newsgroups: This is a very standard and popular dataset used for evaluation of many text applications, data mining methods, machine learning methods, etc.

2. Reuters: This is the most common dataset used for evaluation of document categorization and clustering

3. Keepmedia newsgroups: This is a set of news articles provided by a company.

The similarity (or dissimilarity) between the documents is typically computed based on the distance between document pairs. The similarity measure Euclidean Distance for k mean is given as:

$$D_E(\overrightarrow{t_a}, \overrightarrow{t_b}) = (\sum_{t=1}^{m} |w_{t,a} - w_{t,b}|^2)^{1/2}$$

- DE = Distance between vectors
- wt,a and wt,b are weights as given by tfidf values i.e wt,a= tfidf(da, t)
- **tfidf(d,t) = tf(d, t) * Log(|D| / df(t))**
- |D| = number of documents
- df (t) = number of documents in which 't' appears.

B. EVALUATION MEASURES

For evaluating the clustering algorithms the use of reference partitions is considered as a principal approach. Reference partitions are usually obtained to choose a particular clustering algorithm that is more appropriate for a given application, or to calibrate its parameters. In our study, reference partitions were obtained by expert examiners and the expectation about number of clusters that should be found in the dataset are reflected. The evaluation method is based on the adjusted Rank Index which we use to access the obtained data partition. Adjusted Rank Index measures the agreement between a partition P, obtained from running a clustering algorithm, and the reference partition R given by the expert examiner.

VI. CONCLUSION

Clustering has been adopted as an important technique in various data mining applications. Clustering is considered as the main step in text analysis. We concluded that the K-means algorithm yields good results when properly initialized. We also emphasized that clustering algorithms will reduce the number of clusters formed by relevant or irrelevant documents, thus contributing to enhance the expert examiner's job. Considering the approaches for estimating the number of clusters, the relative validity criterion known as *silhouette* has shown to be more accurate than its (more computationally efficient) simplified version.

VII. FUTURE SCOPE:

Aimed at further leveraging the use of data clustering algorithms in similar applications, a promising venue 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. Finally, the study of algorithms that induce overlapping partitions (e.g., Fuzzy C-Means and Expectation-Maximization for Gaussian Mixture Models) is worth of investigation.

REFERENCES

1. Agrawal, Rakesh, Gehrke, Johannes, Gunopulos, Dimitrios, Raghavan and Prabhakar, "Automatic subspace clustering of high dimensional data", Data Mining and Knowledge Discovery (Springer Netherlands) Vol. 11, pp. 5-33, DOI:10.1007/s10618-005-1396-1, 2005.

2. Alam, S., Dobbie, G., Riddle, P. and Naeem, M.A. "Particle Swarm Optimization Based Hierarchical Agglomerative Clustering", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol. 2, pp. 64-68, 2010.

3. Baeza-Yates, R.A. "Introduction to Data Structures and Algorithms Related to Information Retrieval", In Information Retrieval: Data Structures and Algorithms, W. B. Frakes and R. Baeza-Yates, Eds. Prentice- Hall, Inc., Upper Saddle River, New Jersey, pp. 13-27, 1992.

4. Charu C. Aggarwal, Jiawei Han, Jianyong Wang and Philip S. Yu, "A Framework for Clustering Evolving Data Streams", Proceedings of the 29th international conference on Very Large Data Bases (VLDB), pp. 81-92, 2003.

5. Crescenzi valter, Giansalvatore Mecca, Paolo Merialdo and Paolo Missier, "An Automatic Data Grabber for Large Web Sites", VLDB , pp. 1321-1324, 2004
6. Grcar, M., Mladenic, D., Grobelnik, M., Fortuna, B. and Brank, J. "Ontology Learning Implementation", Project report IST-2004-026460 TAO, WP 2, D2.2, 2006.
7. Guo-Yan Huang, Da-Peng Liang, Chang-Zhen Hu and Jia-Dong Ren, "An algorithm for clustering heterogeneous data streams with uncertainty", 2010 International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 4, pp. 2059-2064, 2010.
8. Li Taoying, Chne Yan, Qu Lili and Mu Xiangwei, "Incremental clustering for categorical data using clustering ensemble", 29th Chinese Control Conference (CCC), pp. 2519-2524, 2010.
9. Likas, A., Vlassis, N. and Verbeek, J.J. "The Global k-means Clustering algorithm", Pattern Recognition , Vol. 36, No. 2, pp. 451-461, 2003.
10. Lijuan Jiao and Liping Feng, "Text Classification Based on Ant Colony Optimization", Third International Conference on Information and Computing (ICIC), Vol. 3, pp.229 - 232, 2010.
11. Macskassy, S.A., Banerjee, A. Davison, B.D. and Hirsh, H. "Human Performance On Clustering Web Pages: A Preliminary Study", In Proc. of KDD-1998, New York, USA, pp. 264-268, Menlo Park, CA, USA, 1998.
12. Malay K. Pakhira, "A Modified k-means Algorithm to Avoid Empty", International Journal of Recent Trends in Engineering, Vol. 1, No. 1, pp. 220-226, 2009.
13. Meila, M. and Heckerman, D. "An experimental comparison of model-based clustering methods", Machine Learning, kluwer Academic publishers, Vol. 42, pp. 9-29, 2001.
14. Miha Grcar, Marko Grobelnik and Dunja Mladenic, "Using Text Mining and Link Analysis for Software Mining", Lecture Notes in Computer Science, Vol. 4944, pp. 1-12, 2008.
15. Murtagh, F. "A Survey of Recent Advances in Hierarchical Clustering Algorithms Which Use Cluster Centers", Comput. J, Vol. 26, pp. 354-359, 1984
16. Pallav Roxy and Durga Toshniwal, "Clustering Unstructured Text Documents Using Fading Function", International Journal of Information and Mathematical Sciences, Vol. 5, No. 3, pp. 149-156, 2009
17. Shehroz S. Khan and Amir Ahmad, "Cluster Center Initialization Algorithm for K-means Clustering", Pattern Recognition Letters, Vol. 25, No. 11, pp. 1293-1302, 2004.
18. Shin-Jye Lee and Xiao-Jun Zeng, "A three-part input-output clustering-based approach to fuzzy system identification", 2010 10th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 55-60, 2010.
19. Ward Jr, J.H. "Hierarchical grouping to optimize an objective function", J. Am. Stat. Association, Vol. 58, pp. 236-244, 1963.