HIERARCHICAL DOCUMENT CLUSTERING USING CLOSED ITEMSETS

KAVITA NAGAR

ABSTRACT: As we know that a large amount of information is generated and stored in text databases. As the numbers of document files increases we need a robust and efficient way to group this large amount of data. Clustering is the finest tool or technique of data mining for managing and organizing information . Clustering maps the similar objects or data into one cluster and different objects into another one based on their inter similarity and intra dissimilarity. However most of the clustering techniques faces many issues like high dimensionality, scalability, accuracy, etc. Document clustering is the another unsupervised clustering technique for document organization, fast information retrival or filtering. This paper will present a review on some document clustering methods and proposal of a new one approach for hierarchical document clustering using closed itemsets.

Keywords : Document clustering, hierarchical clustering, frequent itemsets, closed itemsets, literature review, Similarity, Dissimilarity.

———————————— ◆ ————————————

Student of Master of Technology,
Department of Computer science and Engineering
Utter Pradesh Technical University,
Gr. Noida, India
Kavitanagar87@gmail.com

1. INTRODUCTION

Clustering is an unsupervised learning technique which groups the text data into same cluster based on their similarity basis. Now a days Hierarchical document clustering is widely being used to organize and browse the information on the internet or network. A large number of clustering algorithms are available but most of them suffers from issues like high dimensionality, accuracy, scalability etc. Clustering algorithms mainly described as Hierarchical and partitioning . In hierarchical clustering a tree like structure is generated it can be bottom up manner or top down manner depending on which hierarchical approach like divisive or agglomerative is applied on the dataset. Where as in partitioning method data is divided into several sub datasets until the condition meet the cretiria or threshold. Document clustering is the automatic organization of documents into clusters in which grouping is done on the basis of maximizing intra –cluster similarity and minimizing inter- cluster similarity. Document clustering algorithm is different from classification it is based on unsupervised learning in which we learn by observations rather then by some given examples. The accuracy of the document clustering algorithm is measured by F-measure which is an external evaluation method. F- measure first of all selects of identifies the best cluster among the given natural classes presented in the data set. Then it measures the accuracy of best cluster and calculate the weighted average accuracy of all natural classes.

2. RELATED LITERATURE SURVEY

AS we know thatat clustering performs the grouping of same data or objects based on their similarity and dissimilarity measures. Hierarchical clustering groups the data into a tree like formation or cluster which can be further divided into two parts i.e Divisive and Agglomerative which formed the cluster in splitting and merging fashion. For a particular split or merge if decision is not taken carefully then it can lead to wrong output or the results can not be changed. Where as in partitioning method data set is divided into sub data sets and each data set exactly belong to one one data sets.

In 2002, Beil, Ester and Xu [1] addressed the problem of finding the relevant content or data from the intranet and state that algorithm like bisecting K-Means did not satisfy the requirement of high dimensionality and large data sets size. An algorithm called Hierarchical Frequent Term Based(HFTC) was developed in which frequent item-sets have been used on the association rule mining basis. According to [1] frequent- term only provide the description about the cluster but does not form the cluster. The algorithm proposed by[1] was a greedy algorithm. It has been stated that dynamic programming of HFTC algorithm may be used to solving the frequent term based clustering for future work.

In 2006 Hassan H. Malik , John R, Kender [2] proposed the new method which was a step ahead from the HFTC algorithm they used the sub-linearly scalable notion along with closed interesting item-sets for hierarchical document clustering method. It has been shown that if we use same threshold for the first level of it will results in small numbers of closed interesting item-sets as compared to number of closed frequent item-se t generated.

In 2007, Yanjun Li, And Chung[3] find that the most of the pre-existing text clustering algorithm used the vector space model in which document is treated like a bag of words, but it ignores the order of the words.[37] proposed a new text clustering algoritm called clustering based on Frequent Word Meaning Sequences(CFWMS) . It uses the synonyms and hyponyms provided by the Word Net Ontology for document pre-processing.

In 2008, Yehang Zhu, Fung, Dejun Mu and Yangling[4] proposed a novel hierarchical clustering method which was a hybrid version of partitioning and agglomerative clustering technique. Experiments results show that the proposed method was effective and efficient but accuracy was not so impressevie for some real life large datasets.

In 2009 Xiaoke Su, Yang Lan, Renxia Wan and Yuming Qin[5] suggested a fast incremental hierarchical clustering algorithm which was feasible and effective. Theoretical analysis and experiments results shows that it not only overcome the inadequate impact of memory while clustering large data set but also reflect the accurate features of the data set.

In 2010 M. Sriniwas , C. Krishna Mohan [6] proposed a new hybrid clustering algorithm called Leaders Complete Linkage

algorithm(LCL) which was the combination of the hierarchical and incremental clustering . In this at each iteration objects are grouped into one cluster to another by splitting and merging two clusters. Clustering be started with one cluster containing all the objects.At every step partition quality is checked when splitting nad merging operations are performed on the two clusters. If the resulting quality is good only then the next or other level splitting and merging can be performed Otherwise not . Currerent partition will be the final clustering result.

In 2010 Rekha Baghel, Dr. Renu Dhir [7] proposed a frequent concept document clustering algorithm(FCDC) in which semantic relationship between the words was used to create the concepts. It uses the Hierarchical clustering with Word Net Ontology to create dimensional feature vector which allows to develop efficient clustering algorithm. FCDC found more efficient accurate and scalable when compared with existing algorithms like, FIHC, K-Mean, UPGMA.

In 2013, Ms. Devika Deshmukh, Mr. Sandip kamble[9] proposed an effective Fuzzy Frequent Item-set Based Hierarchical Clustering(F2HC)which was based on the fuzzy association rule mining. Which perform good in term of cluster quality and accuracy. In Ing It perform the clustering in three steps .In the first step it findouts the document and processed them into designated representation for mining. Then in second step it makes the relevant frequent item-sets by predefined membership functions like low, mid, high. And in the final step it document the clusters into hierarchical tree cluster by assigning one document to exactly one cluster.

In 2014 M.S Patil, M.S Bewoore, S.H. Patil[10] proposed an extractive text summarization method which is based on the Support Vector Mechine(SVM). It shows improving results in performance and quality of the summary generated by cascading with SVM. In this method text summarization contains three steps i.e pre-processing which contains the representation of original text or document, then processing steps which converts the text information into the summary and the last step is summary generation which generate the full summary .

(i)
(ii)
(iii)
(iv)
(v)
(vi)

3. PROPOESED METHOD

From the literature survey we found that accuracy and scalability are the two basic algorithm in which each document will be clustered by using frequent closed itemset which occur in a sufficient numbers in the document. Each document will represent the transaction and each word wil be depicted as a closed item set.

The main objective of this method is to design a document clustering method based on the given architecture:-

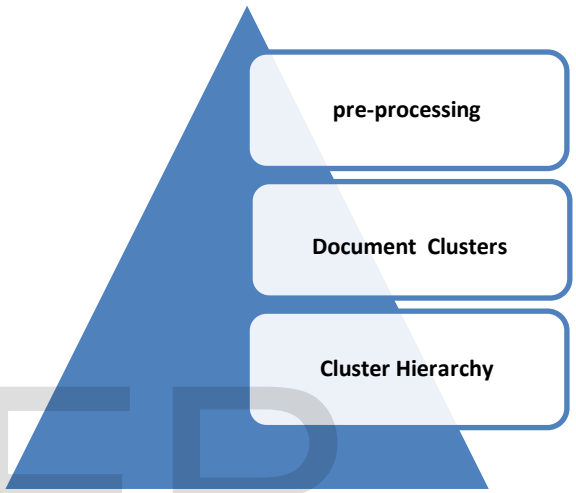
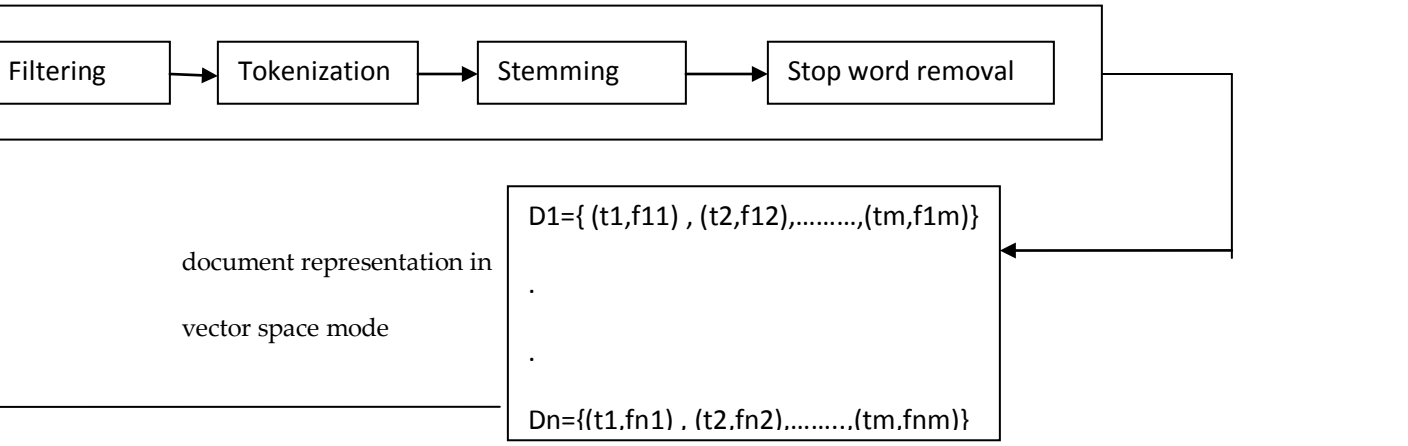


fig 1. overview of document clustering architechture

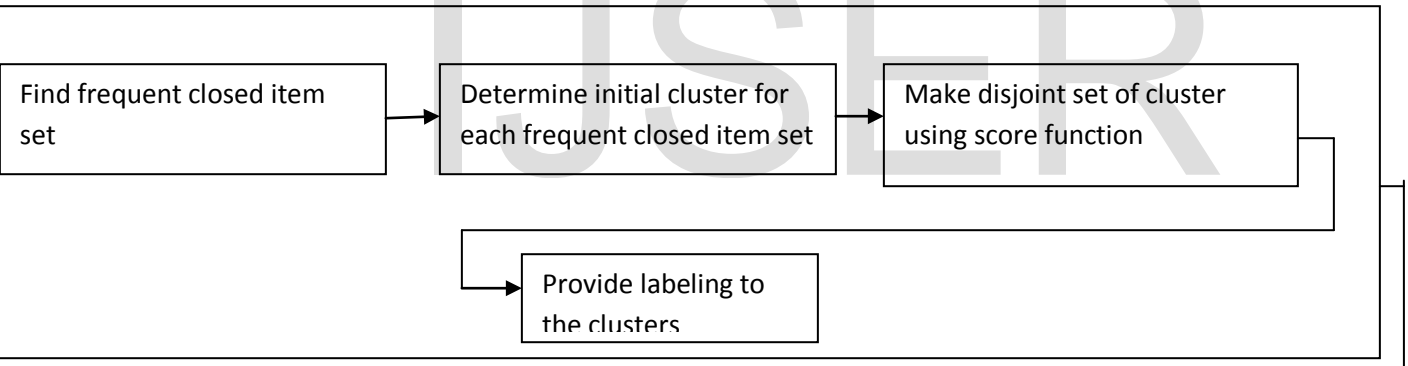
The proposed clustering method will contain the following steps or phases :-

1. First of all preprocessing of the documents is done which consists the Filtering, Tokenization, Stemming, and Stopword Removal which creates the normal document vectors.
2. Closed item-sets for normal vector document are generated specified by user.
3. Initialization of closed item set clusters is prepared.
4. Create disjoint of clusters by using score functions.
5. Build tree by using bottom-up approach and compute the score function for each parent.
6. Prune the tree to form a hierarchy of clusters.

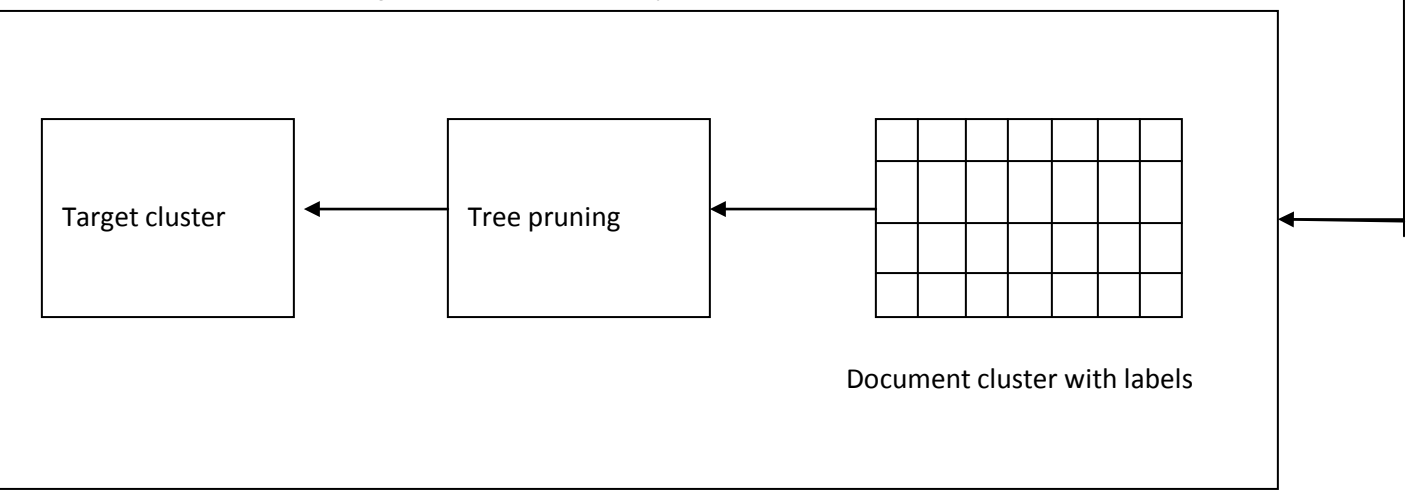
Stage 1: Document pre-processing

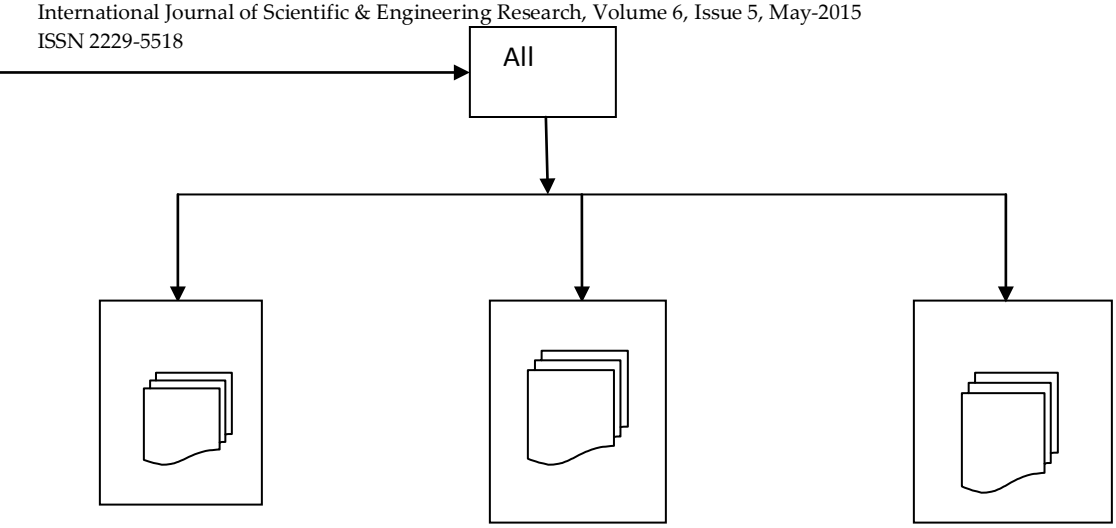


stage 2: Document Clustering



Stage 3: cluster tree hierarchy





Hierarchical cluster tree

fig 2: frame work of HDCUCI approach

. FUTURE WORK

Our focus will be to reduce the height of tree and also the proposed algorithm can be furthure modified to get the clustering results in other than English language. Latent Semantic Indexing or word Net Ontology can be used to improve the accuracy.

. COMPARISION

Following metrics will be used to evaluate the Quality and performance of clusters:

) F-Measure
) Purity
) Overall Similarity
) Entropy

After evaluating the metrics for quality and performance comparision will be done.

. CONCLUSION

Document clustering is widely used in various areas like web mining, information retrival , search engines etc. Most of the traditional algorithms does not satify the special requirement like high dimensionality, acuuracy , scalability,etc. This proposed method is hierarchical document clustering method by using closed itemsets,which may enhance the performance and quality of the clusters generated by the clustering method or technique.

ACKNOWLEDGMENT

I would like to express great pleasure and gratitude to Prof. Rajesh Pathak and Mr. Yatin Agrwal for their invaluable guidance and constant encouragement for my work. I would like express my gratitude to all my friends in Department of Computer Science & Engineering of GNIOT GR. Noida ,UTTERPRADESH.

REFERENCES

1. Beil, M. Ester, and X. Xu," Frequent term based text clustering". In Proc. 8th Int. Conf. on knowledge Discovery and Data Mining(KDD)' 2002, Edmonton, Alberta, Canada, 2002.
2. Hassan H. malik and John R. Kender," High Quality Efficient Hierarchical Document Clustering using Closed Interesting Itemsets". In Proc. Of the IEEE Int. Conf. on Data Mining(ICDM,2006), Hong Kong, 2006.

5. Y. Li and S. M. Chung," Parallel Bisecting K-Mean with Pradiction Clustering Algorithm", The Journals of Supercomputing, 39(1), Springer, pp. 19-37, January 2007.
4. Y. Zhu, B. C. M. Fung, D. Mu, Y. Li, " An Efficient Hybrid Hierarchical Document Clustering Method," FSKD, vol. 2, pp.395-399, 2008 Fifth Int. Conf. on Fuzzy Systems And Knowledge Discovery, 2008.
3. Xiaoke Su,Yang Lan, Renxia Wan and Yuming Qin," A fast Incremental Clustering Algorithm," proceedings of the 2009 Int. Sumposium on Information processing(ISIP'09), Huangshan, P. R. China, August 21-23, 2009,pp. 175-178.
6. M. Shriniwas and C. Krishna Mohan ," Efficient Clustering Apporoach using Incremental and Hierarchical Clustering Methods", 2010 IEEE.
7. Rekha Baghel, DR. Renu Dhir," A Frequent Concepts Based Document Clustering Algorithm", International Journal of Computer Application(0975-8887) vol. 4-no. 5 july 2010. Ashish Jaiswal and Prof. Nitin Janwe," Hierarchical Document Clustering : A review", In 2nd National Conference on Information and Communication Technology(NCICT) 2011 Proceedings Published in International Journal of Computer Applications
9. Ms. Devika Deshmukh, Mr. Sandip Kamble," Survey on Hierarchical Document Clustering techniques Fihc and F2hc", In International journal of Advanced Research in Computer Science and Software Engineering (ISSN: 2277-128X) vol. 3, Issue 7, July 2013.
10. M. S. Patil, M. S. Bewoor, S. H. Patil," A Hybrid Apporoach for Extractive Document Summarization Using machine Learning and Clustering Technique", In International Journal of Computer Science and Information Technology(ISSN: 0975-9646) , vol. 5(2), 2014.
11. Martin Mehlitz, Christian Bauckhage, Sahin ALbayrak," A Fast , Feature-based Cluster Algorithm For Information Retrieval".
12. M. Steinbach, G. Karypis, and V. Kumar," A comparision of Document Clustering Techniques, KDD Workshop On Text Mining '00, 2000Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.
13. Benjamin C. M. Fung, Ke Wang , and Martin Ester, Simon Fraser University, Canada,"Hierarchical Document Clustering".