Volume 5, Issue 12, December 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Hierarchical Document Clustering Using Closed Itemsets with Comparision Using Weka Tools

Kavita Nagar
Student of Master of Technology,
Department of Computer science and Engineering
Utter Pradesh Technical University,
Gr. Noida, India

Yatin Agarwal
Associate Professor
Department of Computer science and Engineering
Utter Pradesh Technical University,
Gr. Noida, India

***Abstract:** In Today's world huge amount of knowledge is propagated and stored in text databases over the large networks. This leads to increment in the numbers of document files. So we need a vigorous and skillful way to group this large amount of data. Clustering is the finest tool of data mining for regulating and harmonizing information. Clustering outline the similar objects or data into one cluster and different objects into another one based on their measurement of diminishing inter similarity and overestimating intra dissimilarity. However most of the clustering techniques face many issues like high dimensionality, scalability, accuracy, etc. Document clustering is an unsupervised clustering method for organizing documents, and providing fast information retrieval 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 item-sets.*

***Keywords:** Document clustering, hierarchical clustering, closed item--sets, literature review, Weka,K-Mean , OPTICS, Expectation Maximization(EM).*

I. INTRODUCTION

Clustering is an unsupervised learning method that binds the text data into same cluster situated on their similarity basis. Recently Hierarchical document clustering is widely being used to regulate and browse the information on the internet or network. A large number of clustering algorithms are available but most of them suffer from issues like high dimensionality, accuracy, scalability etc. Clustering algorithms mainly illuminated as Hierarchical and partitioning. In hierarchical clustering a tree like skeleton is created it can be of two types like bottom up or top down depending on which hierarchical approach like divisive or agglomerative is adapted. In partitioning method data is bisected into several sub datasets until the condition meet the criteriaor point of departure. Document clustering is the automatic organization of documents into clusters in which grouping is done on the basis of overestimating intra –cluster similarity and diminishinginter- cluster similarity. Document clustering algorithm is different fromclassification as it isbased on unsupervised learning in which we learn by observations rather than by some given examples. I am using Weka Data mining tool for clustering comparison.

II. WEKA

Weka is a data mining tool which contains various algorithms for data miningtasks. User can use the built in algorithms directly or can callthem from their own java code. Weka provides various tools like Classification, Clustering ,Regression, Data pre-processing ,Association rule mining and visualization.Weka is open source software and freely available. It is platform-independent.

Waikato Environment for Knowledge Analysis
Version 3.6.7
(c) 1999 - 2012
The University of Waikato
Hamilton, New Zealand

Figure 1: WEKA Tool

III. RELATED LITERATURE SURVEY

AS we know that clustering delivers the grouping of same data or objects based on their diminishing intersimilarity and overestimating dissimilaritymeasurements. Hierarchical clustering aggregate the data into a tree like skeleton or cluster which can be further dissolvedinto two parts i.e. Divisive and Agglomerative which outline the cluster in splitting and merging fashion. If decision is not taken carefully for specific split or merge then it can lead to wrong output and the results cannot be changed. On the other hand in partitioning method data set is dissolved into sub data sets and each data set specifically belong to one data sets.

In 2002, Beil, Ester and Xu [1]addressed the problem of finding the cognate content or data from the intranet and state that algorithm like bisecting K-Means did not indulge the obligation 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 volunteered by[1] was a greedy algorithm and declared 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 track ahead from the HFTC algorithm in which they used the sub-linearly scalable notion with closed interesting item-sets for hierarchical document clustering method.It has been notified that if same dawn is used for the first level of it will results in small numbers of closed interesting item-sets as examine to number of closed frequent item-set originated.

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

In 2008, Yehang Zhu, Fung, Dejun Mu and Yangling[4] volunteered a novel hierarchical clustering method which was a combination of partitioning and agglomerative clustering approach. Experiments results displays that the proposed method was serviceable and profitable but accuracy was not so striking for some real life large datasets.

In 2009 Xiaoke Su, Yang Lan, Renxia Wan and YumingQin[5] recommended a fast incremental hierarchical clustering algorithm which was performable and Profitable. Transcendent analysis and experiments results displays the authentic features of the data set andblown away the inadequate brunt of memory but also reflect

In 2010 M. Sriniwas, C. Krishna Mohan [6] proposed a clusteringalgorithm calledLeaders Complete Linkage algorithm (LCL) which was the reinforcement of the hierarchical and incremental clustering. In this at each iterationobjects are accumulatedinto one cluster to another by splitting and merging two clusters.When splitting and merging operations are performed on the two cluster then partition quality is checked. If the derived quality is good only then the next level ofsplitting and merging can be achieved. Otherwise Current partition will be the final clustering result.

In 2010 RekhaBaghel, Dr. RenuDhir[7] proposed a frequent concept document clusteringalgorithm (FCDC) in which was based on the semantic relationship between the words. Word Net Ontology concept was used to create dimensional feature vector which permitsdeveloping virtuous clustering algorithm. FCDC found more economical rigid and elastic when compared with other existing algorithms like, FIHC, K-Mean, and UPGMA.

In 2013, Ms. DevikaDeshmukh, Mr. Sandipkamble[9] proposed an effective Fuzzy Frequent Item-set Based Hierarchical Clustering(F2HC) which was based on the fuzzy association rule mining. Which performwell in term of cluster quality and accuracy? It performs the clustering in three steps .In the first step it find outs 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] advisedan extort text epitomize method which is based on the Support Vector Machine (SVM). It displays refined results in performance and quality of the summary spawn 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, and thenconverts the text information into the summary and in the last step generates the full summary for the text.

In 2015 Twinkle Svadas, JasminJha [50] proposed a system to categorize the text documents and form the clusters. This system consists two components i.e. Pre-processing and Ontology. Pre-processing involves all processes, methods that are required to prepare data for text mining. It converts data from original form to machine readable format before applying feature extraction methods to generate new collection of documents represented by the concepts. Techniques like stop word removal, stemming and tokenization are involved in preprocessing. Ontology is considered as a repository of knowledge in which concepts and terms are defined and also the relationships between these terms and concepts are given. It is a set of concepts and relationships that describe a domain of interests and represents an overview of the domain. Ontology automates information processing and improve text mining for a particular domain.

In 2015 Latika [51] proposed an effective and efficient algorithm for clustering text documents. This algorithm is formulated by using the concept of well known k-means algorithm. The standard k-means algorithm suffers from the problem of random initialization of initial cluster centers. The proposed algorithm eliminates this problem byintroducing a new approach for selection of initial cluster centroids. Several experiments are conducted on mini_newsgroups dataset to measure the performance of proposed algorithm and find that the results obtained were very promising when compared to two other algorithms: k-means and enhanced k-means.

IV. PROPOESED METHOD

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

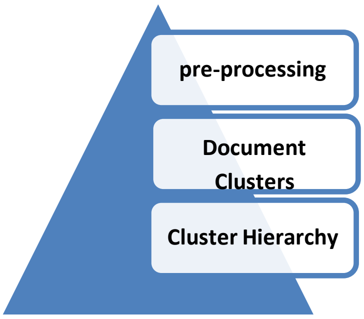
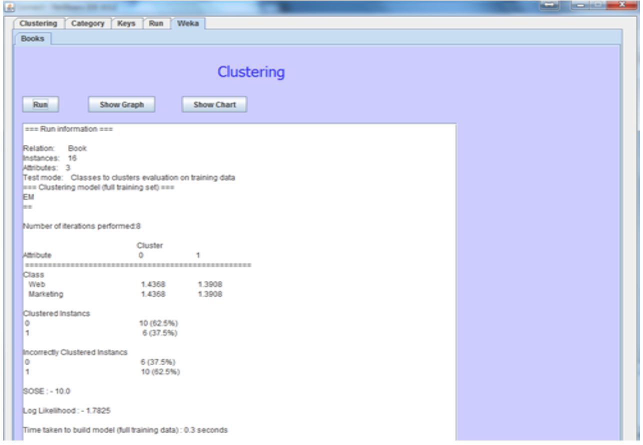


Figure 2: DocumentClustering Architecture

The proposed clustering method will contain the following steps orphases:-

(i) First of all preprocessing of the documents is done which consists the Filtering, Tokenization, Stemming, and Stop word Removal which creates the normal document vectors space for documents in terms of frequency.
(ii) Closed item-sets for normal vector document are generated specified by user using some sill.
(iii) Initialization of closed item set and then clusters will be prepared.
(iv) Actualize disjoint of clusters by using score functions.
(v) Build tree by using bottom-up approach and compute the score function for each parent.
(vi) Prune the tree to mold a hierarchy of clusters.



```
=== Run information ===

Relation:    Book
Instances:   16
Attributes:  3
Test mode:   Classes to clusters evaluation on training data
=== Clustering model (full training set) ===
EM
==

Number of iterations performed:8

                                 Cluster
Attribute                        0                1
===================================================
Class
  Web                            1.4368           1.3908
  Marketing                      1.4368           1.3908

Clustered Instancs
0                                10 (62.5%)
1                                 6 (37.5%)

Incorrectly Clustered Instancs
0                                6 (37.5%)
1                                10 (62.5%)

SOSE : - 10.0

Log Likelihood : - 1.7825

Time taken to build model (full training data) : 0.3 seconds
```

Figure 3: Result of proposed clustering algorithm HDCUCI

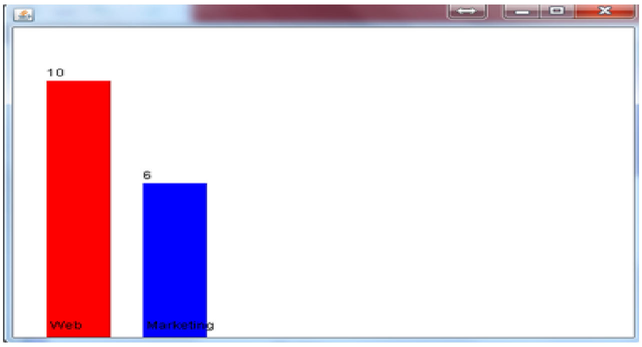


Figure 4: Data processed by HDCUCI

V. OTHER CLUSTERING METHODS

❖ K-means Clustering

The K-means algorithm is the most generally algorithm which usespartitioned clustering algorithm concept. It can be easily implemented and is the most qualified one in terms of the execution time. The algorithm works like as follows :

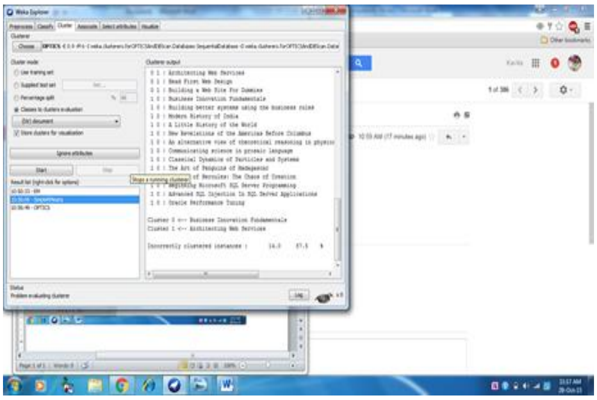
K-Means Algorithm: The algorithm for partitioning, In which each cluster's center point is represented by mean value ofobjects in the cluster.

Input: k: the number of clusters. D: a data set containing n objects.

Output: A set of k clusters.

Method:

1. Randomly choose any k objects from D as the initial cluster centers point.
2. Repeat.
3. Based on the mean value of theobjects which is same to the cluster reassign object to that cluster and update them until no changes are required.
4. Update the cluster means, i.e. calculate the mean value of the objects for each cluster.



```
0 1 | Architecting Web Services
0 1 | Head First Web Design
0 1 | Building a Web Site For Dummies
1 0 | Business Innovation Fundamentals
1 0 | Building better systems using the business rules
1 0 | Modern History of India
1 0 | A Little History of the World
1 0 | New Revelations of the Americas Before Columbus
1 0 | An alternative view of theoretical reasoning in physic
1 0 | Communicating science in prosaic language
1 0 | Classical Dynamics of Particles and Systems
1 0 | The Art of Penguins of Madagascar
        of Hercules: The Chaos of Creation
1 0 | Beginning Microsoft SQL Server Programming
1 0 | Advanced SQL Injection In SQL Server Applications
1 0 | Oracle Performance Tuning

Cluster 0 <-- Business Innovation Fundamentals
Cluster 1 <-- Architecting Web Services

Incorrectly clustered instances :       14.0     87.5   %
```

Figure 5: Result of K-Mean clustering

❖ EM Algorithm

EM ALGORITHM is an iterative method for searching out the parameter estimates. Every cluster or group is specified by parametric probability distribution. The EM forms an expectation (E) step, which enumerates the expectation of the log-likelihood by using the current estimate for the parameters, and maximization (M) step, enumerates the expected log-likelihood found on the Estep. Thesetwo parameter-estimates are then used to resolve the distribution of the latent variables in the next E step.

1. **Expectation**: Fix model and finds out missing labels.
2. **Maximization**: Find out the model that maximizes the expected log-likelihood of the data.

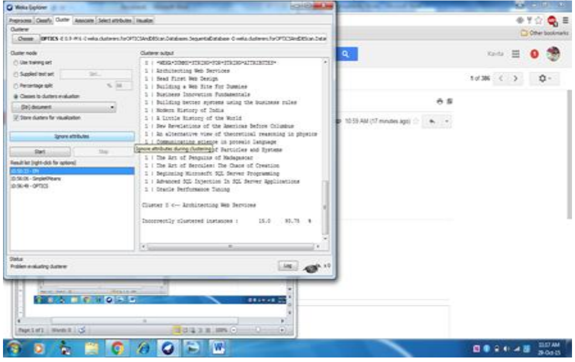
General EM Algorithm in:

E step:

Assessment the distribution over labels for given a certain fixed model.

M step:

Discriminate new parameters for model to maximize expected log-likelihood of recognized data and hidden variables.

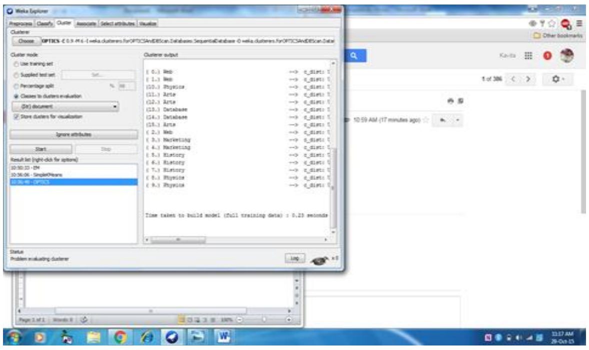


```
1 | <MERGE>ITEMS<STRING>FOR<STRING>ATTRIBUTES>
1 | Architecting Web Services
1 | Head First Web Design
1 | Building a Web Site For Dummies
1 | Business Innovation Fundamentals
1 | Building better systems using the business rules
1 | Modern History of India
1 | A Little History of the World
1 | New Revelations of the Americas Before Columbus
1 | An alternative view of theoretical reasoning in physics
1 | Communicating science in prosaic language
    f Particles and Systems
1 | The Art of Penguins of Madagascar
1 | The Art of Hercules: The Chaos of Creation
1 | Beginning Microsoft SQL Server Programming
1 | Advanced SQL Injection In SQL Server Applications
1 | Oracle Performance Tuning

Cluster 0 <-- Architecting Web Services

Incorrectly clustered instances :       15.0     93.75  %
```

Figure 6: Result of EM clustering

❖ OPTICS

OPTICS ("Ordering Points to Identify the ClusteringStructure") creates liner ordering of objects in the database. Its basic idea is collateral to DBSCAN, but it abodeone of DBSCAN's major enervation i.e: the problem of revealing indicative clusters in data of varying density. For this the points of the database are (linearly) ordered in such a way that points which are spatially closest will becomes the neighbors in the ordering. Additionally, a special distance is stocked for and every point which epitomizes the density that is further needed to sanctioned for a cluster in order to have both points belong to the same cluster. This distance is required to cut off the density of clusters It contains two points like DBSCAN which are "e" and "Minpt" where "e " describes the maximum distance and " Minpt" describes the number of objects which are required to form the cluster.



```
( 0.) Web                                    --> c_dist:
( 1.) Web                                    --> c_dist:
(10.) Physics                                --> c_dist:
(11.) Arts                                   --> c_dist:
(12.) Arts                                   --> c_dist:
(13.) Database                               --> c_dist:
(14.) Database                               --> c_dist:
(15.) Arts                                   --> c_dist:
( 2.) Web                                    --> c_dist:
( 3.) Marketing                              --> c_dist:
( 4.) Marketing                              --> c_dist:
( 5.) History                                --> c_dist:
( 6.) History                                --> c_dist:
( 7.) History                                --> c_dist:
( 8.) Physics                                --> c_dist:
( 9.) Physics                                --> c_dist:

Time taken to build model (full training data) : 0.23 seconds
```

Figure 7: Result of OPTICS clustering

VI. FUTURE WORK

Our focus will be to reduce the height of tree and also the proposed algorithm can be furthers 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.

VII. COMPARISION

Above section involve the proposed method and three other techniques introduced previously using WekaClustering Tool on a document data set. Clustering of the data set is done with each of the clustering algorithm using Weka tool and the results are:-

Table 1: Comparison result of algorithms

Name	No. of Clusters	No. of Iterations	Cluster Instances	Time Taken to Build Model	Log Likelihood	Sum of error mean	Unrelated Cluster Instances
K-Mean	2	2	0:13(81% VIII.1:3(19%))	0 seconds	---------	10.0	14.0%, 87.5%

OPTICS	2	2	0:13(81%)
			1:3(19%)
				0.22seconds	---------	-------
						-
							14.0%, 87.5%

EM	2	5	0:16(100%)	0.17seconds	1.78257	0	15.0%, 93.5%

HDCUCI	2	-----------	0:10(62.5%)
			1:6(37.5%)
				0.3 seconds	---------	0	37.5%, 62.5%

VIII. CONCLUSION

Document clustering is widely used in various areas like web mining, information retrieval, search engines etc. Most of the traditional algorithms do not satisfythe special requirement like high dimensionality, accuracy, scalability,etc. This proposed method is hierarchical document clustering methodby using closed item-sets, which may enhance the performance and quality of the clusters obtained by the clustering method or . And the comparison results shows that proposed method performance is better than other clustering algorithm. All the algorithms have some ambiguity in some data when clustered.

ACKNOWLEDGMENT

I would like to express great pleasure and gratitude to Prof. Rajesh Pathak and Mr. YatinAgrwal 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. 8thInt. 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.

[3] Y. Li and S. M. Chung," Parallel Bisecting K-Mean with Prediction 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.

[5] XiaokeSu,YangLan, Renxia Wan and Yuming Qin," A fast Incremental Clustering Algorithm," proceedings of the 2009 Int. Symposium 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 Approach using Incremental and Hierarchical Clustering Methods", 2010 IEEE.

[7] RekhaBaghel, DR. RenuDhir," A Frequent Concepts Based Document Clustering Algorithm", International Journal of Computer Application(0975-8887) vol. 4-no. 5 July 2010.AshishJaiswal and Prof. NitinJanwe," Hierarchical Document Clustering : A review", In 2nd National Conference on Information and Communication Technology(NCICT) 2011 Proceedings Published in International Journal of Computer Applications

[8] Ms. DevikaDeshmukh, Mr. SandipKamble," 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.

[9] M. S. Patil, M. S. Bewoor, S. H. Patil,” A Hybrid Approach 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.

[10] Martin Mehlitz, Christian Bauckhage, SahinALbayrak,” A Fast, Feature-based Cluster Algorithm ForInformation Retrieval”.

[11] M. Steinbach, G. Karypis, and V. Kumar,” A comparisons of Document Clustering Techniques, KDD Workshop On Text Mining ’00, 2000Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[12] Benjamin C. M. Fung, KeWang , and Martin Ester, Simon Fraser University, Canada,” Hierarchical Document Clustering”.