Journal of Data Mining and Knowledge Discovery
ISSN: 2229–6662 & ISSN: 2229–6670, Volume 3, Issue 2, 2012, pp.-65-68.
Available online at http://www. bioinfo. in/contents. php?id=42

A REVIEW ON HIERARCHICAL DOCUMENT CLUSTERING

DESHMUKH D.B. AND PANDEY Y.

Department of Computer Science & Engineering, Sagar Institute of Research & Technology, Bhopal, India.
*Corresponding Author: Email- devikabdeshmukh@gmail.com, anilpandey45@gmail.com

Received: March 15, 2012; Accepted: April 12, 2012

Abstract- As text documents are largely increasing in the internet, the process of grouping similar documents for versatile applications have put the eye of researchers in this area. However most clustering methods suffer from challenges in dealing with problems of high dimensionality, scalability, accuracy and meaningful cluster labels. This paper presents a review on all these well known methods of document clustering. Hierarchical document clustering method is explained in detail. Study shows that hierarchical document clustering performs well but still there is a scope to improve above mentioned problems.

Keywords- Document clustering, Hierarchical clustering, Frequent item sets.

Citation: Deshmukh D.B. and Pandey Y. (2012) A Review on Hierarchical Document Clustering. Journal of Data Mining and Knowledge Discovery, ISSN: 2229–6662 & ISSN: 2229–6670, Volume 3, Issue 2, pp.-65-68.

Introduction

Document Cluster is a set of similar documents and automatic grouping of text documents is called Document Clustering. The documents within a cluster have high similarity in comparison to one another but are dissimilar to documents in other clusters. Thousands of electronic documents are being added on World Wide Web or internet and to browse them efficiently or search for the relevant data effectively, the concept of Document Clustering is important today. Document clustering is widely applicable in areas such as web mining, information retrieval, search engines and topological analysis [1].

Many algorithms have been proposed for clustering the documents but most of them do not satisfy the special requirements for clustering documents: high dimensionality, scalability, accuracy, easy to browse and prior domain knowledge.

High dimensionality

Each term in the document can be regarded as a dimension and there are thousands of terms in a document. Clustering algorithms can handle the small data sets but for large data sets it's challenging.

Scalability

Many clustering algorithms work well on small data sets but fail to work on large data sets containing millions of objects. Thus high scalable clustering algorithms are needed to resolve this problem

Accuracy

A good clustering solution should have high intra cluster similarity and low inter cluster similarity. The accuracy of the Clustering algorithm is measured by F- measure, which is an evaluation method to check the performance.

Easy of browsing

To browse efficiently, documents should be placed in hierarchy with meaningful cluster description.

Knowledge of input parameters

Many clustering algorithms require users to provide prior knowledge, for example, number of clusters. The result of the clustering algorithm may be sensitive to such input parameters but it's impossible for the user to determine the number of clusters. Thus this may degrade the clustering accuracy [1].

Clustering algorithms are mainly categorized into hierarchical and partitioning methods. A hierarchical clustering method works by grouping data objects into a tree of clusters. These methods can further be classified into agglomerative and divisive hierarchical clustering depending on whether the hierarchical decomposition is formed in a bottom-up or top-down fashion. Steinbach showed that Unweighted Pair Group Method with Arithmetic Mean (UPGMA) is the most accurate one in agglomerative category [2].

K-means and its variants are the most well-known partitioning methods that create a flat, non-hierarchical clustering consisting of k clusters. The k-means algorithm iteratively refines a randomly chosen set of k initial centroids, minimizing the average distance (i.e., maximizing the similarity) of documents to their closest (most similar) centroid. The bisecting k-means algorithm first selects a cluster to split, and then employs basic k-means to create two sub -clusters, repeating these two steps until the desired number k of clusters is reached [1]. Steinbach shows that the bisecting k-means algorithm outperforms basic k-means as well as agglomerative hierarchical clustering in terms of accuracy and efficiency [3].

Wang [2] introduces a new criterion for clustering transactions using frequent itemsets. In principle, this method can also be applied to document clustering by treating a document as a transaction; however, the method does not create a hierarchy for browsing. The HFTC proposed by Beil [4] attempts to address the special requirements in document clustering using the notion of frequent itemsets. HFTC greedily picks up the next frequent itemset (representing the next cluster) to minimize the overlapping of the documents that contain both the itemset and some remaining itemsets. The clustering result depends on the order of picking up itemsets, which in turn depends on the greedy heuristic used. HFTC was the first algorithm in this class and achieves accuracy comparable to 9-secting k-means, and worst than bisecting k-means. Fung showed that HFTC is not scalable for large document collections and proposed FIHC; a frequent itemset based clustering approach that claims to outperform HFTC and the best-known agglomerative and partitional methods, both in terms of accuracy and scalability [5].

FIHC is "cluster-centered" in that it measures the cohesiveness of a cluster directly using frequent itemsets: documents in the same cluster are expected to share more common itemsets than those in different clusters. A frequent itemset is a set of terms that occur together in some minimum fraction of documents [1]. This approach is very different from HFTC where the clustering solution greatly depends on the order of selected itemsets. Instead, FIHC assigns documents to the best cluster from among all available clusters (frequent itemsets) [1].

FIHC uses only the global frequent items in document vectors, drastically reducing the dimensionality of the document set. Experiments show that clustering with reduced dimensionality is significantly more efficient and scalable. FIHC is not only scalable, but also accurate. The clustering accuracy of FIHC consistently outperforms other methods. FIHC allows the user to specify an optional parameter, the desired number of clusters in the solution. However, close-to-optimal accuracy can still be achieved even if the user does not specify this
parameter [2].

The rest of the paper is organized as follows: Section 2 represents the related work in the field; Section 3 represents the algorithm for hierarchical document clustering. Section 4 describes the comparison of algorithms, and Section 5 outlines the future direction to the hierarchical document clustering. We conclude the paper in section 6.

Related Work

In order to solve the problem of high dimensionality, scalability, accuracy various researchers put their efforts. Amaud Ribert, Abdel Ennaji, Yves Lecourtier [6] in 2009 proposed an incremental hierarchical clustering method which is an alternative to partitional clustering technique. The proposed method proceeds by updating the hierarchical representation of the data instead of recomputing the whole tree when new patterns have to be taken into account. Tests have shown that using this algorithm allow to progressively perform the hierarchical clustering for big sets of data which can then contain seven time more elements than using the classical algorithms.

Benjamin C.M. Fung, Ke Wang, Martin Ester [2] in 2003 proposed an algorithm to use the notion of frequent itemsets which comes from association rule mining for document clustering. Each cluster is identified by some words called frequent itemsets for the document in the cluster. Frequent itemsets are also used to produce hierarchical topic tree structure for clusters. By focusing on frequent items the dimensionality of the document set is reduced. This method outperforms best in terms of both clustering accuracy and scalability.

Hassan H. Malik, John R. Kender in 2006 proposed [5] a method which is one step ahead from the previous algorithm. They introduced the notion of closed interesting itemsets. Using closed interesting itemsets they proposed new, sublinearly scalable, hierarchical document clustering method. Result show that using the same support threshold for first level itemsets results in significantly smaller number of closed interesting itemsets as compared to the number of closed frequent itemsets generated.

Chun- Ling Chen, Frank S.C. Tseng, Tyne Liang [7] in 2008 proposed an effective fuzzy frequent itemset based hierarchical clustering approach which uses fuzzy frequent itemsets discovered by fuzzy association rule mining to improve the clustering accuracy of FIHC. Algorithm works in three stages. In the first stage the key terms will be retrieved from the document set for removing noise, and each document is pre-processed into the designated representation for the following mining process. In the second stage, a fuzzy association rule mining algorithm is employed to discover a set of highly relevant fuzzy frequent itemsets, which contains key terms to be regarded as the labels of candidate clusters. In the final stage, the documents will be clustered into a hierarchical cluster tree based on these candidate clusters. The obtained hierarchical cluster tree with meaningful cluster descriptions can offer users a more flexible ability in document management.

Anuj Sharma, Renu Dhir [8] in 2009 proposed a wordset based document clustering algorithm for large datasets. WDC uses a wordsets based approach to build clusters. It first searches frequent closed wordsets by association rule mining and then form initial cluster of documents with each cluster representing single closed wordsets. Then the algorithm refines the initial clusters and make final results as a clustering tree like representations. The idea is to do clustering of documents by using the wordsets that occur in sufficient number of documents. Each document in this approach corresponds to transaction and each word corresponds to an item as in association rule mining. WDC performs well in terms of quality of cluster form.

Xiaoke Su, Yang Lan. Renxia Wan and Yuming Qin [9] in 2009 proposed a fast incremental hierarchical clustering algorithm which is found to be feasible and effective. The existed incremental clustering algorithm does not take the memory constraint into account and it is difficult to obtain a satisfy result when it is used for large-scale data sets. A fast clustering algorithm is presented by changing the radius threshold value dynamically. The clustering result is no longer spherical shape. At the same time an inter-cluster dissimilarity measure is put forward which is capable of handling the categorical data. Theoretical analysis and experimental results show the algorithm can not only overcome the impact of the inadequate of the memory when clustering the large scale data set, but also accurately reflect the characteristics of the data set. Both of these indicate the effectiveness of the algorithm. Clustering with the fixed final clusters number will show a reliable rationality, and can be used for ultra-large-scale data set, particularly for the data stream environment.

M. Srinivas, C. Krishna Mohan [10] in 2010 present a hybrid clustering algorithm namely, Leaders complete linkage algorithm (LCL) that combines the advantages of hierarchical clustering and incremental clustering techniques. Instead of rearranging all objects like what partitional algorithms usually do, in each iteration of the clustering, some objects but all are moved from one cluster to another by the way of splitting a cluster or merging two clusters. It can start with a single cluster containing all objects or start with each object in a distinct cluster. At each step during the clustering, the quality of the current partition is examined as well as the quality of the partition after splitting one cluster or merging two clusters. If the quality of the partition is improved after the splitting or the merging, then a further splitting or merging will be performed. Otherwise, the clustering will terminate and the current partition is the final clustering result. The idea behind the mix of splitting and merging is to allow amendment to the previous clustering result so that high quality clustering can be achieved. The basic technique used is to find suitable prototype from large datasets and then apply the clustering method using the prototype. This algorithm leads to good clustering results in shorter type.

Rekha Baghel Dr. Renu Dhir [11] in 2010 proposed a Frequent Concept based document clustering (FCDC) algorithm which utilizes the semantic relationship between words to create concepts. It exploits the WordNet ontology in turn to create low dimensional feature vector which allows us to develop a efficient clustering algorithm. It uses a hierarchical approach to cluster text documents having common concepts. FCDC found more accurate, scalable and effective when compared with existing clustering algorithms like Bisecting K-means, UPGMA and FISC.

Hierarchical Document Clustering

Hierarchical document clustering is found to be better than the partitioning methods. The main propose of hierarchical document clustering is to build a hierarchical tree of clusters whose leaf nodes represent the subset of a document collection. Moreover, this method can be further classified into agglomerative and divisive approaches, which work in a bottom-up and top-down fashion, respectively. An agglomerative clustering iteratively merges two most similar clusters until a terminative condition is satisfied. On the other hand, a divisive method starts with one cluster, which consists of all documents, and recursively splits one cluster into smaller sub-clusters until some termination criterion is fulfilled.

Proposed work is an extension of mining fuzzy frequent itemsets for hierarchical document clustering. Fuzzy frequent itemset based hierarchical clustering F^{2}IHC proposed as follows [7]:

1. In the first stage, the key terms will be extracted from the document set, and each document is pre-processed into the designated representation for the following mining process. In this stage, a hybrid feature selection method will be used to effectively reduce the unimportant terms for each document.
2. In the second stage, to discover a set of relevant fuzzy frequent itemsets efficiently, we will propose a fuzzy association rule mining algorithm for text. In this algorithm, we revise the method devised by Hong by regarding a document as a transaction, and those term frequency values in a document as the quantitative values (i.e., the number of purchased items in a transaction). A frequent itemset, as defined by Fung [2], is a set of words that occur together in some minimum fraction of documents in a cluster. By employing pre-defined membership functions, our algorithm calculates three fuzzy values, i.e., Low, Mid, and High regions, for each term based on its frequency to discriminate the degree of importance of the term within a document in the mining process. The derived fuzzy frequent itemsets contain key terms to be regarded as the labels of candidate clusters.
3. In the final stage, the documents will be clustered into a hierarchical cluster tree based on these candidate clusters. The cluster tree will be built in a top-down fashion to recursively select the parent clusters at level k - 1 for dividing the documents into its suitable children clusters at level k. Notice that the clusters generated by our algorithm are crisp partitions for assigning a document to exactly one cluster.

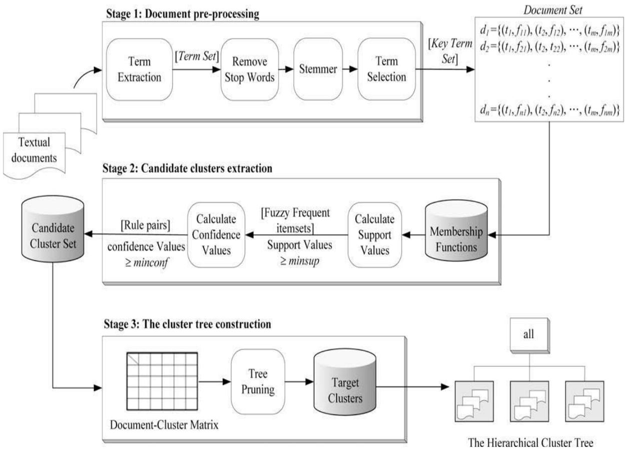
Stage 1: Document pre-processing
Term Extraction
[Term Set]
Remove Stop Words
Stemmer
Term Selection
[Key Term Set]
Document Set
$d_1=\{(t_1,f_{11}),(t_2,f_{12}),\cdots,(t_m,f_{1m})\}$
$d_2=\{(t_1,f_{21}),(t_2,t_{22}),\cdots,(t_m,f_{2m})\}$
$d_n=\{(t_1,f_{n1}),(t_2,f_{n2}),\cdots,(t_m,f_{nm})\}$
Textual documents
Stage 2: Candidate clusters extraction
Candidate Cluster Set
[Rule pairs]
confidence Values ≥ minconf
Calculate Confidence Values
[Fuzzy Frequent itemsets]
Support Values ≥ minsup
Calculate Support Values
Membership Functions
Stage 3: The cluster tree construction
Document-Cluster Matrix
Tree Pruning
Target Clusters
all
The Hierarchical Cluster Tree

Fig. 1- Framework of F2IHC approach

Experiments show that the accuracy of the algorithm is higher than that of FIHC method, UPGMA, and Bisecting k-means when compared on the five standard datasets [7]. Moreover, the experiment results show that the use of fuzzy association rule mining discovery important candidate clusters for document clustering to increase the accuracy quality of document clustering. Therefore, it is worthy extending in reality for concentrating on huge text documents management.

Comparison of Algorithms

Steinbach [3] showed that UPGMA is the most accurate one in Agglomerative hierarchical clustering. K means and its variants represent the category of partitioning clustering methods. Steinbach's experimental result shows that bisecting K-means technique is better than standard k-means as well as agglomerative approach in terms of accuracy and efficiency [3].

HFTC is comparable to bisecting k- means in terms of clustering accuracy but experiments shows that HFTC is not scalable [1, 2].

The Experimental results of Frequent Itemset based Hierarchical Clustering (FIHC) is compared with UPGMA, bisecting k-means and HFTC in terms of F-measure, sensitivity to parameters, efficiency and scalability. The FIHC approach outperforms its competitors in terms of accuracy, efficiency and scalability [2]. This approach fails when the number of frequent sets of terms is large.

Experiment in [5] shows that there is a significant dimensionality reduction with closed interesting itemsets when compared with bisecting k- means and FIHC. It outperforms state of the art approaches in terms of accuracy and run time performance.

Experimental results of [8] are compared with UPGMA, bisecting k - means and FIHC. It is found that WDC outperforms other algorithms including FIHC in terms of accuracy. It is less sensitive to input parameter.

Experimental results of [11] shows that the FCDC algorithm outperforms other algorithms in terms of accuracy. FCDC is compared with FIHC, UPGMA and bisecting K- means and FCDC has better accuracy than UPGMA which is regarded as the best in hierarchical document clustering algorithm. FCDC is more insensitive to number of clusters and produces better results than FIHC and bisecting k- means.

Experimental results in [7] shows that the accuracy of F^{2}IHC is higher than that of FIHC method, UPGMA and bisecting k- means. This approach not only retains the merits of FIHC but also improves the document clustering accuracy quality as compared with the FIHC method.

Future Work

Our focus is for Reduction of height of tree which plays an important role in hierarchical document clustering. WordNet can be used to reduce the dimensionality.

Conclusion

Document clustering is widely applicable in areas such as search engines, web mining, information retrieval and topological analysis. Here we studied number of document clustering algorithms. Mining fuzzy frequent itemset method is considered as the best which solves the problem of high dimensionality, scalability and accuracy.

References

[1] Benjamin C. M. Fung, Ke Wang, and Martin Ester. *Hierarchical Document Clustering*.

[2] Fung B., Wang K. and Ester M. (2003) *SIAM International Conference on Data Mining*, 59-70.

[3] Steinbach M., Karypis G. and Kumar V. (2000) *A comparison of document clustering techniques*.

[4] Beil M. Ester and Xu X. (2002) *8th Int. Conf. on Knowledge Discovery and Data Mining* (KDD).

[5] Malik H.H. and Kender J.R. (2006) *IEEE International Conference on Data Mining (ICDM)*.

[6] Arnaud Ribert, Abdel Ennaji and Yves Lecourtier (1999) *An incremental Hierarchical Clustering,* 19-21.

[7] Chun-Ling Chen, Frank S.C. Tseng and Tyne Liang (2010) *International Journal of Information Processing and Management,* 46, 193-211.

[8] Anuj Sharma and Renu Dhir (2009) *International Conference on Methods and Models in Computer Science*.

[9] Xiaoke Su, Yang Lan, Renxia Wan and Yuming Qin (2009) *International Sumposium on Information Processing* (ISIP), 175-178.

[10] Shriniwas M. and Krishna Mohan C. (2010) *Efficient Clustering Approach using Incremental and Hierarchical Clustering Methods*.

[11] Rekha Baghel and Renu Dhir (2010) *International Journal of Computer Applications*, 4(5), 0975-8887.

[12] Chun-Ling Chen, Frank S.C. Tseng, Tyne Liang (2008) *IEEE, Third international conference on Innovative Computing Information and Control* (ICICIC).

[13] Jain A.K., Murty M.N. and Flynn P.J. (1999) *ACM Computing Surveys*, 31(3), 264-323.

[14] Rui X. (2005) *IEEE Transactions on Neural Networks*, 16(3), 634-678.

[15] Agrawal R. and Srikant R. (1994) *20th Int. Conf. Very Large Data Bases,* 12-15.