Automatic News Extraction Using Hierarchical Clustering

Anish Gupta[#1],Bhanu Prakash Lohani[*2],Pradeep Kumar Kushwaha[@3]

[#1] Assistant Professor,Amity University Gr Noida Campus

[1] agupta@gn.amity.edu

[*2] Senior Lecturer, Amity University Gr Noida Campus

[2] bhanuplohani@gmail.com

[@3] Senior Lecturer, Amity University Gr Noida Campus

[3] pkkushwaha@gn.amity.edu

***Abstract*- In today's era, web newspapers are a rich source of information. User's need is to find the specific information content. But users sometimes may not exactly know what they are looking for. In order to find the interesting news article from the large database, various studies have been devoted to data mining, text mining, web news mining and various other knowledge discovery techniques in the past years. To benefit more from the available information, term mining techniques are also applied. With the increasing amount of number of news sources, utilizing extracted information in deciding about the relevant news becomes increasingly urgent and difficult. Frequent term clustering [2] can be used to cluster large sets of news articles. The well-known methods of clustering do not really address the special problems of clustering: very high dimensionality of the data, very large size of the databases and understandability of the cluster description. This paper will discuss the novel approach of hierarchical clustering algorithm for frequent term based clustering. Hierarchical clustering organizes news (clusters) into a tree or a hierarchy. The parent-child relationship among the nodes in the tree can be viewed as a topic-subtopic relationship.**

***Keywords*- Data mining, Clustering, Frequent term sets**

I. INTRODUCTION

Clustering is an automatic grouping of news articles into clusters so that news within a cluster has high similarity in comparison to one another, but is dissimilar to news in other clusters [5]. Due to massive amount of unstructured news articles, clustering techniques have been widely used to retrieve, filter, and categorize news available on the World Wide Web. However, traditional text clustering techniques do not really address the following major challenges for clustering news [2]:

• Very high dimensionality of the data. The number of relevant news in a document is very large. Each of these news forms a dimension in the vector space.

• Very large size of the databases. Clustering g algorithms are easily applicable to small size of database, but do not form fine clusters in the large database.

• Understandable description of the clusters. Cluster descriptions guide the user interactive browsing.

In this paper we are using frequent term based hierarchical clustering method for news extraction. Hierarchical clustering is based on the formation of nodes where each node represents a term [1]. A term is generally a sequence of characters separated from others terms by a delimiter. Term is made of single /several words. News articles contain a large set of important terms that are retrieved and organized in a hierarchical manner. This form of hierarchy helps in formation of clusters based on the similarity of news terms within a cluster and dissimilarity of terms among intra clusters.

© 2014 IJAIR. All Rights Reserved

But the number of news articles is growing continuously in a database, and it is infeasible to rebuild the cluster tree upon every arrival of a news article. Frequent Item based Hierarchical clustering (FIHC) [1] incrementally updates the cluster tree. It simply assigns the news document to the most similar existing news cluster. Frequent Item based Hierarchical clustering (FIHC) follows the concept of hierarchical clustering.

Hierarchical clustering forms the hierarchy of news articles bottom-up by iteratively computing the similarity between all pairs of clusters (news) and then merging the most similar pair. The hierarchy can also be built top-down which is known as the divisive approach. It starts with all the news terms in the same cluster and iteratively splits a cluster into smaller clusters until a certain termination condition is fulfilled.

The rest of this paper is organized as follows. Section 2 briefly introduces the related approach of frequent term clustering including preprocessing and discusses the need of hierarchical strategy. In section 3, we introduce our novel method of Frequent Item based Hierarchical clustering (FIHC) frequent term-based text clustering and present the algorithm. An experimental evaluation on real text data was conducted, and section 4 summarizes the paper and outlines some interesting directions for future research.

II. RELATED WORK

The process of news extraction using frequent term based clustering requires certain preprocessing steps [7]. Firstly, news articles stored in the database are in unstructured form that is converted into structured form for further processing. Any non textual information like HTML tags and punctuation are removed from the news articles. Unimportant words are removed from news articles by referring to the list of stop words. Stemming dictionary is used for stemming process that removes word's prefixes and suffixes i.e. it reduces the words to their basic stem.

Preprocessed document is then indexed using a weighting scheme [7]. Each news article is represented by a set of important terms called index terms. These index terms are useful in remembering the news articles' main themes. These index terms also have relevance with each other to represent content within a news article. Weights are assigned to each index term to capture the association of news articles that are similar.

A news article is represented by a vector space model [2] which is used in determining the similarity between two news articles. Each news article is represented by the vector of frequencies of terms contained in the article. But the vector space has very high dimensionality because even after preprocessing, there are still thousands of terms in each news article to deal with. If n is the news article containing m terms, it is represented by the vector of frequencies as:

$$n= (tf_1,\ldots tf_m).$$

To obtain frequent term sets within a news article, Apriori algorithm [3] is used based on support count of frequent terms. Apriori algorithm is very much efficient in mining large number of news articles. Frequent term sets based approach is helpful in reducing the high dimensionality of vector space model because it only considers the low dimensional frequent term sets as cluster candidates. This frequent term set is not a cluster but the description of a cluster.

The Hierarchical frequent term clustering proposed is based on frequent term sets in news articles. It greedily selects the next frequent term set which represents the next news cluster. The result of clustering is based on the order in which frequent term sets are selected which depends on greedy heuristic used. Frequent Item based Hierarchical clustering (FIHC) [1] is proposed for clustering of news article using greedy approach.

III. PROPOSED WORK

Frequent Item based Hierarchical clustering (FIHC) is a cluster centric approach that measures the interdependency among clusters of news articles using frequent term sets. We know that frequent term set is a set of terms that occur together in some minimum fraction of news articles. For example, consider two frequent items, "middle" and "class". "Middle" may refer to mid position and "class" mar refer to class of students. If these two words come together in news articles, it refers to news about

© 2014 IJAIR. All Rights Reserved

middle class people that should be identified. By identifying these terms in news articles and then forming the cluster based on them improves the quality of clustering. After this, clusters are organized in topic hierarchy. The overall Hierarchical News Mining and Clustering process is shown in fig 1 below.

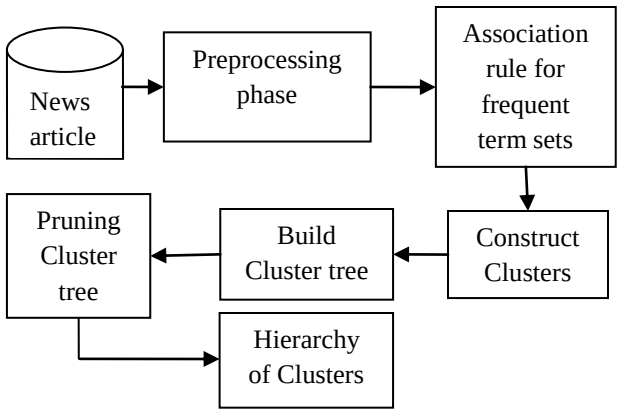

Fig 1 Hierarchical News Mining and Clustering process

The following definitions are introduced:

- Global frequent term set: It is a set of terms that appear together in more than a minimum fraction of the whole news article set.
- Global frequent term: It refers to a term that belongs to some global frequent item set.
- Global frequent k-term set: A global frequent term set containing k terms.
- A global frequent term is cluster frequent in a cluster Ci if the item is contained in some minimum fraction of news article in Ci. FIHC uses only the global frequent items in vector space; thus, the dimensionality is significantly reduced.

Phases in FIHC algorithm are:

1. Constructing Clusters: An initial cluster is created that includes all those news articles containing frequent term set. A news article contains many frequent term sets and thus, the clusters formed are overlapping. Global Frequent term set is used as the cluster label to identify clusters. A best cluster is identified for each news article on the basis of measurement of goodness of cluster. Thus, each article has its exactly one cluster.
2. Building cluster: each cluster has exactly one parent in the hierarchical cluster tree. Clusters uses global frequent term set to identify themselves by cluster label. Tree is formed by choosing the best parent at level k-1 for cluster at level k.
3. Pruning Cluster tree: If two clusters are similar, they are merged into one cluster. If a child cluster is similar to parent cluster, it is replaced by parent cluster. The parent cluster then includes all the article of child cluster replaced.

The algorithm will work as follows:

Algorithm:-

Step 1:- for each Global frequent term set initialize cluster containing term set.

Step 2:- for each news article

a. Check score function of cluster of the news article
b. Assign the best cluster to the news article

Step 3:- Build cluster tree from bottom to top

Step 4:- for each cluster

a. Determine best parent at level k-1 from child cluster at level k

Step 5:- Determine inter cluster similarity,

If child cluster and parent cluster are similar,

i. merge child cluster to parent cluster

Else do not change.

Example: Consider recent news based on cleanup of the coal block allocation process in the country. A news article n contains global frequent terms “Coal blocks”, “Allocation”, and “Scam”. It is assigned to the clusters {Coal blocks}, {Coal blocks, Scam}, {Coal blocks, Allocation}, {Coal blocks, Allocation, Scam} and {Coal blocks, Allocation, CBI}. Let {Coal blocks, Allocation, Scam} is the “best” cluster for article n. n is then removed from clusters {Coal blocks}, {Coal blocks, Scam}, {Coal blocks, Allocation} and {Coal blocks, Allocation, CBI} shown in fig 2.

© 2014 IJAIR. All Rights Reserved

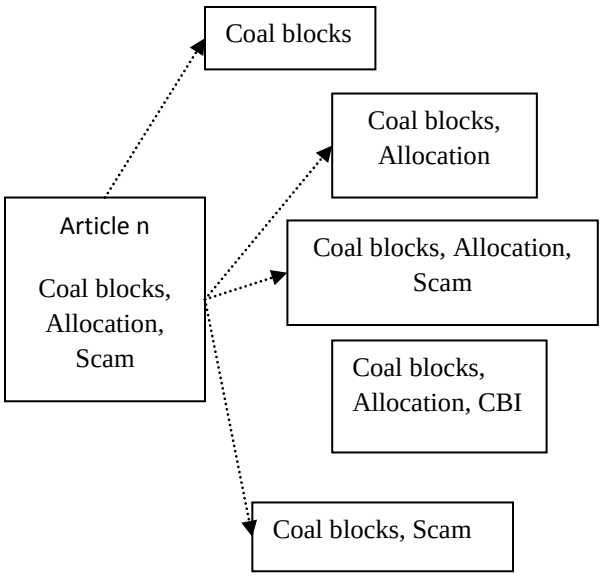

Fig 2 Cluster Construction

Cluster {Coal blocks, Allocation, Scam} has global frequent 3-itemset label. Its parents are {Coal blocks, Scam} and {Coal blocks, Allocation}. Let {Coal blocks, Allocation} has higher score, it becomes the parent of {Coal blocks, Allocation, Scam}.
Let the cluster {Coal blocks, Allocation, Scam}. It is similar to its parent {Coal blocks, Allocation}. {Coal blocks, Allocation, Scam} is pruned and article n is moved up to the cluster {Coal blocks, Allocation} as shown in fig 3.

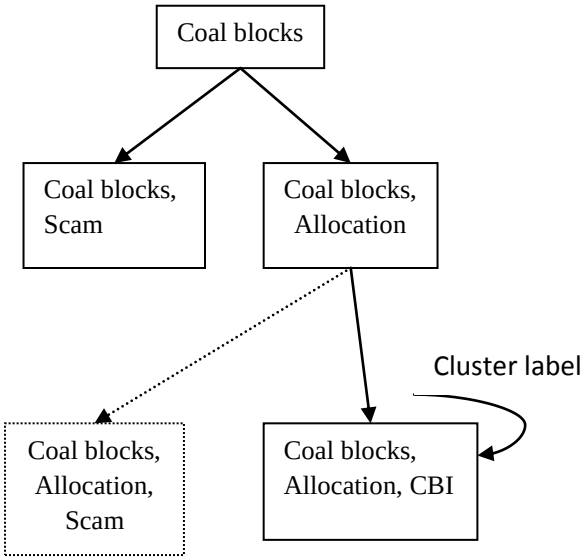

Fig 3 Building Tree and Pruning

IV Conclusion

Most frequent term based clustering methods do not specify requirements of news extraction as high dimensionality and cluster formation as news articles arrive. Due to large amount of unstructured news articles, Frequent Item based Hierarchical clustering(FIHC) is introduced that incrementally updates the cluster tree without rebuilding the entire cluster tree upon arrival of news article. In future we can derive a dynamic approach for the clustering process.

REFERENCES

[1] Benjamin C. M. Fung, Ke Wang, and Martin Ester. "Hierarchical Document Clustering Using Frequent Itemsets", In Proceedings of the 2003 SIAM International Conference on Data Mining (SDM'03), San Francisco, CA, May 1-3, 2003, pp. 59-70.

[2] Florian Beil, Martin Ester, and Xiaowei Xu. “Frequent Term-Based Text Clustering”, In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining New York, NY, USA.

[3] Zhongmin Shi. “Performance Improvement for Frequent Term-based Text Clustering Algorithm”, Technique report on Computing Science, Simon Fraser University(April 2003).

[4] Agrawal, R., Srikant R. Fast Algorithms for Mining Association Rules in Large Databases, In Proceedings of the. VLDB 94, Santiago de Chile, Chile, 1994, pp. 487-499.

[5] Ji-Rui LI, Kai YANG. “News clustering System Based on Text Mining”. In Proceedings of the Advanced Management Science (ICAMS), 2010 IEEE International Conference , July 2010.

[6] Kjetil Nørvag, Randi Øyri. “News Item Extraction for Text Mining inWeb Newspapers”. In Proceedings of the 2005 International Workshop on Challenges in Web Information Retrieval and Integration (WIRI’05).

[7] Hany Mahgoub, Dietmar Rösner, Nabil Ismail and Fawzy Torkey. “A Text Mining Technique Using Association
Rules Extraction”. In proceedings of the International Journal of Information and Mathematical Sciences 4:1 2008.

© 2014 IJAIR. All Rights Reserved