Technical journals online.com

Research Paper

AN EFFICIENT ASSOCIATION RULE BASED HIERARCHICAL ALGORITHM FOR TEXT CLUSTERING

J. Dafni Rose

Address for Correspondence

Associate Professor & HOD, Department of Computer Science and Engineering,
St. Joseph's Institute of Technology, Chennai, India

ABSTRACT

In this modern era, the amount of information available has become too large. But are we getting useful information still remain a question. Text clustering is one of the techniques that helps organize information and hence obtain information in a more efficient manner. This paper presents a new technique for clustering text documents based on association rule based systems. In this approach, the text documents are preprocessed and the association between the text files are found using Apriori algorithm. The associated text files are clustered using hierarchical clustering algorithm. The text files are also clustered using hierarchical algorithm. The results of both the methods are evaluated. The algorithms are tested on benchmark data set Reuters-21578. The experimental results prove that the Association Rule Based Hierarchical clustering method (ARBHC) produce better results and also improved cluster quality over hierarchical method.

KEYWORDS- Text clustering, Association rule, Hierarchical algorithm, Apriori.

I. INTRODUCTION

The increase in the amount of documents in digital libraries, blogs, mails have led to the development of effective and efficient organization of text documents. Text Clustering is a technique that is used to classify texts or passages in natural categories that arise from statistical, lexical, and semantic analysis rather than the arbitrarily pre-determined categories of traditional manual indexing systems. In the context of text mining, it is the derivation of the categories which is of interest, since this is a form of theme finding. Text Clustering algorithms are generally divided into hierarchical methods such as agglomerative, divisive and partition based algorithms such as K-means. Agglomerative algorithms, such as UPGMA [1], single link [2] and Chameleon [3], find the clusters by considering each data as one cluster and then repeatedly merging pairs of clusters until a termination criterion is met, while partitional algorithms, such as *k*-means [4], bisection-k-means [5] and graph-based [6], find the clusters by partitioning the dataset into a number of small clusters. Partitional algorithms are often sensitive to the initial cluster centroids. Efficiency of partition based methods also depends on the number of clusters specified. Moreover, they fail to produce satisfactory clustering results due to the sparsity and high-dimensionality of document datasets. Hierarchical clustering outputs a hierarchy, a structure that is more informative than the unstructured set of clusters returned by flat clustering algorithm such as k-means. Hierarchical clustering does not require the user to pre specify the number of clusters and most hierarchical algorithms that have been used in IR are deterministic. Hierarchical algorithm produces hierarchical solution even though it is inefficient for high dimensional databases. In general, flat clustering such as k-means is selected when efficiency is important and hierarchical clustering is selected when one of the potential problems of flat clustering such as not enough structure, predetermined number of clusters, non-determinism is a concern [7].

Hierarchical algorithm is usually applied on large number of documents. But most of the documents are not really related with each other. So when the hierarchical algorithm is applied on the whole document the performance of the algorithm is reduced. So, in this paper, we propose an improved hierarchical algorithm. In this method the related text documents are found using Apriori algorithm. The remaining documents are removed from the database. This reduces the size of the database to a large extent and the documents that are in the database will be similar. The associated documents alone are then clustered using hierarchical algorithm. This improves the efficiency of the algorithm. The clustered documents are evaluated using cophenetic correlation matrix. Experimental results show that using association rule for clustering outperforms the traditional agglomerative hierarchical algorithm.

The rest of this paper is organized as follows. The next section reviews some related work on document clustering. In section 3, a detailed description of the proposed approach is presented. In section 4, experimental results that evaluate the proposed approach are presented. Finally, the paper is concluded with conclusion.

II. RELATED WORK

In [8], Yehang Zhu proposes a novel hierarchical clustering method which is a hybrid version of both partitioning and agglomerative clustering approaches. This method combines the merits of agglomerative and partition clustering methods. Partitioning clustering is first applied to determine the initial clusters and then hierarchical clustering algorithm is applied to build a hierarchical output. S. S. Bedi [9] presents two new clustering algorithms that cluster documents effectively in high dimensional space. In this paper, the set of items that occur frequently together in transactions are found using association rule discovery methods. The frequent items are then grouped into hyper graph edges and the clusters are found using hyper graph partitioning algorithm. Alisa Kongthon [10], presents a new algorithm called "concept grouping", that adapts an association rule mining technique to construct term thesaurus for data preprocessing purpose. In this paper, similar terms, but written differently, are grouped together into the same concept based on their associations before they are used for subsequent analysis. This technique is used for data preprocessing process. This new Concept grouping algorithm is based on "tree structured networks" to construct thesaurus from related terms. In [11], an improved association rule algorithm is proposed for intelligent QA system. In this work an improved text cluster algorithm, along with the improved association rules algorithm is

proposed. This algorithm classifies the data present in the database accurately, easily locates the learner's question and increases the speed of the QA system.

III. PROPOSED METHOD

In the proposed method, the documents are clustered based on association rule generated by the association algorithm namely Apriori algorithm. The block diagram of the proposed system is depicted in Fig 1.

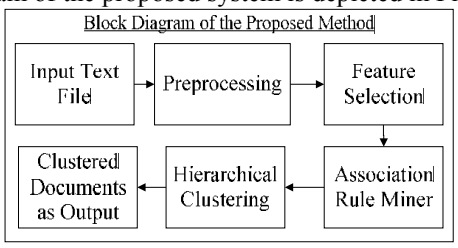

Fig. 1. Block Diagram of the proposed method

A. Preprocessing

In this method, the text documents are first preprocessed. The preprocessing involves stop word removal and stemming. Stop words are defined as common words which have less important meaning than keywords. Documents containing many stop words make the content less important. Stemming is the process of reducing the derived words to their base, stem or root form. As a result of stemming the key terms of the document are represented by stems rather than by words. In this paper, we have used porter stemming algorithm to stem the words.

B. Feature Selection method:

The preprocessed text is then analyzed for feature selection. Even though the traditional methods like term-frequency, bag-of-words, are used in many research works, they do not express the abundant information present in the document. Xiao-Bing Xue and Zhi-Hua Zhou [12] have proposed a new feature selection method. In this paper, distribution of word in a document is explored. It is based on the distributional features like compactness of the appearance of the word and the position of the first appearance of the word.

The distribution of the word is calculated in two steps. First the document is divided into several parts, and then it is entered in an array. The entry in the array corresponds to the count of appearance of that word in the particular part. In this paper, the number of sentences present in the document is related to the number of parts.

Suppose a document d contains n sentences, the word t is represented as distributional array as follows (t, d) = $[c_0, c_1 \ldots c_{n-1}]$. Compactness measures whether the word is constrained in a specific part of a document or spread over the whole document.The compactness (ComPact) of the word t and first appearance (FirstApp) of the word t are defined, respectively, as follows:

$$FirstApp\ (t,d) = \min_{i\in\{0..n-1\}} c_i > 0\ ?\ i : n \quad (1)$$

$$CompPact\ _{PartNum}\ (t,d) = \sum_{i=0}^{n-1} c_i > 0 : 1 : 0 \quad (2)$$

$$LastApp\ (t,d) = \max_{i\in\{0..n-1\}} c_i > 0\ ?\ i : -1 \quad (3)$$

$$ComPact\ _{FLDist}\ (t,d) = LastApp\ (t,d) - FirstApp\ (t,d) \quad (4)$$

$$centroid\ (t,d) = \frac{\sum_{i=0}^{n-1} c_i \times i}{count\ (t,d)}\ count\ (t,d) = \sum_{i=0}^{n-1} c_i \quad (5)$$

$$ComPact\ _{PosVar}\ (t,d) = \frac{\sum_{i=0} c_i \times |\, i - centroid\ (t,d)\,|}{count\ (t,d)} \quad (6)$$

The frequency of a word is calculated using Term Frequency (TF) feature selection method. The compactness of the appearances of a word is calculated, using ComPactness (CP) method. The first appearance of a word is found, using FirstAppearance (FA) method.TF, CP and FA are calculated as follows:

$$TF\ (t,d) = \frac{count\ (t,d)}{size\ (d)} \quad (7)$$

$$CP_{PN}\ (t,d) = \frac{ComPact\ _{ParNum}\ (t,d)}{len\ (d)} \quad (8)$$

$$CP_{FLD}\ (t,d) = \frac{ComPact\ _{FLDist}\ (t,d) + 1}{len\ (d)} \quad (9)$$

$$CP_{PV}\ (t,d) = \frac{ComPact\ _{PosVar}\ (t,d) + 1}{len\ (d)} \quad (10)$$

$$FA(t,d) = f(FirstApp\ (t,d), len(d)) \quad (11)$$

$$f(p, len\ (d) = \frac{len\ (d) - p}{len\ (d)} \quad (12)$$

The importance of each word is calculated as a summation of TF, CP and FA. The words are then ranked according to their values, with the highest value on the top. From this the topmost four words are selected for each text file. All the text files are processed in a similar manner and a final output is obtained.

C. Association Rule Based Hierarchical clustering method:

The features that are selected are passed into association rule miner. The association rule miner uses Apriori algorithm to find the association rules between the text documents. The association algorithm is defined as follows: Let I = $\{i_1, i_2 \ldots i_m\}$ be a set of words. Let D be a set of documents, where each document T contains a set of words [13]. An association rule is an implication of the form X ⇒ Y, where X ⊂ I, Y ⊂ I, and X ∩ Y = φ. The association rule X ⇒ Y holds in the set of documents D with confidence c if c% of transactions in D that contain X also contain Y. The association rule X ⇒ Y has support s if s% of transactions in D contain X∪Y. Mining association rules is to find all association rules that have support and confidence greater than or equal to the user-specified minimum support (called minsup) and minimum confidence (called minconf), respectively [13]**.** The threshold vale for support and confidence lies between 0 to 1. In this work the threshold value is taken as 0.01 as constant for all the documents so that the evaluation criteria remains the same throughout the process.

The documents that are associated with each other are obtained from the association rule miner and the remaining documents are discarded. As a result of this, the database size is reduced to a great extent. This gives the advantage of improved cluster quality and also reduces the time taken for clustering.

The documents that are associated with each other are selected and given as input to the hierarchical algorithm. The hierarchical algorithm performs

agglomerative clustering and the final results are obtained. The clusters that are obtained as a result of ARBHC are found to have better quality than the clusters produced using traditional hierarchical clustering alone.

IV. EXPERIMENTAL RESULTS

The experiments are tested on reuters-21578 dataset. Out of the entire document four Reuters were taken for testing purposes. Each one had 180 documents. The topics include commodity codes, corporate codes, currency codes, energy codes, economic indicator codes and subject codes. The clusters are evaluated using the measures F-measure and Recall. The cluster quality is evaluated using cophenetic correlation coefficient. The time taken to cluster the documents using both the methods is also calculated.

A. Cophenetic correlation Coefficient

This coefficient is used to evaluate the quality of the cluster. The clusters are evaluated and the results are tabulated as below.

TABLE I: CLUSTER QUALITY

	D0	D1	D2	D3
Hier	0.73	-1.52	-0.52	-1.29
ARM	4.13	3.54	0.18	5.17

B. F-measure

F-measure is the weighted harmonic mean of precision and recall. The formula of F-measure is as follows:

$$F = \frac{2.precision\ .recall}{(precision\ + recall\ )} \quad (13)$$

Precision is the fraction of the documents retrieved that are relevant to the user's information need. Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$precision\quad (i,k) = \frac{N_{ik}}{N_k}$$

$$\text{Re}\ call\ (i,k) = \frac{N_{ik}}{NC_i} \quad (14)$$

where N is the total number of documents, i is the number of classes (predefined), K is the number of clusters in unsupervised classification, NCi is the number of documents of class i, NK is the number of documents of cluster CK, Nik is the number of documents of class i in the cluster CK.

TABLE II: F-MEASURE VALUES

	D0	D1	D2	D3
Hier	0.04	0.04	0.045	0.048
ARM	0.29	0.22	0.25	0.28

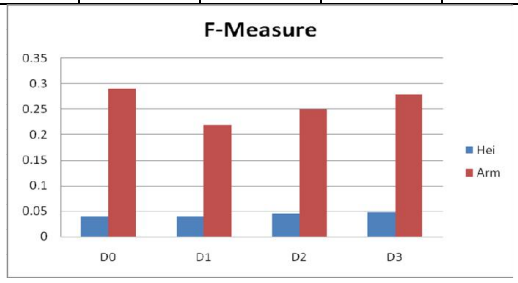

Fig 2. F-Measure

C. Entropy

In information theory, entropy is a measure of the uncertainty associated with a random variable.

TABLE III: ENTROPY VALUES

	D0	D1	D2	D3
Hier	5.66	5.8	6.1	5.9
ARM	3.4	3.5	3.66	3.57

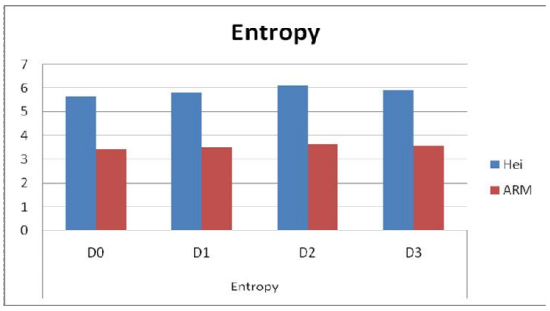

Fig 3.Entropy

D. Time.

The time taken to cluster the documents using the normal hierarchical method and the ARBHC method are calculated and the results are tabulated.

TABLE IV: TIME VALUES

	D0	D1	D2	D3
Hier	178	178	178	178
ARM	54	46	50	38

V. CONCLUSION

In this paper an improved hierarchical clustering algorithm has been developed based on association rules. The results produced as a result of ARBHC method are evaluated using traditional hierarchical algorithm. The results show that the proposed method provides better results that the other. The quality of clusters produced is also measured and is found that the new method produces better clusters than the other. As an extension it is being planned to implement with K-Means algorithm and check the results.

REFERENCES

1. A.K. Jain and R. C. Dubes, Algorithms for Clustering Data. Prentice Hall, 1988.
2. P. H. Sneath and R. R. Sokal, Numerical Taxonomy. Freeman, London, UK, 1973.
3. G. Karypis, E.H. Han, and V. Kumar, "Chameleon: A hierarchical clustering algorithm using dynamic modeling." IEEE Computer, 32(8):68–75, 1999.
4. A.K. Jain and R. C. Dubes," Algorithms for Clustering Data,".Prentice Hall, 1988.
5. E.H. Han, G. Karypis, V. Kumar, and B. Mobasher,"Hypergraph based clustering in high-dimensional data sets: A summary of results," Bulletin of the Technical Committee on Data Engineering, 21(1), 1998.
6. M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," Proc. Text mining workshop, KDD, 2000.
7. http://nlp.stanford.edu/IR-book/html/htmledition/hierarchical-clustering-1.html.
8. An Efficient Hybrid Hierarchical Document Clustering Method, "Fifth International Conference on Fuzzy Systems and Knowledge Discovery, "2008 IEEE DOI 10.1109/FSKD.2008.159
9. S. S. Bedi, Hemant Yadav and Pooja Yadav, "Categorization, Clustering and Association Rule Mining on WWW", IEEE IMPACT2009.
10. Alisa Kongthon Choochart Haruechaiyasak Santipong Thaiprayoon,"Constructing Term Thesaurus using Text Association Rule Mining." in the Proceedings of ECTI-CON 2008.
11. Youfu Du,Ming Zhao,Guanjun Fan ,"Research on Application of Improved Association rules Algorithm in intelligent QA system,", in the proceedings of Second International Conference on Genetic and Evolutionary Computing 2008.
12. Xiao-Bing Xue and Zhi-Hua Zhou, "Distributional Features for Text Categorization," IEEE Transactions on Knowledge and Data Engineering, Apr 2007.
13. R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," *Proc. of the 20th VLDB Conf.*, 1994, pp. 487–499.