Automatic Document Clustering and Knowledge Discovery

Ms. Khureshi Rukhsar Afreen Intizar Ahmed , Mrs. Vishwakarma Pinki

***Abstract*— There always exists a need to organize a large set of documents into categories through clustering, so, document clustering is done such that intra cluster documents are similar to each other than inter cluster documents. Once clusters are formed now reading each document manually to identify what the documents are actually all about is a time consuming task therefore knowledge can be discovered in the form of association rules so that the topic for the cluster can be discovered by just having a glance at the rules. Hence, our paper basically deals with two phases : clustering and knowledge discovery. In the first phase of clustering, we introduce a algorithm where the number of clusters are not to be given as input as in other partitioning algorithms and in second phase of knowledge discovery, we have used GARW algorithm to generate association rules so that we can come to know what the cluster is all about .**

Section 1 is introduction about the topic, section 2 gives details about the literature survey, section 3 is about the proposed system, section 4 is all about the experimental results.

***Index Terms*— About four key words or phrases in alphabetical order, separated by commas.**

I. Introduction

Internet is very important in today’s life. Finding relevant information on the WWW is challenging. To respond to a user query, it is difficult to search through the large number of returned documents with the presence of today’s search engines. There is a need to organize a large set of documents into categories through clustering. Document clustering (or Text clustering) is automatic document organization, and fast information retrieval . To read all the available text documents in text databases and group them effectively and then extract knowledge from them on manual basis is a quite difficult and time consuming task and hence text mining techniques such as clustering are used. After the documents are clustered into various categories , now it’s the time to discover knowledge from them in the form of association rules. Finding association rules in text documents can be useful in a number of contexts . For example, criminal investigations, market trends, intrusion detection , etc. by just having a glance at these association rules we can come to know that the documents of a particular cluster are what all about i.e we can discover a topic for the cluster, and if we are interested in the topic of that cluster we can have a detailed view of those documents without concerning with the documents of other clusters.

II. LITERATURE SURVEY:

Document clustering is the task of arranging a set of documents into clusters so that intra cluster documents are similar to each other than inter cluster documents. There are two common clustering algorithms. 1) Partitioning algorithms in which clusters are computed directly. Clustering is done by iteratively swapping objects or groups of objects between the clusters. 2) The hierarchical based algorithms in which a hierarchy of clusters is build. Every cluster is subdivided into child clusters, which form a partition of their parent cluster. Different clustering algorithms produce different results with different features. Hierarchical algorithms are slower than partitioning algorithms but they give better accuracy. Therefore, a clustering algorithm should be chosen based on the applications, as the desirable features are application dependent.

The author of [1] has compared kmeans algorithm with his proposed algorithm and proved that his algorithm is much better than k means because it shows better results on large number of documents and also resolved zero clustering issue.

The author of [2] has proposed a novel k – nearest neighbor algorithm As the name suggests the method is dependent on the parameter ‘k’ which can drastically change the output as we vary its values. When the training set contains classes of unequal sizes, the test data is likely to get classified to a class which has more samples than the actual class it belongs to, if that actual class has less number of samples. The parameter k in this method depends on the size of the smallest class sample. The proposed algorithm tries to take care of this limitation by first considering the size of the smallest class and then selecting the k nearest neighbors.

The author of [3] has used term frequency –inverse document frequency weighting scheme. Here weighting scheme plays very important role as it helps in removing number of words which are less important leading to the dimensionality reduction.

According to the author of [4], the disadvantage in k-means algorithm is that, the accuracy and efficiency is varied with the choice of initial clustering centers on choosing it randomly. So in his paper, less similarity based clustering method is proposed for finding the better initial centroids and to provide an efficient way of assigning the data points to suitable clusters with reduced time complexity.

Existing algorithms for k-means clustering are slow and do not scale to large number of data points and converge to different local minima based on the initializations. A fast greedy k-means algorithm can attack both these drawbacks, but it is a limitation when the algorithm is used for large number of data points, So the author of [5] introduced an

Manuscript received November 30, 2014.

Ms.Khureshi Rukhsar Afreen Intizar Ahmed, Computer Engineering, Shah and Anchor Kutchchi Engineering College, Mumbai University,Mumbai, India,8286203753.

Mrs.VishwakarmaPinki, Computer Engineering, Shah and Anchor Kutchchi Engineering College, MumbaiUniversity

efficient method to compute the distortion for this algorithm and his experiment results show that the fast greedy algorithm is superior to other methods and can help users to find the relevant documents more easily than by relevance ranking.

The author of [8][9] has introduced EART (extraction of association rules from text) technique , which uses term frequency-inverse document frequency as weighting scheme. It helps in removing the words which are of least importance thereby leading to important rules which helps in knowledge discovery.

The author of [10] has used Association rules discovery techniques to compare the student's performance in the subjects common at Graduation and Post Graduation level and will predict the factors which can explain their success or failure. The mined association rules reveal various factors like student's interest, curriculum design; teaching and assessment methodologies that can affect students who have failed to attain a satisfactory level of performance in the Post Graduation level.

A new semantic-based model that analyzes documents based on their meaning is introduced by the author of [11]. He proposed the model that analyzes terms and their corresponding synonyms and/or hypernyms on the sentence and document levels. In this model, if two documents contain different words and these words are semantically related, the proposed model can measure the semantic-based similarity between the two documents. The similarity

between documents relies on a new semantic-based similarity measure which is applied to the matching concepts between documents.

The author of [12] has discussed the structural and semantic similarity of various XML DTDs for efficient clustering mechanism. As weIl as by analysing the structure of DTDs, the nested and repeated nodes in the tree has been eliminated in the pre-processing mode and efficient clusters has been made based on similarities of XML trees. This study proves group of structurally similar XML documents provides most precise results in large document collections and described an efficient clustering strategy which group most similar DTDs where preliminary evaluation made with the synthetic data.

III. PROPOSED SYSTEM:

Our proposed system consists of 2 phases:

A. *Clustering.*

B. *Knowledge Discovery.*

A. *CLUSTERING:*

Document or Text Clustering is an unsupervised technique in which no input output patterns are pre - defined. The clustering is done in an efficient manner if the documents of intra cluster are more similar than the inter-cluster documents. Clustering differs from categorization as the documents are clustered on the fly instead of having trained datasets.

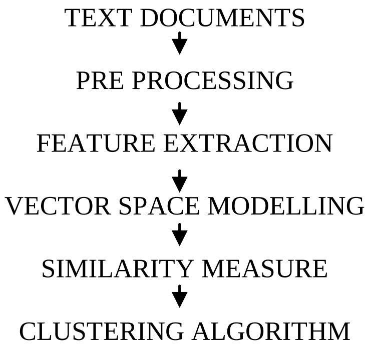


Fig 1: Clustering Phase

a) *Preprocessing :*

Preprocessing steps take as input a plain text document and output a set of tokens to be included in the vector model. The following are the steps of preprocessing [3][4]:

Filtering: Remove special characters and punctuation marks from the plain text document.
Tokenization: Split sentences into individual tokens or words.
Stop word removal: The words (e.g. "and", "the" etc.) which do not convey any meaning as a dimension in the vector space are removed.
Stemming: Reduce the words to their base form, or stem. For example, the words "computer", "computing", are reduced to the stem "compute" using Porter's algorithm.
Pruning: Remove the words with very low frequency in the dataset.

b) *Feature Vector:*

To create Vector space model, a subset of the tokens from the dataset is required which is known as Feature vector. We have used frequency based method for feature vector extraction.

c) *Vector space Model and Similarity Measure:*

It is also known as TF-IDF model i.e. term frequency inverse document frequency model. It is the standard retrieval technique used in text mining area. In this model, each document is represented as an n-dimensional vector using the feature vector. The value of each element in the vector reflects the importance of the corresponding feature in the document. Text documents are converted into machine acceptable, mathematical representations after the transformation. The similarity between documents can be measured by calculating the distance between document vectors [1]. Now that D is a vector => Given a doc, find vectors (docs) "near" it.
Intuition: Documents that are "close together" in vector space talk about the same things.

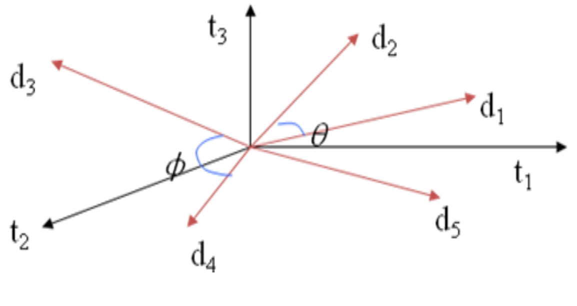


Fig 2 : Vector Space Model

(where, ti= terms and di=documents…..i = 1,2,3,…)
A weighting scheme for document weights is defined by considering the frequency of terms within a document. If the Documents contain the same keywords they are similar Therefore we can use the term frequency tf(i,j) i.e. number of times a term i occurs in a document j. The term frequency is normalized with respect to the maximal frequency of all terms occurring in a document [1].

$$tf\ (i,\ j) = \frac{freq\ (i,j)}{max\ \{f(k,j)\}}$$

We should also consider how frequent a term is in the document collection of size n. The Document frequency (dfi) of a term is the number of documents in which term i occurs. If D is the number of documents in a database then, idf (i, j) = log (D/ dfi)
Term weight is given by: Wi = tfi * log (D/dfi).
Now, the Cosine similarity (the cosine of the angle of two vectors) has to be calculated by the following formula:
Cosine similarity between two vectors x and y is calculated as:

$$\cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Hence, for any two given documents d_j and d_k, their similarity is:

$$sim(d_j, d_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{|\vec{d}_j||\vec{d}_k|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,k}^2}}$$

It has the following properties :
a) If two vectors coincide completely their similarity should be maximal, i.e. 1.
b) If two vectors have no keywords in common, the similarity should be minimal, i.e. 0.
c) In other cases the similarity should be between 0 and 1.

d) Proposed Algorithm:

Input : Document set D = { d1, d2, d3, … , dn}
Output : No of clusters and Set of cluster numbers C along with documents associated.

1. Preprocessing

2. CS ixj =

$$\begin{bmatrix} 1 & d(1,2) & d(1,3) \ldots & d(1,n) \\ d(2,1) & 1 & d(2,3)\ldots & d(2,n) \\ d(3,1) & d(3,2) & 1 \quad \ldots & d(3,n) \\ & . & . & . \\ & . & . & . \\ d(n,1) & d(n,2) & d(n,3) \ldots & 1 \end{bmatrix}$$

3. /* clustering */
   i. Start with document d[i,j] in the matrix with CS value 1, where (i=j).
   ii. Take the mean of row or column the obtained value will be a threshold.
   iii. Now search for documents in the row or column whose value satisfies the above threshold.
   iv. Cluster the documents who satisfy the threshold in the category Ci = {d1,d2, dm,…. | where m is any number between 1 to n}.
   v. For all the documents that are clustered in first category , their CS value will be changed to some other value, and now these documents will not be considered further at any stage to avoid redundancy. Documents with CS value 1 are considered for the formation of the clusters.
   vi. Repeat steps i to v for making more clusters.
3. /* generation of number of clusters */
   i. Initially count =0
   ii. After formation of each cluster count is incremented by 1
   iii. Finally display count
4. Display output in the form c1={d1, d2, d6, d8…..} , and so on.

B. KNOWLEDGE DISCOVERY:

Knowledge discovery from textual database refers generally to the process of extracting interesting information from unstructured text documents. Here, knowledge is discovered in the form of association rules. These rules can be extracted from texts using GARW (Generation of association rules based on weighting scheme) algorithm. These rules would give us the idea regarding the documents of the clusters, and would help us in making correct choice for the topics in which we are interested.

This phase describe a way for finding from a collection of documents by automatically extracting association rules from them . Given a set of keywords A ={ w1 ,w2 ,..., wn} and a collection of indexed documents D = {d1,d2,..., dm }, where each document di is a set of keywords such that di→ A. Let Wi be a set of keywords. A document di is said to contain Wi if and only if Wi → di . An association rule is an implication of the form Wi →Wj where Wi → A , Wj→ A and Wi ∩Wj =φ . There are two important basic measures for association rules, support(s) and confidence(c). The rule Wi →Wj has support s in the collection of documents D if s% of documents in D contain Wi →Wj . The support is calculated by the following formula:
Support (Wi,Wj) =Support count of Wi Wj / Total number of documents D.
and the confidence of a rule is defined as: conf(x→y) = supp(X u Y) / supp(X). [2][3]

An association rule-mining problem is broken into two steps:
1) generate all the keyword combinations (keywordsets) whose support is greater than the user specified minimum support (called minsup). Such sets are called the frequent keywordsets .
2) use the identified frequent keywordsets to generate the rules that satisfy a user specified minimum confidence (called minconf). The frequent keywords generation requires more effort and the rule generation is straightforward.

We have an algorithm for Generating Association Rules based on Weighting scheme (GARW).
The GARW algorithm is as follows:

- Let N be the number of keywords that satisfy the threshold value of weight.
- Store the n keywords in a hashmap along with their frequencies in all document and their weight values TF-IDF.
- Scan the hashmap and find all keywords that satisfy the threshold minimum support.
- Generate candidate keywords from large frequent keywordset..
- Compare the frequencies of candidate keyword sets with minimum support
- Generate different association rules from candidate keywordset , that satisfy the threshold minimum confidence.

The extracted association rules can be reviewed in textual format or tables. The extracted association rules contain important features and describe the informative news included in the documents collection.

The GARW algorithm reduces the execution time in comparable to the Apriori algorithm because it does not make multiple scanning on the original documents like Apriori but it scan only the file or hashmap which contains all the keywords that satisfy the threshold weight value and their frequencies in each document [8]. Apriori-based system generates all frequent keywordsets from all keywords in the documents that are important and unimportant. This leads to extract interesting and uninteresting rules. In contrast, the system based on the GARW algorithm generates all frequent keywordsets from mostly important keywords in the documents based on the weighting scheme. Here, the weighting scheme plays an important role for selecting important keywords for association rules generation. This leads to extract the more interesting rules in short time.

IV. EXPERIMENTAL RESULTS:

For experimental purpose we have used the 20 newsgroup dataset, from that we retrieved some documents and used them for our experiments.

For analysis of results, we applied k means and our proposed algorithm on these dataset. We got the following observations:

Phase 1: Clustering

a) K means:

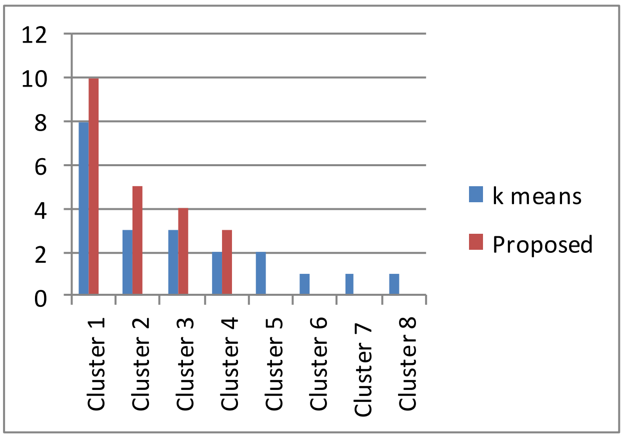
If we give k = 10 , our output consists of 10 clusters whereas maximum clusters are 4 only (according to the documents taken for experiment) , similarly if we give k = 15, 15 clusters will be obtained. Some of the clusters would be singleton(i.e. clusters with just one document) as well.

Therefore, it means time required to make so many clusters and the space required in memory for storing so many clusters is also more.

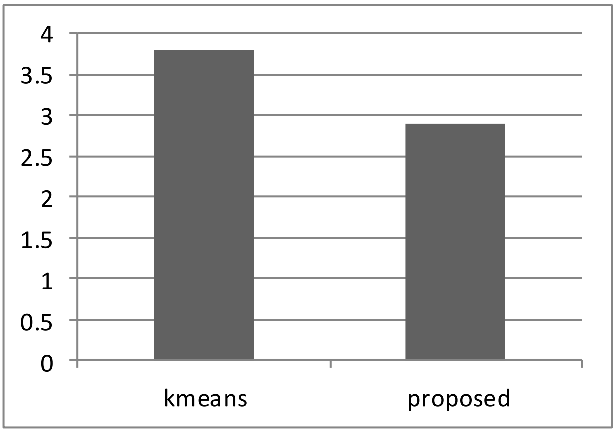
b) Proposed algorithm:

Here, there is no need to specify k (the number of clusters) and we get 4 clusters from the input text dataset.

Therefore, it means time required to make clusters and the space required in memory for storing these clusters is also less.



X axis : cluster no
Y axis : no of documents

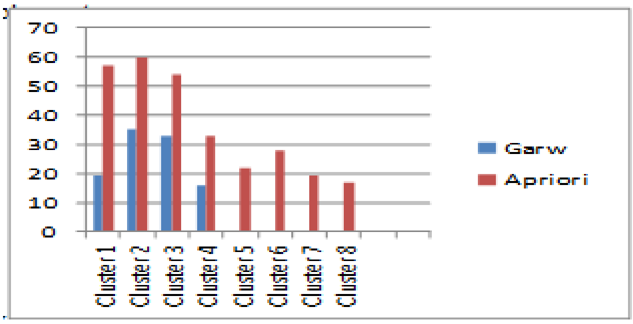


X axis: algorithm type
Y axis : time in seconds for cluster generation

Phase 2: Knowledge Discovery
Here, we have used GARW algorithm for generation of association rules, in this algorithm first the keywords who donot satisfy the threshold weight is removed thus, only important keywords will be remained hence we get important rules.

However,if apriori algorithm is used for generation of rules it would lead to all important and unimportant rules thereby, consuming time and memory as well.

So, on each cluster obtained in phase 1, we apply both these algorithm for analysis purpose and we got the following observations.



X axis : cluster no
Y axis : no of rules generated.

V. CONCLUSION:

Our proposed system automatically groups unknown text dataset in optimum clusters as it does not takes number of clusters as input. It also generates association rules for each cluster so that it would discover the knowledge or the topic about the documents of the clusters and would give the idea to the user about them, so that if the user is interested in that topic he would have a detailed information about that cluster without concerning the other clusters, thereby manual effort and time will be reduced.

REFERENCES

[1] Madhura Phatak , Ranjana Agrawal , "A Novel Algorithm for Automatic Document Clustering" , *2013 3rd IEEE International Advance Computing Conference (IACC)*.

[2] Ms. Anjali Ganesh Jivani, "The Novel k Nearest Neighbor Algorithm" , 2013 International Conference on Computer Communication and Informatics (*ICCCI* -2013).

[3] Ms. Vaishali Bhujade, Prof. N. J. Janwe, Ms. Chhaya Meshram , "Discriminative Features Selection in Text Mining Using TF-IDF Scheme" , International Journal of Computer Trends and Technology- July to Aug Issue 2011.

[4] Manjot Kaur , Navjot Kaur "Web Document Clustering Approaches Using K-Means Algorithm" , International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3, Issue 5, May 2013 ISSN: 2277 128X.

[5] Hongwei Yang , "A Document Clustering Algorithm for Web Search Engine Retrieval System" , 2010 International Conference on e-Education, e-Business, e-Management and e-Learning.

[6] http://pyevolve.sourceforge.net/wordpress/?p=97

[7] http://people.csail.mit.edu/jrennie/20Newsgroup.

[8] Vaishali Bhujade , N.J. janwe , "Knowledge discovery in text mining using association rules extraction" , 2011 IEEE computer society.

[9] Hany Mahgoub, Dietmar Rösner, Nabil Ismail and Fawzy Torkey ," A Text Mining Technique Using Association Rules Extraction" , International Journal of Information and Mathematical Sciences 4:1 2008.

[10] Dr. Varun Kumar1, Anupama Chadha2 , "Mining Association Rules in Student's Assessment Data" , IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.

[11] Shady Shehata , "AWordNet-based Semantic Model for Enhancing Text Clustering" , 2009 IEEE International Conference on Data Mining Workshops .

[12] Mary Posonia A , Dr V LJyothi , "Structural- based Clustering Technique of XML

[13] Documents" 2013 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2013].

Ms.Khureshi Rukhsar Afreen Intizar Ahmed, ME pursuing, Computer Engineering, Shah and Anchor Kutchchi Engineering College, Mumbai University,Mumbai, India.

Assistant Professor Mrs.Vishwakarma Pinki , Computer Engineering, Shah and Anchor Kutchchi Engineering College, MumbaiUniversity