International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 08 | Aug - 2021 ISSN: 2582-3930

Efficient Keyword Based Document Clustering Using Fuzzy C – Means Algorithm

Mr.K.Vivekanandan[1], Mr.R.Karthikeyan[2], Mr.A.Charles Mahimainathan[3]

Assistant Professsor
Department of Computer Technology
Sri Krishna Adithya College of Arts and Science
Kovaipudur, Coimbatore – 641 042.

Abstract

Text Mining has become an important research area. Clustering is an useful technique in the field of textual data mining. Cluster analysis divides objects into meaningful groups based on similarity between objects. The existing clustering approaches face the issues like practical applicability, very less accuracy, more classification time etc. In recent times, inclusion of fuzzy logic in clustering results in better clustering results. In order to further improve the performance of clustering, the Fuzzy C-Means (FCMA) Algorithm is used . The keywords are extracted from the documents using LSA based document extraction. It is responsible for extracting the important keywords from the documents after the preprocessing. Using the Fuzzy clustering technique the documents are clustered using the Fuzzy C-Means Algorithm. The Fuzzy partition matrix is created for the clustering process and the performance of the document clustering is greater based on the keyword when compared to the Existing K-Means Clustering and EM Algorithm. The proposed technique will be highly useful in the text mining process to increase the accuracy and performance of the text extraction process.

Keywords – Fuzzy cluster, k-means clustering, EM algorithm, Fuzzy C-means , text mining, document clustering

INTRODUCTION

Data mining is a process of the knowledge discovery[2] in databases and the goal is to find out the hidden and interesting information . The technology includes the association rules, classification, clustering, and evolution analysis etc. Text Mining[1] is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. Clustering algorithms are used as the essential tools to group analogous patterns and separate outliers according to its principles that elements in the same cluster are homogenous while elements in the different ones are more dissimilar. It is a technique that when they are run, there is not a particular reason for the creation of the models to perform predication. In clustering, there is no particular sense of why certain records are near each other or why they all fall into the same cluster.

Data mining with large data bases the most common challenges are the noisy, imprecise, vague data, therefore by suitable extracting the relevant characteristics of fuzzy sets the data mining techniques can be made more efficient. The use of fuzzy techniques has been considered to be one of the key components of data mining systems because of the affinity with human knowledge representation. In this paper we try to devise an algorithm which makes word-based retrieval more robust. We will investigate how data mining algorithms based on keywords affects retrieval effectiveness in the document. We will try to answer the following research question in this paper "How can the effectiveness of keyword-based document clustering be improved using fuzzy c means algorithm?"

LITERATURE SURVEY

Several attempts were made by researchers to improve the effectiveness and efficiency of the K-means algorithm. Shreya Jain et al.discussed that clustering is a powerful technique for large scale topic discovery from text. It involves two phases: first, feature extraction maps each document or record to a point in a high dimensional space, then clustering algorithms automatically group the points into a hierarchy of clusters. Hence to improve the efficiency & accuracy of mining task on high dimensional data the data must be pre-processed by

an efficient dimensionality reduction method. M.E.S.Mendes Rodrigues, et.al proposed that most of the document clustering has been widely applied in the field of information retrieval for improving search and retrieval performance. Topics that characterize a given knowledge domain are somehow associated with each other. Those topics may also be related to topics of other domains. Hence, documents may contain information that is relevant to different domains to some degree. With fuzzy clustering methods documents are attributed to several clusters simultaneously and thus, useful relationships between domains may be uncovered, which would otherwise be neglected by hard clustering methods

KEYWORD EXTRACTION

we gradually realized that traditional keyword extraction schemes were not necessarily designed to find keywords[6] that are good for doing subsequent clustering. In this connection, it is important to emphasize that the keywords we intend to extract are not merely keywords in the ordinary sense. They should rather be understood as a form of index terms, document categories or "core-terms" for the given document. But we use the term "keywords" since we still think that this is the notion that comes closest in describing the terms we want to extract – terms that both describe the essence/main topic of the document and at the same time are general enough to ensure the overlap with keywords of other documents, necessary for clustering. Our goal is to describe the entire document collection by a relatively small set of good keywords that links related documents together and unlink unrelated documents.

LSA Based Keyword Extraction

Input:

- *A M × N term-document sparse matrix A (terms as rows and documents as columns) with z non-zero elements.*
- *The desired number of keywords per document k.*
- *The desired number of leading singular triplets to retrieve D (LSA-dimension).*

Step 1: Normalize the document vectors (columns) of A by dividing the values of each document vector with its length |aj |:

$$|aj| = \sqrt{\sum_{I=0}^{M} a^2 i, j}$$

j refers to the current column vector (document).
Step 2: Perform (reduced) SVD on sparse matrix A, requesting the D leading singular triplets:

$$A \sim U' \sum' V'^{T}$$

Step 3: Calculate the approximation (A0) to A, one column at a time, by multiplying the returned leading singular triplets together. For each column (document vector), only save the row number and the value of the k largest values. These row number/value pairs represent the found keywords and their "closeness" to the given document (weight).

$$A \sim A' = U' \sum' V'^{T}$$

Returns: A table with k keywords (with corresponding weights) for each document.

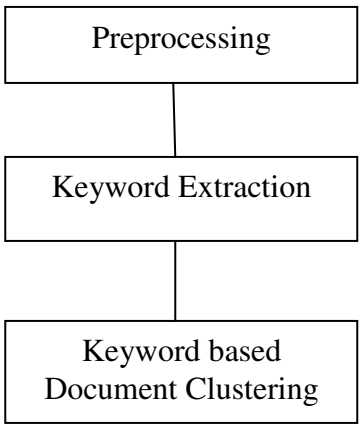

Fig 1. Process Flow in Keyword based Document Clustering

There are two main clustering algorithms are used for the clustering process. They are

- EM Algorithm(EMA)
- K-Means clustering Algorithm(KMA)

The proposed algorithm for the keyword based document clustering is Fuzzy C-Means Algorithm(FCMA).

K means Clustering Algorithm

K means clustering initially takes the number of components of the population equal to the final required number of clusters. In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart. Next, it examines each component in the population and

assigns it to one of the clusters depending on the minimum[7] distance. The centroid's position is recalculated everytime a component is added to the cluster and this continues until all the components are grouped into the final required number of clusters

Step 1 : The Euclidean distance between two multidimensional data points

X=(x1, x2, x3….xm) and Y= (y1, y2, y3…ym) is described as follows:

d(X, Y) =Sqrt(x1-y1)²+(x2-y2)²+…+(Xm-Ym)²(1)

Mean = Sum of elements / Number of elements…… (2) =a1+a2+a3+.....+an/n

Input: D:{d1,d2….dn}\\set of n items K //Number of desired clusters

Output: A set of k-clusters.

Arbitrarily choose k-data items from D as initial centroids

Step 2: Repeat assigns each item di to the cluster which has the closest centroid, Calculate new mean for each cluster; until convergence criteria are met.

EM Algorithm

An expectation–maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes[8] parameters maximizing the expected log-likelihood found on the *E* step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step

Step 1: Guess initial values for the five parameters.

Step 2:Use the probability density function for a normal distribution to compute the cluster probability for each instance. In the case of a single independent variable with mean μ and standard deviation σ, the formula is

$$f(x) = \frac{1}{(\sqrt{2\pi}\sigma)e^{\frac{-(x-\mu)^2}{2\sigma^2}}}$$

In the two-cluster case, we will have the two probability distribution formulas each having differing mean and standard deviation values.

Step 3: Use the probability scores to re-estimate the five parameters.

Step 4: Return to Step 2.

Fuzzy C- Means Algorithm

In fuzzy clustering, each point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster, may be in the cluster to a lesser degree than points in the center of cluster. Any point *x* has a set of coefficients giving the degree of being in the *k*th cluster wk(x). With fuzzy *c*-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster

Documents may contain information that is relevant to different domains to some degree. With fuzzy clustering methods documents are attributed to several clusters simultaneously and thus, useful relationships between domains may be uncovered, which would otherwise be neglected by hard clustering methods.

The Fuzzy C Means algorithm is a generalization of the previously K-Means algorithm. The other difference between Fuzzy C Means and K-Means is that new centers are calculated by a weighted average of the records. Finally, the convergence criterion is relaxed so that the algorithm is permitted to end before there are zero changes.

Step 1: Let us suppose that M-dimensional N data points represented by xi

(i = 1, 2, . . . ,N), are to be clustered.

Step 2: Assume the number of clusters to be made, that is, C, where 2 ≤ C ≤ N.

Step 3: Choose an appropriate level of cluster fuzziness f > 1.

Step 4: Initialize the N × C ×M sized membership matrix U, at random, such

that Uijm ∈ [0, 1] and $\sum_{j=1}^{c} Uijm$ *= 1.0, for each i and a fixed value of m.*

Step 5: Determine the cluster centers CCjm, for jth cluster and its m^{th} *dimension by using the expression given below:*

$$CC_{jm} = \frac{\sum_{i=1}^{N} U_{ijm}^{f} x_{im}}{\sum_{i=1}^{N} U_{ijm}^{f} N}$$

Step 6: Calculate the Euclidean distance between ith data point and jth cluster

center with respect to, say mth dimension like the following:

$$Dijm = ||(xim - CCjm)||.$$

Step 7: Update fuzzy membership matrix U according to Dijm. If Dijm= 0, then the data point coincides with the corresponding data point of jth cluster center CCjm and it has the full membership value, that is, Uijm = 1.0.

Step 8: Repeat from Step 5 to Step 7 until the changes in $U \leq \varphi$, where φ is a pre-specified termination criterion.

EXPERIMENT RESULT

A sample of 8 documents are taken to explain the implementation part. These documents broadly categories to two topics namely Social netpaper (D1 to D4) and Computer netpaper (D5 to D8). Words are assigned weights according to significance in the document which can be the number of occurrences of that word in that document. The implementation is done in MATLAB.

Algorithm	Comparison Aspect			
	RMSE	Accuracy %	Regression Line Slope	Time (in secs)
FCMA	0.44	86	0.56	1.5
KMA	0.53	83	0.54	1.8
EMA	0.46	76	0.49	3.7

Word	WF(Sports)	WF(Politics)
Win	10.02	8.90
Stadium	20.23	7.12
Democracy	1.12	14.12
Ball	13.65	30.21
Team	25.63	10.84
Candidate	38.76	40.23
Campaign	8.83	9.42

The documents that want to cluster into two categories -"sports" and "politics". The first two steps that followed as they are pretty simple. Start from the third – feature selection. There is no idea about which words (features) want to use to cluster our documents.

Take the some documents that are related to Sports and Politics respectively. Then the differences in these WF values for same words between documents relating to Sports category and the documents relating to Politics category.

Table 1. Word Frequencies for sports and politics related documents

PERFORMANCE EVALUATION

The performance of the K-means clustering and EM algorithm is compared with the Fuzzy C-Means clustering algorithm. The performance of the Fuzzy C-Means clustering algorithm is higher when compared to the existing K-Means and EM algorithm.

Performance is calculated by using

- RMSE
- Accuracy
- Regression Line Slope
- Time

Based on the results obtained from the parameters the overall Comparison is given in the table 2. It shows that the performance of the Fuzzy C- Means Algorithm will be better when compared to the K – Means and EM algorithm.

Table 2. Efficiency Comparison of clustering algorithms using keyword based document clustering

The Chart representation for the performance comparison is described below

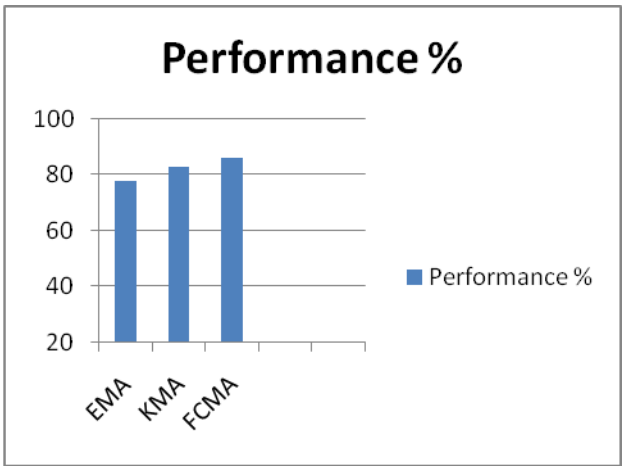

Fig 2. Performance comparison – diagramatic representation

CONCLUSION

Extraction of text is an essential operation. The mining of text is an important process which is performed in many different ways. There have been many text extraction methods developed, this paper presents a novel technique that employs keyword based article clustering to further enhance the text extraction process. Out of those operations clustering is the important technique which is carried out in the data mining process. We have presented the keyword based document clustering process and their performance is analyzed with the help of Fuzzy C-Means, K –Means and EM algorithm.

First the keyword is extracted from the collection of documents and they are used for the clustering process. The two clustering algorithms namely K-Means and EM algorithm are compared with Fuzzy C-Means algorithm and their clustering performance is analyzed. When using the Fuzzy C-Means clustering algorithm for clustering, the performance is greater when compared to K-Means and EM algorithm in the keyword based document clustering. The Performance is measured by using RMSE, Accuracy, Regression Line Slope and Time.

Finally we would conclude that though many algorithms have been proposed for clustering but it is still an open problem and looking at the rate at which the resources is growing, for any application using documents, clustering will become an essential part of the application.

REFERENCES

[1] Vishal Gupta and Gurpreet S. Lehal, “A Survey of Text Mining Techniques and Applications” Journal of Emerging Technologies in Web Intelligence, Vol.1, No.1, August 2009.
[2] Arun K.Pujari, “ Data Mining Techniques”,University Press, First Edition, 2001.
[3] Pavel Berkhin, “Survey of clustering data mining techniques” Technical report, Accrue Software, San Jose, CA, 2002.
[5]Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta,“ A Comparative Study of Various Clustering Algorithms in Data Mining”
[6] Sumit Vashishta,” Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm”.
[7]K.Sathiyakumari, V.Preamsudha and G.Manimekalai, “Unsupervised Approach for Document Clustering Using Modified Fuzzy C mean Algorithm” International Journal of Computer& Organization Trends – Volume1- Issue3 - 2011.
[8]J.A.Hatigan and M.A.Wong, “ K-Means Clustering Algorithm”, -- Applied statistics, 1979.
[9]K means clustering - Wikipedia, the free encyclopedia en.wikipedia.org/wiki/Kmeans_clustering.
[10]Manjot Kaur and 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.
[11]Mrs.Bharati R.Jipkate and Dr. Mrs.V.V.Gohokar, “ A Comparative Analysis of Fuzzy C-Means Clustering and K Means Clustering Algorithms” International Journal Of Computational Engineering Research , March 2011.
[12]EM algorithm - Wikipedia, the free encyclopedia en.wikipedia.org/wiki/EM algorithm
[13] Rahul R.Papalkar,” fuzzy clustering in web text mining and its application in ieee abstract classification”
[14]Sumit Goswami and Mayank Singh Shishodia, “A Fuzzy based Approach to Text Mining and Document Clustering” (IJACSA) International Journal of Advanced Computer Science and Applications, Vol.2, March 2011
[15]Dr. Yogendra Kumar Jain and Sumit Vashishtha, “Efficient Retrieval of Text for Biomedical Domain using Expectation Maximization Algorithm” International Journal of Computer Science Issues (IJCSI), Vol. 8, Issue 6, No 1, November 2011.