Review on Sentence - Level Clustering with Various Fuzzy Clustering Techniques

Christy Maria Joy[1], S. Leela[2]
[1]*PG Student, Computer Science and Engineering, Karunya University*
[2]*Assistant Professor, Computer Science and Engineering, Karunya University*
Tamil Nadu, Coimbatore

Abstract

Clustering plays an important role in the area of data mining. Its application extends to many natural language processing tasks such as text mining, document clustering, discovering the similar groups of linguistic terms, information retrieval etc. There are many available data clustering methods for sentence clustering. But in fuzzy clustering data elements can go to blend in more than one cluster with different cluster membership values, and these values indicate the strength of association between the clusters and data elements. So by incorporating fuzzy clustering techniques in sentence clustering domains can give optimal results than other methods, because a sentence can represent more than one concept within a document. This paper is intended to learn and evaluate various Fuzzy Clustering techniques in the context of sentence- level clustering.

Index Terms *- Fuzzy Clustering, Sentence Level Clustering, FRECCA algorithm, Cluster-Membership, Text Mining.*

I. INTRODUCTION

The tremendous development of information technology in the last two decades paved the way for a world full of data. But most of these data are of potentially not useful. In order to make it useful we need to extract the information or knowledge underlying the data. Data mining is the process of extraction of this core information inside the huge amount of data. Clustering process can help in this data extraction and text analysis. Several clustering algorithms will cluster or group similar data objects and enable us to discern key concepts and thereby summarizing large documents.

The potential application of clustering includes business applications, geology, medical diagnosis, document summarization etc. In general, the text mining process includes the statistical study of terms or phrases which helps us to understand the significance of a word within a document. Even if there are two words with similar frequency, one word will contribute more to the meaning of the sentence.

Clustering text at the sentence level and document level has many differences. Document clustering [1] breakdown the documents into several pieces and cluster those pieces based on the overall theme. It doesn't give much importance to the semantics of each sentence in the document. So there may be content overlap or bad coverage of theme will happen in the case of multidocument summarization. Each data element in hard clustering method belongs to exactly one cluster. But fuzzy clustering allows data elements to fall in all clusters with varying membership values. By analyzing these membership values we can understand the contribution of each data element into a particular cluster. So undoubtedly we can say that incorporating fuzzy clustering into sentence clustering can give optimal results. In fuzzy relational clustering, the data classification problem can be solved by stating a relation that enumerates the similarity, dissimilarity, degree between pair of objects. The purpose of this technical report is to understand how fuzzy relational clustering algorithms work well in sentence clustering domain.

2. FRECCA ALGORITHM

Andrew Skabar et al. proposed the Fuzzy Relational Eigen Vector Centrality Based clustering algorithm (FRECCA) [2], motivated by the mixture model approach in which the data is represented as a combination of components. A graph representation of data objects is used here along with the PageRank algorithm. It operates within an Expectation–Maximization (EM) [3] framework which is general purpose method for learning from incomplete data.

Each sentence in a document is represented by a node in the directed graph and the weighted objects will indicate the object similarity. In order to measure the relative importance of a hyperlinked set of documents PageRank will assigns numerical weighting to each element. And by using this importance we can easily determine the centrality of the graph.

PageRank [4] also assigns a score between 0 and 1 to each node and is then treated as likelihood. Cluster membership values for each node indicate the contribution of each data object into a particular cluster and the mixing coefficients will point out the probability of an object being generated from a component. These two parameters are needed to determine to start on with the FRECCA algorithm and will be optimized by Expectation Maximization. The algorithm proceeds through three main steps. They are Initialization , Expectation, and Maximization. Cluster membership values are chosen randomly and the mixing coefficients are determined in the initialization step. PageRank score for each object is calculated in the Expectation step. It is shown in [2] as follows.

$$PR(V_i) = (1-d) + d \times \sum_{j=1}^{N} \left(w_{ji} \frac{PR(V_j)}{\sum_{n=1}^{N} w_{jn}} \right) \quad (1)$$

Where d is the damping factor which indicates probability of the random surfer get bored and jumping to any other node in the graph. V_i and V_j represents the set of vertices, w_{ji} is the similarity between V_j and V_i. Maximization step includes updating the mixing coefficients based on the Expectation step. The input to the algorithm is pairwise similarities between the sentences and the required number of output clusters. The semantic similarity between sentences can be measured by using cosine similarity. FRECCA algorithm performs significantly better results in identifying overlapping clusters and is not sensitive in cluster membership initialization values. Another striking feature is its ability to converge to an appropriate number of clusters.

3..ARCA ALGORITHM

P.Corsini et al introduced Any Relational Clustering Algorithm (ARCA) [5] which is based on the Fuzzy C-means (FCM) algorithm.ARCA is very stable and it represents clusters with high membership value in terms of the mutual relationship between the objects. If the objects can be shown as points in a multidimensional space, then we can apply the FCM algorithm which is the starting point of ARCA.

In a data set with many objects, the ARCA represent each object or data element by the vector of its relation strength with other data elements. The algorithm doesn't require any particular restriction on the relational matrix. ARCA partitions the dataset with the intent of minimizing the Euclidean distance between each data element in a cluster and the prototype of the cluster. When we represent objects in terms of their similarity with other objects a high dimensionality will be introduced.

The algorithm initially choose a partition $U(0)$ and at the step l, $l=0,1,2$,etc it will calculate the prototype vector $V(l)$ .After that $U(l)$ will be updated to $U(l+1)$ and compare its value in a suitable matrix norm[6]. Depending upon the value of the predetermined threshold, algorithm will either stop or return to a particular point. ARCA is less prone to the dimensionality curse problem and is capable of producing crisper partitions. By complimenting the relationship degree we can convert the initial matrix into dissimilarity matrix, so that it is easy to compare ARCA algorithm with other fuzzy relational clustering algorithms.

4.FUZZY K MEANS ALGORITHM WITH CLUSTER DISPLACEMENT

Chih Tang Chang et al. introduced Fuzzy K –means algorithm with cluster displacement (CDFKM) [7]. This method is capable of reducing the computational complexity and the number of distance calculation in conventional Fuzzy K Means (FKM) algorithm. Like FKM this method also partitions the data points into k clusters. Cluster membership value is also initialized here. There is a fuzzy relationship between cluster representative and a particular data point. By partitioning the data points the FKM process maps a given set of vectors into an improved set. The condition is that no two clusters can have similar cluster representative. The mapping process will end when it reaches the particular stopping criterion. In the cluster displacement FKM method a displacement value is assigned between two cluster centers and if this value is smaller than a particular threshold value, then we can called the cluster centers as stable or active cluster center.

CDFKM method omit the distance calculations for stable clusters in the iterative process When the number of proceeding iterations increases the cluster center number will also increases. By using an algorithm to determine the initial cluster centers for CDFKM, convergence can also speed up. Hamming distance, Euclidean distance etc can be used as distortion measure.

5. CONTRADICTION ANALYSIS

Vasant Kumar Metha et al. pioneered the fuzzy based sentence level document clustering for Micro-level contradiction analysis [8]. This method helps to find out the outlier documents. Document level clustering may often lead to ignorance of some vital information. The algorithm based on this approach begins with the pre-processing operations for data cleansing process such as removal of stem words, stop words etc. And its input is set of documents and number of words in each document.

Sentence similarity is measured for each sentence in the document, and this similarity is measured with every other sentence in the two documents. This content similarity lies within [-1, 1] range. Then separate clusters are generated for similar sentences and contradictory sentences. The sentence

level similarity relationships are then used to infer about the document similarities. It is computed as follows in [8].

$$DS = \frac{1}{n}\sum_{m=0}^{n} CS(Sk, PSk) \qquad (2)$$

Where n is the number of sentences in the document, S represents the sentence and the PS indicate the paired sentences. Paired sentence is the one which is most likely or most contradictory sentence from other documents. This method is best for giving completely correct and completely wrong answer.

6. MEDIAN FUZZY C-MEANS ALGORITHM

Tina Geweniger et al proposed Median Fuzzy C-Means Algorithm (M-FCM) [9], which is a combination of median c-means (M-CM) [10] and fuzzy c-means (FCM) [11]. By keeping the concept of prototype based clustering, median c means also considers the dissimilarity between data points. It offers greater flexibility because of the fuzzy assignment of the data to cluster prototypes. The metric used here is Euclidean distance and it performs a two step iteration using EM algorithm in order to minimize the cost function. In the preceding steps of the algorithm there is an Assignment update and Prototype update step which is similar FCM assignment. To obtain a good performance several runs have to be performed in M-FCM even though it will converge in a finite number of steps. M-FCM bridges the gap between FCM and M-CM. This algorithm can be relevant to non vectorial data. The limitation of this approach is the necessity to store and handle the large dissimilarity matrix.

7. EVALUATION AND COMPARISON OF FUZZY CLUSTERING METHODS

The ultimate performance of clustering techniques depends on the quality of the input data set and the similarity measure used. Good Clustering is regarded as one with high purity and low entropy. Every clustering technique will intend to achieve clusters with high intra-cluster similarity and low inter cluster similarity.

To compute purity, each cluster is allocated to the class which is most repeated in the cluster. Entropy is the measure of how mixed the objects in the cluster [12]. Perfect clustering has purity value related to one. When we applied the famous quotation dataset in the various clustering techniques described above, the FRECCA algorithm showed high purity and low entropy value than others. Even though FRECCA has high time complexity, it is capable of identifying non compact clusters. But the other prototype based algorithms such as ARCA, MFCM can find out only compact clusters. When we compare the clustering with different number of clusters V-measure is more reliable [13].It can be also defined as the harmonic mean of homogeneity and completeness. The other two performance metrics used in clustering are Rand Index [14] and F Measure which is based on a combinatorial approach.

8. CONCLUSION

Fuzzy Clustering techniques have gained great success in many substantive areas. Sentence clustering is one among them. The effective feature selection and proper choice of algorithm will give good clustering of texts. From the study of incorporating various fuzzy clustering techniques in sentence clustering domain, it is clear that optimal clustering with high intra-cluster similarity and low inter cluster similarity can be achieved . Among the different fuzzy clustering techniques FRECCA algorithm is superior to others. It can apply to asymmetric matrices and is not sensitive to the cluster membership value initialization. But the time complexity of FRECCA algorithm is higher than ARCA, M-FCM and contradiction analysis approach. In short we can say that fuzzy clustering techniques are able to overcome the problems in sentence-level clustering, and will generate optimum results

9. REFERENCES

[1] I Dhillon " Feature selection and document clustering. Survey of Text Mining: Clustering, classification and retrieval", Vol. 1 . 74-100 ,2004
[2] Andrew Skabar and Khaled Abdalgader," Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm", IEEE Transactions On Knowledge and Data engineering, vol. 25, 2013
[3] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," J. the Royal Statistical Soc. Series B (Methodological), vol. 39, no. 1, pp. 1-38, 1977.
[4] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," Computer Networks and ISDN Systems, vol. 30, pp. 107-117, 1998
[5] P. Corsini B. Lazzerini F. Marcelloni, "A new fuzzy relational clustering algorithm based on the fuzzy C-means algorithm", Soft Computing, vol. 9, pp. 439-447, 2005.
[6] Windham MP "Numerical classification of proximity datawith assignment measures." J Class 2:157–172,1985
[7] Chih-Tang Chang, Jim Z. C. Lai and Mu-Der Jeng, "A Fuzzy K-means Clustering Algorithm Using Cluster Center Displacement", Journal of information science and engineering 27, 995-1009 ,2011

[8] R. Vasanth Kumar Mehta, B. Sankarasubramaniam, and S. Rajalakshmi "An algorithm for fuzzy-based sentence-level document clustering for micro-level contradiction analysis" ,Proceedings of the International Conference on Advances in Computing, Communications and Informatics , 2012

[9] Tina Geweniger , DietlindZ ulke , Barabara Hammer, and Thomas Villmann" Median fuzzy c-means for clustering dissimilarity data ",Neurocomputing 73, pp.1109–1116,2010.

[10] M.Cottrell, B.Hammer, A.Hasenfu, T.Villmann, Batch and median neural gas, Neural Networks 19.pp.762–771,2006.

[11] J.Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, NewYork,1981.

[12] C.D. Manning, P. Raghavan, and H. Schtze,"Introduction to Information Retrieval. Cambridge Univ. Press", 2008.

[13] A. Rosenberg and J. Hirschberg, "V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure," Proc Empirical Methods in Natural Language Processing (EMNLP '07), pp. 410-420, 2007.

[14] W.M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," Am. Statistical Assoc. J., vol. 66, no. 338, pp. 846-850,1971.

[15] K.Sathishkumar, E.Balamurugan, and D.Kavin , "Sentence Level Clustering Approaches and its Issues in Various Applications", International Journal of Applied Research and Studies, 2278-9480 Volume 2 Issue 9,2013