Review on Sentence-Level Clustering Using Fuzzy Relational Clustering Algorithm

Kamaljit Kaur[1] Shruti Aggarwal[2]
[1]M.Tech Research Scholar [2]Assistant Professor
[1,2]Department of computer science engineering
[1,2]Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India

Abstract— Clustering is a widely studied data mining problem in the text domains. In text processing, clustering the sentence is one of the processes and used within general text mining tasks. Many clustering methods and algorithms are used for clustering the documents at sentence level. In this paper, the sentence level based clustering algorithm is discussed. It is explained that there are the no. of problems in clustering in sentence level and the solutions to overcome these problems. It is related with soft clustering. As in hard clustering methods, pattern belongs to a single cluster, means objects similar to each other are placed in one cluster where dissimilar objects are placed into another one. But fuzzy clustering algorithms allow patterns to belong to all clusters with differing degrees of membership. This is important in case of sentence clustering, since a sentence is likely to be related to more than one theme or topic present within a document or set of documents. This paper presents a novel fuzzy clustering algorithm that operates on relational input data; i.e., data in the form of a square matrix of pair wise similarities between data objects. After, clustering optimized using different algorithms.

Key words: text processing, clustering, EM, FRECCA

I. INTRODUCTION

Clustering is the ability to automatically group similar textual objects together so that one can discover hidden similarity. It plays an important role in many text processing activities. However, sentence clustering can also be used within more general text mining tasks. For example, consider web mining [1], where the specific objective might be to discover some novel information from a set of documents. By clustering the sentences of those documents we can make an estimation that it may contain at least one of the clusters to be closely related to the concepts described by the query terms. Sentence-level clustering technique along with an optimization feature is used to identifying contradictory documents. Contradictory document is a document in which content is contradictory to the theme of a set of documents. Sentence clustering plays an important role in theme-based summarization, which discovers topic themes defined as the clusters of highly related sentences to avoid redundancy and cover more diverse information [2]. As sentences are of short length having limited content the bag-of-words cosine similarity is not suitable. For this, integrated clustering and interactive clustering—both allowing word and document to play an explicit role in sentence clustering. However, some sentences may not be important to eliminate such sentences ranking is done which gives highest ranks to the sentences containing related terms. Sentence having highest rank is selected for summarization. Document summarization is a process of automatically creating a compressed version of a given document that provides useful information to users, and multi-document summarization is to produce a summary delivering the majority of information content from a set of documents about a topic.

Median c-means algorithm is applicable only for metric data. Then new algorithm that is fuzzy relational algorithm is developed with the help of fuzzy C-means for sentence clustering. Fuzzy clustering algorithms allow patterns to belong to all clusters with differing degrees of membership. This is important in domains such as sentence clustering, since a sentence is likely to be related to more than one theme or topic present within a document or set of documents. Fuzzy relational algorithm that is FRECCA that operates on relational input data; i.e., data in the form of a square matrix of pair wise similarities between data objects. It provide significant performance.

II. SENTENCE CLUSTERING USING FUZZY RELATIONAL ALGORITHMS

Clustering of sentence poses various problems as it contains limited no of terms than the clustering of documents. In document clustering document is treated as a data points in a high dimensional vector space in which each dimension corresponds to a unique keyword [3], leading to a rectangular representation in which rows represent documents and columns represent attributes of those documents Since data points lie in a metric space, we can readily apply prototype-based algorithms such as k-Means [4], Isodata [5], Fuzzy c Means (FCM) [6], [7] and the closely related mixture model approach [8], all of which represent clusters in terms of parameters such as means and covariances, and therefore assume a common metric input space. Since pair wise similarities or dissimilarities between data points can readily be calculated from the attribute data using similarity measures such as cosine similarity, we can also apply relational clustering algorithms such as Spectral Clustering [9] and Affinity Propagation [10], which take input data in the form of a square matrix W = {w i,j} (often referred to as the "affinity matrix"). This is applicable only for documents but not for sentences. There are no. of sentence similarity measures these measures do not represent the sentence in vector space but define sentence similarity as some function of inter sentence word-to-word similarities. The above discussed algorithms are prototype based which are not efficient for sentence clustering.

For sentence clustering novel fuzzy relational clustering algorithm is used, which is a graph representation in which nodes represent objects, and weighted edges represent the similarity between objects. Cluster membership values for each node represent the degree to which the object represented by that node belongs to each of the respective clusters, and mixing coefficients represent the probability of an object having been generated from that component. By applying the Page Rank algorithm [11] to

each cluster, and interpreting the Page-Rank score of an object within some cluster as a likelihood, we can then use the Expectation- Maximization (EM) framework [12] to determine the model parameters (i.e., cluster membership values and mixing coefficients). The result is a fuzzy relational clustering algorithm which is generic in nature, and can be applied to any domain in which the relationship between objects is expressed in terms of pair wise similarities. Then its performance is measured against Spectral Clustering, k- Medoids, and ARCA algorithms.

III. Algorithm for Clustering

Page rank is used to measure graph centrality then it is combined with Expectation Maximization framework to construct a complete relational fuzzy clustering algorithm. Since Page Rank centrality can be viewed as a special case of eigenvector centrality [13], we name the algorithm Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm (FRECCA).

IV. Graph-Based Centrality and Page Rank

The basic idea behind the Page Rank [11] algorithm is that the importance of a node within a graph can be determined. Page Rank assigns to every node in a directed graph a numerical score between 0 and 1, known as its Page Rank score (PR). Page Rank can be used more generally to determine the importance of an object in a network. Then the importance of a sentence is calculated that sentences which are similar to a large number of other important sentences are central. Thus, by ranking sentences according to their centrality, the top ranking sentences can then be extracted for the summary.

V. Fuzzy relational clustering

The algorithm uses the Page Rank score of an object within a cluster as a measure of its centrality to that cluster. These Page Rank values are then treated as likelihoods. Since there is no parameterized likelihood function as such, the only parameters that need to be determined are the cluster membership values and mixing coefficients. The algorithm uses Expectation Maximization to optimize these parameters.

A. Initialization

In this cluster membership values are initialized randomly, and normalized such that cluster membership for an object sums to unity over all clusters

B. Expectation step

The E-step calculates the Page Rank value for each object in each cluster. Page Rank values for each cluster are calculated by affinity matrix weights wij obtained by scaling the similarities by their cluster membership values when page ranks are calculated these are treated as likelihoods and used to calculate cluster membership values.

C. Maximization step

Since there is no parameterized likelihood function, the maximization step involves only the single step of updating the mixing coefficients based on membership values calculated in the Expectation Step.

D. Discussion

In case of space complexity **FRECCA** algorithm is not expensive than other spectral clustering or k-means algorithms. But its time complexity is increased. As spectral clustering decompose single eigen value but FRECCA calls page rank on each cluster during each Expectation step[2]. If there are large no of clusters at initial step then there may be the chance of duplicate clusters so it is checked at the completion of each Maximization step. If duplicate clusters are found, membership values are renormalized, and the algorithm is allowed to proceed until a stage at which convergence has been achieved and no duplicate clusters exist. Then damping factor is calculated which affects the fuzziness of clustering.

Like any clustering algorithm, the performance of FRECCA will ultimately depend on the quality of the input data, and in the case of sentence clustering this performance may be improved through development of better sentence similarity measures, which may in turn be based on improved word sense disambiguation.

FRECCA has a number of attractive features:

First, based on empirical observations, it is not sensitive to the initialization of cluster membership values. Second, the algorithm appears to be able to converge to an appropriate number of clusters, even if the number of initial clusters was set very high.

The algorithm can also be applied to asymmetric matrices.

Although we have applied the algorithm to relational data, it can also be applied to attribute data. This might be done by first calculating pair wise distances between pairs of attribute vectors using some suitable distance measure. FRECCA—because it is based on eigenvector centrality—is inherently capable of identifying noncompact clusters

VI. Enhancement

A. Fuzzy C- Means and FRECCA

In case of Fuzzy relational algorithms Fuzzy c-means plays an important role. The FCM technique is more general and useful in case of overlapping clusters. The FCM is based on minimization of an objective function. The algorithm starts with selecting the number of clusters as defined in the problem and initializing the membership matrix U. This matrix contains the membership value for all points for each cluster. The initialization of U is done randomly and the cluster centers are computed using the membership matrix U. The cluster centers are calculated such that the centre is closer to the points having a greater membership value to one cluster. In other words, the membership values act as weights while calculating the centers. Once the cluster centers have been computed, the membership matrix is updated according to the location of the cluster centers. To calculate the new membership value of a point with respect to a particular cluster, the distance of that point from that cluster centre as well as the distance of the point from all other cluster centers is taken into account. The change in membership matrix is computed. If this change is lower than a predefined threshold, then the process is stopped, otherwise, new cluster centers are calculated and membership matrix are updated with respect to the new

cluster centers. The iteration continues till the change in the membership matrix is minimized.

There are certain drawbacks like -Similarity measure in some clustering algorithm can be measured in terms of word co-occurrence may be valid at the document level, the assumed similarity measures does not hold for small-sized text fragments such as sentences, since two sentences may be semantically related despite having few, if any, words in common. Since fuzzy based algorithm provides better performance but due to drawbacks some new methods needed to be implement. Neuro-fuzzy clustering approaches can be used to improve the overall performance of the clustering approaches. T-norm based fuzzy logics can also be cast in the tradition of algebraic logic. t-norm based fuzzy logic is also algebraizable with equivalent algebraic semantics a corresponding Sub variety of resiuated lattices. Moreover, t-norm based fuzzy logics can be seen as a part of the family of sub structural logics. They have two conjunctions (additive ^ and multiplicative &), an additive disjunction _, an implication ! (which satisfies the residuation law with respect to &) and the truth-constant for falsum 0. It works on completeness results, functional representation, proof theory, decidability, computational complexity, arithmetical hierarchy, expanded systems, and game semantics among others. Thus, it has been considered a specific branch of Mathematical Logic called Mathematical Fuzzy Logic, as opposed to the so-called Fuzzy Logic simpliciter. It works with two important constraints. First, since the intended semantics for fuzzy logic systems has been that of the algebras defined over the real unit interval, the so-called standard semantics, the emphasis has been put on completeness results with respect to these semantics. Second, the literature on fuzzy logic systems is usually related to what we may call truth-preserving deductive systems, i.e. many valued logics where the semantically consequence relation is defined as preservation of 1 as the only designated value. There is fuzzy first order and second order algorithms when no order is given means it represent first order.

VII. Conclusion

It is concluded that since sentence clustering is difficult than document clustering. So for this, The FRECCA algorithm was motivated for fuzzy clustering of sentence-level text, and the need for an algorithm which can accomplish this task based on relational input data. The results we have presented show that the algorithm is able to achieve superior performance to benchmark Spectral Clustering and k-Medoids algorithms when externally evaluated in hard clustering mode. It is implemented in soft clustering capable of identifying overlapping clusters of semantically related sentences. When FRECCA is compared with ARCA algorithm it is determined that FRECCA is a generic fuzzy clustering algorithm that can in principle be applied to any relational clustering problem, and application to several non sentence data sets has shown its performance to be comparable to Spectral Clustering and k- Medoid benchmarks. The performance of FRECCA will ultimately depend on the quality of the input data, and in the case of sentence clustering this performance may be improved through development of better sentence similarity measures. It has number of features that are it is based on empirical observations and the algorithm appears to be able to converge to an appropriate number of clusters, even if the number of initial clusters was set very high. Fuzzy c-means can identify only compact clusters but FRECCA—because it is based on eigenvector centrality—is inherently capable of identifying non compact clusters.

References

[1] R. Kosala and H. Blockeel, "Web Mining Research: A Survey," ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, pp. 1-15, 2000.
[2] Andrew Skabar "Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm" IEEE Transactions on Knowledge and Data Engineering, Volume 25, No. 1, January 2013.
[3] G. Salton, Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, 1989.
[4] J.B MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," Proc. Fifth Berkeley Symp. Math. Statistics and Probability, pp. 281-297, 1967.
[5] G. Ball and D. Hall, "A Clustering Technique for Summarizing Multivariate Data," Behavioural Science, vol. 12, pp. 153-155, 1967.
[6] J.C. Dunn, "A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters," J. Cybernetics, vol. 3, no. 3, pp. 32-57, 1973.
[7] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, 1981.
[8] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, seconded. John Wiley & Sons, 2001.
[9] U.V. Luxburg, "A Tutorial on Spectral Clustering," Statistics and Computing, vol. 17, no. 4, pp. 395-416, 2007.
[10] B.J. Frey and D. Dueck, "Clustering by Passing Messages between Data Points," Science, vol. 315, pp. 972-976, 2007.
[11] S. Brin and L. Page, "The Anatomy of a Large-Scale Hyper textual Web Search Engine," Computer Networks and ISDN Systems, vol. 30, pp. 107-117, 1998.
[12] 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.
[13] U. Brandes and T. Erlebach, Network Analysis: Methodological Foundations. Springer, 2005.

