ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 2, February 2014

An Efficient Clustering Sentence-Level Text Using A Novel Hierarchical Fuzzy Relational Clustering Algorithm

K.Jeyalakshmi[1], R.Deepa[2], M.Manjula[3]

Assistant Professor, PG & Research Department of Computer Science, Hindusthan College of Arts & Science, Coimbatore, India[1]

Research Scholar, PG & Research Department of Computer Science, Hindusthan College of Arts & Science, Coimbatore, India[2,3]

Abstract: In comparison with hard and soft clustering methods, in which a pattern belongs to a single cluster, fuzzy clustering algorithms allow patterns to belong to all clusters with differing degrees of membership. In Existing 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. However, the major disadvantage of the Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm (FRECCA) is its time complexity. The FRECCA lies in its ability to identify fuzzy clusters, and if the objective is to perform only hard clustering. This paper presents a novel hierarchical fuzzy relational 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. The algorithm uses a graph representation of the data, and operates in a Fuzzification Degree framework in which the graph centrality of an object in the graph is interpreted as likelihood. Results of applying the algorithm to sentence clustering tasks demonstrate that the algorithm is capable of identifying overlapping clusters of semantically related sentences, and that it is therefore of potential use in a variety of text mining tasks. We also include results of applying the algorithm to benchmark data sets in several other domains.

Keywords: Hierarchical fuzzy relational clustering, Fuzzification Degree, Hard Clustering, Soft Clustering.

I. INTRODUCTION

Sentence clustering plays an important role in many text-processing activities. For example, various authors have argued that incorporating sentence clustering into extractive multi-document summarization helps avoid problems of content overlap, leading to better coverage [1], [2], [3], [4]. However, sentence clustering can also be used within more general text mining tasks. For example, consider web mining [5], where the specific objective might be to discover some novel information from a set of documents initially retrieved in response to some query. By clustering the sentences of those documents we would intuitively expect at least one of the clusters to be closely related to the concepts described by the query terms; however, other clusters may contain information pertaining to the query in some way hitherto unknown to us, and in such a case we would have successfully mined new information.

The goal of text summarization is to present the most important information in a shorter version of the original text while keeping its main content and helps the user to quickly understand large volumes of information. Text summarization addresses both the problem of selecting the most important sections of text and the problem of generating coherent summaries. This process is significantly different from that of human based text summarization since human can capture and relate deep meanings and themes of text documents while automation of such a skill is very difficult to implement. Automatic text summarization researchers since Luhn work [6], they are trying to solve or at least relieve that problem by proposing techniques for generating summaries.

Irrespective of the specific task (e.g., summarization, text mining, etc.), most documents will contain interrelated topics or themes, and many sentences will be related to some degree to a number of these. The work described in this paper is motivated by the belief that successfully being able to capture such fuzzy relationships will lead to an increase in the breadth and scope of problems to which sentence clustering can be applied. However, clustering text at the sentence level poses specific challenges not present when clustering larger segments of text, such as documents. We now highlight some important differences between clustering at these two levels, and examine some existing approaches to fuzzy clustering.

In early classic summarization system, the important summaries were created according to the most frequent words in the text. Luhn created the first summarization system [6] in 1958. Rath et al. [7] in 1961 proposed empirical evidences for difficulties inherent in the notion of ideal summary. Both studies used thematic features such as term frequency, thus they are characterized by surface-level approaches. In the early 1960s, new approaches called entity-level approaches appeared; the first approach of this kind used syntactic analysis [8]. The location features were used in [9], where key phrases are used dealt with three additional components: pragmatic words (cue words, i.e., words would have positive or negative effect on the respective sentence weight like

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 2, February 2014

significant, key idea, or hardly); title and heading words; and structural indicators (sentence location, where the sentences appearing in initial or final of text unit are more significant to include in the summary.

Clustering is an unsupervised method to divide data into disjoint subsets with high intra-cluster similarity and low inter-cluster similarity. Over the past decades, many clustering algorithms have been proposed, including k-means clustering[10], mixture models [10], spectral clustering [11], and maximum margin clustering [12], [13]. Most of these approaches perform hard clustering, i.e., they assign each item to a single cluster. This works well when clustering compact and well-separated groups of data, but in many real-world situations, clusters overlap. Thus, for items that belong to two or more clusters, it may be more appropriate to assign them with gradual memberships to avoid coarse-grained assignments of data [14]. This class of clustering methods is called soft- or fuzzy-clustering.

II. RELATED WORK

A. vector space model

The vector space model has been successful in IR because it is able to adequately capture much of the semantic content of document-level text. This is because documents that are semantically related are likely to contain many words in common, and thus are found to be similar according to popular vector space measures such as cosine similarity, which are based on word co-occurrence [15]. However, while the assumption that (semantic) similarity can be measured in terms of word co-occurrence may be valid at the document level, the assumption 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. To solve this problem, a number of sentence similarity measures have recently been proposed [16].Rather than representing sentences in a common vector space, these measures define sentence similarity as some function of inter-sentence word-to-word similarities, where these similarities are in turn usually derived either from distributional information from some corpora (corpus-based measures), or semantic information represented in external sources such as Word Net [17] (knowledge-based measures).

B.k-Medoids

Like k-Means, methods based on *k*-Medoids are highly sensitive to the initial (random) selection of centroids, and in practice it is often necessary to run the algorithm several times from different initializations. To overcome these problems, the Affinity Propagation, a technique which simultaneously considers all data points as potential centroids (or exemplars). Treating each data point as a node in a network, Affinity Propagation recursively transmits real-valued messages along the edges of the network until a good set of exemplars (and corresponding clusters) emerges. These messages are then updated using simple formulas that minimize an energy function based on a probability model.

B.Fuzzy C-Means

In the FCM algorithm, a data item may belong to more than one cluster with different degrees of membership. To analyzed a several popular robust clustering methods and established the connection between fuzzy set theory and robust statistics. The rough based fuzzy c-means algorithm to arbitrary (non-Euclidean) dissimilarity data. The fuzzy relational data clustering algorithm can handle datasets containing outliers and can deal with all kinds of relational data. Parameters such as the fuzzification degree greatly affect the performance of FCM.

For kernel methods, the key to success are the formation of a suitable kernel function. However, a single kernel that is selected from a predefined group is sometimes insufficient to represent the data. Different features that are chosen for data can result in different similarity measures corresponding to distinct kernels. The combination of multiple kernels from a set of basis kernels has, therefore, gained acceptance as a way to refine the results of single kernel learning.

III. PROPOSED SYSTEM

The proposed system is based on Hierarchical fuzzy relational clustering algorithm. We first describe the use of distance as a general graph centrality measure, and review the objective function, Optimizing Memberships, Optimizing Weights and Hierarchical Fuzzification Degree clustering approach. We then describe how fuzzification can be used within the hierarchical framework to construct a complete relational fuzzy clustering algorithm.

A.Fuzzy Objective function

The objective function of Fuzzy is to classify a data point, cluster centroid has to be closest to the data point of membership for estimating the centroids, and typicality is used for alleviating the undesirable effect of outliers. The function is composed of two expressions:

- The first is the fuzzy function and uses a distance exponent,
- The second is possibilistic function and uses a typical fuzziness weighting exponent; but the two coefficients in the objective function are only used as exhibitor of membership and typicality.

The objective function is to discover nonlinear relationships among data, kernel methods use embedding mappings that map features of the data to new feature spaces.

Given a image dataset, $X = \{x1,\ldots, xn\} \subset R^p$, the original Fuzzy algorithm partitions X into $c$fuzzy subsets by minimizing the following objective function as

$$J(w,U,V) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \, || \, x_k - v_i \, ||^2 \qquad (1)$$

Where c is the number of clusters and selected as a specified value, n the number of data points, u_{ik} the membership of x_k in class i, satisfying the $\sum_{i=1}^{c} u_{ik} = 1$, m

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 2, February 2014

the quantity controlling clustering fuzziness, and V the set of cluster centers or prototypes ($v_i \in R^p$).

B.Optimizing Memberships

The Hierarchical Fuzzy is to find combination weights *w*, memberships *U*, and cluster centers *V*, which minimize the objective function. The first fix the weights and cluster centers to find the optimal memberships. For brevity, we use D_{ic} to denote the distance between data x_i and cluster center v_c,

$$J(w,U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} D_{ic}^{2} \quad (2)$$

Where $D_{ic}^{2} = (v(xi) - vc)^{T}(v(xi) - vc)$, when the weights and cluster centers are fixed, the distances are constants Similar to Fuzzy.

C.Optimizing Weights

The weights *w* and cluster centers *V* are fixed; the optimal memberships *U* can be obtained. Now, let us assume that the memberships are fixed. We seek to derive the optimal centers and weights to combine the kernels. By taking the derivative of *J* (*w,U, V*) in (1) with respect to *vc* and setting it to zero,

$$\frac{\partial J(w,U,V)}{\partial Vc} = -2\sum_{i=1}^{n} u_{ik}^{m}(v(xi) - vc) = 0 \quad (3)$$

The cluster centers are in the kernel induced distance feature space which might be implicit or even have an infinite dimensionality. Therefore, it may be impossible to directly evaluate these centers.

D.Hierarchical Fuzzification Degree

The Hierarchical fuzzy relational clustering uses the probabilistic constraint that the memberships of a data point across classes sum to one. It is useful to creating partitions, the memberships resulting from FRECCA and its derivatives, however, do not always correspond to the intuitive concept of Fuzzification degree of belonging or compatibility. The Hierarchical fuzzy relational clustering algorithm is to minimizing the object function using Gaussian kernel function. Then the updating of memberships use the Gaussian function as the kernel function, and ηi are estimated using,

$$\eta i = K\frac{\sum_{k=1}^{n} u_{ik}^{m} 2(1 - K(xk, vi))}{\sum_{k=1}^{n} u_{ik}^{m}} \quad (4)$$

The fuzzy membership function *uik* is that the edges connecting the inner data points in a cluster may have a larger "degree of belonging" to a cluster than the "peripheral" edges (which, in a sense, reflects a greater "strength of connectivity" between a pair of data points). For instance, the edges (indexed i) connecting the inner point in a cluster (indexed k) are assigned *uik* = 1 whereas the edges linking the boundary points in a cluster have *uik*< 1.

The objective function in the clustering problem becomes more general so that the weights of data points are being taken into account, as follows:

$$H(C) = \sum_{k=1}^{K}\left(\sum_{i=1}^{|ck|} u_{i(j)k}^{m}, k\lambda_k + \sum_{j=1}^{n}\gamma_{jk}(C_k)\right) \quad (5)$$

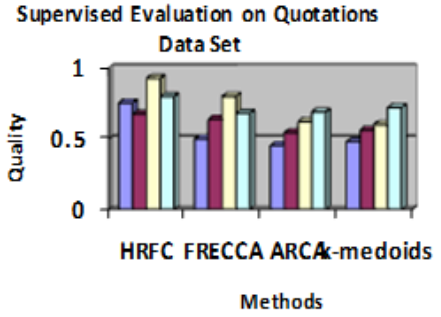
where *C* denotes the decomposition of the given graph *G* into clusters, *C1*, … ,*CK* are not-necessarily disjoint clusters in the decomposition C, H(C) denotes the total strength of connectivity cluster, designates, as in the edge connectivity of cluster, the weight $u_{i(j)}k$, *k* is the membership degree of *i(j)* containing data point *j* in cluster in *k*, and finally, $_{jk}(C_k)$ is the fitness of cluster *j* to cluster *k*.

IV. EXPERIMENTAL RESULTS

Table 1 shows the results of applying the Hierarchical Fuzzy, FRECCA, ARCA and k-Medoids algorithms to the quotations data set and evaluating using the external measures described above. In each case the same affinity matrix was used, with pair-wise similarities calculated as per the method described in Section 3.

Table 1: Supervised Evaluation on Quotations Data Set

N_clust	HRFC	FRECCA	ARCA	k-medoids
3	0.753	0.500	0.452	0.480
4	0.673	0.640	0.543	0.560
5	0.926	0.800	0.622	0.600
6	0.798	0.680	0.690	0.720



Since the four performance measures are not always consistent as to which algorithm achieves best performance for a given number of clusters, we indicate in bold face the value corresponding to the algorithm for which the measure is a maximum. For example, for the quality corresponding to N_clust = 3, HRFC achieves a value of 0.753, which is greater than that achieved by the other algorithms (0.500, 0.452, and 0.480), and hence this value is represented in boldface.

VI. CONCLUSION

In this paper, the HRFC algorithm was motivated by our interest in 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 FRECCA, ARCA Clustering and k-Medoids algorithms when externally evaluated in hard and soft clustering mode on a challenging data set of famous quotations, and applying the algorithm to a recent news article has demonstrated that the algorithm is capable of identifying overlapping

clusters of semantically related sentences. Our main major advantage of the algorithm is its less time complexity.
A Hierarchical Fuzzification degree clustering algorithm is in order to permit overlapping between the obtained clusters. This approach will provide a more flexible use of the mentioned clustering algorithm. We consider that there exist different areas of application for this new clustering algorithm which include not only data analysis but also pattern recognition, spatial databases, production management, etc.

REFERENCES

[1] V. Hatzivassiloglou, J.L. Klavans, M.L. Holcombe, R. Barzilay, M. Kan, and K.R. McKeown, "SIMFINDER: A Flexible Clustering Tool for Summarization," Proc. NAACL Workshop Automatic Summarization, pp. 41-49, 2001.
[2] H. Zha, "Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering," Proc. 25th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 113-120, 2002.
[3] D.R. Radev, H. Jing, M. Stys, and D. Tam, "Centroid-Based Summarization of Multiple Documents," Information Processing and Management: An Int'l J., vol. 40, pp. 919-938, 2004.
[4] R.M. Aliguyev, "A New Sentence Similarity Measure and Sentence Based Extractive Technique for Automatic Text Summarization," Expert Systems with Applications, vol. 36, pp. 7764-7772, 2009.
[5] R. Kosala and H. Blockeel, "Web Mining Research: A Survey," ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, pp. 1-15, 2000.
[6] H. P. Luhn, "The Automatic Creation of Literature Abstracts" IBM Journal of Research and Development, vol. 2, pp.159-165. 1958.
[7] G. J. Rath, A. Resnick, and T. R. Savage, "The formation of abstracts by the selection of sentences" American Documentation, vol. 12, pp.139-143.1961.
[8] Inderjeet Mani and Mark T. Maybury, editors, Advances in automatic text summarization MIT Press. 1999.
[9] H. P. Edmundson., "New methods in automatic extracting" Journal of the Association for Computing Machinery 16 (2). pp.264-285.1969.
[10] R. O. Duda, P. H. Hart, and D. G. Stock, Pattern Classification. New York: Wiley, 2001.
[11] U. von Luxburg, "A tutorial on spectral clustering," Statist. Comput., vol. 17, no. 4, 2007.
[12] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, "Maximum margin clustering," in Proc. Adv. Neural Inf. Process. Syst., 2004, pp. 1537–1544.
[13] K. Zhang, I.W. Tsang, and J. T.Kwok, "Maximum margin clusteringmade practical," in Proc. 24th Int. Conf. Mach. Learning, 2007, pp. 1119–1126.
[14] F.Hoppner, F. Klawonn, R. Kruse, and T. Runkler, Fuzzy Cluster Analysis.New York: Wiley, 1999.
[15] C.D. Manning, P. Raghavan, and H. Schu¨ tze, Introduction to Information Retrieval. Cambridge Univ. Press, 2008.
[16] Y. Li, D. McLean, Z.A. Bandar, J.D. O'Shea, and K. Crockett, "Sentence Similarity Based on Semantic Nets and Corpus Statistics," IEEE Trans. Knowledge and Data Eng., vol. 8, no. 8, pp. 1138-1150, Aug. 2006.
[17] C. Fellbaum, Word Net: An Electronic Lexical Database. MIT Press, 1998.

BIOGRAPHIES

Mrs.K.Jeyalakshmi Pursuing Ph.d in Bharathiar University. Currently she is working as an Assistant Professor of PG & Research Department of Computer Science in Hindusthan college of Arts & Science at Coimbatore. She did her PG degree MCA in Madurai Kamaraj university and also her UG Degree B.Sc (cs) in Bharathiar University. Totally she has 10 years and 8 Months of Experience in Teaching Field.

Ms.R.Deepa, Pursuing Mphil Research Degree in Hindusthan college of Arts & Science at Coimbatore. She did her PG degree MCA in Navarasam Arts & Science College for Women at Erode and also her UG Degree B.Sc (Mathematics) in Bharathidasan University. She had 2 years of Experience as an Assistant Professor in Sasurie College of Arts & Science.

Ms.M.Manjula, Pursuing Mphil Research Degree in Hindusthan college of Arts & Science at Coimbatore. She did her PG degree MCA in Info Institute of Engineering at Coimbatore and also her UG Degree BCA in Bharathiar University. She had 2 years of Experience as a Lecturer in Sasurie College of Engineering.