ISSN:0975-9646

Enhanced Sentence-Level Text Clustering using Semantic Sentence Similarity from Different Aspects

Saranya.J $_{\text{M.Phil}}$[1], Arunpriya.C $_{\text{M.Sc,M.Phil}}$[2]
[1]*Research Scholar, Department of Computer Applications,*
PSGR Krishnammal College for Women, Coimbatore, Tamil nadu, India
[2]*Assistant Professor, Department of Computer Science,*
PSGR Krishnammal College for Women, Coimbatore, Tamil nadu, India

Abstract: **Sentence clustering plays a significant role in many textprocessing activities. For instance, several authors have discussed that integrate sentence clustering into extractive multidocument summarization useful to address issues of content overlap, leading to better coverage. Existing work proposed fuzzy clustering algorithm which is used for relational input data. This existing algorithm uses a graph representation of the data, and performs based on Expectation-Maximization framework. Proposed system improves the result of the clustering by introducing the novel sentence similarity technique. In our proposed system we are propose a new way to determine sentence similarities from different aspects. Probably based on information people can obtain from a sentence, which is *objects* the sentence describes, *properties* of these *objects* and *behaviors* of these *objects*. Four aspects, *Objects-Specified Similarity*, *Objects-Property Similarity*, *Objects-Behavior Similarity* and *Overall Similarity* are calculated to estimate the sentence similarities. First, for each sentence, all nouns in noun phrases are chosen as the *objects specified* in the sentence, all adjectives and adverbs in noun phrases as the *objects properties* and all verb phrases as the *objects behaviors*. Then, the four similarities are calculated based on a semantic vector method. We also conducted an experimental study with that could help us to efficiently clustering the sentence level text. Our study shows that this algorithm generates better quality clusters than traditional algorithms; in other words, it is benefits to increase the accuracy of the clustering result.**

Keywords: **Sentence level clustering, Fuzzy relational clustering, Sentence Similarity and Objects based similarity.**

I. INTRODUCTION

In many text processing activities, Sentence clustering plays a significant role in many textprocessing activities. For instance, several authors have discussed that integrate sentence clustering into extractive multidocument summarization useful to address issues of content overlap, leading to better coverage [1], [2], [3], [4]. On the other hand, sentence clustering can also be used within more general text mining tasks. For instance, regard as web mining [5], where the particular goal might be to find out some novel information from a set of documents primarily recovered 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; though, 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. Irrespective of the specific task (e.g., summarization, text mining, etc.), a large amount documents will hold interconnected topics or themes, and numerous sentences will be related to some degree to a number of these. Nevertheless, clustering text at the sentence level poses specific challenges not present when clustering larger segments of text, such as documents. We now underline some main differences between clustering at these two levels, and analyze some existing methods to fuzzy clustering.

Clustering text at the document level is well established in the Information Retrieval (IR) literature, where documents are typically represented as data points in a highdimensional vector space in which each dimension corresponds to a unique keyword [6], leading to a rectangular representation in which rows represent documents and columns represent attributes of those documents (e.g., tf-idf values of the keywords). This kind of data, which we refer to as “attribute data,” is agreeable to clustering by a large range of techniques.

Given that data points lie in a metric space, we can eagerly perform prototype-based approaches such as k-Means [7], Isodata [8], Fuzzy c-Means (FCM) [9], [10] and the closely related mixture model methodh [11], all of which stand for clusters in terms of parameters, for instance means and covariances, and consequently assume a general metric input space. Because pairwise similarities or dissimilarities between data points can willingly be estimated from the attribute data using similarity approaches like that cosine similarity, we can also use relational clustering techniques such as Spectral Clustering [12] and Affinity Propagation [13], which take input data in the form of a square matrix $W = \{w_{ij}\}$ (often referred to as the “affinity matrix”), where w_{ij} is the (pairwise) relationship between the ith and jth data object.

II. RELATED WORKS

In early traditional summarization system, the significant summaries were created based on the most frequent words in the text. Luhn created the first summarization work [14] in 1958. Rath et al. [15] in 1961 proposed experimental proof for complexity inherent in the notion of ideal summary. Both systems used thematic features such as term frequency; therefore they are illustrated by surface-level techniques. In the early 1960s, new schemes known as entity-level methods appeared; the first technique used syntactic investigation [16]. The location features were used in [17], where key phrases are used dealt with three supplementary components: pragmatic words (cue words, i.e., words would have positive or negative effect on the particular sentence weight like important, 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 approach to categorize data into disjoint subsets with high intra-cluster similarity and low inter-cluster similarity. Recent years, many clustering methods have been proposed, containing kmeans clustering[18], mixture models [18], spectral clustering [19], and maximum margin clustering [20], [21]. Most of these techniques carry out hard clustering, i.e., they give each item to a single cluster. This works better when clustering compact and well-separated groups of data, however in a lot of real-world situations, clusters overlap. Consequently, for items that be owned by two or more clusters, it may be more suitable to assign them with gradual memberships to prevent coarse-grained assignments of data [22]. This class of clustering techniques is called soft- or fuzzy-clustering.

III. Proposed Research Methodology

To bring the sentence semantic meaning more accurately, at the present time more and more applications need not only evaluating the overall similarity between sentences however also the similarity between parts of these sentences. In daily life, people can estimate sentence meaning from various aspects. For two sentences, “Young people like running.” “Old people like walking.” From the common meaning, both sentences say that people like exercises, which states that a strong similarity. However considering subjects and objects, there exists an important difference that different people prefer variouss exercises. To reproduce human’s comprehension to sentence meaning and make sentence similarity comparison more significant, we propose to measure sentence similarities from various aspects.

Owing to the complexity of natural languages, only the minority types of sentences in text have all the three components of subject, predicate verb and object with normal orders, numerous compound and short sentences exist with absent or complemental components, or reversed order. In natural language processing, people regularly use parsing to discover the detailed information in sentences. Presently, the cost of parsing is expensive in time and resources, and the accuracy always proves disappointed. So except those purposes that really require to compare similarities between the subjects, predicate verbs, objects or other components in sentences, it is much inefficient and even impractical to compare sentence similarities according to their fully parsed trees. We propose our sentence similarity definitions, which make the calculating process more be similar to the human’s comprehension to sentence meanings and offer a more levelheaded result in sentence similarity comparison.

Chunking, which is also called as shallow parsing, is a natural language processing approach that attempts to offer a sentence structure which machine can understand. A chunker splits a sentence into series of words that compile a grammatical unit (mostly noun, verb, or preposition phrase). It is an easier natural language processing task than parsing. With the intension of determining the information in sentences that we require to estimate the above four similarities, we chunk each sentence and extract all noun phrases and verb phrases. Then we pick all nouns in noun phrases as the *objects specified* in the sentence, all adjectives and adverbs in noun phrases as the *objects properties* and all verb phrases as the *objects behaviors*.

Generally, people acquire information from a sentence on three aspects, or some of them: *objects* the sentence describes, *properties* of these *objects* and *behaviors* of these *objects*. Here, we try to estimate the sentence similarities from those three aspects. We define *Objects-Specified Similarity* to express the similarity between the *objects* which the two sentences explain; *Objects-Property Similarity* to show the similarity between the *objects properties* of the two sentences; and *Objects-Behavior Similarity* to express the similarity between the *objects behaviors*. After that, we are calculating the *Overall Similarity* to describe the overall similarity of the two sentences, which is defined as the summation of the above three.

A. Objects-Specified Similarity

First, we map all nouns (*objects specified*) which is extracted from noun phrases of a sentence into an *objects specified* vector, which is abstractly similar to a representative vector space demonstration used in a standard IR method, however it only analyzes the nouns from noun phrases of the two compared sentences as the feature set instead of employing all indexed terms in the corpus. Each entry in the vector is derived from calculating the word similarity. After that, the maximum score from the matching words that exceeds certain similarity threshold θ will be chosen. Secondly, the similarity between *objects specified* of two sentences is described from the cosine coefficient between the two vectors. It is defined as,

$$Sim_{os} = \frac{v_{os1} \cdot v_{os2}}{\|v_{os1}\| \cdot \|v_{os2}\|} \qquad (1)$$

Where, Sim_{os} is similarity between *objects specified* of two sentences, v_{os1} is *objects specified* vector s1 and v_{os2} is *objects specified* vector s2.

B. Objects-Property Similarity

First, we map all adjectives and adverbs (*objects property*) which is extracted from noun phrases of a sentence into an *objects specified* vector, which is abstractly similar to a representative vector space demonstration used in a standard IR method, however it only analyzes the adjectives and adverbs from noun phrases of the two compared sentences as the feature set instead of employing all indexed terms in the corpus. Each entry in the vector is derived from calculating the word similarity. After that, the maximum score from the matching words that exceeds certain similarity threshold θ will be chosen. Secondly, the similarity between *objects property* of two sentences is described from the cosine coefficient between the two vectors. It is defined as,

$$Sim_{op} = \frac{v_{op1} \cdot v_{op2}}{\|v_{op1}\| \cdot \|v_{op2}\|} \qquad (2)$$

Where, Sim_{os} is similarity between *objects property* of two sentences, v_{op1} is *objects property* vector s1 and v_{op2} is *objects property* vector s2.

C. Objects-Behavior Similarity

First, we map all verb phrases (*objects behavior*) of a sentence into an *objects behavior* vector, which is abstractly similar to a representative vector space demonstration used in a standard IR method, however it only analyzes the verb phrases of the two compared sentences as the feature set instead of employing all indexed terms in the corpus. Each entry in the vector is derived from calculating the word similarity. After that, the maximum score from the matching words that exceeds certain similarity threshold θ will be chosen. Secondly, the similarity between *objects behavior* of two sentences is described from the cosine coefficient between the two vectors. It is defined as,

$$Sim_{ob} = \frac{v_{ob1} \cdot v_{ob2}}{\|v_{ob1}\| \cdot \|v_{ob2}\|} \qquad (3)$$

Where, Sim_{ob} is similarity between *objects behavior* of two sentences, v_{ob1} is *objects behavior* vector s1 and v_{ob2} is *objects behavior* vector s2.

IV. Experimental Results

We analyze and compare the performance offered by fuzzy relational clustering method and clustering with objects based sentence similarity. The performance is evaluated by the parameters such as accuracy, f-measure, Purity and Entropy, runtime and computational cost. Based on the comparison and the results from the experiment show the proposed approach works better than the existing system.

A. Accuracy

Accuracy can be calculated from formula given as follows

$$\text{Accuracy} = \frac{\text{True positive + True negative}}{\text{True positive + True negative+False positive+ False negative}} \qquad (4)$$

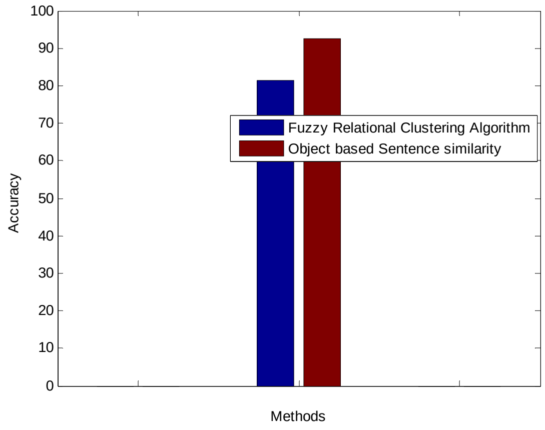

Fig.1. Accuracy comparison

This graph shows the accuracy rate of existing fuzzy relational clustering method and proposed clustering with objects based sentence similarity based on two parameters of accuracy and methods such as existing and proposed system. From the graph we can see that, accuracy of the system is reduced somewhat in existing system than the proposed system. From this graph we can say that the accuracy of proposed system is increased which will be the best one.

B. F-measure comparison

F-measure distinguishes the correct classification of document labels within different classes. In essence, it assesses the effectiveness of the algorithm on a single class, and the higher it is, the better is the clustering. It is defined as follows:

F=2.precision.recall/precision+recall (5)

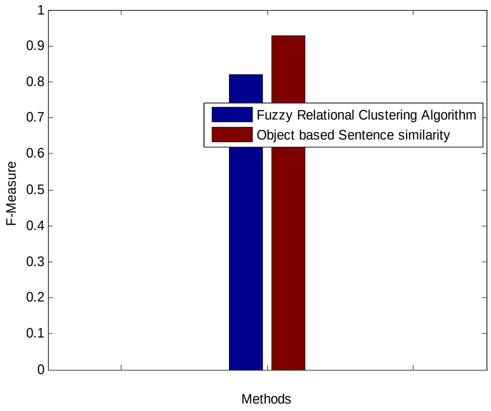

Fig.2. F-measure comparison

In this section, we compare the F-measure parameter between existing fuzzy relational clustering method and proposed clustering with objects based sentence similarity. It is mathematically calculated by using formula. As usual in the graph X-axis will be methods such as existing and proposed system and Y-axis will be F-measure rate. From view of this F-measure comparison graph we obtain conclude as the proposed algorithm has more effective in F-measure performance compare to existing system.

C. Purity

The purity of a cluster is defined as the fraction of the cluster size that the largest class of objects assigned to that cluster represents; thus, the purity of cluster j is

$$P_j = \frac{1}{|w_j|} \max_i (|w_j \cap c_i|) \qquad (6)$$

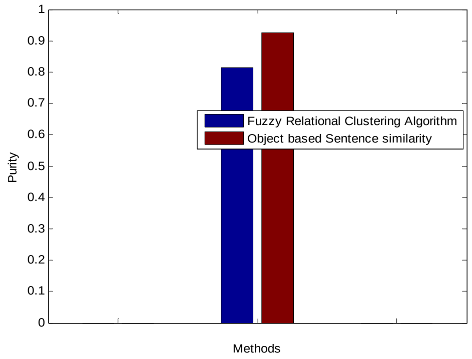

Fig.3. Purity comparison

In this section, we compare the purity parameter between existing fuzzy relational clustering method and proposed clustering with objects based sentence similarity. It is mathematically calculated by using formula. As usual in the graph X-axis will be methods such as existing and proposed system and Y-axis will be purity rate. From view of this purity comparison graph we obtain conclude as the proposed algorithm has more effective in purity performance compare to existing system.

D. Entropy

The entropy of a cluster j is a measure of how mixed the objects within the cluster are, and is defined as

$$E_j = \frac{1}{\log|C|}\sum_{i=1}^{|C|}\frac{|w_j \cap c_i|}{|w_j|}\,log\,\frac{|w_j \cap c_i|}{|w_j|} \qquad (7)$$

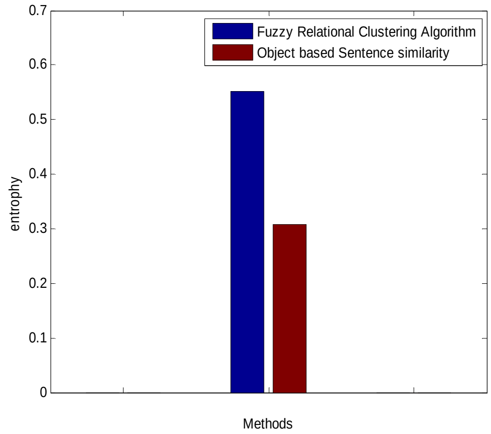

Fig.4. Entropy comparison

In this section, we compare the entropy parameter between existing fuzzy relational clustering method and proposed clustering with objects based sentence similarity. It is mathematically calculated by using formula. As usual in the graph X-axis will be methods such as existing and proposed system and Y-axis will be entropy rate. From view of this entropy comparison graph we obtain conclude as the proposed algorithm has more effective in entropy performance compare to existing system.

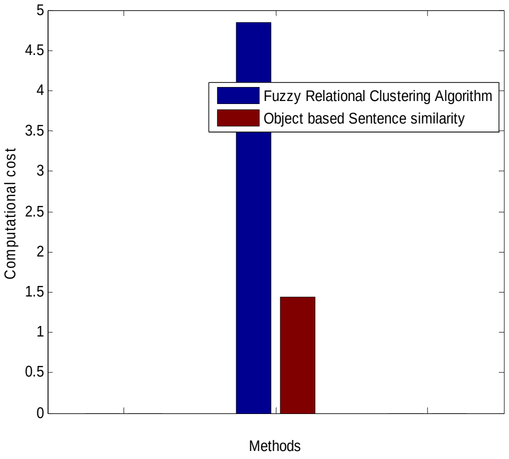

Fig.5. Computational cost comparison

In this section, we compare the computational cost parameter between existing fuzzy relational clustering method and proposed clustering with objects based sentence similarity. As usual in the graph X-axis will be methods such as existing and proposed system and Y-axis will be computational cost rate. From view of this computational cost comparison graph we obtain conclude as the proposed algorithm has more effective in computational cost performance compare to existing system.

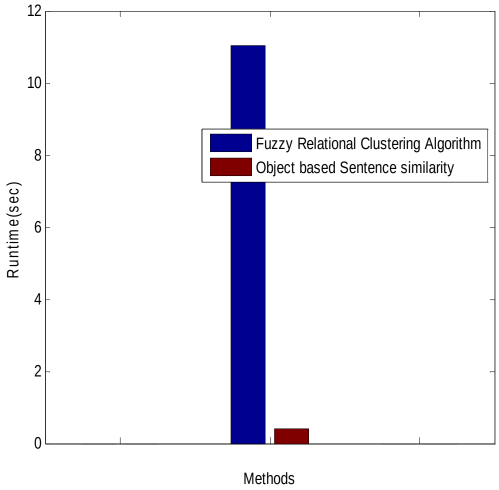

Fig.6. Runtime comparison

In this section, we compare the run time parameter between existing fuzzy relational clustering method and proposed clustering with objects based sentence similarity. As usual in the graph X-axis will be methods such as existing and proposed system and Y-axis will be run time in seconds. From view of this run time comparison graph we obtain conclude as the proposed algorithm has more effective in run time performance compare to existing system.

V. CONCLUSION

Existing work proposed fuzzy clustering algorithm which is used for relational input data. This existing algorithm uses a graph representation of the data, and performs based on Expectation-Maximization framework. Proposed system improves the result of the clustering by introducing the novel sentence similarity technique. In our proposed system we are propose a new way to determine sentence similarities from different aspects. In our proposed system, we are proposing the objects in sentence based sentence similarity. Probably based on information people can obtain from a sentence, which is *objects* the sentence describes, *properties* of these *objects* and *behaviors* of these *objects*. Four aspects, *Objects-Specified Similarity*, *Objects-Property Similarity*, *Objects-Behavior Similarity* and *Overall Similarity* are calculated to estimate the sentence similarities are proposed in our proposed work. Experiments show that the proposed clustering approach makes the sentence similarity comparison more spontaneous and provide a more reasonable result, which imitates the people's knowledge to the meanings of the sentences. Our main future plan is to extend these proposals to the improvement of a probabilistic based fuzzy relational clustering algorithm.

REFERENCES

[1] 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] G. Salton, Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, 1989.

[7] J.B MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," Proc. Fifth Berkeley Symp. Math. Statistics and Probability, pp. 281-297, 1967.

[8] G. Ball and D. Hall, "A Clustering Technique for Summarizing Multivariate Data," Behavioural Science, vol. 12, pp. 153-155, 1967.

[9] 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.

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

[11] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, second ed. John Wiley & Sons, 2001.

[12] U.V. Luxburg, "A Tutorial on Spectral Clustering," Statistics and Computing, vol. 17, no. 4, pp. 395-416, 2007.

[13] B.J. Frey and D. Dueck, "Clustering by Passing Messages between Data Points," Science, vol. 315, pp. 972-976, 2007.

[14] H. P. Luhn, "The Automatic Creation of Literature Abstracts" IBM Journal of Research and Development, vol. 2, pp.159-165. 1958.

[15] 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.

[16] Inderjeet Mani and Mark T. Maybury, editors, Advances in automatic text summarization MIT Press. 1999.

[17] H. P. Edmundson., "New methods in automatic extracting" Journal of the Association for Computing Machinery 16 (2). pp.264-285.1969.

[18] R. O. Duda, P. H. Hart, and D. G. Stock, Pattern Classification. New York: Wiley, 2001.

[19] U. von Luxburg, "A tutorial on spectral clustering," Statist. Comput., vol. 17, no. 4, 2007.

[20] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, "Maximum margin clustering," in Proc. Adv. Neural Inf. Process. Syst., 2004, pp. 1537–1544.

[21] K. Zhang, I.W. Tsang, and J. T.Kwok, "Maximum margin clusteringmade practical," in Proc. 24th Int. Conf. Mach. Learning, 2007, pp. 1119–1126.

[22] F.Hoppner, F. Klawonn, R. Kruse, and T. Runkler, Fuzzy Cluster Analysis.New York: Wiley, 1999.