Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 12, December 2014, pg.11 – 15

RESEARCH ARTICLE

Clustering Sentence-Level Text Using a Hierarchical Fuzzy Relational Clustering Algorithm

Deepika U. Shevatkar

Department of Computer Engineering, Pune University, India

Email: dipikaus@gmail.com

V.K.Bhusari

Department of Computer Engineering, Pune University, India

Email: vrundabhusari82@gmail.com

Abstract— In comparison with hard 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. 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. However, because most sentence similarity measures do not represent sentences in a common metric space, conventional fuzzy clustering approaches based on prototypes or mixtures of Gaussians are generally not applicable to sentence clustering. 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. The algorithm uses a graph representation of the data, and operates in an Expectation-Maximization 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— Fuzzy relational clustering, natural language processing, graph centrality

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. However, sentence clustering can also be used within more General text mining tasks. For example, consider web mining, 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. 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.

II. EXISTING WORK

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 high dimensional vector space in which each dimension corresponds to a unique keyword, 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).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. 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.

2.1. DISADVANTAGES OF EXISTING WORK

The results often suffered from instability in the optimization algorithms that were used.

A limitation of existing approach is the high dimensionality introduced by representing objects in terms of their similarity with all other objects.

III. PROPOSED WORK

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 pairwise similarities between data objects. The algorithm uses a graph representation of the data, and operates in an Expectation-Maximization framework in which the graph centrality of an object

in the graph is interpreted as a 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

3.1. ADVANTAGES OF PROPOSED WORK:

Able to achieve superior performance to benchmark Spectral Clustering and k-Medoids algorithms when externally evaluated in hard 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.

IV. IMPLEMENTATION

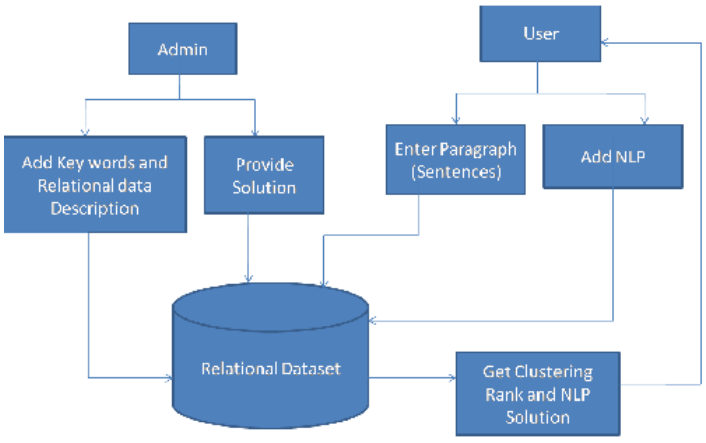
Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

4.1. Graph-Based Centrality and Page Rank

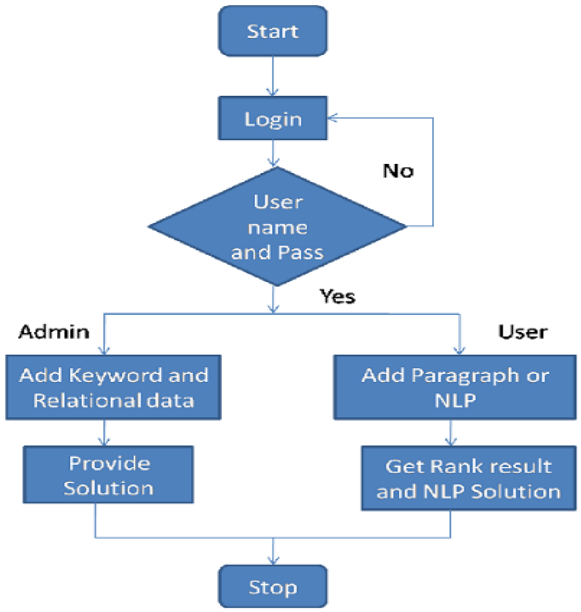
The basic idea behind the Page Rank algorithm is that the importance of a node within a graph can be determined by taking into account global information recursively computed from the entire graph, with connections to high-scoring nodes contributing more to the score of a node than connections to low-scoring nodes. It is this importance that can then be used as a measure of centrality.

4.2.Fuzzy Relational Clustering

Unlike Gaussian mixture models, which use a likelihood function parameterized by the means and covariances of the mixture components, the proposed 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.



Comparisons with the ARCA algorithm on each of these data sets suggest that FRECCA is capable of identifying softer clusters than ARCA, without sacrificing performance as evaluated by external measures.



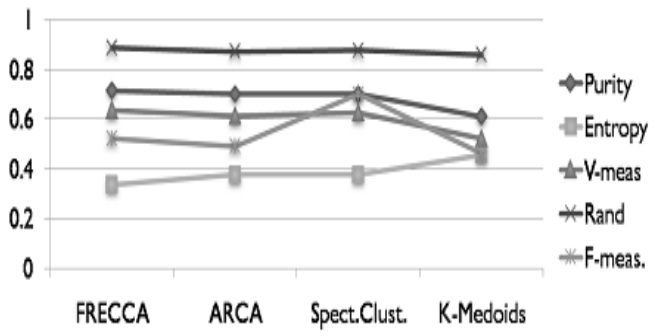
V. RESULT

Clustering performance of FRECCA with that of ARCA, Spectral Clustering, and k-Medoids using the five cluster quality measures described in below, with best performance according to each measure depicted in boldface.

TABLE 5

Hard Clustering Performance on Additional Data Sets

No. Of Algo.	Purity	Entropy	V-meas	Rand	F-meas.
FRECCA	0.713	0.335	0.634	0.885	0.521
ARCA	0699	0.375	0.611	0.871	0.491
Spect. Clust.	0.699	0.375	0.624	0.875	0.699
K-Medoids	0.608	0.456	0.520	0.859	0.462



VI. Conclusion

The FRECCA 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 Spectral Clustering and k-Medoids algorithms when externally evaluated in hard 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. Comparisons with the ARCA algorithm on each of these data sets suggest that FRECCA is capable of identifying softer clusters than ARCA, without sacrificing performance as evaluated by external measures. Although motivated by our interest in text clustering, FRECCA is a generic fuzzy clustering algorithm that can in principle be applied to any relational clustering problem, and application to several no sentence data sets has shown its performance to be comparable to Spectral. Graph-based methods are an exciting area of research within the pattern recognition community. We have already mentioned some of the new work we are conducting in this area; however, what we are most excited about is extending the technique to perform hierarchical clustering. The concepts present in natural language documents usually display some type of hierarchical structure, whereas the algorithm we have presented in this paper identifies only flat clusters. Our main future objective is to extend these ideas to the development of a hierarchical fuzzy relational clustering algorithm.

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.

ACKNOWLEDGEMENT

This is a great pleasure & immense satisfaction to express my deepest sense of gratitude & thanks to everyone who has directly or indirectly helped me in my project work successfully.

I express my gratitude towards project guide **Prof.V.K.Bhusari**,and **Prof. G.M.Bhandari**, Head of Department of Computer Engineering , Bhivarabai Sawant Institute Of Technology and Research College Of Engineering, Pune who guided & encouraged me in my project work in scheduled time. I would like to thanks our **Principal Dr. D.M.Yadav**, for allowing us to pursue my project in this institute.

