New approach of sentence clustering

Daksha chaudhari[1]
[1] Computer Science & Engineering Department
[1]GTU, Gujarat, India.

***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 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. We also include results of applying the algorithm to benchmark data sets in several other domains.

Keywords: similarity measure, graph centrality, EM framework, fuzzy cluster , sentence similarity measure.

I. INTRODUCTION

Advances in sensing, storage technology and rapid growth in applications such as Internet search, digital imaging, etc. have created many high-volume, high-dimensional datasets. Nowadays, most of data is stored digitally in electronic devices, hence improvement in various techniques like automatic data analysis, classification, and retrieval techniques become necessary. Many of these data streams are unstructured hence it is difficult to analyze them. This increases volume as well as variety of the data which requires advances in methodology to understand process and summarize the data automatically. Natural Language Processing (NLP) is a modern computational technology which is a method of inquiring and evaluating claims about human language itself. By applying techniques from natural language processing and data mining, previously unknown information is discovered by text mining. Sometimes Text mining is also alternatively named as text data mining, approximately equivalent to text analytics, refers to the process of extracting high quality information from text [1].

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 [2], [3], [4], [5]. However, sentence clustering can also be used within more general text mining tasks. For example consider web mining [6], where the specific objective might be to discover some novel information from a set of documents initially retrieved in response to some query. 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.

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 kmeans clustering[7], mixture models [7], spectral clustering [8], and maximum margin clustering [9], [10]. 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 [11]. This class of clustering methods is called soft- or fuzzy-clustering.

II. RELATED 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. To solve this problem, a number of sentence similarity measures have recently been proposed .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 (knowledge-based measures).

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.

III. PROPOSED WORK

For the data analysis of the high-volume and high dimensional dataset, clustering is useful technique. To cluster the text data, FRECCA and rough k-means clustering algorithm will be implemented. The comparative study of those techniques is necessary so as to make it more interactive as per the user's point of the view. The main objectives are as follows:

A. To cluster the text data use FRECCA.

B. To cluster the text data use Rough FRECCA clustering algorithm

C. To find the efficient algorithm compare ARCA and s FRECCA clustering algorithm.

To find the efficient algorithm, Results of the both the algorithms will be compared as well as on the basis of parameters such as execution time and memory required, performance of the algorithms will be compared. In some situation user has importance of time or whenever large volume of data present at that time cost of memory may matters. In such situations, result obtained in this module will be useful.

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. FRECCA will be extended as FRECCA using the concept of hierarchical fuzzy clustering. This algorithm is applied for the clustering of the text data which is present in the form of xml files. FRECCA will give the output as clusters which are formed from text data present in a given documents. Prior to the FRECCA, Page Rank algorithm is used as similarity measure.

Page Rank is used as a graph centrality measure. Basic idea behind the Page Rank algorithm is that the importance of a node within a graph can be find out by taking into account global information recursively which is computed by using the entire graph, connections to high-scoring nodes contributes more to the score of a node than connections to low-scoring nodes. Importance of node is used as a measure of centrality. This algorithm assigns numerical score (from 0 to 1) to every node in graph. This score is known as Page Rank Score. Sentence is represented by node on a graph and edges are weighted with value representing similarity between sentences [12, 15].

FUZZY clustering is used to partition the data into number of groups. Each group has a data that is similar in some sense [16]. An expectation–maximization (EM) algorithm is an iterative process, in which the model depends on unobserved latent/hidden variables. This algorithm is used for finding maximum likelihood estimates of parameters. The EM iteration alternates between performing an expectation (E) step, which creates a function to compute the cluster membership probabilities and maximization (M) step, in which these probabilities are then used to re estimate the parameters. These parameter-estimates are then used to figure out the distribution of the latent variables in the next E step. Divisive algorithms begin with just only one cluster. Then, the single cluster splits into two or more clusters that have higher dissimilarity between them until the number of clusters becomes number of samples or as specified by the user.

IV. APPLICATION

A. biology

In biology clustering has many applications. In the area of plant and animal ecology, clustering is used to define and to make spatial and temporal comparisons of organisms communities in heterogeneous environments; it is also used in plant systematic to create clusters of organisms at the species, genus or higher level that share number of parameters. In molecular biology, we are often interested in determining the group structure in, e.g., a population of cells or microarray gene expression data. Clustering methods identify groups of similar observations, but the results can depend on the chosen method's assumptions and starting parameter values.

B. medicine

In medical imaging, such as PET scans, clustering can be used in a three dimensional image to differentiate between different types of tissue and blood cells. In medicine and medical bioinformatics, more and more data arise from clinical measurements such as EEG or fMRI studies for monitoring brain activity, mass spectrometry data for the detection of proteins, peptides and composites, or microarray profiles for the analysis of gene expressions.

C. Recommender system

Recommender systems are designed to recommend new items on the basis of a user's tasks. Sometimes clustering algorithms are used to predict a user's preferences based on the preferences based on the user's in the user's cluster. . We also apply user clustering for organizing users into clusters of users with similar preferences. We propose the use of these cluster to efficeiently locate similar user to given one: these way searching for similar user is restricted within his/her corresponding cluster instead of whole database.

D. Social network

In analysis of social networks, clustering can be used to identify the communities within huge groups of people, people having similar designation, people who worked together in past or working together at present. Networks are

widely used to represent data on relations between interacting actors or nodes They can be used to describe the behaviour of epidemics, the interconnectedness of corporate boards, networks of genetic regulatory interactions and computer networks, among others. In social networks, each actor represents a person or social group, and each link, tie or arc represents the presence or strength of a relationship between two actors. Nodes can be used to represent larger social units , objects or abstract entities .

E. News extraction

To extract news related to the domain in which user is interested or want to know more about specific topic or location, this clustering algorithms can be used. E.g. Person A want to know news related to sports at Kolhapur. To extract the data for the given example, proposed algorithms will be very useful. The similar news headlines were grouped into a single cluster. The real world application of the project study would help people to find the similar news headlines on different news portal from a single platform. This would not have been possible without the use of text mining and clustering techniques. In general, it is not feasible to manually look for similar news in each of the portals and then compare each of them to find similarities between them.

V. CONCLUSIONS

Websites most often uses text-based searches, which only found documents having specific user-defined words or phrases. By using a semantic web, rather than just by a specific word text mining can find content based on meaning and context. Usually, concepts present in natural language documents contain hierarchical structure hence extension of ARCA i.e. Hierarchical Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm (FRECCA) will be useful for natural language documents. In this paper, 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 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 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] Shehata, S., Karray, F. and Mohamed, S. Kamel, 2010. An Efficient Concept-Based Mining Model for
[2] Enhancing Text Clustering IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 10.
[3] 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.
[4] 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.
[5] 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.
[6] 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.
[7] R. Kosala and H. Blockeel, "Web Mining Research: A Survey," ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, pp. 1-15, 2000.
[8] R. O. Duda, P. H. Hart, and D. G. Stock, Pattern Classification. New York: Wiley, 2001.
[9] U. von Luxburg, "A tutorial on spectral clustering," Statist. Comput., vol. 17, no. 4, 2007.
[10] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, "Maximum margin clustering," in Proc. Adv. Neural Inf. Process. Syst., 2004, pp. 1537–1544.
[11] K. Zhang, I.W. Tsang, and J. T.Kwok, "Maximum margin clusteringmade practical," in Proc. 24th Int. Conf. Mach. Learning, 2007, pp. 1119–1126.
[12] F.Hoppner, F. Klawonn, R. Kruse, and T. Runkler, Fuzzy Cluster Analysis.New York: Wiley, 1999.