A NOVEL FUZZY BASED APPROACH FOR CLUSTERING SENTENCE-LEVEL TEXT

[1]MEGHA NAIK, [2]RAJESH BHISE

[1]PG Student, Department of Computer Engineering, PHCET, Rasayani, Raigad, India
[2]Assistant Professor, Department of Computer Engineering, PHCET, Rasayani, Raigad, India
E-mail: [1]meghanaik09@gmail.com, [2]rhbhise@gmail.com

Abstract-Fuzzy clustering algorithms allow patterns to belong to all clusters with differing degrees of membership. In domains such as sentence clustering this is important. It is a novel fuzzy clustering algorithm that operates on relational input data in which the data is in the form of a square matrix of pairwise similarities between data objects. A detailed study about the Fuzzy Clustering Algorithm has been discussed in this paper. The fuzzy clustering 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. The task demonstrates that sentence clustering the algorithm is capable of identifying overlapping clusters of semantically related sentences and it is for a great potential use in a variety of text mining tasks. Here we also include results of applying the algorithm to benchmark data sets in several other domains.

Index Terms- Expectation Maximization, Fuzzy relational clustering, Graph Centrality, Natural Language Processing, Similarity Measure, Sentence Level Clustering.

I. INTRODUCTION

The development of information technology in the last two decades paved the way for a world full of data. But much of these data are of potentially not useful. In order to make it useful one, we need to extract the large amount of information or Knowledge underlying the data. Data mining is a process of extracting the valuable information inside the huge amount of data. Clustering techniques can help in this data discovery and data analysis. Clustering the sentences is mainly useful in Information Retrieval (IR) Process. Clustering text at the sentence level and document level has many differences. Document clustering partitions the documents into several parts and cluster those parts based on the overall theme [5]. It doesn't give much importance to the semantics of each sentence in the document. So there may be content overlap or bad coverage of them will happen in the case of multi document summarization [6]. Each data element in hard clustering method belongs to exactly one cluster.

Cluster Analysis

There are several algorithms available for clustering. By efficient way each algorithm will group or cluster similar data objects [7]. It assigns the task of dividing the data into various groups called clusters. The application of clustering includes Bioinformatics, News Extraction, and Social Network etc. In general, the text mining process focuses on the statistical study of terms or phrases which helps us to understand the significance of a word within a document [1]. Even if the two words didn't have similar meanings, clusters will be formed. The most important unsupervised learning framework for Clustering can be determined as a cluster is declared as a group of data items [2], which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Text mining applications is mainly used in variety of Sentence Clustering in which the user specifies the dataset and the query output of clustering should be related to it.

Similarity Measure

Similarity between the sentences is measured in terms of some distance function; such functions are Euclidean distance or Manhattan distance [9]. The measure is based on the choice of our requirement that induces the cluster size and formulates the success of a clustering algorithm on the specific application domain. Current sentence clustering methods [3],[6],[7] usually represent sentences as a term document matrix and perform the clustering algorithm on it. The documents can be grouped satisfactorily by using these clustering methods, but still it is hard for people to capture the meanings of the documents since for each document cluster there is no satisfactory interpretation. Based on the similarity or dissimilarity values clustering will take place [2].

II. REVIEW OF LITERATURE

In data capture, processing power, data transmission, and storage capabilities lead to dramatic advances in enabling organizations to integrate their various databases into *data warehouses*. In data mining functionalities [7], clustering analysis is the most significant tool for distribution of data. Clustering is a dynamic field of research in data mining concept. It is related to unsupervised learning [3] in machine learning. The process is initiated based on the similarity measures the cluster formation is made. With the help of different notations used in clustering algorithms, unique clusters are formed with the same data set.

Data mining technique based on machine learning is classification. Basically classification is used to classify each item in a set of data into one of a predefined set of classes or groups. In classification, we develop the software that can learn how to classify the data items into groups. To explain in more detail how we can apply classification in an application with the help of example, such as "given all records of employees who left the company; predict who will probably leave the company in a future period." In this case, we fragment the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.

Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique [4]. To make the concept clearer, we can take book management in library as an example. In a library, there is a wide range of books on various topics available. The challenge is how to keep those books in a way that readers can take several books on a particular topic without hassle. By using clustering techniques, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books on that topic, they would only have to go to that shelf instead of looking for the entire library.

Text mining [12] is the analysis of data contained in natural language text. The application of text mining techniques to solve business problems is called *text analytics*. Text mining can help an organization derive potentially valuable business insights from text-based content such as word documents, email and postings on social media streams like Facebook, Twitter and LinkedIn. Mining unstructured data with Natural Language Processing (NLP) [8], statistical modeling and machine learning techniques can be challenging, however, because the natural language text is often inconsistent. It contains ambiguities caused by inconsistent syntax and semantics, including slang, language specific to vertical industries and age groups, double engenders and sarcasm.

Natural language processing (NLP) [8] is the ability of a computer to understand what a human is saying to it. The goal of NLP evaluation is to measure one or more *qualities* of an algorithm or a system, in order to determine whether (or to what extent) the system answers the goals of its designers, or meets the needs of its users. Intrinsic evaluation considers an isolated NLP system and characterizes its performance mainly with respect to a *gold standard* result, pre-defined by the evaluators. Extrinsic evaluation, also called *evaluation in use* considers the NLP system in a more complex setting, either as an embedded system or serving a precise function for a human user. Based on text analyses, semantic relatedness between units of language (e.g., words, sentences) can also be estimated using statistical means such as a vector space model to correlate words and textual contexts from a suitable text corpus.

Efficient conceptual rule mining on text clusters is a modern and computational approach in text mining, which attempts to determine new, formerly unidentified information by applying various techniques from normal language processing and data mining domains. Clustering is one of the conventional data mining techniques that is used as an unsubstantiated learning pattern, where clustering techniques attempt to groupings of the text documents with different data items. Clusters have high intra-cluster similarity and low inter-cluster similarity. Most of the current document clustering methods and algorithms are fully based on the Vector Space Model (VSM), which is a widely used as a data representation for text classification and clustering. Weighting these feature words accurately affects the result of the clustering algorithm and improves its efficiency. In this paper, we are going to present a Conceptual rule mining, which is generated for the sentence meaning and related sentences in the document [3]. The sentences with appropriate weights in the entire document, the topic having higher contribution within.

Semantic similarity based on corpus statistics and lexical taxonomy is a new approach for measuring the semantic similarity between the words and concepts. The characteristics of polysemy and synonymy that exist in words of natural language have always been a challenge in the fields of Natural Language Processing (NLP) and Information Retrieval (IR) techniques. There are certain advantages in the work of semantic association discovery by combining a taxonomy structure with corpus statistics [3]. The proposed approach outperforms other computational models. With a benchmark resulting from human similarity judgments, it gives the highest correlation value.

Some experiments on clustering similar sentences of texts by identifying similar text passages or sentences plays an important role in many of the text mining applications. The proposal is based on some experiments on clustering similar sentences of texts in the documents [3]. The proposed framework is based on an incremental and unsupervised clustering method which is combined with statistical similarity metrics to measure the semantic distance between sentences. It aims at identifying sets of highly semantically-related sentences from a collection of documents. The Sentence Splitting is performed by a textual-segmentation tool called SENTER, which is based on a list of abbreviations and some sentence delimiters.

Clustering algorithms for text summarization using expectation maximization is very difficult for human beings to manually find out useful and significant data from a large amount of text data. With the help of text summarization algorithms this problem can be solved. Text Summarization [4] is the process of

condensing the input text file into shorter versions by preserving its overall content and meaning by using natural language processing, text summarization is obtained [6]. The two steps which consist in proposal describe a system. In the first step, we are implementing the phases of natural language processing such as dividing, tokenization, and applying tags to part of speech [8], and parsing. In the second step, we are implementing Expectation Maximization (EM) Clustering Algorithm to find out a sentence similarity between the sentences. Summarize text can easily obtain based on the value of sentences similarity.

III. EXISTING SYSTEM

In [9], [10] the Information Retrieval (IR) literature clustering text at the document level is well established, 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 such as TF-IDF (Term Frequency – Inverse Document Frequency) values of the keywords.

The vector space model [11] has been successful in IR because it is able to adequately capture much of the semantic content of document-level text. This is because semantically related documents that 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 [6]. However, while the assumption that (semantic) the similarity can be measured in terms of word co - occurrence may be valid at the document level, since two sentences may be semantically related despite having few, if any, words in common these assumptions do not hold for small-sized text fragments such as sentences.

The optimization algorithms [3], [7] that were used often suffered from instability in the results [6], [11].

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

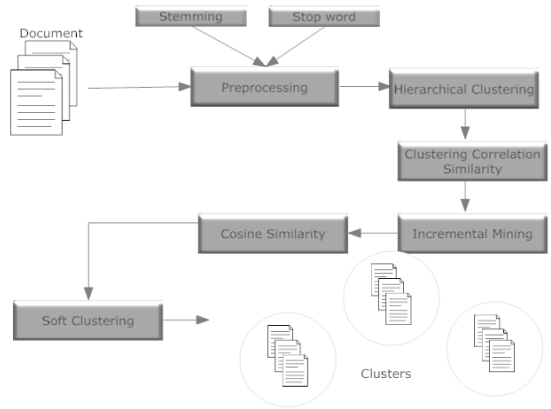

Fig. 1 Data Clustering and Processing

IV. PROPOSED SYSTEM

This paper presents [1], 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 [2]. 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. By applying FRECCA, the results of 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.

ADVANTAGES OF PROPOSED SYSTEM

The superior performance of the proposed system is capable to achieve the 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 [4] clusters of semantically related sentences.

Comparisons performed on each of these data sets suggest that FRECCA is capable of identifying softer clusters than ARCA [2], without sacrificing performance as evaluated by external measures.

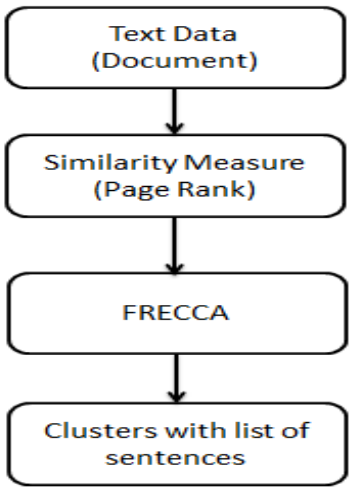

Fig. 2 Concept-Based Block Diagram Page Rank

A. Page Rank

We describe the application of the algorithm to data sets, and show that our algorithm FRECCA performs better than other fuzzy clustering algorithms. By describing the use of PageRank in the proposed algorithm and use the Gaussian mixture model approaches. Based on a graph centrality measure Page Rank is used. A Page Rank algorithm is used to determine the importance of a particular node within a graph. Depending on importance of node a measure of centrality is used. This algorithm

assigns a numerical score (from 0 to 1) to every node in the graph. This score is known as the Page Rank Score. The representatives of sentence are done with a node in a graph and edges are weighted with value representing similarity between sentences.
PageRank can be used within the Expectation-Maximization algorithm to optimize the parameter values and to formulate the clusters. With the help of the PageRank algorithm, a graph representation of data objects is used. It operates within an Expectation-Maximization; it is a framework which is a general purpose method for learning knowledge from the incomplete data. Each sentence in a document is represented by a node in the directed graph and the objects with weights indicate the object similarity.

B. EM algorithm
It is an unsupervised method, which does not need any training phase; it tries to find the parameters of the probability distribution that has the maximum likelihood of its parameters. Its main role is to parameter estimation. It is an iterative method, which is mainly used to finding the maximum likelihood parameters of the model. The E-step involves the computation of cluster membership probabilities. The probabilities calculated from an E - step are re-estimated with the parameters in the M - step.

C. Fuzzy Relational Clustering – FRECCA
A fuzzy relational clustering approach [1] is used to produce clusters with sentences, where each of them corresponds to some content. The strength of the association among the data elements is indicated by the output of clustering. A novel fuzzy relational clustering algorithm called FRECCA (Fuzzy Relational Eigen Vector Centrality based Clustering Algorithm).

The algorithm [1] involves the following steps:

- *Initialization:* Cluster membership values are initialized randomly, and normalized. Mixing coefficients are initialized.
- *Expectation*: Calculates the PageRank value for each object in each cluster.
- *Maximization*: The mixing coefficients are updated based on membership values calculated in the Expectation Step.

The data set used in this approach is Famous Quotations Data set. Quotations are able to provide a rich and challenging content for evaluating sentence clustering because contains a rich set of semantic information. The quotations of different categories like Marriage, Peace, Friendship, Food, Knowledge, and Nature. We have designed a database of 160 famous quotations from five different classes. There is some degree of overlap between the words in the quotations, this is not sufficient to allow to measure of similarity. To calculate similarity between the words, we are using a common vector space representation for all sentences.

System Architecture

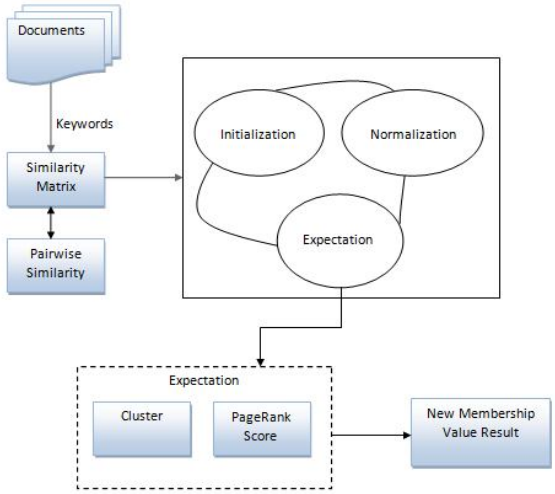

Fig. 3 Proposed Architecture

ALGORITHM USED

The proposed algorithm Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm (FRECCA), presents the proposed clustering algorithm. Here the algorithm begins by describing the use of PageRank as a general graph centrality measure, and review the Gaussian mixture model approach. How the PageRank can be described within an Expectation-Maximization framework to construct a complete relational fuzzy clustering algorithm. The final discussion of this section has various issues relating to convergence and complexity, duplicate clusters, and other implementation issues. Since as a special case of eigenvector centrality, PageRank centrality can be viewed.

CONCLUSION

Sentence Clustering is one of the clustering techniques. The performance of clustering techniques mainly depends on the quality of the input data set and the similarity measure that we choose. This kind of clustering techniques has gained great success in many areas. The effectiveness of the algorithm is based on the feature selection and it leads to good clustering of texts. From the study of analyzing various fuzzy clustering techniques in sentence clustering domain, it is clear that the algorithm can apply to asymmetric matrices and is not sensitive to the cluster membership value initialization. The results we got from the clusters is not unique and it is strongly depends upon the algorithm taken. We analyzed how it is possible to combine different results in order to obtain the stable clusters, not depending too much on the criteria selected to analyze data. The proposed

RECCA algorithm can also work with any type of relational clustering algorithms. The algorithm can also be used in general text mining applications. In the case of sentence clustering, the algorithm is not sensitive to the initialization of cluster membership values. Our main future objective is to extend these ideas to the development of a probabilistic based fuzzy relational clustering algorithm and the output will be clustered with its respective sentences.

REFERENCES

[1] Andrew Skabar, Khaled Abdalgader, "Clustering Sentence - Level Text Using a Novel Fuzzy Relational Clustering Algorithm" IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 1, January 2013.

[2] P. Corsini, F. Lazzerini, and F. Marcelloni, "A New Fuzzy Relational Clustering Algorithm Based on the Fuzzy C -Means Algorithm," Soft Computing, vol. 9, pp. 439-447, 2005.

[3] J. Durga, D. Sunitha, S.P.Narasimha, B. Tejeswini Sunand "A Survey on Concept Based Mining Model using Various Clustering Techniques" International Journal of Advanced Research in Computer Science and Software Engineering 2012.

[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. (4)

[5] Ms. Seema V. Wazarkar, Ms. Amrita A. Manjrekar, "Text Clustering Using HFRECCA and Rough K-Means Clustering Algorithm", International Conference on Advances in Computer Engineering & Applications (ICACEA-2014) at IMSEC, GZB.

[6] K.Sathishkumar, E. Balamurugan, and D. Kavin, "Sentence Level Clustering Approaches and its Issues in Various Applications", International Journal of Applied Research and Studies, 2278-9480 Volume 2 Issue 9, 2013.

[7] Yaminee S. Patil and M.B.Vaidya, "A Technical Survey on Cluster Analysis in Data Mining" - ISSN 2250-2459, Volume 2, Issue 9, September 2012.

[8] Richard Khoury "Sentence Clustering Using Parts-of-Speech" I. J. Information Engineering and Electronic Business, 2012, 1, 1-9.

[9] T. Aswani, A. Nageswara Rao, "Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm", International Journal of Research Studies in Science, Engineering and Technology, 2349-4751 Volume 1 Issue 8, November 2014.

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

[11] G. Salton, A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing" Comm. ACM, vol. 18, no. 11, pp.112-117, 1975

[12] http://www.dqglossary.com/textmining.html

★★★