Sentence Level Clustering using Fuzzy Relational Algorithm

Snehal Raundal
Computer Department,
G. E. S'S R. H. Sapat College 0f Engineering,
Management Studies and Research,Nashik,India

C. R. Barde
Computer Department,
G. E. S'S R. H. Sapat College of Engineering,
Management Studies and Research,Nashik,India

ABSTRACT

Clustering is an extremely studied in data mining problem for the text mining domains. Difficulty finds in various applications like customer segmentation, visualization, and collaborative filtering, classification, indexing and document organization. In text mining, clustering the sentence is the processes and it is used within a general text mining tasks. Some of the clustering algorithms and methods are used for clustering the documents at sentence level. In text clustering, sentence level clustering plays a important role this is used in text mining activities. The cluster size may change from one cluster to another and traditional clustering algorithms have some problems in clustering while the input in the form of data set. This gives problems such as, instability of clusters, sensitivity and complexity. To overcome the limitations of those clustering algorithms, in this paper proposes a algorithm called Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm (FRECCA) which is used for the clustering of sentences. We can obtain the more efficient method to overcome the problems in these existing approaches.

Keywords:

data mining, information retrieve, fuzzy relational clustering, sentence clustering.

1. INTRODUCTION

The development of information technology in the last two decades paved the way for a world full of data. But many of these data are not useful. Data mining is a process which extraction the valuable information from the huge amount of data. Clustering techniques are useful for data discovery and data analysis.
Clustering the sentences is useful in Information Retrieval (IR) Process for the better results. Clustering document at the sentence level and document level has many differences [1] [2] [3]. Document clustering is used to partitions the documents into parts and cluster this parts based on the overall theme. Each data element of the hard clustering is belongs to only one cluster.

1.1 Cluster Analysis

There are many algorithms available for clustering the data. The algorithm will cluster the data or make a group of similar data objects in a useful manner. This task involves dividing the data into various groups called clusterskey1. The various applications of clustering are Bioinformatics, Data mining, Business modelling, image processing 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. Even if the two words didn't have similar meanings then clustering will takes place. Clustering is to be considered the most important learning framework, a cluster is known as a group of similar data items, which are similar between them and are dissimilar to the objects belonging to other clusters. Sentence Clustering mainly used in variety of text mining applications. Output of clustering should be related to the query, which is specified by the user.

1.2 Similarity Measure

Distance functions are used to measure the similarity between the sentences these are Manhattan distance or Euclidean distance. The choice of the measure is based on the requirement of the cluster size and clustering algorithms are formulating the success of the specific application domain. The current clustering methods are representing sentences as a term matrix and perform clustering algorithm. Based on the similarity and dissimilarity of clustering values will take place.
In the following sections proposed system is described in detail. Section I gives introduction to the technique need and applications of the proposed system. Section II describes the related work i.e. study of existing systems, analysis of existing systems. Section III describes the motivation and problem definition of the system. In section IV the system design is presented using system architecture, methodology. Section V gives the algorithms for that requires and the results comparison of system. Finally, conclusions and acknowledgement are stated and also references are given at the end of the report.

2. RELATED STUDY

2.1 Efficient Conceptual Rule Mining on Text Clusters

Text mining is a modern and computational approach, which attempts to determine new, formerly unidentified information by applying various techniques from normal language processing and data mining domains. The clustering is one of the conventional data mining system that is used us an unsubstantiated learning pattern and where clustering are 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), [4][5] which is a widely used as a data representation for text classification and clustering. Words are accurately affects the result by weighting features of the clustering algorithm and improves its efficiency. In this paper, they are going to present the Conceptual rules which are used to generate for the related sentences and sentence meaning in the document. Weights of the sentences are appropriated having higher contribution to the topic of the entire document.

2.2 Clustering algorithm based on the fuzzy C-means algorithm

In this work, showed at which point one can take advantage of the efficiency and stability of data clustering algorithm when data to be clustered are ready to be drawn in the consist of of mutual numerical relationships between pairs of objects More particularly, here propose a new fuzzy relational algorithm, based on the popular fuzzy C-means(FCM) algorithm[6], which does not demand any particular restriction on the relation matrix. Here describe the review of the algorithm to four real and four synthetic data sets, and prove that this algorithm performs better than popular fuzzy relational clustering algorithms on all these sets.

2.3 A Clustering Algorithm for Text Summarization using Expectation Maximization

Large amount of data is available in the form of texts. It is very difficult for human beings to manually find out useful and significant data. This problem can be solved with the help of text summarization algorithms[7]. Text Summarization is a process which is condensing the input file as a text into small version by preserving its content and meaning. The paper is about called text summarization using natural language processing. The proposal describes a system, which consists of two steps. In first step, we are implementing the phases of natural language processing that are splitting, tokenization, and part of speech tagging, and parsing. In second step, we are implementing Expectation Maximization Clustering is to find out similarity of sentence between the sentences. Value of sentences similarity we can easily summarize the text.

2.4 Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy

This proposal is give the approach for calculating the semantic similarity between the concepts and words. In the fields of Natural Language Processing and Information Retrieval techniques the characteristics of synonymy and polysemy that exist in words of language have always been a challenge the proposed approach outperforms other computational models. That gives the highest correlation value that with a resulting from human similarity and judgements.
Sarkar [8] proposed cluster which represent the sentence in multi-document text summarization depend on the factors such as clustering the sentences, cluster ordering. The uni-gram matchingbase similarity measures after a preprocessing in a similar sentence. During pre-processing stemming was not applied on input and properties such as length, sentence position, and cue phrase are not incorporated to make system effective and portable in domain and language. Wazarkar Majrekar [9] proposed Rough set clustering whose exact border line cannot be defined due to incomplete information gives another way of representing datasets. Rough sets have been conventional used and can be equally useful in clustering for classification of a sets. The crisp boundary line did not necessary in data mining. According to Pimpalkar [10], this system collects the number of reviews from various online websites. The given text sentences at document level was checked by all the detail of that particular product. It clusters the contents of the documents +ve, -ve or neutral. The output for any product reviews Rule based method approach was used for proper filter. Sentiment of the product was used for selecting directly and it can also accept the smiley?s of the product. To select the best product between the two it compares two products. Li et al. [11] proposed the semantic and word order information present method for measuring the similarity between sentences or very short text. The lexical knowledge base and corpus has given by semantic similarity. Word order similarity measures the number.

3. MOTIVATION

This paper contributes a novel fuzzy relational clustering algorithm. which is inspired by the mix model approach, data modelling is a combination of components unlike the conventional mixture models, which can be operated in a Euclidean space and use as a likelihood function parametrized by the means and covariance's of Gaussian components, The explicit density model (e.g., Gaussian) represents the clusters key1. Instead, we will use the graph representation in which nodes will represent the objects and the weighted edges will represent the similarity between objects.
Cluster membership values for each node represent the degree to which the object represented by that node belongs to each of the respective cluster and mixing coefficients are representing the probability of an object having been generated from that component. By applying the Page Rank algorithm to every cluster and interprets the Page-Rank score of an object within some cluster as a likelihood, we can also use the Expectation-Maximization (EM) framework for determining the model parameters (i.e., cluster membership values and mixing coefficients) [1]. The result is a fuzzy relational clustering algorithm which is generic and can be applied to any domain in which the relationship between objects is expressed in terms of pair wise similarities.

3.1 Problem Definition

Sentence clustering is a type of text classification since they are small sized texts, it posses many challenges than other text classification process. Most of the standard clustering algorithm can be used to cluster sentences. But fuzzy based sentence clustering algorithms shows improved performance evaluation results. Consider web mining, where the specific objective might be discover some information from the set of documents that initially retrieved is response to some query.
Input: Set of document, News articles and Text documents.
Output: Gives the data in to the clustering form.

4. SYSTEM ARCHITECTURE

In this wok, we analyse how one can take advantage of the stability and efficiency of clusters, when data to be clustered are available in the form of similarity relationships between pairs of objects. We propose a new fuzzy relational clustering algorithm, based on the

existing fuzzy C-means algorithm [1], which does not require any restriction on the relation matrix.

4.1 Page rank

We describe the application of the algorithm to data sets and the result show that our algorithm performs better than other clustering algorithms. The proposed algorithm describes the use of Page Rank and use the Gaussian mixture model approach. This algorithm is used as a graph centrality measure. Algorithm is to determine the importance of a particular node within a graph. Importance is to measure a centrality by using node. The algorithm assigns score (from 0 to 1) to every node in graph and this score known as the Page Rank Score. The sentence is represented by node on a graph and Edges are weighted with value representing similarity between sentences.
It is used within the Expectation- Maximization algorithm to minimize the parameter values and to formulate the clusters. A graph representation is used in along with this algorithm of data objects. It operates within an Expectation-Maximization algorithm; this 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.

4.2 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.

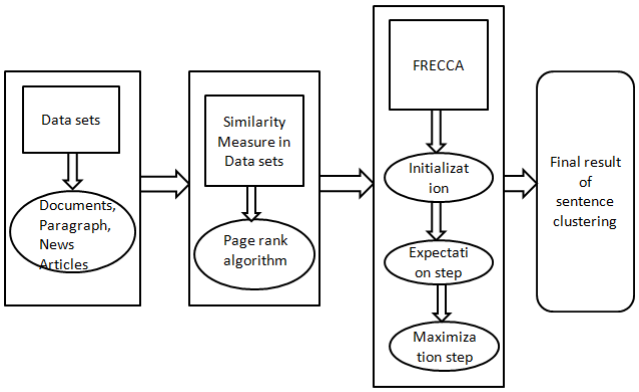

Fig. 1. System Architecture

Its main role is to parameter estimation. It is an iterative method, which is mainly used to finding the likelihood parameters of the model. The E-step involves the computation of cluster membership probabilities. The probabilities are calculated from E-step are re estimated with the parameters in M-step.

4.3 Fuzzy Relational Clustering FRECCA Algorithm

A fuzzy relational clustering approach is used to produce clusters with sentences, where each of them corresponds to some content. The clustering indicates the strength of the association among the elements. Andrew Skabar and Khaled Abdalgader proposed a novel fuzzy relational clustering algorithm called FRECCA [1].

The general idea of the clustering is the partitioning of the data items into the clusters. The data points shows the assigned membership values for every clusters. Many existing clustering technique has difficulties in handling extreme outliers but fuzzy clustering algorithm is give them very small membership degree in surrounding clusters. An expectation?maximization algorithm is an iterative process, in which the model mainly depends on some unobserved latent/hidden variables. This algorithm is particularly used in finding maximum likelihood estimates of parameters. The EM algorithm's iteration alternates between performing an expectation Step; this creates a function to compute the cluster membership probabilities and maximization step, in which these probabilities are then used to re-estimate the parameters. These parameter estimates are then used to find out the distribution of the latent variables in the next E step.

5. ALGORITHM

The algorithm has the following steps.
1. Initialization: cluster membership values are randomly initialized, and normalized. Initialize Mixing coefficient.
2. Expectation: Page Rank values are calculated for each object in each cluster.
3. Maximization: Membership values calculated in the Expectation Step based on updating the mixing coefficients.

INITIALIZATION
// initialize and normalize membership values
for i = 1 to N
for m = 1 to C
p_i^m =rnd // random number on [0, 1]
end for
for m = 1 to C
$p_i^m = p_i^m / \sum_{j=1}^{c} p_{ji}$ //normalize
end for
end for
for m = 1 to C
π_m=1 / C // equal priors
end for
repeat until convergence

EXPECTATION STEP
for m = 1 to C
// create weighted affinity matrix for cluster m for i = 1 to N
for j = 1 to N
$w_{ij}^m = s_{ij} \times p_i^m \times p_j^m$ end for
end for
// calculate PageRank scores for cluster m
repeat until convergence
$PR_i^m = (1-d) + d \times \sum_{j=1}^{N} w_{ij}^m (PR_i^m / \sum_{k=1}^{N} w_{ik}^m)$
end repeat
// assign PageRank scores to likelihoods
$l_i^m = PR_i^m$
end for
// calculate new cluster membership values
for i = 1 to N
for m = 1 to C
$p_i^m = (\pi_m \times l_i^m) / \sum_{j=1}^{c} (\pi_j \times l_i^j)$
end for
end for

MAXIMIZATION STEP

Table 1. Clustering Performance Evaluation

Techniques	Purity	Entropy	V-meas	Rand	F-meas
FRECCA	0.800	0.324	0.646	0.863	0.601
ARCA	0.622	0.451	0.524	0.815	0.462
Kmedoids	0.720	0.457	0.546	0.779	0.459
Spec Clus	0.690	0.475	0.508	0.800	0.444

// Update mixing coefficients
for m = 1 to C
$\pi_m = \frac{1}{N} \sum_{i=1}^{N} p_i^m$
end for
end repeat

5.1 Result

The performance evaluation of the proposed FRECCA clustering algorithm is based on certain performance metrics. The results shown in Table 1.
The performance metrics used in this paper are Partition Entropy Coefficient (PE), Purity and Entropy, V-Measure, Rand Index and F-Measure.

$$PE = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{|L|} (u_{ij} \log_a u_{ij}) \quad (1)$$

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

$$OverallPurity = \frac{1}{N} \sum_{j=1}^{|L|} (|w_j| \times P_j) \quad (2)$$

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

$$OverallEntropy = \frac{1}{N} \sum_{j=1}^{|L|} (|w_j| \times E_j) \quad (3)$$

V -measure also known as the Normalized Mutual Information (NMI) which is defined as the harmonic mean of homogeneity (h) and completeness (c); i.e.,

$$V = hc/(h + c) \quad (4)$$

where h and c are defined as
$h = 1 - \frac{H(C|L)}{H(C)}$ and $c = 1 - \frac{H(L|C)}{H(L)}$

Rand Index and F-Measure
Each pair can fall into one of four groups: if both objects belong to the same class and same cluster then the pair is a true positive (TP); if objects belong to the same cluster but different classes the pair is a false positive (FP); if objects belong to the same class but different clusters the pair is a false negative (FN); otherwise the objects belong to different classes and different clusters, and the pair is a true negative (TN).

$$RI = (TP + TN)/(TP + FP + FN + TN) \quad (5)$$

$$F - measure = 2PR/(P + R) \quad (6)$$

where $P = TP/(TP + FP)$ and $R = TP/(TP + FN)$

6. CONCLUSION

The algorithm was motivated by 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 given that the algorithm is able to find overlapping clusters of semantically related sentences. We have already mentioned some of the new work we are conducting in this area; what we are most excited about is extending the technique to perform hierarchical clustering. The concepts are present in NLP documents usually display some type of hierarchical structure, whereas the algorithm we have presented in this paper identifies only at clusters. The main objective of the project is to extend this idea to the development of a fuzzy relational clustering algorithm, 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.

Acknowledgment

I would like to thank all people who help me in different way. Especially I am thankful to my guide Prof. C. R. Barde for his continuous support and guidance in my work. Also, I would like thank our H.O.D. of Computer Engineering Prof. N. V. Alone and our PG. Coordinator Prof. A. S. Vaidya for motivating me. Lastly, I thank to IJCA who have given opportunity to present my paper.

7. REFERENCES

[1] Andrew Skabar and Khaled Abdalgader. Clustering sentence-level text using a novel fuzzy relational clustering algorithm. *IEEE Transl. on Knowledge and data Engineering*, 25(1):62–75, January 2013.

[2] H. Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. pages 113–120, 2002.

[3] R.M. Aliguyev. A new sentence similarity measure and sentence based extractive technique for automatic text summarization. *Expert Systems with Applications*, 36:7764–7772, 2009.

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

[5] C.D. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval*. Cambridge Univ. Press, 2008.

[6] P. Corsini, F. Lazzerini, and F. Marcelloni. A new fuzzy relational clustering algorithm based on the fuzzy c-means algorithm. *Soft Computing*, 9:439–447, 2005.

[7] D. Wang, T. Li, S. Zhu, and C. Ding. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. *Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization*, pages 307–314, 2008.

[8] K. Sarkar. Sentence clustering-based summarization of multiple text documents. *TECHNIAInternational Journal of Computing Science and Communication Technologies*, 2(1):325 – 335, July 2009.

[9] Seema V. Wazarkar and Amrita A. Manjrekar. Text clustering using hfrecca and rough k-means clustering algorithm. *Discovery*, 15(40):44– 47, April 2014.

[10] A. Pimpalkar, T. Wandhe, M. Swati Rao, and M. Kene. Review of online product using rule based and fuzzy logic with smileys. *International Journal of Computing and Technology*, 1(1):39 – 44, February 2014.

[11] Yuhan Li., D. McLean, Z. A. Bandar, J. D. OShea, and K. Crockett. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18:1138 – 1150, June 2006.