Clustering Sentence-Level Text Using Novel Fuzzy Approach

Gopi G [1], Pandiarajan [2], Vanitha E [3], Kala P [4]

[1, 2, 3, 4] *Assistant Professor*
Department of Computer Science and Engineering
PTR College of Engineering and Technology, Madurai, Tamilnadu, India

***Abstract*—** Fuzzy clustering algorithms permit patterns to belong to all clusters with differing degrees of membership. As a sentence is likely to be related to more than theme or topic present with in a document or set of documents sentence clustering becomes important. This paper presents a 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 algorithm is capable of identifying overlapping clusters of semantically related sentences. In this paper it is proposed to develop an Emotional term model to predict the sentence emotions. Emotional term model with fuzzy relationship algorithm is used for variety of text mining process. This method is also used in several domains for further classification.

Keywords— Sentence, Fuzzy Clustering, Text Mining, Expectation-Maximization

I.INTRODUCTION

Mining is the process of inferring for patterns with in a structured or unstructured data. There are various mining methods out of which they differ in the context and type of dataset that is applied.

The process of extracting information and knowledge from unstructured text led to the need for various mining techniques for useful pattern discovery. Data Mining (DM) and Text Mining (TM) is similar in that both techniques mine large amounts of data, looking for meaningful patterns. Some of the mining types are data, text, web, business Process and service mining.

Text Mining (TM) refers some informational content included in any of the items such as: newspaper; articles; books; reports; stories; manuals; blogs; email, and articles in the WWW. The quantum of text of the present day is pretty vast with ever-growing incremental power. The prime aim of the text mining is to identify the useful information without duplication from various documents with synonymous understanding. TM is an empirical tool that has a capacity of identifying new information that is not apparent from a document collection.

A.CLUSTERING

Clustering refers to the process of unsupervised partitioning of a data set based on a dissimilarity measure, which determines the cluster shape. Considering that cluster shapes may change from one cluster to another, it would be of the utmost importance to extract the dissimilarity measure directly from the data by means of a data model. The aim of cluster analysis is to organize a collection of patterns (usually represented as vectors of measurements, or points in a multidimensional space) into homogeneous groups (called *clusters*) based on pattern similarity.

a. Sentence Clustering

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 project 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 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. This type of data, which we refer to as “attribute data,” is amenable to clustering by a large range of algorithms. Since data points lie in a metric space, we can readily apply prototype-based algorithms such as k-Means , Isodata, Fuzzy c-Means (FCM) and the closely related mixture model approach , all of which represent clusters in terms of parameters such as means and covariance’s, and therefore assume a common metric input space.

Emotional sentence clustering provides the larch application in much text processing activity. It is effectively used in multi document summarization Process. This application used in documents which contain large segments. It provides the relationship between objects (cluster document) expressed in terms of pair wise similarities. The Gaussian used to represent the cluster in the graph. Fuzzy relational clustering gives the cluster that contains the famous quotations of given input. An emotional term gets the famous quotation from Fuzzy relational clustering module. Then it produces the emotional term such as happy and sad related sentences.

II.METHODS

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 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. This type of data, which we refer to as “attribute data,” is amenable to clustering by a large range of algorithms. Since data points lie in a metric space, we can readily apply prototype-based algorithms such as k-Means , Isodata, Fuzzy c-Means (FCM) and the closely related mixture model approach , all of which represent clusters in terms of parameters such as means and covariance’s, and therefore assume a common metric input space.

Emotional sentence clustering provides the larch application in much text processing activity. It is effectively used in multi document summarization Process. This application used in documents which contain the large segments. It provides the relationship between objects (cluster document) is expressed in terms of pair wise similarities. The Gaussian used to represent the cluster in the graph. Fuzzy relational clustering give the cluster that contain the famous quotations of given input An emotional term get the famous quotation from Fuzzy relational clustering module. Then it produces the emotional term such as happy and sad related sentences.

This paper presents a 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.

In this paper we also propose the Emotional term Model by using Fuzzy Relational Clustering Algorithm. Emotional term model is mining social emotions from text and more documents then produce emotion such as happiness, sadness, and surprise. A joint emotion-topic model by augmenting Latent Dirichlet Allocation with an additional layer for emotion modeling is developed.

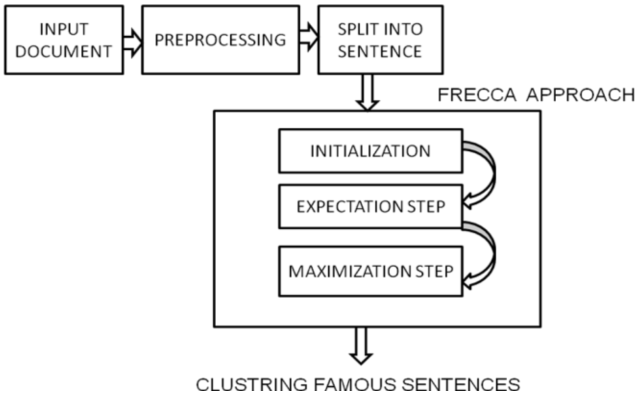


Figure1: Architecture diagram

PROPOSED ALGORITHM

In this paper we are using the FRECCA Algorithm (Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm) Proposed first describe the use of Page Rank as a general graph centrality measure, and review the Gaussian mixture model approach.

Then describe how Page Rank can be used within an Expectation-Maximization framework to construct a complete relational fuzzy clustering algorithm.

FRECCA algorithm contains following steps

i) Initialization
ii) Expectation Step
iii) Maximization Step

i)Initialization

Here that clusters membership values are initialized randomly, and normalized such that cluster membership for an object sums to unity over all clusters. Mixing coefficients are initialized such that priors for all clusters are equal.

- Initialize and normalize membership values.

$p_i^m = p_i^m / \sum_{j=1}^{C} p_i^j$ ………(1)

- Random number on [0, 1]
- Equal priors

ii) Expectation step

The E-step calculates the Page Rank value for each object in each cluster. Page Rank values for each cluster are calculated as described in with the affinity matrix weights obtained by scaling the similarities by their cluster membership values.

- Create weighted affinity matrix for cluster

$w_{ij}^m = \mathrm{s_{ij}} \mathrm{x}\, \mathrm{p_i^m} \mathrm{x}\, \mathrm{p_j^m}$ ………(2)

- Calculate Page Rank scores for cluster

$PR_i^m = (1-d) + d \mathrm{X} \sum_{j=1}^{N} w_{ij}^m \left(PR_j^m / \sum_{j=1}^{N} w_{ij}^m\right)$

……(3)

- Assign Page Rank scores to likelihoods

$l_i^m = PR_i^m$ ………(4)

- Calculate new cluster membership values

$p_i^m = (\pi_m \mathrm{x}\, l_i^m) / \sum_{j=1}^{C} (\pi_j \mathrm{x}\, l_i^j)$ ………(5)

iii) Maximization step:

Since there is no parameterized likelihood function, the maximization step involves only the single step of updating the mixing coefficients based on membership values calculated in the Expectation Step.

- Update mixing coefficients

$$\pi_m = \frac{1}{N} \sum_{j=1}^{N} (p_i^m)$$ ………(6)

III. RESULT AND DISCUSSION

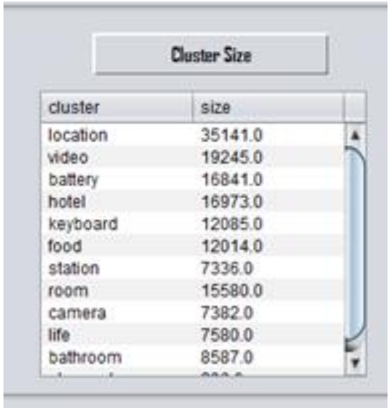
FRECCA has a number of attractive features. First, based on empirical observations, it is not sensitive to the initialization of cluster membership values, with repeated trials on all data sets converging to exactly the same values, irrespective of initialization. Second, the algorithm appears to be able to converge to an appropriate number of clusters, even if the number of initial clusters was set very high. Finally, while we have applied the algorithm using symmetric similarity measures, the algorithm can also be applied to asymmetric matrices.

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.

Page Rank assigns to every node in a directed graph a numerical score between 0 and 1, known as its Page Rank score (PR).FRECCA requires that an initial number of clusters is specified. This number was varied from 3 to 15, running 50 trials for each case, each trial commencing from a different random initialization of membership values. Interestingly, only twelve unique clustering's were found, each containing a different number of clusters, which ranged from three to eight. Emphasize that the tabulated results for FRECCA are not averages these were the only clustering found.

The famous quotations data set was constructed in order that we could evaluate performance of the algorithm using standard external cluster quality criteria. To demonstrate how the algorithm may be of more general use in activities related to text mining, we now apply the algorithm to clustering sentences from a document.

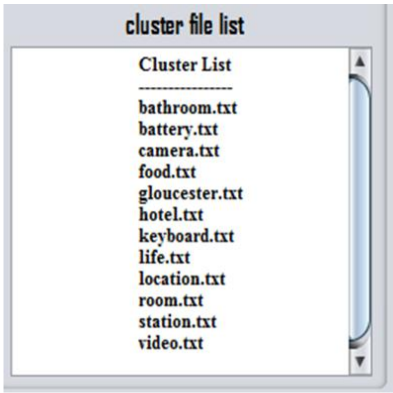
Cluster evaluation may be either supervised, in which case external information (usually known class labels associated with the instances) is used to measure the goodness of the clustering; or unsupervised, in which case no external information is used.



Cluster Size

cluster	size
location	35141.0
video	19245.0
battery	16841.0
hotel	16973.0
keyboard	12085.0
food	12014.0
station	7336.0
room	15580.0
camera	7382.0
life	7580.0
bathroom	8587.0

Figure 2: Famous sentence cluster and its size



cluster file list

Cluster List
----
bathroom.txt
battery.txt
camera.txt
food.txt
gloucester.txt
hotel.txt
keyboard.txt
life.txt
location.txt
room.txt
station.txt
video.txt

Figure 3: Cluster file list

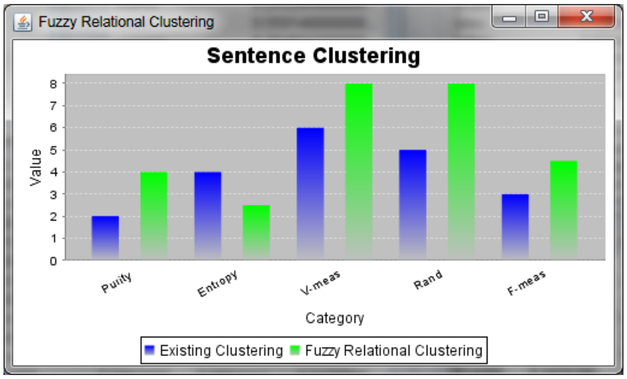


Figure 4: Cluster evaluation criteria chart

Many unsupervised evaluation measures have been defined, but most are only applicable to clusters represented using prototypes. Two exceptions are the Partition Coefficient (PC) and the closely related Partition Entropy Coefficient.

Two widely used external clustering evaluation criteria are purity and entropy. 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 . The entropy of a cluster j is a measure of how mixed the objects within the cluster. Good clustering is thus characterized by a high purity and low entropy. Because entropy and purity measure how the classes of objects are distributed within each cluster, they measure homogeneity; i.e., the extent to which clusters contain only objects from a single class. However, we are also interested in completeness; i.e., the extent to which all objects from a single class are assigned to a single cluster. While high purity and low entropy are generally easy to achieve when the number of clusters is large, this will result in low completeness, and in practice we are usually interested in achieving an acceptable balance between the two.

This problem with purity and entropy is overcome by the V -measure, also known as the Normalized Mutual Information (NMI), which is defined as the harmonic mean of homogeneity and completeness;

Unlike purity, entropy, and V - measure, which are based on statistics, Rand Index and F-measure are based on a combinatorial approach which considers each possible pair of objects. 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). The Rand index is simply the accuracy;

All the clusters are evaluated by finding the purity, entropy, f-measure and random index and compared with previous literature result. In future we are going to extend in this project in the domain of emotion analysis.

IVCONCLUSION AND FUTURE WORK

In our paper presented a novel fuzzy relational sentence clustering scheme and evaluated its implementation, showing significantly superior performance over common sentence clustering techniques. Our work focuses on the design of a successful clustering based fuzzy relational clustering algorithm and the related issues such as how to cluster sentences, how to order clusters and how to select representative sentences from the clusters. The performance of our system can be improved by improving its different components. How to measure similarity between sentences is also a crucial issue in sentence clustering based clustering approach. The better similarity measure will improve the clustering performance. Experimental results demonstrate the effectiveness and the robustness of the proposed approach.

We plan to further explore the suggested scheme by utilizing emotional term model to determining intensity of sentence emotion and recognizing emotion of sentence or document used in document clustering.

REFERENCES

[1]. Andrew Skabar, Member, IEEE, and 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]. 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.

[3]. Jesús Andrés-Ferrer , Germán Sanchis-Trilles, Francisco Casacuberta "Similarity word-sequence kernels for sentence clustering" Proceeding SSPR&SPR'10 Proceedings of the 2010 joint IAPR international conference on Structural, syntactic, and statistical pattern recognition Pages 610-619 .

[4]. Jian-Ping Mei, Lihui Chen "SumCR: A new subtopic-based extractive approach for

text summarization" Journal of Knowledge and Information Systems Volume 31, Issue 3 , pp 527-545 June 2012.

[5]. Lili Kotlerman, Ido Dagan, Maya Gorodetsky, Ezra Daya "Sentence Clustering via Projection over Term Clusters" Proceeding SemEval Martina Naughton, Nicola Stokes, and Joe [6]. Carthy "Sentence-Level Event Classification in Unstructured Texts" Journal Information Retrieval archive Volume 13 Issue 2, April 2010 Pages 132-156.

[7]. 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.

[8]. R. Kosala and H. Blockeel, "Web Mining Research: A Survey," ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, pp. 1-15, 2000.

[9]. R. Krishnapuram, A. Joshi, and Y. Liyu, "A Fuzzy Relative of the k-Medoids Algorithm with Application to Web Document and Snippet Clustering," Proc. IEEE Fuzzy Systems Conf., pp. 1281-1286, 1999.

[10]. R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity," Proc. 21st Nat'l Conf. Artificial Intelligence, pp. 775-780, 2006.

[11].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.

[12].Ramiz M. Aliguliyev "A new sentence similarity measure and sentence based extractive technique for automatic text summarization" Journal Expert Systems with Applications: An International Journal archive Volume 36 Issue 4, May, 2009.

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

[14].S. Theodoridis and K. Koutroumbas, Pattern Recognition, fourth ed. Academic Press, 2008.

[15].T. Geweniger, D. Zu¨ hlke, B. Hammer, and T. Villmann, "Median Fuzzy C-Means for Clustering Dissimilarity Data," Neurocomputing, vol. 73, nos. 7-9, pp. 1109-1116, 2010.

[16].Trappey, A.J.C.; Trappey, C.V.; Fu-Chiang Hsu ; Hsiao, D.W. "A Fuzzy Ontological Knowledge Document Clustering Methodology" Systems, Man,

and Cybernetics, Part B: Cybernetics, IEEE Transactions on 2009.

[17].Xiaoyan Cai, Wenjie Li “Enhancing sentence-level clustering with integrated and interactive frameworks for theme-based summarization” Journal of the American Society for Information Science and Technology archive Volume 62 Issue 10, October 2011

[18].Xiaoyan Cai, Wenjie Li, You Ouyang, Hong Yan, “Simultaneous ranking and clustering of sentences: a reinforcement approach to multi-document summarization” Proceeding COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics Pages 134-142 2010.