CLUSTERING TEXT IN SENTENCE LEVEL

[1]G.Nivetha,[2] K.S.Gunavathy
[1]Assistant professor ,[2]Assistant Professor
[1]Commerce With Computer Application
Saiva Bhanu Kshatriya College, Aruppukottai, India

ABSTRACT The sentence clustering is very important in some application of text mining such as single and documents summarization in which a sentence is selected based on information contribution by sentence scoring. The document's sentences are semantically fully related or some degree of overlapping exists among other sentences. Our proposed clustering algorithm utilizes sentence overlapping (relation) in terms of fuzzy relational measurements. Results of applying this algorithm to sentence clustering tasks demonstrate that the algorithm is capable of identifying overlapping clusters of semantically related sentences.

INDEX TERMS **–**Clustering,fuzzy relation,Sentence overlapping,Text mining

I.INTRODUCTION

Sentence clustering plays an important role in many text processing activities. Sentence clustering used for text mining task.By clustering the sentences of the 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.

II.EXISTING SYSTEM

K-means clustering is a method of vector quantization. It is used for cluster analysis in data mining

2.1 K-means Clustering

Sentences are grouped with respect to distance between sentence and centroid of the cluster. A sentence become member of a cluster if it has minimum distance to the cluster than others. A sentence member should be only single cluster. Cluster boundary is crisp. (Sentence cannot be member of more than one cluster) .The clustering quality output depends on initial cluster's centroid value. The cluster output is not consistent. Output changes at every execution.

III.PROPOSED SYSTEM

3.1 Fuzzy relational Clustering

Sentences are grouped with respect to fuzzy membership value between sentence and centroid of the cluster. A sentence membership value measured from parameter probability distribution (expectation maximization) technique. A sentence has a membership value to each cluster. The value refers degree of relationship between the sentence and cluster. Cluster centroid updated with membership values. Cluster boundary is fuzzy. (Sentence can have member to more than one cluster) So that, every centroid's updated proposonal to the membership values simultaneously. The clustering quality output not depends on initial cluster's centroid value. The cluster output is consistent. Even after every execution. The cluster quality will be improved.

IV.PROPOSED ALGORITHM

4.1 SENTENCE SIMILARITY

Here it measures semantic similarity among sentences. The Word Net is a lexical database which returns collection of synsets (synonyms sets) given a word and their pos (parts of speech). The synsets are sentences that have high probability the word being in it. The word hypernyms tree extracted from synsets and measure distance between root word to given word. The Minimum Distance Length(MDL) algorithm is used to measure distance.
Word : car
Hypernyms: car: vehicle : automobile
Number of Sentences in Dataset Collection: N
Similarity matrix : N x N

4.2 FUZZY RELATION MEMBERSHIP & PARAMETER INITIALIZATION

Here we need to initialize some parameters and matrix to start before clustering process. The following three terms need be initialize in clustering: No of clusters denoted by C. Prior assumption value to Expectation maximum.

4.3 FUZZY MEMBERSHIP MATRIX.

The sentence collection is clustered into C number. Prior value for EM denoted by

Pi (i = 1,2,..C) are set by value of 1/C.

Fuzzy membership matrix initialized with random variable of normal distribution value between 0 to 1. And then the values are normalized by column-wise to determine fuzziness of sentence in the clusters.

Membership matrix size = Number of clusters X similarity matrix

N x C x N **(4.1)**

4.4 SENTENCE RANKING

Sentence Ranking is method that helps to find a sentence and arrange their related sentences in order from sentence collection dataset. Vi and Vj the two sentences and their similarity is wij also called weight.

PR(Vi) = (1- d)+d(Sum (Wij x (PR(Vj)/ Sum(Wjk)) **(4.2)**

The ranking gives output measurement how much similarity a sentence have to be member in a cluster.

4.5 CLUSTER LIKELIHOOD ESTIMATION

Each cluster is modeled with a Gaussian normal distribution. The sentences fuzzy membership values to each cluster are measured from these normal distributions. The distribution model parameters mean and standard deviation are updated at every iteration. It will stop iteration when no change in parameter values between current and previous. Each cluster is modeled with a Gaussian normal distribution. The sentences fuzzy membership values to each cluster are measured from these normal distributions. The distribution model parameters mean and standard deviation are updated at every iteration. It will stop iteration when no change in parameter values between current and previous. A sentence has a fuzzy membership value to each cluster. The values which is maximum, is chosen and its corresponding cluster label is assigned to the sentence.

V.RESULT AND ANALYSIS

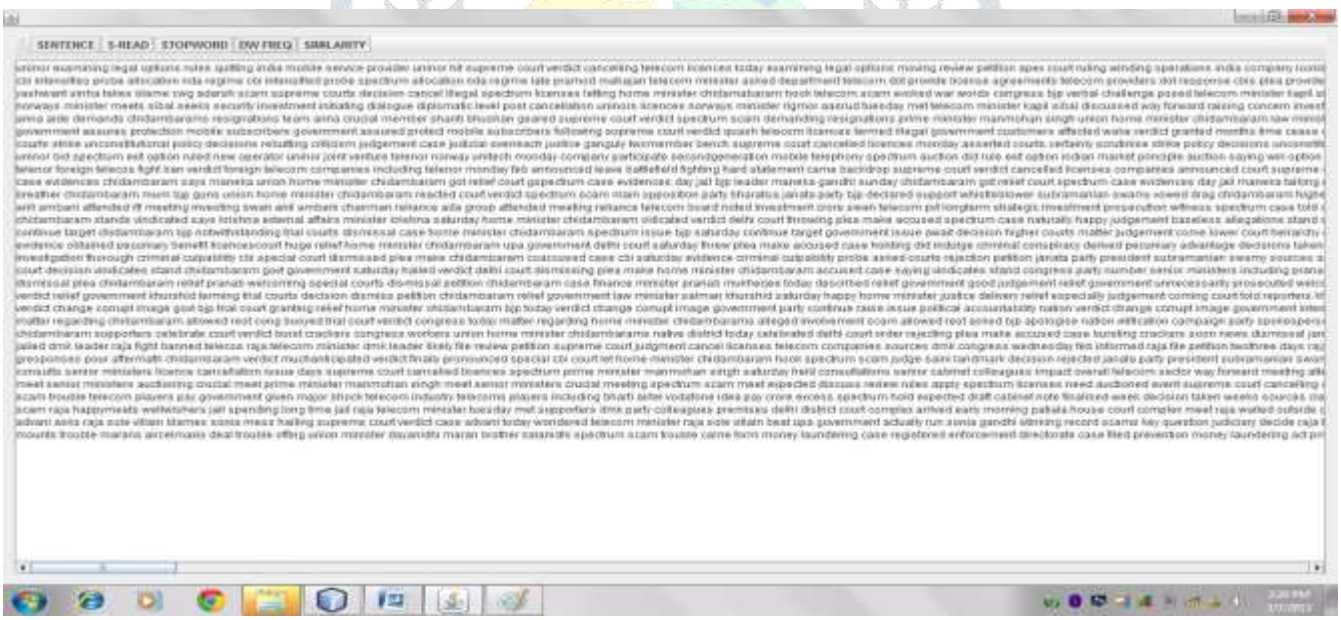

Fig 1. Stop words removal

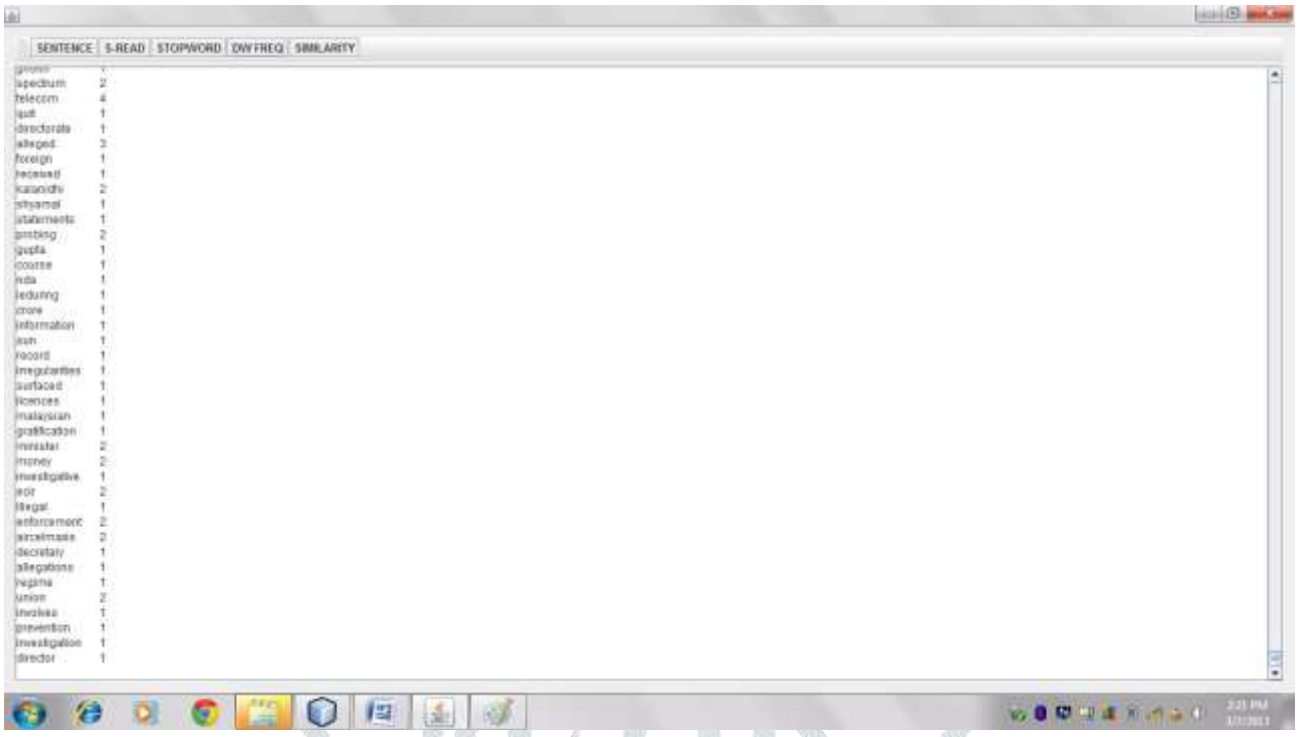

Fig 2. Word Frequency

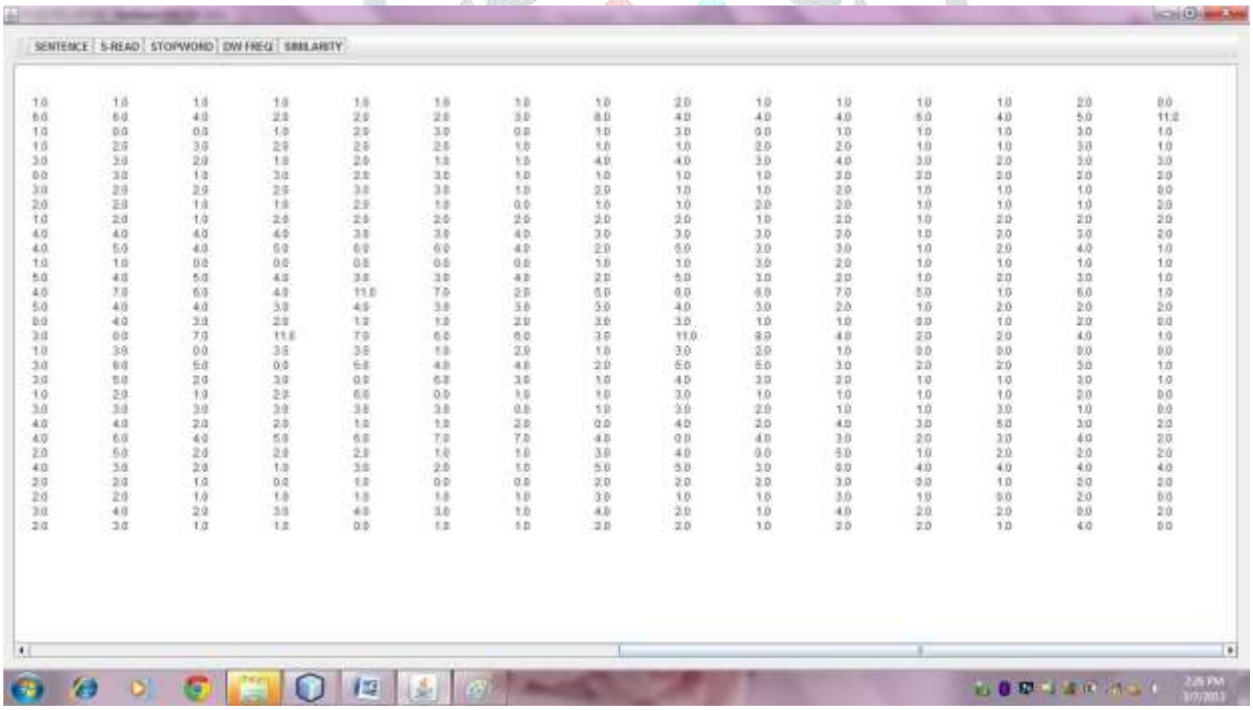

Fig 3. Similarity matrix

VI.CONCLUSION

An obvious potential application of the algorithm is to document summarization; however, the algorithm can also be used within more general text mining settings such as query-directed text mining. Like any clustering algorithm, the performance will ultimately depend on the quality of the input data, and in the case of sentence clustering this performance may be improved through development of better sentence similarity measures, which may in turn be based on improved word sense disambiguation, etc. Any such improvements are orthogonal to the clustering model, and can be easily integrated into it.

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.