IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 04, Issue 05 (May. 2014), ||V6|| PP 45-50

Multi Document summarization using EM Clustering

Kalyani Bhagat[1], M.D.Ingle[2]
[1](PG Student Department of computer engineering JSCOE Pune University,
[2](Associate professor Department of computer engineering)

Abstract: **-** In today's world Users deals with the different data over internet, Document summarization is the process for summarizing the data from the different files from the single folder without losing their semantic content as per user query. Till now researchers has discovered various techniques to summarize the document to achieve the best output .Cluster identification is the one of the most important step for identifying the most relevant sentences from the different files which is further ranked using the ranking algorithm for ranking the top ranked sentences. This process is then further followed by the post processing. Existing clustering identification technique used for clustering does not shown the accuracy while fetching the sentence, hence we are introducing the new technique for cluster identification called EM (Expectation Maximization) which helps to identify the unobserved latent variables from the sentences. Here we are using the manifold ranking based on relevance propagation via mutual reinforcement between sentences and cluster for more.

Keywords: **-** *Clustering, Mutual Reinforcement, Manifold ranking, Relevance propagation, Query based Summarization*

I. INTRODUCTION

With the rapid growing popularity of the web and a variety of information services, obtaining the meaningful information within a short time is in demand. This has becomes a serious problem in the information age. New technologies that can process information efficiently are in great need. Multi document summarization, which is a process of reducing the size of the original documents while preserving their important semantic meaning, is an essential technology to overcome this problem. The main goal for automatic summarization techniques is to produce condensed summary from a set of source documents [1][2]. It aims to create the meaning full summary of the files into its essential content and to assist in filtering and selection of necessary information [2].Several problem faced while fetching the huge data over the web like increase in data complexity, performance degradation, and time consumption while extracting meaningful information, dirty and unorganized structure because data is not filtered properly.

In document summarization process filtering of data is very important as data needs to be fetched from the various files and user query. User query requires well structured data in input files so that final summary will contain the rich information containing the pocket of fruits. User Query/sentences and clusters are mutually reinforced to find the best solution from the given input files [5][12]. Here we are using the new clustering algorithm called Expectation–maximization to improve the clustering accuracy for fetching the sentences.

Importance of data filtration in multi document summarization

Let us discuss some major issues faced while dealing with the web data:-

1. Millions of data over web: - As we know internet is the large source of data ,it contains important information as well as unnecessary files/document which causes side effect like performance degradation, data complexity due to huge size, time consuming while extracting the information.
2. Duplicate / Noisy data: - There are huge records of data present over the web which consist of random input from different sources .for e.g. real time data from debit cards, online payment transaction etc. such data is collected every minute which leads to duplicate ,dirty and unorganized structure because data is not organized properly.
3. Filtering technique for noisy data: - For dirty data from the internet needs to be filtered by good data stage technique which will help system to find best result for the given query. Hence cleaning of data is Importance aspect.

Here in Multi document summarization filtering of data is very important as data needs to be collected from the different files and user query. For user based query, well structured data should be maintained in input files so that summary will contain the rich information for the user. User Query and the sentences are mutually reinforced to find the accurate solution from the given input files [9][1]. This techniques help to reduce the human effort to great extend.

II. LITERATURE SURVEY

Different clustering techniques are discovered for multi document summarization, let us discuss some of them:-

1.1 Summarization Using Cluster-Based Link Analysis

Multi-document summarization by making use of the clusters, this is done by linking the relationships between the sentences in the given document, precondition is that all the sentences are in different from each other. In this, system first constructs a directed or undirected graph to reflect the relationships between the sentences, after this applies the graph-based ranking algorithm which will compute the ranking scores for the sentences. The sentences with large rank scores are chosen for the summary [6].Also the model makes use of the sentences in the bunch of documents, i.e. all sentences are ranked without considering the higher- level information beyond the sentence-level information [5]. The theme clusters close to the main topic of the document set are usually more important than the far away from the main topic of the document set.

Drawback of this approach:

- Relationship between cluster and the ranking sequence is not present [6]
- Accuracy is very less

2.2 Document summarization using spectral analysis clustering approach

A spectral analysis approach developed for simultaneously clustering and ranking of sentences. Datasets demonstrate the improvement of the proposed approach over the other existing clustering-based approaches [10]. This approach ranks sentences simultaneously based on the spectral analysis. This new approach explores the clustering Structure of sentences before the actual clustering algorithm is performed. The special cluster identification structure, called the structure of beams [10], is discovered by analyzing the spectral characteristics of the sentence similarity network. This method defines a natural and healthy relationship between the information necessary for clustering and ranking.

Drawback of this approach:

- Due to this approach ranking performance will be inevitably influenced by the clustering result.[6]

2.3 Summarization using reinforcement approach

In this approach it will tightly integrates ranking and clustering by simultaneously updating each other so that the performance of both can be improved [6]. This approach has shown its robustness and effectiveness. In this approach ranking and clustering are regarded as two independent processes ,although the cluster-level information has been incorporated into the sentence & ranking process ,this results the ranking performance is inevitably influenced by the clustering result .The quality of ranking and clustering both improved when the two processes are mutually enhanced.

2.4 Summarization using mutual reinforcement principle using K –Means clustering

It randomly selects K sentences as the initial centroids of the K clusters and then iteratively assigns all sentences to the closest cluster and re-computes the centroid of each cluster until the centroids do not change. The similarity between the sentence and the cluster centroid is computed by the standard cosine measure.

Drawback of this approach:

- In the initial phase of the clustering user needs to input no of cluster
- It limits the clusters coverage

2.5 Summarization using mutual reinforcement principle using Affinity propagation clustering approach

This approach is different from the above clustering algorithms in this we do not need to provide the cluster number & graph based. The algorithm takes each sentence as a vertex in a graph and considers all the vertices as potential exemplars. After this processes it recursively transmits the real valued messages along edges of the graph until a good set of exemplars and corresponding clusters emerges.

Drawback of this approach:

- This approach does not find the semantic meaning between the sentences in the file and identified clusters

III. PROPOSED MODEL

In proposed model we are using expectation–maximization (EM) algorithm for cluster identification, it is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models ,the model depends on unobserved latent variables. The expectation–maximization iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the is evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found in E step. These parameter estimates are then used to determine the distribution of the latent variables in the next E step.

A new wiki tool called Weka is used for finding the semantic meaning of the sentences can help in cluster identification .For example if the file contains word Countries then all the semantic words of countries like India, UK, and Canada are also fetched in summarization process. Another example of electronics can have different semantic words like Television, Fridge, Microwave or Mixer is also consider in the cluster identification. This has shown the great difference in the result of cluster identification. Experimental results show that the total number of cluster identified by Expectation maximization algorithm is more than Affinity propagation. This increases the accuracy of the cluster result and also identifies the relevant sentences from the given dataset.

3.1 Architecture Diagram:

Based on above technique following proposed framework is discovered shown in figure 3.1 below

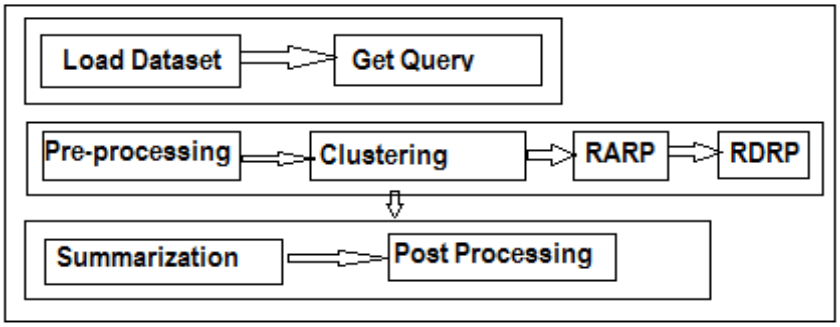

Fig 3.1 Architectural Diagram

Let us discuss the steps in details:-

1. Load database and user Query :- User inputs the query and the dataset which will load all the files
2. Pre processing: - In this process sentences are separated from file, meaning full words are identified. Also the removal of support words, stop words, e.g. commas, full stop.
3. Clustering: - Expectation–maximization clustering is used to identify unobserved latent variables are discovered.
4. RARP :- It stands for Reinforcement after Relevance Propagation (RARP) algorithm. It performs the internal relevance propagation in the sentence set and the cluster set separately until the stable states for both is reached. The obtained cluster and sentences ranking scores are then updated using external mutual reinforcement until all the scores are converged.
5. RDRP :- The second ranking algorithm is called the Reinforcement During Relevance Propagation (RDRP) algorithm, which alternatively performs first round of internal relevance propagation in the cluster set (or the sentence set), and another round of external mutual reinforcement, which will update the current ranking scores of the sentences set (or the cluster set). The same process is repeated until an global stable state is reached.
6. Summarization: -Based on the output from above ranking algorithm the top ranked sentences are identified.
7. Post Processing: - As we are fetching the different documents for summarizing information redundancy problem appear to be more serious in this compare to single-document summarization, hence removal of duplication is done in post processing.

IV. EXPERIMENTAL RESULTS

Experimental results shows that the cluster identified by AF algorithms is less than the EM clustering as shown below:-

4.1Comparison between EM Clustering and AP Clustering

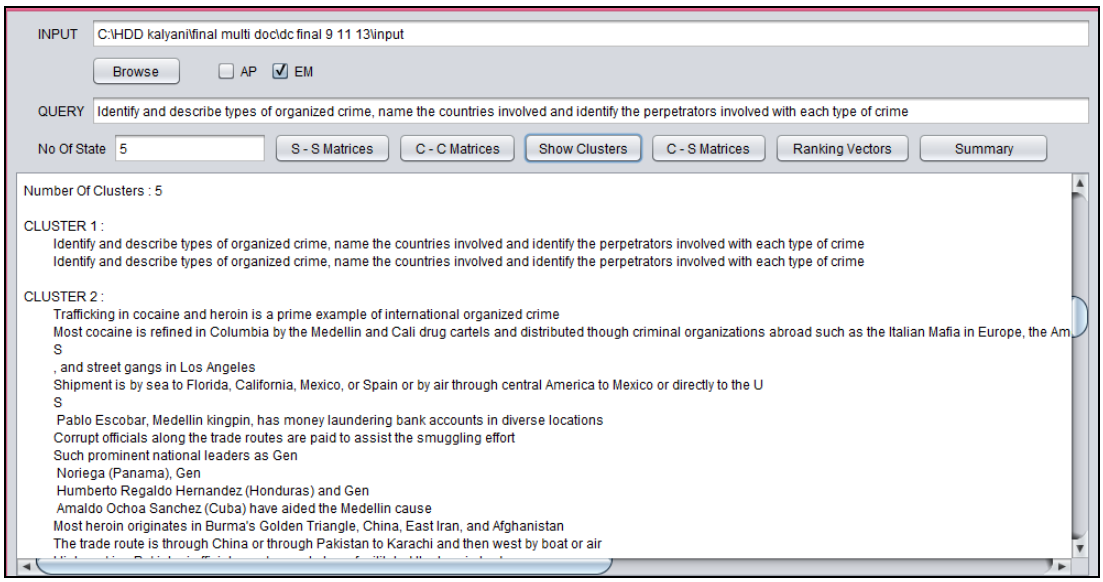

Fig4.1.1 Clusters identified by Expectation–maximization

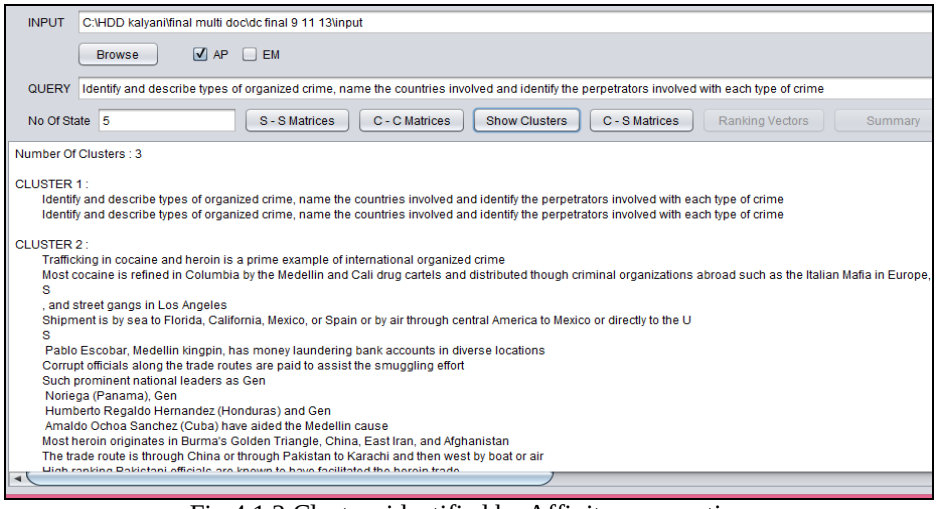

Fig 4.1.2 Clusters identified by Affinity propagation

Both the results are examined on the DUC2007 dataset, and it is observed that the number of cluster are more in Expectation–maximization clustering which is very help full to identify the sentences of semantic meaning

4.2 Result Table for RARP and RDRP

No of statements	RARP		RDRP	
	Precision	Recall	Precision	Recall
5	0.192307692	0.094339623	0.185185185	0.094339623
10	0.384615385	0.188679245	0.37037037	0.188679245
3	0.115384615	0.056603774	0.111111111	0.056603774
8	0.307692308	0.150943396	0.296296296	0.150943396

Fig. 4.2.1 Precision and Recall functions for RARP and RDRP algorithms

4.3 Ranking algorithm comparison between RARP and RDRD

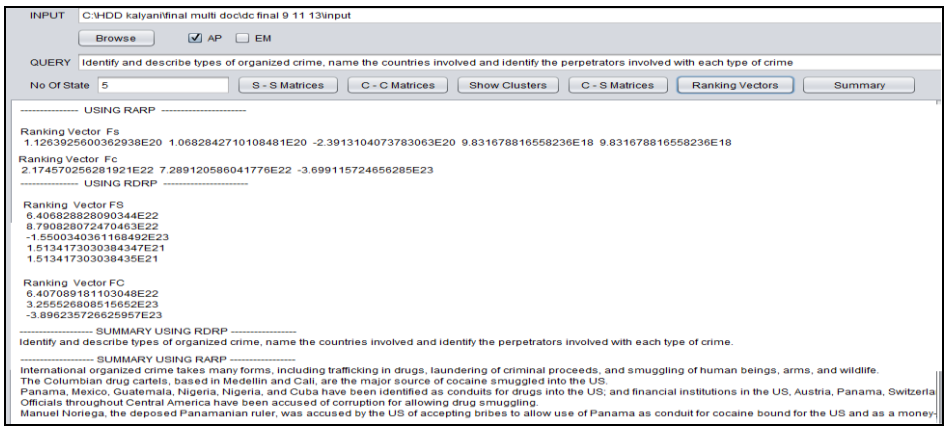

Fig 4.3.1 Ranking algorithm (RARP & RDRP) using AF Clustering

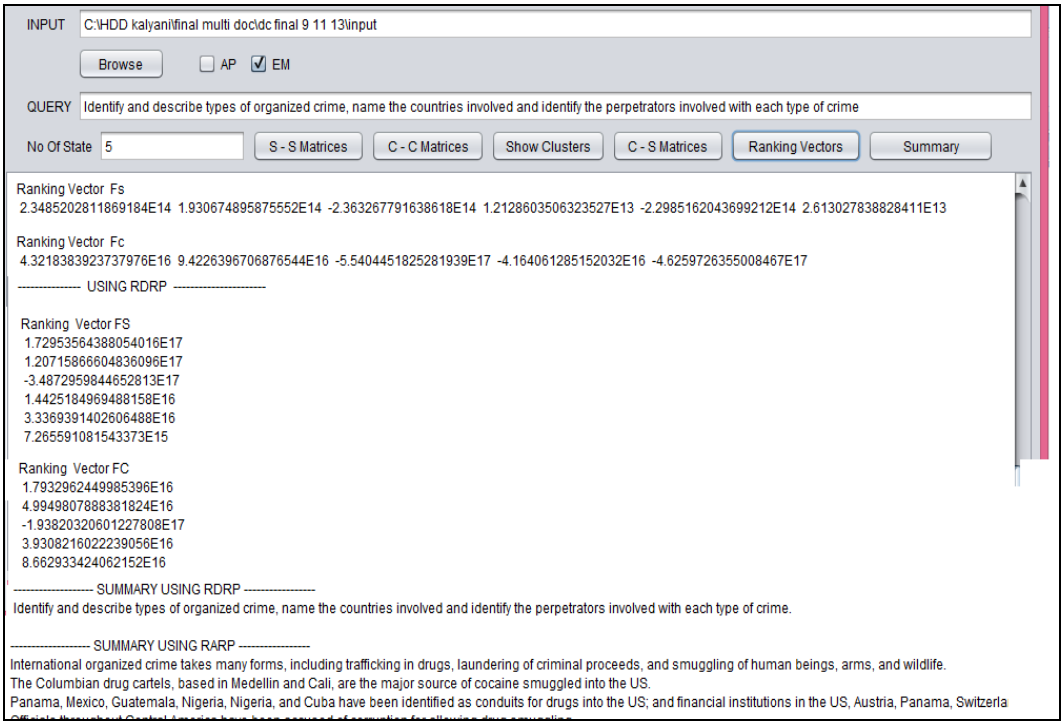

Fig 4.3.2 ranking algorithm (RARP & RDRP) using EM Clustering

4.4 Result table for two clustering algorithms

Sr. No	Cluster method	No of cluster	Matrix dimension	Laplacian Matrix (matching sentences)
1	Affinity propagation	3	3-3	498
2	Expectation maximization	5	5-5	1242

Table 4.4 Clustering algorithm comparison

4.5 Result Graph

Graphs shown below shows the clustering algorithm and total number of sentences identified against them.

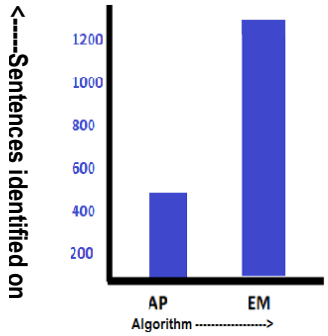

Fig 4.5 Result Graph for sentences identification

V. CONCLUSION

This paper presents a new clustering approach for multi-document summarization system using manifold ranking and mutual reinforcement principle. In this study, expectation–maximization clustering algorithm is used for cluster identification which gives better results than affinity propagation clustering algorithm. RARP and RDRP are two ranking algorithm used to rank the sentence as per user request. Also time taken by EM algorithm is more. In future we will other effective machine learning technique for more accurate results.

VI. ACKNOWLEDGEMENTS

I would like to express my gratitude towards my guide and ME Coordinator Computer Engineering Department Prof. Madhav.D. Ingle for his valuable guidance and encouragement throughout the period this work was carried out. I also thank to the Principal of JSCOE, Dr. M. G. Jadhav for providing me all the necessary facilities to carry out this work

REFERENCES

[1] X. J.Wan, J. W. Yang, and J. G. Xiao, “Manifold-ranking based topic focused multi-document summarization,” in Proc. 18th IJCAI Conf., 2007, pp.2903–2908

[2] S. Harabagiu and F. Lacatusu, “Topic themes for multi document summarization,” in Proc. 28th SIGIR Conf., 2005, pp. 202–209.

[3] Wan X. and Yang J. 2006 "Improved Affinity Graph based Multi-Document Summarization."

[4] R. X.Y. Cai, W.J. Li, in "Simultaneous ranking and clustering of sentences: a reinforcement approach to multi-document summarization, 2010,".

[5] Xiaojun Wan and Jianwu Yang “Multi-Document Summarization Using Cluster-Based Link Analysis 2008”

[6] Xiaoyan Cai, Wenjie Li “A spectral analysis approach to document summarization: Clustering and ranking sentences simultaneously ”inproc X. Cai W. Li / Information Sciences 181 (2011) 3816–3827.

[7] Xiaoyan Cai and Wenjie Li ,” Mutually Reinforced Manifold-Ranking Based Relevance Propagation Model for Query-Focused Multi-Document Summarization” in proc IEEE transaction on audio, speech and language processing , vol 20,no 5 july 2012.

[8] J. F. Bredan and D. Delbert, “Clustering by passing messages between data points,” Science, vol. 315, no. 5814, pp. 972–976, Jan. 2007.

[9] K. F. Wong, M. L. Wu, and W. J. Li, “Extractive summarization using supervised and semi-supervised learning,” in Proc. 22nd OLING Conf., 2008, pp. 985–992.

[10] H. Y. Zha, “Generic summarization and key phrase extraction using mutual reinforcement principle and sentence clustering,” in Proc. 25th SIGIR Conf., 2002, pp. 113–120

[11] R. Barzilay, K.R. Mckeown, in "Sentence fusion for multi-document news summarization, Computational Linguistics 31 (3) (2005) 297327.".

[12] A. Haghighi and L. Vanderwende, "Exploring content models for multi document summarization, in Proc. 10th NAACL-HLT, 2009