International Refereed Journal of Engineering and Science (IRJES)
ISSN (Online) 2319-183X, (Print) 2319-1821
Volume 3, Issue 4(April 2014), PP.75-79

Comparison of clustering technique used in automatic Document summarization

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

Abstract:- In today's world Users deals with the different data over the web, multi document summarization is the process for summarizing the data from the different files without losing their semantic content as per user query. Various techniques has been discovered 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 then supplied to the ranking algorithm for ranking the top ranked sentences followed by the post processing technique. Existing clustering technique used for clustering does not shown the accurate result while sentence fetching hence new technique for cluster identification is introduced called EM (Expectation Maximization) which helps to identify the unobserved latent variables. Here we are using the manifold ranking based on relevance propagation via mutual reinforcement between sentences and cluster.

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

I. INTRODUCTION

With the rapid growing popularity of the internet and a variety of information services, obtaining the meaningful information within a short time is the need. This has becomes a serious problem in the information age. New technologies that can process information efficiently are in great demand. Automatic document summarization, which is a process of reducing the size of documents while preserving their important semantic content, is an essential technology to overcome this problem. The main goal for Multi-document 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 original text into its essential content and to assist in filtering and selection of necessary information [2].Various problem faced while dealing with the huge data over the web are like performance degradation, increase in data complexity, time consuming while extracting the information, dirty and unorganized structure because data is not filtered properly.

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

II. RELATED WORK

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

2.1 Summarization Using Cluster-Based Link Analysis

Multi-document summarization by making use of the cluster, it can be achieved by link relationships between sentences in the given document, assumption that all the sentences are in different from each other. In this model system first constructs a directed or undirected graph to reflect the relationships between the sentences & then applies the graph-based ranking algorithm to compute the rank scores for the sentences. The sentences with large rank scores are chosen for the summary [6].Also the model makes uniform use of the sentences in the document set, i.e. all the 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 theme clusters far away from the main topic of the document set.
Drawback of this approach:
- It does not link relationship between cluster and the ranking sequence [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 clustering structure, called the structure of beams [10], is discovered by analyzing the spectral characteristics of the sentence similarity network. This method reveals a natural 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 Multi-document Summarization using reinforcement approach

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

2.4 Multi document 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:

- Number of cluster needs to input in the initial phase of the clustering
- It limits the clusters coverage

2.5 Multi document 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. It is also graph based. The algorithm takes each sentence as a vertex in a graph and considers all the vertices as potential exemplars. Then 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 similarity 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, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

Weka tool is used for finding the semantic meaning of the sentences can help in cluster identification. For example if the file contains word Flower then all the semantic words of flowers like Rose, Jasmine, and Lily are also fetched in summarization process. Another example of colors can have different semantic words like Blue, Green, White or red are also consider in the cluster identification. This has shown the great difference in the cluster identification. Experimental results conclude that the total number of cluster identified by Expectation maximization algorithm is more than Affinity propagation. This increases the accuracy of the clusters and also identifies the relevant sentences from the given dataset.

3.1 Architecture Diagram:

Based on above technique we proposed framework as shown in figure 3.1 below.

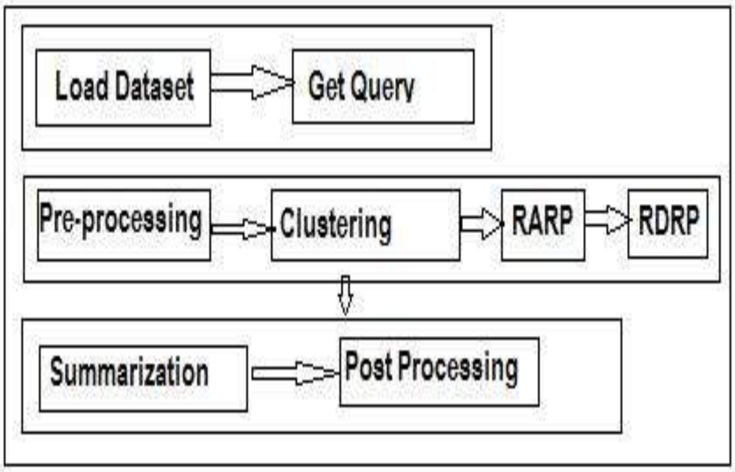


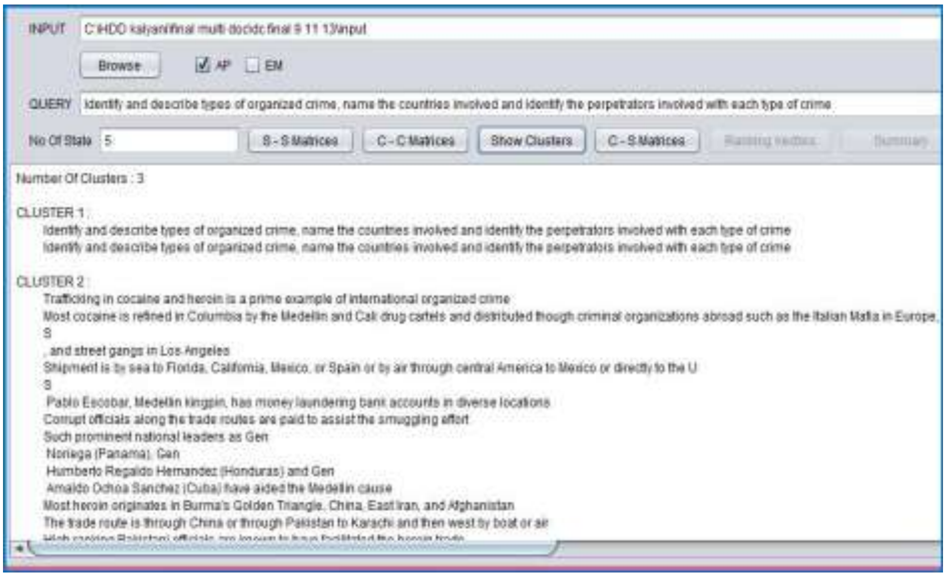
Fig 3.1 Architectural Diagram

Let us discuss the steps in details:-

1. Load dataset / Query :- Initially user inputs the query ,also the dataset which will load all the files
2. Pre processing: - In this process separation of sentence from file, meaning full words are identified. Also the removal of stop words, like commas, full stop takes place.
3. Clustering: - EM clustering is used to identify the clusters in which the unobserved latent variables are discovered.
4. RARP (Ranking algorithm):-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 of both are reached. The obtained sentence and cluster ranking scores are then updated via external mutual reinforcement until all the scores are converged.
5. RDRP(Ranking algorithm):- The second algorithm is called the Reinforcement During Relevance Propagation (RDRP) algorithm, which alternatively performs one round of internal relevance propagation in the sentence set (or the cluster set), and one round of external mutual reinforcement to update the current ranking scores of the cluster set (or the sentence set). The whole process is iterated until an overall global stable state is reached.
6. Summarization: - Depends on the output from Ranking algorithm the top ranked sentences are identified.
7. Post Processing: - The number of the documents to be summarized can be very large. This makes information redundancy problem appear to be more serious in multi-document summarization than in single-document summarization ,hence removal of duplication needed.

IV. EXPERIMENTAL RESULTS

Experimental results shown that the cluster identified by EM algorithms is more than the affinity propagation as shown below:-



INPUT C:\HDD kalyani\final multi doc\dc final 9 11 13\input

☑ AP □ EM

QUERY Identify and describe types of organized crime, name the countries involved and identify the perpetrators involved with each type of crime

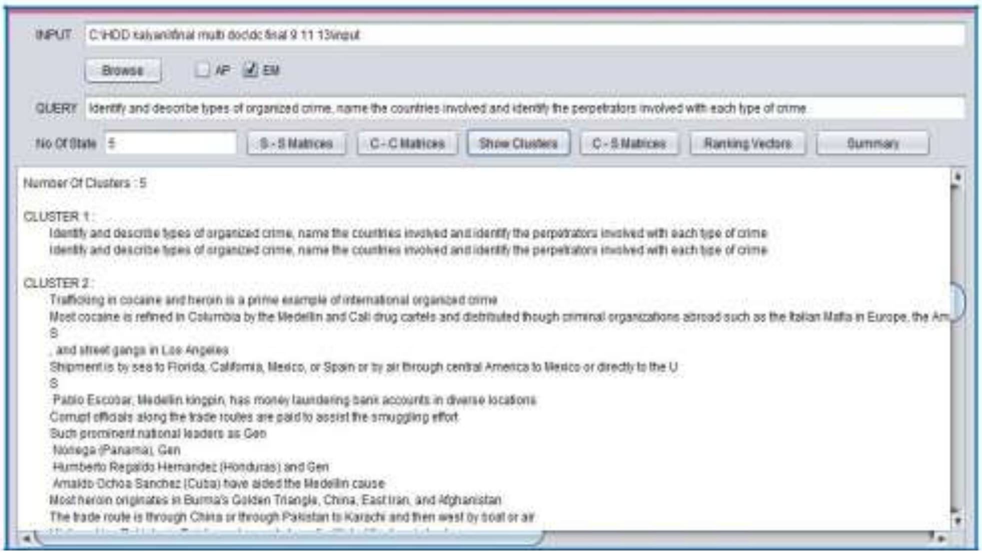
No Of State 5

```
Number Of Clusters : 3

CLUSTER 1 :
     Identify and describe types of organized crime, name the countries involved and identify the perpetrators involved with each type of crime
     Identify and describe types of organized crime, name the countries involved and identify the perpetrators involved with each type of crime

CLUSTER 2 :
     Trafficking in cocaine and heroin is a prime example of international organized crime
     Most cocaine is refined in Columbia by the Medellin and Cali drug cartels and distributed though criminal organizations abroad such as the Italian Mafia in Europe,
     S
     , and street gangs in Los Angeles
     Shipment is by sea to Florida, California, Mexico, or Spain or by air through central America to Mexico or directly to the U
     S
      Pablo Escobar, Medellin kingpin, has money laundering bank accounts in diverse locations
     Corrupt officials along the trade routes are paid to assist the smuggling effort
     Such prominent national leaders as Gen
      Noriega (Panama), Gen
      Humberto Regaldo Hernandez (Honduras) and Gen
      Arnaldo Ochoa Sanchez (Cuba) have aided the Medellin cause
     Most heroin originates in Burma's Golden Triangle, China, East Iran, and Afghanistan
     The trade route is through China or through Pakistan to Karachi and then west by boat or air
     High ranking Pakistani officials are known to have facilitated the heroin trade
```

Fig 4.1 Clusters identified by Affinity propagation



INPUT: C:\HOD kalyani\final multi doc\dc final 9 11 13\input

☐ AP ☑ EM

QUERY: Identify and describe types of organized crime, name the countries involved and identify the perpetrators involved with each type of crime

No Of State: 5

```
Number Of Clusters : 5

CLUSTER 1 :
     Identify and describe types of organized crime, name the countries involved and identify the perpetrators involved with each type of crime
     Identify and describe types of organized crime, name the countries involved and identify the perpetrators involved with each type of crime

CLUSTER 2 :
     Trafficking in cocaine and heroin is a prime example of international organized crime
     Most cocaine is refined in Columbia by the Medellin and Cali drug cartels and distributed though criminal organizations abroad such as the Italian Mafia in Europe, the Am
     S
     , and street gangs in Los Angeles
     Shipment is by sea to Florida, California, Mexico, or Spain or by air through central America to Mexico or directly to the U
     S
      Pablo Escobar, Medellin kingpin, has money laundering bank accounts in diverse locations
     Corrupt officials along the trade routes are paid to assist the smuggling effort
     Such prominent national leaders as Gen
      Noriega (Panama), Gen
      Humberto Regaldo Hernandez (Honduras) and Gen
      Arnaldo Ochoa Sanchez (Cuba) have aided the Medellin cause
     Most heroin originates in Burma's Golden Triangle, China, East Iran, and Afghanistan
     The trade route is through China or through Pakistan to Karachi and then west by boat or air
```

Fig4.2 Clusters identified by Expectation–maximization

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

4.3 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

Table 4.3 Ranking algorithm comparison

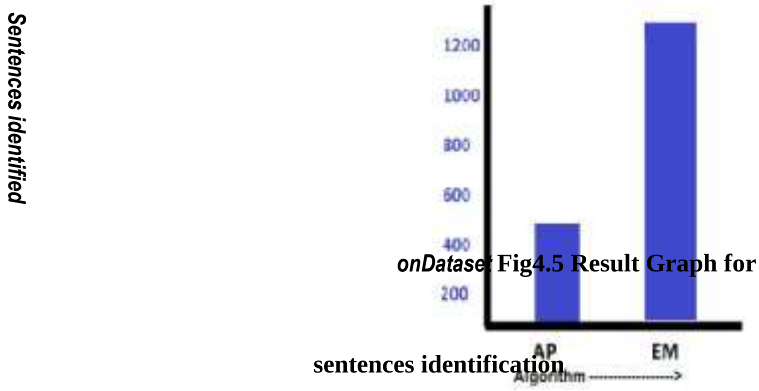
4.4 Result table for two clustering algorithms

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 total number of sentences identified by each algorithm while identifying the clusters



onDataset Fig4.5 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.

ACKNOWLEDGEMENTS

I would like to express our profound sense of gratitude towards my guide and ME Coordinator Computer Engineering Department Prof. M.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

Journal Papers:

[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. Highlight and L. Vanderwende, "Exploring content models for multidocument summarization, in Proc. 10th NAACL-HLT, 2009".