Query Based Multi-Document Summarization Using Manifold Ranking Approach

Ms. Savita Badhe[1]
[1] Department of Information Technology
[1]AISSMS Institute of Information Technology

Abstract- *Query based multi-document summarization aims to generate a summary related to given query or topic. This paper presents a new multi-document summarization system using manifold ranking and mutual reinforcement approach. Manifold-ranking has been recently conquered for query-based summarization. This approach uses relationship between given query and the sentences and sentence to sentence. In this paper, a model is proposed to improve manifold-ranking based relevance propagation via mutual reinforcement between sentences and clusters. The document consists of number of sentences which can be grouped into theme cluster of related sentences. The proposed model uses affinity propagation clustering algorithm for cluster identification and sentence ranking algorithms to rank the sentences based on mutual reinforcement approach. The summary is created by extracting sentences with high ranking score and by removing redundancy.*

Keywords- Multi-document summarization, manifold ranking, cluster identification, mutual reinforcement, sentence ranking.

I. INTRODUCTION

Multi-document summarization generates a summary of related documents. Researchers mainly focus on extracting and presenting the most important content from documents. In recent years, rapid growth of the web, the huge amounts of information make it more difficult to efficiently access the useful information. Thus, there is need to automatically compress the information covering multiple documents and present the summary to the users would help out to solve this problem. Query-based multi-document summarization aims to form a summary from document set related to given query. As compared with generic multi-document summarization, the challenge for query-based multi-document summarization is that a query based summary expected to give important information contained in the document set as well as expected to guarantee that the information is related to the given query. Therefore, we need effective method to take into account this query-based technique during summarization process. The main objective of this paper is to produce an effective summary relevant to the user's query from the given set of the documents. While search engines were developed to deal with this huge volume of documents, even they output a large number of documents for a given user's query. With the rapid growing popularity of the Internet obtaining the desired information within a short amount of time becomes a serious issue in the information era. Automatic multi-document summarization, i.e. a process of reducing the size of documents while preserving their important content, is an essential technology to overcome this problem.

Among the existing methods for query based multi-document summarization, manifold ranking based approach is an efficient way to enforce the query's impact on sentence ranking. In recent times, manifold ranking approach has been subjugated for query based summarization. This ranking approach builds a weighted graph that represents query and sentences as vertices. Some predefined positive ranking score is assigned to related sentences and the query. Each and every sentence is ranked according to its ranking score. Higher is the ranking score higher is the chance for extraction. This approach also performs the relevance propagation between the sentences. Actually document set consist of number of clusters of related sentences. Theses clusters may have different size and importance. In the whole document set , cluster which is related to user's query is more important than cluster which is not related to user's query. So the cluster information has more impact on sentence ranking. Based on the analysis of cluster, we prove that the ranking score of sentence rely on its relationship with user's query as well as relationship between clusters to the query. So in this paper we apply mutual reinforcement principle.

In short, the main contributions of this paper are two folds: First, we propose an affinity propagation clustering algorithm for cluster identification. Second, two query based sentence ranking algorithms based on model.

The rest of the paper organized as follows. Section II briefly reviews related work on query based summarization. Section III describes the proposed system. Section IV concludes this paper.

II. RELATED WORK

A. *Extractive summarization methods*

The various summarization approaches are either abstractive or extractive. In abstractive summarization abstract of the document is created. It creates a summary which involves

sentence fusion, sentence compression and reformulation. Extractive summarization on other hand assigns a score to sentences and extracts sentences with highest score to create a summary. It usually ranks the sentences according to their scores calculated by statistical and linguistic features such as sentence position or term position, term frequency inverse sentence frequency (tf.isf). Extractive summarization falls into two categories generic summarization and query based summarization. Generic summarization extracts a summary about the general topics in the documents and query based summarization extracts important information from the documents and guarantees that the extracted information is related to the given query.

B. *Manifold Ranking Approach*

The manifold ranking method makes uniform use of sentence to sentence relationship and the sentence to query relationship. The manifold-ranking [1] based summarization approach consists of two steps: In the first step, manifold-ranking score is computed for every sentence and the score denotes the biased information richness of sentence. In second step, the diversity penalty is imposed on each sentence and the overall ranking score of each sentence is obtained to return both the biased information richness and the information uniqueness of the sentence. The sentences with high overall ranking scores are chosen for the summary. The manifold-ranking method is a universal ranking algorithm and it is initially used to rank data points along their underlying manifold structure. The prior assumption of manifold-ranking is: (1) nearby points have the same ranking scores; (2) points on the same structure have the same ranking scores. An intuitive description of manifold-ranking [11] is as follows:

1. Manifold-ranking based summarization approach builds a graph that represents query and sentences as vertices.
2. The pre-defined positive score of query is then propagated to nearby vertices via the graph iteratively until a global stable state is achieved.
3. At the last part, all the sentences are ranked according to their final scores. Higher is the ranking score higher is the chance for extraction
4. This approach performed relevance propagation among the sentences. In a given document set, there usually exist a number of topics with each theme represented by a cluster of related sentences.
5. The clusters are of different size and especially different importance to assist the users in understanding the content in the whole document set. So the cluster level information is supposed to have great influence on sentence ranking.
6. Based on the above analysis, we argue that the ranking score of a sentence depends on its relevance to the given query as well as on the relevance of its belonging cluster to the query.

C. *Mutual Reinforcement Principle*

Zha proposed a mutual reinforcement principle that was competent of extracting significant sentences and key phrases at the same time [14]. In Zha's work, a weighted document graph was constructed by connecting together the sentences in a document and the terms appearing in those sentences. We apply mutual reinforcement principle to query-based sentence and cluster ranking. The mutual reinforcement principle says that "A sentence should be ranked higher if it is contained in the cluster which is more relevant to the given query while a cluster should be ranked higher if it contains many sentences which are more relevant to the query."

III. THE PROPOSED SYSTEM

This paper proposes a new approach for multi-document summarization using manifold ranking and mutual reinforcement principle.

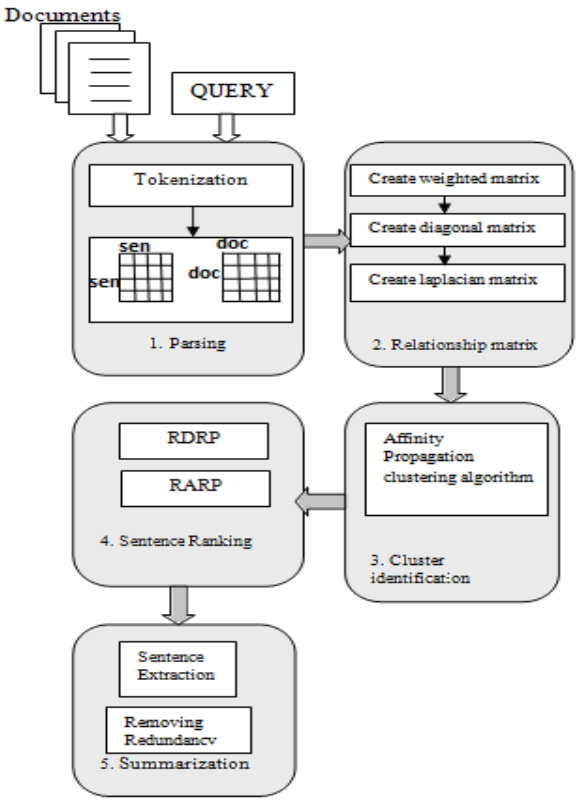

Figure 1. Overall process of the proposed system

The summarization system is designed with extractive framework. The overall process is shown in Figure 1. The query based summarization consists of the following major phases:

A. Parsing

Parse the document by using delimiter. Preprocess the document by removing stop words and stemming. Tokenize each sentence of the document and construct the weighted matrix of sentence to sentence and document to document.

B. Relationship Matrix

This phase creates the weighted matrix, diagonal matrix and laplacian matrix of sentence to sentence and cluster to cluster relationship.

C. Cluster Identification

Cluster identification groups the sentences in the documents into a number of clusters. Affinity propogation algorithm is used for cluster identification. This algorithm identifies the set of "exemplars" between the data points and forms clusters of the data points. This algorithm operates by taking into consideration all data points as potential exemplars and exchanges messages between data points until convergence. The pair-wise similarity between data points is the input for affinity propagation clustering algorithm. Similarities compose of N X N similarity matrix S such: s[i, j] denotes the similarity between data point i and j. In affinity propagation algorithm, two kinds of messages are exchanged between data points Responsibity and Availability. Responsibity r[i, j] is the message from data point i to data point j. It reflects the accumulated evidence that data point j should be exemplar for data point i. Availability a[i, j] is the message from data point j to data point i. It reflects the accumulated evidence that data point i should choose the data point j to be its exemplar and consider the values for other data points that j should be an exemplar. All values for responsibilities and availabilities are set to zero, and calculation of each iterates until convergence.

$$r[i, j] = (1- \lambda)\, \rho[i, j] + \lambda\, r[i, j] \quad (1)$$
$$a[i, j] = (1- \lambda)\, \alpha[i, j] + \lambda\, a[i, j] \quad (2)$$

where, λ is a damping factor to avoid numerical oscillations, and $\rho[i, j]$ and $\alpha[i, j]$ are, propagating responsibility and propagating availability, respectively. The values of $\rho[i, j]$and $\alpha[i, j]$ are computed by the following equations:

$$\rho[i, j] = \begin{cases} s[i,j] - \max_{k \neq j} \{a[i,k] + s[i,k]\} & (i \neq j) \\ s[i,j] - \max_{k \neq j} \{s[i,k]\} & (i = j) \end{cases} \quad (3)$$

$$\rho[i, j] = \begin{cases} s[i,j] - \max_{k \neq j} \{a[i,k] + s[i,k]\} & (i \neq j) \\ s[i,j] - \max_{k \neq j} \{s[i,k]\} & (i = j) \end{cases} \quad (4)$$

That is, messages between data points are computed from the corresponding propagating messages. The final exemplar of data point i defined as:

$$\arg\max \{r[i, j] + a[i, j] : j = 1, 2, \ldots\ldots\ldots, N\} \quad (5)$$

This algorithm requires $O(N^2T)$ time to update messages.

Here in above algorithm,
N = Number of data points
T = Number of iterations
λ = Damping factor, $0 \leq \lambda < 1$
$s[i, j]$ = Similarity between data point i and j
$r[i, j]$ = Responsibility between data point i and j
$a[i, j]$ = Availability between data point i and j
$\rho[i, j]$ = Propagating responsibility between data point i and j
$\alpha[i, j]$ = Propagating availability between data point i and j

D. *Sentence Ranking*

Sentence ranking is a vital module in the extractive summarization system. The proposed model consists of both internal relevance propogation and external mutual reinforcement. As for internal relevance propagation, the manifold-ranking approach is applied to either the set of sentences or the set of clusters, i.e., we construct a weighted network for each set, where the vertices represent the query and the sentences (or the clusters). Initially, a positive rank score is assigned to the query point and zeros to the remaining sentence (or cluster) points. All the sentence (or cluster) points then spread their ranking scores to their nearby neighbors via the weighted network. As for external mutual reinforcement, the ranking scores of one set are refined by the ranking scores of the other set via their formulated links. These two processes can be carried out sequentially or in combination until a global stable state is achieved, in which all the sentence points obtain their final ranking scores. On this basis, Xiaoyan Cai and Wenjie Li develop two corresponding ranking algorithms [1]. The first one is called the 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. 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.

RARP Algorithm Steps

1. Get sentence S, get query Q
2. Create theme cluster C
3. Create constant vector Yc and Ys
4. Create laplacian matrix for cluster and sentence
5. Create symmetric matrix for cluster and sentence
6. Create diagonal matrix for cluster and sentence
7. Calculate initial ranking score vector
8. Normalized cluster-to-sentence adjacency matrix Lcs
9. Get first sentence score Fs

10. Calculate variance v
11. Get transpose of Lcs and multiply with v and previous cluster score and add into first cluster score
12. Get first cluster score Fc
13. Calculate variance v
14. Get Lcs and multiply with v and new sentence score and add into first cluster score
15. Get difference among previous and current sentence score
16. Get difference among previous and current cluster score
17. Find maximum from difference
18. If this difference is less than threshold goto step 12
19. Else stop

RDRP Algorithm

1. Get sentence S, get query Q
2. Create theme cluster C
3. Create constant vector Yc and Ys
4. Create laplacian matrix for cluster and sentence
5. Create symmetric matrix for cluster and sentence
6. Create diagonal matrix for cluster and sentence
7. Calculate initial ranking score vector
8. Normalized cluster-to-sentence adjacency matrix Lcc
9. Normalized cluster-to-sentence adjacency matrix Lss
10. Normalized cluster-to-sentence adjacency matrix Lcs
11. Calculate new sentence score
    a. User normalized matrix for sentence to sentence
    b. First sentence score for sentence
12. Calculate new cluster score
    a. User normalized matrix for cluster to cluster
    b. First sentence score for cluster
13. Get difference between previous and current sentence score
14. Get difference between previous and current cluster score
15. Find maximum from difference
16. If this difference is less than threshold goto step 12
17. Else stop

E. *Summarization*

Summarization phase consist of sentence extraction and redundancy removal. In multi-document summarization there are large numbers of documents to be summarized. This causes information redundancy problem. At the beginning, we choose the first sentence from the ranking list. Then we examine the next one and compare it with the sentence(s) already present in the summary. The sentence that is not too similar to any sentence in the summary (i.e., the cosine similarity between them is lower than a threshold) is selected into the summary. This process is repetitive until the length of the sentences in the summary reaches the length limitation.

IV. CONCLUSION

This paper presents a query based multi-document summarization system using manifold ranking and mutual reinforcement principle. In this study, graph based affinity propogation clustering algorithm is used for cluster identification which gives better results than other clustering algorithm. Also the RDRP algorithm works better than the RARP algorithm.

REFERENCES

[1] Xiaoyan Cai and Wenjie Li "Mutually Reinforced Manifold-Ranking Based Relevance Propagation Model for Query-Focused Multi-Document Summarization". IEEE Transactions On Audio, Speech, And Language Processing, Vol. 20, No. 5, July 2012.

[2] Savita Badhe , Prof K.S. Korabu. "A New Approach for Multi-document Summarization" The International Journal of Engineering And Science (IJES) Volume1, Issue,2,Pages 280-282, 2012.

[3] Adam L. Berger and Vibhu O. Mittal. Query-Relevant Summarization Using FAQs. In Proceedings of Association for Computational Linguistics ACL 2000, pages 294{301, 2000.

[4] Jiayin Ge, Xuanjing Huang, and LideWu. Approaches to Event-Focused Summarization Based on Named Entities and Query Words. In Prceedings of Document Understanding Conferences, 2003.

[5] Judith D. Schlesinger and Deborah J. Baker. Using Document Features and Statistical Modeling to Improve Query-based Summarization. In Proceedings of Workshop on Document Understanding Conferences, DUC01, New Orleans, LA, 2001.

[6] Anastasios Tombros and Mark Sanderson. Advantages of Query Biased Summaries in Information Retrieval. In Research and Development in Information Retrieval, pages 2{10, 1998.

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

[8] D. R. Radev, H. Y. Jing, M. Stys, and D. Tam, "Centroid-based summarization of multiple documents," Inf. Process. Manage., vol. 40, pp. 919–938, Nov. 2004.

[9] Yasuhiro Fujiwara, Go Irie, Tomoe Kitahara. "Fast Algorithm for Affinity Propagation", In Proceedings Of The Twenty Second International Joint Conference In Artificial Intelligence.

[10] S. Harabagiu and F. Lacatusu, "Topic themes for multi-document summarization," in Proc. 28th SIGIR Conf., 2005, pp. 202–209.[10] K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," ACM Comput. Surv., vol. 31, no. 3, pp. 264–323, Sep. 1999.

[11] X. J. Wan and J. G. Xiao, "Graph-based multi-modality learning for topic-focused multi-document summarization," in Proc. 20th IJCAI Conf., 2009, pp. 1586–1591.

[12] 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.

[13] F. R.Wei, W. J. Li, Q. Lu, and Y. X. He, "Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization," In Proc. 31st SIGIR Conf., 2009, pp. 283–290.

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

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

[16] D. Y. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Scholkopf, "Ranking on data manifolds," in Proc. 17th NIPS Conf., 2003, pp. 169–176.

[17] A. Celikyilmaz and D. Hakkani-Tur, "A hybrid hierarchical model for multi-document summarization," in Proc. 48th Annu. Meeting Assoc. Comput. Linguist., ACL, 2010.