The International Journal of Engineering And Science (IJES)
||Volume|| 1 ||Issue|| 2 ||Pages|| 280-282 ||2012||
ISSN: 2319 – 1813 ISBN: 2319 – 1805

A New Approach For Multi-Document Summarization

[1] Savita P. Badhe, [2] Prof. K. S. Korabu
[1, 2]Department of Information Technology, Sinhgad College of Engg, Pune-041

----------ABSTRACT----------
This paper presents a new multi-document summarization system using manifold ranking and mutual reinforcement approach. Manifold-ranking has been recently subjugated for query-based summarization. It propagates the relevance from query to the document sentences by making use of both the relationships among the sentences and the relationships between the given query and the sentences. Set of document covers a number of topic themes. In this a model is proposed to enhance manifold-ranking based relevance propagation via mutual reinforcement between sentences and clusters. The proposed model uses two new sentence ranking algorithms, namely the reinforcement after relevance propagation algorithm and the reinforcement during relevance propagation algorithm.

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

Date of Submission: 11, December, 2012 ⟺ Date of Publication: 25, December 2012

1. INTRODUCTION

Multi-document summarization produces a summary of collection of related documents. Researchers mainly focus on extracting and presenting the most important content from documents. However in recent, with the rapid growth of the Internet, the massive amounts of information make it more difficult to efficiently access the usable information. Thus, the ability to automatically compress the information covering multiple documents and present the summary to the users would help to solve this problem. Query-based multi-document summarization aims to create a summary from document set which answers the need for information articulated in given query. As compared with generic multi-document summarization, the challenge for query-based multi-document summarization is that a query focused summary is not only expected to give important information contained in the document set but also is expected to guarantee that the information is related to the given topic. 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.

A huge amount of on-line information is available on the web, and is still growing. 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. Under these circumstances it became very difficult for the user to find the document he actually needs, because most of the users are reluctant to make the cumbersome effort of going through each of these documents. Therefore systems that can automatically summarize one or more documents are becoming increasingly desirable. With the rapid growing popularity of the Internet and a variety of information services, obtaining the desired information within a short amount of time becomes a serious problem in the information age. Automatic document summarization, i.e. a process of reducing the size of documents while preserving their important semantic content, is an essential technology to overcome this obstacle.

The rest of the paper organized as follows. Section 2 briefly describes related work. Section 3 introduces the proposed system. Section 4 concludes the paper.

2. RELATED WORK

A. Extraction- based Document summarization methods

Ranks sentences according to various criteria and selects the top-ranked sentences from the original documents to form the summaries. Extraction-based summarization falls into two basic categories:

- *Generic summarization*: Extract a summary about the general ideas (or topics) in the documents
- *Query-focused summarization:* extracts the most important information conveyed in the documents but also guarantees that the extracted information is biased towards the given query.

B. Manifold Ranking Approach:

The manifold-ranking [1] based summarization approach consists of two steps*:*

1. The manifold-ranking score is computed for each sentence in the manifold-ranking process where the score denotes the biased information richness of a sentence;
2. Based on the manifold-ranking scores, the diversity penalty is imposed on each sentence and the overall ranking score of each sentence is obtained to reflect both the biased information richness and the information novelty 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 are likely to have the same ranking scores; (2) points on the same structure (typically referred to as a cluster or a manifold) are likely to have the same ranking scores. An intuitive description of manifold-ranking [10] is as follows:

1. Manifold-ranking based summarization approach constructs a weighted graph that explicitly represents both query and sentences as vertices.
2. The pre-specified positive ranking score of query is then propagated to nearby vertices via the graph iteratively until a global stable state is achieved.
3. At the end, all the sentences are ranked according to their final scores, with a larger score indicating a higher chance to be extracted.
4. However, this approach performed relevance propagation among the sentences. The information beyond the sentence level is totally ignored. Actually, in a given document set, there usually exist a number of themes (or topics) with each theme represented by a cluster of highly related sentences.
5. The theme 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 not only on its relevance to the given query, but also on the relevance of its belonging cluster to the query. We apply mutual reinforcement principle to query-focused sentence and theme cluster ranking.

C. Mutual Reinforcement principle:

Zha proposed a mutual reinforcement principle that was capable of extracting significant sentences and key phrases at the same time [13]. In his work, a weighted bipartite document graph was constructed by linking together the sentences in a document and the terms appearing in those sentences. We apply mutual reinforcement principle to query-based sentence and theme cluster ranking i.e.,

"A sentence should be ranked higher if it is contained in the theme cluster which is more relevant to the given query while a theme cluster should be ranked higher if it contains many sentences which are more relevant to the given query."

3. THE PROPOSED SYSTEMS

This paper proposes a new approach for multi-document summarization using manifold ranking and mutual reinforcement principle. Our summarization system is designed with extractive framework. Summarization system is divided into four phases: Pre-processing, cluster Identification, sentence ranking, and summarization. The overall process is shown in Figure 1.

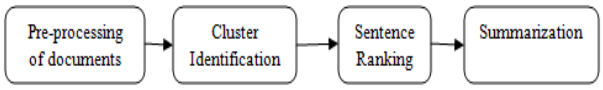

Figure1: Summarization System.

A. Pre-processing:

Pre-processing of multiple documents plays important role for improving the accuracy. In this we performed parsing of the documents related to the given query by user. After parsing document-query relationship matrix is constructed. Different weighting scheme can be used at this stage.

B. Cluster Identification:

Cluster identification groups the sentences in the documents into a number of theme clusters. Affinity propogation algorithm is used for cluster identification.

- ***Affinity Propogation algorithm:***

Affinity Propagation (AP) [6] is different from k-means clustering algorithms in that it does not have to predefine the cluster number. It is a 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.

- Take data point as a node.
- Consider all data points as potential cluster centers .
- Start the clustering with similarity between pair of data points.
- Exchange messages between data points until good cluster centers are found.

The all limitations of k-means clustering algorithm is overcome by affinity propogation algorithm. Thus it is better for cluster identification.

C. Sentence Ranking:

Sentence ranking algorithm is a vital component 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 based algorithm 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 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. The above 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.

D.Summarization

Summarization phase consist of sentence extraction and redundancy control. In multi-document summarization there are large numbers of documents to be summarized. This makes information redundancy problem. At the beginning, we choose the first sentence from the ranking list into the summary. Then we examine the next one and compare it with the sentence(s) already included in the summary. Only 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 repeated until the length of the sentences in the summary reaches the length limitation.

4. CONCLISION AND FUTURE SCOPE

This paper presents a new 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 k-means clustering algorithm. In future we will other effective machine learning technique for more accurate results.

.

ACKNOWLEDGMENT

This is a small review of my M.E. thesis work that I am going to implement. I specially thank to my guide for her assistance.

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] 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.

[3] 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.

[4] 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.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

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

Biographies

Prof. K. S. Korabu

Is currently working as an Assistant Professor in Sinhgad College of Engineering, Pune, India. Her research interests are Data Structures, Database Management Systems, Software Engineering, Document summarization, Data Mining etc.

Savita P. Badhe

Pursuing M.E.(IT) from Sinhgad College of Engineering, Pune, India. Her research interests are Document summarization, Data Mining, Network Security etc.