ISSN:0975-9646

Ranking With Cluster-Based Non-Segregated Approach to Multi-Document Categorization

T.Sathish kumar, V.Sharmila

Department of computer Science and Engineering,
K.S.R College of Engineering ,Tamilnadu ,India

*Abstract:***To summarization of one or more document aims to create a strong summary while retaining the main characteristics of the original set of documents. To cover a number of topic with each theme represented by a cluster of highly related sentences. Sentence clustering is used, it directly generates clusters integrated with ranking. Ranking distribution for sentence in each and every cluster is different in nature which may serve as features of clusters and new clustering measures of sentences can be calculated. To improve the performance of summarization, we will focus on the Influence of document and Proper information Such as Document cluster and Topic query.**

Keywords:SentenceRanking(globalranking,conditonalranking),sentence clustering, MultiDocument summarization

I. INTRODUCTION

The exponential growth in the volume of documents available on the Internet brings the problem of finding out whether a single document can meet a user's complex information need. In order to solve this problem, multi-document summarization which reduces the length of a collection of documents while preserving their important semantic content is highly demanded. Most of the summarization work done till date follow the sentence extraction framework [1], which is governed by importance of information and coherence. We focus on the former issue in this paper. Sentence ranking is a technique of detecting importance of information in the sentence extraction framework. It ranks sentences according to various pre-specified criteria and selects the most salient sentences from the original documents to form summaries. In other words, sentence ranking is one of the most important issues in extraction- based document summarization framework.

The cluster-based ranking approaches fall into two basic categories. The first one is the "isolation."These approaches apply a clustering algorithm to obtain the theme clusters first, and then either rank the sentences within each cluster or explore the interaction between sentences and obtained clusters. In other words, clustering and ranking are regarded as two independent processes in this category although the cluster-level information has been incorporated into the cascaded approach. As a result, the ranking performance is inevitably influenced by the clustering result. The second one is the "mutuality," which uses clustering results to improve or refine the sentence ranking results.

A. Ranking Functions

A ranking is a relationship between a set of items such that, for any two items, the first is either 'ranked higher than', 'ranked lower than' or 'ranked equal to' the second

The ranking are classified into three types such as,

1. Global ranking
2. Local ranking
3. Conditional ranking

1. Global Ranking (without clustering)

Ranking Score of the sentences and terms in the whole document set. The whole document set is taken from DUC dataset.

2. Local Ranking (with in clustering)

Decompose the whole document set into sentences and obtain k sentence clusters (theme clusters) by certain clustering algorithm.

3. Conditional Ranking (across clustering)

To simplify the discovery of rank distribution of terms and sentences overall theme clusters.

Two conditional ranking one for sentence and another for term.

3.1 Ranking score

- In this module sentences then are reassigned to the nearest cluster under the new measure space to improve clustering. In this module values are summarized.
- The number of documents to be summarized can be very large.
- This makes information redundancy appears to be more serious in multi-document summarization than in single-document summarization.

Advantages

-This novel method is very efficient compared to existing method.

-Accurate summarization because initial ranking for sentence and terms.

B. Clustering

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.

A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute “best” criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.

Classification

Clustering algorithms may be classified as listed below:

1. Exclusive Clustering
2. Overlapping Clustering
3. Hierarchical Clustering
4. Probabilistic Clustering

Cluster Algorithm

The Cluster CMRW (Cluster-based Conditional Markov Random Walk) model incorporates the cluster-level information into the text graph and manipulates clusters and sentences equally, the Cluster-HITS model treats clusters and sentences as hubs and authorities in the HITS algorithm

Cluster Function

Traditional feature-based ranking approach employ quite different techniques to rank sentences, they have at least one point in common, i.e., all of them focus on sentences only, but ignoring the information beyond the sentence level.

Clustering is done with most related sentences in different themes of data.

These theme clusters are of different size and especially different importance to assist users in understanding the content in the whole document set.

Isolation approach is apply a clustering algorithm to obtain the theme clusters first.

Mutuality uses clustering results to improve or refine the sentence ranking results.

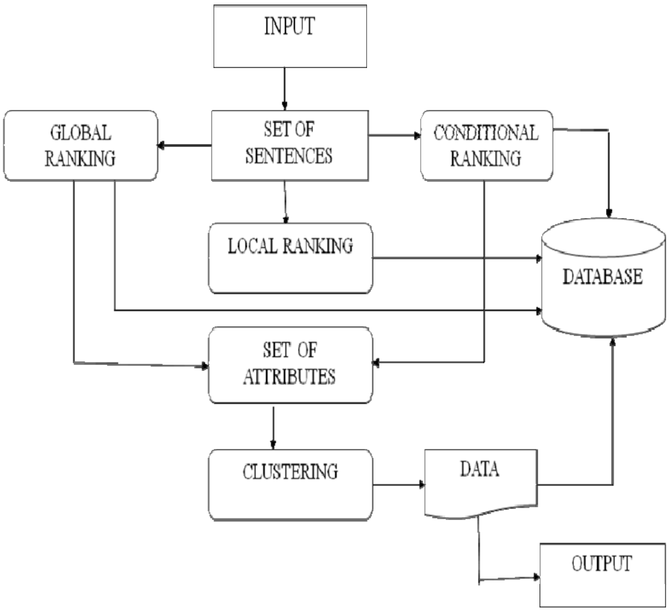

fig1Architecture Diagram

II. PROPOSED SYSTEM

In novel approach for sentence clustering. That directly generates clusters integrated with ranking

To produce ranking to the document (sentence and term) before clustering. .

The updates ranking and clustering sentences interactively and iteratively to multi-document summarization

A support approach is proposed to securely integrate ranking and clustering of sentences by exploring term rank distributions over the clusters.

Systematic experimental studies are conducted to verify the effectiveness and robustness of the proposed approach

A. Sentence Ranking Algorithm

Initially calculate cluster center and also conditional ranking for each cluster.

Create new attribute and center for each sentence, and then calculate similarity value for sentence and cluster.

The overall sentence ranking function is defined as the ensemble of all the sentence conditional ranking scores on the clusters.

*III.*MULTI DOCUMENT SUMMARIZATION:

It is an automatic procedure aimed at extraction of information from multiple texts written about the same topic. Resulting summary report allows individual users, such as professional information consumers, to quickly familiarize themselves with information contained in a large cluster of documents. In such a way, multi-document summarization systems are complementing the news aggregators performing the next step down the road of coping with information overload.

A good summarization technology aims to combine the main themes with completeness, readability, and conciseness. Document under Conferencing conducted annually by NIST, have developed sophisticated evaluation criteria for techniques accepting the multi-document summarization challenge.

An ideal multi-document summarization system does not simply shorten the source texts but presents information organized around the key aspects to represent a wider diversity of views on the topic. When such quality is achieved, an automatic multi-document summary is perceived more like an overview of a given topic. The latter implies that such text compilations should also meet other basic requirements for an overview text compiled by a human. The multi-document summary quality criteria are as follows:

1. clear structure, including an outline of the main content, from which it is easy to navigate to the full text sections
2. text within sections is divided into meaningful paragraphs
3. gradual transition from more general to more specific thematic aspects
4. good readability

IV.CONCLUSION AND FUTURE WORK

The three different ranking functions in a bi-type document graph constructed from the given document set. The k-Cluster ranking is applied and it gives good measure for each cluster. Conditional ranking function distributions are necessary for the parameter estimation during the reinforcement process. Finally, each sentence is re-assigned to a cluster that is the most similar to the sentence. Based on the updated clusters, within-cluster ranking is updated accordingly, which triggers the next round of clustering refinement.

REFERENCE

[1] L. Antiqueris, O. N. Oliveira, L. F. Costa, and M. G. Nunes, “A complex network approach to text summarization,” *Inf. Sci.*, vol. 175, Feb. 2009 , pp. 297–327

[2] R. Barzilay and K. R. Mckeown, “Sentence fusion for multi-document news summarization,” *Comput Linguist.*, vol. 31, pp. 297–327,2005, pp. 297–327

[3] R. Barzilay and L. Lee, “Catching the drift: Probabilistic content models, with applications to generation and summarization,” in *Proc.HLT-NAACL ’04*, 2004, pp. 113–120.

[4] J. Bilmes, “A Gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models,” Univ. of Berkeley, Berkeley, CA, USA, Tech. Rep.ICSI-TR-97-02, 1997.

[5] X. Y. Cai and W. J. Li, “A spectral analysis approach to document summarization: Clustering and ranking sentences simultaneously,” *Inf.Sci.*, vol. 181, no. 18, May 2011, pp. 3816–3827

[6] X. Y. CaiandW. J. Li, “Enhancing sentence level clustering with integrated and interactive frameworks for theme—Based summarization,”*J. Amer. Soc Inf. Sci. Tech.*, vol. 62, Oct. 2011 , pp. 2067–2082

[7] X. Y. Cai, W. J. Li, Y. Ouyang, and Y. Hong, “Simultaneous ranking and clustering of sentences: A reinforcement approach to multi-document summarization,” in *Proc. 23rd COLING Conf. ’10*, 2010, pp. 134–142.

[8] X. Y. Cai and W. J. Li, “Mutually reinforced manifold-ranking based relevance propagation model for query-focused multi-document summarization,”*IEEE Tran. Audio, Speech, Lang. Process.*, vol. 20, no. 5, Jul. 2012, pp. 1597–1607

[9] P. Fung and G. Ngai, “One story, one Folw: Hidden Markov story models for multilingual multi-document summarization,” *ACM Trans.Speech Lang. Process.*, vol. 3, no. 2, pp. 1–16, 2006.

[10]D. Gillick, B. Favre, and D. Hakkani-Tur, “The ICSI summarization system at TAC 2008,” in *Proc. Text Analysis Conf.*, 2008.