ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Ranking Through Clustering for Webdocument Using Semantic Similarity

[I]K.Vanisri, [II]P.Ponnila
[I]PG Student, Dept of CSE, Kalasalingam Institute of Technology, Krishnankovil, India
[II]Assistant Professor, Dept. of CSE,Kalasalingam Institute of Technology, Krishnankovil, India

Abstract
Multidocument summarization is a set of documents on the same topic, the output will be a paragraph length summary. So the documents often cover a number of topic themes with each theme represented by a clustering of highly related sentences, sentence cluster has been explored in order to provide more informative summaries. An existing cluster-based ranking approach that directly generates clusters integrated with ranking. We proposed an integrated approach that overcomes the drawback that we provide ranking for same meaning of different words. The basic idea of the approach is that ranking distribution of sentences in each cluster should be quite different from others, which may serve as features of clusters and new clustering measures of sentences can be calculated accordingly. As a clustering result to improve or refine the sentence ranking results. The proposed approach is demonstrated by both the cluster quality analysis and the summarization evaluation conducted on the DUC 2004-2007 datasets.

Keywords
Document summarization, sentence clustering, sentence ranking, Semantic similarity, Web Mining

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 [2], [3], 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. Sentence ranking is a technique of detecting importance of information in the sentence extraction framework. Though traditional feature-based ranking approaches [4], [5], [6], [7], [8] 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 (referring to Fig. 1(a)).In order to enhance the performance of summarization, recently cluster-based ranking approaches are proposed in the literature [9], [10], [11]. 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 (referring to Fig. 1(b)).The second one is the "mutuality," which uses clustering results to improve or refine the sentence ranking results (referring to Fig. 1(c)). The mutuality category can alleviate the problem occurring in the first category. Based on the latter one, we propose a reinforcement approach that updates ranking and clustering interactively and iteratively to multi-document summarization. The basic idea is follows, first collects some set of documents and then spilt in to some set of sentences then sentences can be spitted in to set of terms. Next ranking the documents then clustering should be performed to find the term ranking. As a result, the quality of sentence clustering is improved. In addition, sentence ranking results can thus be enhanced further by these high quality sentence clusters. Combining ranking and clustering in a two stage procedure like the first category, isolation, we propose an approach which can mutually enhance the quality of clustering and ranking. That is, sentence ranking can enhance the performance of sentence clustering and the obtained result of sentence clustering can further enhance the performance of sentence ranking. The motivation of the approach is that, for each sentence cluster, which forms a topic theme, the rank of terms conditional on this topic theme should be very distinct, quite different from the rank terms in other topic themes.

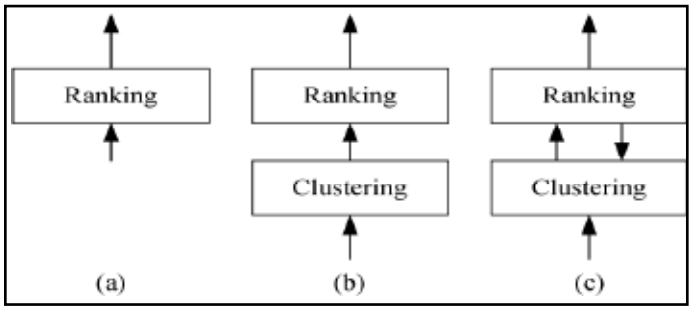

Fig. 1: Ranking vs. Clustering.

II. MOTIVATION

The motivation of the approach is that, for each sentence cluster, which forms a topic theme, the rank of terms conditional on this topic theme should be very distinct, and quite different from the rank of terms in other topic themes. Therefore, applying either clustering or ranking over the whole document set often leads to incomplete, or sometimes rather biased, analytical results.
The two main contributions of the paper are:

1. Three different ranking functions are defined in a bi-type document graph [12] constructed from the given document set, namely global, within-cluster and conditional rankings, respectively.
2. A reinforcement approach is proposed to tightly integrate ranking and clustering of sentences by exploring term rank distributions over the clusters.

III. Multi-Document Summarization

Multi-document summarization aims to produce a summary delivering the majority of information content from a set of documents about an explicit or implicit main topic. Multi-document summary can be used to concisely describe the information contained in a cluster of documents and facilitate the users to understand the document cluster. Finally to provide the ranking for web documents.

IV. Modules

A. Preprocessing

Given a collection of documents, we first decompose them into sentences. Then the stop-words are removed and words stemming is performed. After these steps, a sentence-term matrix is constructed and each element is the term frequency.

B. Document bi-type graph

In this section, we first present the sentence-term bi-type graph model for a set of given documents D, based on which the algorithm of reinforced ranking and clustering is developed. Let G = {V,E,W},where v is the set of vertices that consists of the sentences set S = $\{s_1,s_2,\ldots\ldots s_n\}$ and the term set T=$\{t_1,t_2,\ldots\ldots t_n\}$, i.e., V = S U T, "n" is the number of sentences and "m" is the number of terms. "E" is the set of edges that connect the vertices. The graph G is presented in fig. 2. "W" is the adjacency matrix in which the element w_{ij} represents the weight of the edge connecting v_i and v_j. Formally can be decomposed into four blocks, i.e. W_{SS}, W_{ST}, W_{TS}, W_{TT}.

$$W = \begin{pmatrix} W_{SS} & W_{ST} \\ W_{TS} & W_{TT} \end{pmatrix}$$

S = SENTENCES
T = TERMS

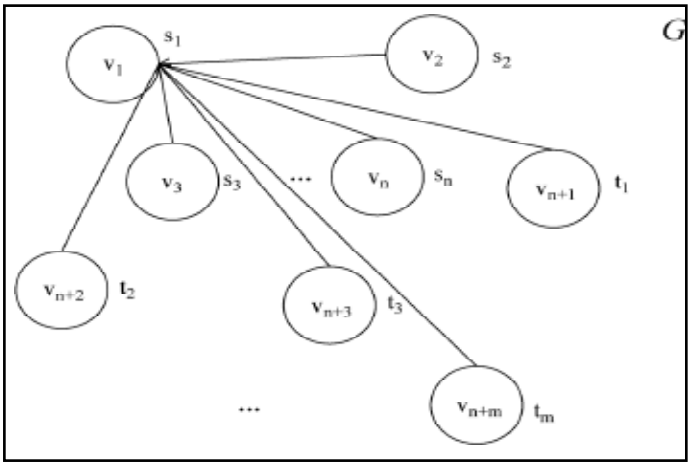

Fig. 2: Illustration of Graph

C. Ranking function

Recall that our ultimate goal is sentence ranking. More importantly, in this paper, conditional ranks of terms are served as features for each cluster. Each sentence is composed of terms, so the sentence can be considered as a mixture model over these rank distributions. The component coefficients can thus be used to improve clustering. In this section, we propose three ranking functions.

1. Global Ranking (Without Clustering):

A sentence should be ranked higher if it contains highly ranked terms and it is similar to the other highly ranked sentences, while a term should be ranked higher if it appears in highly ranked sentences and it is similar to the other highly ranked terms.

2. Local Ranking (Within Clusters):

We decompose the whole document set into sentences, and obtain K sentence clusters (also known as theme clusters) by certain clustering algorithm. The V theme clusters is denoted as C = $\{C_1$,C_2,…….. ,$C_K\}$ where CK (K = 1,2,3,…..,K) represents a cluster of highly related sentences S_{Ck}, which contains the terms T_{Ck}.

3 Conditional Ranking (Across Clusters):

To facilitate the discovery of rank distributions of terms and sentences over all the theme clusters, we further define two "conditional ranking functions" r(S|Ck) and r(T|Ck). sentence and term conditional ranks over all the theme clusters and are ready to introduce the reinforcement process. These two rank distributions are necessary for the parameter estimation during the reinforcement process.

Term Ranking

$$r(t_j|C_k) = \frac{r(t_j|C_k)}{\sum_{j=1}^{m} r(t_j|C_k)}$$

Sentence Ranking

$$r(s_i|C_k) = \frac{\sum_{j=1}^{m} W_{ST}(i,j) \cdot r(t_j|C_k)}{\sum_{i=1}^{n}\sum_{j=1}^{m} W_{ST}(i,j) \cdot r(t_j|C_k)}$$

D. Similarity measures

The similarity between a sentence and a cluster can be calculated as the cosine similarity between them. Where $W_{ST}(i,j)$ is the cosine similarity between the sentence Si and the term T_j. Thus the value of $W_{ST}(i,j)$ is between 0 and 1. If $W_{ST}(i,j)$ is near to 1, it means the sentence Si and the term T_j are semantically similar. If $W_{ST}(i,j)$ is near to 0, it means the sentence and the term are semantic different. $W_{SS}(i,j)$ is the cosine similarity between the sentences S_i and S_i. $W_{TT}(i,j)$ is the cosine similarity between the terms T_j and T_j. First we calculate the center of each cluster can thus be calculated accordingly, which is the mean of Si for all in the same cluster, i.e.,

$$\overrightarrow{\text{Center}}_{C_k} = \frac{\sum_{s_i \in C_k} \overrightarrow{s_i}}{|C_k|}$$

Where is the size C_k is cluster size .

Then the similarity between a sentence and a cluster can be calculated as the cosine similarity between them, i.e.,

$$sim(s_i, C_k) = \frac{\left\langle \overrightarrow{s_i}, \overrightarrow{\text{Center}}_{C_k} \right\rangle}{\sqrt{\|\overrightarrow{s_i}\|^2} \cdot \sqrt{\left\|\overrightarrow{\text{Center}}_{C_k}\right\|^2}}$$

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.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

The Overall Sentence Ranking Algorithm

Input: The bi-type document graph $G = < S \cup T, E, W >$, ranking functions, the cluster number K, $\delta = 1$, $Tre = 0.001$, $IterNum = 10$.
Output: sentence final ensemble ranking vector $f(S)$.

1. $t \leftarrow 0$;
2. Get the initial partition for S, i.e. C_k^t, $k = 1,2,\ldots K$, calculate cluster centers $\overrightarrow{Center_{C_k^t}}$ accordingly.
3. **For** (t=1; t<IterNum && $\varepsilon > Tre$; t++)
4. Calculate the within-cluster ranking $r_{C_k}(T_{C_k})$, $r_{C_k}(S_{C_k})$ and the conditional ranking $r(s_i \mid C_k)$;
5. Get new attribute $\vec{s}_i$ for each sentence s_i, and new attribute $\overrightarrow{Center_{C_k^t}}$ for each cluster C_k^t;
6. **For** each sentence s_i in S
7. **For** k=1 to K
8. Calculate similarity value $sim(s_i, C_k^t)$
9. **End For**
10. Assign s_i to $C_{k_0}^{t+1}$, $k_0 = \arg\max_k sim(s_i, C_k^t)$
11. **End For**
12. $\delta = \max_k | \overrightarrow{Center_{C_k^{t+1}}} - \overrightarrow{Center_{C_k^t}} |$
13. $t \leftarrow t + 1$
14. **End For**
15. For each sentence s_i in S
16. **For** k=1 to K
17. $f(s_i) = \sum_{k=1}^{K} \alpha_k \cdot r(s_i \mid C_k)$
18. **End For**
19. **End For**

V. Results

The following result show the preprocessing of documents

Fig. 3: Preprocessing the document sets.

The following fig shows the ranking with in cluster

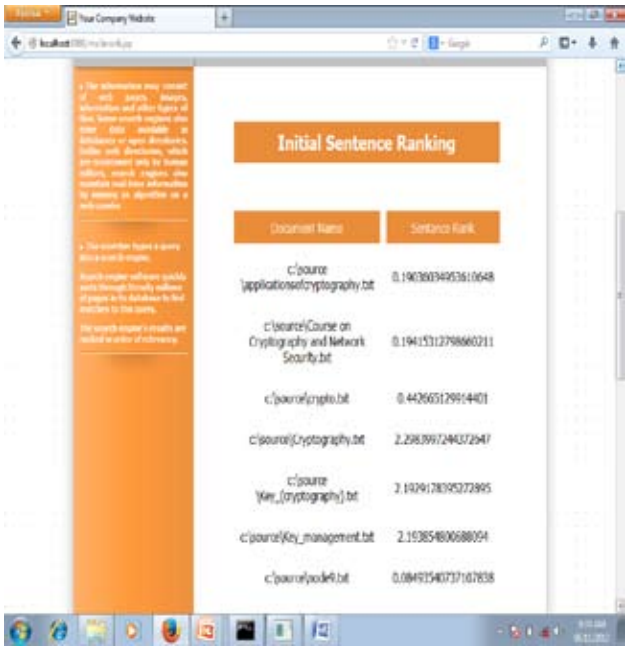

Fig. 4: Ranking the documents.

The following fig shows the ranking across the cluster

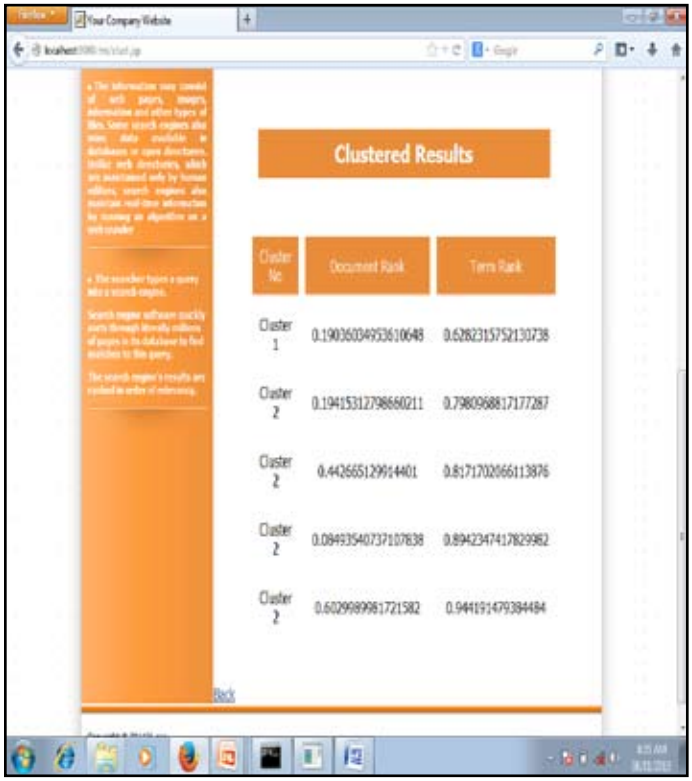

Fig. 5: Clustering the Documents

VI. Conclusion

In this paper, we first define three different ranking functions in a bi-type document graph constructed from the given document set. Based on initial K clusters, ranking is applied separately, which serves as a good measure for each cluster. Sentences then

are reassigned to the nearest cluster under the new measure space to improve clustering. As a result, quality of clustering and ranking are mutually enhanced. But Not consider the words that are different but in same meaning. Because cluster based summarization approach directly generates cluster first and with ranking next.

In the future, we plan to be applied to provide Integrating Clustering and ranking simultaneously terms and sentences and to improve the efficiency of document retrieval. In future studies, we will focus on the influence of document or other proper information, such as document cluster and topic query, to further improve the performance of summarization.

References

[1] *L. Antiqueris, O. N. Oliveira, L. F. Costa, and M. G. Nunes, "A complex network approach to text summarization," Inf. Sci., vol. 175, no.5, pp. 297–327, Feb. 2009.*

[2] *K. S. Jones, "Automatic summarising: The state of the art," Inf. Process Manag., vol. 43, no. 6, pp. 1449–1481, 2007.*

[3] *C. D. Manning and H. Schutze, Foundations of Statistical Natural Language Processing. Cambridge, MA, USA: MIT Press, 1999.*

[4] *S. Fisher and B. Roark, "Query-focused summarization by supervise sentence ranking and skewed word distributions," in Proc. DUC'06, 2006.*

[5] *W. J. Li,W. Li, Q. Chen, and M. L.Wu, "The Hong Kong Polytechnic University at DUC2005," in Proc. DUC'05, 2005.*

[6] *V. Qazvinian and D. R. Radev, "Scientific paper summarization using citation summary networks," in Proc. 17th COLING Conf., 2008, pp. 689–696.*

[7] *D. R. Radev, J. Otterbacher, H. Qi, and D. Tam, "MEAD ReDUCs: Michigan at DUC2003," in Proc. DUC'03, 2003.*

[8] *L. Zhao, X. J. Huang, and L. D.Wu, "Fudan University at DUC2005," in Proc. DUC2005.*

[9] *D. R. Radev, H. Y. Jing, M. Stys, and D. Tam, "Centroid-based summarization of multiple documents," Inform Process Manag, vol. 40, no. 6, pp. 919–938, 2004.*

[10] *Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T.Wu, "Rankclus: Integrating clustering with ranking for heterogenous information network analysis," in Proc. 12th EDBT Conf., 2009, pp. 79–85.*

[11] *X. J.Wan and J. W. Yang, "Improved affinity graph based multi-document summarization," in Proc.HLT-ANNCL Conf., 2006, pp. 362–370.*

[12] *A. Celikyilmaz and D. Hakkani-Tur, "Discovery of topically coherent sentences for extractive summarization," in Proc. 49th ACL Conf. '11, 2011, pp. 491–499.*

Mrs. K .Vanisri, the author is currently pursuing a Master of Engineering in Computer Science and Engineering at Kalasalingam Institute of Technology, affiliated to Anna University Chennai. She had complete B.E degree from Kamaraj College of Engineering and Technology, affiliated to Anna University Chennai.

Ms. P.Ponnila, the author is an Assistant Professor in Computer Science Engineering Department at Kalasalingam Institute of Technology. He received his B.TECH from SCAD College of Engineering and Technology, affiliated To Anna University, and M.E. Degree from Jaya Engineering College. Her Research interests are in the areas of Data Mining and cloud computing Security.