An Integrated Approach to Web Document Summarization Using Semantic Similarity

K .VANISRI [1], P .PONNILA [2] & J .JEEJOVETHARAJ [3]
[1]*PG Student, Dept of Computer Science and Engineering,*
[2]*Assistant Professor, Dept of Computer Science and Engineering,*
[3]*PG Student, Dept of Computer Science and Engineering,*
Kalasalingam Institute of Technology, Krishnankovil.

ABSTRACT

Multi document summarization is an automatic procedure aimed at extraction of information from multiple texts written about the same topic. The output will be a paragraph summary. Multi document summarization is very useful for presenting and organizing search results. An existing cluster-based summarization approaches utilize the clustering results to select the representative sentences in order to generate summaries. But it not considers providing the ranking for same meaning with different words or terms in the given set of documents. We proposed integrated approach that overcomes the drawback that we provide ranking for same meaning of different terms. The proposed approach aims to improve the clustering results compared to the existing methods.

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 spited 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, and quite different from the rank terms in other topic themes.

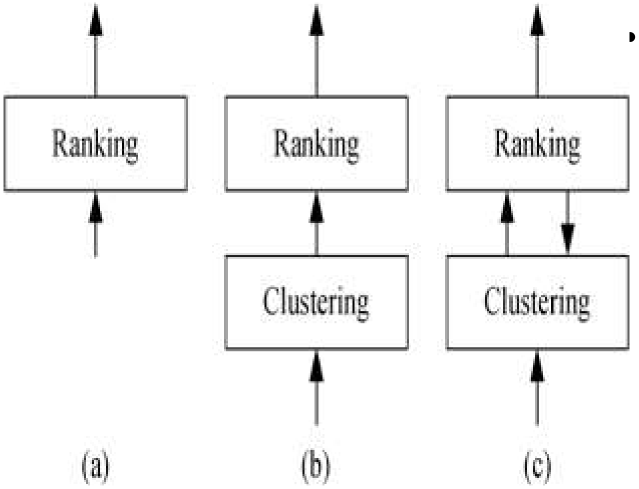


Fig. 1. Ranking vs. Clustering.

II. PROPOSED SYSTEM

The proposed system includes

- Integrating Clustering and ranking simultaneously terms and sentences
- A Bipartite Graph is used to show their relationships.
- A conditional ranking is integrated used to perform better results.

It also provides the ranking for same meaning with different words or terms by using Word Net tool in the given set of documents.

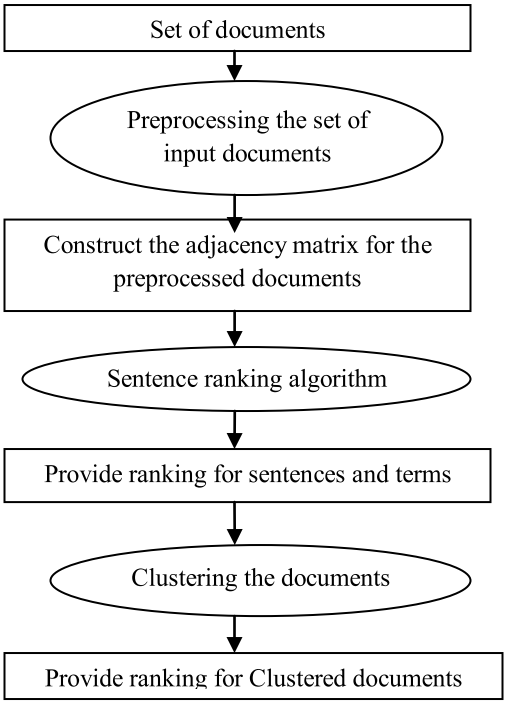
(i) Word Net tool

Word Net is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. Word Net superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, Word Net interlinks not just word forms strings of letters, but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, Word Net labels the semantic relations among words, whereas the grouping of words in a thesaurus does not follow any explicit pattern other than meaning similarity.

(ii) Advantages

- Ranking functions are defined as bi-type document graph.
- Tightly integrates clustering of sentences and term rank distributions.
- More effective and robust.

III. FLOW DIAGRAM


Set of documents
Preprocessing the set of input documents
Construct the adjacency matrix for the preprocessed documents
Sentence ranking algorithm
Provide ranking for sentences and terms
Clustering the documents
Provide ranking for Clustered documents

IV. MODULES

(i). 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.

(ii). 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 = $\{s_{1,}s_{2,........}s_{n\}}$ and the term set T=$\{t_{1,}t_{2,........}t_{n\}}$, 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

ISSN: 2278 – 7798

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. $\mathbf{W_{SS}}$, $\mathbf{W_{ST}}$, $\mathbf{W_{TS}}$, $\mathbf{W_{TT}}$.

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

S = SENTENCES
T = TERMS

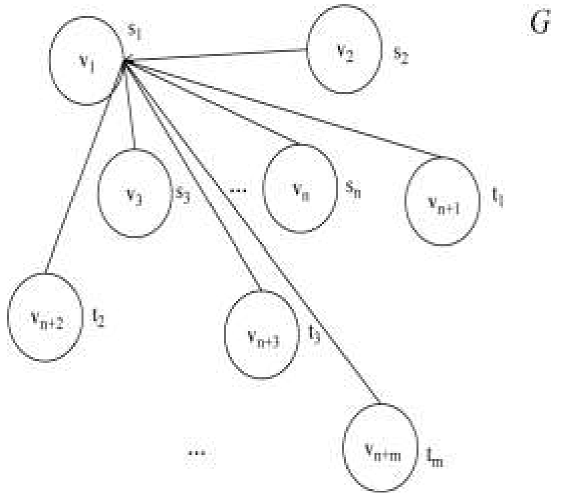


Fig. 2. Illustration of Graph .

(iii). 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 C_K (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" $r(S|C_k)$ and $r(T|C_k)$. 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.

1. **Term Ranking**

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

2. **Sentence Ranking**

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

(iv). 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 S_i 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 S_i 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 S_i for all in the same cluster, i.e.,

$$\overrightarrow{\mathbf{Center}}_{C_k} = \frac{\sum\limits_{s_i \in C_k} \overrightarrow{\mathbf{s_i}}}{|C_k|}$$

Where is the size $_{Ck\ is}$ cluster size .

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

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

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.

V. RESULTS

The following result show the preprocessing of documents

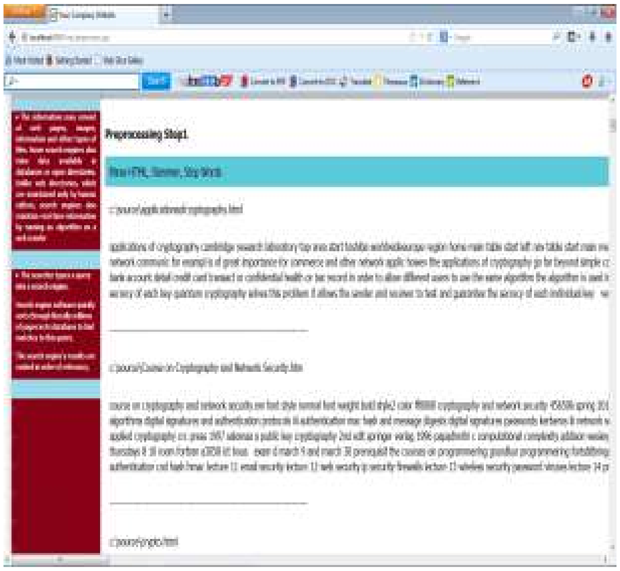


Fig. 3.Preprocessing the document sets.

The following fig shows the ranking with in cluster

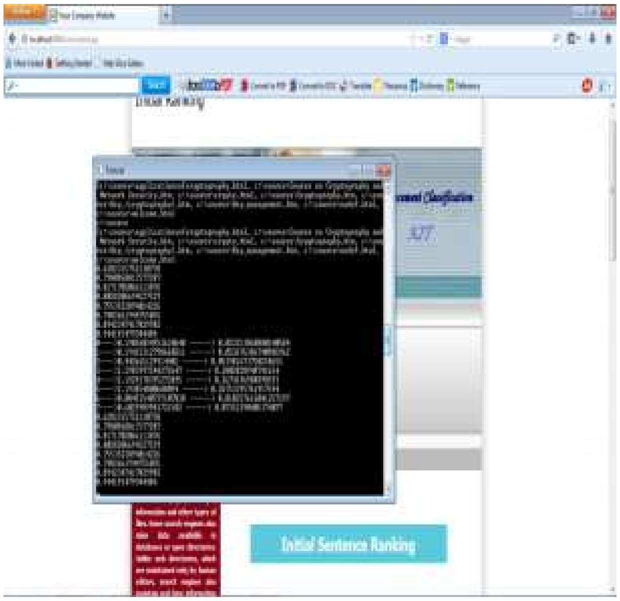


Fig. 4.Ranking the documents.

The following fig shows the ranking across the cluster

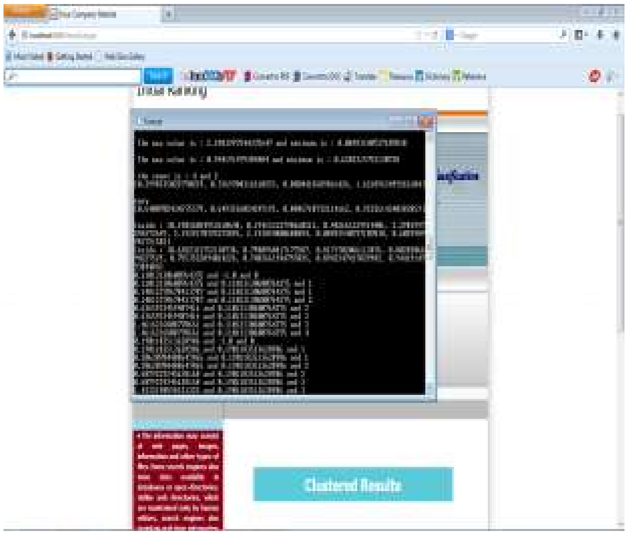


Fig 5.Clustering the documents.

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.

[13] G. Erkan and D. R. Radev, "LexRank: Graph-based centrality as salience in text summarization," *J Artif. Intell. Res*, vol. 22, no. 1, pp. 457–479, 2004.

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. She received her 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.