Ranking Through Clustering: An Integrated Approach to Multi-Document Summarization

Xiaoyan Cai and Wenjie Li

***Abstract*—Multi-document summarization aims to create a condensed summary while retaining the main characteristics of the original set of documents. Under such background, sentence ranking has hitherto been the issue of most concern. Since documents often cover a number of topic themes with each theme represented by a cluster of highly related sentences, sentence clustering has been explored in the literature in order to provide more informative summaries. For each 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. Existing cluster-based summarization approaches apply clustering and ranking in isolation, which leads to incomplete, or sometimes rather biased, analytical results. A newly emerged framework uses sentence clustering results to improve or refine the sentence ranking results. Under this framework, we propose a novel approach that directly generates clusters integrated with ranking in this paper. The basic idea of the approach is that ranking distribution of sentences in each cluster should be quite different from each other, which may serve as features of clusters and new clustering measures of sentences can be calculated accordingly. Meanwhile, better clustering results can achieve better ranking results. As a result, ranking and clustering by mutually and simultaneously updating each other so that the performance of both can be improved. The effectiveness of the proposed approach is demonstrated by both the cluster quality analysis and the summarization evaluation conducted on the DUC 2004–2007 datasets.**

***Index Terms*—Document summarization, sentence clustering, sentence ranking.**

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 [20], [25], 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.

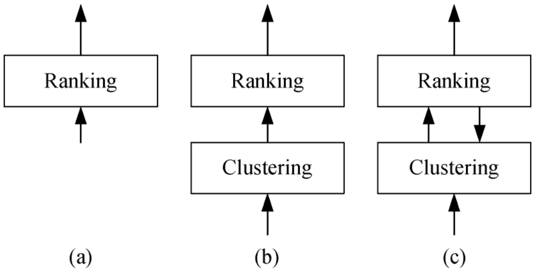


Fig. 1. Ranking vs. clustering.

Though traditional feature-based ranking approaches [15], [22], [29], [31], [38] 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)). 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 [13], [18], [19], [27], [33]. These theme clusters are of different size and especially different importance to assist users in understanding the content in the whole document set. The cluster level information is supposed to have foreseeable influence on sentence ranking [6].

In order to enhance the performance of summarization, recently cluster-based ranking approaches are proposed in the literature [5]–[7], [30], [33], [34]. 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)). 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 (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

Manuscript received May 31, 2012; revised September 04, 2012, December 11, 2012, and February 24, 2013; accepted February 25, 2013. Date of publication March 15, 2013; date of current version March 29, 2013. This work was supported in part by a Hong Kong RGC Project (Project Number PolyU 5230/08E) and in part by Central Universities under Grant (Project Number Z109021108). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yang Liu.

X. Cai is with the College of Information Engineering, Northwest A&F University, Xi'an 712100, China (e-mail: xiaoyanc@mail.nwpu.edu.cn).

W. Li is with the Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (e-mail: cswjli@comp.polyu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2013.2253098

ranking and clustering interactively and iteratively to multi-document summarization. The basic idea is as follows. First the documents are clustered into K clusters. Then the sentences are ranked within each cluster. After that, a mixture model is used to decompose each sentence into a K-dimensional vector, where each dimension is a component coefficient with respect to a cluster. Each dimension is measured by rank distribution. Sentences then are reassigned to the nearest cluster under the new measure space. 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. In all, instead of 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 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. For example, ranking sentences over the whole document set without considering which clusters they belong to often leads to insignificant results. Alternatively, clustering sentences in one cluster without distinction is meaningless as well. However, combining both functions together may lead to more comprehensible results.

The three 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.
3) Thorough experimental studies are conducted to verify the effectiveness and robustness of the proposed approach.

The rest of this paper is organized as follows. Section II reviews related work in cluster-based ranking. Section III defines three new ranking functions and explains the reinforced ranking and clustering processes and their application to multi-document summarization. Section IV presents experiments and evaluations. Conclusions are presented in Section V.

II. Related Work

A variety of summarization approaches have been proposed in the literature. These approaches are either extractive or abstractive. Extractive summarization assigns a significance score to each sentence and extracts the sentences with highest scores to form the summaries. Abstraction summarization, on the other hand, involves a certain degree of understanding of the content conveyed in the original documents and creates the summaries based on information fusion and/or language generation techniques [2]. Like most researchers in this field, we follow the extractive summarization framework [1], [28] in this work.

Under the framework of extractive summarization, sentence ranking [9], [17], [24] is the issue of most concern. Beyond that, when the given documents are all supposed to be about the same topic, they are very likely to repeat some important information in different documents or in different places in the same document. Therefore, effectively clustering the sentences with the same or very similar content is necessary. Recently it has been successfully applied in cluster-based summarization [6], [8], [13], [29]. These cluster-based summarization approaches utilize the clustering results to select the representative sentences in order to generate summaries. Alternatively, the clustering results could be used to improve or refine the sentence ranking results. Newly emerged cluster-based summarization approaches are of this nature. For example, Wan and Yang [34] propose two models to incorporate the cluster-level information into the process of sentence ranking for generic multi-document summarization, namely ClusterCMRW and Cluster-HITS. The ClusterCMRW (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.

Meanwhile, Wang *et al.* [35] propose a language model to simultaneously cluster and summarize multi-documents. Non-negative factorization is performed on the term-document matrix using the term-sentence matrix as the base so that the document-topic and sentence-topic matrices could be constructed, from which the document clusters and the corresponding summary sentences are generated simultaneously. Similarly, Huang *et al.* [37] propose to use ensemble non-negative matrix factorization to enhance clustering of multiple biomedical documents. Joris [39] develops the pairwise-adaptive dissimilarity measure to cluster documents. In addition, Wang *et al.* [36] apply symmetric matrix factorization to generate sentence clusters. Each sentence's score is based on the linear combination of two elements. One is the average similarity score between a sentence and all the other sentences in the same cluster. The other is the similarity between the sentence and the given query. Celikyilmaz and Hakkani-Tur [10] formulate extractive summarization as a two step learning problem building a generative model for pattern discovery and a regression model for inference. The above approaches aim to obtain more accurate sentence clusters in various ways. Besides, Cai and Li [5] firstly explore the "clustering structure" of sentences before the actual clustering algorithm is performed. They call the spectral clustering structure "beam," which is discovered by analyzing the spectral characteristics of the sentence similarity network. The structure of beams illustrates the distribution of sentences, where each beam represents a cluster. Based on it, they deem the sentences projected on the same beam share similar content, while the sentences projected on the different beams have less content overlap with each other. The generated multi-document summary can be obtained using the above principle. The spectral analysis is also applied to identify topics in document sets.

Rather than clustering and ranking sentences only once, we propose a novel approach in this paper that updates ranking and clustering sentences interactively and iteratively to multi-document summarization. This is similar to the RankClus framework

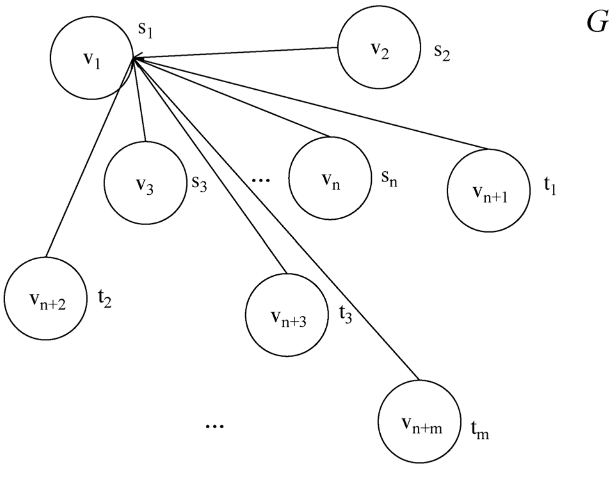


Fig. 2. Illustration of graph G.

[7], which uses frequency relationships between sentences, sentences and terms, ignoring the relationship between terms. In our approach, semantic relationships between sentences, sentences and terms are used. In addition, semantic relationship between terms is also captured. In this way we can obtain much more meaningful information, helping to generate summaries which are much more similar to human-generated summaries.

Hidden Markov Model is also applied in multi-document summarization [3], [16]. But the approaches iteratively learn the parameters of Hidden Markov Model to obtain final sentence clusters. They do not use any ranking information. Our proposed approach uses EM model to estimate the component coefficients, and represents the sentences in a new measure space, so the quality of clustering and ranking are mutually enhanced.

III. An Integrated Approach to Multi-Document Summarization

A. 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 $G = \langle V, E, W \rangle$, where V is the set of vertices that consists of the sentence set $S = \{s_1, s_2, \ldots, s_n\}$ and the term set $T = \{t_1, t_2, \ldots, t_m\}$, i.e., $V = S \cup T$, n is the number of sentences and m is the number of terms. Each term vertex is the sentence that is given in the WordNet as the description of the term. It extracts the first sense used from WordNet instead of the word itself.[1] E is the set of edges that connect the vertices. An edge can connect a sentence to a word, a sentence to a sentence, or a word to a word, i.e. $E = \{\langle v_i, v_j \rangle | v_i, v_j \in V\}$. The graph G is presented in Fig. 2. For ease of illustration, we only demonstrate the edges between v1 and other vertices.

$\boldsymbol{W}$ is the adjacency matrix in which the element w_{ij} represents the weight of the edge connecting v_i and v_j. Formally, $\boldsymbol{W}$ can be decomposed into four blocks, i.e. $\mathbf{W_{SS}}$, $\mathbf{W_{ST}}$, $\mathbf{W_{TS}}$ and $\mathbf{W_{TT}}$, each representing a sub-graph of the textual objects indicated by the subscripts. $\mathbf{W}$ can be written as

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

where $\mathbf{W_{ST}}(i,j)$ is the cosine similarity between the sentence s_i and the term t_j. Thus the value of $\mathbf{W_{ST}}(i,j)$ is between 0 and 1. If $\mathbf{W_{ST}}(i,j)$ is near to 1, it means the sentence s_i and the term t_j are semantically similar. If $\mathbf{W_{ST}}(i,j)$ is near to 0, it means the sentence s_i and the term t_j are semantic different. $\mathbf{W_{SS}}(i,j)$ is the cosine similarity between the sentences s_i and s_j. $\mathbf{W_{TS}}$ is equal to $\mathbf{W_{ST}^{T}}$ as the relationships between terms and sentences are symmetric. $\mathbf{W_{TT}}(i,j)$ is the cosine similarity between the terms t_i and t_j.

B. Basic Ranking Functions

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): Let $r(s_i)$ $(i = 1, 2, \ldots, n)$ and $r(t_j)$ $(j = 1, 2, \ldots, m)$ denote the ranking scores of the sentence s_i and the term t_j in the whole document set, respectively. Based on the assumptions that

> *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.*

We define

$$r(s_i) = \alpha \cdot \sum_{i=1}^{m} \mathbf{W_{ST}}(i,j) \cdot r(t_j) + (1-\alpha) \sum_{i=1}^{n} \mathbf{W_{SS}}(i,j) \cdot r(s_j), \tag{1}$$

and

$$r(t_j) = \beta \cdot \sum_{i=1}^{n} \mathbf{W_{TS}}(j,i) \cdot r(s_i) + (1-\beta) \sum_{i=1}^{m} \mathbf{W_{TT}}(i,j) \cdot r(t_i). \tag{2}$$

where α and β are weighting parameters, ranging from 0 to 1.

For calculation purpose, $r(s_i)$ and $r(t_j)$ are normalized by

$$r(s_i) \leftarrow \frac{r(s_i)}{\sum_{i'=1}^{n} r(s_{i'})} \quad \text{and} \quad r(t_j) \leftarrow \frac{r(t_j)}{\sum_{j'=1}^{m} r(t_{j'})}$$

Equations (1) and (2) can be rewritten using the matrix form, i.e.

$$\begin{cases} \boldsymbol{r}(\boldsymbol{S}) = \alpha \cdot \frac{\mathbf{W_{ST}} \cdot \boldsymbol{r}(\boldsymbol{T})}{\|\mathbf{W_{ST}} \cdot \boldsymbol{r}(\boldsymbol{T})\|} + (1-\alpha) \cdot \frac{\mathbf{W_{SS}} \cdot \boldsymbol{r}(\boldsymbol{S})}{\|\mathbf{W_{SS}} \cdot \boldsymbol{r}(\boldsymbol{S})\|} \\ \boldsymbol{r}(\boldsymbol{T}) = \beta \cdot \frac{\mathbf{W_{TS}} \cdot \boldsymbol{r}(\boldsymbol{S})}{\|\mathbf{W_{TS}} \cdot \boldsymbol{r}(\boldsymbol{S})\|} + (1-\beta) \cdot \frac{\mathbf{W_{TT}} \cdot \boldsymbol{r}(\boldsymbol{T})}{\|\mathbf{W_{TT}} \cdot r(T)\|}. \end{cases} \tag{3}$$

We call $\boldsymbol{r}(\boldsymbol{S})$ and $\boldsymbol{r}(\boldsymbol{T})$ the "**global ranking functions**," because at this moment sentence clustering is not yet involved and

[1] For example, if the term is 'cat', then the corresponding term vertex in the graph should be 'feline mammal usually having thick soft fur and being unable to roar; domestic cats; wildcats' extracted from the WordNet.

all the sentences/terms in the whole document set are ranked together.

Theorem 1: The solutions to $\boldsymbol{r}(\boldsymbol{S})$ and $\boldsymbol{r}(\boldsymbol{T})$ given by (3) are the primary eigenvectors of $\alpha\beta \cdot \mathbf{W_{ST}}(I-(1-\beta)\cdot \mathbf{W_{TT}})^{-1} \cdot \mathbf{W_{TS}} + (1-\alpha)\cdot \mathbf{W_{SS}}$ and $\alpha\beta \cdot \mathbf{W_{TS}}(I-(1-\alpha)\cdot \mathbf{W_{SS}})^{-1} \cdot \mathbf{W_{ST}} + (1-\beta)\cdot \mathbf{W_{TT}}$, respectively.

Proof of the THEOREM 1 is provided in Appendix.

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 $C = \{C_1, C_2, \ldots, C_K\}$ where C_k $(k = 1, 2, \ldots, K)$ represents a cluster of highly related sentences $S_{C_k}(\in C_k)$ which contains the terms $T_{C_k}(\in C_k)$. The sentences and terms within the cluster C_k form a cluster bi-type graph with the adjacency matrix W_{C_k}. Let $\boldsymbol{r_{C_k}}(\boldsymbol{S_{C_k}})$ and $\boldsymbol{r_{C_k}}(\boldsymbol{T_{C_k}})$ denote the ranking scores of S_{C_k} and T_{C_k} within C_k. They are calculated by an equation similar to (3) by replacing the document level adjacency matrix W with the cluster level adjacency matrix W_{C_k}. We call $\boldsymbol{r_{C_k}}(\boldsymbol{S_{C_k}})$ and $\boldsymbol{r_{C_k}}(\boldsymbol{T_{C_k}})$ the "**within-cluster ranking functions**" with respect to the cluster C_k. They are the local ranking functions, in contrast to $\boldsymbol{r}(\boldsymbol{S})$ and $\boldsymbol{r}(\boldsymbol{T})$ that rank all the sentences and terms in the whole document set D. For each sentence cluster which could form 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. Similarly, ranking sentences over the whole document set without considering which clusters they belong to often leads to meaningless results. Thus, it will benefit sentence overall ranking when knowing more details about the ranking results at the finer granularity of theme clusters, instead of at the coarse granularity of the whole document set.

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" $\boldsymbol{r}(\boldsymbol{S}|\boldsymbol{C_k})$ and $\boldsymbol{r}(\boldsymbol{T}|\boldsymbol{C_k})$. These two rank distributions are necessary for the parameter estimation during the reinforcement process introduced later. The conditional ranking score of the term t_j for the cluster C_k, i.e., $\boldsymbol{r}(\boldsymbol{T}|\boldsymbol{C_k})$ is directly derived from T_{C_k},

$$\boldsymbol{r}(\boldsymbol{t_j}|\boldsymbol{C_k}) = \begin{cases} \boldsymbol{r_{C_k}}(\boldsymbol{t_j}), & t_j \in C_k \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

It is further normalized as

$$\boldsymbol{r}(\boldsymbol{t_j}|\boldsymbol{C_k}) = \frac{\boldsymbol{r}(\boldsymbol{t_j}|\boldsymbol{C_k})}{\sum_{j=1}^{m} \boldsymbol{r}(\boldsymbol{t_j}|\boldsymbol{C_k})}. \tag{5}$$

Then the conditional ranking score of the sentence s_i on the cluster C_k is deduced from the terms that are included in s_i, i.e.,

$$\boldsymbol{r}(\boldsymbol{s_i}|\boldsymbol{C_k}) = \frac{\sum_{j=1}^{m} \mathbf{W_{ST}}(i,j) \cdot \boldsymbol{r}(\boldsymbol{t_j}|\boldsymbol{C_k})}{\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{W_{ST}}(i,j) \cdot \boldsymbol{r}(\boldsymbol{t_j}|\boldsymbol{C_k})}. \tag{6}$$

Equation (6) can be interpreted as that the conditional rank of s_i on C_k is higher if many terms in s_i are ranked higher in C_k. Now we have sentence and term conditional ranks over all the theme clusters and are ready to introduce the reinforcement process.

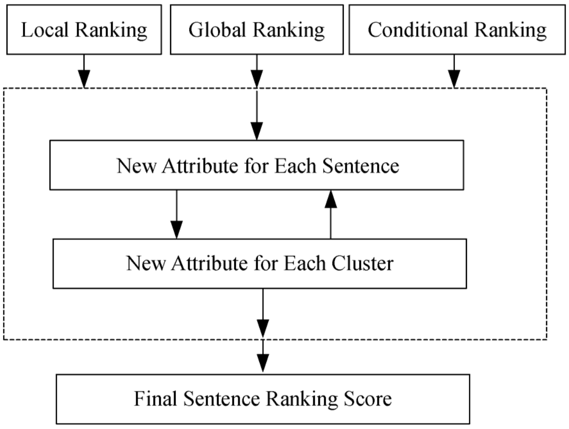

Local Ranking
Global Ranking
Conditional Ranking
New Attribute for Each Sentence
New Attribute for Each Cluster
Final Sentence Ranking Score

Fig. 3. The sentence ranking process.

C. Reinforcement Between Within-Cluster Ranking and Clustering

The conditional ranks of the term t_j across the K theme clusters can be viewed as a rank distribution. Then the rank distribution of the sentence s_i can be considered as a mixture model over K conditional rank distributions of the terms contained in the sentence s_i. And the sentence s_i can be represented as a K-dimensional vector in the new measure space, in which the vectors can be used to guide the sentence clustering update. Next, we will explain the mixture model of sentence and use EM algorithm [4] to get the component coefficients of the model. Then, we will present the similarity measure between sentence and cluster, which is used to adjust the clusters that the sentences belong to and in turn modify within-cluster ranking for the sentences in the updated clusters. The sentence ranking process can be illustrated in Fig. 3 as follow:

Based on the Fig. 3, we develop the new attribute for each sentence in Section III-C-1) and Section III-C-2). The new attribute for each cluster is illustrate in Section III-C-3). The above two processes are repeated until both of them are stable. Then we can get the final sentence ranking score which is presented in Section III-D.

1) Sentence Mixture Model: For each sentence s_i, we assume that it follows the distribution $\boldsymbol{r}(\boldsymbol{T}|\boldsymbol{s_i})$ to generate the relationship between the sentence s_i and the term set T. This distribution can be considered as a mixture model over K component distributions, i.e. the term conditional rank distributions across K theme clusters. We use $\gamma_{\mathrm{i},k}$ to denote the probability that s_i belongs to C_k, then $\boldsymbol{r}(\boldsymbol{T}|\boldsymbol{s_i})$ can be modeled as:

$$\boldsymbol{r}(\boldsymbol{T}|\boldsymbol{s_i}) = \sum_{k=1}^{K} \gamma_{\mathrm{i},k} \cdot \boldsymbol{r}(\boldsymbol{T}|\boldsymbol{C_k}) \quad \text{and} \quad \sum_{k=1}^{K} \gamma_{\mathrm{i},k} = 1. \tag{7}$$

$\gamma_{\mathrm{i},k}$ can be explained as $\boldsymbol{p}(\boldsymbol{C_k}|\boldsymbol{s_i})$ and calculated by the Bayesian equation $\boldsymbol{p}(\boldsymbol{C_k}|\boldsymbol{s_i}) \propto \boldsymbol{p}(\boldsymbol{s_i}|\boldsymbol{C_k}) \cdot \boldsymbol{p}(\boldsymbol{C_k})$, where $\boldsymbol{p}(\boldsymbol{s_i}|\boldsymbol{C_k})$ is assumed to be $\boldsymbol{r}(\boldsymbol{s_i}|\boldsymbol{C_k})$ obtained from the conditional rank of s_i on C_k as introduced before and $p(C_k)$ is the prior probability.

2) Parameter Estimation: We use the EM algorithm to estimate the component coefficients $\gamma_{\mathrm{i},k}$ along with $\{p(C_k)\}$. A hidden variable C_z, $z \in \{1, 2, \ldots, K\}$ is used to denote the cluster label that a sentence term pair (s_i, t_j) are from. In addition, we make the independent assumption that the probability of s_i belonging to C_k and the probability of t_j belonging to C_k are independent, i.e. $p(s_i, t_j|C_k) = p(s_i|C_k) \cdot p(t_j|C_k)$, where $p(s_i, t_j|C_k)$ is the probability of s_i and t_j both belonging to C_k. Similarly, $p(t_j|C_k)$ is assumed to be $r(t_j|C_k)$.

Let Θ be the parameter matrix, which is a $n \times K$ matrix $\boldsymbol{\Theta}_{n\times K} = \{\gamma_{i,k}\}$ $(i = 1, \ldots, n;\ k = 1, \ldots, K)$. The best $\boldsymbol{\Theta}$ is estimated from the relationships observed in the document bi-type graph. The likelihood of generating all the relationships under the parameter $\boldsymbol{\Theta}$ can be calculated as:

$$\begin{aligned} L'(\boldsymbol{\Theta}|\mathbf{W_{ST}}, \mathbf{W_{SS}}) &= p(\mathbf{W_{ST}}|\boldsymbol{\Theta}) \cdot p(\mathbf{W_{SS}}|\boldsymbol{\Theta}) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{m} p(s_i, t_j|\boldsymbol{\Theta})^{\mathbf{W_{ST}}(i,j)} \\ &\quad \cdot \prod_{i=1}^{n}\prod_{j=1}^{n} p(s_i, s_j|\boldsymbol{\Theta})^{\mathbf{W_{SS}}(i,j)}, \end{aligned} \tag{8}$$

where $p(s_i, t_j|\boldsymbol{\Theta})$ is the probability that s_i and t_j both belong to the same cluster, given the current parameter. As $p(s_i, s_j|\boldsymbol{\Theta})$ does not contain variables from $\boldsymbol{\Theta}$, we only need to consider maximizing the first part of the likelihood in order to get the best estimation of $\boldsymbol{\Theta}$. Let $L(\boldsymbol{\Theta}|\mathbf{W_{ST}})$ be the first part of likelihood.

Taking into account the hidden variable C_z, the complete log-likelihood can be written as

$$\begin{aligned} &\log L(\boldsymbol{\Theta}|\mathbf{W_{ST}}, C_Z) \\ &\quad = \log \prod_{i=1}^{n}\prod_{j=1}^{m} \left(p(s_i, t_j, C_z|\boldsymbol{\Theta})\right)^{\mathbf{W_{ST}}(i,j)} \\ &\quad = \log \prod_{i=1}^{n}\prod_{j=1}^{m} \left(p(s_i, t_j|C_z, \boldsymbol{\Theta}) \cdot p(C_z|\boldsymbol{\Theta})\right)^{W_{ST}(i,j)} \\ &\quad = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{W_{ST}}(i,j) \cdot \log\left(p_Z(s_i, t_j) \cdot p(C_z|\boldsymbol{\Theta})\right). \end{aligned} \tag{9}$$

In the E-step, given the initial parameter $\boldsymbol{\Theta}^0$, which is set to $\gamma_{i,k}^0 = 1/K$ for all i and k, the expectation of log-likelihood under the current distribution of C_Z is:

$$\begin{aligned} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) &= E_{f(C_Z|W_{ST}, \boldsymbol{\Theta}^0)}(\log L(\boldsymbol{\Theta}|\mathbf{W_{ST}}, C_Z) \\ &= \sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{W_{ST}}(i,j) \\ &\quad \cdot \log\left(p_k(s_i, t_j)\right) \cdot p(C_z = C_k|s_i, t_j, \boldsymbol{\Theta}^0) \\ &\quad + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{m} \mathbf{W_{ST}}(i,j) \cdot \log\left(p(C_z = C_k|\boldsymbol{\Theta})\right) \\ &\quad \cdot p(C_z = C_k|s_i, t_j, \boldsymbol{\Theta}^0). \end{aligned} \tag{10}$$

The conditional distribution in the above equation, i.e.,$p(C_z = C_k|s_i, t_j, \boldsymbol{\Theta}^0)$, can be calculated using the Bayesian rule as follows:

$$\begin{aligned} &p(C_z = C_k|s_i, t_j, \boldsymbol{\Theta}^0) \\ &\quad \propto p(s_i, t_j|C_z = C_k, \boldsymbol{\Theta}^0) p(C_z = C_k|\boldsymbol{\Theta}^0) \\ &\quad \propto p^0(s_i|C_k) p^0(t_j|C_k) p^0(C_z = C_k). \end{aligned} \tag{11}$$

In the M-Step, we first get the estimation of $p(C_z = C_k)$ by maximizing the expectation $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0)$. By introducing a Lagrange multiplier λ, we get the equation below.

$$\begin{aligned} &\frac{\partial}{\partial p(C_z = C_k)}\left[Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) + \lambda\left(\sum_{k=1}^{K} p(C_z = C_k) - 1\right)\right] = 0 \Rightarrow \\ &\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{W_{ST}}(i,j) \frac{1}{p(C_z = C_k)} p(C_z = C_k|s_i, t_j, \boldsymbol{\Theta}^0) + \lambda = 0. \end{aligned} \tag{12}$$

Thus, the estimation of $p(C_z = C_k)$ given previous $\boldsymbol{\Theta}^0$ is

$$p(C_z = C_k) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{W_{ST}}(i,j) p(C_z = C_k|s_i, t_j, \boldsymbol{\Theta}^0)}{\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{W_{ST}}(i,j)}. \tag{13}$$

Then, the parameters $\gamma_{i,k}$ can be calculated with the Bayesian rule as

$$\gamma_{i,k} = \frac{p(s_i|C_k) p(C_z = C_k)}{\sum_{l=1}^{K} p(s_i|C_l) p(C_z = C_l)}. \tag{14}$$

By setting $\boldsymbol{\Theta}^0 = \boldsymbol{\Theta}$, the whole process can be repeated. The updating rules provided in (11)–(14) are applied at each iteration. Finally, $\boldsymbol{\Theta}$ will converge to a local maximum. A similar estimation process has been adopted in [32], which is used to estimate the component coefficients for author-conference networks.

3) Similarity Measure: After we get the estimations of the component coefficients $\gamma_{\mathrm{i},k}$ for s_i, s_i will be represented as a K dimensional vector $\vec{\mathbf{s_i}} = \{\gamma_{i,1}, \gamma_{i,2}, \cdots, \gamma_{i,K}\}$. The center of each cluster can thus be calculated accordingly, which is the mean of $\vec{\mathbf{s_i}}$ for all s_i in the same cluster, i.e.,

$$\overrightarrow{\mathbf{Center}}_{C_k} = \frac{\sum_{s_i \in C_k} \overrightarrow{\mathbf{s_i}}}{|C_k|}, \tag{15}$$

where $|C_k|$ is the size of C_k.

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

$$sim(s_i, C_k) = \frac{\left\langle \overrightarrow{\mathbf{s_i}}, \overrightarrow{\mathbf{Center}}_{\mathbf{C_k}} \right\rangle}{\sqrt{\|\overrightarrow{\mathbf{s_i}}\|^2} \cdot \sqrt{\left\|\overrightarrow{\mathbf{Center}}_{\mathbf{C_k}}\right\|^2}}. \tag{16}$$

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. It is expected that the quality of clusters can be improved during this iterative update process since the similar sentences under new attributes will be grouped together, and meanwhile the quality of ranking will be improved along with the better clusters and thus offers better attributes for further clustering.

TABLE I
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 $f(S)$.
t ← 0;
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.
For ($t$=1; $t$<IterNum && $\varepsilon > Tre$; $t$++)
    Calculate the within-cluster ranking $r_{C_k}(T_{C_k})$, $r_{C_k}(S_{C_k})$ and
    the conditional ranking $r(s_i|C_k)$;
    Get new attribute $\vec{s_i}$ for each sentence $s_i$, and new
    attribute $\overrightarrow{Center_{C_k^t}}$ for each cluster $C_k^t$;
    For each sentence $s_i$ in $S$
        For $k$=1 to $K$
            Calculate similarity value $sim(s_i, C_k^t)$
        End For
        Assign $s_i$ to $C_{k_0}^{t+1}$, $k_0=\arg\max_k sim(s_i, C_k^t)$
    End For
    $\delta=\max_k|\overrightarrow{Center_{C_k^{t+1}}}-\overrightarrow{Center_{C_k^t}}|$
    $t \leftarrow t+1$
End For
For each sentence $s_i$ in $S$
    For $k$=1 to $K$
        $f(s_i)=\sum_{k=1}^{K}\alpha_k \cdot r(s_i|C_k)$
    End For
End For
```

D. Ensemble Ranking

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

$$f(s_i)=\sum_{k=1}^{K}\alpha_k \cdot r(s_i|C_k), \tag{17}$$

where α_k is a coefficient evaluating the importance of C_k. It can be formulated as the normalized cosine similarity between a theme cluster and the whole document set for generic summarization, or between a theme cluster and a given query for query-based summarization. $\alpha_k \in [0,1]$ and $\sum_{k=1}^{K}\alpha_k=1$.

Table I summarizes the whole process that determines the overall sentence ensemble ranking scores.

E. Summary Generation

In multi-document summarization, 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. Redundancy control is necessary. Two popular techniques for avoiding redundancy in summarization are Maximal Marginal Relevance (MMR) [11] and clustering [26], [30]. In MMR, the determination of redundancy is based mainly on the textual overlap between the sentence that is about to be added to the output and the sentences that are already in the generated summary text. MMR has been modified by many researchers [1], [14]. On the other hand, clustering offers an alternative that the summarization system clusters the input textual units before starting the selection process. This step allows analyzing one or a few number of representative units from each cluster instead of all textual units.

We apply a simple yet effective way to choose summary sentences, which is a modified MMR-like approach. 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 sentences in the summary reaches the length limitation. In our experiment, the threshold is set to 0.9.

IV. EXPERIMENTS AND EVALUATION

We conduct the experiments on the DUC 2004 generic multi-document summarization dataset and the DUC 2005, DUC 2006 and DUC 2007 query-based multi-document summarization datasets. According to task definitions, systems are required to produce a concise summary for each document set (without or with a given query description) and the length of summaries is limited to 665 bytes in DUC 2004 and 250 words in DUC 2005–2007.

A well-recognized automatic evaluation toolkit ROUGE [23] is used in evaluation. It measures summary quality by counting overlapping units between system-generated summaries and human-written reference summaries. We report two common ROUGE scores in this paper, namely ROUGE-1 and ROUGE-2, which base on Uni-gram match and Bi-gram match, respectively. Documents and queries are pre-processed by segmenting sentences and splitting words. Stop words[2] are removed and the remaining words are stemmed using Porter stemmer.[3]

A. Comparison With Other Summarization Approaches

In order to evaluate the performance of reinforced clustering and ranking approach, we compare it with the other five ranking approaches: (1) Global-Rank, which does not apply clustering and simply relies on the sentence global ranking scores to select summary sentences; (2) Local-Rank, which clusters sentences first and then ranks sentences within each cluster. A summary is generated in the same way as presented in [30]. The clusters are ordered by decreasing size; (3) Cluster-HITS [36], which also clusters sentences first, but then regards clusters as hubs and sentences as authorities in the HITS algorithm and uses the obtained authority scores to rank and select sentences. (4) Spectral-Analysis [5], which discovers special clustering structure called beams by analyzing the spectral characteristics of the sentence similarity network first, and then deems the sentences projected on the same beam share similar content while the sentences projected on the different beams have less content overlap with each other. Meanwhile, the sentences with the

[2] Words which are filtered out prior to, or after, processing of natural language data (text).

[3] http://tartarus.org/~martin/PorterStemmer/index.html

TABLE II
RESULTS ON THE DUC 2004 DATASET

DUC 2004	ROUGE-1	ROUGE-2
Ours	**0.37475**	**0.08973**
RankClus	0.37082	0.08531
Spectral-Analysis	0.36793	0.08034
Cluster-HITS	0.36463	0.07632
Local-Rank	0.36294	0.07351
Global-Rank	0.35729	0.06893

TABLE III
RESULTS ON THE DUC 2005 DATASET

DUC 2005	ROUGE-1	ROUGE-2
Ours	**0.36451**	**0.07375**
RankClus	0.36117	0.06968
Spectral-Analysis	0.36040	0.06809
Cluster-HITS	0.35822	0.06927
Local-Rank	0.35233	0.06407
Global-Rank	0.34488	0.06009

TABLE IV
RESULTS ON THE DUC 2006 DATASET

DUC 2006	ROUGE-1	ROUGE-2
Ours	**0.40581**	**0.09396**
RankClus	0.40153	0.08957
Spectral-Analysis	0.39922	0.08700
Cluster-HITS	0.38315	0.08632
Local-Rank	0.38104	0.08841
Global-Rank	0.37478	0.08531

large projection lengths on a beam play a leading role in the corresponding cluster. The summary is generated in the same way as Local-Rank. (5) RankClus [7], which uses frequency relationships between sentences, sentences and terms, ignoring the relationship between terms and terms. Our proposed approach uses semantic relationships between sentences and sentences, sentences and terms and terms and terms. RankClus clusters and ranks sentences simultaneously in the framework similar to ours. The classical clustering algorithm K-means is used in the above approaches, where clustering algorithm is needed. For query-based summarization, the additional query-relevance (i.e. the cosine similarity between sentences and query) is involved to re-rank the candidate sentences chosen by the ranking approaches for generic summarization.

Note that K-means requires a predefined cluster number K. To avoid exhaustive search for a proper cluster number for each document set, we employ the spectra approach introduced in [21] to predict the number of the expected clusters. Based on the sentence similarity matrix W_{SS} using the normalized 1-norm, for its eigenvalues λ_i $(i = 1, 2, \ldots, n)$, the ratio $\theta_i = \lambda_{i+1}/\lambda_2 (\lambda \geq 1)$ is defined. If $\theta_i - \theta_{i+1} > 0.05$ and θ_i is still close to 1, then set $K = i + 1$. With this approach, it is flexible to determine the numbers of clusters that are most adaptive to the topic distributions in different document sets. Tables II–V compare the performance of the six approaches on DUC 2004–2007 according to the calculated K.

TABLE V
RESULTS ON THE DUC 2007 DATASET

DUC 2007	ROUGE-1	ROUGE-2
Ours	**0.41622**	**0.12013**
RankClus	0.41047	0.11579
Spectral-Analysis	0.40906	0.10341
Cluster-HITS	0.39816	0.09104
Local-Rank	0.39489	0.08848
Global-Rank	0.38759	0.08012

It is not surprised to find that “Global-Rank” shows the poorest performance, when it utilizes the sentence level information only whereas the other five approaches all integrate the additional cluster level information in various ways. In addition, as results illustrate, the performance of “Cluster-HITS” is better than the performance of “Local-Rank.” This can be mainly credited to the ability of “Cluster-HITS” to consider not only the cluster-level information, but also the sentence-to-cluster relationships, which are ignored in “Local-Rank.” The “Spectral-Analysis” shows better performance than “Cluster-HITS,” because it uses spectral analysis to detect the clustering structure which can cluster and rank sentences simultaneously and effectively than any of the aforementioned document summarization approaches. It is glad to see that the proposed approach shows the best performance, because it updates sentence ranking and clustering interactively and iteratively, while “Spectral-Analysis” ranks and clusters sentences only once. Although “RankClus” also updates sentence ranking and clustering interactively and iteratively, it only uses frequency relationships between sentences, sentences and terms, but does not take the relationship between terms into consideration. The proposed approach uses cosine similarity relationships between sentences, terms, as well as sentences and terms, so it performs best among the six approaches.

Besides the above numerical analysis, we take the DUC2004 D30002 document set as an example to further illustrate the limitations of existing cluster-based summarization approaches. According to the spectral approach introduced in [21], there exist four sentence clusters. For ease of illustration, we list the generated summary by “Spectral-Analysis” and the proposed reinforced approach in Tables VI and VII, respectively.

When we compare the above two system-generated summaries with the human-generated summaries, the second one appears much more similar to the human-generated summaries. This can be credited to that the proposed reinforced approach can get high quality sentence clusters.

B. T-Test Evaluation

The significance of the improvement is always of concern. To examine it, we further conduct the paired t-test evaluation using ROUGE-2 scores, the primary DUC evaluation criterion, on all 50 DUC2004 document set, DUC2005 document sets, 50 DUC2006 document sets and 45 DUC 2007 document sets. The hypothesis here is that “the first approach is equal to or inferior to the second one in ROUGE-2” and the significance level is 5%.

The P-values presented in Table VIII suggests that all the hypotheses are rejected, which means the first approach is superior

TABLE VI
THE GENERATED DUC2004 D3002 SUMMARY WITH THE SPECTRAL-ANALYSIS METHOD

Hurricane Mitch paused in its whirl through the western Caribbean on Wednesday to punish Honduras with 120-mph (205-kph) winds, topping trees, sweeping away bridges, flooding neighborhoods and killing at least 32 people.

Aid workers struggled Friday to reach survivors of Hurricane Mitch, who are in danger of dying from starvation and disease in the wake of the storm that officials estimate killed more than 10,000 people.

Better information from Honduras"ravaged countryside enabled officials to lower the confirmed death toll from Hurricane Mitch from 7,000 to about 6,100 on Thursday, but leaders insisted the need for help was growing".

At least 231 people have been confirmed dead in Honduras from former- hurricane Mitch, bringing the storm's death toll in the region to 357, the

TABLE VII
THE GENERATED DUC2004 D3002 SUMMARY WITH THE PROPOSED REINFORCED APPROACH

The 350-mile (560-kilometer) wide hurricane was moving west at 8 mph (12 kph).

The strongest hurricane to hit Honduras in recent memory was Fifi in 1974, which ravaged Honduras' Caribbean coast, killing at least 2,000 people.

Mitch was drifting west at only 2 mph (3 kph) over the Bay Islands, Honduras' most popular tourist area.

With the storm seemingly anchored off Honduras where hundreds of people remained in shelters as a precaution Wednesday night.

More than 72,000 people had been evacuated to shelters.

In Washington on Thursday, President Bill Clinton ordered dlrs 30 million in Defense Department equipment and services and dlrs 36 million in food, fuel and other aid be sent to Honduras, Nicaragua, El Salvador and Guatemala.

Hurricane Mitch cut through the Honduran coast like l

TABLE VIII
T-TEST EVALUATION ON DUC2004–DUC2007

	DUC2004	DUC2005	DUC2006	DUC2007
Ours vs. RankClus	0.01387	0.01740	0.01452	0.01632
Ours vs. Spectral-Analysis	0.02569	0.02332	0.02556	0.02621
Ours vs. Cluster-HITS	0.02834	0.02974	0.02834	0.02990
Ours vs. Local-Rank	0.03153	0.03329	0.03440	0.03489
Ours vs. Global-Rank	0.03951	0.03839	0.03668	0.03953

to the second one. Table VIII also demonstrates our proposed reinforced approach outperforms the other five approaches.

C. Comparison With DUC Systems

TABLE IX
COMPARISON WITH DUC PARTICIPATING SYSTEMS ON THE DUC2004 DATASET

	ROUGE-1	ROUGE-2
System 67	0.36684	0.08995
Ours	0.37475	0.08973
System 66	0.36382	0.08814
System 104	0.37045	0.08527
NIST Average	0.33220	0.06796

TABLE X
COMPARISON WITH DUC PARTICIPATING SYSTEMS ON THE DUC2005 DATASET

	ROUGE-1	ROUGE-2
System 15	0.37665	0.07381
Ours	0.36451	0.07375
System 17	0.36930	0.07256
System 10	0.36298	0.07042
NIST Average	0.33366	0.05852

TABLE XI
COMPARISON WITH DUC PARTICIPATING SYSTEMS ON THE DUC2006 DATASET

	ROUGE-1	ROUGE-2
System 24	0.41108	0.09558
Ours	0.40581	0.09396
System 15	0.40279	0.09097
System 12	0.40100	0.08921
NIST Average	0.37256	0.07391

Next, we compare the proposed approach with 34 DUC2004 participating systems, 32 DUC2005 participating systems, 35 DUC2006 participating systems and 32 DUC2007 participating systems, respectively. Due to the page limitation, only the results of the top three systems are presented in Table IX–XII. As ROUGE-2 is the primary DUC evaluation criterion, the performances of the systems are ordered according to the ROUGE-2 values. For reference, the average ROUGE scores of all the submitted systems from the official results of NIST are also provided (denoted as "NIST Average").

The advantage of the proposed reinforced approach is clearly demonstrated in the above tables. They produce very competitive results, which significantly outperform the NIST averages and are comparable to the top-performing systems in those years.

D. Discussion on Influence of Cluster Number on Summarization

In previous experiments, the cluster number is predicted through the eigenvalues of 1-norm normalized sentence similarity matrix. This number is just the estimated number. The actual number is hard to predict accurately. To further examine how the cluster number influences summarization, we conduct the following additional experiments by varying the cluster number. Given a document set, we let S denote the sentence set in the document set, and set K in the following way:

$$K = \varepsilon \times |S|, \tag{18}$$

where $\varepsilon \in (0, 1)$ is a ratio controlling the expected cluster number. The larger ε is, the more clusters will be produced.

TABLE XII
COMPARISON WITH DUC PARTICIPATING SYSTEMS ON THE DUC2007 DATASET

	ROUGE-1	ROUGE-2
System 15	0.44100	0.12392
Ours	0.41622	0.12013
System 29	0.42983	0.12005
System 4	0.43005	0.11809
NIST Average	0.39378	0.09443

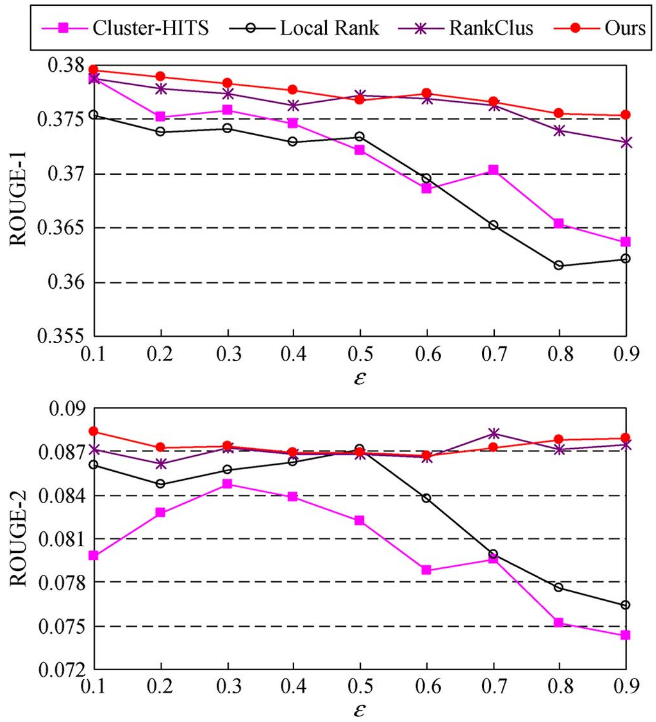


Fig. 4. ROUGEs vs. ε on DUC 2004.

ε ranges from 0.1 to 0.9 in the experiments. Due to page limitation, we only provide the ROUGE-1 and ROUGE-2 results of the proposed approach, "RankClus," "Spectral-Analysis," "Cluster-HITS" and "Local-Rank" on the DUC 2004 dataset in Fig. 4. The similar curves are also observed on the DUC2005–2007 datasets.

It is shown that (1) the proposed approach outperforms "RankClus," "Spectral-Analysis," "Cluster-HITS" and "Local-Rank" in almost all the cases no matter how the cluster number is set; (2) the performances of "Cluster-HITS" and "Local-Rank" are more sensitive to the cluster number and a large number of clusters appears to deteriorate the performances of both. This is reasonable. Actually when ε getting close to 1, "Local-Rank" approaches to "Global-Rank." These results demonstrate the robustness of the proposed approach. The advantages can be credited to: (1) our proposed reinforced approach clusters and ranks sentences simultaneously, while either "Cluster-HITS" or "Local-Rank" clusters and ranks sentences separately. Thus the performance of the latter two approaches will be highly affected by the detected theme clusters; (2) our proposed reinforced approach not only explores sentence-to-sentence, sentence-to-term and term-to-term relationships, but also explores distributions of terms and sentences within and across each theme cluster, while "Spectral-Analysis" makes use of sentence-to-cluster relationships and sentence-to-sentence relationships within each cluster only. Effectively utilizing multi-faceted associated relationships and distributions of terms and sentences will certainly release the negative impact of the undesired inaccurate clustering results.

V. CONCLUSION AND FUTURE WORK

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. Then, we use a mixture model to decompose each sentence into a K-dimensional vector, where each dimension is a component coefficient with respect to a cluster, which is measured by rank distribution. 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. Experimental results on the DUC2004–2007 datasets demonstrate the effectiveness of the proposed approach, which clearly outperforms the existing cluster-based approaches in the literature. We explore information of terms and sentences in this study.

The proposed approach can be applied to multi-human conversation analysis and automatic speech summarization as well. As effective speech summarization would also be expected to reduce the time required to review speech documents 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.

APPENDIX

Proof of Theorem 1: From the second line of (3), we get

$$\boldsymbol{r}(\boldsymbol{T})=(1-\beta)\cdot\frac{\mathbf{W}_{\mathbf{TT}}\cdot\boldsymbol{r}(\boldsymbol{T})}{\|\mathbf{W}_{\mathbf{TT}}\cdot\boldsymbol{r}(\boldsymbol{T})\|}+\beta\cdot\frac{\mathbf{W}_{\mathbf{TS}}\cdot\boldsymbol{r}(\boldsymbol{S})}{\|\mathbf{W}_{\mathbf{TS}}\cdot\boldsymbol{r}(\boldsymbol{S})\|}$$
$$\left[I-(1-\beta)\cdot\frac{\mathbf{W}_{\mathbf{TT}}}{\|\mathbf{W}_{\mathbf{TT}}\cdot\boldsymbol{r}(\boldsymbol{T})\|}\right]\cdot\boldsymbol{r}(\boldsymbol{T})=\beta\cdot\frac{\mathbf{W}_{\mathbf{TS}}\cdot\boldsymbol{r}(\boldsymbol{S})}{\|\mathbf{W}_{\mathbf{TS}}\cdot\boldsymbol{r}(\boldsymbol{S})\|}$$

thus,

$$\boldsymbol{r}(\boldsymbol{T})=\left[I-(1-\beta)\cdot\frac{\mathbf{W}_{\mathbf{TT}}}{\|\mathbf{W}_{\mathbf{TT}}\cdot\boldsymbol{r}(\boldsymbol{T})\|}\right]^{-1}\cdot\beta\cdot\frac{\mathbf{W}_{\mathbf{TS}}\cdot\boldsymbol{r}(\boldsymbol{S})}{\|\mathbf{W}_{\mathbf{TS}}\cdot\boldsymbol{r}(\boldsymbol{S})\|}$$

Combine (1) and (2), we obtain

$$\begin{aligned}\boldsymbol{r}(\boldsymbol{S})=&\alpha\cdot\frac{\mathbf{W}_{\mathbf{ST}}\cdot\left(\mathbf{I}-(1-\beta)\cdot\frac{\mathbf{W}_{\mathbf{TT}}}{\|\mathbf{W}_{\mathbf{TT}}\cdot\boldsymbol{r}(\boldsymbol{T})\|}\right)^{-1}\cdot\beta\cdot\frac{\mathbf{W}_{\mathbf{TS}}\cdot\boldsymbol{r}(\boldsymbol{S})}{\|\mathbf{W}_{\mathbf{TS}}\cdot\boldsymbol{r}(\boldsymbol{S})\|}}{\left\|\mathbf{W}_{\mathbf{ST}}\cdot\left(\mathbf{I}-(1-\beta)\cdot\frac{\mathbf{W}_{\mathbf{TT}}}{\|\mathbf{W}_{\mathbf{TT}}\cdot\boldsymbol{r}(\boldsymbol{T})\|}\right)^{-1}\cdot\beta\cdot\frac{W_{TS}\cdot\boldsymbol{r}(\boldsymbol{S})}{\|W_{TS}\cdot\boldsymbol{r}(\boldsymbol{S})\|}\right\|}\\&+(1-\alpha)\cdot\frac{\mathbf{W}_{\mathbf{SS}}\cdot\boldsymbol{r}(\boldsymbol{S})}{\|\mathbf{W}_{\mathbf{SS}}\cdot\boldsymbol{r}(\boldsymbol{S})\|}\\=&\alpha\cdot\frac{\beta\cdot\mathbf{W}_{\mathbf{ST}}\cdot(\mathbf{I}-(1-\beta)\cdot\mathbf{W}_{\mathbf{TT}})^{-1}\cdot\mathbf{W}_{\mathbf{TS}}\cdot\boldsymbol{r}(\boldsymbol{S})}{\left\|\beta\cdot\mathbf{W}_{\mathbf{ST}}\cdot(\mathbf{I}-(1-\beta)\cdot\mathbf{W}_{\mathbf{TT}})^{-1}\cdot\mathbf{W}_{\mathbf{TS}}\cdot\boldsymbol{r}(\boldsymbol{S})\right\|}\\&+(1-\alpha)\cdot\frac{\mathbf{W}_{\mathbf{SS}}\cdot\boldsymbol{r}(\boldsymbol{S})}{\|\mathbf{W}_{\mathbf{SS}}\cdot\boldsymbol{r}(\boldsymbol{S})\|}\end{aligned}$$

As the iterative process is a power method, it is guaranteed that $\boldsymbol{r}(\boldsymbol{S})$ converges to the primary eigenvector of $\alpha\beta \cdot \mathbf{W_{ST}} \cdot (\mathbf{I} - (1-\beta) \cdot \mathbf{W_{TT}})^{-1} \cdot \mathbf{W_{TS}} + (1-\alpha) \cdot \mathbf{W_{SS}}$. Similarly, $\boldsymbol{r}(\boldsymbol{T})$ is guaranteed to converge to the primary eigenvector of $\alpha\beta \cdot \mathbf{W_{TS}} \cdot (\mathbf{I} - (1-\alpha) \cdot \mathbf{W_{SS}})^{-1} \cdot \mathbf{W_{ST}} + (1-\beta) \cdot \mathbf{W_{TT}}$.■

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] R. Barzilay and K. R. Mckeown, “Sentence fusion for multi-document news summarization,” *Comput Linguist.*, vol. 31, no. 3, pp. 297–327, 2005.

[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, pp. 3816–3827, May 2011.

[6] X. Y. Cai and W. J. Li, “Enhancing sentence level clustering with integrated and interactive frameworks for theme—Based summarization,” *J. Amer. Soc Inf. Sci. Tech.*, vol. 62, no. 10, pp. 2067–2082, Oct. 2011.

[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, pp. 1597–1607, Jul. 2012.

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

[10] A. Celikyilmaz and D. Hakkani-Tur, “A hybrid hierarchical model for multi-document summarization,” in *Proc. 48th Annu. Meeting Assoc. Comput. Linguist.*, 2010, pp. 815–824.

[11] J. G. Corbonell and J. Goldstein, “The use of MMR, diversity-based reranking for reordering documents and producing summaries,” in *Proc. 21st SIGIR Conf.*, 1998, pp. 335–336.

[12] I. S. Dhillon, “Co-clustering documents and words using Bipartite spectral graph partitioning,” in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2001, pp. 269–274.

[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.

[14] E. Filatova and V. Hatzivassiloglou, “Event-based extractive summarization,” in *Proc. 42nd ACL Conf.*, 2004, pp. 104–111.

[15] S. Fisher and B. Roark, “Query-focused summarization by supervised sentence ranking and skewed word distributions,” in *Proc. DUC’06*, 2006.

[16] 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.

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

[18] S. Harabagiu and F. Lacatusu, “Topic themes for multi-document summarization,” in *Proc. 28th SIGIR Conf.*, 2005, pp. 202–209.

[19] H. Hardy, N. Shimizu, T. Strzalkowski, L. Ting, G. B. Wise, and X. Zhang, “Cross-document summarization by concept classification,” in *Proc. 25th SIGIR Conf. ’02*, 2002, pp. 121–128.

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

[21] W. J. Li, W. K. Ny, Y. Liu, and K. L. Ong, “Enhancing the effectiveness of clustering with spectra analysis,” *IEEE T Knowl Data Eng.*, vol. 19, no. 7, pp. 887–902, 2007.

[22] W. J. Li, W. Li, Q. Chen, and M. L. Wu, “The Hong Kong Polytechnic University at DUC2005,” in *Proc. DUC’05*, 2005.

[23] C. Y. Lin and E. Hovy, “The automated acquisition of topic signature for text summarization,” in *Proc. 18th COLING Conf.*, 2000, pp. 495–501.

[24] H. Lin and J. Bilmes, “A class of submodular functions for document summarization,” in *Proc. 49th ACL Conf.*, 2011.

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

[26] K. R. Mckeown, J. L. Kalvans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin, “Towards multi-document summarization by reformulation: Progress and prospects,” in *Proc. 13th AAAI Conf.*, 1999, pp. 121–128.

[27] R. Mihalcea, “Graph-based ranking algorithms for sentence extraction, applied to text summarization,” in *Proc. 42nd ACL Conf.*, 2004, pp. 20–24.

[28] R. Mihalcea, “Language independent extractive summarization,” in *Proc. 42nd ACL Conf.*, 2004, pp. 49–52.

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

[30] 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.

[31] D. R. Radev, J. Otterbacher, H. Qi, and D. Tam, “MEAD ReDUCs: Michigan at DUC2003,” in *Proc. DUC’03*, 2003.

[32] P. Sun, J. H. Lee, D. H. Kim, and C. M. Ahn, “Multi-document using weighted similarity between topic and clustering-based non-negative semantic feature,” in *Proc. APWEB/WAIM Conf.*, 2007, pp. 60–63.

[33] 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.

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

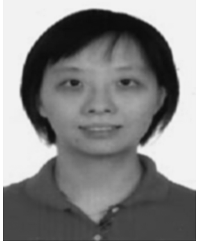
[35] X. J. Wan and J. W. Yang, “Multi-document summarization using cluster-based link analysis,” in *Proc. 31st SIGIR Conf.*, 2008, pp. 299–306.

[36] D. D. Wang, S. H. Zhu, T. Li, Y. Chi, and Y. H. Gong, “Integrating clustering and multi-document summarization to improve document understanding,” in *Proc. 17th CIKM Conf.*, 2008, pp. 1435–1436.

[37] D. D. Wang, T. Li, S. H. Zhu, and C. Ding, “Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization,” in *Proc. 31st SIGIR Conf.*, 2008, pp. 307–314.

[38] L. Zhao, X. J. Huang, and L. D. Wu, “Fudan University at DUC2005,” in *Proc. DUC2005*.

[39] D. Joris, V. Joris, V. Paul-Amand, and C. Dirk, “Pairwise-adaptive dissimilarity measure for document clustering,” *Inf. Sci.: An Int. J.*, vol. 180, no. 12, pp. 2341–2358.



Xiaoyan Cai is currently an assistant professor in College of Information Engineering, Northwest A&F University. She was a research associate in the Department of Computing, the Hong Kong Polytechnic University, Hong Kong, from May 2012 to June 2012, June 2009 to June 2011. She received the PHD degrees from Northwestern Polytechnical University, China, in 2009. Her current research interests include document summarization, information retrieval and machine learning.

Wenjie Li is currently an associate professor in Department of Computing, the Hong Kong Polytechnic University, Hong Kong. She received her PhD degree from Department of Systems Engineering and Engineering Management in the Chinese University of Hong Kong, Hong Kong, in 1997. Her main research topics include natural language processing, information extraction and document summarization.