Information Sciences 181 (2011) 3816–3827

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A spectral analysis approach to document summarization: Clustering and ranking sentences simultaneously

Xiaoyan Cai, Wenjie Li *

Department of Computing, The Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Article history:
Received 16 August 2010
Received in revised form 29 March 2011
Accepted 27 April 2011
Available online 10 May 2011

Keywords:
Document summarization
Sentence clustering
Sentence ranking
Spectral analysis

ABSTRACT

Automatic document summarization aims to create a compressed summary that preserves the main content of the original documents. It is a well-recognized fact that a document set often covers a number of topic themes with each theme represented by a cluster of highly related sentences. More important, topic themes are not equally important. The sentences in an important theme cluster are generally deemed more salient than the sentences in a trivial theme cluster. Existing clustering-based summarization approaches integrate clustering and ranking in sequence, which unavoidably ignore the interaction between them. In this paper, we propose a novel approach developed based on the spectral analysis to simultaneously clustering and ranking of sentences. Experimental results on the DUC generic summarization datasets demonstrate the improvement of the proposed approach over the other existing clustering-based approaches.

1. Introduction

With the rapid growing popularity of the Internet and a variety of information services, obtaining the desired information within a short amount of time becomes a serious problem in the information age. As such, new technologies that can process information efficiently are in great need. Automatic document summarization, i.e., a process of reducing the size of documents while preserving their important semantic content, is an essential technology to overcome this obstacle. 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. Abstractive 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 [1]. Like most researchers in this field, we follow the extractive summarization framework [24] in this work.

Graph-based ranking approaches, such as PageRank [2,20] and HITS [11], have achieved much success in extractive summarization [5,28] in the past few years. Nevertheless, these approaches all make uniform use of the sentences [33] in the document sets. The information beyond the sentence level is totally ignored. Actually, in a given document set, there usually exist a number of themes (or topics) with each theme represented by a cluster of highly related sentences [9,10]. The 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 great influence on sentence ranking. Clustering-based approaches for document summarization attract more and more attention. Most of these approaches first cluster sentences and then rank sentences within each cluster [5,25,26,30,31]. Although the clustering-based HITS model [23] considered the

* Corresponding author. Tel.: +852 2766 7297.
E-mail addresses: csxcai@comp.polyu.edu.hk (X. Cai), cswjli@comp.polyu.edu.hk (W. Li).

0020-0255/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2011.04.052

cluster-level information and the sentence-to-cluster relationship, it still does not concern how to integrate sentence ranking and clustering together.

In this paper, we propose a novel approach that clusters and ranks sentences simultaneously based on the spectral analysis. Different from other existing clustering-based summarization approaches, this new approach explores the "clustering structure" of sentences before the actual clustering algorithm is performed. The special clustering structure, called the structure of beams, is discovered by analyzing the spectral characteristics of the sentence similarity network. It reveals a natural relationship between the information necessary for clustering and ranking, but is hidden in most sentences. The structure of beams illustrates the distribution of sentences, where each beam represents a cluster [6]. That is, 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. At the same time, the sentences with the large projection lengths on a beam play a leading role in the corresponding cluster. To generate a summary, we extract the most salient sentences from each cluster (according to the order of cluster size) until the size limit is reached. Experimental results show that the proposed approach is able to achieve a competing performance compared to other clustering-based approaches.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 then analyzes the spectral geometry of a similarity network and its connection to the clustering structure. Next, Section 4 proposes the new spectral-based summarization approach. Section 5 presents experiments and evaluations. Finally, Section 6 concludes the paper.

2. Related work

Under the framework of extractive summarization, sentence ranking is the issue of most concern. Traditional feature-based approaches evaluated sentence significance and ranked sentences relying on the features that were well-designed to characterize the different aspects of the sentences. The centroid-based approach [22] was among the most popular feature-based approaches. Other statistical features and linguistic features, such as term frequency, sentence position, and sentence dependency structure, have also been extensively investigated in the past.

The composite effects of the features were often linearly combined. The weights of them were either experimentally tuned or automatically derived by applying learning-based mechanisms [19,27]. Learning-based models were quite popular in DUC/TAC competitions [4,7], such as the discriminative training model that learnt the feature weights by co-training a probabilistic Support Vector Machine and a Naïve Bayesian classifier [27], the support vector regression model that predicted composite sentence scores [19] and the log-linear model learned by maximizing the self-defined metrics of sentence goodness [4].

In contrast, graph-based approaches like LexRank [5] and TextRank [17,18] modeled a document or a set of documents as a weighted text graph constructed by taking sentences as vertices and the similarity between sentences as edge weights. They took into account the global information and recursively calculated sentence significance from the entire text graph rather than simply relying on unconnected individual sentences. These approaches were inspired by PageRank [20] that has been successfully applied to rank Web pages in the Web graph. Besides, Ye et al. [29] proposed a document concept lattice that indexes the hierarchy of local topics tied to a set of frequent concepts and the corresponding sentences containing these topics. Once words that represent unified concepts in the documents are linked, they represented the sources as a document concept lattice. This data structure organizes the set of all concepts presented in the input into a direct acyclic graph, where nodes represent overlapping sets of concepts.

Clustering-based approaches were explored in recent years [3,12,16,21,23,26]. For example, Qazvinian and Radev [21] applied hierarchical agglomerative clustering algorithm to obtain sentence clusters, and then developed two strategies to extract sentences from the clusters to build a summary. One was to extract the first sentence in the order it appeared in the original documents from the largest to the smallest cluster, then the second ones and so on, until the summary length limit is reached. The other was to rank sentences within each cluster with LexRank and then choose the most salient sentence from each cluster, then the second most salient sentence of each cluster, and so on. Wan and Yang [23], on the other hand, proposed a clustering-based HITS model which formalized the sentence-cluster relationships as the authority-hub relationships in the HITS algorithm. Finally sentences which had high authority scores were selected to form a summary. Besides, Wang et al. [26] proposed a language model to simultaneously cluster and summarize documents. Nonnegative factorization was 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 were generated simultaneously. A flaw of these clustering-based approaches is that clustering and ranking are independent of each other and thus they cannot share the information that is useful for both, e.g. the spectral information of sentence similarity matrix. A new approach that can really couple clustering and ranking together is required in order to improve the performance of each other.

3. Similarity network and its spectral geometry characteristics

3.1. Preliminaries of similarity network

Most clustering algorithms, actually graph-based ranking algorithms as well, embed the data in a similarity space determined by a certain similarity measure, for example, the widely used cosine similarity. Given n data points, we can construct a similarity network $N(S) = (V, E, S)$, where V is the set of n nodes (i.e. data points) and E is the set of weighted edges, $S = (s_{ij})_{n\times n}$

is a similarity matrix, where s_{ij} denotes cosine similarity between the ith and jth nodes in the network, $0 \leqslant s_{ij} \leqslant 1$, $s_{ii} = 1$ and $s_{ij} = s_{ji}$, i.e. S is symmetric. Each node v_i of $N(S)$ corresponds to the ith row (or column) of S, and the weight of each edge e_{ij} corresponds to the non-diagonal entry s_{ij}. For any two nodes (v_i, v_j), a larger value of s_{ij} indicates a higher connectivity between them, and vice versa.

Liu et al. [15] defined the "total path" $\varphi(v_i, v_j)$ between two nodes v_i and v_j as the number of all paths starting from the ith node and ending at the jth node in a similarity network. $\varphi_k(v_i, v_j)$ indicates the number of all k-length paths between the nodes v_i and v_j, then $\varphi(v_i, v_j) = \sum_{k=1}^{\infty} \varphi_k(v_i, v_j)$. Estrada and Rodríguez-Velázquez [6] pointed out that the metric $\varphi(v_i, v_j)$ can be used to measure the neighborhood relationship of two nodes. There are two characteristics of these paths: (a) the contribution of the paths to the cluster decreases as the length of the paths increases, and (b) the sum of the paths with different length presupposes a mathematical problem as the series $\sum_{k=1}^{\infty} \varphi_k(v_i, v_j) = \infty$ diverges. In order to avoid this problem, they scaled the contribution of the paths to the cluster by dividing them by the factorial of the order of the spectral moment k, that is,

$$\varphi(v_i, v_j) = \sum_{k=1}^{\infty} \frac{\varphi_k(v_i, v_j)}{k!} \tag{1}$$

This is called the normalized path between the two nodes v_i and v_j.

3.2. Spectral geometry of similarity network and clustering structure

The original objective to define the normalized path in [6] is to determine the sub-graph centrality of a node, which is used as a metric to evaluate if a node plays an important role in a cluster. Nevertheless, it can not identify which cluster the node belongs to. We will prove one of the important properties of the normalized path, i.e. the large length of the normalized path between two nodes implies that the two nodes most probably belong to the same cluster; otherwise, they may be in different clusters. Based on this property, we further make use of the normalized path to characterize the neighborhood relationship of the two nodes in terms of the clustering structure. The similar idea was also used by Liu et al. [15] to analyze query search results.

Following the above analysis, $\varphi_k(v_i, v_j)$ can also be simply defined as the value in the ith row and the jth column of the kth power of the similarity matrix S, i.e.,

$$\varphi_k(v_i, v_j) = (S^k)_{ij} \tag{2}$$

Based on the close relationship between the power of a matrix and its eigenvalues and eigenvectors, i.e. spectral information, the following theorem shows that $\varphi(v_i, v_j)$ can be obtained mathematically from the spectra of the similarity matrix S.

Theorem 1. *Let $N(S) = (V, E, S)$ be a similarity network and $x_1, x_2, \ldots, x_n$ be eigenvectors of S associated with the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ $(\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n)$. Let $x_i(t)$ denote the tth component of x_i. Then for any two nodes v_i and v_j, $\varphi(v_i, v_j)$ can be expressed as*

$$\varphi(v_i, v_j) = \sum_{t=1}^{n} x_i(t) x_j(t) e^{\lambda_t} \tag{3}$$

which reveals the degree to which two nodes v_i and v_j belong to the same cluster.

Proof. The orthogonal projection of the unit vector e_i (whose ith component is 1 and the rest of components are 0) on x_t is

$$proj_{x_t}(e_i) = \frac{\langle e_i, x_t \rangle}{\|x_t\|^2} x_t = \langle e_i, x_t \rangle x_t = x_t(i) \cdot x_t \tag{4}$$

where $\langle \cdot, \cdot \rangle$ is the dot product of the two vectors. As n eigenvectors form an orthonormal basis of n-dimensional space, each vector e_i can be formulated as $e_i = \sum_{t=1}^{n} proj_{x_t}(e_i)$. Therefore, $\varphi_k(v_i, v_j)$ can be expressed in terms of the spectral information of the similarity matrix as follows,

$$\varphi_k(v_i, v_j) = (S^k)_{ij} = \left\langle S^k e_i, e_j \right\rangle = \left\langle S^k \sum_{t=1}^{n} proj_{x_t}(e_i), \sum_{t=1}^{n} proj_{x_t}(e_j) \right\rangle = \sum_{t=1}^{n} \lambda_t^k x_i(t) x_j(t) \tag{5}$$

Replacing Eq. (1) by Eq. (5), we obtain

$$\varphi(v_i, v_j) = \sum_{k=1}^{\infty} \left(\sum_{t=1}^{n} \frac{x_i(t) x_j(t) \lambda_t^k}{k!} \right) \tag{6}$$

By reordering the terms of series in Equation (6), we can get the absolutely convergent series:

$$\varphi(v_i, v_j) = \sum_{t=1}^{n} \left(x_i(t) x_j(t) \sum_{k=1}^{\infty} \frac{\lambda_t^k}{k!} \right) = \sum_{t=1}^{n} x_i(t) x_j(t) e^{\lambda_t} \tag{7}$$

After taking a closer look, Eq. (3) can then be rewritten as

$$\varphi(v_i, v_j) = \sum_{t=1}^{n} e^{\frac{1}{2}\lambda_t} x_i(t) e^{\frac{1}{2}\lambda_t} x_j(t) = \langle \hat{x}_i, \hat{x}_j \rangle \tag{8}$$

The dot product $\langle \hat{x}_i, \hat{x}_j \rangle$ is the multiplication of $\langle \hat{x}_i, \hat{x}_i \rangle, \langle \hat{x}_j, \hat{x}_j \rangle$ and the cosine of their angle. As indicated in [6], $\langle \hat{x}_i, \hat{x}_i \rangle$ or $\langle \hat{x}_j, \hat{x}_j \rangle$ is the sub-graph centrality of the node v_i or $v_j \cdot \varphi(v_i, v_j)$ reflects the geometric relationships of the two vectors and it also reveals the degree to which how the two nodes v_i and v_j belong to the same cluster. □

Moreover, the dot product is fundamentally projection. The dot product between $\hat{x}_i$ and $\hat{x}_j$ is the projection of $\hat{x}_i$ in the direction given by $\hat{x}_j$ multiplying $\hat{x}_j$ and so the normalized path $\varphi(v_i, v_j)$ can also be deemed as the multiplication of $\langle \hat{x}_i, \hat{x}_j \rangle$ and the projection of $\hat{x}_i$ on $\hat{x}_j$. According to the spectral geometry of the similarity network, we can deduce that, given a cluster in a network, the data points corresponding to the nodes within the cluster are distributed in the shape of a beam, where few nodes connect most nodes in the cluster, at the same time the remaining nodes connect relatively less nodes.

We also observe that the difference between the eigenvalues is significantly enlarged by an exponential function. It means that $|e^{\lambda_i} - e^{\lambda_j}| \gg |\lambda_i - \lambda_j|$. Thus in Eq. (8), larger λ_t contributes more to the dot product $\varphi(v_i, v_j)$ of any two points. So in the above theorem, as $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$, we can use the first k largest eigenvalues and corresponding eigenvectors to approximate the dot product over the original space, such as $\sum_{t=1}^{k} e^{\frac{1}{2}\lambda_t} x_i(t) e^{\frac{1}{2}\lambda_t} x_j(t) \approx \sum_{t=1}^{n} e^{\frac{1}{2}\lambda_t} x_i(t) e^{\frac{1}{2}\lambda_t} x_j(t)$. Based on this analysis, we can conclude that the possible beam number in the first k-dimensional space is still k.

4. A spectral-based approach to document summarization

Based on the spectral analysis introduced in the above section, we propose a novel spectral-based approach to document summarization which allows sentence clustering and sentence ranking within each cluster to be achieved simultaneously. Now, let us describe the proposed approach in detail.

4.1. Construction of sentence similarity network

According to the definition of similarity network provided in Section 3, given a document set D, we construct a sentence similarity network of D as $G = (V, E, W)$, where $V = \{v_i | 1 \leqslant i \leqslant n\}$ is the set of nodes and each node v_i in V represents a sentence, n is the sentence number in document set D. $E = \{e_{ij} | v_i, v_j \in V,\ i \neq j\}$ is the set of edges and each edge e_{ij} in E is associated with a weight w_{ij} between the sentences v_i and v_j The weight is computed using the standard cosine measure

$$w_{ij} = \frac{\vec{v}_i \cdot \vec{v}_j}{|\vec{v}_i| \cdot |\vec{v}_j|} \tag{9}$$

where $\vec{v}_i$ and $\vec{v}_j$ are the corresponding term vectors of the sentences v_i and v_j The term weight in the vector is set in terms of the TFISF of the term in the sentence, where TFISF [32] is the term frequency inverse sentence frequency. Based on sentence cosine similarity, a sentence similarity matrix W can be constructed, where $0 \leqslant w_{ij} \leqslant 1,\ w_{ii} = 1,\ w_{ij} = w_{ji},$ and $1 \leqslant i, j \leqslant n$.

In order to speed up the analysis process, some weighted edges are removed if the similarity between the nodes is less than a given threshold (we set the threshold to 0.1 in our experiments).

4.2. Detection of direction beams in the sentence similarity network

Based on the constructed sentence similarity network $G = (V, E, W)$, we can get the eigenvectors $x_1, x_2, \ldots, x_n$ of W associated with the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ $(\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n)$. As illustrated in Section 3.2, each beam is a description of the special clustering structure in the sentence similarity network, and the position where the sentence nodes are located in each beam shows the importance of the sentences in forming the corresponding cluster. So if we can get beams of the sentence similarity network, we could obtain corresponding sentence clusters.

We suppose there are k beams in the graph and initialize k vectors c_j $(j = 1, \ldots, k)$ as random unit vectors. Then normalized path between a beam and a node can be computed as the projection length of the node on the beam when the beam is a unit vector. So each node x_i can be projected to these k vectors. We then get $l_{x_i}^{c_j} = |proj_{c_j}(x_i)|$, where $l_{x_i}^{c_j}$ is the projection length of the node x_i on the vector c_j Based on it, the maximum projection length for the node x_i is $l_{x_i} = \max_{j=1}^{k} l_{x_i}^{c_j}$ Therefore, l_{x_i} is the projection length of the vector which the node x_i belongs to. Then we sum the squares of all the nodes' maximum projection length on the vector c, i.e., $L = \sum_{i=1}^{n} l_{x_i}^2$ It is expected that L is maximized when the k vectors c_j $(j = 1, \ldots, k)$ is the direction vector of the k beams. In this way, the detection of beams is then formulated to a maximization problem. As the function L is the sum of squared $l_{x_i}^c$ we convert the maximization problem of L to be a nonlinear least-square minimization problem, formulated as $L_e = \sum_{i=1}^{n} \exp^2(-l_{x_i})$ instead of $L = \sum_{i=1}^{n} l_{x_i}^2$ The algorithm for detecting direction vectors of beams in the sentence similarity network is described in Table 1.

Table 1
Detection of direction beams in the sentence similarity network.

Input: (1) $G(V, E, W)$, sentence similarity network ($n = |V|$); and (2) $X_{n\times k}$: the matrix whose column is the first k eigenvectors corresponding to the first k largest eigenvalues of the adjacency matrix of the undirected graph
Output: $c_1, \ldots, c_k$, direction vectors of the k beams
// **Step 1: Intialization**
1 For $i = 1$ to k
2 Initialize the ith vector c_i to be a random unit vector
3 End for
// **Step 2: Finding the direction vectors of k beams**
// **Repeat step 4–11 until termination tolerance on L_e is reached**
// **or the maximum number of iteration is reached.**
4. For $i = 1$ to n
5. For $j = 1$ to k
6. Calculate the projection length of the ith point x_i on the jth vector c_j, i.e. $l_{x_i}^{c_j}$
7. End for
8. Get the maximum value for the ith sentence's projection length of the k direction vectors, i.e., $l_{x_i} = \max_{j=1}^{k} l_{x_i}^{c_j}$
9. End for
10. Return $L_e = \sum_{i=1}^{n} \exp^2(-l_{x_i})$
11. Call nonlinear least-squares optimization to update $c_1, \ldots, c_k$

4.3. Simultaneous clustering and ranking of sentences

Once obtain beams of the sentence similarity network, we can get each sentence's cluster label, i.e., a sentence should belong to a beam (cluster label) if its projection length is maximal on the beam compared to the other beams. After getting all sentences' cluster labels, we can sort the projection length of each sentence of each beam in decreasing order, i.e. ranking position of all sentences in corresponding clusters.

4.4. Sentence extraction and redundancy removal

It is widely agreed that the facts with higher weights appear in greater number of sentences and clustering should be able to cluster such fact-sharing sentences in the same communities [21]. Thus, starting with the largest cluster is important to ensure that the system-generated summary first covers the facts that have higher frequencies and therefore higher weights. Following this idea, we start with the largest cluster and extract sentences in the order according to their ranking position in each cluster to form a ranking list. That is, we extract the first sentences from each cluster, then the second ones, and so on.

Since the number of the documents to be summarized can be large in multi-document summarization, it makes information redundancy problem appear to be more serious in multi-document summarization than in single-document summarization. To alleviate this problem, we apply a simple yet effective way to choose summary sentences. At the beginning, we choose the first sentence from the sentences ranking list into the summary. Then we examine the next one and compare it with the sentence(s) already included in the summary. Only the sentence that is not too similar to any sentence in the summary (i.e. the cosine similarity between them is lower than a threshold) is selected into the summary. This process is repeated until the length of the sentences in the summary reaches the length limitation. The threshold is set to 0.9 in this paper.

5. Experiment and evaluation

5.1. Experiment setup

Generic multi-document summarization has been one of the fundamental tasks in DUC2001 and DUC2004 (i.e. task 2 in DUC2001 and task 2 in DUC2004). We use the datasets of these two tasks for evaluation. Table 2 presents basic statistics of the two datasets. Systems are required to create from the documents a brief, well-organized and fluent summary which can express the main information conveyed in the document sets.

We conduct a large number of experiments to evaluate the effectiveness of our proposed approach by comparing it with three existing clustering-based approaches, including Cluster Round-Robin (CRR) [21], Cluster LexRank (C-LexRank) [21] and ClusterHITS [23] on the same data set.

A well-recognized automatic evaluation toolkit ROUGE [14] is used in evaluation. It measures summary quality by counting the overlapping units between system-generated summaries and human-written reference summaries. We report three ROUGE scores in this paper, namely ROUGE-1, ROUGE-2 and ROUGE-SU4, which base on Uni-gram match, Bi-gram match and Skip-Bi-gram match, respectively.

Table 2
Summary of the two DUC datasets.

	DUC2001	DUC2004
Number of document sets	30	50
Average number of documents in each document set	10.3	10
Average number of sentences in each document set	383.9	276
Average number of sentences in each document	37.2	27.6
Average number of words in each document set	3279	4534
Average number of words in each document	468.43	453.4
Average number of words in each sentence	22.30	24.62
Summary length	100 words	665 bytes

5.2. Estimation of cluster number

All of the above clustering-based approaches require pre-defining a cluster number. To avoid exhaustive search for a proper cluster number for each document set, we employ the spectra approach introduced in [13] to predict the number of the expected clusters. Based on the sentence similarity matrix W using the normalized 1-norm, for its eigenvalues λ_i^{norm} $(i = 1, 2, \ldots, n)$, the ratio $\alpha_i = \lambda_{i+1}^{norm}/\lambda_2 (i \geqslant 1)$is defined. If $\alpha_i - \alpha_{i+1} > 0.05$ and α_i is still close to 1, then set the cluster number $k = i + 1$. With this approach, it is flexible to determine the numbers of clusters that are most adaptive to the topic distributions in different document sets.

Fig. 1 visualizes the sentence adjacency matrix for DUC2004 D30015 document set. It is clear that there are four dense clusters in Fig. 1, which corresponds to the cluster number estimated by the above spectra approach. When we take a closer look at the sentences in these clusters, four theme topics are found. They are about "NATO, airstrikes", "Yugoslavia, force, Kosovo", "Russian, strike occur" and "U.S. envoy, Milosevic, pull back military", respectively and correspond to the four clusters illustrated in the figure. From the figure, we see that the size of the second cluster is the biggest one, and then the forth cluster and the first cluster. The size of the third cluster is the smallest. So the summary sentence selection process will follow this sequence. The generated summary of D30015 looks like:

> "On Oct. 4, 1998 Yugoslav President Milosevic ordered his forces in Kosovo back to their barracks. Yugoslav President Slbodan Milosevic does not appear to be complying completely with UN demands to withdraw his troops and stop anti-Albanian activity in Kosovo. US envoy tells Milosevic to pull back his military and let Albanian refugees return home. U.S. special envoy Richard Holbrooke said the level of fighting may have abated but the situation is such that it could resume. Kosovo to suppress the ethnic Albanian separatist uprising be withdrawn and is now threatened with NATO airstrikes Russia. NATO threatened airstrikes unless hostilities ceased and peace talks began. Russia, previously against a NATO attack, said the strikes could occur if steps aren't taken to end the crisis."

Notice that in order to generate a coherent summary, so far we re-order the sentences selected in this way: (1) we group the sentences within the same cluster together and order them according to their projection length in that cluster; and (2) the different clusters are ordered according to their size. This ordering strategy is by no mean a perfect one. The semantics or discourse connection among the sentences will be studied in our future work.

5.3. Comparison with Other clustering-based approaches

As indicated in [8], bisecting K-means and hierarchical agglomerative algorithms are two competing text clustering techniques. So we apply both of them in CRR, C-LexRank and ClusterHITS algorithms. The average recalls of ROUGE-1, ROUGE-2 and ROUGE-SU4 are illustrated in Tables 3 and 4.

From Tables 3 and 4, we can seen that C-LexRank shows better performance than CRR, because LexRank can find the most salient sentence in each cluster, while CRR only select sentences in each cluster without considering their centrality. ClusterHITS performs better than C-LexRank, 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 LexRank. We also observe that the bisecting K-means-based approaches and agglomerative-based approaches show relatively similar performances, this is equivalent to the conclusion that "the bisecting K-means technique is as good as the hierarchical agglomerative technique", which is indicated in [8]. It is happy to see that the proposed approach shows the best performance, because it uses spectral analysis to detect the clustering structure which can cluster and rank sentences more effectively than any of the aforementioned document summarization approaches.

5.4. Comparison with DUC systems

Next, we compare the proposed approach with a coverage baseline, which takes the first sentence from the first document to the last document if the summary length permits, where documents are assumed to be ordered chronologically.

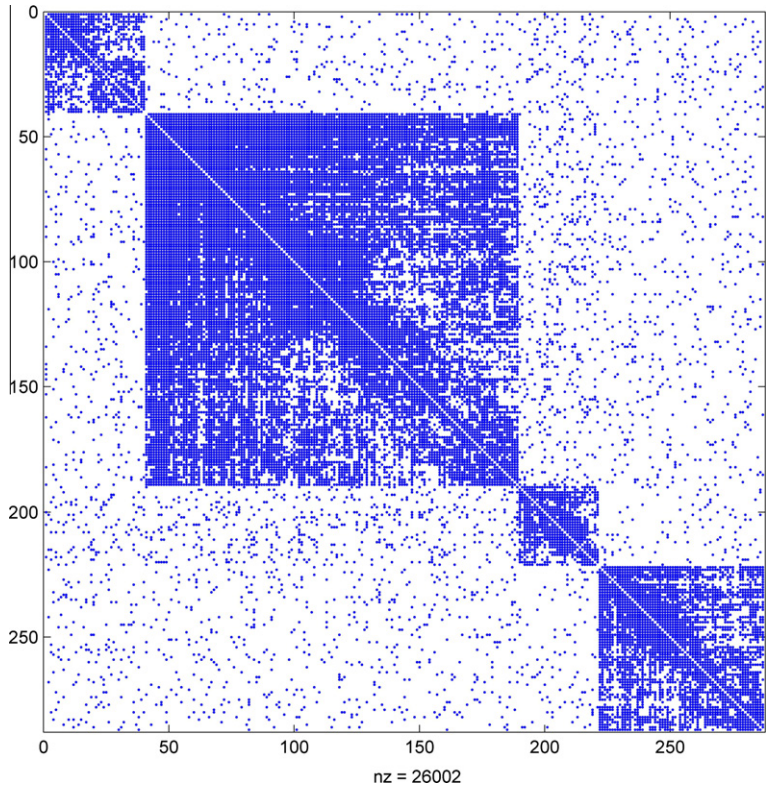

Fig. 1. Clustering structure of the sentence adjacency matrix.

Table 3
Evaluation of the approaches on the DUC2001 dataset.

	ROUGE-1	ROUGE-2	ROUGE-SU4
Ours	0.35732	0.06293	0.11215
ClusterHITS (agglomerative)	0.34845	0.06074	0.10407
ClusterHITS (bisecting K-means)	0.34730	0.05957	0.10380
C-LexRank (agglomerative)	0.34294	0.05852	0.09973
C-LexRank (bisecting K-means)	0.34275	0.05736	0.09839
CRR (bisecting K-means)	0.33798	0.05631	0.09538
CRR (agglomerative)	0.33583	0.05583	0.09306

Table 4
Evaluation of the approaches on the DUC2004 dataset.

	ROUGE-1	ROUGE-2	ROUGE-SU4
Ours	0.37875	0.09435	0.13263
ClusterHITS (agglomerative)	0.37594	0.08413	0.13022
ClusterHITS (bisecting K-means)	0.37418	0.08326	0.12987
C-LexRank (agglomerative)	0.37362	0.08293	0.12557
C-LexRank (bisecting K-means)	0.37293	0.08152	0.12359
CRR (bisecting K-means)	0.36911	0.08351	0.12179
CRR (agglomerative)	0.36874	0.08266	0.11986

For comparison purpose, the ROUGEs results of top three DUC systems participating in DUC2001 and DUC2004 are also included. For the sake of analysis, we only concentrated on agglomerative based approaches here. Tables 5 and 6 present the comparison results.

Table 5
Comparison with DUC participating systems on the DUC2001 dataset.

	ROUGE-1	ROUGE-2	ROUGE-SU4
Ours	0.35732	0.06293	0.11215
ClusterHITS (agglomerative)	0.34845	0.06074	0.10407
C-LexRank (agglomerative)	0.34294	0.05852	0.09973
CRR (agglomerative)	0.33583	0.05583	0.09306
System O	0.33082	0.06771	0.09231
System N	0.33045	0.06439	0.09307
System T	0.33005	0.06180	0.09422
Coverage	0.30632	0.04317	0.08269

Table 6
Comparison with DUC participating systems on the DUC2004 dataset.

	ROUGE-1	ROUGE-2	ROUGE-SU4
System 65	0.37816	0.09147	0.13178
Ours	0.37793	0.09034	0.13106
ClusterHITS (agglomerative)	0.37594	0.08413	0.13022
C-LexRank (agglomerative)	0.37362	0.08293	0.12557
System 35	0.37076	0.08335	0.12733
System 104	0.37045	0.08527	0.12763
CRR (agglomerative)	0.36874	0.08266	0.11986
Coverage	0.34729	0.06983	0.10498

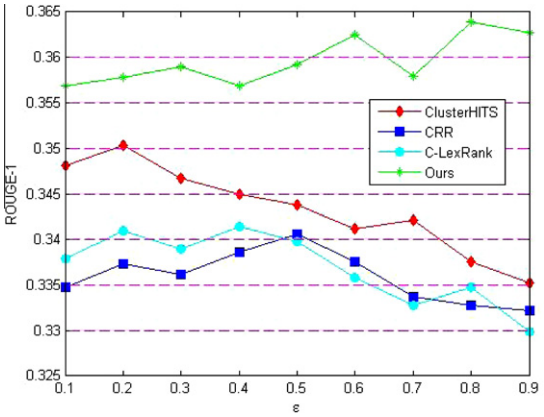

Fig. 2. ROUGE-1 vs. ε on DUC2001.

The advantages of the proposed approach are clearly demonstrated in Tables 5 and 6. It produces very competitive results, which apparently outperforms the coverage baseline in both years. More important, it is ahead of the best system in DUC2001 on ROUGE-1, and ranks the second in DUC2004. It is quite encouraging to us. Notice that in our current experiments which focus on the ranking and clustering relationships, the position of a sentence in the document is not considered yet. Nevertheless the position feature has been employed in all the participating systems as one of the most significant features.

5.5. Discussion on influence of cluster number on summarization

In previous experiments, the cluster number is predicated through the eigenvalues of the 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 the summarization performance, we conduct the following additional experiments by varying the cluster number. Given a document set, we let V denote the sentence collection for the document set, and k is set to the following way:

$$k = \varepsilon \times |V| \tag{10}$$

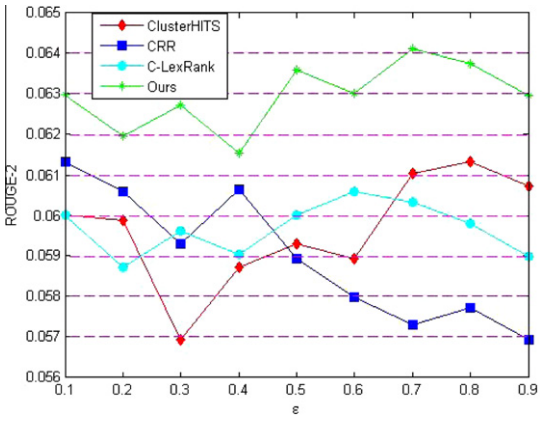

Fig. 3. ROUGE-2 vs. ε on DUC2001.

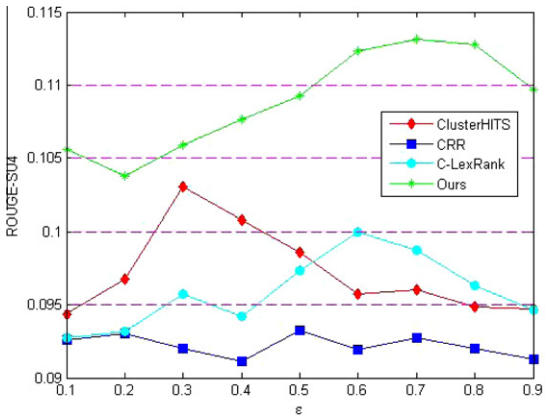

Fig. 4. ROUGE-SU4 vs. ε on DUC2001.

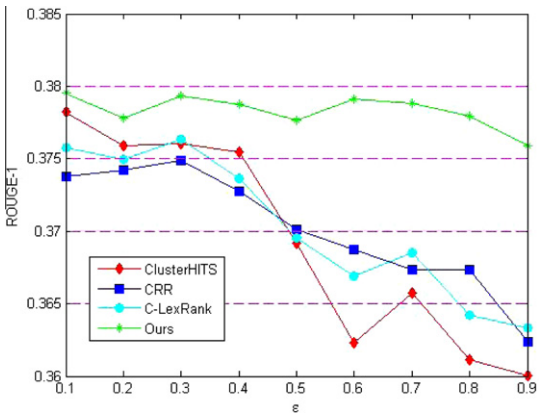

Fig. 5. ROUGE-1 vs. ε on DUC2004.

where $\varepsilon \in (0, 1)$ is a ratio controlling the expected cluster number for the document set. The larger ε is, the more clusters will be produced and used in corresponding clustering algorithms. ε ranges from 0.1 to 0.9 in the experiments. Figs. 2–7 plot ROUGE-1, ROUGE-2 and ROUGE-SU4 curves of our spectral-based approach, ClusterHITS, C-LexRank and CRR on the DUC2001 and DUC2004 datasets, respectively.

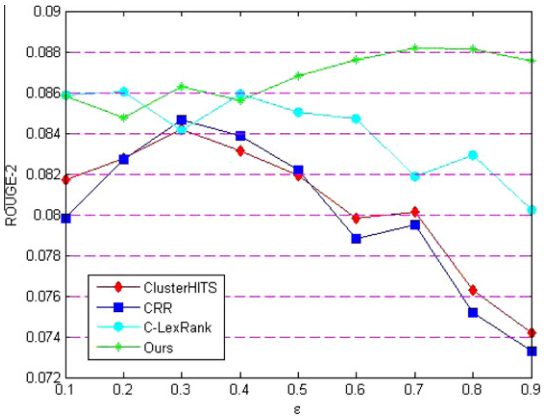

Fig. 6. ROUGE-2 vs. ε on DUC2004.

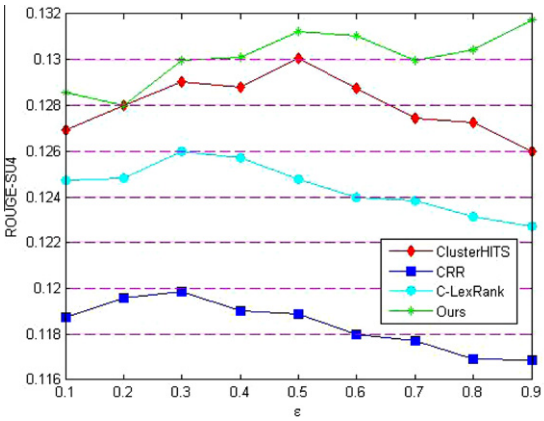

Fig. 7. ROUGE-SU4 vs. ε on DUC2004.

It is shown from the figures that (1) our approach can outperform ClusterHITS, C-LexRank and CRR approaches in most cases, no matter how the cluster number is set; (2) the performances of ClusterHITS, C-LexRank and CRR algorithms are more sensitive to the cluster number and a large number of clusters appears to deteriorate the performances of them. The performances of C-LexRank and CRR are even worse than the ClusterHITS, when ε is set to a large value. These results demonstrate that our proposed approach is more robust than ClusterHITS, C-LexRank and CRR with respect to different cluster numbers. The advantages can be credited to: (1) Our proposed approach clusters and ranks sentences simultaneously, while either ClusterHITS or C-LexRank or CRR clusters and ranks sentences separately. Thus the performance of the latter three approaches will be highly affected by the detected theme clusters. (2) Our proposed approach not only explores sentence-to-sentence relationship, but also explores sentence-to-cluster relationship, while ClusterHITS or C-LexRank or CRR makes use of the sentence-to-cluster relationships or sentence-to-sentence relationships within each cluster only. Effectively utilizing multi-faceted associated relationships will certainly releases the negative impact of the undesired inaccurate clustering results.

6. Conclusion

In this paper, we develop a new summarization approach which can simultaneously cluster and rank sentences by investigating the spectral characteristics of the similarity network which is constructed upon the document(s). The proposed approach identifies the beams which represent the clustering structure. The sentence points projected onto one beam direction will belong to the same cluster. At the same time, the sentence points with a larger projection length play an important role in the corresponding cluster as they have much larger similarity with the beam. Experimental results on the DUC2001 and DUC2004 datasets demonstrate the effectiveness of the proposed approach, which clearly outperforms the existing clustering-based approaches in the literature. During the study, we notice that if we could differentiate the importance of beams

and to compare the different project length in different beams, we would be able to make the proposed approach more robust and more effective. In the future, we will focus on these issues. As the position sentence of a sentence in a document is a rather important factor for sentence selection in generic summarization, we will also consider to explore the possibility to integrate the position information in the simultaneous sentence clustering and ranking algorithm in our future work.

Acknowledgements

The work described in this paper was partially supported by a Hong Kong RGC Project (Project No. PolyU5217/07E) and partially supported by the university internal grants (Account Nos. G-YG80 and G-YH53).

References

[1] R. Barzilay, K.R. Mckeown, Sentence fusion for multi-document news summarization, Computational Linguistics 31 (3) (2005) 297–327.
[2] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Computer Networks 30 (1–7) (1998) 107–117.
[3] X.Y. Cai, W.J. Li, Simultaneous ranking and clustering of sentences: a reinforcement approach to multi-document summarization, in: Proceedings of 23rd International Conference on Computational Linguistics, 2010, pp. 134–421.
[4] J.M. Conroy, J.D. Schlesinger, CLASSY query-based multi-document summarization, in: Document Understanding Conferences, 2005.
[5] G. Erkan, D.R. Radev, LexRank: graph-based lexical centrality as salience in text summarization, Journal of Artificial Intelligence Research 22 (2004) 457–479.
[6] E. Estrada, J.A. Rodŕıguez-Veĺazquez, Subgraph centrality in complex networks, Physical Review E 71 (2005) 1–9.
[7] J. Ge, X. Huang, L. Wu, Approaches to event-focused summarization based on named entities and query words, in: Document Understanding Conferences, 2003.
[8] K. George, V. Kumar, M. Steinbach, A comparison of document clustering techniques, in: Proceedings of Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000, pp. 109–111.
[9] S. Harabagiu, F. Lacatusu, Topic themes for multi-document summarization, in: Proceedings of the 28th Annual International Conference on Research and Development in Information Retrieval, 2005, pp. 202–209.
[10] H. Hardy, N. Shimizu, T.L. Ting, G.B. Wise, X. Zhang, Cross-document summarization by concept classification, in: Proceedings of the 28th Annual International Conference on Research and Development in Information Retrieval, 2005, pp. 121–128.
[11] J. Kleinberg, Authoritative sources in a hyperlinked environment, in: Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 1998, pp. 668–677.
[12] J. Larocca Neto, A.D. Santos, C.A.A. Kaestner, A.A. Freitas, Document clustering and text summarization, in: Proceedings of the 4th International Conference on Practical Applications of Knowledge Discovery and Data Mining, 2000, pp. 41–55.
[13] W.Y. Li, W.K. Ng, Y. Liu, K.L. Ong, Enhancing the effectiveness of clustering with spectra analysis, IEEE Transactions on Knowledge and Data Engineering (TKDE) 19 (7) (2007) 887–902.
[14] C.Y. Lin, E. Hovy, Automatic evaluation of summaries using N-gram co-occurrence statistics, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003, pp. 71–78.
[15] Y. Liu, W.Y. Li, Y.J. Lin, L.P. Jing, Spectral geometry for simultaneously clustering and ranking query search results, in: Proceedings of the 31st Annual International Conference on Research and Development in Information Retrieval, 2008, pp. 539–546.
[16] D.X. Liu, Y.X. He, D.H. Ji, H. Yang, Z. Wu, Chinese multi-document summarization using adaptive clustering and global search strategy, Lecture Notes on Computer Science (LNCS) (2006) 1135–1139.
[17] R. Mihalcea, Graph-based ranking algorithms for sentence extraction, applied to text summarization, in: Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions, 2004, pp. 170–173.
[18] R. Mihalcea, Language independent extractive summarization, in: Proceedings of the 20th National Conference on Artificial Intelligence, 2005, pp. 1688–1689.
[19] Y. Ouyang, S.J. Li, W.J. Li, Developing learning strategies for topic-based summarization, in: Proceedings of ACM 16th Conference on Information and Knowledge Management, 2007, pp. 79–86.
[20] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: bringing order to the web, Technical Report, Stanford Digital Library Technologies Porject, 1998.
[21] V. Qazvinian, D.R. Radev, Scientific paper summarization using citation summary networks, in: Proceedings of 22nd International Conference on Computational Linguistics, 2008, pp. 689–696.
[22] D.R. Radev, H.Y. Jing, M. Stys, D. Tam, Centroid-based summarization of multiple documents, Information Processing and Management 40 (2004) 919–938.
[23] X. Wan, J. Yang, Multi-document summarization using cluster-based link analysis, in: Proceedings of the 31st Annual International Conference on Research and Development in Information Retrieval, 2008, pp. 299–306.
[24] L. Antiqueris, O.N. Oliveira, L.F. Costa, M.G. Nunes, A complex network approach to text summarization, Information Sciences 179 (5) (2009) 584–599.
[25] X. Wan, J. Yang, Improved affinity graph based multi-document summarization, in: Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2006, pp. 181–184.
[26] D.D. Wang, S.H. Zhu, T. Li, Y. Chi, Y.H. Gong, Integrating clustering and multi-document summarization to improve document understanding, in: Proceedings of ACM 17th Conference on Information and Knowledge Management, 2008, pp. 1435–1436.
[27] K.F. Wong, M.L. Wu, W.J. Li, Extractive summarization using supervised and semi-supervised learning, in: Proceedings of 22nd International Conference on Computational Linguistics, 2008, pp. 985–992.
[28] H. Zha, Generic summarization and key phrase extraction using mutual reinforcement principle and sentence clustering, in: Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval, 2002, pp. 113–120.
[29] S.R. Ye, T.S. Chua, M.Y. Kan, L. Qiu, Document concept lattice for text understanding and summarization, Information Processing and Management 43 (2007) 1643–1662.
[30] P. Sun, J.H. Lee, D.H. Kim, C.M. Ahn, Multi-document using weighted similarity between topic and clustering-based non-negative semantic feature, Lecture Notes in Computer Science (2007) 108–115.
[31] D.D. Wang, T. Li, S.H. Zhu, C. Ding, Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization, in: Proceedings of the 31st Annual International Conference on Research and Development in Information Retrieval, 2008, pp. 307–314.
[32] F.R. Wei, W.J. Li, Q. Lu, Y.X. He, Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization, in: Proceedings of the 31st Annual International Conference on Research and Development in Information Retrieval, 2008, pp. 283–290.
[33] W.Y. Liu, X.J. Quan, F. Min, Q. Bite, A short text modeling method combining semantic and statistical information, Information Sciences 180 (2010) 4031–4041.

Xiaoyan Cai is a research associate in department of computing, the Hong Kong Polytechnic University, Hong Kong. She received the PHD degree 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.