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, $S = (s_{ij})_{n\times n}$