ARTICLE IN PRESS

Information Sciences xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

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

Enhancing diversity and coverage of document summaries through subspace clustering and clustering-based optimization

Xiaoyan Cai [a,*], Wenjie Li [b], Renxian Zhang [c]

[a] College of Information Engineering, Northwest Agricultural and Forestry University, Shaanxi, China
[b] Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[c] Department of Computer Science and Technology, Tongji University, Shanghai, China

ARTICLE INFO

Article history:
Received 8 April 2013
Received in revised form 26 March 2014
Accepted 11 April 2014
Available online xxxx

Keywords:
Document summarization
Information diversity
Information coverage
Subspace clustering

ABSTRACT

Sentence clustering has been successfully applied in document summarization to discover the topics conveyed in a collection of documents. However, existing clustering-based summarization approaches are seldom targeted for both diversity and coverage of summaries, which are believed to be the two key issues to determine the quality of summaries. The focus of this work is to explore a systematic approach that allows diversity and coverage to be tackled within an integrated clustering-based summarization framework. Given the fact that normally each topic can be described by a set of keywords and the choice of the keywords among the topics is topic-dependent, we take the advantage of the newly emerged subspace clustering to enable the flexibility of keyword selection and the improved quality of sentence clustering. On this basis, we develop two clustering-based optimization strategies, namely local optimization and global optimization to pursue our targets. Experimental results on the DUC datasets demonstrate effectiveness and robustness of the proposed approach.

1. Introduction

With the rapid growth of the Internet and information explosion, automatic document summarization, which aims to condense the original text into its essential content and to assist in filtering and selection of necessary information has drawn increasing attention in the past years.

In this field, sentence ranking that determines the importance of each individual sentence has long been the emphasis of the existing researches, based on which the summaries can be generated by extracting the most salient sentences from the original documents. However, such generated summaries are not guaranteed to be the best summaries. When many competing sentences are available, given the summary length limit, the strategy of selecting good summaries rather than ranking sentences becomes important. A good summary is expected to be the one with extensive coverage of the focuses presented in documents or specified by users and with maximum diversity. Or to say, information diversity and information coverage of an entire summary, to a large degree, determine the quality of the summary. While information diversity helps to provide a summary that contains as much diversified information as possible, information coverage recommends a summary that contains many, if not every, important aspects in the documents. These two issues are what we are concerned in this study.

* Corresponding author. Tel.: +86 29 8849 1812.
E-mail addresses: xiaoyanc@mail.nwpu.edu.cn (X. Cai), cswjli@comp.polyu.edu.hk (W. Li), rxzhanggm@gmail.com (R. Zhang).

http://dx.doi.org/10.1016/j.ins.2014.04.028
0020-0255/

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 different places in the same document. Enforcing diversity in summarization can effectively reduce redundancy. This is normally achieved by clustering highly related sentences into topic themes. With the standard vector space representation, sentence clustering inevitably confronts two challenging problems. One is high dimensionality, i.e., there exist a large number of words in a given document collection. The other is high sparsity, i.e., compared with the words in a document, the words in a sentence is quite limited due to the shorter length of it. In a sparse and high dimensional space, the similarity between sentences is very high. Consequently, sentence clusters cannot be effectively detected by the traditional clustering algorithms.

If we take an in-depth look at the topic themes, it will not be difficult to observe that actually only a few number of keywords are required to describe a topic theme while most other words within the topic theme are irrelevant or at least not crucial. If we can cluster the sentences with some selected keywords that can better illustrate each topic theme and meanwhile maximize the difference among all the topic themes, the reasonably good sentence clusters can be expected. Newly emerged subspace clustering is just well-suited for such a scenario, where the correlated keywords vary with different sentence clusters. Another advantage of subspace clustering is that it allows the keywords to be selected freely for each topic theme, i.e., the keywords in the different topic themes can be overlapped. This property naturally reflects a very realistic situation that a word often has multiple senses and thus it should be allowed to be involved in describing multiple topic themes.

Besides information diversity, information coverage is also important in summarization. An existing solution is to identify a set of keywords and then select the sentences covering as many of the keywords as possible to form a summary. However, while the keywords are identified for the whole document set rather for a particular topic theme, this solution is not able to give consideration to both coverage and diversity and thus the selected sentences may focus on a few important topic themes only. The motivation of this paper is to explore a systematic approach that allows information diversity and information coverage of a summary to be tackled within an integrated clustering-based summarization framework. The main idea is to take advantage of the subspace clustering to select keywords for each topic theme and then to assign sentences to the corresponding topic themes. On this basis, two optimization strategies, namely local optimization strategy and global optimization strategy are proposed to select the sentences that contain as many keywords as possible and cover as many topic themes as possible.

The main contributions of the paper are three fold. (1) A subspace clustering algorithm is introduced to generate keywords and sentences of topic themes in the given document collection. (2) Two optimization strategies are developed for summary generation, namely local optimization strategy and global optimization strategy. The local optimization strategy maximizes the coverage of keywords within each topic theme, while the global optimization strategy simultaneously optimizes coverage and diversity. (3) Thorough experimental studies are conducted to verify the effectiveness and robustness of the proposed approach.

The rest of the paper is organized as follows. Section 2 reviews related work. Subspace sentence clustering and diversity-optimized summary generation are explained in Section 3 and Section 4, respectively. Section 5 presents the experimental results. The paper is concluded in Section 6.

2. Related work

Clustering has become an increasingly important topic with the explosion of information available via the Internet. It is an important tool in text mining and knowledge discovery. Recently, it has been successfully applied in clustering-based summarization.

In terms of the roles of clustering in summarization, one could take advantage of the clustering results to select the representative sentences in order to generate diverse summaries. The typical examples of such use are C-RR and C-LexRank proposed by Qazvinian and Radev [28], which selected the important citation sentences from the sentence clusters generated by a hierarchical agglomeration algorithm. Alternatively, the clustering results could be used to improve or refine the sentence ranking results. Most of the clustering-based summarization approaches are of this nature [34,37]. There were also some research works focusing on simultaneous sentence ranking and keyword extraction, such as [8,36,39]. The keywords in these approaches were selected for the whole document set and used to enhance sentence ranking. Our objective differs from theirs in that we would like to generate different sets of keywords for different topic themes and to cluster sentences based on them.

Subspace clustering [27] provides an effective way to achieve this purpose. Subspace clustering is an extension of traditional clustering that seeks to find clusters in different subspaces within a dataset. The subspace clustering algorithms localize the search for relevant dimensions allowing them to find clusters that exist in multiple, possibly overlapping subspaces. It is especially effective in domains where one can expect to find relationships across a variety of perspectives. Such technique has been successfully applied in text mining [14] and gene expression analysis [12,16]. Because of these advantages, we believe that it can be applied to generate more accurate sentence clusters, which in turn can further improve the performance of summarization.

There were also approaches that attempted to handle diversity by reducing information redundancy. Earlier proposed Maximum Marginal Relevance (MMR) method [5] used the full similarity matrix to choose the sentences that are the least

Please cite this article in press as: X. Cai et al., Enhancing diversity and coverage of document summaries through subspace clustering and clustering-based optimization, Inform. Sci. (2014), http://dx.doi.org/10.1016/j.ins.2014.04.028