Neurocomputing 157 (2015) 356–366

Contents lists available at ScienceDirect

Neurocomputing

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

Unsupervised document summarization from data reconstruction perspective

Zhanying He [a], Chun Chen [a,*], Jiajun Bu [a], Can Wang [a], Lijun Zhang [a], Deng Cai [b], Xiaofei He [b]

[a] Zhejiang Provincial Key Laboratory of Service Robot, College of Computer Science, Zhejiang University, Hangzhou 310027, China
[b] State Key Lab of CAD&CG, College of Computer Science, Zhejiang University, Hangzhou 310058, China

ARTICLE INFO

Article history:
Received 2 April 2014
Received in revised form
2 July 2014
Accepted 20 July 2014
Communicated by M. Wang
Available online 28 September 2014

Keywords:
Document summarization
Data reconstruction
Manifold adaptive kernel

ABSTRACT

Due to its wide applications in information retrieval, document summarization is attracting increasing attention in natural language processing. A large body of recent literature has implemented document summarization by extracting sentences that cover the main topics of a document with a minimum redundancy. In this paper, we take a different perspective from data reconstruction and propose a novel unsupervised framework named *Document Summarization based on Data Reconstruction* (DSDR). Specifically, our approach generates a summary which consist of those sentences that can best reconstruct the original document. To model the relationship among sentences, we firstly introduce the linear reconstruction which approximates the document by linear combinations of the selected sentences. We then extend it into the non-negative reconstruction which allows only additive, not subtractive, linear combinations. In order to handle the nonlinear cases and respect the geometrical structure of sentence space, we also extend the linear reconstruction in the manifold adaptive kernel space which incorporates the manifold structure by using graph Laplacian. Extensive experiments on summarization benchmark data sets demonstrate that our proposed framework outperform state of the art.

1. Introduction

With the explosion of the textual information on the World Wide Web, people are overwhelmed by innumerable accessible documents. This means that we are in great need for technologies like document summarization that can better help users digest the information on the Web. Summarization techniques address this problem by condensing the document into a short piece of text covering the main topics. For example, search engines can provide users with snippets as the previews of the document contents, and help them to find the desired document. News sites usually describe hot news topics in concise headlines to facilitate browsing all news. Both the snippets and headlines are specific forms of document summary in real applications. Especially in the micro-blogging services, such as Twitter, Weibo and Tumblr, a hot topic can yield millions of short massages including enormous noises and redundancies. The possible solution is to summarize the massive tweets into a set of short text pieces covering the main topics [1].

Document summarization can be categorized as abstractive summaries or extractive summaries. Given a document, the abstractive summary is generated from complex natural language processing like information fusion, sentence compression and reformulation. Obviously, it is a difficult task for computer to automatically generate a satisfactory summary by abstraction. So the common practice is to perform extractive summarization in which a subset of existing sentences is used to form a final summary. Most of the existing generic summarization approaches use a ranking model to select sentences from a candidate set [2–4]. But these methods suffer from the redundancy problem in that top ranked sentences usually share much information in common. Although there are some methods [5–7] trying to reduce the redundancy, selecting sentences which have both good information coverage and minimum redundancy is a non-trivial task.

The motivation of our work is that the traditional methods usually solve the document summarization as a natural language problem rather than a data reconstruction problem although the second has been explored greatly in the literature of machine learning such as dimension reduction and feature selection. So in this paper, we propose a novel unsupervised summarization framework from the perspective of data reconstruction. As far as we know, our work is the first to treat the document summarization as a data reconstruction problem. We argue that a good summary should consist of those sentences that can best reconstruct the original document. Therefore, the reconstruction error becomes a natural criterion for measu ring the quality of summary. The new framework, namely *Document*

* Corresponding author.
E-mail addresses: hezhanying@zju.edu.cn (Z. He), chenc@zju.edu.cn (C. Chen), bjj@zju.edu.cn (J. Bu), wcan@zju.edu.cn (C. Wang), zljzju@zju.edu.cn (L. Zhang), dengcai@cad.zju.edu.cn (D. Cai), xiaofeihe@cad.zju.edu.cn (X. He).

http://dx.doi.org/10.1016/j.neucom.2014.07.046
0925-2312/

Summarization based on Data Reconstruction (DSDR), finds the summary sentences by minimizing the reconstruction error. DSDR learns a reconstruction function for each candidate sentence of an input document and then formulates an objective function minimizing the error to obtain an optimal summary. The geometric interpretation is that DSDR tends to select sentences that span the intrinsic subspace of candidate sentence space, so that it is able to cover the core information of the document.

We firstly introduce the linear reconstruction to model the relationship between the document and the summary. The linear reconstruction aims to approximate the document by linear combinations of the selected summary sentences. Further, inspired by previous studies which indicate the existence of psychological and physiological evidence for parts-based representation in the human brain [8–10], we assume that document summary should consist of the parts of sentences, and introduce the non-negative constraints into the DSDR framework. With the non-negative constraints, our method leads to parts-based representation so that no redundant information needs to be subtracted from the combination. Still another issue to be addressed in document summary is the nonlinearity of the sentence space, as recent research [11] shows that the raw sentences are supposed to be highly nonlinear in distribution. The linear functions therefore lead to suboptimal fit in that neither the linear reconstruction nor the non-negative linear reconstruction respect the nonlinear manifold structure of sentence space. So we propose a novel nonlinear reconstruction which is performed in the manifold adaptive kernel space by using graph Laplacian [11–13]. By extracting sentences which can reconstruct the document in the kernel space, we are able to produce a better summary than the classical methods.

It is worthwhile to highlight the following three contributions of our proposed DSDR framework in this paper:

- We propose a novel unsupervised summarization framework from the perspective of data reconstruction which as we known is the first work to treat the document summarization from such a perspective.
- We firstly introduce the linear reconstruction and a greedy optimization method to solve the problem efficiently and effectively. Further, we propose the non-negative reconstruction and the corresponding iterative method to get a global optimum. To handle the nonlinearity, we finally propose the nonlinear reconstruction based on the manifold adaptive kernel.
- The proposed framework should not be restricted to the three types of reconstruction mentioned in this paper. Actually it is suitable for any other data reconstruction types. Since DSDR is unsupervised and language independent, it can be extended to summarize non-English document easily and even multi-language document.

This work is an extended and improved follow-up to our earlier work [14]. In comparison, we add a substantially theoretical analysis about extending DSDR in the manifold adaptive kernel space. For both linear reconstruction and non-negative linear reconstruction, the details of the mathematical translations are introduced additionally. We also extend the experiments here, such as implementing DSDR in the manifold adaptive kernel space and comparing it with existing approaches.

Our paper is organized as follows. We briefly review the related work in Section 2. In Section 3, we introduce the details of the *Document Summarization based on Data Reconstruction* (DSDR) including the optimization algorithms. Finally, we experimentally demonstrate the effectiveness of our proposed approaches in Section 4 and conclude in Section 5.

2. Related work

Recently, lots of extractive document summarization methods have been studied. Most of them involve assigning salient scores to sentences or paragraphs of the original document and composing the result summary of the top units with the highest scores. The computation rules of salient scores can be categorized into three groups [15]: feature based measurements, lexical chain based measurements and graph based measurements [4]. Salient scores in feature based measurements are usually related with various features such as term frequency, position, length, and topic presentation. The first method proposed in [16] ranks the sentences which are represented by the weighted term frequency vectors according to the relevance scores to the whole document. In the second type of measurements, a lexical chain is defined by a coherent sequence of related nouns, verbs and others. Sentence scores are then computed according to the lexical chain. In [17], the semantic relations of terms in the same semantic role are discovered by using the WordNet [18]. The relations are finally used in pairwise semantic similarity calculations which serve for the construction of their semantic similarity matrix. A tree pattern expression for extracting information from syntactically parsed text is proposed in [19]. In the graph based measurements, the sentence scores propagate around the graph on the basic idea that the score of one sentence affects scores of its neighbor sentences in the graph. Algorithms like PageRank [2] and HITS [3] are used in the sentence score propagation based on the graph constructed through the semantic affinity among sentences. In [4], it is also shown that this kind of measurements can improve single-document summarization by integrating multiple documents of the same topic.

Almost all the mentioned document summarization methods based on sentence scores have to incorporate with the adjustment of term weights which is one of the most important factors that influence summarization performance [20]. The adjustment process is used to eliminate the redundant information while it is not necessary when methods without saliency scores are applied in summarization. For extracting sentences, the methods without saliency scores include classification-based methods [21,22], clustering-based methods [23], as well as model-based methods [5–7]. Inspired by the latent semantic indexing (LSA), Ref. [16] applies the singular value decomposition (SVD) to select highly ranked sentences for generic document summarization. Besides, to improve summarization performance, there are some other studies like clustering sentences into topic themes, improving the topic representation and also time series text. Ref. [17] uses symmetric non-negative matrix factorization (SNMF) to cluster sentences into groups and selects sentences from each group for summarization. And [24] analyzes five different topic representations and proposes a novel topic representation based on topic themes. In [24], authors propose a novel symbolic representation of time series for text processing.

However, all the above summarization methods aim to obtain the summary which covers the core information, but few conduct the extractive task from the data reconstruction perspective. We believe that a good generic summary should contain those sentences that can best reconstruct the document. So how to best reconstruct the original document by the selected sentences is the main focus of the proposed DSDR in this study.

Notation: Small letters (*e.g.* x) denote scalars. Lowercase bold letters (*e.g.* $\mathbf{x}$) denote column vectors and $\|\cdot\|$ denotes the vector l_2-norm. Uppercase letters (*e.g.* X) denote matrices or graphs. The matrix trace is denoted by $\mathrm{Tr}(\cdot)$ and the Forbenius norm of a matrix is denoted by $\|\cdot\|_F$. Script uppercase letters (*e.g.* $\mathcal{X}$) denote ordinary sets and $|\mathcal{X}|$ is the size of the set. Blackboard bold capital letters (*e.g.* $\mathbb{R}$) denote number sets.