Information Sciences 349–350 (2016) 12–24

Contents lists available at ScienceDirect

Information Sciences

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

Multi-document summarization via group sparse learning

Ruifang He[a,*], Jiliang Tang[b], Pinghua Gong[b], Qinghua Hu[a,*], Bo Wang[a]

[a] School of Computer Science and Technology, Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, 300072 Tianjin, China

[b] School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, 85281 Tempe, AZ, USA

ARTICLE INFO

Article history:
Received 17 January 2015
Revised 7 February 2016
Accepted 16 February 2016
Available online 23 February 2016

Keywords:
Multi-document summarization
Group sparse learning
Nesterov's method
Accelerated projected gradient algorithm

ABSTRACT

Multi-document summarization (MDS) aims to capture the core information from a set of topic-specific documents. Most existing extractive methods evaluate sentences individually and select summary sentences one by one, which may ignore the hidden structure patterns among sentences and fail to keep less redundancy from the global perspective. We study this task from the perspective of compressive sensing, consider sentences in documents as a kind of signals, which are usually sparse or compressible in the sense that they have concise representation pattern expressed in the proper sentence basis and transform it to a group sparse representation issue. A novel multi-document **S**ummarization with **G**roup **S**parse learning (SGS) framework is proposed, which can maximally reconstruct the original documents via minimizing the approximation error and jointly select summary sentences with the learnt group structure information among sentences. The summary relatedness can be modeled by constraining the reconstruction models to be close to each other, and make multiple sentences share a common underlying structure to form the summary content. With this model, we take the global information into account in evaluating the importance of sentences and further reduce the redundancy. In order to solve this group sparse convex optimization problem for MDS, we also develop an efficient algorithm based on the Nesterov's method, which leads to much faster convergence rate than some traditional methods. Experimental results on DUC 2006 and TAC 2007 main task corpora show the effectiveness of our proposed framework. The relevant experiments are conducted to demonstrate the working mechanism of main components in the SGS framework.

1. Introduction

The explosive growth of Internet makes the information overload problem increasingly sever. People are flooded by a vast amount of accessible documents. MDS aims to produce a summary delivering the core information from documents and help online users acquire information efficiently. It has attracted increasing attention from various research fields such as information retrieval, natural language processing and machine learning, and can be used in various applications such as search engines [1,17], Q&A systems [32,37] and hand-held devices [5].

* Corresponding authors. Tel.: +86 15802296249, +86 15122108020.
E-mail addresses: he.ruifang@gmail.com (R. He), jiliang.tang@asu.edu (J. Tang), pinghua.gong@asu.edu (P. Gong), huqinghua@tju.edu.cn (Q. Hu), bo.wang.1979@gmail.com (B. Wang).

http://dx.doi.org/10.1016/j.ins.2016.02.032
0020-0255/

MDS is essentially a kind of information compression technique. Most of the existing generic extractive summarization methods can be roughly divided into two groups—unsupervised methods and supervised methods. Unsupervised methods often use the ranking models to select sentences from a candidate set, and include centroid-based method [30], language model based methods [15,18,29] and graph-based methods [12,26,43]. Supervised methods consider document summarization as a classification problem [41] or a sequence labeling problem [10,31]. However, most of these methods may suffer from a severe problem that top ranked sentences or labeled candidate summary sentences usually share much redundant information, and could not discriminate the salience and the redundancy simultaneously. Also the majority of these methods evaluate sentences individually and select sentences one by one, which may neglect the structure information among sentences and bring redundancy. It is still a daunting task for the extractive summarization methods to select sentences which have both high importance and minimum redundancy.

Compressive sensing indicates that many natural signals are sparse or compressible in the sense that they have concise representations when expressed in the proper basis [7], and provides a new perspective for MDS [2,7,11,34]. Even though there are many sentences in the document set, there may be a small number of them salient and informative. If we view sentences as a kind of signals, MDS can be considered to extract a subset of sentences that can best reconstruct the full set of sentences in the original documents. That is we transform MDS to a sparse representation problem. These intuitions motivate us to study the MDS problem from the compressive sensing perspective. In essence, we aim to jointly extract a small number of sentences which can maximally reconstruct the full set of sentences in the document set. Particularly, we investigate—(1) how to model the group sparse representation problem mathematically; and (2) how to jointly select summary sentences for MDS. In our attempt to answer these questions, we propose a novel compressive sensing based multi-document **S**ummarization framework via **G**roup **S**parse learning (SGS). Our contributions are summarized below:

- Formulate the compressive sensing problem for multi-document summarization into an optimization problem via group sparse learning;
- Provide a methodology to efficiently solve the optimization problem based on the accelerated projected gradient algorithm;
- Propose a novel compressive sensing based multi-document summarization framework which can select summary sentences in batch mode by considering group structure information among sentences; and
- Evaluate the proposed framework on two summarization benchmarks to understand the working mechanism of the proposed framework.

The rest of paper is organized as follows: Section 2 formulates how to model the sparse representation for MDS as a group sparse convex optimization problem and gives the proposed framework; Section 3 introduces how to solve the optimization issue with the accelerated projected gradient algorithm; Section 4 shows the experiments with details about data sets, metric, results and discussions. Section 5 briefly reviews the related work. Finally, we conclude the paper with future work in Section 6.

2. The proposed compressive sensing based multi-document summarization framework

Before going into the details, we give the notations used in this paper. Scalars are denoted by lower case letters, vectors by bold-faced lower case letters, and matrices by bold-faced capital letters. Let $\|\cdot\|_2$ and $\|\cdot\|_1$ denote the ℓ_2 norm and the ℓ_1 norm of a vector, respectively, and $\|\cdot\|_F$ and $\|\cdot\|_{2,1}$ represent the Frobenius norm and $\ell_{2,1}$-norm of a matrix, respectively. Given a set of documents, which are segmented into n sentences $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$. Let $\mathcal{D} = \{w_1, w_2, \ldots, w_d\}$ be the dictionary where d is the size of the dictionary. Then we can represent the n sentences $\mathcal{S}$ into a matrix $\mathbf{S} \in \mathbb{R}^{d\times n}$ where the ith column of $\mathbf{S}$ $\mathbf{S}_i \in \mathbb{R}^d$ denotes the ith sentence. $\mathbf{S}_{ij}$ is the weight of the jth word in the ith sentence, which is calculated as the TFIDF like weight. With the above notations and definitions, next we will give details about how to model the sparse representation problem for MDS, correspondingly answering the first question in the introduction part.

2.1. Modeling the group sparse representation problem

Sparse coding is to model data vectors as the sparse linear combinations of basis elements (also called dictionary), and widely used in machine learning, signal processing, image and video applications. Although sparse learning models based on the ℓ_1 norm such as the Lasso [33] have achieved great success in many applications, they do not take the existing feature structure into consideration. However, the features exhibit the natural group structure in some applications. For example, the problem includes multi-factors, each factor may have several levels and can be represented using a group of dummy variables [42]. The group Lasso based on the $\ell_{q,1}$-norm penalty can capture the group structure information and perform the selection of group structures. The group selection distinguishes the group Lasso from the Lasso which does not take group information into account and does not support group selection [40]. It well fits the essential requirements of MDS. In this paper, we explore MDS with group sparse learning to perform the content selection of summarization from the different perspectives.

MDS is to find the core information from documents. We consider sentences in documents as a kind of signal, which are sparse since only small amount of them are informative. The crucial observation for compressive sensing is that one can design efficient sensing or sampling protocols that capture useful information embedded in a sparse signal and condense it