Information Processing and Management 52 (2016) 698–719

Contents lists available at ScienceDirect

Information Processing and Management

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

Amplifying scientific paper's abstract by leveraging data-weighted reconstruction

Shansong Yang, Weiming Lu*, Zhanjiang Zhang, Baogang Wei, Wenjia An

College of Computer Science and Technology, Zhejiang University, Hangzhou, 310027,China

ARTICLE INFO

Article history:
Received 10 December 2014
Revised 21 December 2015
Accepted 25 December 2015
Available online 18 January 2016

Keywords:
Document summarization
Citation analysis
Scientific literature
Data-weighted reconstruction

ABSTRACT

In this paper, we focus on the problem of automatically generating *amplified scientific paper's abstract* which represents the most influential aspects of scientific paper. The influential aspects can be illustrated by the target scientific paper's abstract and citation sentences discussing the target paper, which are provided in papers citing the target paper. In this paper, we extract representative sentences through data-weighted reconstruction approach(DWR) by jointly leveraging target scientific paper's abstract and citation sentences' content and structure. In our study, we make two-folded contributions.

Firstly, sentence's weight was learned by exploiting regularization for ranking on heterogeneous bibliographic network. Specially, *Sentences-similar-Sentences* relationship was identified by language modeling-based approach and added to the bibliographic network. Secondly, a data-weighted reconstruction objective function is optimized to select the most representative sentences which reconstructs the original sentence set with minimum error. In this process, sentences' weight plays a critical role. Experimental evaluation over real dataset confirms the effectiveness of our approach.

1. Introduction

With the rapidly expanding scientific literature, identifying and digesting valuable knowledge is a challenging task. The explosive growth of the publications makes it rather difficult to quickly understand large amounts of scientific papers. Given a query, there are too many scientific papers for researcher to understand. The best approach to this problem is to select the most influential and representative papers which meanwhile are close to the researcher's research interests.

Although most researchers grasp a scientific paper's general outline through its abstract, the aspects described by abstract are frequently biased and incomplete. The abstract reflects the author's viewpoint about its key characteristic, which is subjective. Intuitively, the influential aspects or contributions of papers should be identified and evaluated by researchers in the same field, especially the authors who cited the target paper.

Based on the above analysis, in this paper, our goal is to generate *amplified scientific paper's abstract*, which can illustrate the most influential aspects of paper. In this paper, we achieve this goal through data-weighted reconstruction approach which consists of weight learning and salient sentence selection. Citation sentences' semantic information and social structure are taken into consideration in the process of sentences' weight learning.

* Corresponding author. Tel.: +86 57187953779; fax: +86 57187952300.
E-mail address: luwm@zju.edu.cn (W. Lu).

http://dx.doi.org/10.1016/j.ipm.2015.12.014
0306-4573/

Table 1
Notations used in this paper.

Symbol	Description
P_t	Target scientific paper
$\{P_c\}$	Scientific paper set citing P_t
$T = \{t_1, t_2, \ldots, t_K\}$	Sentence set from P_t's abstract
$C = \{c_1, c_2, \ldots, c_L\}$	Citation sentences from $\{P_c\}$
$X = T \cup C$	The whole sentences sets
$K = \|T\|$	The number of sentences in R
$L = \|C\|$	The number of sentences in C
$N = K + L$	The number of sentences in X
w	Sentence words
$G = \{G_i\}$	Semantic group relations
V	Vertices of hypergraph
E	Hyperedges of hypergraph
$\boldsymbol{H}$	Hypergraph's incidence matrix
$\boldsymbol{D}_e$	Diagonal matrix of hyperedge's degree
$\boldsymbol{D}_v$	Diagonal matrix of vertex's degree
$\boldsymbol{y} = [y_1, \ldots, y_{\|V\|}]^T$	The initial score of all vertices
$\boldsymbol{f}^* = [f_1, \ldots, f_{\|V\|}]^T$	The final ranking score of all vertices
$\boldsymbol{U}$	Diagonal matrix of all sentences's weight
$S = \{s_1, s_2, \ldots, s_m\}$	The summarization

In our study, the sentences of *amplified scientific paper's abstract* is extracted from the target scientific paper's abstract and citation sentences provided in papers which cite the target paper. So document summarization technique is a natural choice to work out this problem.

Inspired by document summarization based on data reconstruction (DSDR) (He et al., 2012), which selects a subset of sentences to best reconstruct the original document, we optimize a data-weighted reconstruction objective function for salient sentence selection. 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. The drawback of DSDR is that it treats all sentences equally important, which violates realistic cases obviously. Some sentences appear as the decorated or transitionary role in document, especially in scientific literature, the citation sentences are inherently informal, noisy and not well structured. Many citation sentences may contain information irrelevant to the target scientific paper. For these noisy sentences, we should not reconstruct them or reconstruct them at a little cost.

Especially to generate *amplified scientific paper's abstract*, each citation sentence's authority is another consideration other than the summarization's coverage. In other words, citation sentences from influential papers are probably more important than others. Intuitively, the reconstruction of important sentences should be assigned to high priority.

Based on the above analysis, we make such assumption:

- Different sentences should be reconstructed with different priority.

So this paper proposes a data-weighted reconstruction approach(DWR) to generate *amplified scientific paper's abstract*, which is designed to take the assumption into consideration. DWR first learns sentence's weight and then selects salient sentences from data-weighted reconstruction perspective.

For data-weighted reconstruction objective function, sentences' weight need to be first learned. In our study, sentences' weight are learned based two factors: sentence's semantic information and scientific paper's social structure (Table 1) .

Those two factors are embodied in the heterogeneous bibliographic network. In the field of scientific literature, there are various kinds of social media information, including different types of objects and relations among these objects. For example, a typical bibliographic information network contains objects in four types of entities: *paper*(P), *venue*(i.e., conference or journal)(V), *author*(A), and *sentence*(S). For each paper, it has links to a set of authors, a venue, and a set of sentences, belonging to a set of link types. Intuitively these abundant heterogeneous information should be leveraged to measure sentence's importance. In the process of evaluating sentence's importance, we suppose close ranking scores should be assigned to similar sentences. Specifically, hypergraph is utilized to model the various objects and relations, and then regularization cost function is introduced to evaluate sentence's importance considering the social media information and sentence's semantic similarity.

After sentence's weight learning process, the data-weighted reconstruction objective function can be solved to select salient sentences.

The contributions of this paper are summarized as follows: Firstly, we exploit semi-supervised PLSA and regularization for ranking on heterogeneous bibliographic network to compute sentence's weight. Secondly, a data-weighted reconstruction objective function is proposed to generate *amplified scientific paper's abstract*.