Future Generation Computer Systems 32 (2014) 246–252

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Summarization of scientific documents by detecting common facts in citations

Jingqiang Chen, Hai Zhuge *

Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, 100190, Beijing, China
Nanjing University of Posts and Telecommunications, 210003, Nanjing, China
Aston University, Birmingham, B4 7ET, UK

HIGHLIGHTS

- A summarization system that expands citations through common fact.
- Explore common fact phenomenon in the scientific literature.
- An approach to expand terms to associated terms.

ARTICLE INFO

Article history:
Received 27 March 2013
Received in revised form
18 July 2013
Accepted 31 July 2013
Available online 22 August 2013

Keywords:
Summarization
Semantic link network
Natural language processing

ABSTRACT

Reading scientific articles is more time-consuming than reading news because readers need to search and read many citations. This paper proposes a citation guided method for summarizing multiple scientific papers. A phenomenon we can observe is that citation sentences in one paragraph or section usually talk about a common fact, which is usually represented as a set of noun phrases co-occurring in citation texts and it is usually discussed from different aspects. We design a multi-document summarization system based on common fact detection. One challenge is that citations may not use the same terms to refer to a common fact. We thus use term association discovering algorithm to expand terms based on a large set of scientific article abstracts. Then, citations can be clustered based on common facts. The common fact is used as a salient term set to get relevant sentences from the corresponding cited articles to form a summary. Experiments show that our method outperforms three baseline methods by *ROUGE* metric.

1. Introduction

Researchers have to read many papers relevant to their research, especially for new comers of areas. Reading one paper, a reader needs to search and read a big number of cited papers.

A shortcut is to develop a system that automatically retrieves the most relevant content from the cited papers to complement the citations.

Previous studies either exploit single citation [1] or summarize the articles co-cited in one citation sentence [2]. We make a further progress by exploiting the information contained in multiple citations that co-occur in one paragraph or section.

Each citation may talk about some facts. For example in [3], "*It is generally agreed upon that manually written abstracts are good summaries of individual papers. More recently, Qazvinian and Radev (2008) argue that citation texts are useful in creating a summary of the important contributions of a research paper. The citation text of a target paper is the set of sentences in other technical papers that explicitly refer to it (Elkiss et al., 2008a). However, Teufel (2005) argues that using citation text directly is not suitable for document summarization.*" has three citation sentences, all of which mention the noun phrase of '*citation text*'. This can be regarded as the common fact of the three citation sentences. The three citations and their corresponding cited papers discuss the common fact from different aspects. We can use it to create a summary by retrieving relevant sentences from the cited papers to enrich the citation paragraph.

However, sometimes citation sentences may not use the same words explicitly to refer to a fact. Two citation sentences may not share terms with each other but they do talk about the same fact. In the following citation paragraph [4]: "*In (Elmacioglu and Lee, 2005), it was shown that the DBLP network resembles a small world network due to the presence of a high number of clusters with a small average distance between any two authors. This average distance is compared to (Milgram, 1967)'s six degrees of separation' experiments, resulting in the DBLP measure of average distance between two authors stabilizing at approximately six*", there are two citation

[*] Corresponding author at: Nanjing University of Posts and Telecommunications, 210003, Nanjing, China.
E-mail address: zhuge@ict.ac.cn (H. Zhuge).

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.07.018

sentences, the first one talks about *small world network*, and the second talks about *six degrees of separation* phenomenon. They are totally different terms. However, the two citations are actually about the same fact. The noun phrase *small world network* is highly associated with the other noun phrase *six degrees of separation* in the domain of *complex network* as they co-occur frequently.

Therefore, it is necessary to expand the terms in a citation first to find the common fact. One way to expand a term is to find the ones that co-occur frequently with it according to corpus. After term expanding, the terms that co-occur in the citations are taken as the common fact. Here we can simply regard common fact as a set of terms with weights. The weight associated with a term reflects its significance in the common fact. Sometimes we need to cluster the citation sentences first for that despite they are in the same paragraph or section there may be different common facts existing in different subsets of the citations. Such common fact is then used as a query to get relevant sentences from the cited papers to form a summary.

Our main contribution is to exploit the common fact phenomenon to design a multi-document summarization system called *CFDSumm* to expand citations in an article. The first key technique is to expand terms in citations. We construct a term co-occurrence base based on 18 514 scientific abstracts in the domain of *Computational Linguistics*. The second key technique is to detect common facts in citations. We find out the common facts in the citations first and then cluster the citations based on the common facts. The third key technique is to find a subset of the most relevant sentences and form a summary. We treat common fact as a saliency term set where each member term is weighted and is used to score sentences. To evaluate our method, we create gold standard summaries by collecting 13 citation paragraphs from 11 papers manually. Experiments show that our method outperforms three baseline methods *MEAD* [5], *SciSumm* [2] and *CSIBS* [1] by *ROUGE*. The improvement significance is at p-value <0.05 and p-value <0.01.

2. Related work

Lots of work have been done on document summarization [6–14]. Analyzing citations in scientific articles is a feasible approach to summarize documents. Although there are many studies on treating citations as guidance for summarization, little work exploits common facts in citations.

The system *CSIBS* uses nouns in one citation as query to get more information from Ref. [1]. It is based on single citations rather than multiple citations. Another similar system *SciSumm* [2] aims to summarize documents co-cited within the same citation using surrounding text as query. *SciSumm* first applies *TextTiling* [15] technique to split the text into tiles which are then clustered. The clusters most relevant to the query are extracted to form a summary. Our work differs from *SciSumm* in that we work with multiple citations which may talk about the same fact and use such information to do summarization.

In recent years, citations are also used to create summary directly, called citation-based summarization [16,17,3,18–20]. Elkiss pointed out that citation summaries contain information that does not existing in abstracts and main contexts [16]. Hence, Qazvinian used citation sentences to create scientific paper summaries through citation summary networks and apply keyphrase extraction techniques to extract the key information contained in each citation which are then used to get high-quality citation-based summaries [19]. Meanwhile, Mohammad and Dorr think that citations can also be used to generate surveys of the scientific literature [3]. Our work is a reverse process of citation-based summarization.

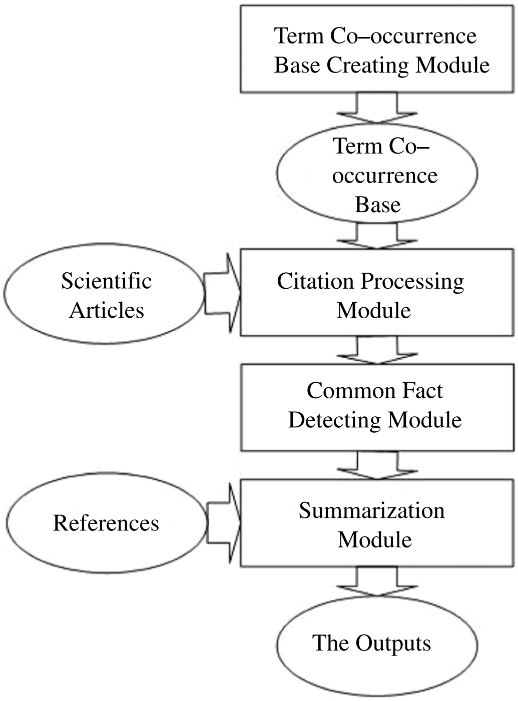
Term Co–occurrence Base Creating Module
Term Co–occurrence Base
Scientific Articles
Citation Processing Module
Common Fact Detecting Module
References
Summarization Module
The Outputs

Fig. 1. The architecture of CFDSumm System.

Analyzing the function of citations of the scientific literature has also received a lot of attention in the past years. *Argumentative Zone Theory* is a rhetorical classification task that classify functions of citations into seven categories (Aim, Textual, Own, Background, Contrast, Basis and Other) [21,22]. Teufel pointed out that particularly important labels of Aim, Contrast and Basis are more suitable for creating extractive summaries [21]. In contrast, Nanba and Okumura classified citation functions into three main categories [23], i.e. type B, type C and type O. Here type B refers to references based on other researchers' theories or methods, type C refers to references that compare with related works or to point out their problems, and type O refers to references other than types B and C. They find that bibliographic coupling using citation type C is more accurate and efficient to classify scientific articles, which are then used to create review articles automatically.

A methodology for semantic linking through spaces for cyber-physical society was proposed [24]. A document is created by writers using language units according to semantic images in mental space. A probabilistic resource space model is proposed to manage resources in the cyber-physical society [25]. A text scanning mechanism simulating human reading process was proposed in [26].

3. The system

The architecture of the system is shown in Fig. 1, which is composed of the following four modules:

(1) The Term Co-occurrence Base Creating Module takes scientific abstracts, titles or even conclusions of a domain as input, preprocesses these resources and computes the frequently co-occurring terms to create the term co-occurrence base.
(2) The Citation Processing Module takes a scientific article as input, extracts citation sentences, parses the citation sentences