Information Processing and Management 51 (2015) 1–24

Contents lists available at ScienceDirect

Information Processing and Management

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

Summarization based on bi-directional citation analysis

Filippo Galgani *, Paul Compton, Achim Hoffmann

School of Computer Science and Engineering, The University of New South Wales, Australia

A R T I C L E   I N F O

Article history:
Received 20 January 2014
Received in revised form 3 August 2014
Accepted 18 August 2014
Available online 17 September 2014

Keywords:
Natural language
Information extraction
Citations in information extraction
Summarization
Summarising legal documents

A B S T R A C T

Automatic document summarization using citations is based on summarizing what others explicitly say about the document, by extracting a summary from text around the citations (citances). While this technique works quite well for summarizing the impact of scientific articles, other genres of documents as well as other types of summaries require different approaches. In this paper, we introduce a new family of methods that we developed for legal documents summarization to generate catchphrases for legal cases (where catchphrases are a form of legal summary). Our methods use both incoming and outgoing citations, and we show how citances can be combined with other elements of cited and citing documents, including the full text of the target document, and catchphrases of cited and citing cases. On a legal summarization corpus, our methods outperform competitive baselines. The combination of full text sentences and catchphrases from cited and citing cases is particularly successful. We also apply and evaluate the methods on scientific paper summarization, where they perform at the level of state-of-the-art techniques. Our family of citation-based summarization methods is powerful and flexible enough to target successfully a range of different domains and summarization tasks.

1. Introduction

Citations have long been used to characterize and obtain information on different types of documents: for example for scientific papers, the number of citations received has been used to establish the impact of a paper (Cole, 2000; Garfield & Merton, 1979); for web pages the number and quality of incoming links contributes to ranking in information retrieval systems (Brin & Page, 1998). Internet technology makes it possible to track citations bi-directionally in a network, and links can be given an associated meaning. For web pages, anchor text is used to ascertain the content of a page. More recently citation text has been used also for summarization of scientific articles: the set of citations to a target article – sentences about the cited paper, called *citances* (Nakov, Schwartz, & Hearst, 2004) – are used to ascertain its important contributions, and to form a summary (Divoli, Nakov, & Hearst, 2012; Elkiss et al., 2008; Mei & Zhai, 2008; Mohammad et al., 2009; Qazvinian & Radev, 2008; Qazvinian et al., 2013).

While in the last decade several studies have explored citations for summarization of scientific articles, we suggest that they suffer a limitation in that they extract a summary only from the set of citing sentences. We propose that the citances (or anchor text) should not be the only text used as an indication of the content of the target document, rather the citances, full text, and where available the summaries of the citing documents should all be taken into account to build a global view of the target paper, and thus a better summary. In this paper we introduce a new family of methods that use both incoming and

* Corresponding author.
E-mail addresses: galganif@cse.unsw.edu.au (F. Galgani), compton@cse.unsw.edu.au (P. Compton), achim@cse.unsw.edu.au (A. Hoffmann).

http://dx.doi.org/10.1016/j.ipm.2014.08.001

outgoing citations, and combine citances with other elements of cited and citing documents, including the full text of the target document.

We apply our approach to a particular summarization problem: creating catchphrases for legal case reports. The field of law is one where automatic summarization can greatly enhance access to legal repositories, as legal professionals are confronted by large bodies of documents that must be scrutinized with great precision (Moens, 2007). Legal cases, rather than summaries, often contain a list of what the legal profession call catchphrases. Catchphrases present the important legal points of the case, their function is to "*give a summary classification of the matters dealt with in a case. [...] Their purpose is to tell the researcher whether there is likely to be anything in the case relevant to their research topic*" (Olsson, 1999). Like discursive summaries, catchphrases give a quick impression of what the case is about, but they have an indicative function rather than informative: they indicate all the legal points considered instead of just summarizing the key point(s) of a decision.

Despite the large quantity of material stored in textual format, and the strong need for intelligent text processing, little has been done for automatic summarization of legal documents. Given the importance of citations in the legal system, it is also surprising that citation analysis has not been incorporated in previous attempts at summarization of case reports. In the work describe here, we use the network of citations (Zhang & Koppaka, 2007) to assign catchphrases using both incoming and outgoing citations, and considering not only citation sentences but also citation catchphrases and full text sentences. Our methods substantially improve performance over full-text-only methods. We also show that this approach is suitable for scientific-article summarization, so the one method can be used for both.

The paper is organized as follows: related work on citation summarization is presented in Section 2, legal catchphrases are described in Section 3, and our evaluation framework is presented in Section 4. We first outline and evaluate a range of full-text-only baseline techniques, developed to identify important sentences in legal cases (Section 5). Traditional frequency-based approaches, popular in other domains such as news, perform quite poorly for legal cases. We also introduce a centrality-based approach using HITS (Kleinberg, 1999). We then move to citation-based methods and show how these substantially increase performance. We introduce a family of methods which use citation sentences, summaries of related papers and full text, both for incoming and outgoing citations (Section 6). In Section 7 we compare all the techniques and show how the citation-based methods obtain higher performance when compared to methods based only on the full text of documents. We then apply our methods to the domain of scientific articles in Section 8, and show how we obtain results comparable to state-of-the-art tools.

2. Related work

The use of citations for summarization has been mainly applied to scientific articles. Bradshaw (2003) proposed Reference-Directed Indexing (RDI), in which citations are used to determine the content of articles for information retrieval. Nakov et al. (2004) introduced the term "*citances*", defined as the sentences surrounding citations, and pointed out the possibility of using citances directly for text summarization, as they provide information on the important facts contained in a paper.

Elkiss et al. (2008) provided a quantitative analysis of the advantages of using citation contexts in applications such as summarization and information retrieval. In particular, they examined the relationship between the abstract and the citation contexts for a given scientific paper. Their experiments showed that citation contexts tend to have further focused information that is not present in the abstract, therefore citation contexts can be utilized as a different kind of summary, supplementary to the abstract. These results are also supported by Divoli et al. (2012), who compared citances to abstracts of biomedical papers, finding that the set of citances for a target article covers all information found in its abstract, and provides about 20% more concepts (provided that the article has already accumulated enough citations).

A first method for citation summarization was given by Mei and Zhai (2008), who proposed the use of citation contexts to produce an impact-based summary of a single research paper, using language modeling methods; i.e. the set of citing sentences is used to understand the impact, and sentences that talk about that impact are searched in the original paper and extracted.

Qazvinian and Radev (2008) presented an application of citation analysis to summarization, where they sought to extract, from among the set of sentences that constitute the citation summary, a subset that gives the main contributions of that paper. The proposed method, C-LexRank, first clusters citing sentences using Tf-idf vector similarity to find the different contributions of the target paper. Then LexRank (Erkan & Radev, 2004) is used to extract sentences from each cluster. Follow-up studies explored different ways of selecting sentences from the citation summaries: Qazvinian, Radev, and Ozgur (2010) identified important keyphrases from the set of citation sentences (statistically significant N-grams extracted using point-wise Kullback–Leibler divergence), then a greedy algorithm picked sentences that covered more (non-redundant) nuggets. Mei, Guo, and Radev (2010) introduced DivRank, based on a reinforced random walk in an information network. This ranking method focused on balancing the prestige and the diversity of the top ranked vertices. Qazvinian and Radev (2011) tried to maximize diversity using an approach based on distributional similarity. They applied this idea not only to scientific citations but also to a set of news articles. The work of Abu-Jbara and Radev (2011) aimed at producing more readable and cohesive summaries. Mohammad et al. (2009) focused on multi-document summarization, to obtain a survey on a given topic, given a collection of research papers. The authors compared summaries obtained from the full text of the papers, the abstract of the papers, and the citances of the papers. Qazvinian et al. (2013) compared different methods for creating technical summaries from citations, both for single-document summarization (using C-LexRank) as well as multi-document surveys of scientific paradigms.