Information Processing and Management 50 (2014) 443–461

Contents lists available at ScienceDirect

Information Processing and Management

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

Text summarization using Wikipedia

Yogesh Sankarasubramaniam [a], Krishnan Ramanathan [a,*], Subhankar Ghosh [b,1]

[a] HP Labs India, Bangalore, India
[b] SAS Institute, San Diego, CA, United States

ARTICLE INFO

Article history:
Received 18 January 2013
Received in revised form 20 January 2014
Accepted 4 February 2014
Available online 6 March 2014

Keywords:
Summarization
Wikipedia
Sentence ranking
Personalization

ABSTRACT

Automatic text summarization has been an active field of research for many years. Several approaches have been proposed, ranging from simple position and word-frequency methods, to learning and graph based algorithms. The advent of human-generated knowledge bases like Wikipedia offer a further possibility in text summarization – they can be used to understand the input text in terms of salient concepts from the knowledge base. In this paper, we study a novel approach that leverages Wikipedia in conjunction with graph-based ranking. Our approach is to first construct a bipartite sentence–concept graph, and then rank the input sentences using iterative updates on this graph. We consider several models for the bipartite graph, and derive convergence properties under each model. Then, we take up personalized and query-focused summarization, where the sentence ranks additionally depend on user interests and queries, respectively. Finally, we present a Wikipedia-based multi-document summarization algorithm. An important feature of the proposed algorithms is that they enable real-time *incremental summarization* – users can first view an initial summary, and then request additional content if interested. We evaluate the performance of our proposed summarizer using the ROUGE metric, and the results show that leveraging Wikipedia can significantly improve summary quality. We also present results from a user study, which suggests that using incremental summarization can help in better understanding news articles.

1. Introduction

Text summarization has seen renewed interest recently. The reason for this is twofold: first, summarization can help cope with the information overload, and second, small form-factor devices are becoming increasingly popular. Internet access on the move often happens in an attention-deficit situation where the user is capable of assimilating lesser content. Hence, it becomes important to present only the most relevant information. For example, while reading a news article, the user might first want to look at a short summary (about 50–100 words), and then request the full article if interested.

Radev, Hovy, and McKeown (2002) define a document summary as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that". In other words, a good summary: (1) is short and (2) preserves important information. There is a rich history of literature, dating back to the 1950s, that aims to achieve these two objectives. While the earliest efforts were restricted to simple position and word-frequency methods (perhaps due to the limited computation available at the time), more recent work

* Corresponding author. Tel.: +91 80 33829145; fax: +91 80 26129434.
E-mail addresses: yogesh@hp.com (Y. Sankarasubramaniam), krishnan_ramanathan@hp.com (K. Ramanathan), subhankar.ghosh@gmail.com (S. Ghosh).
[1] Work done while the author was with HP Labs India.

http://dx.doi.org/10.1016/j.ipm.2014.02.001
0306-4573/

has leveraged learning and graph-based algorithms for generating better summaries. This paper aims to study a further possibility: utilizing a human-generated knowledge base, like Wikipedia, in conjunction with graph-based summarization.

Wikipedia is perhaps the best example of collaborative knowledge creation and sharing, whereby the entire information is readily available to anyone–anywhere–anytime. The fact that it contains topics and entities of interest to humans makes it especially useful for summarization tasks. Moreover, Wikipedia is constantly updated, and provides the topic quality required for generating good summaries. Thus, Wikipdia can serve as the basis for understanding salient concepts from the input text, which can then be used to extract summary sentences.

The main contributions of this work are: (1) We cast the Wikipedia-based summarization problem into a general sentence–concept bipartite framework, and propose an iterative ranking algorithm for selecting summary sentences. (2) We provide precise mathematical definitions and analysis of the iterative ranking algorithm, and derive several convergence results. (3) We study generalizations of the basic bipartite setup, including directed/weighted edges, personalization and query focusing of summaries, and extensions to multi-document summarization. Furthermore, our algorithms provide *incremental summarization* in real-time, so that users can first view an initial summary, and if interested, can then request additional content. We also provide novel connections of our proposed algorithms with latent topic models and other known optimization problems.

The rest of this paper is organized as follows. We first review related summarization literature in Section 2, and state our research objective in Section 3. Then, in Section 4, we present the Wikipedia-based single document summarizer which is based on a novel iterative sentence–concept ranking. We provide precise mathematical definitions and convergence analysis under different scenarios, starting with undirected binary sentence–concept mappings in Section 4.2, followed by a generalization to weighted and directed mappings in Section 4.4, and then personalized and query-focused summarization in Section 4.5. Next, the Wikipedia-based multi-document summarizer is presented in Section 5. Experiments and performance evaluation are presented in Section 6, and the paper is concluded in Section 7.

2. Related work

A few recent efforts have leveraged Wikipedia for summarization tasks.[2] This includes our early work Ramanathan, Sankarasubramaniam, Mathur, and Gupta (2009), where summary sentences are selected based on Wikipedia concept frequency thresholds; the work of Ye, Chua, and Lu (2009) on summarizing definitions from Wikipedia pages; Wikipedia-based feature selection approaches (Bawakid & Oussalah, 2010; Gong, Qu, & Tian, 2010); Wikipedia-based sentence similarity approaches Miao and Li (2010); and the recent work of (Pourvali & Abadeh, 2012a, 2012b) which leverage Wikipedia to form multiple independent graphs, and then use graph importance and lexical cohesion features for summarization. However, one shortcoming of the above approaches is that they do not leverage graph-based ranking algorithms, which can capture the inter-sentence dependence and also utilize the entire graph structure for ranking sentences.

On the other hand, there exist several graph-based ranking algorithms that have independently been proposed for summarization tasks. Prominent among these are LexRank proposed by Erkan and Radev (2004), the iSpreadRank of Yeh, Ke, and Yang (2008), and the work of (Mihalcea et al., 2004; Mihalcea & Tarau, 2005). A particularly interesting approach is that of (Mihalcea & Tarau, 2005), who seek to rank sentences using well-known algorithms like HITS (Kleinberg, 1999) or PageRank Brin and Page (1998). Their evaluations show that graph-based methods can outperform baseline summarizers, and are in fact, competitive with the best supervised summarization algorithms. However, the document graph in their case is formed by connecting sentences based on word overlap, and they do not leverage knowledge bases like Wikipedia.

Thus, we see that there are two parallel lines of work: one that leverages Wikipedia, but does not utilize graph-based ranking algorithms; and the other that uses graph-based ranking algorithms, but without leveraging a knowledge base like Wikipedia. In this paper, we aim to bridge this gap by considering graph-based summarization in conjunction with Wikipedia. Moreover, our proposed ranking algorithms run directly on a bipartite graph, which is a departure from prior work which typically use a document graph. We also present several novel convergence results, and evaluations from experiments and user studies.

3. Research objective

Our aim in this work is to consider graph-based summarization in conjunction with Wikipedia. The specific research objectives are fourfold: (1) to present a unified sentence–concept bipartite framework for Wikipedia-based summarization, and propose a novel iterative ranking algorithm for summarization within this framework; (2) to provide mathematical formulation and analysis of the iterative ranking algorithm, and derive convergence properties; (3) to generalize the above framework and analysis to include directed/weighted edges, personalization and query-focusing of summaries, and multi-document summarization; and (4) to evaluate the performance of the proposed algorithms, experimentally using the ROUGE metric, and also directly by involving the user.

This paper also introduces the *incremental summarization* property (see Definitions 1 and 2 in Section 4), which is key to providing additional summary content in real-time. Results from our user study (see Section 6.4) suggest that this feature

[2] There is a long history of summarization literature, and the aim of this discussion is not to provide an exhaustive survey. We only seek to highlight the key ideas that help motivate our present work. The interested reader is referred to Das and Martins (2007) and Mani and Maybury (1999) for literature surveys.