Information Processing and Management 52 (2016) 670–681

Contents lists available at ScienceDirect

Information Processing and Management

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

Query-focused multi-document summarization using hypergraph-based ranking

Shufeng Xiong[a,b], Donghong Ji[a,*]

[a] Computer School, Wuhan University, 430072 Wuhan, China
[b] PingDingShan University, 467099 PingDingShan, China

ARTICLE INFO

Article history:
Received 13 January 2015
Revised 29 November 2015
Accepted 22 December 2015
Available online 19 January 2016

Keywords:
Multi-document summarization
Hypergraph-based ranking
HDP

ABSTRACT

General graph random walk has been successfully applied in multi-document summarization, but it has some limitations to process documents by this way. In this paper, we propose a novel hypergraph based vertex-reinforced random walk framework for multi-document summarization. The framework first exploits the *Hierarchical Dirichlet Process* (HDP) topic model to learn a word-topic probability distribution in sentences. Then the hypergraph is used to capture both cluster relationship based on the word-topic probability distribution and pairwise similarity among sentences. Finally, a time-variant random walk algorithm for hypergraphs is developed to rank sentences which ensures sentence diversity by vertex-reinforcement in summaries. Experimental results on the public available dataset demonstrate the effectiveness of our framework.

1. Introduction

The task of query-focused multi-document summarization is to create a summary for a document set, which aims to provide an answer for a given query. Since the task has been initiated in DUC (Document Understanding Conferences), it has attracted more and more attention.

The main method for query-focused multi-document summarization is based on sentence selection, and the selected sentences both summarize the documents and answer the query. In general, there are three steps to select sentences. First, the document is divided into sentences. Second, an abstractive or extractive summarizer is used to get some representative sentences for the query. The final step is to select the most informative ones as a summary. In this paper, we concentrate on extractive multi-document summarization.

For extractive summarization, the main strategies fall into several categories. Feature-based approaches use features like term frequency, sentence position, length, etc., to rank sentences (Ouyang, Li, Lu, & Zhang, 2010). Graph-based approaches rank sentences by a random walk which explores the relations between sentences. The main graph-based ranking approaches used in summarization include LexRank (Erkan & Radev, 2004b), Manifold-Ranking (Wan, Yang, & Xiao, 2007), Hyperlink-Induced Topic Search (HITS) (Kleinberg, Kumar, Raghavan, Rajagopalan, & Tomkins, 1999) and DivRank (Mei, Guo, & Radev, 2010).

For graph-based approaches, it has been demonstrated that a cluster-based method can effectively improve the quality of extractive summarization (Cai & Li, 2012; Li & Li, 2012; Wan & Yang, 2008; Zhang, Ge, & He, 2012). The method generally

* Corresponding author at: Computer School, Wuhan University, 430072 Wuhan, China.
E-mail addresses: xsf@whu.edu.cn, pdsujnow@163.com (S. Xiong), dhji@whu.edu.cn (D. Ji).

http://dx.doi.org/10.1016/j.ipm.2015.12.012
0306-4573/

focuses on sentence-level or cluster-level similarity, as well as the cluster-level relationship of sentences. For example, Wang, Li, Li, Li, and Wei (2013) and Wang, Wei, Li, and Li (2009) used hypergraphs to model both pairwise similarity and sentence clusters simultaneously, and they employed a hypergraph-based significance score propagation process to rank sentences. A good query-focused summary is expected to meet two factors: (1) high query relevance, (2) high diversity or minimum redundancy. High query relevance indicates the summary accounts for the given query. High diversity or minimum redundancy indicates the summary is able to present information without any convoluted. For these two aspects, existing graph-based methods have two limitations: (1) they cluster sentences simply based on co-occurrence lexical similarity, it may put semantically similar sentences into different topics, if they share few common words. For example:

- S1: A forest is a large area where trees grow close together.
- S2: Woodland is land with a lot of trees.

Since S1 and S2 share very few words, the methods based on "word co-occurrence" may result in a decision that S1 and S2 are discussing different topics. However, it is clear for a human interpreter that S1 and S2 are both talking about "forest". (2) The highest ranked sentences may be those ones with higher similarity to each other. In other words, it must provide an extra algorithm to remove redundancy.

In order to address the above limitations, we attempt to integrate probabilistic topic cluster and lexical similarity of sentences and develop a sentence ranking approach for achieving both diversity and centrality on the hypergraph model. Recently, Yin, Pei, Zhang, and Huang (2012) and Hennig and Labor (2009) exploited Probabilistic Latent Semantic Analysis (PLSA) to represent documents as a mixture of topics and clustered sentences by their topic distribution. Mei et al. (2010) proposed a reinforced random walk algorithm in an information network. Inspired by these work, we propose a Hypergraph-based vertex-rEinforced Ranking Framework (denoted as HERF) for query-focused summarization. First, we get a word-topic distribution by the HDP model and then cluster sentences by a hybrid clustering method in which it measures similarity based on word-topic distribution. Second, we build a hypergraph to capture both topical cluster relationship and pairwise cosine similarity of sentences. Then a sentence ranking approach, which balances the centrality and the diversity of the sentences by a vertex-reinforced strategy, is developed for scoring sentences. In practice, enforcing diversity in summarization can effectively reduce redundancy among the sentences. i.e., two sentences providing similar information should not be both present in the summary. Finally, a naive sentence selection and redundancy removal strategy is used to generate a summary. The experiments showed that our HERF framework performs better than other baseline summarizers on widely used benchmark datasets.

Two basic issues addressed in this paper are: (1) how to cluster semantically similar sentences into one topic even if they share few common words and (2) how to integrate redundancy removing policy into a hypergraph-based sentence ranking algorithm. Then, the main contributions of our work are summarized as follows.

- We proposed the HERF framework in which an adaptive vertex reinforcement random walk process is used to model the query similarity, the centrality and the diversity of sentences in hypergraph based model.
- We propose a hybrid method to construct a hypergraph to integrate topic distribution and word co-occurrence of sentences.
- To verify the effectiveness of our framework, we implement and evaluate our proposed framework HERF over widely used benchmark datasets, empirically verifying improvement over similar methods and systems.

The remainder of this paper is organized as follows. In Section 2, we describe our proposed summarization framework and details of constructing hypergraph, ranking sentences and generating summary. Section 3 gives the experiments and results. Section 4 briefly reviews the related work on graph/hypergraph based summarization. Finally, Section 5 concludes the paper.

2. Our summarization framework

In this section, we discuss our summarization framework which consists of four crucial components, as shown in Fig. 1. In HERF, we first cluster sentences using a HDP-based approach based on sentence topic similarity. The topic similarity is calculated by the transformed radius (TR) based on the KL divergence, and KL divergence is based on word-topic probability distribution learned by the HDP topic model. We then construct a hypergraph based on sentence clusters as well as pairwise relationship between sentences. Then, we score sentences with a vertex-reinforced ranking approach which considers both the topic sensitivity and the diversity of sentences. Finally, the summary generation follows a greedy approach for selecting ranked sentences.

2.1. HDP-based sentences topic clustering approach

In state-of-the-art methods, one usually estimates the topic distribution of sentence and cluster sentences into the topic which has the highest probability among its topic distribution. In our method, we use the whole topic distribution but not the main topic of sentence as a similarity measure to group sentences. And we argue that only using topic-based similarity is not enough for selecting important sentences. So, we simultaneously compute similarity of sentences by their cosine distance and integrate both of them into hypergraph. In our work, the topic model HDP (Teh, Jordan, Beal, & Blei, 2006) is