Knowledge-Based Systems 99 (2016) 28–38

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Multi-document summarization using closed patterns

Ji-Peng Qiang[a,b], Ping Chen[b], Wei Ding[b], Fei Xie[a,c], Xindong Wu[a,d,*]

[a] *Department of Computer Science, Hefei University of Technology, Hefei 230009, China*
[b] *Department of Computer Science, University of Massachusetts Boston, Boston 02125, USA*
[c] *Department of Computer Science and Technology, Hefei Normal University, Hefei 230601, China*
[d] *Department of Computer Science, University of Vermont, Burlington, VT 05405, USA*

ARTICLE INFO

Article history:
Received 6 March 2015
Revised 20 January 2016
Accepted 21 January 2016
Available online 2 March 2016

Keywords:
Multi-document summarization
Closed patterns
Text mining
Diversity
Content coverage

ABSTRACT

There are two main categories of multi-document summarization: term-based and ontology-based methods. A term-based method cannot deal with the problems of polysemy and synonymy. An ontology-based approach addresses such problems by taking into account of the semantic information of document content, but the construction of ontology requires lots of manpower. To overcome these open problems, this paper presents a pattern-based model for generic multi-document summarization, which exploits closed patterns to extract the most salient sentences from a document collection and reduce redundancy in the summary. Our method calculates the weight of each sentence of a document collection by accumulating the weights of its covering closed patterns with respect to this sentence, and iteratively selects one sentence that owns the highest weight and less similarity to the previously selected sentences, until reaching the length limitation. The sentence weight calculation by patterns reduces the dimension and captures more relevant information. Our method combines the advantages of the term-based and ontology-based models while avoiding their weaknesses. Empirical studies on the benchmark DUC2004 datasets demonstrate that our pattern-based method significantly outperforms the state-of-the-art methods. Multi-document summarization can be used to extract a particular individual's opinions in the form of closed patterns, from this individual's documents shared in social networks, hence provides a useful tool for further analyzing the individual's behavior and influence in group activities.

1. Introduction

Multi-document summarization has attracted much attention in recent years. With the rapid development of the World Wide Web, the explosion of electronic documents presents a serious challenge for readers to extract useful information from many relevant and similar documents. The Internet provides access to a huge volume of documents on a variety of topics with a considerable amount of redundancy. It calls for a robust multi-document summarization system, which can generate a succinct representation of a document collection by reducing information redundancy.

A large number of multi-document summarization systems have been presented in the literature. For example, the centroid-based methods [23,37] use clustering algorithms to generate sentences' clusters by calculating sentence similarity, and then select the most representative sentences from different clusters. The graph-based approaches [45,50] build a graph-based model, and then select sentences by means of voting from their neighbors using ideas like the well-known PageRank algorithm [7]. By considering latent semantics of document content, many methods based on latent semantic analysis [14] and non-negative matrix factorization [22,36] have been proposed. In addition, some ontology-based approaches [5,16] have also been used to produce summaries using lexical semantics.

Existing approaches basically fall into two major categories: term-based and ontology-based methods. A term-based method has the advantages of efficiency and maturity for term weight calculation. However, the main drawback is that it only focuses on single word significance without considering the problems of polysemy and synonymy, where polysemy means multiple meanings for a given word, and synonymy means multiple words express the same meanings. To solve these problems, ontology-based approaches take into account of meanings of lexicons. But they are restricted in some specific application domains where ontologies are available, and they cannot attain the semantic meanings of terms that do not exist in the ontology. Meanwhile, the construction of ontology is usually prohibitively expensive. To overcome these inherent weaknesses and keep the advantages of both term-based and ontology-based methods, we propose to generate a

[*] Corresponding author at: Department of Computer Science, Hefei University of Technology, China. Tel.: +86 055162902373.
E-mail address: xwu@hfut.edu.cn (X. Wu).

http://dx.doi.org/10.1016/j.knosys.2016.01.030
0950-7051/© 2016 Elsevier B.V. All rights reserved.

summary based on closed patterns, because closed patterns can capture the associations among the words, and no additional resources are required.

Over the past decade, a large number of association mining technologies have been proposed for a variety of tasks, including association rule mining, frequent itemset mining, sequential pattern mining, closed pattern mining, and maximum pattern mining [25]. In this paper, we will discuss how to effectively use these patterns in multi-document summarization. There are many types of sequential patterns, including frequent patterns, closed patterns, and so on [11,33]. As closed patterns are more compact and contain more information than frequent patterns without losing any information, we choose closed patterns for term weight calculation. To be specific, this paper will discuss a novel method for multi-document summarization using closed patterns, namely pattern-based summarization, which simultaneously considers content coverage and non-redundancy. It holds good statistical properties and captures more relevant information relative to the term-based methods. Compared with the ontology-based approaches, our multi-document summarization using closed patterns can capture informative terms in the document collection. The method does not rely on any external resources such as lexical knowledge bases. It only relies on the information in a document collection from which the summary is to be created.

Our pattern-based method includes the following steps. First, we mine all closed sequential patterns from a corpus. Then, we present a novel method that represents all sentences using these closed patterns. The model of sentence representation covers the main content of the document collection by calculating pattern weights with respect to the distribution of the closed patterns. Finally, we iteratively choose informative and non-redundant sentences by adopting a variant of the maximal marginal relevance evaluation strategy [8]. Experiments on the standard benchmark DUC2004 data sets demonstrate that the proposed algorithm outperforms the state-of-the-art term-based and ontology-based methods.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 presents pattern-based summarization. Section 4 shows experimental results. Finally, Section 5 concludes the paper.

2. Related work

Depending on the number of documents, automatic document summarization includes single-document summarization and multi-document summarization [3,46]. Single-document summarization only condenses one document into a summary, whereas multi-document summarization condenses a document collection into a single shorter representation. Multi-document summarization is considered as an extension of single-document summarization, and needs more sophisticated technologies and attracts much attention [29,31].

Multi-document summarization methods can be classified into two classes: extractive summarization and abstractive summarization [24,26]. Extractive summarization extracts the most informative document components, and abstractive summarization involves reformulation of contents. Extractive summarization is a simple but robust method without the requirement for advanced post-processing steps. Abstractive summarization requires deep natural language processing techniques for understating the documents [30]. Extractive summarization is more feasible and has become the standard in multi-document summarization. Summarization techniques also can also be categorized into query-based or generic (given a query or not), supervised or unsupervised methods (with a training set or not).

In this paper, we focus on unsupervised, extractive, generic, multi-document summarization. Unsupervised, extractive, and generic methods usually adopt the bag-of-words model for term weight calculation, also called term-based methods, which often use term frequency/inverse sentence frequency (TF∗ISF) weighting model and some extended schemes [20].

Term-based methods can be divided into the following categories. The centroid-based methods, as one of the most popular extractive methods, group document sentences into homogeneous clusters, and then select the representative sentences through computing the similarity values between sentences and the centroids of the clusters. For example, MEAD [23] computes the average cosine similarity between sentences and the rest of the sentences in the document collection as the centroid value of a sentence. Gong and Liu [14] presented a method to identify semantically important sentences using the latent semantic analysis (LSA). They first created a term-sentence matrix with each entry representing the weight of a term in its documents. Then they derived the latent semantic information by applying singular value decomposition (SVD). Some methods were proposed based on non-negative matrix factorization (NMF) [22,36]. The NMF-based methods also first create a term-sentence matrix to select meaningful sentences. Other methods were also developed including conditional random fields [32] and hidden Markov model [9].

The graph-based approaches [12,45,50] also belong to extractive summarization. They first produce a similarity graph, in which each node represents a sentence. When the cosine similarity value between a pair of sentences exceeds a threshold, these two sentences are connected by an edge. Erkan and Radev [12] proposed a method, called LexPageRank, which ranks the sentences based on the similarity graph following the well-known PageRank algorithm. Other improved graph-based algorithms have been proposed [6,45,50]. Bollegala et al. [6] presented a bottom-up approach to arrange sentences extracted for multi-document summarization. They defined four criteria, chronology, topical-closeness, precedence and succession, for capturing the association and order of two sentences.

Compared to the term-based methods, some ontology-based approaches [5,16] have been used to produce summaries. Specifically, ontologies have been used to (i) identify the concepts that are either most pertinent to a query [17,43] or most suitable for performing query expansion [27], (ii) model the context in which summaries are generated in a variety of domains, such as business domain [40], disaster management domain [20] and so on. Baralis et al. [5] proposed an ontology-based approach, called Yago-based summarization, which relied on Wikipedia [39] to map the words to non-ambiguous ontological concepts called entities. Yago-based summarization selects document sentences according to the previously assigned entities. Ontology-based approaches are limited in specific application domains, and also it takes much effort to construct the ontologies.

Pattern-mining techniques have been extensively studied for many years in data mining, such as frequent itemset algorithms (Apriori [1], FP-tree [15]), sequential pattern algorithms (PrefixSpan [28], SPADE [47]), closed sequential pattern algorithms (CloSpan [44], AGraP[13], BIDE [38]). As frequent itemsets or frequent patterns include more contextual semantic information than an individual term, they can improve the effectiveness of text mining applications, for example, text classification [2,41,49] and text clustering [19,48]. Algarni and Li [2] calculated term weights based on both their frequencies in documents and their distributions in sequential patterns. Zhang et al. [48] proposed maximum capturing (MC) for text clustering using frequent itemsets. MC can be divided into two components: constructing document clusters and assigning document topics.