Journal of Informetrics xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Journal of Informetrics

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

Finding a representative subset from large-scale documents

Jin Zhang [a], Guannan Liu [b], Ming Ren [c,*]

[a] School of Business, Renmin University of China, Beijing 100872, China
[b] School of Economics and Management, Beihang University, Beijing 100191, China
[c] School of Information Resource Management, Renmin University of China, Beijing 100872, China

ARTICLE INFO

Article history:
Received 4 March 2016
Received in revised form 23 May 2016
Accepted 23 May 2016

Keywords:
Information extraction method
Coverage
Redundancy
Distribution consistency

ABSTRACT

Large-scale information, especially in the form of documents, is potentially useful for decision-making but intensifies the information overload problem. To cope with this problem, this paper proposes a method named *RepExtract* to extract a representative subset from large-scale documents. The extracted representative subset possesses three desirable features: high coverage of the content of the original document set, low redundancy within the extracted subset, and consistent distribution with the original set. Extensive experiments were conducted on benchmark datasets, demonstrating the superiority of *RepExtract* over the benchmark methods in terms of the three features above. A user study was also conducted by collecting human evaluations of different methods, and the results indicate that users can gain an understanding of large-scale documents precisely and efficiently through a representative subset extracted by the proposed method.

1. Introduction

Recent years have witnessed the explosion of information, especially in the form of web documents, such as search results and online reviews. This enormous information pool is undoubtedly valuable for decision-making (Archak, Ghose, & Ipeirotis, 2011; Mudambi & Schuff, 2010). Users seeking certain information usually turn to search engines, e.g., Google, entering keywords and browsing the web pages returned by the search engines. Scholars usually need to process an abundance of academic articles returned by academic databases when they want to explore a certain topic or a specific research field. Customers prefer to read a number of online reviews prior to making purchase decisions, and enterprises also attempt to exploit such reviews on e-commerce platforms for gaining a competitive advantage.

The abundance of documents is potentially useful, but they contain redundant content and intensify the information overload problem (Chen, Huang, Hsieh, & Lin, 2011; Farhoomand & Drury, 2002). Users with limited time and information processing capability can easily get overwhelmed and obtain an incomplete understanding that deviates from the totality of the original document set, leading to a lower response rate and decisions of poor quality (Chen, Shang, & Kao, 2009).

Previous research efforts revealed that reducing the document volume can help mitigate the information overload problem (Badenoch, Reid, Burton, Gibb, & Oppenheim, 1994; Cook, 1993). Along this direction of investigation, both scientists and engineers have begun to explore information extraction methods to assist users in understanding large-scale documents comprehensively and easily. A commonly used type of method is to extract a subset of documents according to certain evaluation functions, which is called the top-*k* methods (Bruno, Chaudhri, & Gravand, 2002; Dai & Davison, 2011;

* Corresponding author.
E-mail address: renm@ruc.edu.cn (M. Ren).

http://dx.doi.org/10.1016/j.joi.2016.05.003
1751-1577/

Ilyas, Beskales, & Soliman, 2008; Mamoulis, Cheng, Yiu, & Cheung, 2007; Marian, Bruno, & Gravano, 2004). The top-*k* results perform well on the evaluation functions, but they are often quite similar, sometimes even identical, leading to a rather high redundancy in the results. In other words, top-*k* results cover only a few aspects of the content and may miss some aspects that are important to users. Furthermore, top-*k* results cannot reflect the content distribution, which is crucial to the decision-making process. For example, customers may make a different decision if they read the top-*k* reviews in which the distribution of positive and negative opinions is not consistent with that of the original set (Lei, Dawar, & Gurhan-Canli, 2012).

Therefore, it is desirable to provide users with a representative subset to aid them in grasping the main ideas of the original documents. Hochbaum and Pathria (1998) have proposed a method called Maximum Coverage to generate a subset in light of the coverage. However, this method only performed well on a balanced dataset and hardly considered the redundancy. Pan, Wang, Anthony, and Yang (2005) have proposed a greedy method that aims to extract a representative subset in terms of both coverage and redundancy. This method has limitations in two aspects: one is that it does not take into account the content distribution of the representative subset (i.e., consistent distribution with the original set of documents), and the other is that the measures the authors used need to be further improved upon.

This paper proposes a novel method called *RepExtract* to extract a representative subset from large-scale documents, with good performance on three desirable features, i.e., high coverage of the content of the original document set, low redundancy within the extracted subset, and distribution consistent with the original set of documents. First, the original set is clustered so that documents with similar content are assigned to the same cluster. Then, an algorithm is designed to extract a certain number of representative documents from each cluster according to the cluster sizes. The larger a cluster, the more documents are extracted from it. In doing so, the representative documents from various clusters cover different content of the original set. The representative subset has low redundancy because the documents extracted from different clusters are dissimilar. Moreover, the content distribution of the representative subset is consistent with that of the original set. Extensive data experiments on benchmark datasets and a user study were conducted to demonstrate the advantages of *RepExtract* over other commonly used extraction methods, showing that the representative subset enables users to have a precise and comprehensive understanding of the large-scale documents and thus make informed decisions.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 introduces the extraction method in detail. Section 4 shows experimental results that demonstrate the superior of the proposed method. Further discussion of the proposed method is provided in Section 5, and Section 6 concludes this paper and highlights future research.

2. Literature review

The related studies with respect to the information overload problem can be categorized into four streams, i.e., top-*k* methods, representative extraction methods, text summarization methods, and associative retrieval methods, and they are discussed in this section.

2.1. Top-k methods

Top-*k* methods are commonly used in information retrieval (Bruno et al., 2002; Dai & Davison, 2011; Fagin, Lotem, & Naor, 2003; Guntzer, Balke, & Kießling, 2000; Ilyas et al., 2008; Mamoulis et al., 2007; Marian et al., 2004). They identify the *k* data objects with the largest values with respect to certain ranking functions. The basic assumption is that every data object is mutually independent. In other words, the ranking function value is calculated for each data object separately. Guntzer et al. (2000) studied the retrieval of the top-*k* results from a multi-feature image database and presented a method called Quick-Combine for combining the multi-feature result lists, guaranteeing the correct retrieval of the top *k* results. Bruno et al. (2002) attempted to optimize top-*k* queries by mapping a top-*k* selection query to a traditional range selection query, so that it can be executed by an RDBMS. Fagin et al. (2003) used an aggregation function to combine the values under each attribute and obtain an overall evaluation. They presented the "threshold algorithm" for obtaining the top-*k* answers with the overall scores from the aggregation function. Marian et al. (2004) studied top-*k* queries for web applications. They designed a sequential algorithm to process these queries and parallelized the source access to minimize the query response time. Mamoulis et al. (2007) identified two phases in the traditional top-*k* algorithms and proposed a new algorithm by exploiting appropriate data structures to improve the efficiency of the top-*k* search. Ilyas et al. (2008) conducted a survey on the top-*k* query processing techniques in relational database systems and classified the processing techniques into query models, data access methods, implementation levels, data and query uncertainty, and ranking functions. Dai and Davison (2011) demonstrated that the ranking performance was sensitive to the topic distribution of queries, and they concluded that the topic distributions of queries should approximate real-world search log topic distributions in ranking evaluations.

As discussed in Section 1, the results provided by the top-*k* methods are quite similar, and sometimes even identical, leading to high redundancy and low coverage degrees in the top-*k* results. Furthermore, the content distribution of the top-*k* results does not closely reflect the original distribution, which largely motivates our attempt at developing an effective method for extracting a representative subset.

Please cite this article in press as: Zhang, J., et al. Finding a representative subset from large-scale documents. *Journal of Informetrics* (2016), http://dx.doi.org/10.1016/j.joi.2016.05.003