Big Data Summarization Using Novel Clustering Algorithm and Semantic Feature Approach

Shilpa G. Kolte, University of Mumbai, Terna Engineering College, Navi Mumbai, India
Jagdish W. Bakal, Shivajirao S. Jondhale College of Engineering, Dombivli, Thane, India

ABSTRACT

This paper proposes a big data (i.e., documents, texts) summarization method using proposed clustering and semantic features. This paper proposes a novel clustering algorithm which is used for big data summarization. The proposed system works in four phases and provides a modular implementation of multiple documents summarization. The experimental results using Iris dataset show that the proposed clustering algorithm performs better than K-means and K-medodis algorithm. The performance of big data (i.e., documents, texts) summarization is evaluated using Australian legal cases from the Federal Court of Australia (FCA) database. The experimental results demonstrate that the proposed method can summarize big data document superior as compared with existing systems.

KEYWORDS

Data Generalization, Data Summarization, Hadoop, Semantic Term Identification

1. INTRODUCTION

In recent years, with the rapid development of the Internet, network service has become one of the most frequently used computer applications. Search engine, webmail, and social network services are presently crucial data intensive applications. Search engine, webmail, and social network services are presently crucial data intensive applications. Because increasingly more people use web services, processing a big amount of data efficiently can be a significant problem. The very word “Big” indicates size. Big data have rapidly developed into a hotspot that attracts great attention to academia, industry, and even governments around the world (Mayer-Schonberger, & Cukier, 2013). Big data is of great value, which is beyond all doubt. From the perspective of the information industry, big data is a strong impetus to the next generation of IT industry, which is essentially built on the third platform, mainly referring to big data, cloud computing, mobile Internet, and social business. IDC predicted that by 2020 the market size of the third IT platform will reach US$ 5.3 trillion; and from 2013 to 2020, 90% of the growth in the IT industry would be driven by the third IT platform. From the socio-economic point of view, big data is the core connotation and critical support of the so-called second economy, a concept proposed by the American economist (W.B. Arthur, 2011), which refers to the economic activities running on processors, connectors, sensors, and executors. It is estimated that at 2030 the size of the second economy will approach that of the first economy (namely, the traditional physical economy). The main support of the second economy is big data, as it is an inexhaustible and constantly enriching resource. In the future, by virtue of big data, the competence under the second economy will no longer be that of labor productivity but of knowledge productivity.

DOI: 10.4018/IJRSDA.2017070108

Copyright © 2017, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
Compare with traditional datasets, big data typically includes masses of unstructured data that need more real-time analysis. In addition, big data also bring about new opportunities for discovering new values, helps us to gain an in-depth understanding of the hidden values, and also incurs new challenges, e.g. how effectively organized and manage such datasets. Recently, industries become interested in the high potential of big data, and many government agencies announced major plans to accelerate big data research and applications (Federal, 2012). Nowadays, big data related to the service of Internet companies grow rapidly. For example, Google processes data of hundreds of Petabyte (PB), Facebook generates log data of over 10 PB per month, Baidu, a Chinese company, processes data of 10 PB per day, and Taobao, a subsidiary of Alibaba, generates data of tens of Terabyte (TB) for online trading per day. The features of big data can be characterized by 5V, namely, huge Volume, high Velocity, high Variety, low Veracity, and high Value. The main difficulty in coping with big data does not only lie in its huge volume, as we may alleviate to some extent this issue by reasonably expanding or extending our computing systems. Actually, the real challenges center on the diversified data types (Variety), timely response requirements (Velocity), and uncertainties in the data (Veracity). Because of the diversified data types, an application often needs to deal with not only traditional structured data, but also semi-structured or unstructured data (including text, images, video, and voice). Timely responses are also challenging because there may not be enough resources to collect, store, and process the big data within a reasonable amount of time.

Big document summarization method is an important technique for data management of Internet of Things (IoT). Document summarization is the process of reducing the sizes of documents while maintaining their basic outlines. That is, it should extract the most significant information from the document. The summarization method can involve either generic summaries or query-based summaries. A generic summary distills an overall wisdom of a document’s contents, whereas a query-based summary distills only the contents of a document that is relevant to a user’s query. Traditional document summarization methods are restricted for summarizing suitable information from the big document data, since it has been proposed for enhancing the summarization precision which it uses various statistical or natural language processing methods based on single node computer environment. The objective of this paper is proposed big document summarization method which the information is summarized from a big document data. The proposed method uses the novel clustering algorithm and extracted semantic feature of document using Haddop.

The reminder of the paper is organized as follows. Section 2 presents related work for big data summarization, Section 3 explores Proposed Framework for big data summarization; Section 4 illustrates experimental setup of the proposed data summarization system. This section also gives performance evaluation with the existing algorithms. Finally, the section 5 concludes the paper.

2. RELATED WORK

Gong and Liu proposed summarization method using LSA (Latent Semantic Analysis). This method extracts the important sentence which has the largest index value with respect to the important singular vector by LSA (Gong & Liu, 2011). Zha employed the mutual reinforcement principle (MRP) and sentence clustering for the generic summarization. Their method clusters sentences of documents into several topical groups by sentence clustering method. And then, sentences are extracted from each topical group by saliency scores using the MRP (i.e., modified LSA method) (Zha, 2002). Yeh et al. proposed the summarization method using LSA and the text relationship map (TRM). Their method finds semantic sentences using LSA. TRM is constructed by the semantic sentences, and the important sentences are extracted by the number of links in TRM (Yeh et al., 2005). Li et al. extended
A Novel Aspect Based Framework for Tourism Sector with Improvised Aspect and Opinion Mining Algorithm

Secure Mechanisms for Key Shares in Cloud Computing
[www.igi-global.com/article/secure-mechanisms-for-key-shares-in-cloud-computing/206875?camid=4v1a](www.igi-global.com/article/secure-mechanisms-for-key-shares-in-cloud-computing/206875?camid=4v1a)