Knowledge-Based Systems 118 (2017) 15–30

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

An overview on the roles of fuzzy set techniques in big data processing: Trends, challenges and opportunities

Hai Wang[a], Zeshui Xu[a,b,*], Witold Pedrycz[c,d,e]

[a] School of Economics and Management, Southeast University, Nanjing, Jiangsu 211189, China
[b] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China
[c] Department of Electrical & Computer Engineering, University of Alberta, Edmonton, AB T6R 2V4, Canada
[d] Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[e] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

ARTICLE INFO

Article history:
Received 8 July 2016
Revised 29 October 2016
Accepted 13 November 2016
Available online 14 November 2016

Keywords:
Big data
Data-intensive science
Fuzzy sets
Fuzzy logic
Granular computing

ABSTRACT

In the era of big data, we are facing with an immense volume and high velocity of data with complex structures. Data can be produced by online and offline transactions, social networks, sensors and through our daily life activities. A proper processing of big data can result in informative, intelligent and relevant decision making completed in various areas, such as medical and healthcare, business, management and government. To handle big data more efficiently, new research paradigm has been engaged but the ways of thinking about big data call for further long-term innovative pursuits. Fuzzy sets have been employed for big data processing due to their abilities to represent and quantify aspects of uncertainty. Several innovative approaches within the framework of Granular Computing have been proposed. To summarize the current contributions and present an outlook of further developments, this overview addresses three aspects: (1) We review the recent studies from two distinct views. The first point of view focuses on what types of fuzzy set techniques have been adopted. It identifies clear trends as to the usage of fuzzy sets in big data processing. Another viewpoint focuses on the explanation of the benefits of fuzzy sets in big data problems. We analyze when and why fuzzy sets work in these problems. (2) We present a critical review of the existing problems and discuss the current challenges of big data, which could be potentially and partially solved in the framework of fuzzy sets. (3) Based on some principles, we infer the possible trends of using fuzzy sets in big data processing. We stress that some more sophisticated augmentations of fuzzy sets and their integrations with other tools could offer a novel promising processing environment.

1. Introduction

We are in the era of big data [1]. Every day, data are generated and grown from lots of sources including retail transactions, social media and sensors, with the unprecedented rate which exceeds the Moore's law [2]. For instance, it was reported that 665 terabytes of data were created by a typical hospital in 2015 [3]. This volume is much greater than that of the web archive of the US Library of Congress. Even in our daily life, we use search engines, send and receive e-mails and texts, celebrate our newborns on social media networks, and navigate cars by geo-tracking systems. Consequently, big data have influenced our daily behaviors, revolutionized scientific developments, even affected the planning and policies of the governments [4]. However, the phrase "big data" focuses on not only the volume but also on their velocity and variety, which are known as 3Vs [5]. Moreover, other characteristics, such as value and veracity, are also frequently considered [2]. A brief explanation of these characteristics is shown in Fig. 1.

It has been demonstrated that big data can help a lot with businesses, management, medicine, health care, engineering, scientific research, to name but a few examples. One prominent example, reported by Nature [6], is that the Google Flu Trends (GFT) can predict more than double the proportion of doctor visits for influenza-like illness than the Centers for Disease Control. Although Lazer et al. [7] remarked some limitations of GFT, many of which have been eliminated by other systems, such as Twitter [8]. In medicine, analyzing a large amount of tumors would reveal general patterns to improve diagnosis and treatment [9]. Also, the appearance of big data is leading to a revolution of statistics because data can be collected with universal or near-universal population coverage instead of relatively small-sample surveys [10]. Economic and management

* Corresponding author at: School of Economics and Management, Southeast University, Nanjing, Jiangsu 211189, China.
E-mail addresses: wanghai17@sina.com (H. Wang), xuzeshui@263.net (Z. Xu), wpedrycz@ualberta.ca (W. Pedrycz).

http://dx.doi.org/10.1016/j.knosys.2016.11.008
0950-7051/

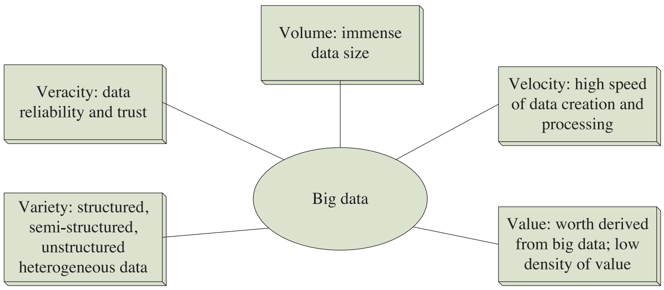

Fig. 1. The 5Vs view at big data.

research has evidently benefitted from this fact. For example, Blumenstock et al. [11] demonstrated that the use of individual's past history of mobile phone can infer the socioeconomic status and accurately reconstruct the distribution of wealth of an entire nation.

Apparently, the discovery of knowledge from big data calls for the support of certain techniques and technologies. It is commonly acknowledged that a novel paradigm of scientific research, i.e., data- intensive science [12], has emerged along with the appearance of big data. In this paradigm, some basic activities, such as data capture, curation, analysis and interpretation/visualization, are usually considered as the value chain of big data [13]. To achieve these activities, several categories of distinct techniques have been considered, including mathematical tools, data analysis techniques, visualization tools and Granular Computing (GrC) techniques. These techniques are usually implemented by specific big data technologies, which involve batch processing, stream processing (or real-time processing) or hybrid processing with the Lambda architecture [2,14].

However, although many technologies, such as MapReduce [15] and Hadoop [16], have been released, those are far from meeting the ideal requirements of each processing step. There are challenges presented in almost every aspect of big data processing and applications, including technical challenges and non-technical ones. A visible general technical challenge is that the speed of data generation has overtaken our capacity of processing [17]. It is essential to improve capabilities of data management and programming, develop creative and scalable techniques to analyze and understand large-scale data sets with complex structures [10]. When analyzing big data, another essential topic is how to access and explore data without sacrificing privacy and confidentiality concerns [10,18-20]. Frequently, consumers and clients reveal information to others including commercial entities and governments knowingly or unwittingly. But the erosion of privacy is alarming now [18]. Researchers have argued that policy should focus more on how big data is used and less on how it is collected and analyzed because the abuse of personal data may threaten our autonomy [19]. One commonly mentioned non-technical challenge is that research budgets are flat or declining in inflation-adjusted terms [3,21].

In order to figure out current challenges, various solutions are being sought for supplying many possible directions. Sejdic [3] suggested adapting classical information processing techniques. Whereas Heinis [22] insisted that scalable approximation algorithms with tight error boundaries would be more efficient than those yielding "conventional" precise computations for interpreting data and refining the explorative phase of the analysis.

Fuzzy set techniques, including extensions and generalizations of fuzzy sets, fuzzy logic, fuzzy systems, has become an interesting and viable methodologies and tools for GrC [23]. Since introduced by Zadeh [24], fuzzy sets have been applied to various areas such as control systems, pattern recognition and machine learning. Fuzzy sets enable us to represent and process information at distinct levels of information granularity. So far there have been a number of contributions focusing on the use of fuzzy sets to process and/or understand big data. There are at least four reasons why fuzzy set techniques offer some promise or have already demonstrated some advantages in the context of big data:

(1) Uncertainties not only exist in the data themselves but occur at each phase of big data processing. For instance, the collected data may be created by faulty sensors or provided by not fully informed customers; the outputs of specific artificial intelligent algorithms also contain uncertainties. In these cases, fuzzy set techniques could be one of the most efficient tools to handle various types of uncertainties.
(2) Handling uncertainties can come with different flavors. Most frequently, an excessively precise solution to a problem could be very expensive, or may not be required. It might be sufficient to go at a certain level of detail to discover necessary knowledge and provide required solutions. In fact, fuzzy set techniques (and other GrC techniques) can be employed so that a problem can be reconstructed at certain granular level. For instance, when developing advertising strategy, it is significant to recognize the purchase preferences of a community. But it is not always necessary (maybe not accessible) to differentiate among exact preferences of individuals. In this case, it would be more efficient to mine preferences from the view of communities instead of individuals. In other words, it is better to solve the problem at a coarse granular level, i.e., communities.
(3) Especially, fuzzy set techniques would be more efficient if they are used associated with other decision making techniques, such as probability, rough sets, neural networks, etc., because each type of techniques exhibit their own strengths of representing and handling information granularity.
(4) It has been acknowledged that information granules are considered instead of numbers for communication with users in systems and platforms. Fuzzy set techniques, e.g., computing with words (CWW), could be instrumental for interacting with users in an understandable and interpretable manner.

Therefore, fuzzy sets can improve the current big data techniques and alleviate the existing big data challenges, including the ones raised by the 5Vs, by pre-processing data or by reconstructing the problem at a certain granular level. However, it should be mentioned that, different from other hot techniques for big data, like deep learning, the role of fuzzy set techniques (as well as other GrC techniques) is a kind of methodology that provides new strategy for knowledge abstraction (granulation) and knowledge representation. As will be seen in the coming parts, fuzzy set techniques help us handle data in a new manner. Thus we do not anticipate solving big data problems independently by using only GrC techniques.

This paper is aimed to offer a systematic review on the existing contributions of big data processing based on fuzzy set techniques. To do so, the taxonomy of this review is conducted by two perspectives. In the first taxonomy, we classify the existing contributions by the specific fuzzy set techniques to illustrate what techniques have been employed. Then, in the second taxonomy, the literature is categorized by the focused big data problems in order to explain when and why the fuzzy set techniques are useful. Furthermore, we discuss the current challenges, which might be partially solved or mitigated by fuzzy set techniques. Based on the existing trends and challenges, we present some possible opportunities to guide further developments. Therefore, this review could be useful and meaningful for designing new and state-of-the-art big data techniques.

The paper is organized as follows. Section 2 recalls some necessary preliminaries. The current contributions are reviewed by