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/

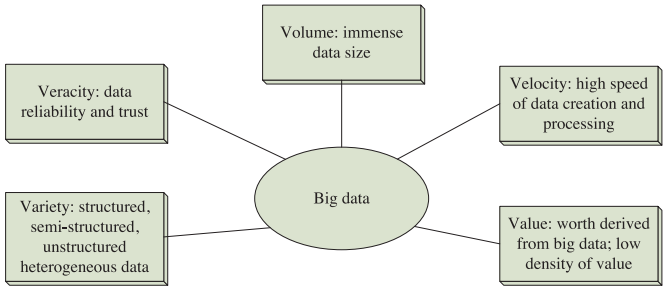

Fig. 1. The 5Vs view at big data.

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

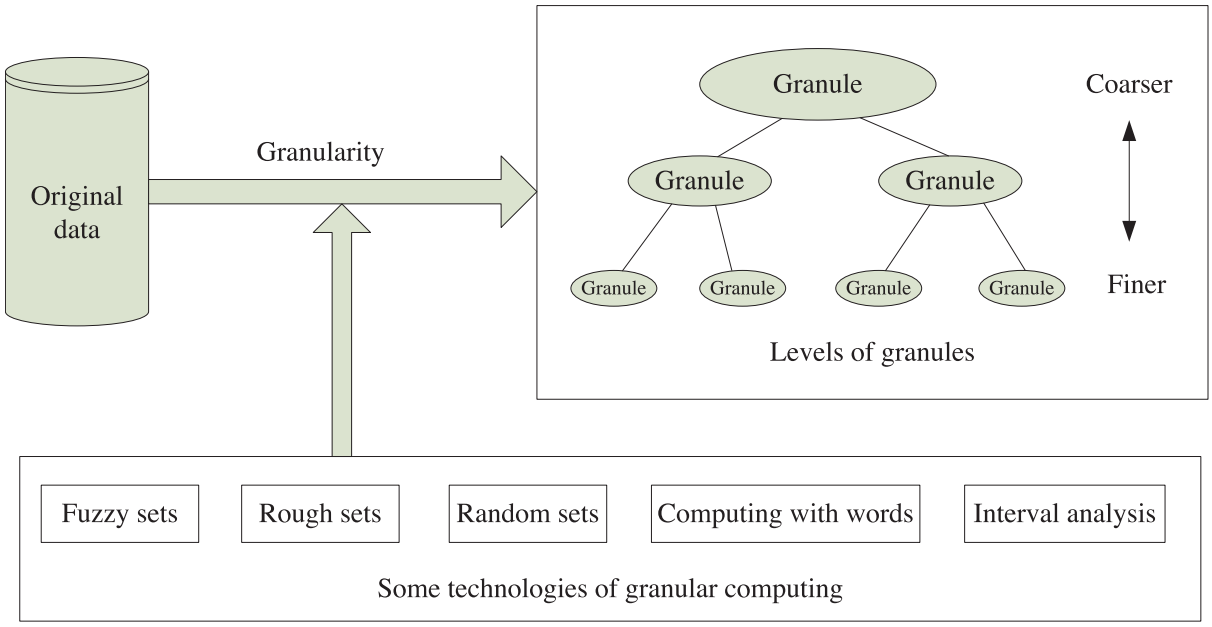

Fig. 2. The framework and selected components of GrC.

two distinct taxonomies in Section 3. The summarization on the emerged fuzzy set techniques, as well as their roles in big data problems, is presented in Section 4. A critical review is also given in this section. Section 5 summarizes current challenges related to our topic. Then possible directions and opportunities are discussed in Section 6. Finally, Section 7 presents some concluding remarks.

2. From granular computing to fuzzy set techniques

We start with some necessary prerequisites, involving the generic idea and framework of GrC, as well as elaborate on some specific fuzzy set techniques.

GrC is a category of computing paradigm of information processing involving theories, methods and techniques of information granulation. As shown in Fig. 2, this paradigm is, frequently, implemented by several technologies including fuzzy sets, rough sets, random sets, CWWs, etc. [4]. The core concepts are granule and granulation/abstraction. The role of a granule in GrC is similar to that of subset, class, or cluster in a universe. Generally, a granule is a building block, formed by fuzzy sets, rough sets, random set, and interval set [25]. It is a clump of entities drawn together by the inter and intra relationships among granules such as indistinguishability, similarity and functionality. In practice, distinct aspects of the problem can be characterized by granules with different levels of granularity and different sizes and shapes. Granulation, which refers to the process of construction, representation and interpretation of granules, can be performed by integrating and dispersing the structure of granules [26]. By integration, larger and/or higher level granules are developed based on smaller and/or lower level granules; while by dispersion, things are done in the opposite way.

Among those GrC tools, fuzzy set techniques are the popular one to deal with big data. As will be seen in the coming section, more than 50 contributions related to big data and fuzzy set techniques have been published in the recent 5 years. The major reason is that granulation is inherent in the concept of fuzzy set. Fuzzy set techniques provide a novel way to investigate and represent the relation between a set and its members by considering the continuum degree of belonging, namely membership functions, which is similar to the way of human recognition. Moreover, fuzzy information granulation is about a pool of fuzzy granules derived by granulating objects, rather than a single fuzzy granule [4]. Generally, a fuzzy set A in X is defined by a membership function, which maps each $x \in X$ into a value in the interval [0, 1]. The membership function refers to the degree of compatibility of x to an imprecise concept.

Fuzzy logic, as a form of many-valued logic, was introduced based on fuzzy set to handle the partial truth between completely true and completely false. When applied to control theory and artificial intelligence, all input values should be fuzzified into membership degrees, and then the computational processes are usually executed by the IF-THEN rules, associated with some fuzzy operators on fuzzy sets including the operations of AND, OR and NOT. Especially, when linguistic variables are used, the corresponding fuzzy logic processes are also called CWW. In the setting of big data, there are several studies focusing on the use of fuzzy logic.

Type-2 fuzzy sets generalize the common fuzzy sets by incorporating the uncertainties of defining membership functions into fuzzy sets. That is, the membership degree of an element belonging to a set can be also a fuzzy set. There are a lot of specific type-2 fuzzy sets, such as interval fuzzy sets, intuitionistic fuzzy sets, interval-valued type-2 fuzzy sets, hesitant fuzzy sets, and etc. Generally, type-2 fuzzy sets and systems are more complicated than the type-1 version, but they are more efficient and sophisticated to handle uncertainties.

As will be seen in the coming section, theories based on fuzzy sets have been applied in various big data applications, such as medical and healthcare, intelligent transportation systems and social networks. Fig. 3 shows the numbers of publications, which involve fuzzy set techniques in big data processing. The number of publications exhibits a growing tendency.

3. Fuzzy set techniques in big data processing: current developments

In order to understand the roles of fuzzy set techniques easily, we first recall the commonly acknowledged paradigm of big data processing, i.e., data-intensive science. The central issue of the paradigm is data, instead of computation, and results in thinking with data. Fig. 4 shows the chain of value-generating in the presence of big data. The first activity, namely data capture, collects

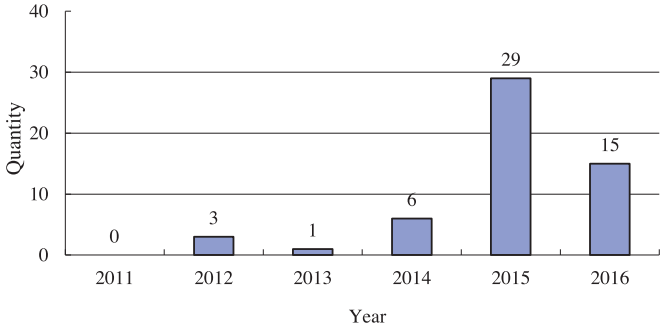

Fig. 3. The numbers of publications regarding fuzzy set techniques presented in big data processing.

65

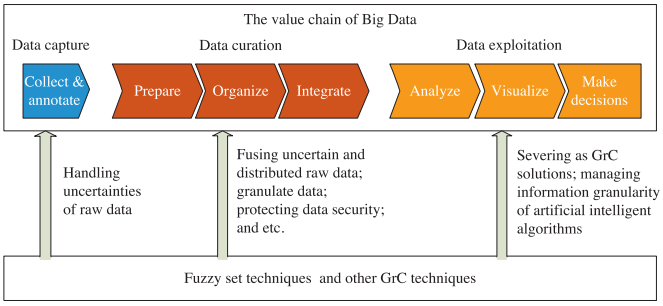

Fig. 4. Roles of fuzzy set techniques in big data processing.

Table 1
An overview of contributions using fuzzy set techniques in big data processing.

Theory	#Articles	Articles' references
Fuzzy sets	26	[15,25,27-50]
Fuzzy logic	16	[31,37,51-64]
Type-2 fuzzy sets	4	[65-68]
Integration with other GrC techniques	7	[16,32,68-72]

available data resources so that the analysis can be performed. Data annotation is also considered in this phase [13]. Data curation covers a wide range of procedures, including data preparation (which enables access to sources and sets up access-control rules), organization (which identifies syntax, structure and semantics for each data source) and integration (which establishes a common data representation and maintains data provenance). Data Exploitation includes the whole range of activities throughout the workflow pipeline, moving from analyzing the prepared data to presenting the analytic result as a static report or an interactive application that supports exploration and refinement and then arriving at making the informative decisions.

The main roles of fuzzy set techniques, as well as other GrC techniques, are summarized in Fig. 4.

As will be reviewed in this section, these techniques can serve as effective tools in big data processing. In all phases of the value chain, the functions of fuzzy set techniques move from handling the uncertainties of raw data and annotating data to preparing specific granular representation of data for artificial intelligent algorithms. In what follows, we summarize the emerged fuzzy set techniques at first, and then analyze when and why these techniques can work well.

3.1. The emergence of fuzzy sets in big data

We will review the existing big data processing techniques involved with specific fuzzy set techniques in this section. Table 1 shows an overview on the application of various genres of fuzzy set techniques in big data processing.

Generally, the listed classes of theories may overlap each other. Fuzzy sets refer to the contributions using the membership functions of fuzzy sets directly for computations; Fuzzy logic focuses on fuzzy reasoning based on membership degrees or linguistic terms, fuzzy control, etc., based on fuzzy sets.

As can be seen in Table 1, most existing contributions take use of the basic concepts of fuzzy sets, such as membership functions and alpha-cuts [27-28,30,34,41-42,44,47-48,50]. The most significant role of fuzzy sets is to represent various kinds of uncertainties in the data themselves and in the procedure of processing data. Especially, the introduction of fuzzy sets in big data enhances the capability of representing information granules. Several other GrC tools, such as clustering algorithms, have been used and/or developed for big data. Among which, the fuzzy c-means (FCM) algorithm is definitely the most popular one for processing big data [15,33,38,43,45-46]. Some new developments of clustering algorithms [36,73] have also been proposed. Besides, fuzzy sets have been adopted for granularity directly [25,29,35]. Reasoning with membership degrees are usually considered in fuzzy control, fuzzy inference system and fuzzy classification in big data processing [31,37,51-54]. CWW techniques are more popular in big data applications. Most of the contributions consider the use of the existing CWW techniques (with slight changes to fit the specific cases), including the fuzzy rules reasoning system [58], the fuzzy logic based approach [61], linguistic fuzzy rule-based classification systems [56,59] and the fuzzy linguistic summarization technique [63] and neuro-fuzzy reasoning systems [55,62]. New CWW techniques were presented for the causal combination of variables [64], which represent the embedded uncertainty and ambiguity [60]. Multi-granular CWW was considered in [57].

Big data applications have benefitted from type-2 fuzzy sets as well. Chang et al. [68] proposed a novel large-scale type-2 neuro-fuzzy multi-rule based system to explore and assemble knowledge granules. Li et al. [67] put forward an interval fuzzy sets-based classification with continuous valued attributes, in which type-2 fuzzy sets are considered to fuzzify class labels and discrete conditional attributes. Wang [65] developed a type-2 fuzzy event parallel computing system to overcome computer int index limitation in big data. In [66], a type-1 interval number represents a distribution of features and type-2 interval numbers emerge when the distributions of type-1 interval numbers are learnt by a fuzzy reasoning-based classifier.

Fuzzy set techniques can perform much better for some special big data applications if it is integrated with other GrC techniques. For instance, the approach proposed in [68] is formed based on the type-2 fuzzy sets and the traditional fuzzy rule based system. Several studies resolve big data problems based on the combination of fuzzy sets and rough sets. Fuzzy rough sets were used for incremental feature selection [71] and community detection [69]. A more general version, i.e., rough intuitionistic fuzzy sets, were applied to improve the performance of FCM algorithm [16]. Different from which, the decision making approach proposed by Bai et al. [70] was constructed by using rough sets and FCM algorithms separately. More complicatedly, He et al. [72] designed the decision making system based on fuzzy sets, rough sets and technique for order preference by similarity to ideal solution (TOPSIS). But it is a pity that their system has not been validated by real big data problems. Besides, fuzzy sets was applied to distill big data, associated with Bayesian inference [32].

3.2. When and why fuzzy set techniques work

We have seen that various genres of fuzzy set techniques have been introduced to handle big data in Section 3.1. We shall review

the successful big data application area of using fuzzy set techniques to show which fields have been benefitted. To do so, we will: (1) ignore the specific techniques of fuzzy set techniques in the literature; (2) review the existing contributions of using fuzzy set techniques from two aspects: big data processing and big data applications. Notice that it is generally hard to separate big data processing from applications. We try to do this based on the major focuses of the contributions, regardless of some inevitable overlaps. To begin with, we summarize the contributions, which include the phrase "big data" in their title or keywords, based on the aspects of the nature of problems, the roles of fuzzy set techniques, the advantages of using this technique, datasets, and the 5Vs of big data. The summary is listed in Table 2.

3.2.1. Improving the capacity of processing big data

Several contributions focus on improving the existing approaches and/or algorithms for handling big data by means of fuzzy set techniques. The involved area includes machine learning, query refining, data pre-processing, etc.

3.2.1.1. Machine learning and pattern recognition in big data setting. The roles of fuzzy set techniques in classification of big data do not make tremendous changes, comparing with that in the traditional artificial intelligence. The fuzzy linguistic rule based classification system (FRBCS) was improved to a cost-intensive one [56] so as to meet the requirement of the imbalanced big data. Four Vs are considered in their study. FRBCS can not only manage uncertainty, ambiguity and vagueness in an efficient manner but also supply interpretable results for users by using fuzzy linguistic labels. In the situation of big data, the usage of the improved version FRBCS provides an effective way to treat the inherent uncertainties such as the borderline samples, the noise within raw data, etc. The improved version possesses good performance of classifying the imbalanced datasets and the generated rules have the same structure, and thus, it can be implemented in a parallel approach. Similarly, under the framework of MapReduce, Del Rio et al. [59] implemented the linguistic fuzzy rule based classification system in parallel. Their proposal was tested by various large data sets. But no other feature of big data has been focused. In [62], three adaptive neuro-fuzzy inference systems were implemented for short term photo-voltaic power prediction. Here, volume is the only focused V.

Considering that fuzziness in clustering has been started for decades, the scholars have improved this type of algorithms a lot in the big data era. Havens et al. [33] extended the traditional FCM algorithm to suit very large data in 2012. After that, the use of fuzziness in FCM has not been improved. Most of the subsequent investigations consider its application in big data, such as implementation and scalability in the framework of MapReduce and Hadoop, ensemble learning [46], and business intelligence [68]. Moreover, Peters and Weber [74] presented a framework of guiding further developments of fuzzy clustering algorithms, especially dynamic fuzzy clustering algorithms. It is a pity that only Chang et al. [68] concentrated on more than one V, i.e., volume and velocity.

Feature selection is vital for further processing including classification and clustering. It has been demonstrated that feature selection can profit from fuzzy set techniques. Azar and Hassanien [55] trained neuro-fuzzy classifier to select features for medical big data classification. The employed fuzzy feature selection algorithm [75] takes use of the values of linguistic hedges to denote the importance of degree of input features. Zeng et al. [71] presented an incremental feature selection algorithm to handle the volume, velocity and variety of big data. Based on the hybrid distance measure for various kinds of data, a novel fuzzy rough set was constructed, and then the conditional attributes were selected by means of the idea traditional rough set model. Considering the velocity of big data, the incremental version of feature selection algorithm was developed at last. Besides, Kaburlasos and Papakostas [66] studied the distributions of image features by means of fuzzy logic, where an interval number represents a distribution of image features including orthogonal moments and type-2 interval numbers describe the distribution of interval numbers. Different from the traditional selection strategy, Mendel and Korjani [64] established the nonlinear combinations of conditional attributes by a causal combination method. Each combination, i.e., a subset of attributes, has a linguistic term whose semantics is a fuzzy set. The method focuses on interconnections, which is modeled by the minimum operation of fuzzy sets with a causal condition or its complement. Comparing with classical Cantor sets, fuzzy sets are more efficient to model closeness to corners in a vector space.

Sampling also benefits from fuzzy set techniques. The sampling approach proposed by He et al. [27] depends on the fact that the distribution of the original large data set, the boundaries of all possible separating hyper surfaces and the elements in a minimal consistent subset are all uncertain. Especially, all the possible separating hyper surfaces result in a fuzzy set. The samples in the fuzzy set, with different membership degrees, will be useful in division process. Except for this, Li et al. [67] reduced the size of large data samples by treating the "similar" samples as one record. The original samples are granulated by discretizing the values of conditional attributes into a number of intervals, which are represented by the means and standard deviations. Then, the samples having the same granular values with respect to all conditional attributes are merged. The class labels are decoded by fuzzy linguistic terms as well. As a result, the given dataset is transferred into a relatively small version where the values of conditional attributes are interval and the labels are fuzzy classes.

3.2.1.2. Query refining and data privacy. Privacy is one of the most significant issues of this big data age. But privacy protection may conflict with the requirement of query sometimes. Some scholars even argue that the technologies may be unnecessary if they contain the risk of violating privacy [20,76]. Till now, there are some researches that focus on adopting fuzzy set techniques to improve the ability of query refining and/or privacy protection.

Generally, fuzzy measures are considered to understand the users' query requirements so that the quality of query can be refined. Fuzzy query has been acknowledged as an efficient strategy because it considers the users' query by a fuzzy similarity measure or an optimization of users' satisfaction. Liu et al. [48] improved the traditional SQL-based fuzzy query using format-preserving encryption as the underlying primitive. The novel SQL-based fuzzy query mechanism can query over encrypted data directly. Cai et al. [42] presented a query refining system in collaborative classification systems. In their study, a collaborative tagging system classifies resources by a multi-class classification algorithm. A query is then treated as a fuzzy satisfaction problem. Specifically, the relevance measurement between a resource and a user's query is defined by the fuzzy mapping from the Cartesian product of the set of resources and the set of queries to the interval [0, 1]. A user's query is regarded as fuzzy requirements of the user on resources content. Similarly, the relevance measurement between a resource and a user's interest requirement is treated as a fuzzy satisfaction problem as well. Then, the final resources ranking score is represented by another fuzzy satisfaction problem of both query relevance and the user's interest relevance. Novikov et al. [34] described a systematic approach for complex query processing in the layered system architecture. By means of the adaptive abstract operations based on fuzzy sets, the approach can handle complex, semi-structured and unstructured search and query. For implementation, however, approximate algorithms are necessary. Prasad et al. [31] stated that their data-driven neural fuzzy system can pre-

Table 2
A detailed summarization of the contributions.

Article	Nature of problem	The role of fuzzy set technique	Advantages of using fuzzy set technique	Character of data	Vs			
					Volume	Velocity	Variety	Veracity
[15]	Clustering	FCM	N/A	The cover type data of a forest	581,012 samples, 54 attributes	N/A	N/A	N/A
[16]	Image segmentation	Rough intuitionistic FCM	Improve the performance of FCM	Four kinds of images	6 images	N/A	N/A	N/A
[25]	Community detection	Fuzzy granularity	Model a network in terms of granules	Two real-world social networks	34 nodes with 78 edges; 62 nodes with 159 edges	N/A	N/A	N/A
[27]	Parallel sampling	Handle uncertainties of the boundaries of hyper surfaces	Represent granules by fuzzy boundary, the algorithm maintains identical distribution	Two UCI data sets	Up to 24 million samples by replicating	N/A	N/A	N/A
[28]	Manage fuzzy attributes in queries	Model the vagueness of attributes	Embrace the inherent vagueness in the linked open data cloud	Linked Open Data from DBpedia data repository[a]	Not clear	N/A	N/A	N/A
[29]	Microblog summarization	Fuzzy formal concept analysis	Arrange the tweets into a hierarchical conceptual structure of topics	Four tweet streams of Twitter	Microblog of 477,819 tweets	Data stream	N/A	The existence of noisy & redundant data
[30]	Chaotic cyber-physical systems (CPS)	Fuzzy feedback linearization	Resolve the chaotic status associated with chaotic time series prediction	Simulated data of a chaotic CPS	Not clear	N/A	N/A	N/A
[31]	Time series prediction	Neural fuzzy inference, collaborative fuzzy clustering	Improve the interpretation results; reduce computational complexity	chaotic Mackey glass time-series prediction	Benchmark data sets	N/A	N/A	N/A
[32]	Distilling data	Fuzzy Bayesian inference	An interpretable model with good performance	Generated data for simulation	100 sources and 1000 assertions	N/A	N/A	Fuzzy Bayesian scheme to refine data
[35]	Intelligent transportation	represent the granules by fuzzy ontology	Can support heterogeneous retrieval	Media databases of a city	40,000 media documents	N/A	Include images, videos, audios and text	N/A
[37]	Identify protein-protein interfacial residues	Fuzzy reasoning	Neuro-fuzzy classifiers are fast and accurate than other tested classifiers	Protein structural data	154,993 residues	N/A	N/A	N/A
[45]	Traffic flow state prediction	FCM	The improved FCM integrates the strong local and global search ability	Traffic data of a section of Beijing	One month's traffic data of the section	N/A	N/A	N/A
[51]	Intelligent transportation	Fuzzy reasoning	Alleviate burdens of systems, can be easily distributed	Traffic data	Data from poisson simulation	Constantly simulated data	N/A	N/A
[53]	Heart arrhythmia detection	Fuzzy partition rules based reasoning	Merge and manage three types of knowledge	Free scientific papers from PubMed; experimental data	Very large, but not specified	N/A	N/A	N/A
[54]	Web new mining	Fuzzy reasoning	The evolving fuzzy systems can update in real time, create interpretable models	News articles from New York Times	3500 articles	Sequential updated with new articles	N/A	N/A
[55]	Dimensionality reduction	Linguistic hedges neuro-fuzzy reasoning	Reduce the dimensions, improve classification performances	Four UCI data sets	569, 198, 366 and 215 samples	N/A	N/A	N/A
[56]	Imbalanced classification	Linguistic fuzzy rule-based classification	An interpretable model with reasonable accuracy	24 UCI imbalanced data sets from	Max size: 5 million samples with 41 attributes	Sequential data	Represented by uncertainties of data	Represented by uncertainties of data

(*continued on next page*)

Table 2 (*continued*)

Article	Nature of problem	The role of fuzzy set technique	Advantages of using fuzzy set technique	Character of data	Vs			
					Volume	Velocity	Variety	Veracity
[58]	Fuzzy diagnosis systems	fuzzy rules-based reasoning	A simple and descriptive way for control	Generated cases	2000 cases for training, 5000 cases for learning	N/A	N/A	N/A
[59]	Classification	Linguistic fuzzy rule-based classification	A descriptive model with good accuracy	Six UCI data sets	Max size: 5 million samples with 41 attributes	N/A	N/A	N/A
[60]	State transition in social network communication	Represent the embedded uncertainty and ambiguity	Model embedded uncertainty and ambiguity within the conversation blocks	Social big data	Not clear	Data stream	N/A	N/A
[63]	Health-shocks prediction	Fuzzy linguistic summarization	Provide interpretable linguistic rules to explain the causal factors	Collected data	Collected from 1000 households	N/A	N/A	N/A
[65]	Overcome the int index limitation	Type-2 fuzzy event parallel computing	Incorporate statistical inference with type-2 fuzzy events, retain the specialties of modeling uncertainty	WSS CAPS (Wireless soft-switch call attempts per second)	7 weeks WSS CAPS time series	N/A	N/A	N/A
[67]	Dimensionality reduction	Fuzzy classification	Original data set is transferred into a smaller one with fuzzy classes	A UCI data set	245,057 samples with 3 attributes	N/A	N/A	N/A
[68]	Business intelligence	Type-2 neuro-fuzzy multi-rule based reasoning	Comparing with standard fuzzy approach, it can discover complex knowledge granules	Internet of thing data from a B2C organization	Not clear	Real-time consumer's data	N/A	N/A
[69]	Community detection	Fuzzy granularity	Based on [25], bounds of communities can be viewed by lower and upper approximations	Two real-world social networks	34 nodes with 78 edges; 62 nodes with 159 edges	N/A	N/A	N/A
[71]	Feature selection	Fuzzy granularity	Generate fuzzy information granules in hybrid information systems	6 UCI data sets	Up to 2800 samples; up to 40 features	N/A	Different kinds of data (attributes)	N/A

[a] http://dbpedia.org/.

serve privacy and security because only one half of the samples of the dataset are used for training. In fact, their system may lessen the risk of violating privacy but it is not enough to say preserving privacy.

3.2.1.3. Query refining and data privacy. We have also realized that fuzzy set techniques present a new way to various tasks of big data curation, such as data clean. For instance, Ramachandramurthy et al. [32] improved the quality of data by utilizing the fuzzy Bayesian process, which was implemented by adding fuzzy rules and weights to the traditional Bayesian process. It is an interesting contribution because fuzzy set techniques have been introduced to deal with the veracity of big data. Other studies focus on data preprocessing like datasets management in a distributed environment. Wang and Su [77] considered the problem regarding allocating resources in multiple clouds. Tasks and nodes were divided into different levels, represented by fuzzy linguistic terms dynamically. When a new task arrives, only the nodes, whose level is the same as the task, join in the bid. Here, fuzzy set techniques help speed up data access.

3.2.1.4. New techniques for GrC. Except for the above mentioned literature, some contributions focus on the development of novel GrC techniques by means of fuzzy set techniques.

In [25], a uniform framework for representing social networks is provided based on GrC and fuzzy set techniques. A social network is considered as a set of relations between individuals and their interactions. Closely operative groups formed by individuals resemble the concept of granules and can be modeled by fuzzy sets. Some granules are used for representing a social network. Different from the former studies [78], a group of individuals are treated as an actor, i.e., a granule. In addition, some basic concepts of a social network, such as "between nodes" and "cluster of nodes" have ill-defined boundaries, and thus, they can be represented in terms of fuzzy granules. Consequently, they construct a granule around a node with fuzzy boundary. Finally, some well-known measures for social networks, like the entropy of the network to measure uncertainties arising from fuzziness, are defined in the proposed model as well.

The fuzzy GrC approach for social network was improved by themselves in their later paper [69]. Here, they further consid-

ered the situation where a node can belong to more than one group (community) with different memberships of associations. After getting all communities by the fuzzy GrC approach, the nodes definitely belonging to a community and the others possibly belonging to the community are identified as the lower approximation and boundary regions of a rough set, respectively. Thus, rough sets emerged. The nodes in boundary region are assigned fuzzy membership degrees based on their connectivity with the cores. The advantages of both approaches have been demonstrated by running experiments involving benchmark data sets.

These studies are interesting because they take use of fuzzy set techniques in a novel way. It is worth stressing that the role of fuzzy sets is more than just a tool to represent uncertainties. They contribute to the theory of GrC.

3.2.2. Enhancing the capability of big data application

In what follows, we will analyze how fuzzy set techniques work towards better knowledge discovery and decision making in several different kinds of big data, viz. medical and healthcare big data, transportation big data and social big data.

3.2.2.1. Medical treatment and healthcare. Srivathsam and Yogesh [79] provided a computing framework, Prognotive Computing, for health monitoring using big data techniques. Artificial intelligent algorithms based on fuzzy logic were suggested to compute results of atypical, active and discreet prognosis. But they did not focus on the detail of implementation. On the contrary, the cloud-based health care system proposed by Sundharakumar et al. [52] is more specific. This system encompasses fuzzy logic, neural networks and genetic algorithms to build a knowledge-based system, in which fuzzy logic converts the system into an expert system. Especially, the fuzzy inference system is incorporated with a real time analyzer, namely Storm.[1] Wang et al. [37] paid their attention to the identification of protein-protein interfacial residues by analyzing 154,993 residues. They did not focus on developing new algorithms. The empirical results showed that the neuro-fuzzy classifiers are one of the most efficient tools for handling this type of data. In these studies, the algorithms regarding fuzzy set techniques are the existing ones and the authors did not present any new thinking about the role of fuzziness.

Aiming at a powerful and accurate insight into cardiac arrhythmia, Behadada et al. [53] developed a novel method which can define fuzzy partition rules semi-automatically. As a vital procedure of the method, the rule base is defined and integrated by two parts: the expert rules are expressed as fuzzy linguistic rules by the invited experts; and the induced rules are induced from scientific articles using the same fuzzy linguistic terms. Consequently, both classes of rules can be merged into a sole knowledge base.

Mahmud et al. [63] provided a predictive model for health-shocks based on large-scale health informatics datasets, in which a fuzzy rule summarization technique is used. The role of the adopted fuzzy technique is to present interpretable linguistic rules to stakeholders and then to explain the causal factors that affect health-shocks.

3.2.2.2. Intelligent transportation. Intelligent transportation is a typical application of big data because it involves huge volumes, and sometimes heterogeneous types, of data to be processed in an acceptable time. In order to reach an efficient decision making, like real-time prediction, fuzzy set techniques have been introduced to handle uncertainties and/or solve the problem in a certain level of granules.

The real-time system for traffic flow state identification and prediction, developed by Lu et al. [45], employs a simulated annealing genetic algorithm based FCM for traffic flow quantification. The traffic flow state is an uncertain social phenomenon and is, frequently, characterized by fuzzy descriptions such as “congested” and “uncongested” . Thus, it is more rational and natural to represent the states by membership functions. Consequently, the FCM algorithm is designed and further improved by genetic algorithm to enhance the capability of global searching. In their next study [47], an additional algorithm for real-time traffic flow state correction is developed, where the fuzzy neural network is employed to calculate the error term for correction and overcome the nonlinear mapping problems.

Wang et al. [51] presented an intelligent transportation control system to reduce the averaging waiting time. Fuzzy control rules are considered to avoid unnecessary computational complexities. Another advantage of fuzzy logic is that it can be easily distributed to different kinds of intersections with distinct traffic flow states.

Guo et al. [35] focused on another topic of intelligent transportation, viz. information retrieval for various and heterogeneous media data. Their semantic-based approach extracts semantic fields from heterogeneous documents and datasets and stores them by ontology. The similarity between two ontologies is arduous because semantic fields are imprecise to represent the users' intention. Thus, fuzzy matching is adopted to measure the similar degree between the users' intention and media documents, in which an ontology is mapped to a fuzzy set.

3.2.2.3. Social networks. Big data collected from social networks are usually referred to as social big data. The existing tools for big data, such as techniques, technologies, systems and platforms, have provided better understanding of social big data and intelligent decision making for organizations. It is interesting that fuzzy set techniques have been introduced to process social big data.

Wang et al. [38] tried to understand human behaviors, especially the behaviors of users in mobile networks, to benefit Internet application designing and service expansion. First of all, they granulated the large dataset by clustering users based on the behavior patterns. The FCM algorithm is employed because it allows observations to be loaded in two or more clusters, and finally, the most suitable one can be chosen.

Meng et al. [50] provided some scalable clustering algorithms based on the fuzzy adaptive resonance theory (FART). In the framework of FART, the input pattern is a fuzzy vector. The fuzzy AND operator is adopted for the definitions of choice function, similarity measure, and learning function. By real-time searching, FART can process input samples incrementally. Its performance has been demonstrated by some heterogeneous and large datasets. The theoretical analysis demonstrates that the FART leads to linear computational complexity of the proposed algorithms, and the only one parameter of a cluster is used to form a hyper octagon region for the cluster.

Wei et al. [44] evaluated hotel quality by online comment. Considering the veracity of the comment, the proposed approach improves the traditional fuzzy comprehensive evaluation by importing trust- worthy degree to it. Then, a fuzzy cognitive mapping method was developed to uncover the causal relations among evaluation indices so that the problematic areas of hotel quality can be revealed.

To facilitate the discovery of information of interest from huge amount and often noisy and redundant data of social networks, De Maio et al. [29] presented a microblog summarization algorithm based on temporal extension of fuzzy formal concept analysis. Generally, fuzzy formal concept analysis, defined by introducing fuzzy logic to formal concept analysis, can address fuzzy ontology extraction by considering uncertainties and imprecision in

[1] http://storm-project.net/.

unstructured data and thus can represent fuzzy relations between objects and attributes in a domain [80]. Based on the strength, a temporal extension of fuzzy formal concept analysis was then defined to arrange resources into a timed fuzzy lattice.

Understanding state transition in communication and conversation under social networks is one challenge of social big data. Ghosh et al. [60] provided a fuzzy logic based algorithm to discover the embedded uncertainties and ambiguity involved in the conversation blocks. Here, fuzzy logic is capable because the state transition of conversation is generally vague. Especially, linguistic terms that signify the current state of topic, like "interesting", "not interesting", "motivating" and "deviating", are expressed by fuzzy logic.

In this field, most of the studies have developed new approaches and algorithms based on fuzzy set techniques, except for [38]. Tools like FART, fuzzy formal concept analysis and CWW are considered for reducing the computational complexity, understanding semantics and modeling uncertainties and imprecision.

4. Analysis of the main trend in the use of fuzzy set techniques

Based on the literature review in the above section, we shall present some discussions on the current contributions, including the summary on the emerged techniques, the trend of the roles of these techniques, and the existing limitations.

4.1. A summarization on the emerged techniques

We can see that fuzzy set techniques have been emerged in big data problems frequently in the recent few years. Specifically,

(1) Most of the existing contributions take use of the traditional fuzzy set techniques, such as fuzzy sets, fuzzy reasoning and CWW. This fact may rely on two reasons: (1) The use of fuzzy set techniques usually follows the way that they are used in the traditional artificial intelligent algorithms. Fuzzy set techniques have been integrated within these algorithms. Thus, the role of fuzziness is not improved very much when these algorithms are redesigned for big data. (2) Although fuzzy sets are not convenient enough to represent complex uncertainties, they are very simple and straightforward. Thus, either fuzzy sets or CWW can be found in some novel big data techniques as well.
(2) Several extensions of fuzzy sets have been used in big data processing when fuzzy sets are not efficient enough to model uncertainties and/or represent granules. Till now, we can see several specific type-2 fuzzy sets, including intuitionistic fuzzy sets, type-2 interval fuzzy sets and type-2 fuzzy numbers, are introduced to deal with big data.
(3) Another obvious feature is that fuzzy set techniques are integrated with other decision making tools. Most of the existing contributions focus on the combination of fuzzy sets and rough sets. This phenomenon has become popular since the concept of fuzzy rough set was introduced [81]. Besides, other tools, such as multi-criteria decision making approaches and probabilistic theory, have been applied as well.

4.2. An analysis of the trend on the roles of fuzzy sets

We have analyzed why fuzzy set techniques work when it is applied in handling big data or large data sets. It has been introduced to many distinct fields related to data processing and application. The roles of fuzzy set techniques have moved from improving the existing artificial intelligent algorithms to developing new GrC techniques. The major trends can be summarized as follows:

(1) Artificial intelligent algorithms based on fuzzy set techniques have been adopted most frequently. Because of the advantages of these algorithms, such as efficiently handling uncertainties, this trend will continue. For instance, the FCM will be constantly popular for the purpose of either clustering or reaching a high level of granules.
(2) More and more novel, well-designed, or even revolutionary, fuzzy sets-based algorithms will be donated to figure out specific challenges of big data although fuzzy sets are not convenient enough to represent uncertainties in complex setting. This is mainly because fuzzy sets are the simplest model in the area of fuzzy set techniques. It is commonly acknowledged that good models are usually simple [82]. In order to show further potential power of this type of techniques, more sophisticated methods should be designed to obtain membership functions. In this facet, the CWW techniques including fuzzy linguistic classifiers and new fuzzy linguistic inference systems will become more prevalent than those approaches which reason with membership degrees directly.
(3) The implementation of fuzzy set techniques based algorithms has also drawn the scholars' attention. Especially, if the adopted fuzzy set techniques are a bit complex, then their introduction to big data may raise the computational complexity or lessen the scalability of the system. That is why several contributions focus on the systematic implementation. Current solutions for this issue include parallel implementation and distributing in clouds. It should be pointed out that these possible drawbacks are generally not the characteristics of fuzzy set techniques. In contrast, fuzzification would reduce the computational complexity and enhance the scalability if some shrewd and state-of-the-art techniques are designed under the framework of GrC or if fuzzy set techniques can be used in an innovative way, see [25] for example.
(4) The use of complex models, like type-2 fuzzy sets and fuzzy rough sets, seems to be another explosive trend. The strengths of these techniques are apparent. For example, type-2 fuzzy sets can model more complex cases of uncertainties from more than one source; Fuzzy rough sets combine the advantages of both fuzzy sets and rough sets. Along with the theoretical development of this area, many new and complex models have been emanated. However, their efficiency should be clarified by further deep studies. For more discussions about this issue, please refer to Section 6.
(5) Fuzzy set techniques have benefitted a variety of big data applications. There is no doubt that more and more big data problems will be (partially) solved by the theory. Many vital and specific problems have not been discussed, such as business, management, social security. Although medical big data have been processed by fuzzy sets-based techniques, it is far from enough. For instance, the prediction of a specific disease based on the existing medical big data and/or social networks is a crucial issue in the big data age, and could reach a better solution if fuzzy set techniques are adopted for modeling various and complex uncertainties of the problem. It is rational to forecast that fuzzy set techniques can provide a satisfactory solution for most of the big data problems because: (i) uncertainties always exist in immense volume of data as well as the transformation of various types of data; (ii) current computational capability cannot process such enormous amount of data, and then, fuzzy set techniques can mine necessary and useful knowledge if it is used as a GrC tool.

However, although we can see some interesting results from the above reviewed literature, it can also be realized that there are some vital problems regarding big data and the use of fuzzy set techniques. For example, as shown in Table 2, more than one half articles focus only on the volume of data. In this case, the data sets are not really big. That is why we use the term "large data sets" in some places of this section. For the convenience of high quality developments in the future, we need to discuss the issue and put it in a right perspective. We will present a critical discussion in the next subsection.

4.3. A critical analysis

Maybe it is not surprised that, as can be seen in Table 2, many of the contributions have focused only on the volume. In fact, we deal with large data sets but not real big data if only volumes are focused on. Going through this kind of articles, we realize that their major contribution is the development of new algorithms to handle large data sets within acceptable time and/or spaces and mine valuable knowledge. Due to the nature of specific problems, maybe it is not necessary to focus on all the 5Vs at the same time, but it is definitely not enough to concentrate one V when processing big data. Alternatively, considering volumes might be enough in some problems, but we cannot use the phrase "big data" in the cases. Although we have merely analyzed the literature related to fuzzy set techniques, it should be noticed that the phenomenon of emphasizing the one V is very common in the papers whose topic is big data. In sum, it seems that many authors have overused "big data" in their studies.

Other Vs, like velocity, variety and veracity, should also be included in big data processing. Till now, velocity has been considered in several studies related to social networks, intelligent transportation and internet of thing. However, to facilitate the comparison with benchmark algorithms, most of these studies simulated data stream by the existing data sets. Variety has been considered in intelligent transportation systems due to the nature of the problem. But it is a pity that we cannot find any discussion about variety in social networks. Maybe this is because the approach of handling variety by fuzzy set techniques has not been developed. In addition, veracity is crucial for the cases when data are collected by sensors, smart phones and other terminals. But we can find only one study which focused on distilling quality of data.

Especially, value has not been discussed associated with fuzzy set techniques. It is out of question that data are valuable. However, this V also emphasizes the extremely low density of value. This is somewhat like the idea of pattern recognition with imbalanced data sets. For instance, when mining social networks, it is important to figure out the leading user whose thinking and comments affect a large amount of users; when seeking knowledge from a collection of scientific papers, the most valuable knowledge may come from one or a small amount of papers, or even one picture of a paper. This is generally a big challenge of analyzing big data because it may be necessary to think about our problems in a novel way.

When facing big data, we have to focus on more Vs rather than only volume. Take social big data for example, the period between data generation and obsolete is very short, and thus, velocity should be considered. Various data including texts, images, audios and videos are involved, therefore, variety is natural. What users said maybe cheat and cannot be trusted, so veracity is also a big problem, and as mentioned above, new algorithms are necessary to handle the low density of value. However, current techniques and technologies do not match the requirements of the 5Vs. Challenges exist in almost every phase of big data processing. It can be anticipated that some of the current challenges can be partially solved by fuzzy set techniques. In the next section, we will present a summary on some of the current challenges and illustrate their relevance to fuzzy set techniques.

5. Current big data challenges potentially related to fuzzy set techniques

If big data are gold ores, then we do not have enough capabilities to explore them. Currently, technical and non-technical challenges exist in almost each phase of big data processing. Due to the focus of this review, we shall concentrate on some of the challenges which might be related to uncertainties modeling and GrC. Some other challenges, such as the system imbalance about CPU-heavy but I/O-poor [12] and costs and risks of performing data analytics [83], are not referred to in the section. For more information about the existing challenges, please refer to [2,14,83-85].

5.1. Challenges raised by basic characteristics of big data

Major challenges are caused by the commonly acknowledged characteristics of big data. For instance, the volume challenges scalability of data processing algorithms. In this section, we focus on challenges caused by the characteristics.

5.1.1. Challenges caused by huge volume

Volume is generally the first characteristic when considering big data. The major challenge caused by volume is scalability, especially, the scalability to benefit from distributed processing infrastructure. Scalability involves in several aspects including data store, data curation and computability.

Firstly, the huge volume leads to store big data in the distributed environments like clouds. However, this raises challenges regarding data transformation and data accessing. In data transformation, the volume of communication could be very large whereas the network bandwidth capacity has limitations. Secondly, quick allocation and accessing of specific data in real-time processing framework are also challenging. Thus, we cannot store data using the same strategy as it is used in common distributed systems. Based on the paradigm of GrC, one solution might be to represent and reconstruct the data at distinct granular levels based on the problems and goals in hand. This would be helpful to decrease the volume of data and thus cut down the complexity of computation and communication. In addition, the volume of data, which are necessary to be transformed, can be reduced by a granular representation.

Parallel computing is frequently considered as a systematic implementation solution of big data applications, usually associated with a distributed system. But the biased view of the partial data of each different node or cloud often leads to uncertainties and biased decisions. This challenge requests our capability of enabling information exchanges among nodes and providing fusion mechanisms to ensure that all distributed data sources can work together to achieve a global optimization goal. Fuzzy set techniques would be an excellent technique to fuse local and temporary results because of their capability of handling uncertainties.

The volume also causes another challenge regarding data visualization due to the large size and/or high dimension of big data. Granulation has been considered naturally in data visualization. It is criticized that most of the current big data visualization tools suffer from poor performances in functionalities, scalability and response time [2]. We may need to rethink the strategy for visualization, such as a new framework that can model and characterize the evolution of uncertain information.

5.1.2. Challenges caused by high velocity

Velocity indicates the frequency of big data generating as well as the necessity of outputting results in real-time, especially in

the paradigm of stream processing. This feature implies some challenges, such as: (1) the ability of focusing on and ranking the relevant data; (2) the ability of tackling data in an incremental and sequential way; and (3) the ability of culling, evolving and honing on relevant background knowledge [86]. In addition, incremental algorithms cannot solve this point fundamentally though their scalability is generally good. Non-deterministic algorithm theory may be more suitable for real time big data analysis [85]. It can be expected that this challenge would lead to the swerve of developments of software architecture and sophisticated algorithms to cloud computing. The lambda architecture [87], which synthesizes and extends the paradigm of batch processing and stream processing, has demonstrated its efficiency gradually in recent years. If GrC techniques are well deployed to lessen the volume of data, then the challenges of velocity would be alleviated a lot.

5.1.3. Challenges caused by variety

Variety highlights the different types of data schemas and different sources of information. This characteristic leads to another big challenge: the ability to integrate and interoperate with structured, semi-structured and even unstructured heterogeneous data.

To manage big data, NoSQL databases are usually considered as a solution. Some big data platforms, such as Hbase, utilize NoSQL to break and transcend the rigidity of the traditional relational database schemas. Data storage and management are separated in NoSQL systems. The storage part possesses good scalability and high performances. Low-level access mechanism is provided through the management part so that data management can be implemented in application layer. In addition, most NoSQL systems are schema-free. Thus, NoSQL systems are more flexible than SQL systems, and more convenient to update application developments and deployments. However, NoSQL is less reliable and more complex than SQL. Thus, some other platforms, such as SQLstream, still adopt SQL in the database system.

The various types and complex structures of data also cause data complexity and computational complexity. There is no acknowledged effective and efficient model to handle heterogeneous big data. It is not surprised that the traditional data analysis techniques are difficult or even disabled to tackle big data because the laws of distribution and association relationship, as well as the domain-oriented processing methods of big data have not been fully understood yet. In this sense, we may face a barrier of formulating and representing the complexity of big data quantitatively. Fuzzy set techniques could be considered to transform and fuse heterogeneous data. The powerful ability of handling uncertainties would also be helpful in understanding the semantics of complex data and measuring the complexity of data.

5.1.4. Challenges caused by veracity

Veracity corresponds to what extent the data can be trusted. This issue refers to uncertain, incomplete and false data. For instance, some comments of online transactions may be fabricated artificially and thus do not reflect the real quality of products and services. The following challenges exist when determining the veracity of big data: (1) the ability to detect anomalies and inconsistencies in data sources; (2) the ability to collect evidence and exploit conflict resolution strategies in decision making [86].

The most challenging task is to recognize the false and baseless data from the sources. Usually, this can be considered as pattern recognition based on semantic understanding. Apart from the issue of volume, this challenge would become much severer if various heterogeneous and distributed data sources are involved. In some big data applications, such as social big data, we should understand the semantic association and relationship among texts, images, audios and videos. Till now, we have no clear idea about how to efficiently bridge the semantic gaps of various heterogeneous data.

Uncertain data can be produced in some specific domains, such as data generated by GPS equipment. The values of these data are not deterministic but subject to some random/error distributions. The errors could be created subjectively in some other cases for the purpose of data privacy. Most existing data mining algorithms cannot handle sample distributions directly. Current solutions are to estimate the parameters of distributions, like the mean and variance values, and then to build intelligent models [88].

Incomplete data of some samples are also inevitable, such as the missing data field caused by the malfunction of a sensor node. In this case, data imputation is considered as the strategy to handle missing values. Some data imputation solutions ignore data fields with missing values; others predict possible values for each missing field. The rationality and efficiency of these imputation methods should be verified by further big data applications.

Several soft computing approaches based on fuzzy sets and their extensions have been widely investigated and developed. Some approaches own good performances of handling various kinds of uncertain data and incomplete data. Furthermore, sophisticated fuzzy set techniques would be possible to help understand the semantics of data, especially texts, video and audio, so that the veracity of data can be discovered by considering multiple data sources and their relationships. However, due to the natural of fuzzy logic, one cannot handle probabilistic uncertain directly. In addition, real big data applications should be the only way to demonstrate whether the approaches are effective or falsifiable, because most of the existing soft computing approaches are developed from the theoretical view.

5.1.5. Challenges caused by density of value

The extremely low density of value challenges our ability to mine useful knowledge from huge volume and high velocity of heterogeneous data. The main challenges of extracting value are: (1) the ability to acquire, apply and integrate knowledge from data; (2) the ability to learn and apply models for decision making (such as classification, prediction and personalization).

We have realized that big data do not imply that we have enough data. In some big data applications, the number of valuable samples may be sparse. One cannot mine clear trends or distributions for deriving reliable conclusions. This case is similar to the classification or clustering in imbalance data sets. But it would turn into worse problems if data are in a high dimensional space. High dimensional spare data (such as more than 1000 dimensions) significantly challenge the difficulty and reliability of most of the current machine learning and data mining algorithms derived from the data. Possible solutions include dimension reduction to reduce the data dimensions, sampling to change the distribution of patterns to decrease the data scarcity and cost-based algorithms to emphasize the role of sparse patterns. As have been reviewed in this paper, some of the possible solutions have benefitted from fuzzy set techniques, mainly because of the ability of handling uncertainties caused by not enough information.

5.2. Other related challenges

Besides, other challenges with potential relation to fuzzy set techniques include the challenges caused by data security and the necessity of understanding big data.

5.2.1. Challenges related to data security

Data security refers to in which way data can be accessed, transformed and used. It involves both technical and ethical issues. The ethical issue surpasses the aim of this review. The technical issue leads to challenges regarding protecting private and sensitive data so that all data can be used in a normal and legal way. For example, information related to medical records and banking

transactions is sensitive and not proper for sharing and transmissions. It calls for the strategy and protocol to control and limit the activities on data. The existing strategies contain anonymization, fuzzification of private and sensitive data, designing secured certification mechanisms. Anonymization is definitely the intuitive and straight path. But it still suffers the risk of illegal access. The primitive, namely format-preserving encryption [89-90], could partially meet the challenge. In addition, the single value of private information can be fuzzified and transformed into uncertain data with some special distributions. However, as have mentioned hereinabove, this strategy leads to another challenge.

5.2.2. Challenges related to understandable big data

The aim of understanding big data falls into two facets: (1) data generated by human and disseminated through web tools should be understood for the purpose of security, business intelligence, etc.; (2) connected or similar objects should be recognized so that they can be merged and integrated. Definitely, the ability of processing big data would be improved a lot if we can leverage on semantics of big data.

We had realized that the texts, including word-of-mouth text and comments, are valuable for many applications. To understand the semantics of texts, especially the sentiments, a large amount of contributions have devoted many useful techniques, see [91] for a detailed review of this field. Fuzzy set techniques have been employed in this field. Overall, mining semantics of texts involve various topics, such as reasoning, co-reference resolution, entity linking, information extraction, consolidation, paraphrase resolution and ontology alignment [92]. However, semantic mining becomes exceedingly complex in the circumstance of big data. The scalability of the existing algorithms should be enhanced.

Understanding the semantics of texts is not the end. We may need to understand the intrinsic semantic associations between types of heterogeneous data. In a community, the users may express their feelings about a film by means of texts, pictures or even audios. Mining these complex semantic associations will bring us to a new milestone of big data system performances. However, till now, it is quite challenging to describe semantic features efficiently and to build semantic association models. Thus, we are far from bridging the semantic gaps between various types of heterogeneous data.

6. Possible directions and opportunities of using fuzzy set techniques

Based on the existing contributions and current challenges, we will predict some possible directions of using fuzzy set techniques in big data as well as possible opportunities and trends in the coming years. First of all, we would like to mention some principles to serve as the guideline of future investigations.

6.1. Some principles of developing novel techniques

Fuzzy set techniques have been introduced to handle uncertainties of data and tackle uncertain outputs of artificial intelligent algorithms. In most existing contributions, the role of fuzzy set techniques does not change a lot. Although a few new strategies have been introduced, it is far from enough to meet the big challenges. We may need to reconsider the way that we use fuzziness, and then develop completely new techniques to reach much more excellent performances in big data processing. In this circumstance, we propose the following generic principles. Note that it is not necessary for one technique to fit all the principles.

Principle 1: New techniques should be developed based on the framework of GrC. The advantages of GrC include efficiently handling uncertainties, sharply reducing the volume of data, recognizing data from different levels of granules, to name but a few. If we want to benefit from GrC, then the straightforward and intuitive way is to follow the framework and paradigm of GrC. The vital points are defining granules and granulation. For more guides and research map of GrC, please refer to [4,23,74].

Principle 2: The motivation of new techniques should be application-oriented. It has been widely acknowledged that no size can fits all [93]. When developing new techniques, it is rational and reasonable to start with a specific big data application. Due to the complex feature of distinct data-intensive applications, the only rule to verify new techniques is that if the faced problems can be solved properly. Till now, it is hopeless to reach a generic theory to analyze and mine big data.

Principle 3: New techniques should possess satisfactory scalability. This is the common requirement of handling big data, especially when the new technique is about to implement in a big data platform. In practice, scalability might be a problem if the technique is slightly improved by the traditional machine learning algorithms.

6.2. Possible trends and opportunities

Using fuzzy set techniques is an emerging trend of processing big data. Fuzzy set techniques are taken into account by two separated strategies. The first one is to extend the existing artificial intelligent algorithms related to fuzzy set techniques to the big data applications. In this case, more attention should be paid to the implementation of the existing algorithms, and the role of fuzzy set techniques is generally the same as it was before. The second strategy develops novel GrC techniques to fit specific big data applications. It seems to be more interesting because it develops completely new techniques to deal with certain features of specific big data problems. One representative contribution is the fuzzy granular social networks proposed in [25]. Both strategies will continue in the near future. However, due to the necessity and challenges of handling big data, we anticipate that we should evolutionarily change our way of processing data so as to accelerate our ability of analyzing big data. Therefore, the second strategy may bring us a collection of exciting and excellent techniques. By the way, the role of fuzzy set techniques will be changed a lot, its strength and force will be further exploited.

In this sense, we list some possible further directions as follows:

6.2.1. Using more sophisticated extensions of fuzzy sets

Various techniques have been introduced to process big data, including fuzzy sets (and their special case, namely linguistic fuzzy sets), interval-valued fuzzy sets, intuitionistic fuzzy sets, type-2 fuzzy sets. Recently, the extensions of fuzzy sets have been developed quickly. Several other extensions, such as hesitant fuzzy sets [94], fuzzy multisets [95], complex fuzzy sets [96], neutrosophic sets [97], have been developed to model uncertainties from different perspectives. Their effectiveness in multi-criteria decision making, information fusion and soft computing has been clarified by scholars. One may expect that these extensions can be used in the presence of big data if they are considered as the GrC tools. For instance, hesitant fuzzy sets focus on the situation when the decision makers hesitate among a set of possible membership degrees. This is similar to the situation in ensemble learning. When the outputs of each component classifiers are ready (and can be interpreted by the membership degree), we hesitate among several possible degrees of the given sample belonging to one class and seek for a useful fusion method to determine the class label, see [98] for example.

On the other hand, the use of linguistic fuzzy sets (linguistic terms) is also very popular in fuzzy reasoning and fuzzy linguis-

tic classifiers. However, current consideration of linguistic terms is also quite straightforward. To deal with more complex uncertainties, we may need multi-granularity linguistic term sets [99], hesitant fuzzy linguistic term sets [100-101], and probabilistic linguistic term sets [102-103]. Multi-granularity linguistic term sets provide a collection of linguistic terms with distinct granular levels and different semantics. Hesitant fuzzy linguistic term sets enable us to use a set of continuous linguistic terms to represent uncertainty. They could be the very good GrC tools due to their natures of representing granules. In addition, the traditional fuzzy reasoning takes use of the classical maximum and minimum operators in fuzzy sets, which leads to the loss of information during the reasoning procedure. Two accurate linguistic computational models, i.e., the virtual linguistic model [104] and the 2-tuple linguistic model [105], could be alternatives to improve the performance of linguistic terms based fuzzy reasoning.

However, although these techniques seem to be powerful to model uncertainties, we should notice that their performances in big data should be verified by real big data applications. As stated in Principle 2, the rationality and effectiveness of the abovementioned techniques is doubtful unless they are demonstrated by solving at least one kind of big data problems.

6.2.2. Integration with other GrC tools

Apart from fuzzy sets, the tools for granular representation include rough sets, interval analysis and so on. Due to their natures, all tools can perform well in the GrC framework both in isolation and integration. In the presence of big data, the integration will be the next emerging trend because the combination of strengths of different tools would be more convenient to fight against the existing difficulties.

Rough sets can effectively acquire hided knowledge by their core concepts: the upper approximation and lower approximation based on indistinguishability relations. They can granulate a collection of observations into information granules by selecting attributes and values of each attribute, see [106-107] for examples. The optimized coarser and more abstract information granules can be derived if the number of attributes and the discrete values of each attribute are lager. Till now, we have found that the integration of fuzzy sets and rough sets, either using them in distinct procedures or using rough fuzzy sets, has shown the wonderful capabilities in big data processing. However, we have to say that this is just the beginning. Because most contributions with this strategy only focus on the improvement of the existing algorithms, except for [69]. It is imperative and compelling to develop more sophisticated approaches for different big data applications.

The combination of fuzzy sets and neural networks, known as neural fuzzy networks, has appeared since the last decade, and has been revised to suit big data applications. But the existing versions should be improved to enhance the scalability and reduce the computational complexity in big data environment. Interval analysis has been widely used for the extension of fuzzy sets, which results in interval-valued fuzzy sets, interval-valued intuitionistic fuzzy sets, interval-valued linguistic terms, etc. The capabilities of handling uncertainties of these extensions have been demonstrated. But their effectiveness should be further verified if they serve as a GrC tool for big data.

Finally, we note that the integration of different GrC tools could be diversified. Firstly, fuzzy sets would not be necessary to involve in. The rough neutral networks [108] could also be considered if required. Secondly, more than two tools can be integrated for one special big data application. For instance, we may employ rough sets to define granules of big data, and then utilize interval-valued linguistic terms to improve the ability of handling uncertainties in rule-based classification systems.

Whenever developing and extending GrC tools for big data, it would be sensible and shrewd to exert the advantages of big data platforms adequately. Diverse algorithms related to the idea of GrC have been developed using these platforms. Gu et al. [109] presented a comparable analysis on some popular open-source platforms. Huang et al. [110] implemented the rough set theory by Spark. Some GrC techniques, such as clustering algorithms, were suggested to be implemented in Spark [111] or Hadoop [112]. Bahrami and Singhal [113] summarized the information granularity and GrC infrastructures which could be provided by cloud computing. It could be speculated that another data processing technique, i.e., GPU computing or multi-GPU computing, may also offer some opportunities for new technique design [114-115].

6.3. Other potential tools for big data processing

In this subsection, we present a perspective description on other tools related to fuzzy set techniques that might be employed in the presence of big data.

6.3.1. Information fusion under uncertainties

Information fusion is a generic issue of data analysis. The fusion processes can be considered in low, intermediate and high levels. The low level of information fusion combines raw data of different (and usually distributed) sources to new raw data. Processes of the other two levels produce new information and knowledge at different degrees. In recent years, the theory and methods related to the intermediate and high levels of information fusion have been developed quickly. Especially, information fusion under fuzzy uncertainty fascinates many attentions. Based on specific fuzzy setting, new results of information fusion include aggregation functions, computing with complex linguistic expressions and so on.

When it comes to big data, information fusion at the low level refers to the fusion of data with complex structures in the distributed storages; and information fusion at other levels considers the derivation of new knowledge, associated with certain big data technique and technology. The recently developed theory and methods would help the intermediate and high levels of big data information fusion. For example, in the framework of ensemble learning, information fusion techniques can help fuse the outputs of individual classifiers which contain uncertainties due to the nature of classification. Similarly, in a distributed setting, we may need to fuse heterogeneous data of local datasets or all datasets; or if each node works in isolation, then we have to fuse all the local results to obtain a global result.

The role of information fusion has a close connection with the purpose of GrC. In fact, GrC begins with information representational models, and information fusion focuses on the integration, combination and synthesis of data. Therefore, the marriage of information fusion and GrC would produce great achievements of knowledge mining in big data.

6.3.2. Fuzzy fractal theory

The fractal theory was initially presented by Mandelbort [116] to describe a special category of geometrical objects. A fractal is a geometrical shape whose parts are similar to the whole in some way. Two core concepts of fractal theory are self-similarity and fractal dimension. Self-similarity is a global property of a fractal to describe how a part can be scaled up to a whole. From a generic view, self-similarity comes from not only shapes but also function and information of an object. To quantify this phenomenon, fractal dimension is defined to represent the structural complexity of a fractal.

However, the self-similarity of objects or phenomena may be not absolute but approximate. In this case, fuzzy sets have been introduced to fractal theory to describe approximately similarity,

and the membership functions are used for the calculation of fuzzy fractal dimensions. Till now, fuzzy fractal theory has been applied in many theoretical fields of decision making and prediction, such as time series prediction, fuzzy control, and neural networks [117-118]. The resultant tools perform pretty well in various disciplines including medical, physics, and computer sciences.

When it comes to big data, fractal theory (or fuzzy fractal theory) owns more or less potential to improve the current status. In fact, big data sets possess several characteristics of a fractal. Take the community in social networks for example, if we consider the distinct levels of communities, then the roles of individuals in a refined community are similar to the roles of these refined communities in a coarse community. A fractal dimension defined by a special way may be used for the measurement of the structural complexity of the whole community. In addition, we can see from the example that the determination of fractal dimensions is inherently similar to the use of information granules [119]. Therefore, fractal theory can be considered as another GrC tool. It can be anticipated that fractal theory would be an alternative technique for some big data problems, especially involving self-similarity, if it is adopted under the framework of GrC, or if it is considered associated with fuzzy set techniques and/or other GrC techniques.

However, as stated in Principle 2, the real impact of fuzzy fractal theory on big data can only be justified by practical big data problems. We have pointed out some conceptual relationships between the theory and some big data problems. Till now, no study focuses on the issue. Thus, the above-mentioned potential is waiting for clarification.

7. Conclusions

Nowadays, big data are present in almost everywhere of our daily life including social networks, online and offline transactions, medical records, and sensors. An immense volume of heterogeneous data can be generated at exponential rate. The capabilities of handling big data are vital to many scientific and engineering applications. Fuzzy set techniques play an important role in processing big data as they cannot only model uncertainties of both data sources and results of algorithms but also represent information granules to reduce the volume of data and enhance the scalability of the existing algorithms. There are various contributions, which involve fuzzy set techniques in big data applications. Some of them improve the existing fuzzy set-based algorithms to suit big data environment while others help develop new techniques suitable to handle specific big data problems.

We have gone through the recent contributions within this field from two perspectives: the utilized fuzzy set techniques and the focused big data theoretical and applied problems. We have discussed that many current big data challenges would be solved or alleviated if fuzzy set techniques can be used appropriately. To reach the goal, we have presented some principles of developing novel big data techniques based on fuzzy set techniques. Associated with some possible opportunities inferred from the current challenges and trends, we have provided some featured guidelines for further developments. According to the review, we can draw the following conclusions:

(1) The roles of fuzzy set techniques in the existing contributions are two-folds: (i) handing uncertainties in the procedures; and (ii) serving as a kind of GrC tool. Many theoretical and applied fields have benefitted from fuzzy set techniques. One can anticipate that in a near future, current trends will continue.
(2) Existing contributions overuse the phrase “big data”. Many of them focus only on the large data sets rather than big data. When dealing with big data, several Vs should be taken into account based on the specific problems in hand.
(3) Challenges exist in almost everywhere of big data processing and application. Many of the challenges, such as the ones caused by the Vs, own inherent links to fuzzy set techniques. It can be anticipated that some challenges could be (partially) solved or alleviated by fuzzy set techniques.

Based on the review and these conclusions, we can also summarize some open problems, which might be solved and clarified in the near future:

(1) How to take more than one Vs into account in big data problems should be paid more attention. This drives us to think about real big data problems rather than only on facet of them. Maybe it is not necessary to consider all the 5 Vs in one problem, but big data definitely include more than volumes. Especially, fuzzy set techniques may be a good choice of representing information granules. How to design sophisticated GrC approaches for big data would be the next big and interesting issue.
(2) The listed potential opportunities should be tested by big data problems. Based on the challenges and the trends, several possible tools have been analyzed. It should be justified by practical problems so as to clarify that to what extent these tools, or some of these tools, can match the requirements of the challenges.
(3) The application fields of GrC techniques could be extended. As the framework of knowledge granulation and knowledge representation, GrC techniques are possible to be combined with other big data techniques such as machine learning and deep learning. In this case, GrC would act as the data preprocessing tool.

Acknowledgments

The authors would like to thank the Editor-in-Chief and four anonymous reviewers for their insightful and constructive commendations that have led to an improved version of this paper. The work was supported by the National Natural Science Foundation of China (Nos. 61273209, 71571123, 71601092), the Key University Science Research Project of Jiangsu Province (No. 16KJA520002).

References

[1] E. Miller, Community cleverness required, Nature 455 (2008) 1.
[2] C.P. Chen, C.Y. Zhang, Data-intensive applications, challenges, techniques and technologies: a survey on big data, Inf. Sci. 275 (2014) 314–347.
[3] E. Sejdic, Adapt current tools for handling big data, Nature 507 (2014) 306.
[4] S.K. Pal, S.K. Meher, A. Skowron, Data science, big data and granular mining, Pattern. Recogn. Lett. 67 (2015) 109–112.
[5] D. Laney, 3D data management: controlling data volume, velocity and variety, META Group Res. Note 6 (2001).
[6] D. Butler, When Google got flu wrong, Nature 494 (2013) 155–156.
[7] D. Lazer, R. Kennedy, G. King, A. Vespignani, The parable of Google flu: traps in big data analysis, Science 343 (2014) 1203–1205.
[8] D.A. Broniatowski, M.J. Paul, M. Dredze, Twitter: big data opportunities, Science 345 (2014) 148.
[9] J.U. Adams, Big hopes for big data, Nature 527 (2015) S108–S109.
[10] L. Einav, J. Levin, Economics in the age of big data, Science 346 (2014) 715-+.
[11] J. Blumenstock, G. Cadamuro, R. On, Predicting poverty and wealth from mobile phone metadata, Science 350 (2015) 1073–1076.
[12] A.J. Hey, S. Tansley, K.M. Tolle, The fourth paradigm: data-intensive scientific discovery, Microsoft Research Redmond, WA, 2009.
[13] H.G. Miller, P. Mork, From data to decisions: a value chain for big data, IT Prof 15 (2013) 57–59.
[14] R. Casado, M. Younas, Emerging trends and technologies in big data processing, Concurr. Comp.-Pract. E 27 (2015) 2078–2091.
[15] S.A. Ludwig, MapReduce-based fuzzy c-means clustering algorithm: implementation and scalability, Int. J. Mach. Learn. Cyber. 6 (2015) 923–934.
[16] B.K. Tripathy, D. Mittal, Hadoop based uncertain possibilistic kernelized c-means algorithms for image segmentation and a comparative analysis, Appl. Soft Comput. 46 (2016) 886–923.

[17] H. Esmaeilzadeh, E. Blem, R. St Amant, K. Sankaralingam, D. Burger, Power challenges may end the multicore era, Commun. ACM 56 (2013) 93–102.
[18] A. Acquisti, L. Brandimarte, G. Loewenstein, Privacy and human behavior in the age of information, Science 347 (2015) 509–514.
[19] K. Wren, Big data, big questions, Science 344 (2014) 982–983.
[20] S. Wilson, Big data held to privacy laws, too, Nature 519 (2015) 414.
[21] P.E. Bourne, J.R. Lorsch, E.D. Green, Sustaining the big-data ecosystem, Nature 527 (2015) S16–S17.
[22] T. Heinis, Approximation aids handling of big data, Nature 515 (2014) 198.
[23] W. Pedrycz, Granular Computing: Analysis and Design of Intelligent Systems, CRC Press/Francis Taylor, Boca Raton, 2013.
[24] L.A. Zadeh, Fuzzy sets, Inf. Control 8 (1965) 338–353.
[25] S. Kundu, S.K. Pal, FGSN: fuzzy granular social networks - model and applications, Inf. Sci. 314 (2015) 100–117.
[26] S.F. Su, W. Pedrycz, T.P. Hong, F.D.T. De Carvalho, Special issue on granular/symbolic data processing, IEEE Trans. Cybernet 46 (2016) 342–343.
[27] Q. He, H.C. Wang, F.Z. Zhuang, T.F. Shang, Z.Z. Shi, Parallel sampling from big data with uncertainty distribution, Fuzzy Sets Syst 258 (2015) 117–133.
[28] D.J. Lewis, T.P. Martin, Managing vagueness with fuzzy in hierarchical big data, in: A. Roy, P. Angelov, A. Alimi, K. Venayagamoorthy, T. Trafalis (Eds.), INNS Conference on Big Data 2015 Program, 2015, pp. 19–28.
[29] C. De Maio, G. Fenza, V. Loia, M. Parente, Time aware knowledge extraction for microblog summarization on Twitter, Inf. Fusion 28 (2016) 60–74.
[30] L. Liu, S.L. Zhao, Z.L. Yu, H.J. Dai, A big data inspired chaotic solution for fuzzy feedback linearization model in cyber-physical systems, Ad Hoc Netw 35 (2015) 97–104.
[31] M. Prasad, Y.Y. Lin, C.T. Lin, M.J. Er, O.K. Prasad, A new data-driven neural fuzzy system with collaborative fuzzy clustering mechanism, Neurocomputing 167 (2015) 558–568.
[32] S. Ramachandramurthy, S. Subramaniam, C. Ramasamy, Distilling big data: refining quality information in the era of yottabytes, Sci. World J. (2015) 1–9.
[33] T.C. Havens, J.C. Bezdek, C. Leckie, L.O. Hall, M. Palaniswami, Fuzzy c-means algorithms for very large data, IEEE Trans. Fuzzy Syst. 20 (2012) 1130–1146.
[34] B. Novikov, N. Vassilieva, A. Yarygina, Querying big data, in: International Conference on Computer Systems and Technologies, 2012, pp. 1–10.
[35] K.H. Guo, R.F. Zhang, L. Kuang, TMR: Towards an efficient semantic-based heterogeneous transportation media big data retrieval, Neurocomputing 181 (2016) 122–131.
[36] M. Sato-Ilic, Multidimensional joint scale and cluster analysis, in: C.H. Dagli (Ed.), Complex Adaptive Systems, 2015, 2015, pp. 11–17.
[37] D.D. Wang, W.Q. Zhou, H. Yan, Mining of protein-protein interfacial residues from massive protein sequential and spatial data, Fuzzy Sets Syst 258 (2015) 101–116.
[38] Z.H. Wang, L. Tu, Z. Guo, L.T. Yang, B.X. Huang, Analysis of user behaviors by mining large network data sets, Future Gener. Comput. Syst. 37 (2014) 429–437.
[39] M. Sato-Ilic, P. Ilic, On a multidimensional cluster scaling, Proc. Comput. Sci. 36 (2014) 278–284.
[40] M. Sato-Ilic, P. Ilic, Fuzzy dissimilarity based multidimensional scaling and its application to collaborative learning data, in: C.H. Dagli (Ed.), Complex Adaptive Systems: Emerging Technologies for Evolving Systems: Socio-Technical, Cyber and Big Data, 2013, pp. 490–495.
[41] E. Bou-Harb, M. Debbabi, C. Assi, A novel cyber security capability: Inferring Internet-scale infections by correlating malware and probing activities, Comput. Netw. 94 (2016) 327–343.
[42] Y. Cai, Q. Li, H.R. Xie, H.Q. Min, Exploring personalized searches using tag-based user profiles and resource profiles in folksonomy, Neural Netw. 58 (2014) 98–110.
[43] R. Dutta, A. Morshed, J. Aryal, C. D'Este, A. Das, Development of an intelligent environmental knowledge system for sustainable agricultural decision support, Environ. Modell. Softw. 52 (2014) 264–272.
[44] X. Wei, X. Luo, Q. Li, J. Zhang, Online comment-based hotel quality automatic assessment using improved fuzzy comprehensive evaluation and fuzzy cognitive map, IEEE Trans. Fuzzy Syst. 23 (2015) 72–84.
[45] H.P. Lu, Z.Y. Sun, W.C. Qu, Big data-driven based real-time traffic flow state identification and prediction, Discrete. Dyn. Nat. Soc. (2015) 1–11.
[46] P. Su, C.J. Shang, Q. Shen, A hierarchical fuzzy cluster ensemble approach and its application to big data clustering, J. Intell. Fuzzy Syst. 28 (2015) 2409–2421.
[47] H.P. Lu, Z.Y. Sun, W.C. Qu, L. Wang, Real-time corrected traffic correlation model for traffic flow forecasting, Math. Probl. Eng. (2015) 1–7.
[48] Z.L. Liu, J.W. Li, J. Li, C.F. Jia, J. Yang, K. Yuan, SQL-based fuzzy query mechanism over encrypted database, Int. J. Data Warehous. 10 (2014) 71–87.
[49] Z. Peng, J. Peng, W. Zhao, Z.G. Chen, Research on FCM and NHL based high order mining driven by big data, Math. Probl. Eng. 2015 (2015) 1–7.
[50] L. Meng, A.H. Tan, D.C. Wunsch, Adaptive scaling of cluster boundaries for large-scale social media data clustering, IEEE Trans. Neur. Net. Lear. (2015) 1–14.
[51] C. Wang, X. Li, X.H. Zhou, A.L. Wang, N. Nedjah, Soft computing in big data intelligent transportation systems, Appl. Soft Comput. 38 (2016) 1099–1108.
[52] K.B. Sundharakumar, S. Dhivya, S. Mohanavalli, R.V. Chander, Cloud based fuzzy healthcare system, in: V. Vijayakumar, V. Neelanarayanan (Eds.), Big Data, Cloud and Computing Challenges, 2015, pp. 143–148.
[53] O. Behadada, M. Trovati, M.A. Chikh, N. Bessis, Big data-based extraction of fuzzy partition rules for heart arrhythmia detection: a semi-automated approach, Concurr. Comp-Pract. E. 28 (2016) 360–373.
[54] J.A. Iglesias, A. Tiemblo, A. Ledezma, A. Sanchis, Web news mining in an evolving framework, Inf. Fusion 28 (2016) 90–98.
[55] A.T. Azar, A.E. Hassanien, Dimensionality reduction of medical big data using neural-fuzzy classifier, Soft Comput 19 (2015) 1115–1127.
[56] V. Lopez, S. del Rio, J.M. Benitez, F. Herrera, Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data, Fuzzy Sets Syst 258 (2015) 5–38.
[57] J.A. Morente-Molinera, I.J. Perez, M.R. Urena, E. Herrera-Viedma, Creating knowledge databases for storing and sharing people knowledge automatically using group decision making and fuzzy ontologies, Inf. Sci. 328 (2016) 418–434.
[58] E.J. Khatib, R. Barco, A. Gomez-Andrades, P. Munoz, I. Serrano, Data mining for fuzzy diagnosis systems in LTE networks, Expert Syst. Appl. 42 (2015) 7549–7559.
[59] S. del Rio, V. Lopez, J.M. Benitez, F. Herrera, A MapReduce approach to address big data classification problems based on the fusion of linguistic fuzzy rules, Int. J. Comput. Int. Sys. 8 (2015) 422–437.
[60] G. Ghosh, S. Banerjee, N.Y. Yen, State transition in communication under social network: An analysis using fuzzy logic and Density Based Clustering towards big data paradigm, Future Gener. Comp. Sy. 65 (2016) 207–220.
[61] E. Francalanza, J. Borg, C. Constantinescu, A fuzzy logic based approach to explore manufacturing system changeability level decisions, in: Cirp Cms 2015 - Cirp Conference on Manufacturing Systems, 2015, pp. 3–8.
[62] I. Jayawardene, G.K. Venayagamoorthy, Comparison of adaptive neuro-fuzzy inference systems and echo state networks for PV power prediction, in: A. Roy, P. Angelov, A. Alimi, K. Venayagamoorthy, T. Trafalis (Eds.), Inns Conference on Big Data 2015 Program, 2015, pp. 92–102.
[63] S. Mahmud, R. Iqbal, F. Doctor, Cloud enabled data analytics and visualization framework for health-shocks prediction, Future Gener. Comput. Syst. 65 (2016) 169–181.
[64] J.M. Mendel, M.M. Korjani, On establishing nonlinear combinations of variables from small to big data for use in later processing, Inf. Sci. 280 (2014) 98–110.
[65] Y.Y. Wang, Type-2 fuzzy event parallel computing system: overcoming computer int index limitation in big data, Appl. Soft Comput. 38 (2016) 1076–1087.
[66] V.G. Kaburlasos, G.A. Papakostas, Learning distributions of image features by interactive fuzzy lattice reasoning in pattern recognition applications, IEEE Comput. Intell. Mag. 10 (2015) 42–51.
[67] Y.J. Li, R. Wang, S.C.K. Shiu, Interval extreme learning machine for big data based on uncertainty reduction, J. Intell. Fuzzy Syst. 28 (2015) 2391–2403.
[68] H.T. Chang, N. Mishra, C.C. Lin, IoT big-data centred knowledge granule analytic and cluster framework for BI applications: a case base analysis, Plos One 10 (2015) e0141980.
[69] S. Kundu, S.K. Pal, Fuzzy-rough community in social networks, Pattern Recogn. Lett. 67 (2015) 145–152.
[70] C. Bai, D. Dhavale, J. Sarkis, Complex investment decisions using rough set and fuzzy c-means: an example of investment in green supply chains, Eur. J. Oper. Res. 248 (2015) 507–521.
[71] A.P. Zeng, T.R. Li, D. Liu, J.B. Zhang, H.M. Chen, A fuzzy rough set approach for incremental feature selection on hybrid information systems, Fuzzy Sets Syst 258 (2015) 39–60.
[72] Y.H. He, L.B. Wang, Z.Z. He, M. Xie, A fuzzy TOPSIS and rough set based approach for mechanism analysis of product infant failure, Eng. Appl. Artif. Intell. 47 (2016) 25–37.
[73] I. Timon, J. Soto, H. Perez-Sanchez, J.M. Cecilia, Parallel implementation of fuzzy minimals clustering algorithm, Expert Syst. Appl. 48 (2016) 35–41.
[74] G. Peters, R. Weber, DCC: a framework for dynamic granular clustering, Granul. Comput. 1 (2016) 1–11.
[75] B. Cetisli, The effect of linguistic hedges on feature selection: part 2, Expert Syst. Appl. 37 (2010) 6102–6108.
[76] S. Aftergood, Big data: stealth control, Nature 517 (2015) 435–436.
[77] Z.J. Wang, X.X. Su, Dynamically hierarchical resource-allocation algorithm in cloud computing environment, J. Supercomput. 71 (2015) 2748–2766.
[78] G.B. Davis, K.M. Carley, Clearing the FOG: fuzzy, overlapping groups for social networks, Soc. Netw. 30 (2008) 201–212.
[79] M. Srivathsan, K.Y. Arjun, Health monitoring system by prognotive computing using big data analytics, in: V. Vijayakumar, V. Neelanarayanan (Eds.), Big Data, Cloud and Computing Challenges, 2015, pp. 602–609.
[80] C. De Maio, G. Fenza, V. Loia, S. Senatore, Hierarchical web resources retrieval by exploiting fuzzy formal concept analysis, Inf. Process. Manag. 48 (2012) 399–418.
[81] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, Int. J. Gen. Syst. 17 (1990) 191–209.
[82] M. Hindman, Building better models prediction, replication, and machine learning in the social sciences, ANN. Am. Acad. Polit. Soc. Sci. 659 (2015) 48–62.
[83] M.D. Assunção, R.N. Calheiros, S. Bianchi, M.A. Netto, R. Buyya, Big data computing and clouds: trends and future directions, J. Parallel Distr. Com. 79 (2015) 3–15.
[84] H.G. de Zúñiga, Citizenship, social media, and big data current and future research in the social sciences, Soc. Sci. Comput. Rev. 33 (2015) 1–7.

[85] X. Jin, B.W. Wah, X. Cheng, Y. Wang, Significance and challenges of big data research, Big Data Res 2 (2015) 59–64.
[86] K. Thirunarayan, A. Sheth, Semantics-empowered approaches to big data processing for physical-cyber-social applications, in: Proceedings of the AAAI 2013 Fall Symposium on Semantics for Big Data, 2013, pp. 68–75.
[87] N. Marz, J. Warren, Big Data, Principles and Best Practices of Scalable Realtime Data Systems, Manning Publications Co., 2015.
[88] X. Wu, X. Zhu, Mining with noise knowledge: Error-aware data mining, IEEE Trans. Syst. Man Cybernet. A 38 (2008) 917–932.
[89] J. Black, P. Rogaway, Ciphers with arbitrary finite domains, in: B. Preneel (Ed.), Topics in Cryptology — CT-RSA 2002: The Cryptographers' Track at the RSA Conference 2002 San Jose, CA, USA, 2002, pp. 114–130.
[90] M. Li, Z. Liu, J. Li, C. Jia, Format-preserving encryption for character data, J. Netw. 7 (2012) 1239–1244.
[91] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: tasks, approaches and applications, Knowl. Based Syst. 89 (2015) 14–46.
[92] C. Kacfah Emani, N. Cullot, C. Nicolle, Understandable big data, Comput. Sci. Rev. 17 (2015) 70–81.
[93] C. MolinariNo one size fits all strategy for big data, says IBM, 2012 (accessed July 2016). http://www.bnamericas.com/news/technology/no-one-sizefits-all-strategy-for-big-data-says-ibm .
[94] V. Torra, Hesitant fuzzy sets, Int. J. Intell. Syst. 25 (2010) 529–539.
[95] R.Y. Ronald, On the theory of bags, Int. J. Gen. Syst. 13 (1986) 23–37.
[96] D. Ramot, R. Milo, M. Friedman, A. Kandel, Complex fuzzy sets, IEEE Trans. Fuzzy Syst. 10 (2002) 171–186.
[97] F. Smarandache, A Unifying Field in Logics: Neutrosophic Logic, American Research Press, 1999.
[98] H. Wang, G. Qian, X.Q. Feng, Predicting consumer sentiments using online sequential extreme learning machine and intuitionistic fuzzy sets, Neural Comput. Appl. (2012) 1–11.
[99] Z.S. Xu, H. Wang, Managing multi-granularity linguistic information in qualitative group decision making: an overview, Granul. Comput. 1 (2016) 21–35.
[100] R.M. Rodriguez, L. Martinez, F. Herrera, Hesitant fuzzy linguistic term sets for decision making, IEEE Trans. Fuzzy Syst. 20 (2012) 109–119.
[101] H. Wang, Extended hesitant fuzzy linguistic term sets and their aggregation in group decision making, Int. J. Comput. Int. Sys. 8 (2015) 14–33.
[102] Q. Pang, H. Wang, Z.S. Xu, Probabilistic linguistic term sets in multi-attribute group decision making, Inf. Sci. 369 (2016) 128–143.
[103] Y.L. Zhai, Z.S. Xu, H.C. Liao, Probabilistic linguistic vector-term set and its application in group decision making with multi-granular linguistic information, Appl. Soft Comput. (2016), doi:10.1016/j.asoc.2016.08.044/.
[104] Z.S. Xu, H. Wang, On the syntax and semantics of virtual linguistic terms for information fusion in decision making, Inf. Fusion 34 (2017) 43–48.
[105] F. Herrera, L. Martinez, A 2-tuple fuzzy linguistic representation model for computing with words, IEEE Trans. Fuzzy Syst. 8 (2000) 746–752.
[106] C. Luo, T. Li, Z. Yi, H. Fujita, Matrix approach to decision-theoretic rough sets for evolving data, Knowl. Based Syst. 99 (2016) 123–134.
[107] T. Li, C. Luo, H. Chen, J. Zhang, PICKT: a solution for big data analysis, in: D. Ciucci, G. Wang, S. Mitra, W.Z. Wu (Eds.), Rough Sets and Knowledge Technology: 10th International Conference, RSKT, 2015, pp. 15–25.
[108] J. Rigelsford, Rough neural computing: Techniques for computing with words, Ind. Robot 31 (2013) 534.
[109] G. Gu, Q. Li, X. Wen, Y. Gao, X. Zhang, An overview of newly open-source cloud storage platforms, in: Proceedings of 2012 IEEE International Conference on Granular Computing (GrC), Hangzhou, China, 2012, pp. 142–147.
[110] K.M. Huang, H.Y. Chen, K.L. Hsiung, On realizing rough set algorithms with apache spark, in: Proceedings of The Third International Conference on Data Mining, Internet Computing and Big Data, Konya, Turkey, 2016, pp. 111–112.
[111] D.L. Ding, D.Y. Wu, F.L. Yu, An overview on cloud computing platform Spark for human genome mining, in: Proceedings of 2016 IEEE International Conference on Mechatronics and Automation, Harbin, China, 2016, pp. 2605–2610.
[112] L. Yang, Z. Shi, L.D. Xu, F. Liang, I. Kirsh, DH-TRIE frequent pattern mining on Hadoop using JPA, in: Proceedings of 2011 IEEE International Conference on Granular Computing (GrC), Kaohsiung, 2011, pp. 875–878.
[113] M. Bahrami, M. Singhal, The role of cloud computing architecture in big data, in: W. Pedrycz, S.M. Chen (Eds.), Information Granularity, Big Data, and Computational Intelligence, 8, 2015, pp. 275–295.
[114] P. Richtarik, M. Takac, Parallel coordinate descent methods for big data optimization, Math. Program. 156 (2016) 433–484.
[115] C. Napoli, G. Pappalardo, E. Tramontana, G. Zappala, A cloud-distributed GPU architecture for pattern identification in segmented detectors big-data surveys, Comput. J. 59 (2016) 338–352.
[116] B.B. Mandelbrot, The fractal geometry of nature, Am. J. Phys. 51 (1983) 286–287.
[117] U.R. Acharya, S.V. Sree, P.C.A. Ang, R. Yanti, J.S. Suri, Application of non-linear and wavelet based features for the automated identification of epileptic signals, Int. J. Neural Syst. 22 (2012) 1250002.
[118] O. Castillo, P. Melin, Hybrid intelligent systems for time series prediction using neural networks, fuzzy logic, and fractal theory, IEEE Trans. Neural Netw 13 (2002) 1395–1408.
[119] W. Pedrycz, A. Bargiela, Fuzzy fractal dimensions and fuzzy modeling, Inf. Sci. 153 (2003) 199–216.