Safety Science 101 (2018) 60–71

Contents lists available at ScienceDirect

Safety Science

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

Methodologies, principles and prospects of applying big data in safety science research

Qiumei Ouyang*, Chao Wu, Lang Huang*

School of Resources and Safety Engineering, Central South University, No. 932 South Lushan Road, Changsha 410083, PR China

ARTICLE INFO

Keywords:
Safety big data (SBD)
Safety small data (SSD)
Safety science
Big data application
Method
Principle
Prospect and challenge

ABSTRACT

It is clear that big data has numerous potential impacts in many fields. However, few papers discussed its applications in the field of safety science research. Additionally, there exist many problems that cannot be ignored when big data is applied to safety science, most outstanding of which is lack of universal supporting theory that guides how to apply big data to safety science research like methods, principles and approaches, etc. In other terms, it is not enough for big data to be viewed as a strong enabler for safety science applications mainly due to lack of universal and basic theory from the perspective of safety science. Considering the above analyzes, the two key objectives of this paper are: (1) to propose the connotation of safety big data (SBD) and its applying rules, methods and principles, and (2) to put forward some application prospects and challenges of big data to safety science research seen from theoretical research. First, by comparing SBD and traditional safety small data (SSD) from four aspects including theoretical research, typical research method, specific analysis method and processing mode, this paper puts forward the definition and connotation of SBD. Subsequently this paper further summarizes and extracts the application rules and methods of SBD. And then nine principles of SBD are explored and their relationship and application are addressed from the view of theory architecture and working framework in data processing flow. At last, this paper also discusses the potential applications and some hot issues of SBD. Overall, this paper will play an essential role in supporting the application of SBD. In addition, it will fill in the theory gaps in the field of SBD beyond traditional safety statistics, and further enriches the contents of safety science.

1. Introduction

Computer and Internet technologies have made data digitized and networked, bringing big data new connotations. Arthur (2011) believed big data is an inexhaustible and constantly enriching resource. In recent years, big data has caused the scientific community to reexamine its methodology of scientific research and has triggered a revolution in scientific thinking and methods (Hey, 2012; Chen and Zhang, 2014). All of these changes bring a considerable amount of attention to big data to enhance the security and reliability of research and operation (Mayer-Schonberger and Cukier, 2013; Cuzzocrea, 2014; Thomson et al., 2014). *Nature* and *Science* have published special issues dedicated to discussing the opportunities and challenges brought by big data (Staff, 2011; Lynch, 2008). Huang et al. (2017) thought it is no secret that in big data research and applications, industry is ahead of academia, and Jin et al. (2015) concluded that there could be such two essential driving forces as industrial demand and the requirements on making better use of the data.

It is clear that big data has numerous potential applications (e.g. Srunswicker et al., 2015; Shi and Abdel-Aty, 2015; Nuaimi et al., 2015; Wang et al., 2016), so as in the field of safety science (e.g. Sakthivadivel et al., 2013; Batarseh and Latif, 2016; Guo et al., 2015; Walker and Strathie, 2016). Current works and research projects in this field have generated some literature that highlighted the importance and feasibility of big data in supporting safety science applications and services. For example, new accident causation and prediction models were put forward through mass typical accident statistics and analysis (Cuadras and Arenas, 2011; Al-Shanini et al., 2015). In addition, big data technology enabled risk and emergency management software, enterprise EHS control platform, safety service products like accident emergency equipment. Some paid attention to SBD application to special safety field, such as coal mine safety (El-Nasr and Shaban, 2015), traffic safety (Chen et al., 2016; Shi and Abdel-Aty, 2015), oil and gas safety (Sakthivadivel et al., 2013) and healthcare (Batarseh and Latif, 2016). Besides, some works started from the influence factor of safety data acquisition and sharing to propose some measures or develop new

* Corresponding authors.
E-mail addresses: ouyangqiumei0893@csu.edu.cn (Q. Ouyang), 18971185983@189.cn (L. Huang).

http://dx.doi.org/10.1016/j.ssci.2017.08.012
Received 7 March 2017; Received in revised form 30 July 2017; Accepted 13 August 2017

platforms (Zayed and Mahmoud, 2013; Tao and Jochen, 2013). It is safe to say that there is a fascinating outlook with far-reaching consequences. And in many situations, it is possible to bring safety up to the necessary level by adding more safety data which will smooth out errors of human or machine to avoid or reduce accident.

However, there exist still many problems that cannot be ignored when big data is applied to safety science. It's important to realize that safety data is becoming exponentially un-analyzable with traditional statistics methods, most of whose contributions fail to move beyond existing theories and routine application to tackle the dynamic and complex issues in terms of volume, intensity and complexity, forcing a rethink on how to exploit the vast values of safety data efficiently (Klous and Wielaard, 2016). More specifically, there are four concrete manifestations as below: (i) data plugging, mainly due to inconsistent standard of data format which makes it hard to seize useful information, (ii) data separation, referring to scattered or redundant data storage in different parts with little associated integration, (iii) slow data, mainly result from weak supporting environment in data acquisition, and lack of practical tools and informationization ability to safety data analysis, and (iv) data missing, referring to important and unrecoverable data that fails to acquire and update in time.

Meanwhile, it is not enough for big data to be viewed as a strong enabler for safety science applications. First of all, big data is not mature enough to resolve perfectly some safety dilemma as it may bring noise accumulation, spurious correlation and incidental homogeneity, creating issues in computational feasibility and algorithmic stability (Fan et al., 2014). Then as for those small and simple safety problems it is enough just to apply traditional safety statistics. Furthermore, the most significant aspect of all is that there is lack of universal supporting theory that guides big data application in safety science field. In the context of the development strategy of big data, from the perspective of safety science theory, exploring how to achieve effective SBD application has an important practical significance.

From what has discussed above, the main objective of this study is to explore the methodologies and principles of SBD application. In what follows, we proceed by first comparing SBD with traditional SSD from four perspectives including theoretical research, typical research method, specific analysis method and processing mode. Next, we put forward the new definition of SBD and give a detailed description of its connotation. Then, this paper discusses the universal rules and interoperable methods of SBD and SSD. The following section sums up the nine basic application principles of SBD and constructs their theory architecture and working framework in data processing flow. Finally, this paper expounds the prospects and challenges of SBD application.

2. Connotation of SBD and comparison with SSD

2.1. The connotation of safety data

Although there is a worldwide prevalence of big data application, there is no universally agreed and accepted definition (Jin et al., 2015). In other words, a shared definition of big data is not yet offered, and it has been difficult to pinpoint a standard global meaning.

Data is being generated from multiple sources resulting in the formation of what is currently known as big data (Nuaimi et al., 2015). To obtain a universal definition of SBD, it is more convenient and prime to shed light on safety data, which can be simplified into a kind of special data related to safety. In the past, safety data was widely considered as digitized information records and more closely related to work safety. But in big data era, it is an assemble of all kinds of safety data, including text, audio, video and picture et al., and covers widespread areas such as work safety, safety regulation, and social engagement. Besides, it focuses more on the joint changes of data quality and quantity. The extensive sources of safety data, including internet, government regulation, intermediary organization, enterprise and social individual, make the contents wider and more various.

From the above mentioned, SSD can be described simply as a kind of safety data based on traditional small sample statistics. Corresponding to the big data era, those whole safety data as sample are called SBD. And comparative method is a great approach to getting further understanding of both.

2.2. The comparison between SBD and SSD

By means of literature review and comparative method, it is not hard to find out that SBD has plenty of differences from SSD (Edosio, 2014; Chen et al., 2016; Davenport et al., 2012; Ouyang and Wu, 2016), which can be summed up in the following four tables from the perspectives of theoretical research (e.g. Mayer-Schonberger and Cukier, 2013; Roger, 2015; Sacristán and Dilla, 2015; Jagadish, 2015; Klous and Wielaard, 2016; Lee, 2017), typical research method (Wu and Wang, 2014; Sivarajah et al., 2017; Maunder and Punt, 2013) and specific analysis method (Oussous et al., 2017; Wu, 2011) and processing mode (Luo et al., 2013; Ghit et al., 2013).

The Tables 1–4 above four tables show that SBD has some advantages over SSD, which can be seen from three perspectives: (1) its volume and variety can describe some rules; (2) it can accept some outliers, and (3) it can trace some week and potential information.

However, SBD also has some shortcomings that SSD can make a good defect-remedy, such as: (i) SBD is more likely to come from computer platform which reflects the situation of special groups sometimes (for example, mass safety monitor information of storage tank by linking sensor to computer), instead, SSD collected by random sampling could give expression to full view, the typical means of which is method of mass observation; (ii) rules or results mined by SBD have obvious periodicity and timeliness, and sometimes it depends on human brain to judge, analyze and apply; (iii) lots of false and disloyal information have a significant impact on SBD, moreover, the complexity, indeterminacy and emergency of SBD make it difficult to solve safety problems quickly and efficiently. (iv) As Mayer-Schonberger and Cukier (2013) pointed out that some insolvable problems that traditional method remains, maybe even more serious with the increasing data volume.

SSD plays an indispensable role at the present stage, which could be summed up to the following four concrete manifestations: (i) when all data aren't available, random sampling is still the first choice to have an insight into full view; (ii) the access to useful information is difficult to distinguish because of low value density of SBD. On the contrary, SSD follows the principle of taking demands as orientation, which makes it more likely to obtain the required data; (iii) obviously, only correlation is not enough for safety science, consequently, exploring accident cause is a must to provide safety protection and control measures on similar

Table 1
SBD compared with SSD based on theoretical research.

Perspective	SSD	SBD
Philosophical foundation	Homogeneity	Heterogeneity
Data operation mode	Manpower first, machine next	Machine first, manpower next
Main data type	Structured	Semi-structured and unstructured
Data state	Static	Dynamic
Data dimension	One-dimensional	Multi-dimensional
Emphasis on data collection	Non-scene data	Scene data
Foundation of research tool	Scale theory	Long tail theory
data model relationship	Causality	Correlativity
Main usage	Description	Prediction
Main database	Relational database	Non-relational database
Data orientation	Object-oriented	Subject-oriented
Database language	SQL	NOSQL