Transportation Research Part C 90 (2018) 226–246

Contents lists available at ScienceDirect

Transportation Research Part C

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

Recent applications of big data analytics in railway transportation systems: A survey☆

Faeze Ghofrani[a], Qing He[a,b,*], Rob M.P. Goverde[c], Xiang Liu[d]

[a] *Department of Civil, Structural and Environmental Engineering, University at Buffalo, The State University of New York, 212 Ketter Hall, Buffalo, NY 14260, USA*
[b] *Department of Industrial and Systems Engineering, University at Buffalo, The State University of New York, 342 Bell Hall, Buffalo, NY 14260, USA*
[c] *Department of Transport and Planning, Delft University of Technology, Stevinweg 1, 2628 CN Delft, The Netherlands*
[d] *Department of Civil and Environmental Engineering, Rutgers, The State University of New Jersey, 96 Frelinghuysen Road, Piscataway, NJ 08854, USA*

A R T I C L E I N F O

Keywords:
Survey
Big data
Data analytics
Railway transportation systems

A B S T R A C T

Big data analytics (BDA) has increasingly attracted a strong attention of analysts, researchers and practitioners in railway transportation and engineering. This urges the necessity for a review of recent research development in this field. This survey aims to provide a comprehensive review of the recent applications of big data in the context of railway engineering and transportation by a novel taxonomy framework, proposed by Mayring (2003). The survey covers three *areas* of railway transportation where BDA has been applied, namely operations, maintenance and safety. Also, the *level* of big data analytics, types of big data *models* and a variety of big data *techniques* have been reviewed and summarized. The results of this study identify the existing research gaps and thereby directions of future research in BDA in railway transportation systems.

1. Introduction

The fast-paced development of advanced technologies has made BDA as one of the most focused areas of both academia and industry. The features of big data can be characterized by 5 V, namely, volume, variety, velocity, veracity, and value (Fosso Wamba et al., 2015). The magnitude of data is featured by volume and it is among the most challenging issues specifically in terms of the storage capacity of devices (Emani et al., 2015). Variety refers to the various resources from which data can be generated in the forms of structured, semi-structured or unstructured data (Tan et al., 2015). Speed of generating data is characterized by velocity which, according to Assunção et al. (2015), may be processed in batch, real-time, nearly real-time, or streamlines. Since many data sources contain a specific level of uncertainty, the level at which a data source is trustable is featured by veracity (Gandomi and Haider, 2015). Finally, the process of revealing underexploited values from big data to support decision-making is referred by value (Assunção et al., 2015).

Railways are among the industries in which the application of big data analytics is a topic of big interest. A systematic consideration of the use of data in the context of railway transportation systems (RTS) was firstly provided in Faulkner (2002) in which four categories of data were introduced for the railway control system: (1) *configuration data* which is mostly regarded as static data that represents the entities from the real world, and change only in response to the action of maintenance or modification on these

☆ This article belongs to the Virtual Special Issue on “Big Data Railway”.

* Corresponding author at: Department of Civil, Structural and Environmental Engineering, University at Buffalo, The State University of New York, 212 Ketter Hall, Buffalo, NY 14260, USA.

E-mail address: qinghe@buffalo.edu (Q. He).

https://doi.org/10.1016/j.trc.2018.03.010
Received 11 September 2017; Received in revised form 15 January 2018; Accepted 14 March 2018
0968-090X/

entities, (2) *train schedule* which is used to describe the use of the infrastructure, (3) *status data* which is provided through interfaces to external reporting systems and (4) *operational data* which accounts for the individual operational conditions which are commonly communicated to the railway control system via manual input. In this essence, many aspects of the railway world benefit from today's capability of information technology in collecting, storing, processing, analyzing and visualizing large amounts of data as well as new methods coming from machine learning, artificial intelligence, and computational intelligence to recognize patterns and retrieve useful information (Al-Jarrah et al., 2015). This is in accordance with the growing demand on railway transportation, which necessitates ensuring customer satisfaction by being safe, reliable and service-oriented. In fact, railway industry has been revolutionized by big data analytics (BDA) which contributes to the decision-making processes of railway companies. Studies in the literature demonstrate multiple advantages of applying BDA in RTS in reducing cost and delay, and in parallel maintaining high standards of safety, reliability and customer satisfaction.

Despite the fact that BDA adoption can enhance RTS performance, not many of railway-related enterprises have implemented BDA in one or more RTS areas. This is mainly due to the lack of understanding on how BDA can be implemented in RTS, the inability to collect and process massive data, data security issues, routinization and assimilation of BDA by railway companies. This motivates our exploration of identifying the existing gaps in the applications of BDA in RTS.

There are a number of surveys in the literature on the application and challenges of BDA in RTS context. However, most of these studies tend to focus on a specific aspect of RTS. For instance, in Hodge et al. (2015), a survey of wireless sensors network technology for monitoring and analyzing railway systems, structures, vehicles, and machinery was conducted. A survey of railway-related planning and scheduling issues in Europe was provided in Turner et al. (2016). As another example Nunez and Attoh-Okine (2015) conducted a literature review on the application of metaheuristic optimization in railway engineering. Some other survey articles on the application of data analytics in a specific aspect of RTS can be found in Soleimanmeigouni, et al. (2016), Singh et al. (2015), Hodge et al. (2015), Thaduri et al. (2015), Griffin et al. (2014), Summit (2014), Figueres-Esteban et al. (2015). To the best of authors' knowledge, the literature in this field of study suffers from the lack of a holistic survey which takes a broad perspective of RTS as a whole and cross-maps with BDA.

Our survey develops a taxonomy framework in Section 2, which identifies the areas of RTS and connects them with the level of analytics, BDA models, and techniques. The developed framework aims to provide a complete picture of where and how BDA has been applied in RTS. To obtain this objective the study considers four aspects namely, the *areas* of railway transportation in which big data analytics is applied, the *level* of big data analytics in rail transportation, types of big data *models* and big data *techniques* used to apply these models.

Section 3 specifically studies the material evaluation from the proposed framework. This includes the review of articles based on the three areas of railway transportation systems, i.e. maintenance, operations, and safety, as well as review by BDA models and BDA techniques. It should be mentioned that although the term "operations" is usually referred to a comprehensive spectrum of activities in the RTS context which occasionally include maintenance and safety as well, what we mean in this paper is the activities related to the train traffic and transportation services, thus excluding maintenance and safety activities.

One of the limitations of the current paper is that the categories in the proposed classification framework are interpretative, which is probably to result in subjective bias. This is also one of the main issues of the content analysis method according to Seuring (2013), despite several of validations being carried out.

In Section 4 a discussion on the future direction of the studies of BDA in the context of RTS as well as the advanced big data computational technologies in railway transportation systems around the world is provided and finally the conclusions are provided in Section 5.

2. Methodology

According to Brewerton and Millward (2001), from a methodological point of view, a literature review could be as comprehended as a content analysis which considers both quantitative and qualitative aspects of a context. In this essence, we have considered a survey approach based on content analysis by Mayring (2003). Practices and context of content analysis by Mayring (2003) has been taken up by various scholarly communities and modified by some of them specifically in the field of supply chain management (Seuring and Müller, 2008; Brandenburg et al., 2014; Govindan et al., 2015; Seuring, 2013; Klewitz and Hansen, 2014). With respect to BDA, this methodology has been adopted for big data analytics in supply chain management by Nguyen et al. (2017). Apart from the field of supply chain management, Gläser and Laudel (2010) took Mayring's method as a starting point for developing their own technique for analyzing expert interviews in social science. All the contexts which apply content analysis involve systematic reading or observation of texts, documents or artifacts which are assigned labels to indicate the presence of interesting, meaningful patterns in a specific field of study (Tipaldo, 2014). After labeling a large set of texts, documents, papers (in our case) researcher is able to statistically estimate the proportions of patterns in those documents, as well as correlations between them. This refers to the fact that content analysis is not limited to specific contexts to be applicable. As long as an eligible topic for a literature review is defined, sufficient materials are observed for a topic and the questions on "How the data is collected?", "Which population the data is taken from?", "How the relevant context is defined?", "What are the boundaries of the analysis?" are responded, the content analysis technique could be conducted as a methodology for classification and labeling of available materials (Krippendorff, 2004). It should be noted that in content analysis the analyst can make various decisions about how the paper is to be comprehended and what dimensions/classification are about to be considered. This is such a risk which can be reduced by involving two or more researchers when searching for and analyzing the data (Krippendorff, 2004) which is true for our research. Our survey is conducted according to the four-step iterative process as follows (Mayring, 2003; Seuring and Müller, 2008):