informatics

MDPI

Editorial

Quality Management in Big Data

Mouzhi Ge * and Vlastislav Dohnal

Faculty of Informatics, Masaryk University, Brno 602 00, Czech Republic; dohnal@fi.muni.cz
* Correspondence: mouzhi.ge@muni.cz

Received: 11 April 2018; Accepted: 12 April 2018; Published: 16 April 2018

Abstract: Due to the importance of quality issues in Big Data, Big Data quality management has attracted significant research attention on how to measure, improve and manage the quality for Big Data. This special issue in the Journal of Informatics thus tends to address the quality problems in Big Data as well as promote further research for Big Data quality. Our editorial describes the state-of-the-art research challenges in the Big Data quality research, and highlights the contributions of each paper accepted in this special issue.

Keywords: big data; quality management; data management; data quality

---

In the era of Big Data, organizations are dealing with tremendous amounts of data. These data are fast-moving and can originate from various sources, such as social networks, unstructured data from different websites and multimedia archives, or raw-data feeds from sensors. Big Data solutions are used to optimize business processes and reduce decision-making time, so as to improve operational effectiveness. However, Big Data practitioners are experiencing a huge number of data quality problems, which can be time-consuming to solve, or can even lead to incorrect data analytics. Managing quality in Big Data has become very challenging, and the research published so far can only address limited aspects. In particular, given the complex nature of Big Data, one cannot simply apply traditional data quality management models to Big Data quality management due to their scalability limits or incapability to handle data streams. Therefore, it creates new challenges for researchers and practitioners to address quality management in Big Data. Hereby, this Special Issue in the Journal of *Informatics* intended to attract submissions on the topic of quality issues in Big Data, namely Big Data quality measurement, Big Data governance, and Big Data value.

This Special Issue has accepted three high-quality papers. The acceptance rate is around 40%. Each paper has been reviewed by at least three reviewers. The readers have offered very insightful reviews, and the authors have invested extensive efforts to revise their papers. We appreciate the constructive reviews and efforts from authors to continuously improve their papers.

Firstly, the paper by Ben Evans, Kelsey Druken, Jingbo Wang, Rui Yang, Clare Richards, and Lesley Wyborn has developed a data quality strategy in the context of high-performance computing from Australian National Computational Infrastructure. The proposed strategy covers a method for the assessment of data consistency for a high-performance data platform; a quality control model in compliance with standards recognized within the community; data quality assurance functionality demonstrated across different domains and services; and benchmarking cases to exhibit the operational performance. Besides, the proposed data quality strategy can help increase data re-usability in high-performance environments and identify possible extensions of quality standards for interoperability and programmatic access.

Secondly, Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz propose an approach to assess the relative data quality in Wikipedia, which is considered to be one of the most popular knowledge bases in the world. The research on relative data quality is quite up-to-date. This paper is motivated by data quality variations in Wikipedia articles that are edited independently in different languages. The authors have conducted a comparative analysis of relative quality and

popularity assessment over 28 million articles in 44 selected languages. Their results show that the data quality varies across different language versions of the same Wikipedia article. Based on the correlation between quality and popularity of Wikipedia articles, it recommends how the quality of Wikipedia articles can be improved by using the articles with information of higher quality in other languages.

Thirdly, the paper by Janet G. Baseman, Debra Revere, and Ian Painter summarizes Big Data research challenges in the domain of health information exchanges. Since data exchange in health-care and public health systems facilitates the electronic sharing of data and information between health-care organizations, it is challenging to determine if a public health agency is able to meet the data quality level of disease reporting and surveillance, which largely depends on how timely, complete, accurate, and useful the data are across different health-care and public health systems. This paper presents the assessment results of the Big health-care Data in public health agencies, and derives research challenges related to the initial assessment results.

Each of the articles focuses on different aspects of quality management in Big Data applied to various domains or operated in different computing platforms (e.g., Wikipedia and Health Information Exchanges), and high-performance computing, respectively. All of the articles provide theoretical contributions and practical research results that contribute to the knowledge of Big Data quality management. Since the quality issue in Big Data is considered to be critical in Big Data analytics and Big Data value, this Special Issue brings together the key challenges and subjects of continuous debate on how to define, measure, model, and manage Big Data quality. The accepted papers also facilitate future research in Big Data quality management.

Acknowledgments: This research was supported by the Czech Science Foundation project number GA16-18889S.

Author Contributions: While Mouzhi Ge contributes 60% to the manuscript, Vlastislav Dohnal contributes 40% to the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.