Journal of Information Technology Management

Print ISSN: 2008-5893

Online ISSN: 2423-5059

# Big Data Quality: From Content to Context

**Ahmad Khalilijafarabad**

PhD, Department of Information Technology Management, Faculty of Management, University of Tehran, Tehran, Iran. E-mail: Ahmad.khalili@ut.ac.ir

## Abstract

Over the last 20 years, and particularly with the advent of Big Data and analytics, the research area around Data and Information Quality (DIQ) is still a fast growing research area. There are many views and streams in DIQ research, generally aiming at improving the effectiveness of decision making in organizations. Although there are a lot of researches aimed at clarifying the role of BIG data quality for organizations, there is no comprehensive literature review that shows the main differences between traditional data quality researches and Big Data quality researches. This paper analyzed the papers published in Big data quality and find out that there is almost no new mainstream about Big Data quality. It is shown in this paper that the main concepts of data quality do not changes in Big Data context and that only some new issues have been added to this area.

**Keywords:** Big data, Big data quality, Data quality, Text mining.



DOI: 10.22059/JITM.2019.72762

## Introduction

Big Data has become a very attractive area of research and development for both academia and industries in recent years (Chen, Mao, & Liu, 2014). The Big Data concept was firstly defined in 2005 Tim O'Reilly and since then a lot of researchers have studied its applications, challenges, tools, technologies and quality (Kataria & Mittal, 2014).

The term of Big Data applies to data sets that are beyond the ability of manual techniques and software tools to capture, manage, and process within a tolerable timeframe (Becker, King, & McMullen, 2015). Big data generally refers to social media data, mobile phone call records, commercial website data, volunteering geographical information, search engine data, smart card data, and taxi trajectory data that are linkable, large and with complex structure (Khoury & Ioannidis, 2014).

Big data is defined as the data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it (Dumbill, 2013). But the most popular description of big data is 3V model and its extensions which is volume, variety and velocity (Laney, 2001). Volume means Big Data has large volume, Variety means Big Data have diverse data source and velocity refers to real time data updating.

We have analyzed the papers published in web of science (WOS) and retrieved more than 40 thousand papers with Big Data as the topic. As it is demonstrated in Figure1, the number of publication on Big Data is growing very fast and it will probably continue to grow the future.


**Figure 1.** Number of Published papers about Big Data

As it has mentioned, data quality is critical issue in information management and Big Data. The effectiveness of business information management within an organization is determined by the quality of information (Khalilijafarabad, Helfert, & Ge, 2016).

Furthermore, with the increasing importance of Analytics and the growth of Big Data, the importance of Data quality has been emphasized in many publications and it is going to become a significant research area considering its fast growth (Shankaranarayanan & Blake, 2017). There are some studies about data quality problems and challenges of Big Data. In Big Data, data comes from different aspects, multiples sources. It must be cleaned, filtered, processed, integrated, merged, partitioned, transported, sketched, and stored (Chen et al., 2014). It is also important to note that all steps should be executed in real-time, in batch or in parallel and preferably on the cloud (Chen et al., 2014).

Although there are some studies about the new aspects of data quality in Big Data, there is no comprehensive literature review to show how the data quality issues change in Big Data problems. With today's rapid technological changes such as Big Data (Cai & Zhu, 2015; Saha & Srivastava, 2014), crowdsourcing (Lukyanenko & Parsons, 2015), social information system (Tilly, Posegga, Fischbach, & Schoder, 2015) and semantics web (Fürber, 2015), it becomes critical to identify emerging research directions.

In this paper we want to analyse the papers published in Big Data quality and find out how data quality has been changed in Big Data.

## Literature review

Data quality is related to various areas including statistics, management and computer science. Statistical researchers were the first to address some of the data quality issues. By presenting mathematical theories in the late 1960s, they proposed solutions for finding duplicate data in a dataset. Subsequently, management science researchers in the 1980s focused on eliminating data quality problems in data production processes and related systems. In the 1990s, computer science researchers also began to define, measure, and improve the quality of electronic data in databases and data warehouses (Batini & Scannapieca, 2006).

The term quality has been defined as fitness for use (Juran, 1974) and this definition is widely adopted in the quality literatures (Wang & Strong, 1996). But if we want to have a closer look at this area, the field of data quality management was first introduced in the 1980s by Brodie. He showed that the importance of organizational areas is as important as the technical areas of quality management. He also emphasized that data quality would not occur without regard to both the aforementioned aspects, namely the organizational and technical fields (Brodie, 1980).

After all, the most serious works in this field can be attributed to MIT University, which started in 1990 with the launch of a research team at MIT University in the field of computer science. Wang and Strong (1996) define DIQ in their seminal work as information that is fit for use by information consumers. They extracted different dimensions that are important to data consumers and categorized the dimensions in four classes(Wang & Strong, 1996). This definition is accepted by a wide range of researchers (Breur, 2009; Ofner, Otto, & Österle,

2012). They argue that ultimately it is the consumer who will judge whether or not an information product is fit for use. However, information consumers are not very capable of finding errors in information and altering the way they use the information (Klein, 2001). From a data perspective, DIQ can be defined as the information that meets the specifications or requirements.

After the advent of Big Data, some studies have been done in order to study the differences between traditional data quality and Big Data quality. The first and mostly focused approach is to add some dimensions to Big Data quality. Dimensions are the most important and discussed concepts in data quality management. There are a lot of studies that suggest a list of dimensions for Big Data quality managements. They have suggested a variety of dimensions such as timeliness, latency, scalability, accuracy, consistency, usabiliy (Desai, 2018; Firmani, Mecella, Scannapieco, & Batini, 2016; Gao, Xie, & Tao, 2016; Onyeabor & Ta'a, 2018).

It seems that depending on the usage of data, some dimensions are more critical than others. But generally consistency, reputation and accuracy are the most important dimensions (Juddoo, 2015).

Some studies suggested that the data quality model should be extended to follow Big Data concepts such as the origin, domain, nature, format, and type of the data. They also claimed that the management of these quality schemes is essential when dealing with large datasets (Chen et al., 2014).

It is also argued that Big Data also brings problems in data quality and data usage, which decrease the usability of Big Data. Research based on wrong or incomplete data or data with errors don't meet the requirements of good scientific research in terms of authenticity and accuracy. This type of research will likely result in biased or wrong conclusions if we do not have a deeper understanding of the quality issues of Big Data and its consequent problems (Liu, Li, Li, & Wu, 2016).

Some studies also claimed that Big Data quality problems are the result of a series of problems. According to these studies data quality problems are the result of poor data quality assurance, poor data management, organizational problems, scalability problems, data transformations problems, data conversion problems and data collection problems (Gao et al., 2016).

Although there are some studies about Big Data quality issues, there is no comprehensive literature review study to show the differences between traditional data quality management and Big Data management.

## Methodology

In order to answer the main question of this research, we have used machine learning approach to analyse the data. We used Latent Dirichlet allocation (LDA) in order to find the topics of the data quality and compare it with traditional data quality topics. The main phases

of this research are data gathering, preprocessing, topic modeling and analysis which will be discussed in details.

## Data gathering

In order to gather the proper data, the researchers selected relevant keywords that aim to cover relevant areas of Big Data Quality. The papers with keywords "data quality," "information quality" and Big Data in their topic are selected. With these query on WOS we have retrieved about 381 papers. After the first retrieval, the researchers analysed the papers based on their title and abstract and removed the irrelevant papers. The final data set consist of 271 papers which are relevant to Big Data quality issues.

## Pre-processing

Pre-processing is one the most important phases in all data mining and machine learning projects. In order find the best topics, it is needed to clean the data before LDA. Here we have done different types of methods such as removing missing values, remove punctuation marks, spell checking, case matching, and transformations in order unify the format of the keywords for analysis.

## Topic Modeling

The goal of this part is to automatically classify researches with respect to an underlying topic using the text's content. Here we have used Latent Drichlet Allocation which is widely used for finding the hidden topics of the papers (Blei, Ng, & Jordan, 2003).

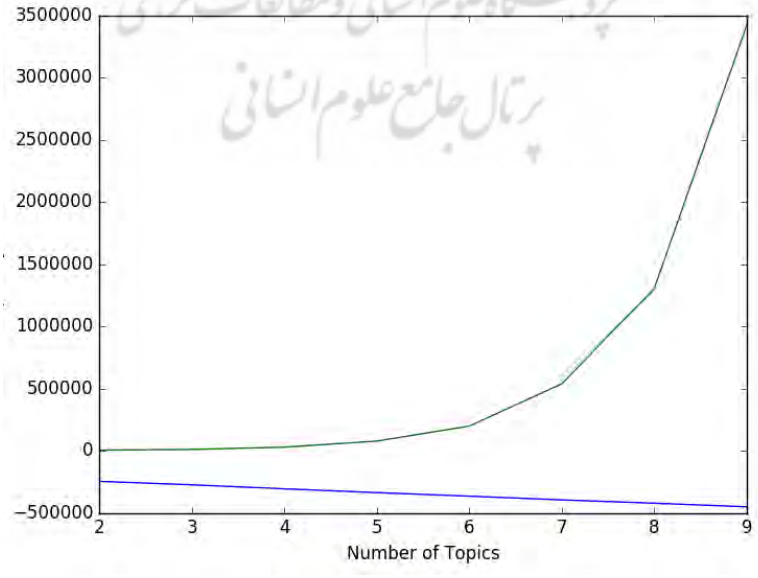


**Figure 2.** Log-liklihood score in order to find the numbe of topics

In order to conduct LDA a Python system is created. LDA was evaluated both quantitatively with Perplexity, a measurement on how well a model represents reality; and qualitatively by subjectively analysing output from the system (Risch, 2016). The number of topics also estimated by log-likelihood. As it is shown in the Figure2, the number of topics should be 5.

### Analysis

In the analysis phase we discuss the findings of the topic's modelling, and according to the findings come up with the main differences between traditional data quality management and Big Data quality management. We also discuss about the reasons of these differences.

## Results

As it has been mentioned, the optimum number of topics is 5. Table 1 shows the keywords related to each of these 5 topics.

**Table 1.** Topics extracted from topics modeling system

| | |
|---|---|
| **Topic1** | quality, information, method, analysis, process, challenge, model, approach, management, business, system, product, application |
| **Topic2** | chain, supply, analytics, performance, capability, resource, operation, management, innovation, literature, decision, system, location |
| **Topic3** | quality, risk, analytics, source, analysis, cloud, query, firm, computing, clinical, database, adoption, information |
| **Topic4** | science, western, educational, approach, teaching, learning, purpose, automated, question, system, helping, generation, author |
| **Topic5** | factor, decision, health, value, accounting, evidence, enablers, implication, making, cvd, scm, program, marine, oscm, mobile |

The first topic is related to managerial approach of the information and data quality in Big Data era. The topics focus in challenges and processes related to Big Data. The second topic is about the supply chain and analytical process of supply chain management in Big Data and the third one is about cloud computing and analytics technologies in Big Data.

It is shown in the Table 1, that there is almost no new concept related to Big Data quality, except the things that are related to the technology of the Big Data. It other words the concepts and topics related to data quality such as the dimensions, metrics, models, governance of data does not change in Big Data technologies and there are only some new concepts that are related to technology. It is also important to note that it does not means that there is no dimension related to Big Data, it means that the concepts is the same as traditional approach with some modifications.

## Conclusion

As it has been mentioned, we found the emerging changes in the context such as Big Data and Cloud Computing. In other words, no new issues are coming up but only old problem in a new context. For example, we noticed that the DIQ metrics have been recently used in Big Data. But here are some papers discussed about the priority or the importance of the dimension related to Big Data Quality.

There are also some old issues that will show up in Big Data contexts. So if there is data quality issues in the traditional data sets, it will be greater if the size of the data increases (Becker et al., 2015). There are also some cases that data quality issues will increasingly grow. When there are some relations between different data, with the increase in volume, the data quality issues will increase exponentially. It is also noteworthy that there some cases that the quality will improve with Big Data.

Another very important difference between Big Data Quality and traditional data quality is that in the traditional approaches the researchers could investigate almost all parts the data, but in Big Data it is not possible to study all elements of the data. Because of this we should select adequate an amount of data for checking its quality (Batini, Rula, Scannapieco, & Viscusi, 2016). This means that although the main issues of the data quality is not changed, the characteristics of Big Data problems changed the way we should think about the data quality.

## References

Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2016). From data quality to big data quality. In *Big Data: Concepts, Methodologies, Tools, and Applications* (pp. 1934-1956): IGI Global.

Batini, C., & Scannapieca, M. (2006). Methodologies for data quality measurement and improvement. *Data Quality: Concepts, Methodologies and Techniques*, 161-200.

Becker, D., King, T. D., & McMullen, B. (2015). Big data, big data quality problem. *Paper presented at the 2015 IEEE International Conference on Big Data (Big Data)*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.

Breur, T. (2009). Data quality is everyone's business—Designing quality into your data warehouse—Part 1. *Journal of Direct, Data and Digital Marketing Practice, 11*(1), 20-29.

Brodie, M. L. (1980). Data quality in information systems. *Information & Management, 3*(6), 245-258.

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal, 14*. DOI: http://doi.org/10.5334/dsj-2015-002.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications, 19*(2), 171-209.

Desai, K. Y. (2018). *Big Data Quality Modeling And Validation*. Master's Theses and Graduate Research. San Jose State University.

Dumbill, E. (2013). Making sense of big data. *In: Mary Ann Liebert, Inc. 140 Huguenot Street,* 3rd Floor New Rochelle, NY 10801 USA.

Firmani, D., Mecella, M., Scannapieco, M., & Batini, C. (2016). On the meaningfulness of "big data quality". *Data Science and Engineering, 1*(1), 6-20.

Fürber, C. (2015). *Data quality management with semantic technologies*: Springer.

Gao, J., Xie, C., & Tao, C. (2016). Big Data Validation and Quality Assurance--Issuses, Challenges, and Needs. *Paper presented at the 2016 IEEE symposium on service-oriented system engineering (SOSE).*

Juddoo, S. (2015). *Overview of data quality challenges in the context of Big Data.* Paper presented at the 2015 International Conference on Computing, Communication and Security (ICCCS).

Juran, J. M. (1974). Basic concepts. *Quality control handbook*, 2.

Kataria, M., & Mittal, M. P. (2014). Big data: a review. *International Journal of Computer Science and Mobile Computing, 3*(7), 106-110.

Khalilijafarabad, A., Helfert, M., & Ge, M. (2016). *Developing a Data Quality Research Taxonomy-an organizational perspective.* Paper presented at the ICIQ.

Khoury, M. J., & Ioannidis, J. P. (2014). Big data meets public health. *Science, 346*(6213), 1054-1055.

Klein, B. D. (2001). User perceptions of data quality: Internet and traditional text sources. *Journal of Computer Information Systems, 41*(4), 9-15.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note, 6*(70), 1.

Liu, J., Li, J., Li, W., & Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS journal of photogrammetry and remote sensing, 115*, 134-142.

Lukyanenko, R., & Parsons, J. (2015). Information quality research challenge: adapting information quality principles to user-generated content. *Journal of Data and Information Quality (JDIQ), 6*(1), 3.

Ofner, M. H., Otto, B., & Österle, H. (2012). Integrating a data quality perspective into business process management. *Business Process Management Journal, 18*(6), 1036-1067.

Onyeabor, G. A., & Ta'a, A. (2018). A Model for Addressing Quality Issues in Big Data. *Paper presented at the International Conference of Reliable Information and Communication Technology.*

Risch, J. (2016). Detecting Twitter topics using Latent Dirichlet Allocation. Available at: http://uu.diva-portal.org/smash/get/diva2:904196/FULLTEXT01.pdf.

Saha, B., & Srivastava, D. (2014). Data quality: The other face of big data. *Paper presented at the Data Engineering (ICDE), 2014 IEEE 30th International Conference on.*

Shankaranarayanan, G., & Blake, R. (2017). From content to context: The evolution and growth of data quality research. *Journal of Data and Information Quality (JDIQ), 8*(2), 9.

Tilly, R., Posegga, O., Fischbach, K., & Schoder, D. (2015). What is Quality of Data and Information in Social Information Systems? Towards a Definition and Ontology. *International Conference on Information Systems,* At Fort Worth, TX, USA, Volume: 36.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems, 12*(4), 5-33.

**Bibliographic information of this paper for citing:**

Khalilijafarabad**,** Ahmad (2018). Big Data Quality: From Content to Context. *Journal of Information Technology Management*, 10(4), 65-71.