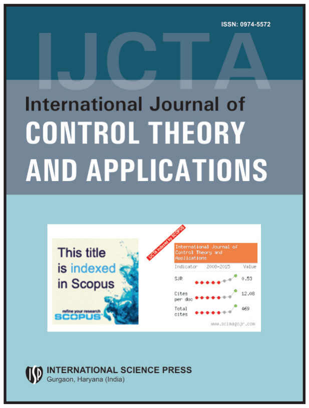

International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 9 • Number 44 • 2016

Survey on an Assessment Model to Evaluate Big Data-Data Quality

Supriya Haribhau Pawar[a] and Devendrasingh Thakore[b]

[a]*Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, Pune, India. Email: pawar45supriya@gmail.com*
[b]*Prof. Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, Pune, India. Email: dmthakore@bvucoep.edu.in*

Abstract: Big data describe volume amount of structured and unstructured data, which can be analyzed computationally to association, trends especially to human behavior and interactions. Social media data is one of big data's most important origins. Social media data are coming from Facebook and Twitter, and this data are important for business decision making. People share their daily experience, activities, thoughts and opinions on social media sites. Interactive interfaces have various data quality challenges. Data quality challenges in social media are accurate, relevant, timely, consistency data. The reason for increasing the technology of big data, building high quality big data services in applications domain is difficult. Researches have been trying to apply more metrics to evaluate the quality of data. This paper presents a comprehensive survey of how to evaluate quality in big data sources using metrics. Their needs to identify advantages and disadvantages are listed in a comparative manner. This paper concludes useful future direction that can help researchers to identify areas where further research is needed.

Keywords: Big data, Data quality, social media data, Metadata, Quality attributes, Quality metrics.

1. INTRODUCTION

Now a days lots of freely accessible data available online and this data is made by different companies, public sectors, organizations, institutes and different forms of social media. Reference [1] consider the social media domain and has a novel goal is evaluating the quality of data. They highlight some metrics, i.e., relevancy and timeliness. The social media data coming from Facebook and Twitter and this data are important for business decision making. Big data describe the volume of structured and unstructured data. Unstructured data is a huge resource for a business person to identify customer issues. The big data dimensions: Volume, variety, velocity, and veracity produce some challenges not only to data analytics but also in big data system like data quality. [1]

The users need to ensure data quality and trustworthiness of data [4]. Consider data quality is a facilitator of data trustworthiness. If data quality is low, users have to less confidence on a piece of data or information [3]. The authors [18] support to express trustworthiness issues includes provenance and quality issue. The user wants to ensure the reliability of the data while collecting. When data is proceeding for analyzing some data, the user wants to ensure that the relevancy and quality of data are appropriate the specific solution. Reliable and valuable data enhance decision making of business. The evaluation of data quality happens in data processing phase in big data architecture, data extraction, data processing, and decision making. Quality evaluation in big data considers while data goes through the pipeline of big data system [2] [6].

When understanding data quality from the user data point of view, then understanding the process of data for information retrieval is also important. Using search engine data are retrieved, a specific set of a keyword which makes up a user's query. User query helps to make decision process where the user makes judgmental value. Judgmental value helps to the user making a choice according to accuracy, usefulness, consistency [17]. Quality attributes are used to assess data quality and using metrics identify which measure improves data quality most? [7].

2. RELATED WORK

Recently, advanced technology and applications are rapidly increases such as smart mobile devices, data analytics, sensors, social network. It is possible to extracting, process and shares a large amount of data referred to as 'Big data.' Big data is relevant to many components like government, healthcare, business management, social media, education, life science. Using big data these components improve their decision making, transparency, quality by providing continuous monitoring. The challenges are arises when the volume of structured and unstructured data coming from different sources. In big data technology many technical challenges which are important to addressed. Discuss such challenges based on the big data processing pipeline [1]. The reason for generating a large volume of data, big data based application introduced new challenges and issues for quality assurance engineers [8]. These challenges are not only limited to data analysis but also to a big data system that manages all the data [1]. Recently social media data such as Twitter, Facebook increase business by providing insights into customer opinions, thoughts, and preferences. Design the platform that supports to monitoring and analyzing customer feedback in social media network and identifies issues which are faced by customers. Internet users communicate and express their thoughts with thousands of other people. People use a social media platform to share their thoughts and experiences with different customer products and services. Using batch version and real time version algorithm monitor and analyze customer feedback. [2]

When data are coming from different sources maintaining or evaluating the quality of data is also important. Data quality is more important in data warehouse system and management support system [6][9]. Data users need to insure quality and trustworthiness of data. When data is collected user want to the reliability of the data, when data is processing and analyzing the user wants to the relevancy of data which are suitable for the specific situation. Data quality is an important characteristic that decides the reliability of data for decision making. Evaluation of quality happens in data processing phase in big data architecture i.e. data extraction, data processing and analyzing and decision making. [1] Present a general framework for quality evaluation in social media. Define methods for identifying high quality content from the community is driven questions/answer sites. Focus on important social media source i.e. Yahoo! It hosts a wide range of questions/answers which are used for inspecting content quality in social media. Liked based method is used for several tasks in social media data. Linked based ranking algorithm is used to evaluate quality in social media data. [11]

Quality metrics are components of an effective quality management plan and measure properties of quality attributes. To evaluate and analyze the quality in any system, first need to characterize any quality attributes which are relevant to that system. Quality attributes such as Accuracy, performance, consistency, timeliness, completeness, relevancy [8] [9] are used to evaluate quality in social media data. Research aim [4] is providing trustworthiness metrics for information provenance and quality evaluation. Consider five metrics namely popularity, competence, corroboration, proximity, recency. The research consisted of 200 radar graphs (one per screen) with each graph present different value of five metrics. Many statistical methods are used to measuring data quality. When the statistical method has used the absence of data nature to measure the quality. Data mining is another method for measure the data quality. Propose data mining algorithm to measure the data quality. Three steps for measuring data quality. (1) Depending upon the input data (D) extract association rules. (2) Separate Compatible rule and incompatible rule. (3) Based on this rule calculate the quality of input data. Describe mathematical formula for measure quality in data. [5]

$$( ) = - + \sum_{=1} - \sum_{=1}$$

where, *r* is number of association rules depend on input data. rc is the number of compatible rules. Is the confident factor of the i^{th} and j^{th} compatible and incompatible rule respectively.

Figure 1 illustrates the economic management of data quality. This loop can be effect via data quality measures. Data quality measures improve the current level of data quality. This process leads to particular economic advantages. Based on the level of data quality and the threshold value, decide on taking a measure or not (Decision making). Data quality level quantifying quality based on data quality dimensions. The classification and identification of data quality dimensions describe based on the relevant situation. [7]

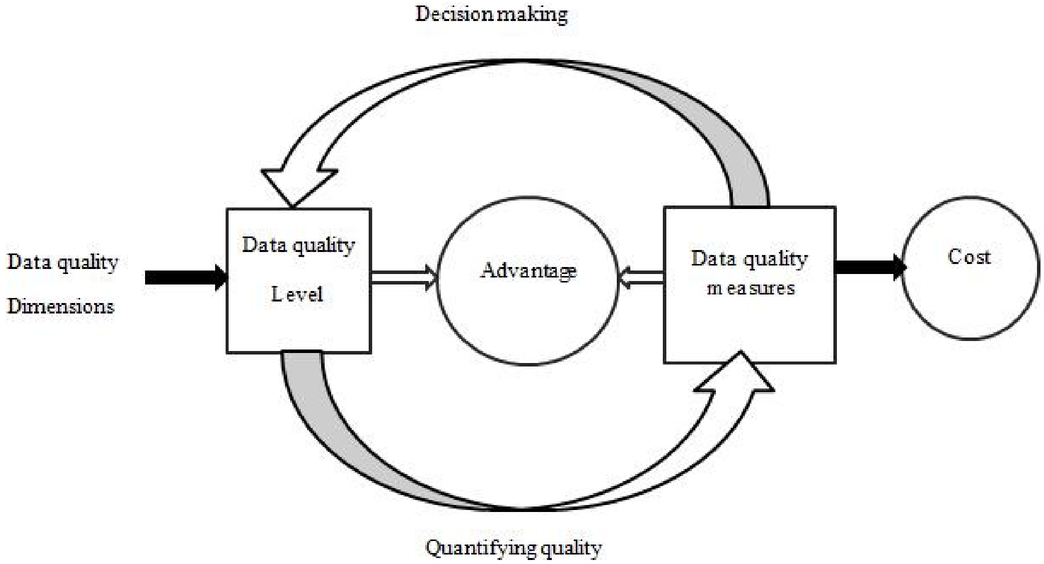

Figure 1: Data quality loop

Quality metrics measure the performance of product and processes. Each metrics has following properties [1].

- **Description:** metric description
- **Purpose:** the purpose of metric
- **Target:** where metric are used
- **Formula:** how the value of metric is achieved
- **Range value:** value of range for the metric evaluation
- **Acceptable value:** minimum value foe accepted quality attribute
- **Rules:** the set of measurement value range and a set of constraints which define a set of target measurement.

The following table describes Quality attributes and its definition. [1] [7]

Quality attribute	*Description of attribute*
Accuracy	Ensure that the error free data
Completeness	Check the data is not missing
Consistency	Implies that not two or more values conflict with each other
Relevancy	Data is relevant or not for particular situation

3. CONCLUSION

Evaluate data quality in big data sources is an interesting research field. Data quality is an important characteristic that decides the reliability of data for decision making. Evaluating data quality happens on unstructured data. The main purpose of this paper is to make researchers familiar with the how to evaluate data quality in the past, the current state of evaluating data quality and possibilities for the future. The survey of this paper carried out from data quality issues in big data sources. Defining data quality is complex and more difficult in the context of data retrieval from non-validated sources of big data. Researcher evaluates the quality using limited quality attributes. For future work extend the quality evaluation by introducing more quality attributes.

Acknowledgment

To prepare the Survey Paper of-" An assessment model to evaluate the big data-data quality" has been prepared by Supriya Haribhau Pawar and Prof.Dr. Devendrasingh Thakore.

I would like to thank my faculty as well as my whole department, parents, and friends for their support and confidence in me. I have obtained a lot of knowledge during the preparation of this document.

REFERENCES

[1] A. Immonen, P. Pääkkönen, and E. Ovaska, "Evaluating the Quality of Social Media Data in Big Data Architecture," vol. 3536, No. c, 2015.

[2] S. Bhatia, J. Li, W. Peng, and T. Sun, "Monitoring and Analyzing Customer Feedback Through Social Media Platforms for Identifying and Remedying Customer Problems," pp. 1147–1154, 2013.

[3] J. R. C. Nurse, S. S. Rahman, S. Creese, M. Goldsmith, and K. Lamberts, "Information Quality and Trustworthiness : A Topical State-of-the-Art Review," No. Iccans, pp. 492–500, 2011.

[4] J. R. C. Nurse, I. Agrafiotis, S. Creese, M. Goldsmith, and K. Lamberts, "Building Confidence in Information-Trustworthiness Metrics for Decision Support," pp. 7–10, 2013.

[5] S. Farzi and A. B. Dastjerdi, "Data Quality Measurement using Data Mining," vol. 2, No. 1, pp. 115– 118, 2010.

[6] E. Bertino, "Big data - Opportunities and challenges: Panel position paper," *Proc. - Int. Comput. Softw. Appl. Conf.*, pp. 479–480, 2013.

[7] P. Hans, U. Buhl, B. Heinrich, M. Kaiser, M. Klier, S. Rivard, and J. Webster, "How to measure data quality ? – a metric based approach by," vol. 4801, No. December 2007.

[8] C. Tao and J. Gao, "Quality Assurance for Big Data Application – Issues, Challenges, and Needs."

[9] G. Cong, W. Fan, F. Geerts, X. Jia, S. Ma, Z. Zhang, S. Wang, G. Cong, W. Fan, F. Geerts, X. Jia, and

[10] S. Ma, "Improving data quality: consistency and accuracy," *Proc. 33rd Int. Conf. Very large data bases*, vol. Vienna, Au, pp. 315–326, 2007.

[11] D. Firmani, M. Mecella, M. Scannapieco, and C. Batini, "On the Meaningfulness of 'Big Data Quality' (Invited Paper)," *Data Sci. Eng.*, vol. 1, No. 1, pp. 6–20, 2015.

[12] E. Agichtein, C. Castillo, D. Donato, A. Gionis, G. Mishne, "Finding high-quality content in social media," *Proceedings of the 2008 International Conference on Web Search and Data Mining WSDM '08*, pp. 183-194, 2008.

[13] S.S. Rahman, S. Creese, M. Goldsmith, "Accepting information with a pinch of salt: Handling untrusted information sources," *Security and Trust Management, Lecture Notes in Computer Science Volume 7170*, pp. 223-238, 2011.

[14] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM, vol. 56*(4), pp. 82-89, 2013.

[15] S. E. Madnick, R. Y. Wang, and Y. W. Lee, "Overview and Framework for Data and Information Quality Research," vol. 1, No. 1, pp. 1–22, 2009.

[16] A. Fabijan, H. H. Olsson, and J. Bosch, "Customer Feedback and Data Collection Techniques in Software R & D : A Literature Review," vol. 1, pp. 139–153, 2015.

[17] C. Castillo, M. Mendoza, and B. Poblete, "Information Credibility on Twitter," pp. 675–684, 2011.

[18] S.-A. Knight and J. Burn, "Developing a framework for assessing information quality on the world wide web," Informing Science Journal, vol. 8, pp. 160–172, 2005.

[19] E.Bertino and H.-S. Lim, "Assuring data trustworthiness – concepts and research challenges," in Secure Data Management, Ser. Lecture Notes in Computer Science, W. Jonker and M. Petkovic, Eds., 2010, vol. 6358, pp. 1–12.

[20] I. Taleb, R. Dssouli, M.A. Serhani, "Big data pre-processing: A quality framework," *IEEE International Congress on Big Data*, New York, pp. 191-198, 2015.

[21] L. Ramaswamy, V. Lawson, S.V. ogineni, "Towards a quality-centric big data architecture for federated sensor services," *IEEE International Congress on Big Data*, Santa Clara, CA, pp. 86-93, 2013.

[22] Supriya Pawar, "A study on Big data security and Data storage infrastructure", IJARCSSE, Volume 6 Issue 7, pp.1-6 publishes on July-2016.