IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Special Issue-2, March 2018 E-ISSN: 2347-2693

Quality Enhancement by Resolving Conflict using Optimization in Big data

P. Bastin Thiyagaraj [1*], A. Aloysius[2]

[1] Department of Information Technology, St. Joseph's College (Autonomous), Trichy – 620002 TamilNadu, India
[2] Department of Computer Science, St. Joseph's College (Autonomous), Trichy – 620002, TamilNadu, India

[*]*Corresponding Author:* bastinstar@gmail.com

Available online at: www.ijcseonline.org

***Abstract*—** Big data is referred as a term that describes volume of data (terabytes to Exabyte's), unstructured (include text and multimedia content), and complex in processing (from Medical data, Business transactions, Data capture by sensors, Social media/networks, Banking, Marketing, Government data, etc.). The traditional technologies are not sufficient to store, process and analyze the data. The unique technologies should be needed to analyze, manage the huge amount and unprocessed data. There are number of sources producing huge volume and variety of data. The number of sources produce amount of various descriptions for same object. This leads to data conflict and source conflict, when various sources generate various descriptions for same objects. Here it is the challenging one to identify which source produces quality information. The source could be identified by discovering the weight of the sources by using optimization method. Here optimization playing an important role to find highest achievable performance under the given constraints, by maximizing desired factors and minimizing undesired ones. In comparison, maximization means trying to attain the highest or maximum result or outcome without regard to cost or expense.

***Keywords*—** Big data, optimization, Reliability, Accuracy, Consistency and Integrity

I. INTRODUCTION

According to the Definition of Gartner "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making"[1]. The huge amount of unstructured data is generated from the various sources. For example, Face book generates over 500 terabytes of data everyday—including uploaded photos, likes, and users' posts [2]. In 2010, the world generated over 1ZB of data; and by 2014, we have generated 7ZB of data. IBM estimates that every day 2.5 quintillion bytes of data are created – so much that 90% of the data in the world today has been created in the last two years. The data unit is no longer the GB and TB, but the PB (1PB = 210TB), EB (1EB = 210PB), and ZB (1ZB = 210EB). According to IDC's "Digital Universe" forecasts (Gantz & Reinsel, 2012), 40 ZB of data will be generated by 2020[7]. Similarly number of sources produce amount of various descriptions for same object. For example Google for the query like ―the height of Mount Everest" include ―29,035 feet", ―29,002 feet" and ―29,029feet"[11]. When various sources generate various descriptions for same objects, the conflict occurred. It leads to data or information conflicts (Which information produced by the variety of sources is correct?) and Quality problems (which source produce quality of information?). The big data quality service applications provide the facilities to enterprises in problem identification, process improvement, productivity increase, efficient customization support, intelligent decisions, and optimized solutions [3]. It is very tedious task to analyze large volumes of data of different varieties. In addition, it raises challenges to design of data storage and database with various data format. Data variety increase with various branches of science and societal systems [4]. It is a challenging one to analyze the selection of quality source from the number of sources. Here this paper focuses on identifying selection of quality of source by reducing the conflict by using optimization method. Optimization implies that maximize the result, maximization means trying to attain the highest or maximum result or outcome without regard to cost or expense.

II. RELATED WORK

Big data Quality

According to ISO 9000:2015[5], data quality can be defined as the degree to which a set of characteristics of data fulfils requirements. *data quality assurance[3]:* Process of discovering inconsistencies, inaccuracy, incompleteness and other anomalies in the data, as well as performing data cleansing activities, data aggregation, to improve big data quality. *Big data quality assurance[3]:* the study and application of various assurance processes, methods, standards, criteria, and systems to ensure the quality of big data in terms of a set of quality parameters. The quality parameters are Availability, Usability, Reliability, relevance and Presentation Quality.

There are certain parameters involved in assessing quality. For example Data Accuracy, Data correctness, Data consistency, Data currency and timeliness are related to information quality and System quality parameters are System Reliability, System Adaptability, System Integration, System Accessibility, System Response Time, System Privacy in Big data Analytics [13]. There are a number of service oriented big data quality factors, which are Data usability, Data security & Data completeness, Data accessibility , Data , Data scalability[3]. The below figure describes the quality framework of big data,

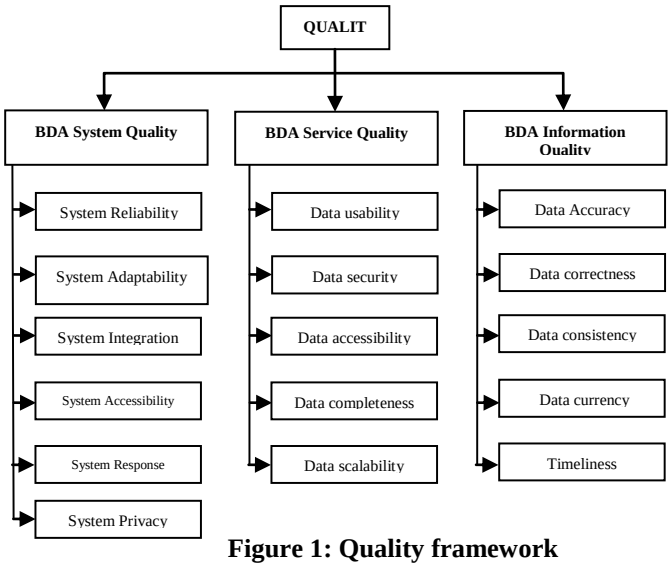

Figure 1: Quality framework

Here the system quality parameter is "Reliability" focused on to find the quality of source from conflicted sources. The elements of reliability are

- ***Accuracy***- Data provided are Accurate.
- ***Integrity***- Data is consistent with structural integrity and content integrity.
- ***Consistency***- During a certain time, data remain consistent and verifiable. Conflict handling strategies are high level strategies on how to handle inconsistent data [12].
- ***Completeness***- Whether the deficiency of a component will impact data accuracy and integrity

It is notable that in the complicated world, it is crucial to **estimate source reliability** to find the quality of source from amount of conflicted information especially when the sources providing low quality information, such as faulty sensors that keep emanating wrong data, and spam users who propagate false information on the Internet. However, there is no oracle telling us which source is more reliable and which piece of information is correct [6].

Optimization techniques

Finding an alternative with the most cost effective or highest achievable performance under the given constraints, by maximizing desired factors and minimizing undesired ones. In computer simulation (modelling) of business problems, optimization is achieved usually by using linear programming techniques of operations research. According to Stephen Boyd and Lieven Vandenberghe[8] "The optimization problem is an abstraction of the problem of making the best possible choice of a vector in R_n from a set of candidate choices when conflict occurred from many choices.One important problem is to identify the quality of source by identifying the true information (i.e., the truths) among conflicting sources of data.

III. METHODOLOGY

The general principle to find the truth by producing the weight of the various sources to select quality of sources for the optimization formulation given by Qi Li et al[6]:

$$f(\mathcal{X}^{(*)}, \mathcal{W}) = \sum_{k=1}^{K} w_k \sum_{i=1}^{N} \sum_{m=1}^{M} d_m(v_{im}^{(*)}, v_{im}^{(k)})$$

Here $\delta(W) = 1; W \in S:$.

Where,

W is the weight of source belong to the domain S

W_k **:** is a weight of the k sources(k_1, k_2, k_3,….,k_K)

N: Number of object that sources produce the information about the objects

M: properties of the object.

d_m: is the distance function, measures the difference between the information of the sources $(v_{im}^{(k)})$, and identified truths$(v_{im}^{(*)})$. d depends on data type of ***m***-th property.

$v_{im}^{(k)}$**:**Observed values of the objects provided by the various sources.

$v_{im}^{(*)}$ identified truths.

The main purpose of using the optimization method is to maximize the weight of the source and minimize the distance between observed values and true values. The quality of source could be identified by maximize the source weight W_k . If the source $W_{k=i}$ has the highest weight compare with other sources, it implies that $W_{k=i}$ produces the quality of information. So, this source is the quality one to produce the correct information than other sources. Here the quality parameter Reliability involved that is Accuracy of each and every sources is very important to identify the source whether provide the accurate information. In order to find the weight of the source, the following formula given by Qi Li et al [6].

$$w_k = -\log\left(\frac{\sum_{i=1}^{N}\sum_{m=1}^{M} d_m(v_{im}^{(*)}, v_{im}^{(k)})}{\sum_{k'=1}^{K}\sum_{i=1}^{N}\sum_{m=1}^{M} d_m(v_{im}^{(*)}, v_{im}^{(k')})}\right)$$

Here the weight of the sources could be identified using inverse log.

The reliability element *consistency* is playing an important role that the information of the objects provided by the various sources should be same. If the information is not same, conflict occurred. So that we should identify the consistency rate of the information to identify the weight of the various sources to select the best source in big data. The reliability element *integrity* also plays an important role that the data type of the properties should be clear. Different properties come in different data types, such as categorical data, continuous data, graph data etc[9].

So that in order to find the weight of the source, the data type is important. For example..
For continuous data, we choose normalized absolute deviation is as given below.

$$d_m(v_{im}^{(*)}, v_{im}^{(k)}) = \frac{|v_{im}^{(*)} - v_{im}^{(k)}|}{std(v_{im}^{(1)}, \cdots, v_{im}^{(k)})}$$

There are other functions used in identifying the deviation between observed values and truth values. The methods are Bregman divergence [10], which includes a variety of loss functions such as squared loss, logistic loss, ItakuraSaito distance, squared Euclidean distance, Mahalanobis distance, KL-divergence and generalized I-divergence[6].
For categorical data, we use ***0-1*** loss function

$$d_m(v_{im}^{(*)}, v_{im}^{(k)}) = \begin{cases} 1, & if\ v_{im}^{(*)} \neq v_{im}^{(k)} \\ 0, & otherwise \end{cases}$$

The selection of source weight depends on the Reliability. So that, by assumption the above formula can be rearranged as by multiplying Accuracy, consistency and integrity rate to identify the weight of the source to select the good quality of source in big data era when conflict occurred.

$$f(\mathcal{X}^{(*)}, \mathcal{W}) = \sum_{k=1}^{K} \quad \boldsymbol{Acc_K.con_K.int_M} \quad \sum_{i=1}^{N}\sum_{m=1}^{M} d_m(v_{im}^{(*)}, v_{im}^{(k)})$$

Where,

Acc_K: Accuracy of K sources, Con_K: Consistancy of K sources, int_M: integrity rate of properties.

IV. CONCLUSION AND FUTURE DIRECTION

The paper is about the big data quality enhancement and is an important part in an analytics and is the study of various assurance processes, methods, standards, criteria, and systems to ensure the quality of big data. This paper focused on that how the quality element "reliability" involved in identifying the selection of source, when producing the conflicted information from the various sources about the same objects. It leads to enhance the quality by using optimization methods in resolving conflicts. This concept is to be applied in streaming data provided by the different sources to identify the source is quality than other sources.

REFERENCES

[1]. Amir Gandomi and Murtaza Haider "Beyond the hype: Big data Concepts, Methods and analytics", International Journal of Information Management (IJIM) ELSEVIER, 2015, pp: 137-144.

[2]. Provost, F., & Fawcett, T. (2013).Data science and its relationship to big data and data driven decision making.Big Data, 1(1), 51–59.

[3]. Jerry Gao, Chunli Xie, Chuanqi Tao, "Big Data Validation and Quality Assurance –Issuses, Challenges, and Needs", 2016 IEEE Symposium on Service-Oriented System Engineering, 978-1-5090-2253-3/16 $31.00 © 2016 IEEE , DOI 10.1109/SOSE.2016.63, 433-41.

[4]. Kushal Patel ,"Big Data, its Issues and Challenges", 2017 IJEDR, Volume 5, Issue 3, ISSN: 2321-9939,123-27

[5].http://www.iso.org/iso/catalogue_detail?csnumber=45481)

[6]. Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han, "Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation," in Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM, 2014, pp. 1187–1198.

[7]. Cai, L and Zhu, Y 2015 The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14: 2, pp. 1-10, DOI: http://dx.doi.org/10.5334/dsj-2015-002.

[8]. S. Boyd and L. Vandenberghe. Convex optimization. Cambridge University Press, 2004

[9]. Fan Zhang, Li Yu, Xiangrui Cai, Ying Zhang, Haiwei Zhang,"Truth Finding from Multiple Data Sources by Source Confidence Estimation", 978-1-4673-9372-0/15 $31.00 © 2015 IEEE DOI 10.1109/WISA.2015.45, 153-56.

[10]. A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. JMLR, 6:1705–1749, 2005.

[11]. Arunima Kumari, Dr. Dinesh Singh, Reviewing Truth Discovery Approaches And Methods For Big Data Integration. International Journal of Science, Engineering and Technology Research (IJSETR), Volume 5, Issue 8, Pages: 2766-2774, August 2016.

[12]. J. Bleiholder and F. Naumann. Conflict handling strategies in an integrated information system. In *Proc. of IIWeb*, 2006.

[13]. Steven Ji-fan Ren, Samuel Fosso Wamba, Shahriar Akter, Rameshwar Dubey & Stephen J. Childe (2016): Modelling quality dynamics, business value and firm performance in a big data analytics environment, International Journal of Production Research, DOI:10.1080 /00207543.2016.1154209

Authors Profile

P. BASTIN THIYAGARAJ is working as an Assistant Professor in the Department of Information Technology, St. Joseph's college(Autonomous), Tiruchirappalli, Tamil Nadu, India. I am having 7 years of experience in teaching and 2 years in research.

Dr. A. ALOYSIUS is working as an Assistant Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has 16 years of experience in teaching and research. He has published many research articles in the National / International conferences and journals. He has acted as a chairperson for many national and international conferences. Currently, eight candidates are pursuing Doctor of Philosophy Programmed under his guidance.

