Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (4): 615-620, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.615.620

A Methodology to Explore Rules and Methods for Data Quality Dimensions Toward Improvement the Quality of Databases

Payam Hassany Shariat Panahy, Fatimah Sidi, Lilly Suriani Affendey, Marzanah A. Jabar, Hamidah Ibrahim and Aida Mustapha
Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor 43400, Malaysia

Abstract: Awareness of data quality dimensions and their relationships, poses new challenges to the database provider during the past two decades. Although, information systems have continuous improvement against their data problems, their success progressively depends on their methodology. This paper presents a methodology to measure, analyze and evaluate data quality dimensions by using subjective and objective measurement. Applying empirical methods and data mining techniques are steps of this methodology to improve database quality in the information systems. The applied rules and methods can be used to visualize and analyze attribute identification of the databases which is powerful and efficient to extract and reduce inconsistencies of the data. This methodology can be applied to compute other measurable quality dimensions and can help the information system providers to have intelligent and highly sophisticated opinions on creating databases.

Key words: Data quality dimensions, methodology, database quality

INTRODUCTION

Data is the most critical resources in various information systems as its quality gains competitive advantages. The effects of poor quality negatively decrease the quality of information which is needed for decision support. There are enormous problems among the data that eventually lead to failure in information systems. Fortunately, defining inconsistent and incoherent data quality problems can be a clue to minimize risks arising in information systems. Nevertheless, some related problems such as duplicate and inconsistent data can be eliminated by different methods and improving process. These methods and techniques can maximize quality of data via minimizing quality problems. However, different information systems have various databases with inconsistent data that need unified planning to build centralized data. As data are a fundamental information asset (Tejay *et al*., 2006) its quality should be identified in the environment from where it is created till it is used. Database provider should identify and eliminate data quality problems to achieve high quality data for decision-making and increasing trust between users and themselves. Data quality methodology and improvement processes have been developed to gain high quality data to reduce failure in information systems. In short, poor data quality can negatively affect on economic, financial and business process. To improve process quality in the database, we need to understand what it means to the data provider and users. For that reason, the purpose of this research is, to present a methodology to measure, analyze, evaluate and improve data quality by selecting the most appropriate data quality dimensions for improving quality in information systems.

Data is group of attributes and relations that alternatively refer to "raw material of information, set of facts and result of the conceptualization of the real world and the relationship" (Wang and Strong, 1996). Data quality refers to usage of the data in the system while the users judge the quality of producing data (Tejay *et al*., 2006). Data has multi-dimensional concept, which means it can be measured by different dimensions (Sidi *et al*., 2012; Slone, 2006; Eckerson, 2002). Wang and Strong (1996) extracted 20 dimensions from a set of 200 data quality attributes and then reduced them to 15 dimensions (Lee *et al*., 2002). Due to the variety of dimensions, selecting suitable dimensions is based on the scope of research (Sidi *et al*., 2012; Slone, 2006). Quality of data is an extended issue of dimensions relationship and is associated among its problems and it is also critical for data providers and users to support and resolve performance-related issues

Corresponding Author: Fatimah Sidi, Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Srdang, 43400 Selangor, Malaysia Tel:+60389471739

(Nyaboga and Mwaura, 2011; Batini *et al*., 2009; Eckerson, 2002) . Quality problems are classified into two categories; problem resolution and changing the production process (Batini *et al*., 2009). Only data with high quality shall build confidence and trust for both data providers and users. Gaining high quality is possible by changing organizational activities and shifting process. Data quality is a multidisciplinary field that covers an extensive variety of topics. The Management Information System (MIS) and Computer Science (CS) are two major disciplines that do research on the impact of data quality (Madnick *et al*., 2009). The relationship between data quality and decision outcome was an extensive research during the last decades. Quality of data is defined by data quality dimensions that contain both subjective and objective information. The subjective information about quality dimension refers to understanding quality criteria based on user requirement and the objective information refers to data types and its process (Li and Osei-Bryson, 2010).

There is no distinction between data and information quality and, usage of theses terms is highly inconsistent in different research (Nyaboga and Mwaura, 2011). In line with this study, the intention of data quality is to technical issues and intention of information quality is in nontechnical issues. In technical issues, data is integrated from different resources while in nontechnical issues there is a lack of a cohesive strategy about stakeholders, information, format, place and time. So, technical solution will improve data quality while information quality will improve through developing standards and strategy. Manual annotation is an example of technical solution and making policies is an example of developing strategy (Nyaboga and Mwaura, 2011). Hence, data quality is a key element for enhancing process quality in database systems. Moreover, the collection of data quality dimensions by survey and measuring them empirically provide hierarchical and dimensional assessment of data quality (Li and Osei-Bryson, 2010). Researchers recommend to focus only on the parts of database and data flows that increase main problems because improving process quality is expensive and complex (Batini *et al*., 2009). Developing database technology is possible through identifying quality problems for assessing, improving and managing quality of data in the database system (Madnick *et al*., 2009).

A data quality methodology is used to measure and improve the quality of data by applying set of strategies and techniques (Madnick *et al*., 2009). There are three approaches to study data quality; initiative, theoretical and empirical (Wang and Strong, 1996):

- **Initiative approach:** It selects important attributes for specific study based on the researchers' understanding or experiences. This approach is commonly used in the field of data quality studies and its effects is on the small set of data quality attributes such as accuracy which is a key attribute in this field
- **Theoretical approach:** It focuses on lack of quality during the manufacturing process and observes inconsistence data in the real world system
- **Empirical approach:** It determines the characteristics of data based on analyzing collected data from the users. The features of data cannot be determined by theoretical or initiative approach (Wang and Strong, 1996)

Based on the literature, product quality refer to data quality that include the tangible measure of information quality and, service quality mention to information quality that include intangible measures and is related to service delivery process. Completeness and accuracy are examples of product quality while security and manipulation are in the category of service quality (Peppard *et al*., 2010). Researchers have identified several quality assessments and models, but most of them are utilized in the context of data warehousing and business domain and their aim is to solve quality problems as a service quality (Martinez, 2007). These methodologies do not fit into the database system because in order to improve quality in the information system we should look at data quality as a product quality not service quality. Therefore, an appropriate methodology is needed to visualize and analyze attribute identification of databases that can reduce inconsistencies of the data by computing the dependencies for improving database quality.

MATERIALS AND METHODS

In academia, the influence of the data quality dimensions of improving processes and efficiency of data quality in information systems is explained widely (Jing-Hua *et al*., 2009). The main objective of this study is to introduce a methodology to; measure and assess data quality dimensions by using subjective and objective measurement. The qualitative and quantitative methods are used to analyze existing dependency among dimensions and to measure data quality dimensions. Further, data mining techniques are applied to improve database quality in databases.

Qualitative and quantitative methods: The initiative approach is chosen to select commonly used data quality dimensions to improve database quality in information systems. As researchers suggest to apply more than one method to measure data quality because data quality has multi-dimensional concept (Pipino *et al*., 2002;

Madnick *et al*., 2009; Lee *et al*., 2002) so, our methodology consists of both subjective and objective measurements to assess the quality of data.

As, the contribution of this research is reducing inconsistencies of the data in databases, we identify and analyze problems of data inconsistencies in database for information system by developing and validating a model for data quality dimensions based on applying IQ instrument (Information Quality). The IQ instrument is used as a qualitative approach to measure and validate selected dimensions with the aim of establishing the meaning of dimensions from the view of the participants who are working with different database in information systems. Next, empirical methods as a quantitative approach apply to data which gathered from the survey and interview to compute, analyze and explore dependency among dimensions which identifies the causes of data inconsistencies. After that, we measure each dimension based on attribute identification and using specific metrics in a database.

Data mining techniques: After measuring dimensions, we use and explore data mining techniques and algorithms to compute the data quality dimensions and evaluate database quality to find inconsistence data. Data mining techniques as part of this methodology has ability to detect inaccurate, inconsistent and incomplete data in order to improve the quality in database through data cleansing and appropriate methods. Generally data preprocessing decreases errors and enhance quality process. Improve data quality inconsistency in the database by proposing and assessing rules and methods is another step to visualize best and appropriate patterns and rules on the database. Implementation and simulation for validation the rules on a database is another step that assists managers to make better decision to create and improve database quality. Finally, the last step is refining and finalizing the rules and methods and put policies and enforcement to define consistence attributes of data for future and to evaluate improvement quality frequently. The flow chart of the methodology is shown in Fig. 1. To

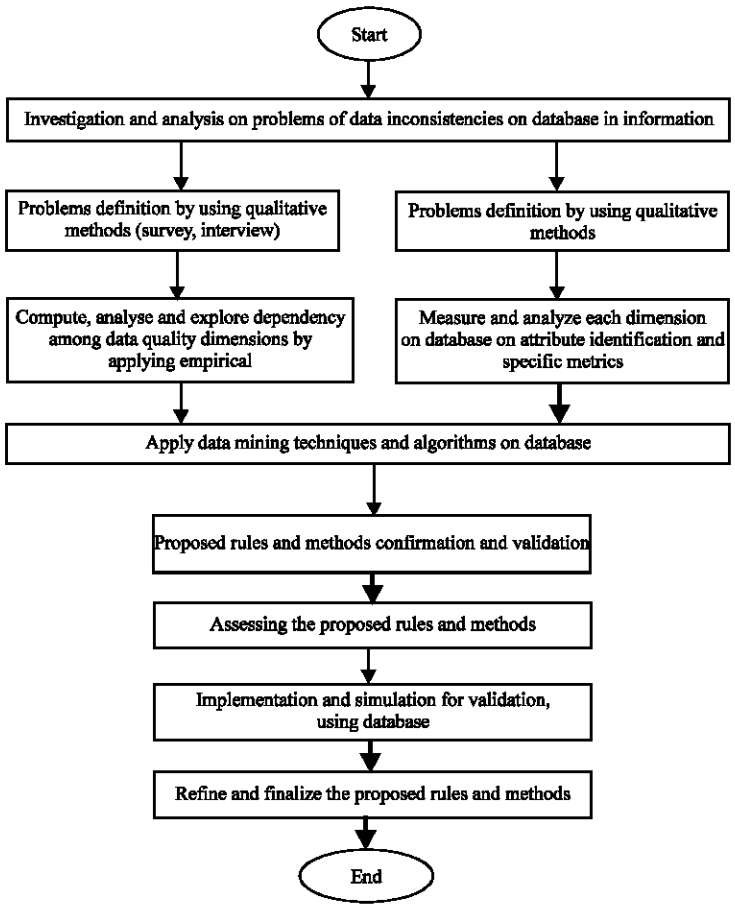

Fig. 1: Flow chart of methodology for measuring and assessing data quality dimensions

conclude, our methodology has both subjective and objective assessments to find and compute the causes of the low quality in the databases which is more effective and efficient than the others' methods.

RESULTS

This methodology provides comprehensive assessment from the view of the database provider and users to evaluate databases. It can determine data quality problems and identify which dimensions has more impact on others. Thus, after gaining a database which has good quality it can be used as a benchmark in other database and can be applied in other information systems or organizations. The proposed method can make intelligent and highly sophisticated opinions and it will carry useful changes in improving processes by minimizing the cost since redundancy and overlapping activities and avoided.

Since, there has been no model and algorithm to evaluate data quality dimensions dependency and nobody applied both empirical methods and data mining techniques to improve quality, our methodology shall be more powerful and efficient to extract, detect and improve inconsistent data.

DISCUSSION

In order to improve data quality, we should know that; important finding for data provider is not equally important for data users. Ranking data quality dimensions of the current state of the competition depend on an individual information systems. Researchers have explained different definition for each dimension based on their domain. Since, there is difficult to give an exact and a common definition for each dimension, measurement has been taken and evaluated over time (Lee *et al*., 2002). Dimensions are major object to measure data quality and some of them are frequently used in database community such as consistency, completeness and accuracy. In order to select suitable quality dimensions that can be assessed by our method , We look for dimensions that could be objectively and subjectively measured.

Nowadays, despite the quality process and control that exist at every stage of production (Guo *et al*., 2010) the improvement process still faces various problems. If these problems are not identified and modified constantly it can lead to failure process in information systems. Finding quality problems and inconsistent data through different techniques and methods is a clue to reduce future failures. Inconsistent data are required to be modified to create a standard database based on their environmental usage. Also, database provider needs to identify and remove data problems to enhance the quality. Improving processes quality increase user satisfaction and leads to encourage data provider to make better decision. So, methodologies are developed to diminish failures in information systems. In brief, low data quality negatively affect on economic, financial and business process. The purpose of this research is, to present the steps of a methodology to measure and improve data quality on databases in information systems. Qualitative and quantitative methods apply to set of dimensions and database to measure the quality of the database and then empirical methods and data mining techniques apply to improve the quality process in the database environment.

Moreover, if this methodology be applied to compute other measurable quality dimensions it can assist Information System (IS) providers to have intelligent and highly sophisticated opinions on creating databases and to make accurate decisions for implementing effective and efficient information by defining and enforcing policies.

Based on the previous studies, we identified data quality as a product quality which involves the tangible item or measure. Tangible item refers to the data that can be produced, store and use later (Kahn *et al*., 2002). Completeness, accuracy, consistency and timeliness are examples of data quality dimensions that can be measured in database systems and are suggested by most database provider and data users (Ge and Helfert, 2008). So, we select these four dimensions that frequently used in information quality researches. These dimensions are in the category of intrinsic and contextual dimensions and can be measured by questionnaire and specific formula. We apply IQ instrument which previously applied in different field such as industry, health care and financial (Kaiser *et al*., 2007) as a subjective measurement. The IQ instrument is used as a qualitative approach to validate the propose dimensions in different information system with the aim of establishing the meaning of dimensions from the view of the participants. The instrument for the quality improvement process is adopted of (Shariat Panahy *et al*., 2013a) which is developed based on comprehensive and extensive literature and perspective of information systems users (Shariat Panahy *et al*., 2013b). Then, we measure value of each dimension based on a previously approved metrics (Heinrich *et al*., 2007; Dong *et al*., 2006). In the next step of the methodology, we apply the empirical methods on data which is collected from questionnaires to validate the model and to analyze existing dependencies among data quality dimensions. Applying data mining techniques on the database, assessing rules and methods to visualize appropriate patterns and rules and, implementation and simulation of validation the rules on a database are the

next steps of the methodology that help managers to make better decision to create and improve database quality. Furthermore, when we refine and finalize the rules and methods, we will enforce some policies for defining consistence attributes of the data and for evaluating improvement quality regularly for the future.

In line with this research, this study sets a methodology to assess dimensions of data quality with the intention to discover relationships among dimensions. In addition to estimate the quality of data based on subjectively techniques and objective measure, our method uses empirical methods and data mining techniques to evaluate and improve data quality process via a set of enforcement policies. So, our methodology shall be more effective and efficient than other methods to measure, analyze, evaluate and improve inconsistent data towards enhancing database quality.

CONCLUSION

This At present, significant progress has been made in research on data quality. However, there is a need to develop new methods and techniques for improving data quality in new forms that can be benchmarked for different database in information systems. In this article, we have presented a methodology to measure and improve data quality in information systems. Applying the steps of this methodology is precious effort to understand the fundamental component of assessment to improve data quality dimensions. Furthermore, an empirical method helps us to investigate the relationships among dimensions and, using data mining techniques can improve efficiency and process quality in database systems.

The key contributions of this research are:

- To provide a methodology to assess, identify and evaluate quality in database systems through interrelated quality dimensions
- To interpret steps of the methodology to assess data quality in the information systems which is helpful for database provider
- To improve quality in information systems through enforcement policies on database systems

This methodology is useful in identifying data quality problems and helps database providers to make intelligent decisions and carry useful changes in the processes. It reduces the process cost since redundancy and overlapping are avoided. However, the effect and relationship among other data quality dimensions are needed to be considered and evaluated to select the most suitable improvement process in information systems.

ACKNOWLEDGMENT

Thanks to Fundamental Research Grant Scheme (FRGS), Ministry of Higher Education (MOHE), Malaysia. (FRGS Number: 03-12-10-999FR).

REFERENCES

Batini, C., C. Cappiello, C. Francalanci and A. Maurino, 2009. Methodologies for data quality assessment and improvement. ACM Comput. Surv., 41: 16-52.

Dong, C., S D.F.M. Sampaio and P.R.F. Sampaio, 2006. Expressing and processing timeliness quality aware queries: The DQ^2L approach. Proceedings of the Workshop on Advances in Conceptual Modeling-Theory and Practice, November 6-9, 2006, Tucson, AZ., USA., pp: 382-391.

Eckerson, W., 2002. Data warehousing special report: Data quality and the bottom line. Applications Development Trends. http://www.estv.ipv.pt/PaginasPessoais/jloureiro/ESI_AID2007_2008/fichas/TP06_anexo1.pdf

Ge, M. and M. Helfert, 2008. Modeling data quality in information chain. Proceedings of the International Conference on Business Innovation and Information Technology, January 24-25, 2008, Ireland.

Guo, W., H. Liang, L. Wang, D. Wang and J. Lv, 2010. Research on multi-dimension model of collaborative quality control in manufacturing network. Proceedings of the International Conference on Information Science and Management Engineering, Volume 2, August 7-8, 2010, Xi'an, China, pp: 331-336.

Heinrich, B., M. Kaiser, M. Klier, S. Rivard and J. Webster, 2007. How to measure data quality? A metric based approach. Proceedings of the 28th International Conference on Information Systems, Volume 4801, December 9-12, 2007, Montreal, Canada, pp: 101-122.

Jing-Hua, X., X. Kang and W. Xiao-Wei, 2009. Factors influencing enterprise to improve data quality in information systems application-An empirical research on 185 enterprises through field study. Proceedings of the International Conference on Management Science and Engineering, September 14-16, 2009, Moscow, Russia, pp: 23-33.

Kahn, B.K., D.M. Strong and R.Y. Wang, 2002. Information quality benchmarks: Product and service performance. Commun. ACM., 45: 184-192.

Kaiser, M., M. Klier and B. Heinrich, 2007. How to measure data quality? A metric-based approach. Proceedings of the 28th International Conference on Information Systems, Volume 4801, December 9-12, 2007, Montreal, Canada, pp: 101-122.

Lee, Y.W., D.M. Strong, B.K. Kahn and R.Y. Wang, 2002. AIMQ: A methodology for information quality assessment. Inform. Manage., 40: 133-146.

Li, Y. and K.M. Osei-Bryson, 2010. Quality factory and quality notification service in data warehouse. Proceedings of the 3rd Workshop on Ph.D. Students in Information and Knowledge Management, October 26-30, 2010, Toronto , ON., Canada, pp: 25-32.

Madnick, S.E., R.Y. Wang, Y.W. Lee and H. Zhu, 2009. Overview and framework for data and information quality research. J. Data Inform. Qual., Vol. 1, No. 1.

Martinez, A., 2007. BIODQ: A model for data quality estimation and management in biological databases. Ph.D. Thesis, University of Florida, USA.

Nyaboga, A.B. and M.F. Mwaura, 2011. Strategies for gaining competitive advantage in a dynamic environment thru data quality. Int. J. Manage. Inf. Syst., 13: 13-22.

Peppard, J., A. Koronios and J. Gao, 2010. The data quality implications of the servitization-theory building. Proceedings of the 4th World Congress on Engineering Asset Lifecycle Management, September 28-30, 2009, Athens, Greece, pp: 230-235.

Pipino, L., Y.W. Lee and Y.W. Richard, 2002. Data quality assessment. Commun. ACM., 45: 211-218.

Shariat Panahy, P.H., F. Sidi, L.S. Affendey, M.A. Jabar, H. Ibrahim and A. Mustapha, 2013a. A framework to construct data quality dimensions relationships. Indian J. Sci. Technol., 6: 4422-4431.

Shariat Panahy, P.H., F. Sidi, L.S. Affendey, M.A. Jabar, H. Ibrahim and A. Mustapha, 2013b. Discovering dependencies among data quality dimensions: A validation of instrument. J. Applied Sci., 13: 95-102.

Sidi, F., P.H. Shariat Panahy, L.S. Affendey, M.A. Jabar, H. Ibrahim and A. Mustapha, 2012. Data quality: A survey of data quality dimensions. Proceedings of the IEEE International Conference on Information Retrieval and Knowledge Management, March 13-15, 2012, Kuala Lumpur, Malaysia, pp: 300-304.

Slone, J.P., 2006. Information quality strategy: Anempirical investigation of the relationship between information quality improvements and organizational outcomes. Ph.D. Thesis, Capella University, Minneapolis, MN., USA.

Tejay, G., G. Dhillon and A.G. Chin, 2006. Data quality dimensions for information systems security: A theoretical exposition (Invited Paper). In: Security Management, Integrity and Internal Control in Information Systems, Dowland, P ., S. Furnell, B. Thuraisingham and X.S. Wang (Eds.). Springer, New York, ISBN-13: 9780387298269, pp: 21-39.

Wang, R.Y. and D.M. Strong, 1996. Beyond accuracy: What data quality means to data consumers. J. Manage. Info. Syst., 12: 5-34.