Contents lists available at GrowingScience

Management Science Letters

homepage: www.GrowingScience.com/msl

A new approach to measure believability dimension of data quality

Seyed Mohammad Hossein Moossavizadeh[a*], **Mehran Mohsenzadeh**[a] and **Nasrin Arshadi**[b]

[a]*Department of Computer Science and Research Branch, Islamic Azad University, Khouzestan, Iran*
[b]*Department of Psychology, Shahid Chamran University of Ahwaz, Ahwaz, Iran*

ARTICLEINFO

Article history:
Received April 28, 2012
Accepted 6 July 2012
Available online
July 11 2012

Keywords:
Data quality
Believability
Quality dimensions
Assessment of data quality dimensions
Data believability

ABSTRACT

Today's methodologies for data quality assessment and improvement are considerably aimed at reducing costs. Data quality comprises different dimensions, each having certain methods and techniques to assess and improve data quality. One of the most controversial dimensions is data believability in which less attention has been paid by scholars and researchers, because of its ambiguous nature. This is categorized under the "intrinsic data quality" dimensions. The current paper offers a precise and comprehensive definition of such quality dimension, and provides some parameters to understand it. In order to calculate these parameters, furthermore, different methods are discussed.

1. Introduction

In recent years, data quality has been received widespread attention for different reasons in the areas of information system management and leadership. Among the most important reasons for this is the high level of costs for production, maintenance and application of poor data quality. Data quality can mean "fitting to use" (Juran & Gryna, 1980). In fact, the higher the data quality is, the higher the analysis accuracy is. Data quality *assessment* is the first step for data quality *improvement*. In general, different methodologies use three different techniques to count up data quality variables; namely, Simple Ratio, Min-Max, and Weighted Average (Pipino et al., 2002).

Data quality comprises different dimensions, which have been frequently studied; however, distinct methodologies apply different number of quality dimensions (Batini et al., 2009). One of the common, important and effective dimensions of data quality is "believability". Owing to being unknown and ambiguous, it can also include various controversial sub-dimensions. Ambiguity is the

* Corresponding author.
E-mail addresses: mh.moossavizadeh@khouzestan.srbiau.ac.ir (S. M. H. Moossavizadeh)

doi: 10.5267/j.msl.2012.07.007

result of its psychological nature, leading to much more difficulties in quantifying and measuring this dimension relative to the rest (Moossavizadeh et al., 2011; Moossavizadeh et al., 2012).

2. Literature review

Many well-known methodologies for data quality such as TDQM, AIMQ, DQA, and AMEQ use the same concept of believability (Batini et al., 2009); although, very little research on this dimension has been yet conducted in details, and the result is just some general guidelines. Prat and Madnick (2008) published the most famous study of this area where the measurement of data believability was done by assessing background information. Indeed, the main idea of the paper was that the background of data production and its resource information can be maintained, and processed and evaluated in order to find whether data in question is believable. In addition to the list of some sub-dimensions of data believability, the paper sought to provide new approach for measuring this dimension, according to a "background-based" model. This approach covers three structures:

1- to define criteria for evaluating the believability of data *resources*,
2- to define criteria for evaluating the believability of data resulted from a *process*, and
3- to determine data believability as a whole.

However, there are some disadvantages; including the high costs of recoding data resource information, and of checking consistency between data items; also the necessity of keeping data backgrounds in order to compare current and previous values.

3. Definition and methodology to measure believability

In the present study, the proposed approach is based on data by its own elements; and to determine the extent of data believability it avoids assessing similar data and keeping additional information, where possible. The reason is the high level of operation costs.

3.1 Definition of Data Believability

By summarizing all definitions of data believability, it can be provided that "Believability is the extent in which the data is accepted, in a specific environment and in accordance with relevant rules; as true, or as item which seems to be true, real and credible".

The above definition includes three parts:

"Acceptability" is a clear term related to the concept of believability.

The phrase of "in a specific environment and in accordance with relevant rules" ensures the generality of the definition based on which local rules and conditions of an operating environment can be covered.

Furthermore, the ontological and psychological conceptions of belief necessitate that believable data in any information system should:

- either be *really* true and consistent with actual data;
- or *appear* as true, real or credible data.

Indeed, the phrase of "as true, or as item which seems to be true, real and credible" refers to such ontological meaning.

Therefore, it can be concluded that the proposed definition has the generality required.

3.2 Criteria for data believability

Through the evaluation of believability concept and its organizational application, relevant parameters can be determined. In order to determine a homogenous range of values, the final values are clamped to the [0 1] range.

3.2.1 Accuracy (A)

Accuracy means data stored in information system is consistent with its external format in the real-world setting (Batini et al., 2009). According to this, one sufficient criterion to make a data item believable is the accuracy. The degree of data believability will be very high and may be called inevitable if the data is accurate. However, there are exceptions where some data are unbelievable, while being accurate.

This criterion in believability area is simply inspired by data integrity assessment techniques and should be evaluated within the accuracy dimension of data range.

There are several methods to measure the above parameter, for example see (Batini et al., 2009, Batini & Scannapieco, 2006; & Olson, 2002).

3.2.2 Consistency with Rational and Organizational Rules (O)

Based on a set of rules, it is necessary to use significant relationships between different data fields in a database and a predetermined rule sets in an environment and to measure data believability in rational and organizational terms. It is known that the rational possibility to believe a data item in any environment shares a very close relationship with environmental and organizational rules, and vice versa. In other words, a data may be rationally accepted, while being unbelievable in a certain environment (due to some environmental and organizational rules). In contrast, data can be acceptable in an environment based on organizational rules; however, it could not be accepted by rational terms. To determine rationally consistency level with rational and organizational rules needs to use insights from organization experts.

Therefore, experts' insight needs to be taken into account when measuring this parameter. To this end, a linear diagram (0 to 1) is used to show the extent of data consistency with rational and organizational rules.

Consistency
Inconsistency

Fig.1. Linear diagram to Determine Data Consistency with Rational and Organizational Rules

3.2.3 Resource Appropriateness (R)

Certainly, the "more appropriate" the resource is, the more believable the resultant data is. To determine the extent of appropriateness in any data resource requires the measurements of its two related components; namely, "reliance of a resource" and "relevance of a resource to data".

It can be better shown as below:

Resource Appropriateness = Reliance of resource × relevance of a resource to data

Prat and Madnick (2008) provided a detailed procedure to assess the appropriateness of data resource.

3.2.4 Consistency with Previous Experiments (X)

Sometimes and in some environments, the extent of un/believability of data can be recognized according to previous experiences. It is probably regarded as the most difficult criterion to measure related to data believability. The only way to determine this parameter is to use insights from organization experts; since experts know about organizational experiences and only they can measure the consistency level for data and experiences. Experts' insight plays a deterministic role in understanding un/believability of any data. Obviously, insights from any organization experts can be reliable just within the same organization.

Similarly, a linear diagram (0 to 1) can be used to determine experts' insight.

Fig. 2. Linear Diagram to Determine Data Consistency with Organizational Experiences

On the other hand, an alternative approach to measure this parameter consists of providing a set of organizational experiences in the form of a rule set, and assesses and report data consistency with the set by using fuzzy methods. This approach has a good potential for future research.

4. Aggregation of believability parameters

The simple way to sum up believability parameters is to obtain average values. However, since the importance of believability parameters differs less or more from an organization to another, the best method to determine the extent of believability for a data item is to use the weighted average method and organization experts determine the parameters of weights. Involving experts' opinion on counting up believability parameters provides generality and integrity requirements to apply on different organizations. There are:

WA: Weight of accuracy

WO: Weight of data consistency with rational and organizational rules

WR: Weight of data resource appropriateness

WX: Weight of data consistency with previous experiences

Therefore, the extent of believability of a data item can be given as the following:

$$Believability(DI) = \frac{(W_A.A) + (W_O.O) + (W_R.R) + (W_X.X)}{\sum_i W_i}; i \in \{A, O, R, X\} \qquad (1)$$

5. Verification

In order to verify the proposed method, three case studies were conducted: two in the National Iranian Drilling Company, and one in Khuzestan Steel Company.

The National Iranian Drilling Company has a working unit to determine the angle for drilling locations. Indeed, according to excavation maps, drilling must be regulated in a certain direction. This unit is responsible for measuring angles. Data items include drill angles, depth of current drilling, necessary depth, etc. Some items have digital and momentary sensors, while others are with mechanical and periodical sensors.

In order to verify the method, believability parameters were first measured on two data items including "the angle of excavated area in north-south direction" and "the drilling angle in east-west direction" and the final values were obtained. Then, the finding was compared with experts' insight.

Case 1: Trough a formal measurement method, the first data item (at angle in a north-south direction=3.2) was obtained for the company. Based on the proposed method, believability parameters here were: (Accuracy= 1), (Data consistency with rational and organizational rules = .7), (Resource appropriateness= .85), (Data consistency with previous experiences = .8). According to the insights from organization experts, the weighted values for these parameters included: (Accuracy= 3), (Data consistency with rational and organizational rules = 2), (Resource appropriateness= 1), (Data consistency with previous experiences = 5). In sum, the extent of believability of data items in question equals .925, which find an approximation to experts' opinion (.9).

Case 2: For the second data item (at angle in east-west direction=5°), the following results were obtained: (Accuracy= 0), (Data consistency with rational and organizational rules = .65), (Resource appropriateness= .9), (Data consistency with previous experiences = .9). According to the weighted values determined by organization experts, therefore, the extent of believability is equal to .67, which acceptable compatibility with experts' opinion is seen (.6).

Case 3: In Khuzestan Steel Company, a data item for "daily production of steel ingots" was studied. The company usually measures the item by using a weighing scale. For the present research, the following results were obtained for the availability of data item (weighting 2370 tones); Accuracy= 0), (Data consistency with rational and organizational rules = .7), (Resource appropriateness= .6), (Data consistency with previous experiences = .3). According to the insights from organization experts, the weighted values for these parameters included: (Accuracy= 2), (Data consistency with rational and organizational rules = 2), (Resource appropriateness= 2), (Data consistency with previous experiences = 4). In sum, the extent of believability of the data item in question equals .38, compatible with experts' opinion.

6. Conclusion

Many data quality methodologies deal with believability as an intrinsic data quality dimension. However, the conception has not been clearly studied and well defined. No reference to parameters the believability dimension includes has been used in the current methodologies. Furthermore, no approach has been provided to measure these parameters in general, and believability in particular, based on data own information. Addressing the necessary comprehensive, the present paper has offered a new definition for data believability and determined those parameters affecting it. In addition, different methods have been discussed to measure these parameters. Results obtained from case studies show that the proposed method to assess data believability (including definition, measurement, and aggregation of parameters) is consistent with the insights from experts of different organization.

Acknowledgment

This research is indebted to our merciful God and the daughter of Great Prophet of Islam, holy Fatima (pbuh). Special thank goes to my wife “Zahra” and my honey daughter “Mohaddeseh”, because of their supports on my researches. Additionally thanks to my dear parents.

References

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3), 1-52.

Batini, C., & Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques*. Springer Verlag.

Juran, J. M., & Gryna, Jr, F. M. (1980). *Quality Planning and Analysis*. 2nd ed., McGraw-Hill, New York, 1980.

Moossavizadeh, S.M.H., Mohsenzadeh, M., & Arshadi, N. (2011). A New Precautionary Method for Measurement and Improvement of the Data Quality. *International Conference on the Software and Knowledge Engineering,* Paris.

Moossavizadeh, S.M.H., Mohsenzadeh, M., & Arshadi, N.A. (2012). New algorithmic approach to detect the good point access in the precautionary process for data quality. *2nd IEEE International Conference on Computer Science and Service System*, China.

Olson, J. E. (2002). *Data Quality - The Accuracy Dimension*. Morgan Kaufmann Publishers.

Prat, N., & Madnick, S. (2008). Measuring data believability: A provenance approach. *Proceedings of the 41st Hawaii International Conference on System Sciences*, 393.

Pipino, L., Lee, Y., & Wang, R. (2002). Data Quality Assessment. *Communications of the ACM,* 45(4), 184-192.