IMPROVING DATA QUALITY IN DATA WAREHOUSING APPLICATIONS

Lin Li, Taoxin Peng, Jessie Kennedy
Edinburgh Napier University, 10 Colinton Road, Edinburgh, UK EH10 5DT
l.li@napier.ac.uk t.peng@napier.ac.uk j.kennedy@napier.ac.uk

Keywords: Data quality, Data quality dimension, Data quality rules, Data warehouses.

Abstract: There is a growing awareness that high quality of data is a key to today's business success and dirty data that exits within data sources is one of the reasons that cause poor data quality. To ensure high quality, enterprises need to have a process, methodologies and resources to monitor and analyze the quality of data, methodologies for preventing and/or detecting and repairing dirty data. However in practice, detecting and cleaning all the dirty data that exists in all data sources is quite expensive and unrealistic. The cost of cleaning dirty data needs to be considered for most of enterprises. Therefore conflicts may arise if an organization intends to clean their data warehouses in that how do they select the most important data to clean based on their business requirements. In this paper, business rules are used to classify dirty data types based on data quality dimensions. The proposed method will be able to help to solve this problem by allowing users to select the appropriate group of dirty data types based on the priority of their business requirements. It also provides guidelines for measuring the data quality with respect to different data quality dimensions and also will be helpful for the development of data cleaning tools.

1 INTRODUCTION

A great number of data warehousing applications have been developed in order to derive useful information from these large quantities of data. However, investigations show that many of such applications fail to work successfully and one of the reasons is due to the dirty data. Due to the 'garbage in, garbage out' principle, dirty data will distort information obtained from it (Mong, 2000). Nevertheless, research shows that many enterprises do not pay adequate attention to the existence of dirty data and have not applied useful methodologies to ensure high quality data for their applications. One of the reasons is a lack of appreciation of the types and extent of dirty data (Kim, 2002). Therefore, in order to improve the data quality, it is necessary to understand the wide variety of dirty data that may exist within the data source as well as how to deal with them. This has already been realized by some research works already (Rahm and Do, 2000, Müller and Freytag, 2003, Kim, Choi, Hong, Kim and Lee, 2003, Oliveira, Rogriques, Henriques and Galhardas, 2005). However, in practice, cleaning all data is unrealistic and simply not cost-effective when taking into account the needs of a business enterprise. The problem then becomes how to make such a selection. In this paper, this problem is referred to as the Dirty Data Selection (DDS) problem. This paper presents a novel method of classifying dirty data types from a data quality dimension angle, embedded with business rules, which has not previously been considered in the literature. The proposed method will help to solve this problem by allowing users to select the appropriate group of dirty data types to

deal with based on the priority of their business requirements.

The rest of the paper is structured as follows: in section 2, data quality, data quality dimensions and data quality rules that are used for the proposed method are discussed. Dirty data types which is used for the classification is presented in section 3. The proposed method is given in section 4. An example of using the method to deal with the DDS problem is demonstrated in section 5. Finally, the paper is concluded and future work is discussed in section 6.

2 DATA QUALITY, DATA QUALITY DIMENSIONS AND DATA QUALITY RULES

2.1 Data Quality

From the literature, data quality can be defined as "fitness for use", i.e., the ability of data to meet the user's requirement. The nature of this definition directly implies that the concept of data quality is relative. For example, an analysis of the financial position of a company may require data in units of thousands of pounds while an auditor requires precision to the pence, i.e., it is the business policy or business rules that determine whether or not the data is of quality.

2.2 Data Quality Dimensions

According to Wang and Strong (Wang and Strong, 1996), the data quality dimension is a set of data quality attributes, which represents a single aspect or construct of data quality. These dimensions represent the measurement of data quality from different angles and classify the measurement of data quality into different categories. Amongst the data quality dimensions considered by researchers, the following four dimensions accuracy, completeness, consistency and currentness have been considered to be the dimensions of data quality involving data values (Fox, Levitin, Redman, 1994). In this paper, these four dimensions will be used for the proposed classification of dirty data.

2.3 Data Quality Rules

According to Adelman *et al*, data quality rules can be categorized into four groups namely business entity rules, business attribute rules, data dependency rules, and data validity rules (Adelman, Moss and Abai, 2005). Among the four categories, data validity rules (R1.1~R6.2) govern the quality of data values. Since the quality dimensions considered in this paper are all data value related, only rules in the data validity category will be considered for the proposed method. It is noticed that data uniqueness rules are associated with the data validity category. Rules R5.1 and R5.2 evaluate a special data quality problem which is caused by duplicate records. Because of the popularity, complexity and difficulty of this problem, it has attracted a large number of researchers (Elmagarmid, Ipeirotis and VeryKios, 2007). Therefore, apart from the four data quality dimensions, an extra data quality dimension "Uniqueness" is introduced for dealing with duplicate records exclusively in the proposed method.

According to David Loshin, it is the assertion embedded within the business polices that determines the quality of data (Loshin, 2006). Business policies can be transferred into a set of data quality rules, each of which can be categorized within the proposed data quality dimensions. In the mean time, these rules can be used to measure the occurrence of data flaws. In this paper, dirty data is defined as these data flaws that break any of the data quality rules. Since these rules are embedded within each of the data quality dimensions, a relationship between data quality dimensions and dirty data is built. The proposed method is formed based on this idea.

3 DIRTY DATA TYPES

A taxonomy of dirty data provides a better understanding of data quality problems. There are several taxonomies/classifications of dirty data existing in the literature (Rahm and Do, 2000, Müller and Freytag, 2003, Kim *et al*, 2003, Oliveira

et al, 2005). Within these works, Oliveira *et al* produced a very complete taxonomy which has identified 35 distinct dirty data types (DT.1~DT.35). Since Oliveira *et al*'s taxonomy is the most complete one existing in the literature, in next section, the proposed method will use the 35 data quality problems collected in their work for the mapping.

4 THE PROPOSED METHOD

Having discussed data quality, data quality dimensions and data quality rules in section 2 together with dirty data set generated based on Oliveira *et al*'s work, a new classification of the dirty data types is introduced beginning with a mapping of data quality rules with data quality dimensions. Table 1 shows the result of the mapping.

Table 1: Data quality dimensions and data quality rules.

Data quality dimensions	Rule No.
Accuracy	R2.1~ R2.5, R3.1, R4.1~R4.5
Completeness	R1.2, R1.4
Currentness	R3.2
Consistency	R5.5, R6.1, R6.2
Uniqueness	R5.1, R5.2

In order to classify dirty data types into data quality dimensions, after mapping data quality rules into data quality dimensions, a mapping from dirty data types to data quality rules is required. The result of this mapping is presented in table 2.

Table 2: Data quality rules and dirty data types.

Rule No.	Dirty data type No.
R1.1	N/A
R1.2	DT.21,
R1.3	N/A
R1.4	DT.1, DT.15
R2.1	DT.4
R2.2	DT.5
R2.3	DT.11, DT.14, DT.17, DT.20, DT.26, DT.35
R2.4	N/A
R2.5	DT.19, DT.34
R3.1	DT.16, DT.24, DT.25
R3.2	DT.3, DT.22
R4.1	DT.8
R4.2	DT.2
R4.3	DT.9
R4.4	DT.7
R4.5	DT.6
R5.1	DT.18, DT.33
R5.2	DT.12
R5.3	N/A
R5.4	N/A
R5.5	DT.10, DT.13,
R6.1	DT.23, DT.27, DT.31,
R6.2	DT.28, DT.29,DT.30, DT.32

The result of Table 2 provides immediate help for the proposed classification of dirty data. Combining the result from table 1 and 2, the classification of dirty data based on data quality dimensions is achieved in table 3.

Table 3: The classification of dirty data

Data quality dimension	Dirty data type
Accuracy	DT.2, DT.4~DT.9, DT.11, DT.14, DT.16, DT.17, DT.19, DT.20, DT.23~DT.26, DT.34, DT.35
Completeness	DT.1, DT.15, DT.21
Currentness	DT.3, DT.22
Consistency	DT.10, DT.13, DT.23, DT.27~DT.32
Uniqueness	DT.12, DT.18, DT.33

A method for dealing with dirty data based on the classification in table 3 is described as follows.

1) Create an order of the five dimensions according to the business priority policy;
2) Identify data quality problems;
3) Map the data types identified in 2) into the dimensions against the classification table;
4) Decide dimensions to be selected based on the budget;
5) Select appropriate algorithms, which can be used to detect dirty data types associated with dimensions identified in 3).
6) Execute the algorithms.

5 AN EXAMPLE

As an example, let's consider an online banking system used by a bank. Customers from the bank could obtain all related banking information via this system. Since the data in the system is comprehensive, it is very likely that dirty data may exist, such as misspelt data (DT.6), Wrong data value range (DT.5), duplicate records (DT.18, DT.33), data entered into a wrong field (DT.7), different formats/patterns for the same attribute (DT.23, DT.27), missing data within a record (DT.1), late updated data (DT.3, DT.22) etc. In this example, suppose cleaning all of the dirty data for this bank is unrealistic. The problem that the bank has to face is how to select a group of types of dirty data to deal with, based on their business priority policy, which is actually a DDS problem. According to the bank's priority policy, firstly the bank needs to make sure that data maintained in the system is accurate enough and up to date to provide correct information. Therefore, the currentness dimension and accuracy dimension are much more urgent than others. The proposed method provides a systematic approach to cope with the problem.

According to table 3, dirty data existing in the system has been found within all of the five data quality dimensions. It is easy to select which of these dirty data types cause accuracy and currentness related problems: DT.3, DT.5, DT.6, DT.7 and DT.22, which need to be dealt with first. Therefore, the data cleaning algorithms or methods designed for these dirty data types should be firstly applied.

6 CONCLUSION AND THE FUTURE WORK

In this paper, a novel method for dealing with dirty data based on the five data quality dimensions is presented. We have shown how the new method builds on and improves existing work on dirty data types and applies them to five data quality dimensions. The resulting method can be used by business to help to solve data quality problems, especially the Dirty Data Selection problem and prioritise the expensive process of data cleaning to maximally benefit their organisations.

Future work will involve the development of a taxonomy from a dimension angle, further more a data cleaning tool to deal with dirty data types based on the proposed method. However, the challenge remains that how to organize the sequence to deal with the dirty data types that are identified as well as selecting suitable methods/algorithms according to different problem domains.

REFERENCES

Adelman, S., Moss, L., Abai, M. (2005). *Data Strategy*. Addison-Wesley Professional.

Elmagarmid, A.K., Ipeirotis, P.G., VeryKios, V.S. (2007). Duplicate Record Detection: A Survey. . *IEEE Trans. on Knowl. and Data Eng.* 19, 1-16.

Fox, C., Levitin, A., Redman, T. (1994). The notion of data and its quality of dimensions. *Information Processing & Management*., vol. 30, no. 1. pp. 9-19

Kim, W., Choi, B., Hong, E.Y., Kim, S.K., Lee, D. (2003). A taxonomy of dirty data. *Data Mining and Knowledge Discovery*, 7,81-99.

Kim,W. (2002). On three major holes in Data Warehousing Today. *Journal of Object Technology*, Vol.1, No.4.

Loshin, D. (2006). *Monitoring Data Quality Performance Using Data Quality Metrics*. Retrived January 10, 2010, from http://www.it.ojp.gov/documents/Informatica_Whitepaper_Monitoring_DQ_Using_Metrics.pdf

Mong, L. (2000). IntelliClean: A knowledge-based intelligent data cleaner. *Proceedings of the ACM SIGKDD*, Boston, USA.

Müller, H., Freytag, J.C. (2003). Problems, Methods, and Challenges in Comprehensive Data Cleansing. Tech. Rep. HUB-1B-164

Oliveira, P., Rodrigues, F.T., Henriques, P., Galhardas, H. (2005). A Taxonomy of Data Quality Problems. *Second International Workshop on Data and Information Quality (in conjunction with CAISE'05)*, Porto, Portugal.

Rahm, E., Do, H. (2000). Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering*. vol.23, 41, No.2.

Wang, R.Y., Strong, D.M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12, 4.