International Journal of Digital Information and Wireless Communications (IJDIWC) 1(2): 536-544
The Society of Digital Information and Wireless Communications, 2011(ISSN 2225-658X)

Open User Involvement in Data Cleaning for Data Warehouse Quality

Louardi BRADJI[1] and Mahmoud BOUFAIDA[2]
LIRE Laboratory, Mentouri University of Constantine
25017 Ain El Bey Constantine Algeria
[1]bradjilouardi@yahoo.fr,[2]mboufaida@umc.deu.dz

ABSTRACT

High quality of data warehouse is a key to make smart strategic decisions. The data cleaning is program that performs to deal with the quality problems of data extracted from operational sources before their loading into data warehouse. As the data cleaning can introduce errors and some data require manually clean, there is a need for an open user involvement in data cleaning for data warehouse quality. This is essential to validate the cleaned data by users and to replace the dirty data in their original sources, and also to correct the poor data that can't be cleaned automatically. In this paper, we extend the data cleaning and extract-transform-load (ETL) processes to better support the user involvement in data quality management. We proposed that the ETL processes include two phases: the transformation to clean data at the operational data sources and the propagation of data cleaned towards their original sources. The major benefits of our proposal are twofold. First, it is the validation of cleaned data by users. Second, it allows the operational data sources quality improvement. Consequently the user involvement based data cleaning leads to a total data quality management and avoids redoing the same clean for future warehousing.

KEYWORDS

Data cleaning, data warehouse, ETL, user involvement, knowledge, data quality, operational data sources.

1 INTRODUCTION

Data Warehouses (DW) integrate data from different various operational data sources all over the organizations. Typically, this helps organizations to create coherent and aggregate data for decision-making purpose [1].
The software process that facilitates the initial loading and the refreshment of DW contents is commonly known as Extract, Transform and Load (ETL) process. ETL is responsible for extracting the data from the different sources, transforming them and then for loading them into the DW with the probability that some of these sources contain dirty data is enough high. So the improvement of data is vital to make perfect and accurate conclusions [2], [3].
Data Quality (DQ) plays a very important role in DW projects where there has been an exponential increase in the quantity of data. As a matter of fact, if the data that is stored and analyzed by end users is inaccurate or/and incomplete, it greatly affects the decisions they make based on this data [4]. The data quality is ensured via on-going audits and continuous DQ improvement efforts within the scope of a proactive DQ management [3], [5].
Data Cleaning (DC) is one of the critically important steps in DQ improvement programs in both the DW and Data Mining (DM) processes. It is

used to improve the quality of data by using various techniques. Many research works use the methods of DM in the DC programs. Missing and incomplete values of data represent especially difficult to build good knowledge and draw any conclusions [6], [7].

In practice, the DQ management in DW has to overcome several inherent problems [2], [8]. We have identified three major problems. First, Most of the data warehouse projects integrate the DQ phase into the ETL process without allocating enough time for efficient data cleaning results (corrected data) validation [4], [8]. Second, the operational data sources quality improvement by the feedback of data cleaning results is not taken into account although its importance to avoid redoing the same data cleaning tasks in futures data integrations. This is, to ensure the mutual coherence of data between various sources. Third, some erroneous data require manual or combined manual and automatic DC approaches [9]. The major cause behind these problems is that the existing DC approaches in DW do not support the involvement of users.

In this perspective, this paper provides an open user involvement data cleaning approach for a better data warehouse quality. We propose that: (1) the extension of DC process to better help users to validate the corrections and then to improve the DQ at the original sources and (ii) the adaptation of ETL process to support our DC.

The rest of this article is organized as follows. In section 2 we present concepts and definitions related to the subject. Section 3 discusses some related works. After giving the fundament and the basis of our approach in section 4, an overview of the proposal is given in section 5. The user involvement based DC process is detailed in section 6, followed by a presentation of the ETL extended to support the proposed approach in section 7. Section 8 presents the first experimental results to illustrate our approach relevance. Finally, we draw our conclusions and our research perspectives in section 9.

2 BACKGROUNDS

As this work aims at improving the data quality in DW environment by involving users in DC process, there is a need to give a brief idea about the concepts and definitions related to this work.

2.1 Data Cleaning

Data cleaning addresses the issues of detecting and removing errors and inconsistencies from data in order to improve the quality of the data [10]. The DC tools roughly distinguish between the single-source and multi-sources problems. These problems can be at the schema level and/or at instance level [1]. In general, the architecture of DC process involves five phases:

- Data Analysis in order to detect the inconsistencies;
- Definition and choice of transformations;
- Verification of the correctness and effectiveness of transformations;
- Execution of the transformation;
- Feedback of cleaned data.

Some DQ problems can't clean automatically. Therefore there is a need to clean them manually [11].

2.2 Extract-Transform-Load Process

Under the general acronym ETL, the Extract, Transform and Load activities cover the most prominent tasks of data preparation before the data warehousing and mining processes [12].
The first part of the ETL process embraces all the actions that are required to extract the data from operational data sources. This also includes pre-cleaning. The second step encompasses all the transformations that the data has to go through in order to fit the DW model. In this phase, data is cleaned, aggregated and transformed so that it could be loaded into a DW. The loading is performed in the last stage. Here also some additional cleaning is possible.
Data cleaning tasks are performed at the transform step in order to improve the accuracy of data before their loading into the DW [13].

3. SOME RELATED WORKS

The most popular DC frameworks include: AJAX, Potter's Wheel, Febrl, FraQL, ARKTOS, IntelliClean, Xclean and ODCF [6], [10], [14], [15], [16]. In all these frameworks, the DC operations will be specified by the end-users through a user interface. The main drawbacks presented by these frameworks in DW are:

- They don't allow the validation of results obtained by the DC process.
- They don't allow the feedback of cleaned data to the operational data sources in order to improve the dirty data at these sources.

Some researches works have addressed the problems of manually DC by involving end users at the intermediaries results [11]. The problem is that this solution can be only performed by the data transformations developers. Therefore, as the end users can only inspect the final results, there is a need to extend the DC process to support the involving of any user.
Although knowledge provides domain independent information that can be used for successful DC, the issue of knowledge use to support DC has not been dealt with [17]. IntelliClean is the first knowledge based data cleaning tool, which scrubs the data for anomalies [18]. However, this tool presents two major drawbacks. First, it can only identify and remove data duplications and therefore cannot deal with all the DQ problems. Second, it does not permit the knowledge validation, which is an important aspect that alters the quality of DC operations and results.
In order to deal with the drawbacks mentioned above, we propose an open user involvement in DC process for both data warehouse quality and the operational data sources quality improvements.

4. BASIC CONCEPTS OF THE PROPOSED APPROACH

After presenting the motivations for our proposal, we firstly detail our mutual coherence mechanism elaborated especially for the propagation of cleaned data. We also describe the knowledge base, which is an essential component of the considered architecture, which supports our approach.

4.1 Motivations

As we have indicated above, the existing DC solutions are a set of data transformations programs that perform automatically to detect and clean data. However, these solutions don't take into account some problems:

- Erroneous results can be produced by data cleaning solutions.
- Some poor data can be only manually cleaned.
- DC can be also made at operational data sources.

To deal with these problems, it is necessary to extend to DC process in order to allow the incorporation of the user involvement. The main benefits of this proposal are threefold. First, it allows the validation of the cleaned data. Second, the incorporation of user provides additional knowledge that are helpful for DC. Third, it allows improving the quality of data at the operational data sources.

4.2 Mutual Coherence

The coherence in the distributed systems and shared databases is guaranteed by the two phases commit protocol [19]. In the context of DW, it is closely related to the mechanisms of data replication and update propagation techniques [20]. This aspect is crucial for our study because the cleaning is an update of data and then there is need to propagate the modified data (cleaned in our case) to the original sources (i.e. operational data sources).

After performing DC, we have three theoretical values for cleaned data. The first value is the initial value (noted V_i) which is stored in a given operational data sources. The second value (noted V_c) is the final result given after the cleaning of poor data that have the initial value V_i. The third value (noted V_{user}) is providing by the user.

As the three values are related to the same data value at the same time, then the mutual coherence will be maintained i.e. the three values must be identical.

Therefore, we have created a mutual coherence strategy to deal with this situation.

The mutual coherence strategy consists of the following elements:

{
Trigger-Analyzing events;
Create dirty data;
Trigger- clean events;
Create Cleaned data;
Trigger-propagate events;
Trigger user Involvement events;
}

The first step of the strategy aims at triggering the data analysis program. After the building of the dirty data set, the cleaning program will be performed in order to correct the dirty data in the third step. The fourth step consists at building the cleaned data set which will be propagating to their original sources in the fifth step. The last step permits the involving of the end users in the data cleaning process.

This strategy is the basis or our proposal i.e. from which we have extended the both data cleaning and ETL processes.

4.3 Knowledge Base

For the purposes of this work, a Knowledge Base (KB) is defined to be a set of texts. These texts are stored by the users through a user interface in order to be used to facilitate the user involvement. The KB is a shared base that may be simultaneously accessed by multiple users and multiple DC programs. We identify two kinds of texts. The first kind is a text that explains the data transformations that are done during the DC task. This helps the user during the involvement to repair

data manually and validate some cleaned data. The second kind is a text that can be transformed into rule and then used by DC programs.

5 OVERVIEW OF THE DIFFERENT PROCESSES

The proposed approach aims at extending the data cleaning and ETL processes in order to facilitate the user involvement for manually data repair and validation of cleaned data. Consequently, the quality of operational data sources can be enhanced. In this section, we give an overview of our proposal.

Figure 1 depicts the first process of a DC. It is inspired from the general architecture of DC process presented in subsection 2.1. It extends by the propagation step which responsible of the feedback of cleaned data to the operational users in order to validate them.

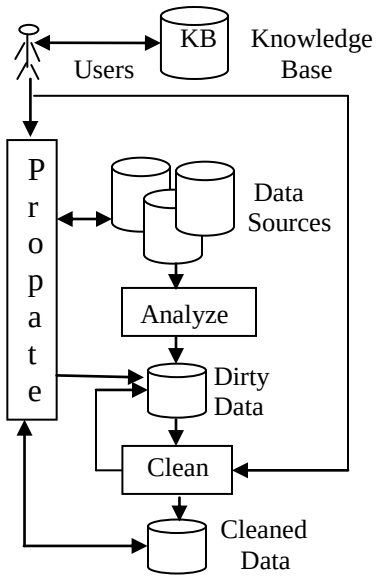
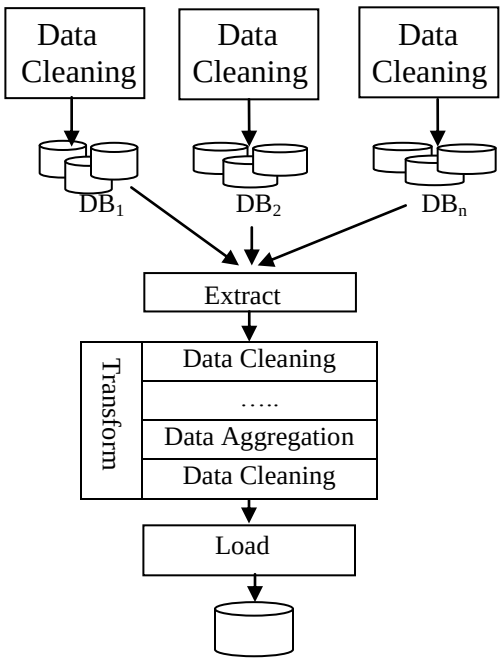


Figure 1. Open User Involvement Based Data Cleaning process

It is logical that the extension of DC process requires also the update of existing steps, especially the clean step in order to facilitate the manually DC.

Figure 2 depicts the second process of Extract-Transform-Load. It is similar to the classical ETL process. This process is extended by a step for cleaning data at the operational data sources independently of the rest of steps. It can be performed before the ETL in order to improve data at the operational data sources and then reduce the time consuming during the transformation step of ETL. As the operational data sources are independently, our system can perform the data cleaning programs at the operational data sources simultaneously. We have also modified the transform step in order to permit the cleaning of data just after their extract because at this point the user can be involved to inspect the data. In the contrary, after the aggregation of data, the involvement is possibly only for the developers and some experts.



DW

Figure 2. Data cleaning process based ETL process

6 USER INVOLVEMENT BASED DATA CLEANING PROCESS.

The proposed approach is open because it allows highly interactive DC process and explicit user involvement. As the analyze step is similar for each DC process where it aims at detecting data quality problems (dirty data), we detailed in this section the clean and propagate steps. Let us notice that the propagation is bidirectional (see figure 1) because it can be from (resp. to) cleaned data set to (resp. from) operational data sources. The output of this step is a set of dirty data. As shown in figure 1, this data set is incrementally built.

6.1 Clean

This step consists of correcting data that evaluate poor in the analyze step. We have added the task that separates the dirty data set into two subsets of data. The first subset is called manually cleaned data (Mcleaned) which are a set of poor data but they can't correct automatically. The user can access to this subset and repair them manually via a user interface at the clean step. The second subset is called cleaned data which contains the data corrected automatically. Let us notice that once a clean step is achieved, the cleaned data set is propagated to the operational data sources users in order to be validated.
As the objective of our proposal is the user involvement, the clean step includes a point where the user can be involved to do manually some cleaning data via a user interface. When the user cleans poor data then automatically it will be deleted from the dirty data set and added to the cleaned data set. At the end of this step, the dirty data set is equal to the Mclean subset i.e. it contains only the poor data that can't be corrected automatically and/or also aren't be yet corrected manually. Therefore both dirty and cleaned data sets data are permanent bases.

6.2 Propagate

This step is the core of our proposal because it involves the user to validate the cleaned data. The Mcleaned is not concerned by this step because the user cleans manually these data during the clean step. The propagate step is part the DC process of the transform step (after the extract step) of ETL process (i.e. it can't performed during the DC of operational data sources).
This step is of high importance because the DC can produce erroneous data and then can lead to catastrophes (for example death and/or epidemic in the medical field).
The propagate step ensures the mutual coherence of data and allows us to improve operational data sources quality. The updates propagation process is a crucial for total data quality assurance i.e. it ensures the quality of data at the operational databases and DW.
The propagation does not treat the mutual coherence as being only the propagation of the corrections but it permits to validate them by the users of operational data sources.
This step has two substeps of propagation: the cleaned data propagation and the validated data one.

6.2.1 Cleaned Data Propagation

This substep aims at transforming cleaned data to the operational data system in order to be validated by the user and to update the operational data sources. It contains three phases: a transmission, a validation and an update. It is used to propagate the cleaned values made during the DC process of ETL to the operational data sources.

The objective of this substep is the mutual coherence of data. As we have presented in the subsection 4.2, at this step we have three values for a given data value. Then the user will be involved to select one of these values.

In this process, the selected replication strategy is asynchronous. Also, the update propagation method is eventual because the cleaned data propagation starts after the integration of cleaned data.

Transmission. As the operational data sources are distributed and heterogeneous, the transmission consists for each source at:

- The construction of the cleaned data sets from the global cleaned data set according to their operational data sources schema;
- The transfer of each cleaned data set to the corresponding operational system.

Validation. The user involves to choice the optimal value. There are three cases:
Case 1: If V_c is retained then V_i: = V_c;
Case 2: If V_i is retained then V_c: =V_i;
Case 3: If V_{user} is retained then V_i: =V_{user} and V_c: =V_{user}.

Update. If the values of cleaned data retained are V_c or V_{user}, then there is a need to update the operational data sources with these values. This update allows improving the operational data sources and the cleaned data if DC has produced erroneous corrections.

At the end of this phase, the system constructs set of validated data (E_v). E_v contains the cleaned values V_i and V_{user} retained by the user in the previous phase. It is allows the correction of the cleaned data evaluated poor by the users.

6.2.2 Validated Data Propagation

If the transform step doesn't achieve then the propagation is performed directly and then the cleaned data is updated by taking into account E_v. But, once the data warehousing is achieved, the propagation can't be made constantly because it is time consuming. From this fact, we propose three different modes of propagation: periodical, rebuilding and on request.

Periodic. The validated data set propagation is done during the refreshing of DW.

Rebuilding. If the volume of E_v is huge, then the DW is judged of poor quality. So the rebuilding of DW is necessary. The evaluation of the hugest of E_v is doing by computing the ratio between the volumes of E_v set and the data extracted set. If this ratio is significant then the DW will be rebuilt.

On request. The propagation is on request of the DW administrator.

7 EXTEND ETL PROCESS

The idea behind the extension of ETL process by an independently step for DC at the operational data sources is twofold. First, as our objective is to validate the cleaned data at the operational systems and to repair data manually at the clean step by involving

users, it is necessary to create a system that allows the user involvement. Therefore this requires cleaning data at operational data sources. As we have indicated, the DC at these data sources is performed without propagation. Second, it reduces the time consuming during the transform step.

The second extension is to perform the DC during the transform step at its start and its end. At the start, DC deals with the data before their aggregations and then the user involvement is realized without difficult. In the contrary, after the aggregation of data the user involvement is difficult because aggregated data can't be evaluated by all the users. Therefore, our proposal deals with the user involvement before the aggregation.

8 EXPERIMENTAL RESULTS

In order to illustrate the applicability and performances of our proposal, this section presents the first results of a case study by applying the proposed approach to health sector. We have selected three departments of this sector (laboratory, pharmacy and medical services) where each department has an operational data source. The manual inspection of the data sets in these departments reveals a wide number of data quality problems (single-source and multi-sources): missing values, empty fields, misspellings, improper generation of data, domain value violations… duplicate entities.

In the first time, we have performed the DC at the operational systems. 47 % of poor data are cleaned. Among them 11% of cleaned are made manually. Therefore we conclude the utility of the user involvement.

During the DC during the transform step, 7 % of dirty data are detected but can't be corrected automatically and then the involving of users permits us to correct 5.2 % of these data. This shows the importance and the benefit of the user's involvement.

After the propagation of cleaned data to the operational systems, 4.7% of cleaned data are evaluated incorrect and then the user keeps the 3.1 % of initial values of these cleaned data and proposed new values for the rest of incorrect data (1.6 %). Some data are cleaned but after having involved the users that they are evaluated them poor.

Besides, the time consuming for the ETL process is reduced because some data are cleaned during the operational data sources cleaning.

Our proposal has given better results for the improvement of DQ but the problem is to validate it with huge DW.

9 CONCLUSIONS

In this paper, we explored the incorporation of the user involvement in DC process for improving the quality of DW. We have showed how the interactivity of the DC is crucial because DC programs can produce poor data and some errors can't be detected and/or corrected automatically. This approach allows one to improve the quality of operational data sources. We have also proposed the propagation of cleaned data that may associated to the cleaned data results to involving the user.

The first experimental results have demonstrated the advantages of our approach. In particular, we show the benefit of DC achieved when the user is involved.

As part of our future work is the application of our proposal to a huge DW.

10 REFERENCES

1. Rahm, E., Do, H.H..: Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bull. Vol 23 No. 4, pp. 3-13 (2000).
2. Vassiliads, P.: A Survey of Extract-Transform-Load Technology. In International Journal of Data Warehousing & Mining, vol. 5, no. 3, pp. 1-27(2009).
3. Nemani, R.R., Konda, R.: A Framework for Data Quality in Data Warehousing. UNISCON 2009, LNBIP 20, pp. 292-297, 2009, , Springer Heidelberg (2011).
4. Adarsh A., Rajendra K.R.: Implementing a Data Quality Module in an ETL Process. A Project Report, Rochester Institute of Technology, Rochester, New York, April (2011).
5. Helfert, M., H., Hermann C.: Proactive Data Quality Management for Data Warehouses Systems. Journal of Data Mining and Data Warehouse (DMDW), Vol. 2002, pp. 97-106 (2002).
6. Matyia, D.: Applications of data mining algorithms to analysis of medical data. Master Thesis, Software Engineering, Thesis no: MSE-2007. Blekinge Institute of Technology. (2007).
7. Lin, J.H., Haug, P.J.: Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. In Journal of Biomedical Informatics, vol. 41, no. 1, pp. 1–14 (2008).
8. Bradji, L., Boufaida, M.: Knowledge based data cleaning for data warehouse quality. In: Proc. 2011 International Conference on Digital Information Processing and Communications, ICDIPC2011, LNCS, Part II, CCIS no 189, pp. 373-384, Springer-Heidelberg (2011).
9. Mahnic, V., Rozanc, I. : Data Quality: A Prerequisite for Successful Data Warehouse Implementation. Journal of Informatica, vol. 25, no. 2, Slovenia (2001),
10. Herbert, K.G., Wang, J.T.L.: Biological data cleaning: a case study. In Int. J. of Information Quality, vol. 1, number. 1, pp. 60–82 (2007).
11. Helena, G., Antonia L., Emanuel S.: Support for User Involvement in Data Cleaning. In: Proc. 2011 13th International Conference Data Warehousing and Knowledge Discovery, DaWaK 2011, Toulouse, France, August 29-September 2,2011. LNCS, vol. 6862, pp. 136-151. Springer (2011).
12. Berti-Equille, L.: Measuring and Modelling Data Quality for Quality-Awareness in Data Mining, Studies in Computational Intelligence (SCI), Vol. 43, pp. 101–126 Springer, Heidelberg. (2007).
13. Vassiliadis, P., Quix, C., Vassiliou, Y., Jarke, M.: Data warehouse process management. In Journal of Information Systems, vol. 26, no..3, pp.205-236 (2001).
14. Berti-Equille, L.., Dasu, T.: Data Quality Mining: New Research Directions. In: Proc. 2009 International Conference on Data Mining,(ICDM'09), (2009).
15. Huanzhuo, Y., Di W., Shuai C.: An Open Data Cleaning Framework Based on Semantic Rules for Continuous Auditing. In: Proc. 2010 2nd International Conference on Computer Engineering and Technology, vol. 2 pp. 158-162, IEEE (2010).
16. Lee, M.L., Ling, T.W., Low, W.L.: IntelliClean : A Knowledge Based Intelligent Data Cleaner. In: Proc. 2000 6th ACM/SIGKDD conference on Knowledge Discovery and Data Mining, 2000, pp.290—294 (2000).
17. Oliveira, P., Rodrigues, F., Henriques, P. : An Ontology-based approach for data cleaning, In: Proc. 2008 11th International Conference on Information Quality (ICIQ'07), pp. 307-320 (2007).
18. Kororoas, A., Lin, S..: Information Quality in Engineering Asset Management. Information Quality Management: Theory and Applications. Ismael C. and Mario P.(ed), pp. 221-251,Idea Group Publishing (2007).
19. Pacitti, E., Simon, E.: Update propagation strategies to improve freshness in lazy master replicated databases. In: the VLDB Journal, vol. 8, pp. 305-318, Springer-Verlag (2000).
20. Pacitti, E.: Improving Data Freshness in Replicated Databases. INRIA. Research report no. 3617 (1999).