A Comparison Study of Data Scrubbing Algorithms and Frameworks in Data Warehousing

Volume 68 - Number 25
Year of Publication: 2013

Authors:
Hamed Ibrahim Housien
Zhang Zuping
Zainab Qays Abdulhadi

10.5120/11752-7406

Abstract

In these days, many organizations tend to use a Data Warehouse to meet the requirements to develop decision-making processes and achieve their goals better and satisfy their customers. It enables Executives to access the information they need in a timely manner for making the right decision for any work. Decision Support System (DSS) is one of the means that applied in data mining. Its robust and better decision depends on an important and conclusive factor called Data Quality (DQ), to obtain a high data quality using Data Scrubbing (DS) which is one of data Extraction Transformation and Loading (ETL) tools. Data Scrubbing is very important and necessary in the Data Warehouse (DW). There are growing relationships to get high DQ and effective DS. The use of DS algorithms is a solution to the constraints that limit the DQ which leads to weak decisions and the burden of the high financial costs. These constraints are: dirty data, noise data, missing values, inconsistency, uncertain data, ambiguous, conflicting, duplicated records and similar columns. The Sources and causes of these constraints are many, including: input error, merge data from different sources, difference in representing the same information, etc. In addition there are more than 35 sources and causes of the poor-quality data that arise at the stage of the ETL process. This paper present
A Comparison Study of Data Scrubbing Algorithms and Frameworks in Data Warehousing

comparison and analysis for DS algorithms and the pros and cons of each algorithm, accuracy and time complexity. Additionally, it present a comparative and analysis of the Data Scrubbing Frameworks and determine the best framework.

References

- Luyi Mo, Reynold Cheng, Xiang Li, David Cheung and Xuan Yang, "Cleaning Uncertain Data for Top-k Queries", Department of Computer Science University of Hong Kong, Hong Kong, 2012.
- Mortadha M. Hamed and Alaa Abdulkhar Jihad, "An Enhanced Technique to
Clean Data in the Data Warehouse; Development in E-systems Engineering, IEEE, 2011.

- Shaofeng Liu, Alex H. B Duffy, Robert Ian Whitifield, Iain M. Boyle, "Integration of decision support systems to improve decision support performance; Springer-Verlag London Limited, February 2009.
- "Figure 1: Basic ETL Functionality; Available at: http://gerardnico.com/wiki/dit/etl_become_di, Accessed on Jan 2013.

**Index Terms**

Computer Science  
Algorithms

**Keywords**

Data scrubbing  Data warehousing  Data Quality  Extract-Transform-Load (ETL)