Abstract

Data Cleansing or (data scrubbing) is an activity involving a process of detecting and correcting the errors and inconsistencies in data warehouse. Thus poor quality data i.e.; dirty data present in a data mart can be avoided using various data cleaning strategies, and thus leading to more accurate and hence reliable decision making. The quality data can only be produced by cleaning the data and pre-processing it prior to loading it in the data warehouse. As not all the algorithms address the problems related to every type of dirty data, one has to prioritize the need of its organization and use the algorithm according to their requirements and occurrence of dirty data. This paper focuses on the two data cleaning algorithms: Alliance Rules and HADCLEAN and their approaches towards the data quality. It also includes a comparison of the various factors and aspects common to both.

References

- Arindam Paul, Varuni Ganesan, “HADCLEAN: A Hybrid Approach to Data Cleaning in
A Comparative Analysis of Data Cleaning Approaches to Dirty Data

- Dr. Mortadha M. Hamad, Alaa Abdulkar Jihad, “An Enhanced Technique to Clean Data in the Data Warehouse”, 2011, IEEE.
- Wai Lup Low, Mong Li Lee, “A Knowledge based Approach for Duplicate Elimination in Data Cleaning”, School of Computing, National University Singapore.

Index Terms

Computer Science  Information Sciences

Keywords

HADCLEAN  PNRS  phonetic algorithm  alliance rules  transitive closure  near miss strategy  scores