Abstract

The paper analyses the problem of data cleaning and automatically identifying the "incorrect and inconsistent data" in the dataset. Extraction, Transformation and Loading (ETL) are the different steps for cleaning a data warehouse. Authors have implemented different algorithms like: cleanString, cleanNumber, hit ratio, check data dictionary, check metadata etc in addition to various existing data cleaning algorithm like PNRS. This paper tries is to improve the quality of data in the database system. This paper emphasizes on the citizen database system to make it errorless. Some of the results along with certain statistics are also provided here.
References

- Erhard Rahm and Hong Hai Do, “Data Cleaning Problems and Current Approaches,” University of Leipzig, Germany.
- Srivatsa Maddodi, Girija V. Attigeri and Dr. Karunakar A. K, “Data Deduplication Techniques and Analysis,” Manipal Institute of Technology, Manipal, India.
- Cleaning the Spurious Links in Data - Mong Li Lee, Wynne Hsu, and Vijay Kothari National University of Singapore.
- An Important Issue in Data Mining - Data Cleaning - Qi Xiao Yang Institute of High Performance Computing Sung Sam Yuan, LuChun School of Computing National University of Singapore, Jay Rajasekera Graduate School of International Management International University of Japan.
- Generic and Declarative Approaches to Data Cleaning: Some Recent Developments – Leopoldo Bertossi and Loreto Bravo.
- Conditional Functional Dependencies for Data Cleaning – Philip Bohannon from Yahoo! Research, Wenfei Fan from Bell Laboratories, Floris Geerts from University of Edinburgh, Xibe Jia from University of Edinburgh, Anastasios Kementsietsidis from Hasselt University/Transnational university Limburg.
- A Study over Problems and Approaches of Data Cleansing/Cleaning by Nidhi Chowdhury, dept. of CS, UPTU, India.
- NADEEF: A Commodity Data Cleansing System Michele Dallachiesa, Amr Ebaid, Ahmed Eldawy, Ahmed Elmagarmid, Hab F. Llyas, Mourad Ouzzani, Nan Tang, OCRI, University of Trento, Purdue University, University of Minnesota.

Index Terms

Computer Science

Databases
Keywords
Data warehouse  ETL  Data Dictionary  Hit Ratio  Dirty Data  Data Cleaning.