Hybrid Technique for Data Cleaning

Role of Engineers in National Building
© 2014 by IJCA Journal

NCRENB

Year of Publication: 2014

Authors:
Ashwini M. Save
Seema Kolkur

{bibtex}ncrenb1402.bib{/bibtex}

Abstract

Data warehouse contains large volume of data. Data quality is an important issue in data warehousing projects. Many business decision processes are based on the data entered in the data warehouse. Hence for accurate data, improving the data quality is necessary. Data may include text errors, quantitative errors or even duplication of the data. There are several ways to remove such errors and inconsistencies from the data. Data cleaning is a process of detecting and correcting inaccurate data. Different types of algorithms such as Improved PNRS algorithm, Quantitative algorithm and Transitive algorithm are used for the data cleaning process. In this paper an attempt has been made to clean the data in the data warehouse by combining different approaches of data cleaning. Text data will be cleaned by Improved PNRS
Hybrid Technique for Data Cleaning

algorithm, Quantitative data will be cleaned by special rules i.e. Enhanced technique. And
lastly duplication of the data will be removed by Transitive closure algorithm. By applying these
algorithms one after other on data sets, the accuracy level of the dataset will get increased.

References

- Arindam Paul, V. Ganesan, and J. Challa, "HADCLEAN: A Hybrid Approach to
Data Cleaning in Data Warehouses" IEEE, 2012.
- Mortadha M. Hamad and AlaaAbdulkhar Jihad, "An Enhanced Technique to
Clean Data in the Data Warehouse" IEEE, 2011.
- K. Ali and M. Warraich, "A framework to implement data cleaning in enterprise
- C. Varol, C. Bayrak, R. Wagner and D. Goff, "Application of the Near Miss
Strategy and Edit Distance to Handle Dirty Data" Data Engineering - International Series
- M. A. Hernández and S J. Stolfo, "Real-world Data is Dirty: Data Cleansing and
The Merge/Purge Problem" IEEE, Data Mining and Knowledge Discovery, Springer
- R. Bheemavaram, J. Zhang and W. N. Li, "Efficient Algorithms for Grouping
Data to Improve Data Quality" Proceedings of the 2006 International Conference on
Information & Knowledge Engineering (IKE 2006), CSREA Press, Las Vegas, Nevada, USA,
- R. Bheemavaram, J. Zhang, W. N. Li, "A Parallel and Distributed Approach for
Finding Transitive Closures of Data Records: A Proposal" Proceedings of the Acxiom
- W. N. Li, R. Bheemavaram, X. Zhang, "Transitive Closure of Data Records:
Application and Computation" Data Engineering - International Series in Operations
- Ballou, D. (1999) "Enhancing data quality in Data Warehousing
- M. Bilenko and R. J. Mooney, "Adaptive duplicate detection using learnable
string similarity measures" ACM SIGKDD, 2003, pp 39-48
- S. Reddy, A. Lavanya, V. Khanna, L. S. S. Reddy, "Research Issues on Data
Warehouse Maintenance" IEEE, ICACC &apos;09. InternationalConference Advanced
Computer Control, Singapore, Jan 2009, Page(s): 623 – 627

Index Terms

Computer Science
Information Science
Keywords
Data Cleaning  Pnrs  Improved Pnrs  Enhance Technique  Transitive.