Abstract

Data cleansing is a process that deals with identification of corrupt and duplicate data inherent in the data sets of a data warehouse to enhance the quality of data. This paper aims to facilitate the data cleaning process by addressing the problem of duplicate records detection pertaining to the ‘name’ attributes of the data sets. It provides a sequence of algorithms through a novel
framework for identifying duplicity in the ‘name’ attribute of the data sets of an already existing
data warehouse. The key features of the research includes its proposal of a novel framework
through a well defined sequence of algorithms and refining the application of alliance rules [1] by incorporating the use of previously existing and well defined similarity computation
measures. The results depicted show the feasibility and validity of the suggested method.

Reference

- Rajiv Arora, Payal Pahwa, Shubha Bansal, “Alliance Rules for Data Warehouse
  Cleansing”, International Conference on Signal Processing Systems IEEE Explore no. D01
  10.1109/ICSPS, 133, pages 743-747, 2009.
- P.Ponniah, “Data Warehousing Fundamentals- A comprehensive guide for IT
  India, 2007.
- A.Marcus, J.I.Maletic,”Utilizing Association Rules For the Identification of Errors in Data”,
  TR-CS-00-04, University of Memphis, 2004.
- A.Marcus, J.I.Maletic, ” Data Cleansing: Beyond Integrity Analysis” Proceedings of the
- T. Redman, "The Impact of Poor Data Quality on the Typical Enterprise",
- A.Marcus, J.I.Maletic, “Automated Identification of Errors in Data Sets”, TR-CS-00-02,
  University of Memphis, 2002.
- A.Marcus, J.I.Maletic and Lin, K.-I., “Association Rules for Error Identification in Data
  Sets”, Proceedings of the 10th ACM Conference on Information and Knowledge Management
- Peter Christen,” A Comparison of Personal Name Matching: Techniques and Practical
- Gérard Bouchard and Christian Pouyez, Name Variations and Computerised Record
- Timothy E. Ohanekwu, C.I. Ezeife," A Token-Based Data Cleaning Technique for Data
- Surajit Chaudhary, Kris Ganjam, Venkatesh Ganti, Rajeev Motwani, " Robust and
  efficient fuzzy match for online data cleaning",ACM SIGMOD,2003
- Amit Rudra, Emilie Yeo, “Key Issues in Achieving Data Quality and Consistency in Data
  Warehousing among Large Organisations in Australia,” Proceedings of the 32nd Hawaii
- Heiko Müller, Johann-Christoph Freytag, Berlin,” Problems, Methods, and Challenges in
  Comprehensive Data Cleansing”, 10099 Berlin, Germany.
- A. D.Chapman, “Principles and Methods of Data Cleaning – Primary Species and
  Species-Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility,
  Copenhagen, 2005.
- Rohit Ananthakrishna (Cornell University) Surajit Chaudhuri Venkatesh Ganti (Microsoft
A Novel Framework and Model for Data Warehouse Cleansing

Research)," Eliminating Fuzzy Duplicates in Data Warehouses”.
- Oktie Hassanzadeh, Mohammad Sadoghi, Ren´ee J. Miller, “Accuracy of Approximate String Joins Using Grams”, University of Toronto 10 King’s College Rd.,Toronto, ON M5S3G4, Canada.
- Jakub Piskorski_, Marcin Sydow, “Usability of String Distance Metrics for Name Matching Tasks in Polish”.

Index Terms

Computer Science

Information Science

Key words

Data warehouse data cleansing fuzzy logic data mining