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

- Surajit Chaudhary, Kris Ganjam, Venkatesh Ganti, Rajeev Motwani, " Robust and efficient fuzzy match for online data cleaning",ACM SIGMOD,2003
- Heiko Müller, Johann-Christoph Freytag, Berlin,” Problems, Methods, and Challenges in Comprehensive Data Cleansing”, 10099 Berlin, Germany.
- Rohit Ananthakrishna (Cornell University) Surajit Chaudhuri Venkatesh Ganti (Microsoft
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