Handling Duplicate Data in Data Warehouse for Data Mining

Abstract

The problem of detecting and eliminating duplicated data is one of the major problems in the broad area of data cleaning and data quality in data warehouse. Many times, the same logical real world entity may have multiple representations in the data warehouse. Duplicate elimination is hard because it is caused by several types of errors like typographical errors, and
different representations of the same logical value. Also, it is important to detect and clean equivalence errors because an equivalence error may result in several duplicate tuples. Recent research efforts have focused on the issue of duplicate elimination in data warehouses. This entails trying to match inexact duplicate records, which are records that refer to the same real-world entity while not being syntactically equivalent. This paper mainly focuses on efficient detection and elimination of duplicate data. The main objective of this research work is to detect exact and inexact duplicates by using duplicate detection and elimination rules. This approach is used to improve the efficiency of the data.

Reference
- Bilenko, M., Mooney, R.J (August 2003), Adaptive Duplicate Detection Using Learnable String Similarity Measures, Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’03), Washington, DC.
- Jiawei Han, Micheline Kamber (March 2006), Data Mining: Concepts and Techniques, Publisher: Elsevier Science & Technology Books, ISBN-13: 9781558609013.
- Judice L.Y.Koh, Mong Li Lee, Asif M. Khan, Paul T.J. Tan and Vladimir (September 24,2004), Duplicate Detection in Biological Data using Association Rule Mining, 2nd European Workshop on Data Mining and Text Mining for Bioinformatics, Pisa, Italy.
- Robert Leland (August 2007), Duplicate Detection with PMC – A Parallel Approach to Pattern Matching Department of Computer and Information Science, Norwegian University of Science and Technology, Ph.D. Thesis.
Index Terms

Computer Science
Data Engineering

Key words

Data Cleaning
Duplicate Data
Data Warehouse
Data Mining