Chapter 17
An Efficient Algorithm for Data Cleaning

Payal Pahwa
Guru Gobind Singh IndraPrastha University, India

Rajiv Arora
Guru Gobind Singh IndraPrastha University, India

Garima Thakur
Guru Gobind Singh IndraPrastha University, India

ABSTRACT

The quality of real world data that is being fed into a data warehouse is a major concern of today. As the data comes from a variety of sources before loading the data in the data warehouse, it must be checked for errors and anomalies. There may be exact duplicate records or approximate duplicate records in the source data. The presence of incorrect or inconsistent data can significantly distort the results of analyses, often negating the potential benefits of information-driven approaches. This paper addresses issues related to detection and correction of such duplicate records. Also, it analyzes data quality and various factors that degrade it. A brief analysis of existing work is discussed, pointing out its major limitations. Thus, a new framework is proposed that is an improvement over the existing technique.

INTRODUCTION

A process of transforming data into information and making it available to users in a timely manner is called Data warehousing.

A *data warehouse* is a central repository of an organization's electronically stored data (http://searchsqlserver.techtarget.com).

DOI: 10.4018/978-1-4666-1873-2.ch017

A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process (http://www.businessdictionary.com). Data from various sub-systems of an organization is accumulated and stored under a unified schema to facilitate reporting and analysis. Centralization of data is needed to maximize user access and analysis. However, the means to retrieve and analyze data, to extract, transform and load data (ETL process), and to manage the data diction-

Copyright © 2013, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

ary are also considered integral components of a data warehousing system (Jarke, Lenzerini, Vassiliou, & Vassiliadis, 2000). The prime aim of data warehouses is to maintain data in a manner that enables its usage for other tasks such as Data mining (Elmasri & Navathe, 2000).

Data mining (often termed as knowledge discovery) is the efficient discovery of valuable, non-obvious information from a large collection of data. Data mining centers around the automated discovery of new facts and relationships in data (Lee, Lu, Ling, & Ko, 1999). With traditional query tools, you search for known information. Data mining tools enable you to uncover hidden information. The assumption is that more useful knowledge lies hidden beneath the surface (Ponniah, 2001). The process of knowledge discovery is meaningful only when it presents data in a useful form i.e. without any errors.

Data cleaning or scrubbing is the process of removing the errors from the data. It is an inherent activity related to database processing, updating and maintenance. Data fed from various operational systems prevailing in the different departments/sub-departments of the organization, has discrepancies in schemas, formats, semantics etc. due to numerous factors (Hang-Hai & Erhard, 2000; Marcus & Maletic, 2000). While integrating data from these heterogeneous sources multiple instances referring to the same real-world entity are generated which need to be pre-processed before loading in the data warehouse (Tamilselvi & Saravanan, 2008; 2009). One of the most difficult tasks is to distinguish between multiple occurrences of the same real-world data sets scattered over different sources (Shahri, Shahri, Hellerstein, & Raman, 2001). In addition, conflict arises when these heterogeneous sources have to be accumulated into a large data warehouse. By this we mean that these representations may contain unnecessary attributes that do not match the target warehouse. They may introduce redundancy leading to exact duplicates of records, interdependence where two or more records contain attributes correlated to each other such that presence of one demands the presence of the other as well, inconsistency where records differ in schemas, formats, abbreviations, etc. and lastly approximate duplicate those records which are replica of each other such that neither they are textually identical nor they point to the same real-world entity (Orr, 1998; Tamilselvi & Saravanan, 2008, 2009). All such unwanted data records are referred to as 'dirty data' (http://en.wikipedia.org). Our approach focuses on the identification of approximate duplicate records before loading them in the data warehouse. Hence, we present a brief overview of various sources of errors that arise due to machine or human intervention (Hernandez & Stolfo, 1995, 1998).

Sources of Erroneous Data

1. Lexical errors name discrepancies between the structure of the data items and the specified format.
2. Syntactical errors represent violations of the overall format.
3. Irregularities are concerned with the non-uniform use of values and abbreviations._
4. Integrity constraint violations Integrity constraints are used to describe our understanding of the mini-world by restricting the set of valid instances (Redman, 1996, 1998). Each constraint is a rule representing knowledge about the domain and the values
5. Duplicates are two or more tuples representing the same entity from the real world. The values of these tuples need not be entirely identical. Inexact duplicates represent the same entity but with different values for all or some of its attributes.
6. Missing values are the result of omissions while collecting the data. This is to some degree a constraint violation if we have null values for attributes where there exists a NOT NULL constraint for them (Winkler, 1999).

14 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/efficient-algorithm-data-cleaning/67730

Related Content

Training Mode of Visual Communication Design Professionals Under the Background of Informationization
Tianjing Yan (2024). *International Journal of Knowledge Management (pp. 1-13).*
www.irma-international.org/article/training-mode-of-visual-communication-design-professionals-under-the-background-of-informationization/355012

Big Data and Knowledge Management in Ancient Chinese Poetry
Zhengyu Yang (2026). *International Journal of Knowledge Management (pp. 1-18).*
www.irma-international.org/article/big-data-and-knowledge-management-in-ancient-chinese-poetry/398858

Impacts from Using Knowledge: A Longitudinal Study from a Nuclear Power Plant
Murray E. Jennex (2010). *Ubiquitous Developments in Knowledge Management: Integrations and Trends (pp. 161-175).*
www.irma-international.org/chapter/impacts-using-knowledge/41862

Using Network Analysis and Visualization to Analyze Problematic Enterprise Information Systems
David Greenwoodand Ian Sommerville (2013). *Multidisciplinary Studies in Knowledge and Systems Science (pp. 291-310).*
www.irma-international.org/chapter/using-network-analysis-visualization-analyze/76236

Exploring the Determinants Affecting Academics' Knowledge-Sharing Behavior in United Arab Emirates Public Universities
Huda Skaikand Roslina Othman (2018). *Contemporary Knowledge and Systems Science (pp. 151-191).*
www.irma-international.org/chapter/exploring-the-determinants-affecting-academics-knowledge-sharing-behavior-in-united-arab-emirates-public-universities/199613