Chapter IV
Discovering Quality Knowledge from Relational Databases

M. Mehdi Owrang O.
American University, USA

ABSTRACT

Current database technology involves processing a large volume of data in order to discover new knowledge. However, knowledge discovery on just the most detailed and recent data does not reveal the long-term trends. Relational databases create new types of problems for knowledge discovery since they are normalized to avoid redundancies and update anomalies, which make them unsuitable for knowledge discovery. A key issue in any discovery system is to ensure the consistency, accuracy, and completeness of the discovered knowledge. We describe the aforementioned problems associated with the quality of the discovered knowledge and provide some solutions to avoid them.

INTRODUCTION

Modern database technology involves processing a large volume of data in databases to discover new knowledge. Knowledge discovery is defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data (Adriaans & Zantinge, 1996; Agrawal, Imielinski, & Swami, 1993; Berry & Linoff, 2000; Brachman & Anand, 1996; Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro, & Simoudis, 1996; Bradley, Gehrke, Ramakrishnan, & Srikant, 2002; Fayad, 1996; Fayad, Piatetsky-Shapiro, & Symth, 1996a, 1996b, 1996c; Fayyad & Uthurusamy, 2002; Frawley, Piatetsky-Shapiro, & Matheus, 1992; Han & Kamber, 2000; Hand, Mannila, & Smyth, 2001; Inmon, 1996; Simoudis, 1996; Uthurusamy, 1996; Keyes, 1990).

Databases contain a variety of patterns, but few of them are of much interest. A pattern is interesting to the degree that it is not only accurate but that it is also useful with respect to the end user's knowledge and objectives (Brachman et al., 1996; Bradley et al., 2002; Hand et al., 2001; Berry & Linoff, 2000; Piatetsky-Shapiro & Matheus, 1994; Silberschatz & Tuzhilin, 1995). A critical issue in knowledge discovery is how well the database is created and maintained. Real-world databases present difficulties as they tend to be dynamic, incomplete, redundant, inaccurate, and very large. Naturally, the efficiency of the discovery process

Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

and the quality of the discovered knowledge are strongly dependent on the quality of data.

To discover useful knowledge from the databases, we need to provide clean data to the discovery process. Most large databases have redundant and inconsistent data, missing data fields, and values, as well as data fields that are not logically related and are stored in the same data relations (Adriaans & Zantinge, 1996; Parsaye & Chingell, 1999; Piatetesky-Shapiro, 1991; Savasere et al. 1995). Subsequently, the databases have to be cleaned before the actual discovery process takes place in order to avoid discovering incomplete, inaccurate, redundant, inconsistent, and uninteresting knowledge. Different tools and techniques have been developed to improve the quality of the databases in recent years, leading to a better discovery environment. There are still problems associated with the discovery techniques/schemes which cause the discovered knowledge to be incorrect, inconsistent, incomplete, and uninteresting.

Most of the knowledge discovery has been done on operational relational databases (Sarawagi et al., 1998). Operational relational databases, built for online transaction processing, are generally regarded as unsuitable for rule discovery since they are designed for maximizing transaction capacity and typically have a lot of tables in order not to lock out users. In addition, the goal of the relational databases is to provide a platform for querying data about uniquely identified objects. However, such uniqueness constraints are not desirable in a knowledge discovery environment. In fact, they are harmful since, from a data mining point of view, we are interested in the frequency with which objects occur (Adriaans & Zantinge, 1996; Berry & Linoff, 2000; Bradley & Gehrke, 2002; Hand et al., 2001). Subsequently, knowledge discovery in an operational environment could lead to inaccurate and incomplete discovered knowledge. The operational data contains the most recent data about the organization and is organized as normalized relations for fast retrieval as well as avoiding update anomalies. Summary and historical data, which are essential for accurate and complete knowledge discovery, are generally absent in the operational databases. Rule discovery based on just the detailed (most recent) data is neither accurate nor complete.

A data warehouse is a better environment for rule discovery since it checks for the quality of data more rigorously than the operational database. It also includes the integrated, summarized, historical, and metadata which complement the detailed data (Bischoff & Alexander, 1997; Bradley & Gehrke, 2002; Hand et al., 2001; Inmon, 1996; Berry & Linoff, 2000; Meredith & Khader, 1996; Parsaye, 1996). Summary tables can provide efficient access to large quantities of data as well as help reduce the size of the database. Summarized data contains patterns that can be discovered. Such discovered patterns can complement the discovery on operational/detail data by verifying the patterns discovered from the detailed data for consistency, accuracy, and completeness. In addition, processing only very recent data (detailed or summarized) can never detect trends and long-term patterns in the data. Historical data (i.e., sales product 1982-1991) is essential in understanding the true nature of the patterns representing the data. The discovered knowledge should be correct over data gathered for a number of years, not just the recent year.
The goals of this chapter are twofold:

1. To show that anomalies (i.e., incorrect, inconsistent, and incomplete rules) do exist in the discovered rules due to:
   a. An inadequate database design
   b. Poor data
   c. The vulnerability/limitations of the tools used for discovery
   d. Flaws in the discovery process (i.e., the process used to obtain and validate the rules using a given tool on a given database)

17 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/discovering-quality-knowledge-relational-databases/28548

Related Content

Design Science Research: The Road Traveled and the Road That Lies Ahead
Matti Rossi, Ola Henfridsson, Kalle Lyytinenand Keng Siau (2013). *Journal of Database Management (pp. 1-8).*
www.irma-international.org/article/design-science-research/94541

Time in Multidimensional Databases
Alberto O. Mendelzonand Alejandro A. Vaisman (2003). *Multidimensional Databases: Problems and Solutions (pp. 166-199).*
www.irma-international.org/chapter/time-multidimensional-databases/26968

Information Analysis in UML and ORM: A Comparison
Terry Halpin (2002). *Advanced Topics in Database Research, Volume 1 (pp. 307-323).*
www.irma-international.org/chapter/information-analysis-uml-orm/4334

Large-Scale Sensor Network Analysis: Applications in Structural Health Monitoring
Joaquin Vanschoren, Ugo Vespier, Shengfa Miao, Marvin Meeng, Ricardo Cachuchoand Arno Knobbe (2014). *Big Data Management, Technologies, and Applications (pp. 314-347).*
www.irma-international.org/chapter/large-scale-sensor-network-analysis/85461

Biometric Databases
Mayank Vatsa, Richa Singh, P. Guptaand A. K. Kaushik (2005). *Encyclopedia of Database Technologies and Applications (pp. 42-46).*
www.irma-international.org/chapter/biometric-databases/11120