

Data Mining Techniques

Jiawei Han*

School of Computing Science, Simon Fraser University, British Columbia, Canada V5A 1S6

Abstract

Data mining, or knowledge discovery in databases, has been popularly recognized as an important research issue with broad applications. We provide a comprehensive survey, in database perspective, on the data mining techniques developed recently. Several major kinds of data mining methods, including generalization, characterization, classification, clustering, association, evolution, pattern matching, data visualization, and meta-rule guided mining, will be reviewed. Techniques for mining knowledge in different kinds of databases, including relational, transaction, object-oriented, spatial, and active databases, as well as global information systems, will be examined. Potential data mining applications and some research issues will also be discussed.

Introduction

Data mining, i.e., mining information and knowledge from large databases or information repositories, has become a highly demanding task, attracted lots of researchers and developers, and made good progress in the past several years.

This tutorial provides a comprehensive survey on the data mining techniques developed in several research communities, including data mining, data warehousing, database systems, machine learning, information retrieval, data visualization, and statistics, with an emphasis on database-oriented techniques and those implemented in applicative data mining systems.

The following issues will be examined.

1. Classification of data mining techniques. Data mining techniques can be classified according to different

*Research was supported in part by the grant NSERC-A3723 from the Natural Sciences and Engineering Research Council of Canada, the grant NCE:IRIS/PRE-CARN-HMI-5 from the Networks of Centres of Excellence of Canada, and grants from MPR Teltech Ltd. and the Hughes Research Laboratories.

views, including the kinds of knowledge to be discovered, the kinds of databases to be mined, and the kinds of techniques to be adopted.

For example, for the kinds of knowledge to be mined, one may classify data mining techniques into generalization, characterization, association, classification, clustering, pattern matching, etc., or based on the level of concepts to be discovered, into primitive level, high level, multiple-level, etc.

2. Data generalization and characterization. Two generalization approaches, *attribute-oriented induction* and *data cube*, are introduced and compared. Techniques for generalization, characterization, “roll-up” and “drill-down”, and extraction of multiple-level, different kinds of rules, including characteristic rules, discriminant rules, multi-dimensional feature tables, etc. are presented.
3. Mining association rules. Techniques for mining association rules in large transaction and relational databases are discussed, including the Apriori algorithm, mining multiple-level, generalized, and quantitative association rules, meta-rule guided mining, parallel and distributed mining, and incremental updating of discovered association rules.
4. Mining other kinds of knowledge. Efficient methods for *classification*, *clustering*, *deviation analysis*, *mining data evolution trends*, *sequential patterns*, and *other patterns* in large databases are examined.
5. Data mining in advanced or specialized database systems. Recent progress on data mining in *object-oriented*, *spatial*, *temporal*, *active*, *textual*, and *heterogeneous databases*, and the *Internet information-base* are briefly discussed.
6. A short overview of several data mining systems, including *Quest*, *DBMiner*, *KDW+*, *INLEN*, *IMACS*, *SKICAT*, *Explora*, *KnowledgeMiner*, etc.

References

The tutorial notes and related papers are accessible with the Internet address: <http://db.cs.sfu.ca/DBMiner>.