Discovery of Frequent Patterns in Large Data Collections

Hannu Toivonen

Department of Computer Science
P.O. Box 26, FIN-00014 University of Helsinki, Finland
Hannu.Toivonen@cs.helsinki.fi, http://www.cs.helsinki.fi/~htoivone/

PhD Thesis, Series of Publications A, Report A-1996-5
Helsinki, November 1996, 116 pages
ISSN 1238-8645, ISBN 951-45-7531-8

Abstract

Data mining, or knowledge discovery in databases, aims at finding useful regularities in large data sets. Interest in the field is motivated by the growth of computerized data collections and by the high potential value of patterns discovered in those collections. For instance, bar code readers at supermarkets produce extensive amounts of data about purchases. An analysis of this data can reveal useful information about the shopping behavior of the customers. Association rules, for instance, are a class of patterns that tell which products tend to be purchased together.

The general data mining task we consider is the following: given a class of patterns that possibly have occurrences in a given data collection, determine which patterns occur frequently and are thus probably the most useful ones. It is characteristic for data mining applications to deal with high volumes of both data and patterns.

We address the algorithmic problems of determining efficiently which patterns are frequent in the given data. Our contributions are new algorithms, analyses of problems, and pattern classes for data mining. We also present extensive experimental results. We start by giving an efficient method for the discovery of all frequent association rules, a well known data mining problem. We then introduce the problem of discovering frequent patterns in general, and show how the association rule algorithm can be extended to cover this problem. We analyze the problem complexity and derive a lower bound for the number of queries in a simple but realistic model. We then show how sampling can be used in the discovery of exact association rules, and we give algorithms that are efficient especially in terms of the amount of database processing. We also show that association rules with negation

and disjunction can be approximated efficiently. Finally, we define episodes, a class of patterns in event sequences such as alarm logs. An episode is a combination of event types that occur often close to each other. We give methods for the discovery of all frequent episodes in a given event sequence.

The algorithm for the discovery of association rules has been used in commercial data mining products, the episode algorithms are used by telecommunication operators, and discovered episodes are used in alarm handling systems.

Computing Reviews (1991) Categories and Subject Descriptors:
H.3.1 Information Storage and Retrieval: Content Analysis and Indexing
I.2.6 Artificial Intelligence: Learning
F.2.2 Analysis of Algorithms and Problem Complexity: Nonnumerical Algorithms and Problems

General Terms:
Algorithms, Theory, Experimentation

Additional Key Words and Phrases:
Data mining, Knowledge discovery, Association rules, Episodes