Information Mining

Rudolf Kruse and Christian Borgelt

Dept. of Knowledge Processing and Language Engineering
School of Computer Science
Otto-von-Guericke-University of Magdeburg
D-39106 Magdeburg, Germany
e-mail: {kruse,borgelt}@iws.cs.uni-magdeburg.de

Due to modern information technology, which produces ever more powerful computers every year, it is possible today to collect, store, transfer, and combine huge amounts of data at very low costs. Thus an ever-increasing number of companies and scientific and governmental institutions can afford to build up large archives of documents and other data like numbers, tables, images, and sounds. However, exploiting the information contained in these archives in an intelligent way turns out to be fairly difficult. Although a user often has a vague understanding of his data and can usually formulate hypotheses and guess dependencies, he rarely knows: where to find the "interesting" or "relevant" pieces of information, whether these pieces of information support his hypotheses and models, whether (other) interesting phenomena are hidden in the data, which methods are best suited to find the needed pieces of information in a fast and reliable way, and how the data can be translated into human notions that are appropriate for the context in which they are needed.

In reply to these challenges a new area of research has emerged, called "knowledge discovery in databases" or "data mining":

> Knowledge discovery in databases (KDD) is a research area that considers the analysis of large databases in order to identify valid, useful, meaningful, unknown, and unexpected relationships.

Often data mining is restricted to the application of discovery and modeling techniques within the KDD process. It is an interdisciplinary field that employs methods from statistics, soft computing, artificial intelligence and machine learning. Usually data mining is defined by a set of tasks, which include at least segmentation (e.g. what kind of customers does a company have?), classification (e.g. is this person a prospective customer?), concept description (e.g. what attributes describe a prospective customer?), prediction (e.g. what value will the stock index have tomorrow?), deviation analysis (e.g. why has the behavior of customers changed?), and dependency analysis (e.g. how does marketing influence customer behavior?)

Although the standard definition of knowledge discovery and data mining only speaks of discovery in *data*, thus not restricting the type and the organization of the data to work on, it has to be admitted that research concentrated mostly on highly structured data. Usually a minimal requirement is relational data. Most methods (e.g. classical methods like decision trees and neural networks) even demand as input a single uniform table, i.e., a set of tuples of attribute values. It is obvious, however, that this paradigm is hardly adequate for mining image or sound data or even textual descriptions, since it is inappropriate to see such data as, say, tuples of picture elements. Although such data can often be treated successfully by transforming them into structured tables using feature extraction, it is not hard to see that methods are needed which yield, for example, descriptions of what an image depicts, and other methods which can make use of such descriptions, e.g., for retrieval purposes.

Another important point to be made is the following: The fact that pure neural networks are often seen as data mining methods, although their learning result (matrices of numbers) is hardly interpretable, shows that in contrast to the standard definition the goal of *understandable* patterns is often neglected. Of course, there are applications where comprehensible results are not needed and, for example, the prediction accuracy of a classifier is the only criterion of success. Therefore interpretable results should not be seen as a *conditio sine qua non*. However, our own experience—gathered in several cooperations with industry—is that modern technologies are accepted more readily if the methods applied are easy to understand and the results can be checked against human intuition. In addition, if we want to gain insight into a domain, training, for instance, a neural network is not of much help.

In a plenary talk at the FUZZ-IEEE conference in Seoul in 1999 we therefore suggested to concentrate on *information mining*, which we see as an extension of data mining and which can be defined in analogy to the KDD definition as follows:

> Information mining is the non-trivial process of identifying valid, novel, potentially useful, and *understandable* patterns in *heterogeneous information sources*.

The term *information* is thus meant to indicate two things: In the first place, it points out that the heterogeneous sources to mine can already provide *information*, understood as expert background knowledge, textual descriptions, images, sounds etc., and not only raw data. Secondly, it emphasizes that the results must be *comprehensible* ("must provide a user with information"), so that a user can check their plausibility and can get insight into the domain the data comes from.

For research this results in the challenges

- to develop theories and scalable techniques that can extract knowledge from large, dynamic, multi-relational, and multimedial information sources,
- to close the semantic gap between structured data and human notions and concepts, i.e., to be able to translate computer representations into human notions and concepts and vice versa.

In this special issue several papers are collected that try to meet these challenges in different application areas—including, for example, text mining, web mining, bio-informatics and data visualization—and with a considerable number of different approaches.