

# 1 What is Data Mining?

**Data mining** is the process of discovering interesting knowledge, such as patterns, associations, changes, anomalies and significant structures, from large amounts of data stored in databases, data warehouses, or other information repositories. Due to the wide availability of huge amounts of data in electronic forms, and the imminent need for turning such data into useful information and knowledge for broad applications including market analysis, business management, and decision support, data mining has attracted a great deal of attention in information industry in recent years [4,5].

Data mining has been popularly treated as a synonym of *knowledge discovery in databases*, although some researchers view data mining as an essential step of knowledge discovery. In general, a knowledge discovery process consists of an iterative sequence of the following steps:

- *data cleaning*, which handles noisy, erroneous, missing, or irrelevant data,
- *data integration*, where multiple, heterogeneous data sources may be integrated into one,
- *data selection*, where data relevant to the analysis task are retrieved from the database,
- *data transformation*, where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations,
- *data mining*, which is an essential process where intelligent methods are applied in order to extract data patterns,
- *pattern evaluation*, which is to identify the truly interesting patterns representing knowledge based on some *interestingness measures*, and
- *knowledge presentation*, where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

With the widely available relational database systems and data warehouses, the first four processes: *data cleaning*, *data integration*, *data selection*, and *data transformation*, can be performed by constructing data warehouses and performing some OLAP operations on the constructed data

warehouses. The *data mining*, *pattern evaluation*, and *knowledge presentation* processes are sometimes integrated into one (possibly iterative) process, referred as *data mining*.

## 2 Major tasks of data mining

In general, data mining tasks can be classified into two categories: *descriptive data mining* and *predictive data mining*. The former describes the data set in a concise and summary manner and presents interesting general properties of the data; whereas the latter constructs one or a set of models, performs inference on the available set of data, and attempts to predict the behavior of new data sets.

A data mining system may accomplish one or more of the following data mining tasks [1, 3].

1. **Class description.** *Class description* provides a concise and succinct summarization of a collection of data and distinguishes it from others. The summarization of a collection of data is called class *characterization*; whereas the comparison between two or more collections of data is called class *comparison* or *discrimination*. Class description should cover not only its summary properties, such as count, sum, and average, but also its properties on data dispersion, such as variance, quartiles, etc.

For example, class description can be used to compare European versus Asian sales of a company, identify the important factors which discriminate the two classes, and present a summarized overview.

2. **Association.** Association is the discovery of *association relationships* or *correlations* among a set of items. They are often expressed in the rule form showing attribute-value conditions that occur frequently together in a given set of data. An association rule in the form of  $X \Rightarrow Y$  is interpreted as “database tuples that satisfy  $X$  are likely to satisfy  $Y$ ”. Association analysis is widely used in transaction data analysis for directed marketing, catalog design, and other business decision making process.

Substantial research has been performed recently on association analysis with efficient algorithms proposed, including the level-wise Apriori search, mining multiple-level, multi-dimensional associations, mining

associations for numerical, categorical, and interval data, meta-pattern directed or constraint-based mining, and mining correlations.

3. **Classification.** Classification analyzes a set of training data (i.e., a set of objects whose class label is known) and constructs a model for each class based on the features in the data. A **decision tree** or a set of **classification rules** is generated by such a classification process, which can be used for better understanding of each class in the database and for classification of future data. For example, one may classify diseases and help predict the kind of diseases based on the symptoms of patients.

There have been many classification methods developed in the fields of machine learning, statistics, database, neural network, rough sets, and others. Classification has been used in customer segmentation, business modeling, and credit analysis.

4. **Prediction.** This mining function predicts the possible values of some missing data or the value distribution of certain attributes in a set of objects. It involves the finding of the set of attributes relevant to the attribute of interest (e.g., by some statistical analysis) and predicting the value distribution based on the set of data similar to the selected object(s). For example, an employee's potential salary can be predicted based on the salary distribution of similar employees in the company. Usually, regression analysis, generalized linear model, correlation analysis and decision trees are useful tools in quality prediction. Genetic algorithms and neural network models are also popularly used in prediction.

5. **Clustering.** Clustering analysis is to identify clusters embedded in the data, where a cluster is a collection of data objects that are "similar" to one another. Similarity can be expressed by distance functions, specified by users or experts. A good clustering method produces high quality clusters to ensure that the inter-cluster similarity is low and the intra-cluster similarity is high. For example, one may cluster the houses in an area according to their house category, floor area, and geographical locations.

Data mining research has been focused on high quality and scalable clustering methods for large databases and multidimensional data warehouses.

6. **Time-series analysis.** Time-series analysis is to analyze large set of time-series data to find certain regularities and interesting characteristics, including search for similar sequences or subsequences, and mining sequential patterns, periodicities, trends and deviations. For example, one may predict the trend of the stock values for a company based on its stock history, business situation, competitors' performance, and current market.

There are also other data mining tasks, such as outlier analysis, etc. Identification of new data mining tasks to make better use of the collected data itself is an interesting research topic.

### 3 Data mining approaches

Data mining is a young interdisciplinary field, drawing from areas such as database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high performance computing. Other contributing areas include neural networks, pattern recognition, spatial data analysis, image databases, signal processing, probabilistic graph theory, and inductive logic programming. Data mining needs the integration of approaches from multiple disciplines.

A large set of data analysis methods have been developed in statistics over many years of studies. Machine learning has also contributed substantially to classification and induction problems. Neural networks have shown its effectiveness in classification, prediction, and clustering analysis tasks. However, with increasingly large amounts of data stored in databases for data mining, these methods face challenges on efficiency and scalability. Efficient data structures, indexing and data accessing techniques developed in database researches contribute to high performance data mining. Many data analysis methods developed in statistics, machine learning, and other disciplines need to be re-examined, and set-oriented, scalable algorithms should be developed for effective data mining.

Another difference between traditional data analysis and data mining is at that traditional data analysis is assumption-driven in the sense that a hypothesis is formed and validated against the data, whereas data mining in contrast is discovery-driven in the sense that patterns are automatically extracted from data, which requires substantial search efforts. Therefore, high performance computing will play an important role in data mining.

Parallel, distributed, and incremental data mining methods should be developed, and parallel computer architectures and other high performance computing techniques should be explored in data mining as well.

Since it is easy for human eyes to identify patterns and regularities in data sets or data mining results, *data and knowledge visualization* is an effective approach for presentation of data and knowledge, exploratory data analysis, and interactive data mining.

With the construction of large data warehouses, data mining in data warehouse is one step beyond on-line analytic processing (OLAP) of data warehouse data [2]. By integration with OLAP and data cube technologies, on-line analytical mining mechanism contributes to interactive mining of multiple abstraction spaces of data cubes.

## 4 Mining complex data in large data and information repositories

Data mining is not confined to relational, transactional, and data warehouse data. There are high demands for mining spatial, text, multimedia, and time-series data, and mining complex, heterogeneous, semi-structured and unstructured data, including the Web-based information repositories [3, 4].

Complex types of data may require advanced data mining techniques. For example, for object-oriented and object-relational databases, object-cube based generalization techniques can be developed for handling complex structured objects, methods, class/subclass hierarchies, etc. Mining can then be performed on the multi-dimensional abstraction spaces provided by object-cubes.

Spatial database stores both *spatial data* which represents points, lines, and regions, and *nonspatial data* which represents other properties of spatial objects and their nonspatial relationships. Spatial data cube can be constructed which consists of both spatial and nonspatial dimensions and/or measures. Since a spatial measure may represent a group of aggregated spatial objects, whereas multi-dimensional spatial aggregation may produce a great number of such aggregated spatial objects, it is impossible to precompute and store all of such spatial aggregations. Therefore, selective materialization of aggregated spatial objects is a good tradeoff between storage space and on-line computation time.

Spatial data mining can be performed in a spatial data cube as well as directly in a spatial database. Because of the high cost of spatial compu-

tation, a *multi-tier computation technique* can be adopted in spatial data mining. For example, at mining spatial association rules, one can first apply rough spatial computation, such as minimal bounding rectangle method, to filter out most of the sets of spatial objects which should be excluded from further consideration (e.g., not spatially close enough), and then apply relatively costly, refined spatial computation only to the set of promising candidates.

Text analysis methods and content-based image retrieval techniques play an important role at mining text and multimedia data, respectively. These techniques can be integrated with data cube and data mining techniques for effective mining of such types of data.

It is challenging to mine knowledge from World-Wide-Web because of the huge amount of unstructured or semi-structured data. However, Web access patterns can be mined from the preprocessed and cleaned Web log records, and hot Web sites can be identified based on their access frequencies and the number of links pointed to the corresponding sites.

## 5 Future research on data mining

There have been many data mining systems developed in recent years, and this trend of research and development on data mining is expected to be flourishing because the huge amounts of data have been collected in databases and the necessity of understanding and making good use of such data in decision making has served as the driving force in data mining.

The diversity of data, data mining tasks, and data mining approaches poses many challenging research issues on data mining. The design of data mining languages, the development of efficient and effective data mining methods and systems, the construction of interactive and integrated data mining environment, and the application of data mining techniques at solving large application problems are the important tasks for data mining researchers and data mining system and application developers.

Moreover, with the fast computerization of the society, the social impact of data mining should not be under-estimated. When a large amount of interrelated data are effectively analyzed from different perspectives, it can pose threats to the goal of protecting data security and guarding against the invasion of privacy. It is a challenging task to develop effective techniques for preventing the disclosure of sensitive information in data mining, especially as the use of data mining systems is rapidly increasing in domains ranging

from business analysis, customer analysis, to medicine and government.

## 6 References

1. R. Agrawal, M. Mehta, J. Shafer, R. Srikant, A. Arning, T. Bollinger. The Quest Data Mining System. Proceedings of 1996 International Conference on Data Mining and Knowledge Discovery (KDD'96), Portland, Oregon, pp. 244-249, August 1996.
2. S. Chaudhuri and U. Dayal. An Overview of Data Warehousing and OLAP Technology. ACM SIGMOD Record, 26(1):65-74, 1997.
3. M. S. Chen, J. Han, and P. S. Yu. Data Mining: An Overview from a Database Perspective. IEEE Transactions on Knowledge and Data Engineering, 8(6):866-883, 1996.
4. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.). Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
5. G. Piatetsky-Shapiro and W. J. Frawley (eds.), Knowledge Discovery in Databases. AAAI/MIT Press, 1991.