CHAPTER 1

Introduction to Data Mining

1.1 TRADITIONAL DATABASE MANAGEMENT SYSTEMS

A traditional Database Management System (DBMS) provides generic software tools and environments that support the development of a database application. A DBMS environment supports the two main functions of database development and operations: data definition and data manipulation. During the data definition stage of database development, the structural components of a database—such as table structures, primary keys, and the number of tables for the database—are determined. The tasks of data manipulation include data storage, data modification, and data retrieval through the use of queries. This concept is depicted in Figure 1.1. Other tools of a DBMS include utilities, report generators, form generators, development tools, design aids, and transaction managers.

In Figure 1.1, the application developers are those who perform logical database design and database programming for physical database development. They design Entity Relationship Diagrams (ERDs), define and construct relational tables, determine primary and foreign keys of each table, and specify data types. Database refinement and maintenance tasks are also performed by application developers. They select a particular software product from a particular vendor that can provide all the necessary tools for the design requirements.

Any DBMS software product should provide all functions of data definition and data manipulation. In addition, a DBMS software product should provide data security and integrity functions. Through these functions, the DMBS can monitor user queries, and any attempts to violate security and

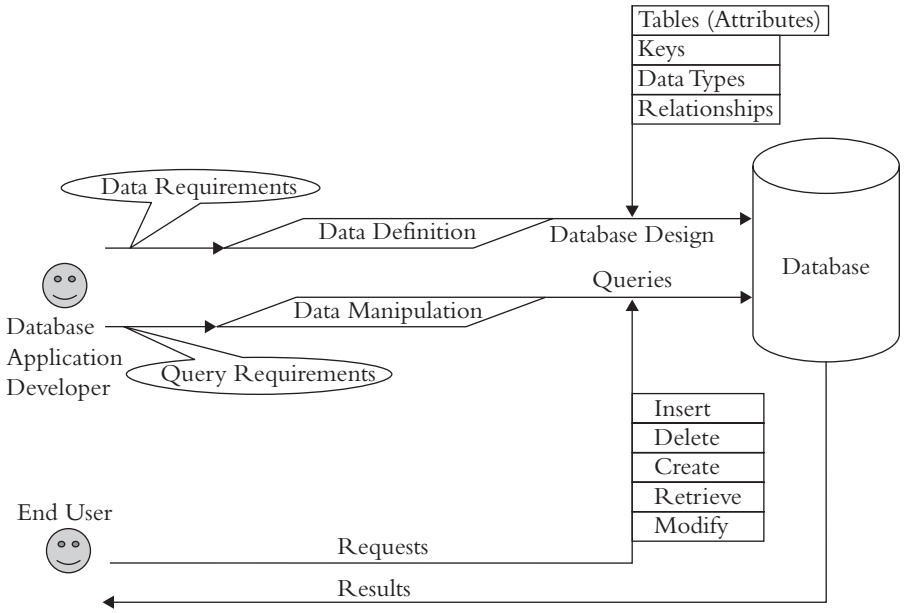

FIGURE 1.1
A simplified functional structure of a DBMS

integrity constraints will be denied. The data dictionary is another function that must be supported by the DBMS. The data dictionary contains metadata, which is data about data, and generally contains various data schemas and mappings, security and integrity constraints, and virtually anything about data that can be useful to a database developer for development and maintenance. Data recovery and concurrency functions must also be supported by the DBMS together with performance-tuning functions, since all DBMS functions should be performed as efficiently as possible.

The task of end users is to simply submit requests. Queries allow users to insert, delete, create, retrieve, and modify data in databases. Hence, user requests are submitted as database queries to a database through the DBMS, which in turn returns the result of the query execution. The efficiency of a database system is often measured by the convenience and the ease of query writing and data manipulation. Users feel more comfortable when it is easier for them to manipulate the system and write the desired queries. Therefore, the system should allow users to access all parts of the database and to form queries as easily as possible. Once the queries are formed and submitted, it is the task of the DBMS to interpret the query and execute it.

A DBMS software component is often available to allow queries to be specified in a flexible manner, such as "forms" or Query By Example (QBE).

The presentation of query results is another area that affects usability. Analysis and interpretation of query results is needed if they are to be meaningful. DBMS software products often provide "report" capabilities to help users with the interpretation of query results. However, there is still one potential issue, namely confirming their validity. If a query returns incorrect results, it must be decided whether the query was wrongly formed or the database contains invalid data. Although many commercially available software products support integrity constraints, the task of proving the correctness of query results has been a major challenge in database application system development.

In general, the effectiveness of database applications depends on the type and accuracy of the data in the database, the flexibility and convenience of user queries, clear and accurate interpretation of query execution results, and validation of query results. If the data in a database is incorrect, then regardless of the query, the result is also incorrect. Data cleansing, noise reduction, and noise removal methods may partially solve this problem, but any missing and/or invalid data must be replaced with correct data in order for the correct result to be obtained. Furthermore, convenient tools for query generation minimizes potential problems and errors. The interpretation of query results is also a very important task because this information will be used for decision making. Finally, providing justification of query results is also very important because it gives confidence and assurance to the user regarding the accuracy of the resulting data.

1.2 KNOWLEDGE DISCOVERY IN DATABASES

The rapid and constant growth of databases in business, government, and science has far outpaced our ability to interpret and make sense of this data avalanche, creating a need for a new generation of tools and techniques for intelligent and automated analysis of databases. These tools and techniques are the subject of the rapidly emerging field of data mining and Knowledge Discovery in Databases (KDD).

KDD is a process that takes a large amount of unprocessed and raw data stored in a data warehouse, transforms it into meaningful patterns,

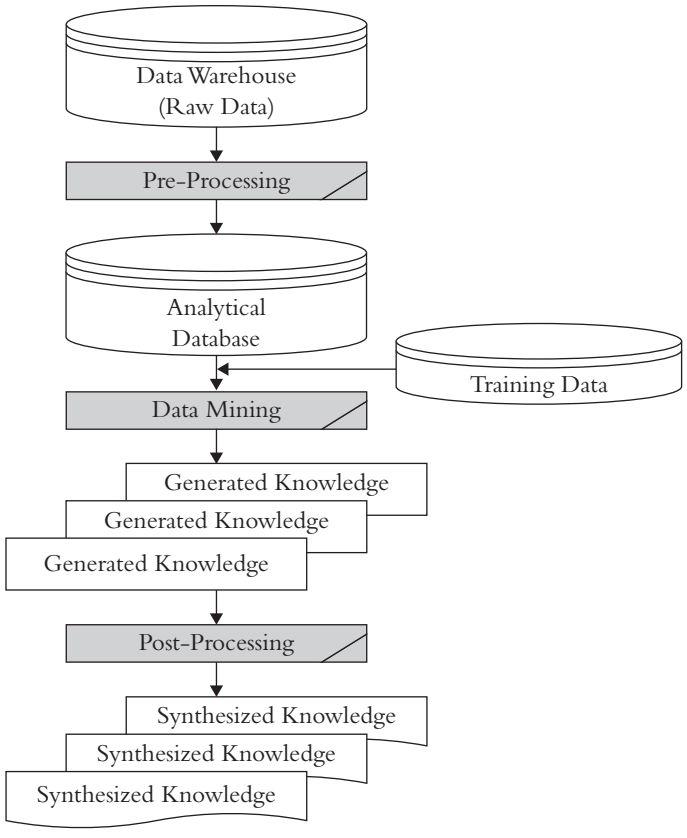

FIGURE 1.2
Knowledge discovery process

and presents them in a form that can easily be interpreted and understood. KDD is a three-step process, as illustrated in Figure 1.2.

The first step is the pre-processing stage, which takes as input a database from the data warehouse with raw data. During pre-processing, the database is cleaned, converted, and prepared for analytical processing in the next step, data mining. During the data-mining step of the KDD process, specific algorithms are applied to extract potentially useful patterns from the raw data. This is the heart of the KDD process because it identifies and exposes interesting and useful patterns and rules hidden in the database. During the data-mining stage, either an entire database or a sample containing a training dataset is used as input to the analytical processing. If the target database is extremely large, it may take too long to process all of the data, and a training

dataset should be used instead of the entire dataset to reduce the time needed for data mining. The training dataset is obtained during pre-processing, and it often contains useful information for data analysis. This dataset can also be used by the data-mining algorithms as a model of the raw data.

Once a certain amount of knowledge is generated by the data-mining stage, the generated pieces of knowledge have to be converted into a form more suitable for interpretation by end users in a post-processing stage. The main purpose of post-processing is to synthesize the generated knowledge into useful and usable information for strategic decision making by end users. During the post-processing stage, irrelevant patterns are eliminated, and relevant patterns are further summarized into more understandable and meaningful expressions. The synthesized knowledge is then usually integrated with or embedded into other systems to help improve the decision-making process.

■ 1.2.1 Pre-Processing

Once the target database has been selected, it has to be cleaned up during the pre-processing stage by eliminating incomplete data and outliers and by filling in missing data. After the dirty data is cleansed, the database has to be prepared for data mining. Depending on the particular data-mining algorithm used, the database might have to be trimmed by being sliced either vertically or horizontally. Occasionally a training dataset of samples may have to be used because the large size of the target database would require an extensive amount of processing time. The collection of data tables prepared during pre-processing constitutes the analytical database. Eventually this analytical database is passed on to the next stage of data mining.

Another task performed during pre-processing is the retrieval of attribute properties from the data dictionary. The attribute-property information helps determine the appropriateness of an attribute for use with a data-mining algorithm. Certain algorithms accept only numeric attributes or only categorical attributes, whereas others can accept a combination of attribute types. Some algorithms deal only with non-key attributes because key attributes generally do not contribute to the identification of patterns and rules hidden in tables.

An additional task of the pre-processing stage is the enforcement of the input requirements of the data-mining algorithm. The raw data in the

data warehouse has to be prepared according to the expectations of the data-mining algorithm used. Sometimes information in addition to the raw data may need to be provided for the execution of the algorithm. Sometimes multiple tables are required by the algorithm. It is the ultimate task of the pre-processor to prepare the input data for use by the data-mining algorithm.

1.2.2 Data Warehousing

Data warehousing is the process of storing as much data as possible (relevant to the task) and retrieving any part or all of it for analysis. It involves the merging of data, the cleaning up of data errors, and the storing of historical information about the data. It is often expensive, and it is a very time-consuming process.

1.2.3 Post-Processing

In predictive data mining, the post-processing step evaluates the discovered models that can be used for the prediction of future processes. In descriptive data mining, on the other hand, the post-processing step evaluates the discovered patterns and presents them in a way that can be easily interpreted and understood by end users. During post-processing, the discovered patterns are interpreted and visualized after any redundant and irrelevant patterns are removed. The useful patterns are presented to end users in a more natural and logical manner. In addition, post-processing helps the decision-making process by converting the generated knowledge into synthesized knowledge. It also checks for and resolves any potential conflicts with previously believed or previously extracted knowledge. The data-mining results synthesized through post-processing are generally integrated with other systems to improve the decision-making processes.

1.3 DATA-MINING METHODS

While KDD refers to the overall process of transforming raw data into useful knowledge, data mining is a large part of this process. Data mining involves the application of specific algorithms to databases to extract potentially useful knowledge.

Data mining is a tool used in various business sectors that provides effective, strategic assistance for decision making. The kind of information

produced by a data-mining system depends on the information needs of the organization using the system. A great variety of information can be extracted from databases using different algorithms. An efficient system is often considered to be one that provides the most assistance to the decision-making process. Simple, blind application of data-mining algorithms can lead to the discovery of meaningless and useless "knowledge" from databases. Hence, data-mining systems have to be carefully customized to fit the individual needs of the intended users.

Sophisticated analytical methods are used for the extraction of mission-critical information from databases. Different algorithms render different types of information. Users are often required to receive extensive training because they have to choose appropriate information to develop the most effective business strategies.

The dynamics of datasets (including size, dimensionality, noise, distributed nature, and diversity) often make data-mining applications difficult to develop. These properties can also make formal problem specifications more difficult to create. Furthermore, solution techniques must deal with the simplification of large volumes of data-mining results and the meaningful interpretation and user-friendly presentation of those results.

While the scope of data-mining applications is extremely wide, typical goals of data-mining applications include the thorough detection, accurate interpretation, and easy-to-understand presentation of meaningful patterns in data. Effective implementations that satisfy these goals use data-mining algorithms employing techniques from a wide variety of disciplines such as artificial intelligence, database development, statistics, and mathematics. Clustering, classification, neural networks, associations, and fuzzy theory are a few examples of algorithmic categories. In the following sections, several data-mining techniques are discussed.

■ 1.3.1 Association Rules

An association rule in a transactional database takes the form $X \Rightarrow Y$, where X and Y are sets of items appearing in transactions. An example of such a rule might be that a customer purchasing a tomato and lettuce will also get salad dressing with 80% likelihood. Data mining in a transactional database is used to find all such rules. Generally, these rules are valuable for cross-marketing, attached mailing applications, add-on sales, catalog design, store layout, and customer segmentation based on purchasing patterns.

The formal definition of the problem of data mining for association rules can be stated as follows: Let Z be a set of items that can be purchased, and let D be a set of transactions for a given period. Let $T \in D$ and $T \subseteq Z$. A unique identifier called a TID is assigned to each transaction. Mining for association rules refers to the process of extracting rules of the form $X \Rightarrow Y$ from databases containing raw data, where $X, Y \subseteq Z$, and $X \cap Y = \emptyset$. There are two factors that affect the significance of the association rules extracted: support and confidence. We say that rule $X \Rightarrow Y$ has support s in transaction set D if $s\%$ of the transactions in D contain $X \cup Y$. On the other hand, we say that rule $X \Rightarrow Y$ holds in transaction set D with confidence c if $c\%$ of transactions in D that contain X also contain Y.

Given the set of transactions D, the goal is to generate all association rules with support and confidence that are greater than a minimum support (called *minsup*) and a minimum confidence (called *minconf*). Generally, both *minsup* and *minconf* are specified by users. The transaction set D can be represented in a flat data file or as a relational table.

Chapter 2 outlines a number of data-mining techniques dealing with association rule extraction. The topics discussed include the Apriori algorithm, attribute-oriented rule induction, association rules in hypertext databases, quantitative association rules, compact association rule mining, and time-constrained association rules.

■ 1.3.2 Classification Learning

Classification learning is a learning scheme that generates a set of rules for classifiying instances into predefined classes from a complete set of independent examples, and then predicts the classes or categories of novel instances according to the generated rules.

The purpose of classification learning is to predict classes of instances, in contrast to other methods. Association learning predicts not only classes but also the attributes used in inducing the classes. In clustering, the classes are not predefined, but rather are unknown at the point of learning, and defining and identifying classes in the database is part of the learning task in clustering. In numeric prediction or regression, the classes are not discrete categories, but continuous numeric values. Regression learning uses techniques very similar to classification learning and is sometimes considered a subtype of classification learning. Therefore, a typical application of classification learning requires the

following characteristics: (1) predefined classes, (2) discrete classes (except in regression learning), (3) a sufficient amount of data (at least as many as the number of classes), and (4) attribute values that are flat rather than structured data such that the values are fixed and each attribute has either a discrete or numeric value.

Among the many algorithms used for classification learning, three major approaches toward inducing classification rules are found. The first approach uses a top-down, "divide and conquer" technique to induce knowledge rules by organizing all instances of the dataset into a "decision tree" based on a series of test outcomes on each attribute. The "divide and conquer" approach recursively selects one attribute at a time to partition the dataset into subsets based on the outcome of a test until pure subsets are obtained (i.e., all members are classified as belonging to only one class). The process of tree creation is in fact the process of heuristically searching for all possible classification rules. The classification rules can be directly generated from the tree by traversing paths from the root to each leaf.

The second approach uses a top-down, "separate and conquer" or "covering" technique to take each class in turn and to directly induce a set of rules, each covering as many instances of the class as possible (and excluding as few instances of other as possible) without erecting the tree first. After a rule is induced, the covered instances are excluded or separated from the dataset. The "separate and conquer" approach takes only one class at a time and performs all the tests to quickly purify the subset, while both subsets may not be pure in the "divide and conquer" approach. Since there are some limitations in the representation of the classified rules, this process is less accurate than the "divide and conquer approach," but it is faster because it does not follow the heuristic tree-searching procedure.

The third approach, the "partial-decision tree approach," is a combination of both the "divide and conquer" and "separate and conquer" approaches and produces rules by the induction of corresponding partial decision trees and separates the covered instances from further induction.

The goal of these approaches is to accurately and efficiently induce classification knowledge from datasets. In Chapter 3, the three classification-learning methods given above are illustrated by algorithms and by examples of how each algorithm is applied, and the advantages and disadvantages of each are compared.

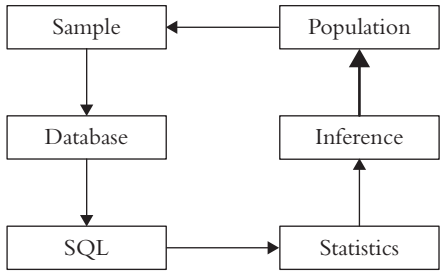

FIGURE 1.3 Role of statistical inferences for query processing

1.3.3 Statistical Data Mining

Statistics provide a useful tool for data mining, and they can be used to analyze or make inferences about data to discover useful patterns from a dataset. The database is integrated with statistical functions to draw statistical conclusions about the dataset in the database. Figure 1.3 shows an illustration of the role of statistical inferences in query processing. The population is a collection of objects from which new facts and patterns are extracted. It may contain either known objects or unknown objects. If the population is unknown, then a sample dataset has to be used to derive any facts or trends about the population. If the population is known, on the other hand, the entire population can be used for statistical processing. If the volume of the known population is very large, a sample dataset can be used for statistical processing. Without a database, statisticians perform statistical processing to make inferences about the population from the sample directly, but the sample dataset can usually be handled in a more efficient manner using SQL-like queries.

In traditional database query processing, data is retrieved using SQL-like queries with built-in database functions or operators to find the exact values of interest. Once a database is integrated with statistical functions, statisticians face two major problems. One is that the structure or format of the data or variables retrieved does not match the statistical variables or functions. The other is that the built-in functions provided by the database systems may not support the statistical calculations that need to be performed.

Statistical query processing can be performed as a two-step process. The first step involves statistical processing and the second step involves the

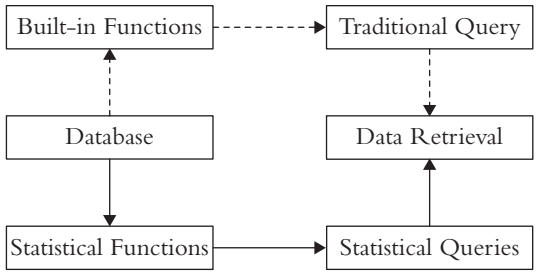

FIGURE 1.4 Statistical vs. traditional query processing

traditional query-processing operation. The transition from the first step to the second step may involve calling functions in other software to meet the needs of the statistical variables or functions in the traditional queries. This fact is illustrated in Figure 1.4. In Chapter 4, several statistical approaches are presented with examples using the data in a house sales database to do statistical analysis, interpretation, and implementation.

1.3.4 Rough Sets for Data Mining

Data mining and knowledge discovery have become increasingly important topics in database discussions. In many applications, the size of the database grows rapidly as transactions continue to occur on a day-to-day basis. These databases can be analytically exploited to discover concepts, patterns, and relationships. But real-life databases often contain data that is imprecise, incomplete, and noisy, which makes it difficult for many data-mining techniques to extract knowledge from the data. Therefore, there is a strong need for a knowledge discovery technique that can identify patterns under noisy conditions.

To deal with uncertain or inaccurate data for knowledge extraction, Bayes' theorem and rough sets are the two known methods widely used for data analysis. Bayes' theorem is based on probability theory. Hence, the interpretation of data analysis is based on the computation of conditional probabilities. Rough-set theory, on the other hand, employs rigorous mathematical techniques for discovering regularities in data and is particularly useful for dealing with ambiguous and inconsistent data. Unlike other methods, such as the fuzzy-set theory of Zadeh and various forms of neural network methods, rough-set analysis requires no external parameters and uses only

the information present in the obtained data. These two techniques have been successfully applied to medical data analysis, decision making in business, industrial design, voice recognition, image processing, and process modeling and identification.

For applications with a great deal of invalid data, it is often difficult to know exactly which features are relevant, important, and useful for the given tasks. Furthermore, certain attributes in the dataset may be undesirable, irrelevant, or unimportant. Some tuples may even contain redundant information. The number of attributes used by practical applications is often greater than 20, and the number of tuples is often greater than several hundred thousand or even more. As the size of databases and the number of attributes grows, effective data analysis techniques such as rough sets and Bayes' are needed to simplify the knowledge extraction process.

In Chapter 5, Bayes' and rough-set theories are introduced as effective methods for analyzing large amounts of data to discover knowledge from a database. The application of Bayes' and rough-set approaches to various data samples in different fields is discussed to describe the process of automated discovery of rules from a database. In rough-set analysis, important and useful information is generated through the removal and separation of redundant tuples and irrelevant attributes.

■ 1.3.5 Neural Networks for Data Mining

Most traditional DBMSs store data in the form of structured records. When a query is submitted, the database system searches for and retrieves records that match the user's query criteria. Artificial neural networks offer an attractive approach for the realization of intelligent query processing in large databases, especially for data retrieval and knowledge extraction based on partial matches. Traditional data analysis techniques make predictions about the future based on a sequence of rules generated from past data, and knowledge is obtained from a database in the form of rules. The traditional system makes predictions and creates classifications based on these rules which contain empirical knowledge.

Neural networks are different. They do not need to identify empirical rules in order to make predictions. Instead, a neural network generates a net by examining a database and by identifying and mapping all significant

patterns and relationships that exist among different attributes. The net then uses a particular pattern to predict an outcome.

The neural net tries to identify an individual mix of attributes that reveals a particular pattern. This process is repeated using a lot of training data, consequently making changes to the weights of the data for more accurate pattern matches. The model is normally built without the need for interactive human participation because the neural network can automatically identify these patterns.

The patterns that exist among the attributes in the database can be identified, and the influence of each attribute can be quantified. Neural networks simply concentrate on identifying these patterns without human guidance, whereas in traditional systems the database and any predictions on it can only be described through the rules that exist behind them.

To support the pattern-generation process of neural networks, three different types of datasets are classified and used as shown below:

Training set: The training set is used for training and for teaching the network to recognize patterns. The training process is done by adjusting the weights according to the input data.

Validation set: A set of examples is used to tune the parameters of a classifier by choosing the number of hidden nodes or hidden layers in a neural network. This set is called a validation set.

Test set: The test set is used to test the performance of a neural network. It consists of a set of examples used only to assess the performance of a fully specified classifier.

Since our goal is to build a network with the best performance based on new data, the simplest approach to the comparison of different networks is to evaluate an error function using the data that is different from the data used for training. Various networks are trained through the minimization process of an appropriate error function defined with respect to a training dataset. The performance of the networks is compared based on the evaluation results of the error function applied to an independent validation set. The network giving the smallest error rate when tested with the validation set is selected. Finally, the performance of the selected network is confirmed by measuring its performance in relation to the third independent set of data called a test set.

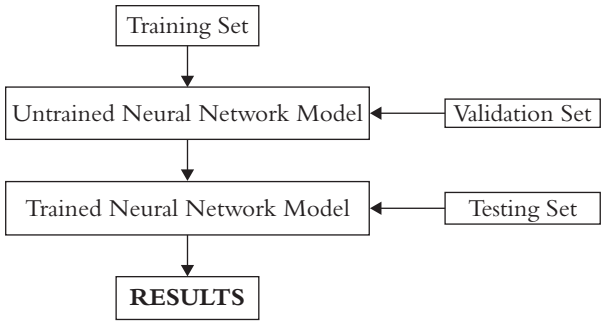

FIGURE 1.5
Neural-network-based query processing

Figure 1.5 shows the role and the relationship of the training set, the validation set, and the test set. First, the training set is used to build an untrained neural network model and to train it. The validation set is then used by the network to validate the trained neural network model and to determine the appropriate number of hidden layers and nodes. After the trained network model is built and validated, the results can be generated in response to the test sets provided. In Chapter 6, various neural network models are presented as tools for data mining and are illustrated with a number of different datasets as examples.

1.3.6 Clustering for Data Mining

Clustering can be used as a data-mining method to group together items in a database with similar characteristics. This methodology on how to group data items is based on the similarities among them. A cluster is a set of data items grouped together according to common properties and is considered an entity separate from other clusters. Hence, a database can be viewed as a set of multiple clusters for simplified processing of data analysis. One advantage of clustering is that the clusters can be profiled according to specific objectives of data analysis so that high-level queries can be formed to achieve objectives such as "identifying critical business values" or "discovering interesting patterns from the database."

As the amount of data stored and managed in a database increases, the need to simplify the vast amount of data also increases. Clustering is defined as the process of classifying a large group of independent data items into smaller groups that share the same or similar data properties. Due to the role of clustering in classifying and simplifying data, it has been extensively studied and has been demonstrated as a tool for any system dealing with massive amounts of data.

A cluster is a basic unit of a classification of initially unclassified data based on common properties. Understanding the various characteristics of clusters will help us to understand the details of the algorithms used for cluster analysis. Unfortunately, explicitly defining a cluster is somewhat difficult due to the diverse goals of clustering, and there is no universal definition. There are many definitions because different researchers define it differently. The following is a list of a few:

- A cluster is a set of entities that are alike and entities from different clusters that are not alike.
- A cluster is an aggregation of points in the test space such that the distance between any two points in the cluster is less than the distance between any point within the cluster and any other point outside the cluster.
- Clusters may also be described as connected regions of a multidimensional space containing a relatively high density of points.
- A cluster is a group of contiguous elements of a statistical population.

Considering these definitions, we can see that even if the clusters consist of entities, points, or regions, the components within the clusters are more similar in some respects than are other components outside of the clusters. A cluster can be considered a set of entities that are more similar in certain aspects within the cluster than the entities classified into other clusters. This definition deals with two important points. One is that similarity can be reflected with distance measures, and the other is that classification suggests the objective of clustering. Therefore, clustering can be defined as a process of identifying those data groups that are similar and of building a classification among them. Hence, the main objective of clustering is to identify a group of data that meets one of the following two conditions:

1. Groups whose members are very similar (similarity-within criterion)
2. Groups that are clearly separated from one another (separation-between criterion)

In Chapter 7, basic clustering concepts and techniques are presented and discussed. Procedures for handling data clusters for data mining are also demonstrated with practical examples. Several different clustering algorithms and methods are examined and compared in four different categories.

■ 1.3.7 Fuzzy Sets for Data Mining

Most of us have had some contact with conventional logic, in which a statement is either true or false with nothing in between. Although this principle of true or false has dominated Western logic for the last 2,000 years, the idea that things are either true or false does not apply in most cases. For example, is the statement "I am rich" completely true or false? The answer is probably neither true nor false since the question we need to consider is how rich *is* rich. A man with a million dollars may or may not be a rich man depending on with whom he is compared. This idea of gradations of truth is familiar to every one of us who faces decision making of any kind.

A fuzzy subset of some universe U is a collection of objects from the universe in which each object is associated with a degree of membership. The degree of membership is always a real number between 0 and 1. It measures the extent to which an element is associated with a particular set. A degree of membership 0 for an element of a fuzzy set is given to an element that is not in an ordinary set, whereas the membership value 1 is given to the elements that are in an ordinary set. Consider the fuzzy set defined as follows:

A = {1/RED, 0.3/BLACK, 0.6/PINK, 0.5/YELLOW, 1/BLUE, 0/GREEN}.

This fuzzy set indicates that:

1. Members RED and BLUE are in the fuzzy set.
2. Member GREEN is not in the fuzzy set.
3. Members B, C, and D are in the fuzzy set with partial membership values of 0.8, 0.2, and 0.7, respectively.

Mathematically speaking, a fuzzy set is a general case of an ordinary set. A fuzzy set is a set without a crisp boundary, which means the transition from "does belong to the set" to "does not belong to the set" is gradual. This gradual transition is characterized by a membership function that gives the fuzzy set flexibility in modeling commonly used linguistic expressions such as "the water is cold" or "the weather is hot." The membership degree of a fuzzy set depends on the problem that needs to be solved and on the information that is to be retrieved. The membership functions can be as simple as a linear relation or as complicated as an arbitrary mathematical function. Furthermore, membership functions can be multidimensional.

Unlike conventional set theory, which uses Boolean values of either 0 or 1, fuzzy sets have a function that admits a degree of membership in the set from complete exclusion, which corresponds to 0, to absolute inclusion, which corresponds to 1. While conventional sets have only two possible values, 0 and 1, fuzzy sets do not have this arbitrary boundary to separate members from nonmembers.

Fuzzy logic can be used to naturally describe our everyday business applications because it presents us with a flexible method to get a high-level abstraction of problem representation. In the real world, problems are often vague and imprecise, so they cannot be described in the conventional dual (true or false) logic ways. On the contrary, fuzzy logic allows a continuous gradation of truth values ranging from false to true in the description process of application models.

In Chapter 8, basic fuzzy-set theory is described. Information retrieval based on a fuzzy set is described as a data-mining example. Furthermore, problem representation with linguistic variables is presented from the viewpoint of fuzzy information retrieval. Problem-solving approaches related to fuzzy information retrieval are also described and reviewed.

■ 1.4 INTEGRATED FRAMEWORK FOR INTELLIGENT DATABASES

The data warehouses of our current market are growing exponentially in number and size, and there are very few tools on the market that fully decipher and process this information into useable information with the ease of use of natural language queries against an array of optimized data-mining algorithms. The majority of tools on the market look into the past, and they effectively use the rear-view-mirror analogy to help decision makers make decisions regarding their future. Data-mining tools are usually used by technically oriented statisticians. Herein lies the gap in time, relevance, technical proficiency, and direct access with which managers are currently faced.

A search mechanism that can find patterns or irregularities in highly data-intensive and ill-structured environments can play a very crucial role in the discovery of information essential for effective management and decision making in a very time-critical operational and strategic environment.

An intelligent database system goes beyond traditional database systems in that it deals not only with structured data but also with unstructured data

such as images, audio clips, movie clips, and hypertexts. It is often integrated with knowledge-based systems and automatic knowledge discovery systems that help convert data into knowledge. A main function of an intelligent database is to provide faster and more automatic access and service to users, even with partial and incomplete data. When user requests are clearly specified, the system's task can be focused and narrowed down for easier and faster retrieval of expected results. On the other hand, requests can be vague and ambiguous when users do not have explicit goals but are simply interested in finding any patterns and/or regularities in data. In that case, filtering of found patterns and regularities may be necessary to determine their usefulness. Intelligent databases can be defined as systems that do the following:

- Manage information in a natural way, making that information easy to store, access, and use
- Provide faster and automatic service to partial and incomplete user requests
- Can handle huge amounts of information in a seamless and transparent fashion, carrying out tasks using appropriate sets of information management tools

From the definitions above, we can see that any database system considered intelligent must deal with partial and incomplete input data and requests and must provide effective ways to manage databases to store, retrieve, modify, access, and use data for appropriate decision making.

In this section, an integrated framework for intelligent databases, called an Embeddable Intelligent Information Retrieval System (EI^2RS), is presented. This framework is used for an embeddable component of intelligent database systems that carries out the task of knowledge extraction from a series of databases containing highly structured or poorly-structured data. A diagrammatic view of the individual components of the EI^2RS structure is shown in Figure 1.6.

The EI^2RS framework can accept two types of input. One input comes from front-end managerial users of the system, and the other input comes from either computer sensors or data packages from various databases. The inputs from front-end users are generally in the form of requests framed in a semi-natural English-language format. The front-end interface accepts these user queries as input and returns the results of the analytical processing performed by the Knowledge Extraction Engine (KEE).

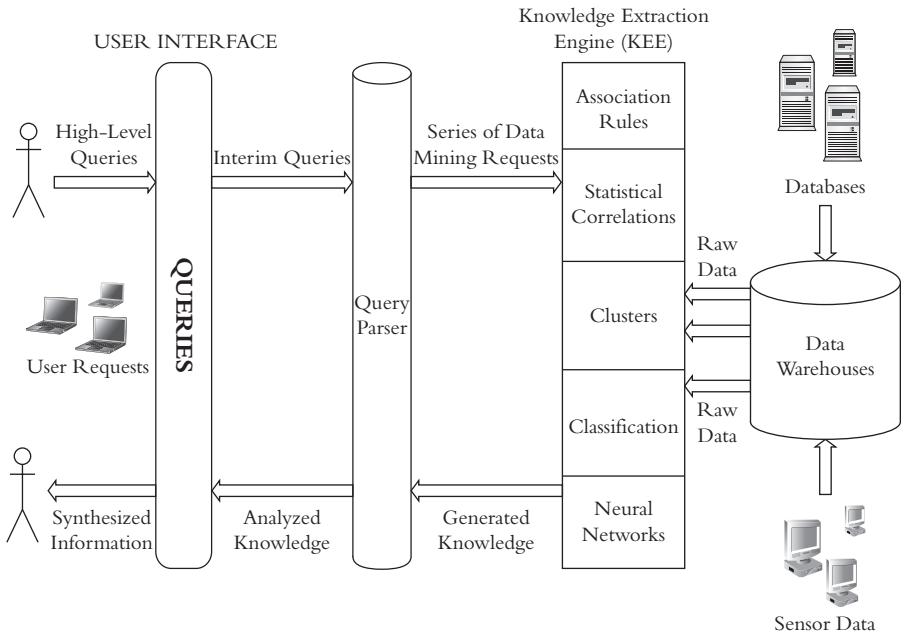

FIGURE 1.6 Embeddable Intelligent Information Retrieval System (EI²RS) Framework

The role of the EI²RS is to more effectively generate synthesized information and to provide it to decision makers by utilizing the effectiveness and efficiency of optimized data-mining algorithms against either sensor-input data or against nontechnical natural-language queries. The main function of the KEE is to support this role by extracting and synthesizing knowledge and/or information gathered from various databases and input sensors. The KEE applies various data-mining techniques such as classification, association rule mining, clustering, statistical correlation, and neural networks to perform the data analysis and synthesis process. Among the many techniques available for data mining, an optimal one is selected and applied by KEE to extract knowledge from raw data and to synthesize it for further processing.

The processing of EI²RS components is as follows. As can be seen from Figure 1.6, it first accepts either a series of user requests or a collection of sensor data as input. The user requests are then processed by a query parser, which will in turn generate a series of internal requests to be used for the selection and execution of data-mining algorithms. These algorithms are applied to extract knowledge from raw data. The KEE will accept data from either input sensors or from various databases. Then it performs an intensive

data analysis process using the optimally selected data-mining algorithms to generate useful, expected, or requested knowledge. The knowledge generated by the KEE is processed further by the query parser and is presented to the end user as synthesized information. This synthesized information will eventually be used for effective decision making.

1.5 PRACTICAL APPLICATIONS OF DATA MINING

In this section, four applications are presented to illustrate the potential implications of data mining. A brief description is given for each application. Next, the implication of data mining in each application is given. The impact and advantages in each application are also illustrated. A list of natural-language queries possible in the specific application is provided to illustrate the potential application of data-mining methods. The various data-mining techniques (described in Chapters 2 through 8) are techniques that can be used to solve these natural-language queries.

1.5.1 Healthcare Services

Data mining has been used intensively and extensively by many healthcare organizations and can greatly benefit all parties involved. For example, data mining can help healthcare insurers detect fraud and abuse, can help healthcare organizations make customer-relationship management decisions, can help physicians identify effective treatments and best practices, and can help patients receive better and more affordable healthcare services. In general, applications of data mining in healthcare services include, but are not limited to, the following:

- Modeling health outcomes and predicting patient outcomes
- Modeling clinical knowledge of decision support systems
- Bioinformatics
- Pharmaceutical research
- Business intelligence such as management of healthcare, customer-relationship management, and the detection of fraud and abuse
- Infection control
- Ranking hospitals
- Identifying high-risk patients
- Evaluation of treatment effectiveness

1.5.1.1 Implications of Data Mining in Healthcare Applications

The large amount of data generated by healthcare transactions is too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information. For example, data-mining algorithms such as the Apriori algorithm can be used to generate sets of association rules to identify relationships among different types of diseases, individual living environments, individual living habits, and individual body indices such as blood pressure, body mass index, and weight.

1.5.1.2 Natural Language Queries

a. How likely is it that a man whose age is more than 60 years old, whose weight is more than 200 pounds, and who has been diagnosed with high blood pressure will have a stroke?
b. How likely is it that an adult whose age is more than 70 years old, whose weight is more than 200 pounds, who has been told by a doctor that both blood pressure and blood cholesterol are high, who is not used to eating vegetables and is not currently taking blood pressure medication will have a heart attack?
c. How likely is it that an adult whose age is more than 70 years old and who has had a stroke will have a heart attack?
d. How likely is it that an adult who is not used to eating fruit and vegetables and does not exercise everyday will be overweight?
e. How likely is it that a man who drinks alcoholic beverages and smokes more than 20 cigarettes every day will be diagnosed with high blood pressure?
f. What are the notable characteristics of patients with a history of at least one occurrence of stroke?
g. Which medications generally have a better curative effect for stroke?
h. According to the current health status of this patient, how long is the patient likely to live?
i. According to current risk factors of a patient, what kind of treatment is likely to make him or her live longer?

j. What hospitals provide patients the best recovery rate, if he or she has a stroke, heart attack, or diabetes?

k. What are the diabetes risk factors for a male senior citizen?

■ 1.5.2 Banking

In today's world, traditional banking has changed for many reasons. Gone are the days when conducting simple surveys would enable banks to make necessary changes in their various marketing, business-process, and customer-relationship strategies. While the emergence of new banks has provided strong competition among them, it has also made it unrealistic for them to rely on only their internal procedures to stay profitable in the market.

Streamlining business procedures, improving customer relationships, detecting fraudulent characters and providing security at all levels of service, and taking other measures to improve business builds trust among not only major players of the market, but also among employees.

1.5.2.1 Implications of Data Mining in Banking

The implementation of data-mining procedures has enabled banks to improve services as described above. At all levels of service, data mining provides the historical data needed to make decisions about the implementation of future strategies. For example, coupling data mining with security involves identifying and flushing out all suspicious individuals before they get away with a crime. Credit-card companies can mine their transaction databases and look for spending patterns that might indicate transactions with a stolen card.

In marketing, data mining helps to detect any changes in customer behavior and to identify factors that may have left customers disgruntled or that may have brought in more customers. These facts will eventually enable banks to make changes that will generate assets.

1.5.2.2 Natural Language Queries

a. What potential factors will draw major industries and investors to the bank?

b. What are the main factors that leave customers unsatisfied and eventually lead them to close their accounts?

c. What are the potential types of loans that might bring profit for the bank?

d. What are the behavioral banking patterns of customers who are the most loyal and profitable for the bank?
e. What information-processing methods often leave customers disgruntled?
f. What are the most effective marketing strategies for bringing in the most customers?
g. What incentives will increase customer satisfaction?
h. What employee-hiring practices can reduce the number of customer complaints?
i. What methods are commonly used to commit fraud against the bank?
j. How easily will employees embrace new banking procedures?
k. How has the change and/or introduction of new information-processing methods affected the overall performance and/or functioning of the bank?

■ 1.5.3 Supermarket Applications

While the success stories of countless retail industries have invariably changed the way traditional businesses were once looked upon, at the same time, they have also brought to light a new range of situations a retail company might encounter. Important decisions have to be made concerning the targeted customer base, transportation of commodities to local shops, estimation of future demands, and marketing strategies, while maintaining low costs and high profits.

1.5.3.1 Implications of Data Mining in Supermarket Applications

Some of the issues faced by a supermarket management team are these: how to increase profits, what items sell faster on certain days, and at which times of year most transactions occur. The management team might also want to know, for example, how likely a customer who purchases a game controller will end up buying a game for the console as well.

To answer these questions and make these decisions, management needs to know a lot about not only the customers coming into the store but also their transaction history. A customer profile can include not only a customer's item preferences but also their age group, ethnic background, gender,

occupation, and education. People in a certain age group, for example, might be more interested in certain items than those not in that age group.

Two important decision-making factors in supermarket applications, where data mining is crucial, are transportation of commodities and manufacturing. The cost of commodities is derived largely from the way they are transported to local stores and affects the sale price of those commodities. There is almost always a trade-off in speedy transportation and price of goods. Hence, in designing data-mining algorithms, certain factors must be taken into account about how the sale of an item will be affected and how it will be transported.

When defining the price of a particular commodity, it makes sense to take a look at the targeted customer base, the overall impression of the company, and the sale of other commodities launched by the company. Data-mining methods not only help a company evaluate overall performance but also identify crucial factors that must be considered when deciding the price of a new commodity.

1.5.3.2 Natural Language Queries

a. What items in the store are popular among teenagers?
b. Do certain branches of the store sell a particular type of item more than others?
c. If an item is purchased by a customer, what other items are likely to be purchased at the same time?
d. At what time of the year are certain products more likely to be sold?
e. Do customers who buy portable MP3 devices also buy ear-bud type headphones, too?
f. How different are the buying patterns of urban customers from rural customers in terms of certain items?
g. Do the buying patterns of customers in certain age groups differ from those in other age groups in terms of frequency and amount of purchase?
h. How likely is it that an item will sell fast enough to make a good profit this year and next?
i. What items are likely to be removed from the shelves or replaced by other items?

j. If a person buys a fishing pole, how likely is it that he will buy fishing tackle and a hunting knife as well?
k. What kind of items should be stocked during holiday seasons such as Christmas, Easter, or Thanksgiving so that even if there is a sale, profits can still be maintained?
l. How likely is it that vegetarian customers will buy non-vegetarian products?

■ 1.5.4 Medical Image Classification

Breast cancer represents the second leading cause of cancer deaths in women today, and it is the most common type of cancer in women. The image classification method used for breast cancer diagnosis is based on digital mammograms. They help detect breast cancer by dividing tumors into two categories: normal and abnormal.

Normal mammograms characterize a healthy patient. Abnormal mammograms include both benign cases with mammograms showing a tumor that is not formed by cancerous cells, and malignant cases with mammograms from patients with cancerous tumors. Digital mammograms are among the most difficult medical images to read due to low contrast and differences in types of tissue. Important visual clues of breast cancer include preliminary signs of masses and calcification clusters. Unfortunately, in the early stages of breast cancer, these signs are very subtle and varied in appearance, making diagnosis difficult. The development of automatic classification systems will assist specialists with early diagnosis of breast cancer.

1.5.4.1 Implications of Data Mining in Medical Image Classification

Association-rule mining has been extensively investigated and presented in the data-mining literature. Association-rule mining typically aims at discovering associations between items in a transactional database. Hence, it can be used to discover association rules among the features extracted from the mammography database and the category to which each mammogram belongs. The association rules are constrained such that the antecedents of the rules are composed of a conjunction of features from the mammogram, while the

consequents of the rules are always the category to which the mammogram belongs. In other words, a rule would describe frequent sets of features per normal and abnormal (benign and malignant) based on an association-rule discovery algorithm.

After all the features are merged and put in a transactional database, the next step is to apply the association-rule mining algorithm to find the relevant association rules in the database. Once the association rules are found, they are used to construct a classification system that categorizes the mammograms as normal, malignant, or benign. In any learning process for building a classifier, there are two steps involved with the classification performed by association rule mining. The first one deals with the training of the system, and the second one deals with the classification of the new images.

During the training phase, the Apriori algorithm is applied to the training data to extract the association rules. During this phase, appropriate support values are used to generate the rules. In the classification phase, the low and high thresholds of confidence are set to reach the maximum recognition rate for the rules selected. Neural networks can also be used to train the system, but the advantage of using an association-rule-based classifier is that it requires less time for training.

1.5.4.2 Natural Language Queries

a. What are the physical characteristics of women who have a higher chance of getting breast cancer?
b. What is the likelihood that a tumor looks like a normal tissue?
c. How likely is it that smoking is a leading cause of breast cancer in women?
d. What are effective ways of preventing breast cancer?
e. Does the region where women live affect their chances of getting breast cancer?
f. What is the probablity that a female child born to a mother suffering from breast cancer will be affected by breast cancer tumors?
g. How likely is it that breast cancer is hereditary?
h. What is the probablity that breast cancer is caused by nutritional habits such as caffeine consumption?
i. What is the probablity that breast cancer tumors will cause other cancer tumors to grow?

j. Does breast cancer have anything to do with body weight?
k. What is the likelihood of women over 75 years of age getting breast cancer?
l. How early should breast cancer be detected to prevent it from getting worse?

■ 1.6 CHAPTER SUMMARY

This chapter began with a brief description of a traditional DBMS, which provides generic software tools and environments that support the development of a database application. The standard functions of a traditional DBMS were described to illustrate the various functions of data definition and manipulation processes in data management. However, a traditional DBMS can only support retrieval operations on data that is physically stored in the database. It does not generally go beyond what an SQL query can offer.

Today's data avalanche has created a need for a new generation of tools and techniques for intelligent and automated analysis of databases. These tools and techniques are the subject of the rapidly emerging field of data mining and Knowledge Discovery in Databases (KDD). KDD is a process that takes a tremendously large amount of unprocessed and raw data that is stored in a data warehouse and transforms it into meaningful patterns and presents them in a form that can easily be interpreted and understood. KDD is a three-step process: pre-processing, data mining, and post-processing.

In Section 1.3, various data-mining methods were described. Data mining involves the application of specific algorithms to databases to extract potentially useful knowledge. Although the scope of data-mining applications is extremely wide, typical goals of data-mining applications are the thorough detection, accurate interpretation, and easy-to-understand presentation of meaningful patterns in data. To effectively satisfy these goals, data-mining algorithms employ techniques from a wide variety of disciplines such as artificial intelligence, databases, statistics, and mathematics. Clustering, classification, neural networks, associations, and fuzzy theory are a few examples of algorithmic categories briefly described in this chapter. Further details of each technique are given in the subsequent chapters of the book.

An intelligent database system goes beyond a traditional database system in that it deals with not only structured data but also with unstructured data such as images, audio clips, movie clips, and hypertexts. It is often integrated with knowledge-based systems and automatic knowledge discovery systems that convert data into knowledge. A main function of an intelligent database is to provide fast and automatic access and service to end users, even with partial and incomplete data. When user requests are clearly specified, the system's task can be focused and narrowed down for easier and faster retrieval of results. On the other hand, sometimes users do not have explicit goals, but are simply interested in finding patterns and regularities from a database. In that case, filtering of results may be necessary to determine the usefulness of the results.

In Section 1.4, an integrated framework for intelligent databases, called an Embeddable Intelligent Information Retrieval System (EI^2RS), was presented. This framework is used as an embeddable component of an intelligent database system to carry out the task of knowledge extraction from a series of databases containing highly structured or poorly-structured data.

In the final section of this chapter, four practical applications of data mining were provided to illustrate potential implications of data mining. For each application, a brief description was first given. Then, the implication of data mining in each application was described. The impact and advantages of data mining in each application was illustrated. A list of natural-language queries that are possible in the specific application was provided to illustrate potential applications of data-mining methods.