DATA MINING

VIKRAM PUDI
Assistant Professor
International Institute of Information Technology
Hyderabad

P. RADHA KRISHNA
Principal Researcher, SET Labs
Infosys Technologies Limited
Hyderabad

OXFORD
UNIVERSITY PRESS

Contents

CHAPTER 1

INTRODUCTION

Pattern discovery is the prime prerequisite to intelligent behaviour.

Arguably, one of the main reasons behind maintaining any database is to enable human users to find interesting patterns and trends in it. While *data warehousing* is a technology that enables users to manually explore data in search of patterns and trends, *data mining* is a technology that automates the process of pattern discovery.

In this chapter, a user's perspective of these pattern discovery technologies is given. The chapter discusses various kinds of patterns that can be discovered and gives examples of their application. It also explores the relationships between these technologies and other fields such as artificial intelligence, statistics, and database systems. In addition, an introduction to various general issues relating to pattern discovery, including the interestingness of the discoverd patterns, the discovery of patterns from changing data, and people's privacy, is also given. Finally, the chapter throws some light on the major challenges related to research in this field.

1.1 MOTIVATION

Imagine that you own a chain of supermarkets in a city. You have digitized data of every sale in the past several years available with you (thanks to computers and bar-code technology!). How can this data be used optimally?

Until recently, when the pattern discovery technologies were not available, the enormous data collected was used mainly by the accounts and inventory departments. These departments used this data to monitor inventory and

finances only. The supermarket would generate megabytes of data everyday that would eventually fill the available hard disk space and be transferred to larger back-up disks or tapes. These tertiary storage spaces also would eventually be filled and the historical data collected (sometimes along with the tertiary storage devices) would be destroyed, as illustrated in Fig. 1.1.

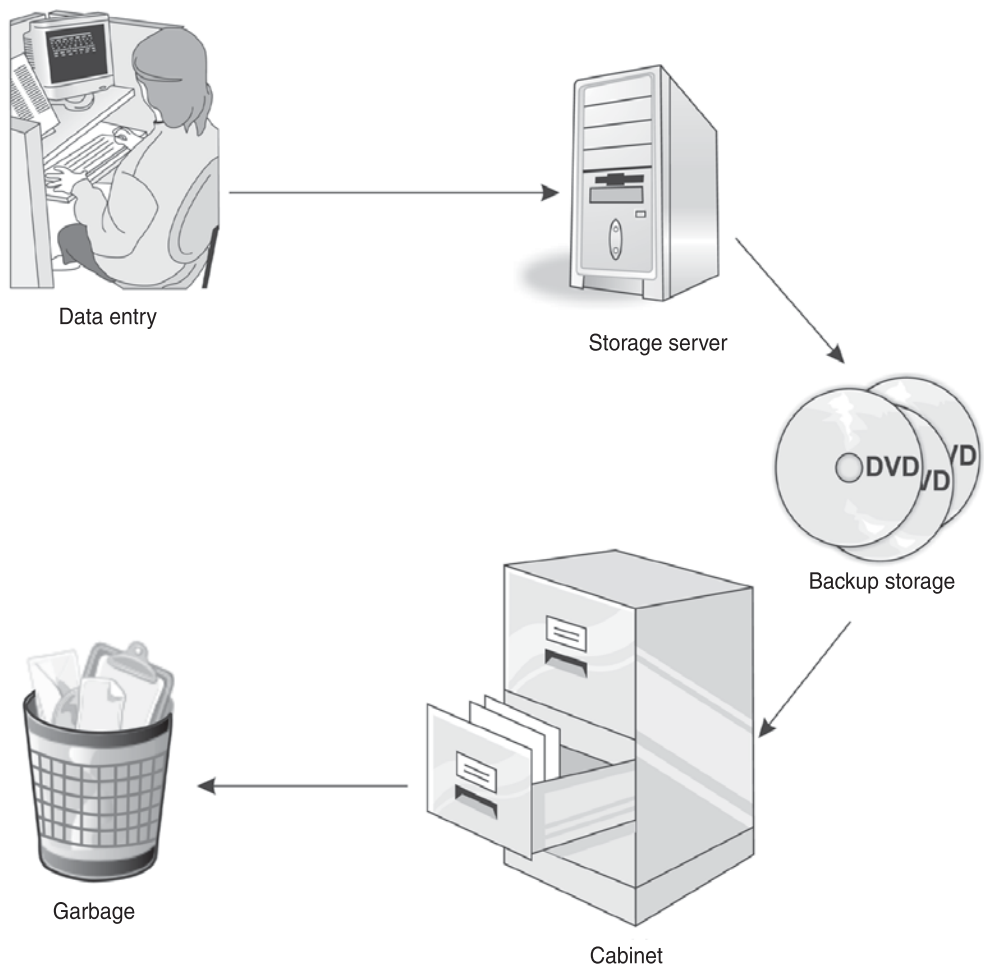

Fig. 1.1 Typical data life cycle before the birth of pattern discovery technologies

The aforementioned scenario did not make optimal use of data. So the question that arises is: Can the obsolete historical data be put to better use? The answer, we now know, is 'yes'—there is gold in the data waiting to be mined. This gold is in the form of useful patterns waiting to be discovered. Today, the worldwide market for pattern discovery technologies exceeds four billion rupees (*source*: www.olapreport.com).

Patterns are considered useful if they are *actionable*, i.e. they suggest decisions that could improve some utility. For business applications, such as the one

pictured above, utility may be measured in terms of total sales or profit. Examples of patterns that the supermarket chain owner can discover are shown in Fig. 1.2.

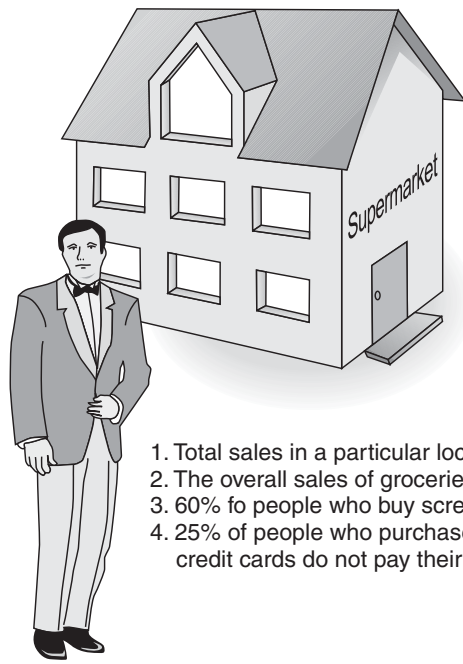

Fig. 1.2 Supermarket patterns

Consider the first pattern (see point 1 in Fig. 1.2) that says that the total sales in a particular location dropped significantly since 1997. Upon investigation, it could turn out that in 1997 a competing supermarket opened up nearby. This additional information could suggest several strategies such as negotiating with the competitor, launching an aggressive advertisement campaign, or cutting down prices.

As another example, imagine that you are the director of one of the leading computer science universities. Your university receives applications from students all over the world. These applications provide the applicant's qualifications, motivation, marks in various subjects, other achievements, and contact information. You also have historical data regarding the performance of your students in previous years such as their programming skills, research publications, marks in various subjects, and their jobs after graduation. By integrating this information, you might be able to discover patterns such as given below:

- 70% of students who scored high in mathematics also scored high in C programming.

- 90% of students from region X did not do well even though they had high marks in their various subjects.
- 80% of students who scored high in mathematics and low in C programming became professors.

The first two patterns can be useful to the admissions department, while the third one may be useful to the faculty recruitment team.

As a final example, consider a database where all the students in some university list food items and rank them on a scale of 1 to 10 (1 = like, 10 = dislike). Discovering the patterns in this dataset can help group the students based on their similarity in taste. This can help the administration decide the number and type of canteens required on the university campus to cater to the students' taste and demand (see Fig. 1.3).

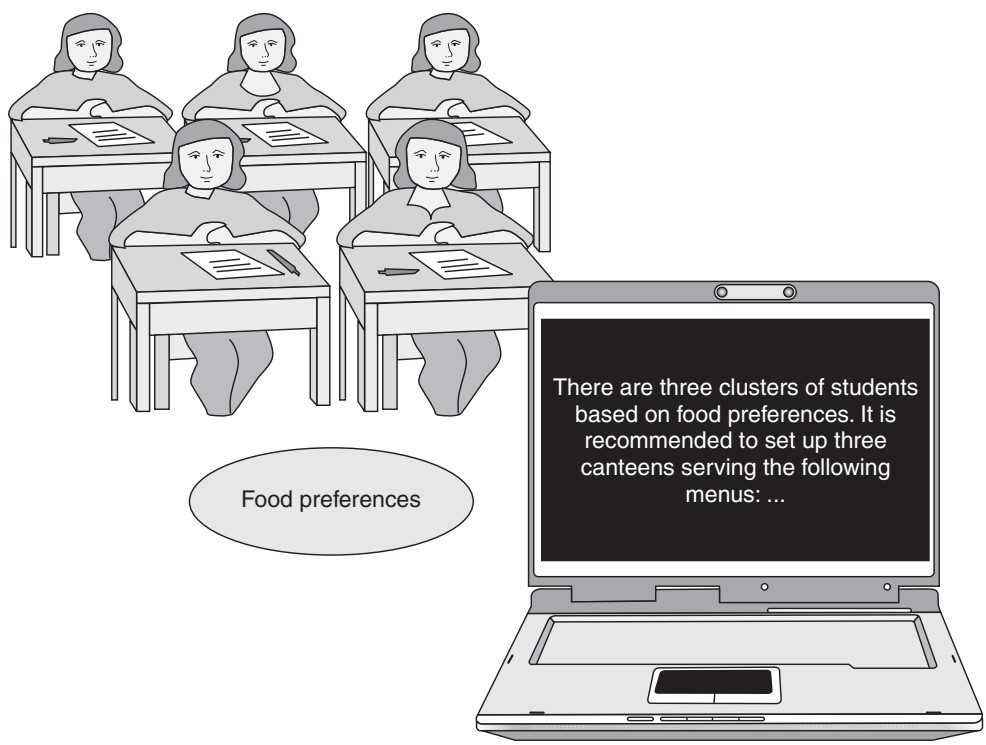

Fig. 1.3 Deciding on canteens in a university campus

1.2 Data Warehousing and Data Mining Technologies

The aforementioned scenarios are just a few simple examples of finding useful patterns. During the past several years, huge amounts of data concerning various fields have been stored, and lots more are being generated and stored in digital form. This has been made possible by the widespread use of computers and the ease with which data can be easily collected using technologies such as bar codes, radio frequency identification (RFID) tags, scanned text, digital cameras, and satellite remote sensing systems. Also, the World Wide Web offers an ever-increasing collection of data and information, most of which is known to be useful. It is humanly impossible to digest and interpret all this data without the help of automated tools.

Data warehousing is a technology that allows one to gather, store, and present data in a form suitable for human exploration. This involves *data cleaning* (removing noise and inconsistent data) and *data integration* (bringing data from multiple sources to a single location and into a common format). *On-line analytical processing* (OLAP) tools then enable us to explore the stored data along multiple dimensions, at any level of granularity, and manually discover patterns. Although this is possible using the standard relational database technology, data warehouses make the process more effective and efficient.

Knowledge discovery in databases (KDD) is the 'automatic' extraction of novel, understandable, and useful patterns from large stores of data. It is a multi-disciplinary field involving artificial intelligence, statistics, information retrieval, database technology, high-performance computing, and data visualization. Data mining, in which intelligent methods are applied to extract patterns, is an essential step in the KDD process. Other steps in the process include pre-mining tasks such as data cleaning and data integration, as well as post-mining tasks such as *pattern evaluation* (identifying the truly interesting patterns representing the knowledge) and *knowledge presentation* (presenting the discovered patterns using visualization and knowledge representation techniques).

As illustrated in Fig. 1.4, a huge amount of data is generated and stored digitally from various domains such as businesses, government organizations, research labs, and the internet. This data contains many rich patterns, which may be discovered using the pattern discovery technologies–data warehousing

and data mining. The discovered patterns can be used for decision-making in businesses and the government, or for generating and testing hypotheses while conducting research. The decisions that are implemented may ultimately have an impact on the data source; for example, it could improve sales in a business. Pattern discovery is thus an iterative feedback process.

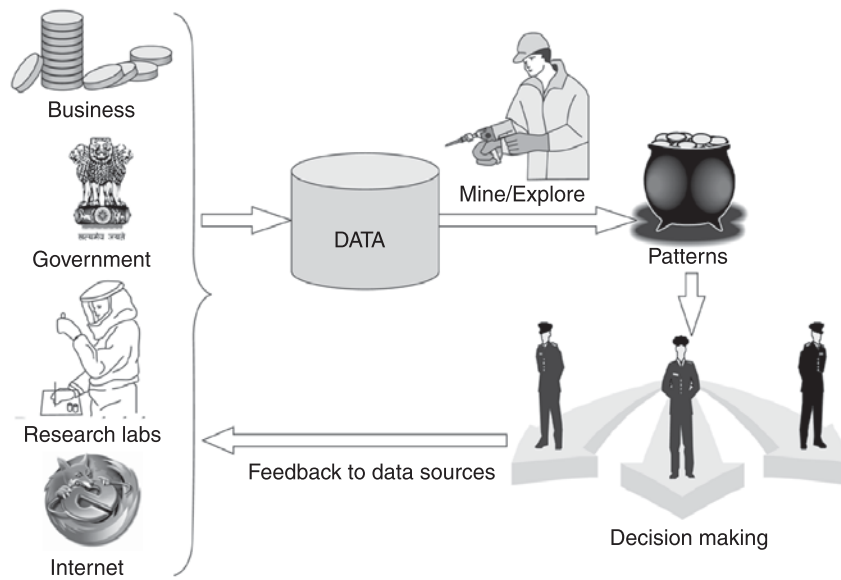

Fig. 1.4 Pattern discovery

1.3 Data Models

A data model is a description of the organization or the structure of data in an information system. The structure of data can be observed at different levels of abstraction––conceptual, logical, and physical—as described below.

Conceptual data model The manner in which users view the overall structure of the data is denoted as a conceptual data model. Entity relationship models and ontologies are examples of conceptual data models.

Logical data model The way in which a database system views the overall structure of the data is denoted as a logical data model. Examples of logical data models include relational, object-oriented, and object-relational models.

Physical data model The way the data is actually stored in a disk (in terms of cylinders and tracks) or in other storage media is denoted as a physical data model.

Real world data is often *unstructured* (e.g., plain text, video data, etc.) or *semistructured* (e.g., web data). Such data may need to be structured according to a specific data model to facilitate data warehousing and mining tasks. Even structured data may need to be restructured in different ways to bring out implicit patterns that may not be apparent otherwise. Data models are discussed in detail in Chapter 2.

1.4 Data Warehousing and OLAP: User's Perspective

A data warehouse is a repository of data *integrated* from multiple data sources. Each data source is a collection of data pertaining to some aspect of day-to-day operations in an enterprise. Data sources may be *volatile*, i.e. their accumulated data is removed periodically. But, the data werehouse is relatively *non-volatile* and accumulates data over several years (*time-variant*).

Data warehouse users are analysts who explore data to find useful patterns. They study how certain attributes of data elements (called *measures*) are related to other attributes (called *dimensions*). It is the user's job to initially specify which attributes of the original data to treat as measures and which to treat as dimensions. The data warehouse is then structured in terms of these *subjects*, i.e., measures and dimensions, to facilitate exploration. The data is conceptually organized as a multi-dimensional array, where each dimension corresponds to a dimension of the warehouse, and the values stored in each cell of the array correspond to the measures of the warehouse. This way of organizing data is referred to as a *multi-dimensional model* and the data repository is said to be *subject-oriented*.

The above description of a data warehouse is summed up in the following definition, originally proposed in 1990 by W.H. Inmon (known as the founder of data warehousing).

Data warehousing A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data to support the decision-making process of an enterprise.

Example 1.1 The owner of a supermarket chain is interested in identifying the factors that affect the sales of items.

There are three broad classes of operations to be carried out for the owner to make this study.

Data integration Each supermarket location maintains its own data. This data needs to be collected and stored in a central repository for analysis. This is not a one-time task. As the data in each location changes, the central repository needs to be updated regularly. Integration may be tedious due to two reasons: (a) different locations may use different codes for the same product; (b) the same product may be sold at different prices at different locations.

Data cleaning The use of bar code technology has made it possible for operators to enter data without errors. However, errors and inconsistencies may occasionally creep in. For example, when a new product is introduced in the supermarket, its code may not be registered in the data entry program. So whenever the product is sold, the operator may manually enter its code and price.

Aggregation The data stored at each location is very detailed; it contains the repetitive details of items in each transaction. The supermarket owner is not interested in such fine-grained data, but instead he wants to obtain a bird's eye view of the data, looking for anything that might necessitate new policies. The data needs to be aggregated (e.g., averaged or totalled) for this purpose. Common aggregation operators include average, total, count, max, and min.

The owner suspects that the supermarket location and product category are the factors that affect the sales. Further, the overall sales seem to be changing year by year. Pondering in this manner, the owner decides that for the multi-dimensional model, the measure should be total sales and the dimensions should be *supermarket location*, *product category*, and *year*.

Once the data warehouse has been built, the owner proceeds to explore several *views* of the data. Some possible views are shown in Fig. 1.5. Here, the dimension *store* represents the specific supermarket location and the dimension *product* represents the product category. *Sales* is the measure that represents the total number of sales. In Fig. 1.5(d) the total sales for each store and each product are given, whereas in Fig. 1.5(a) the overall sales of products (across all stores) are given. In Figs 1.5(b, c), sales are aggregated over the store and product dimensions, respectively. Note that several views including the most detailed view—sales totalled separately over stores, products, and years, have not been shown in this figure. Starting from the most detailed view, the owner can explore any of these views by using OLAP operations. While viewing a particular view, the owner may be interested in more details and may request

for a more detailed view; this operation is called *drilling down*. Alternatively, the owner may be interested in further aggregating along some dimension; this operation is called *rolling up*.

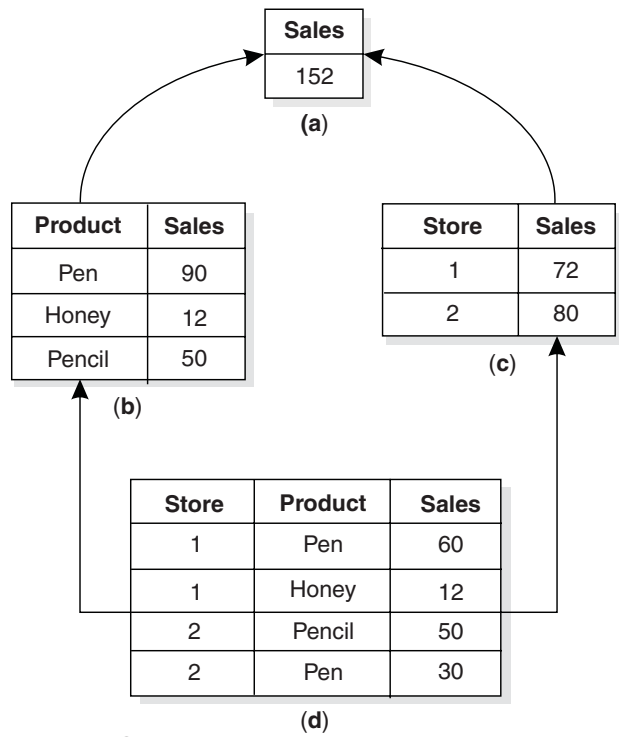

Fig. 1.5 Some possible views of the sales data

Concept hierarchies may be defined over each dimension. For example, product categories may be defined in terms of a specific model (e.g., Parker pen model 75), or an item type (e.g., all pens), or an item category (e.g., all stationary items). The term 'granularity' defines the hierarchy level at which a dimension is specified; more granularity means more detail. The owner can view the data at any level of granularity and may interactively ask to roll up to a lower level of granularity or drill down to a higher level.

The user may also specify an interest in some particular value(s) of specific dimensions. For example, the sales tables should show the total sales of items for the years 2003 and 2005 only. Computing such views is called a *slice and dice* operation. By exploring all of these views, the owner obtains a good understanding of the overall working of his supermarket chain.

Example 1.2 A university director is interested in studying how the number of publications of students is related to their grades in various subjects and their first language.

One view for this example is shown in Fig. 1.6.

Course	Grade	Language	Papers
Math	A	English	10
Math	B	English	2
C++	A	Hindi	3
C++	A	English	5

Fig. 1.6 A view for the publications example

Here, *course* represents the subject taken by students, *grade* is their grade in that subject, *language* represents their first language, and *papers* represent the total number of publications by all these students. As an example, the first row represents the 10 publications that were written by students who scored an 'A' grade in mathematics and whose first language is English.

Example 1.3 A music director in Bollywood is interested in studying the music tastes of people. He hires a programmer to write plug-ins for popular digital media players to maintain a log that keeps a track of how long and how often people listen to specific songs. He collects this log information at various locations and integrates it into one warehouse. By analysing this warehouse, he obtains an idea of the features of the music and the artists involved that make specific songs popular. Note that this exercise can be carried out on a small scale by any music enthusiast on his own data of listening patterns (data warehousing is not for only the rich and famous!). Any sufficiently large collection of data should be probed for its potential in containing interesting patterns.

Example 1.4 An organization wants a web application that can download data from the website of Indian Railways and keep track of the number of available seats in trains going to different destinations. Here the measure is the number of available seats and the dimensions could be the class of the coach (AC-2 tier, sleeper, etc.), the destination, the type of train (express, passenger, etc.), and so on. By exploring this warehouse, users can detect patterns such as 'Express trains travelling from Chennai to New Delhi typically have more vacancies in the sleeper class than the AC classes'.

Traditionally, data warehousing has been prescribed for use in business applications; however, as seen from these examples, these concepts are applicable in more general scenarios as well. In most of these examples, each record (even in the most detailed view) is actually the aggregated sum over several records in the original data sources. Hence, trying to compute views

in a naïve manner directly from the original data would consume a lot of time. The original data may be scattered across several locations, in different formats, constantly updated, partly inaccurate, and incomplete. The technical challenge in data warehousing is to overcome these hurdles and enable the user to have an *interactive* response time in obtaining multi-dimensional views.

1.5 Data Mining: User's Perspective

Human beings are adept at identifying patterns, and data warehousing depends on them for this task. Data mining, on the other hand, is a tougher job since it seeks to *automatically* discover patterns in data. To do this, first the notion of 'pattern' must be precisely defined. Unfortunately, the notion of what constitutes a pattern is unclear and hence many kinds of patterns have been defined in the context of data mining. These are described in the following.

1.5.1 Classification Models

The input data for classification consists of objects each of which belongs to a specific 'class'. For example, customers who have applied for loans from a bank could be classified into three classes: *good* (repay loan on time), *ugly* (repay loan late), and *bad* (do not repay loan). Each object in the input data is defined by a number of features or attributes, which could be numeric (such as the age and annual income) or categorical (such as the gender and occupation).

The classification system is provided with the class labels of some objects known as *exemplars* and these are said to be *labelled*. The data mining task is to compute a model from the labelled objects and use this model to predict the classes of unlabelled objects.

An implicit assumption is that the class value depends, at least to a large extent, only on the features that are actually used in the input data model. If the class value depends heavily on features that are not part of the data model, then the classification model built will not be accurate even when the finest classification techniques are used. Thus the task of selecting the right features for a particular application is important.

1.5.2 Regression Models

In classification, the field being predicted consists of discrete-valued classes. Regression is similar to classification except that the field being predicted comes from a real-valued domain. For instance, we can use regression in the

earlier loan example to predict the *time* when customers will repay loans instead of merely predicting the class of customers.

Regression has numerous applications. In businesses it is commonly used to determine how the volume of sales of products can be affected if one modifies other parameters such as the cost price, quality, etc. Regression is also used to determine how different physical parameters are related (e.g., the temperature and the pressure of a liquid).

Like in classification, regression techniques also assume that the predicted field depends only on the features that are actually used in the input data model. Most regression techniques further assume some specific form for this dependency. For instance, *linear regression* assumes that the predicted field can be described as a weighted sum of the features.

1.5.3 Time-series Patterns

In time-series patterns, the input data consists of values of attributes as they change over time. Examples include stock market data, traces of scientific experiments, medical treatments, etc. The goal of mining such data is to form a model that can be used to predict the time series for future values of time. While this can be considered as a special case of regression, it is often considered separately because it is an important case and techniques have been developed that are suitable only for time-series prediction. There are also other studies that take a sequence as a query and retrieve similar sequences or sub-sequences from a large database of sequences.

Classification, regression, and time-series models are used for prediction and hence, designated as *predictive* models. The other kinds of patterns, described hereafter, are primarily meant to describe the input data in a succinct way and are called *descriptive* models.

1.5.4 Clusters

The input data model for clustering is similar to that in classification, except that there are no class labels for objects. The task here is to cluster or group the data objects in such a manner that the objects in a cluster are very similar to each other (maximize intra-cluster similarity) and the objects that are not in the same cluster are significantly different from each other (minimize inter-cluster similarity).

Defining the exact notion of similarity between two data objects is a non-trivial task. Several similarity measures have been proposed in the research

literature for various types of data objects. Applications of clustering include market segmentation for identifying the target groups of people requiring similar promotion strategies, discovering types of stars in datasets of stellar objects, and so on.

1.5.5 Summaries

The input data model for summarization is typically a relational model. A summary of a dataset is a succinct representation showing data at a low level of granularity. It is obtained by identifying attributes such as the customer name, address, etc., that have too many distinct values and either removing them or performing a roll-up operation identical to that in data warehousing. Summarization is also known as *characterization* or *generalization*.

Alternatively, standard statistics (such as the mean or some other measure of central tendency) can be derived from the data to represent its summary.

1.5.6 Frequent Patterns

In frequent pattern mining, the input data consists of records, each of which contains a set (or sequence) of items. Examples include customers buying sets of items from a supermarket or users visiting a sequence of web pages. The task in this case is to find subsets (or sub-sequences) that occur more frequently than some user-specified threshold.

If X and Y are sets of items, then an association rule $X \rightarrow Y$ is a statement of the form: 'Among the records that contain set X, c% also contain set Y'. Sequential rules are similar except that X and Y are sequences instead of sets. These rules can be easily calculated once the frequent patterns have been mined.

Frequent patterns can be considered as a succinct summary of set- and sequence-valued attributes. They have been found to be useful for classification and clustering tasks especially on large datasets. These broad application areas indicate that frequent pattern mining is an important area of study. Table 1.1 sums up the above discussion on pattern types.

Table 1.1 Pattern types

Pattern Type	Input Data Model	Description	Example Applications
Classification	Data objects consisting of numeric or categorical features. Each object belongs to one among a finite set of classes.	A model learnt from input data that can be used to predict class labels of objects whose class is unknown.	Detecting e-mail spam, computer intrusion, fraud. Speech recognition. OCR.
Regression	Data objects consisting of numeric or categorical features. One real-valued feature is designated as a response variable.	A model learnt from input data that can be used to predict the value of the response variable when it is unknown.	Variation of sales with cost-price (in business). Variation of temperature with pressure (in science)
Time series	Values of attributes as they change over time.	A model learnt from input data to predict the time series for future values of time.	Stock market data, Scientific experiments, Medical prognosis
Clusters	Data objects consisting of numeric or categorical features.	Clusters of objects such that objects in a cluster are similar and objects not in the same cluster are significantly different.	Market segmentation, Spatial data segmentation
Summaries	Relational database table	Table obtained by removing or summarizing attributes with too many distinct values.	Any application having large relational tables
Frequent patterns	Records consisting of sets or sequences of items.	All sets or sequences that frequently occur within records as subsets or sub-sequences	E-commerce, census analysis, sports, Medical diagnosis, Web search

1.6 Related Disciplines

Data mining is a multi-disciplinary area and draws upon resources from many subjects. We will discuss these subject areas one by one in the following sections.

1.6.1 Artificial Intelligence

The task of automatically discovering patterns in data has so far been the domain of artificial intelligence (AI). Two aspects differentiate data mining techniques from those in AI. First, data mining emphasizes the human understandability of discovered patterns; whereas in AI, the discovered patterns are meant to be used by the machine itself. Second, data mining techniques are meant to be scalable to huge stores of data such as the World Wide Web. In contrast, the traditional AI approaches have mostly been researched using small 'toy' datasets that fit in the main memory.

Data mining has borrowed a good deal from AI, especially from the field of *machine learning* which concerns itself with constructing programs that automatically improve with experience. Almost all classification techniques of machine learning have been used in data mining, either directly or adapted to scale to huge datasets. Only those classification models that are not easily understandable by human users, such as some neural network techniques, have been omitted.

1.6.2 Statistics

Research in statistics has produced sophisticated techniques to analyse collections of data. Most of these techniques can be considered as also belonging to data warehousing and mining. Data distributions, measures of central tendency, histograms, and samples are summaries of data and hence can be viewed as descriptive patterns. Warehousing and mining can be viewed as computer-assisted, simplified statistical exploration of data.

1.6.3 Information Retrieval

Information retrieval involves retrieving information from textual data such as web data or digital libraries. Many classification and clustering approaches have had their origins in this field. Notions of information content (entropy) in a collection of data, information gain of an attribute, maximum entropy models, concept hierarchies, and similarity measures have been borrowed from the field of information retrieval.

1.6.4 Data Compression

Data compression is the computing of a concise version of data for the purpose of later retrieval of original data by decompression. In a sense, data warehousing and mining tasks strive to show the human user a concise version of the original data. This concise version is supposed to represent the patterns that

the entire data follows. Due to this common goal, there is much scope for these fields to interact closely.

1.6.5 Database Technology

Data warehousing and mining are considered to be a part of the database technology by most researchers. However, traditional database technology is generally associated with on-line transaction processing (OLTP), which is concerned with providing efficient and concurrent read/write access to the data stored in relational form. Queries in OLTP systems are ordinarily simple views of small parts of the data. These systems are highly optimized to support efficient updates to the data. Data warehouse technology on the other hand is more concerned with OLAP, where complex queries with multiple aggregations over large parts of the data are involved. Data mining also involves complex algorithms that call for the data to be read in non-traditional ways. Efficient updates are not as much an issue with OLAP systems. In spite of these differences, both data warehousing and mining technologies have the option of using relational databases for storing data.

1.6.6 High-performance Computing

Both data warehousing and mining are *highly computation intensive* because they require operating on huge quantities of data in complex ways. Hence researchers have been studying the option of using additional hardware and designing parallel algorithms to simplify these tasks. Often, several machines are required to store the huge amount of raw data itself. In such cases, it would be an added benefit to make use of all these machines for parallel processing of warehousing and mining tasks.

1.6.7 Data Visualization

Since both warehousing and mining are user-driven tasks, the end product should be visually appealing and semantically rich in describing the discovered patterns. Various kinds of diagrams, charts, and graphs have been designed to display the output. However, sometimes neatly formatted text may be desirable. In either case, care needs to taken to ensure that the output shows all the important patterns without burdening the user with too much information. After all, the purpose of warehousing and mining is to enable humans to digest vast amounts of the available information. This is made possible by organizing the output well. Also each visual element on screen can be decorated in different ways to convey additional information. For

example, if the output consists of a tree of nodes, each node can be coloured differently to indicate its type.

1.7 Other Issues

In this section, we will discuss some other issues having some bearing on the understanding, study, and simplification of data mining and warehousing concepts and techniques.

1.7.1 Interestingness of Discovered Patterns

The notion of what constitutes an interesting pattern is important in the context of data mining. Because, if computer programs are to automatically discover interesting patterns, they need to know the meaning of interestingness. Even in data warehousing, the notion of interestingness is important because if it is well defined, then the interesting parts from the whole data can be highlighted to enable better exploration by human users.

Different metrics of interestingness have been developed for different kinds of patterns such as classification models, clusters, and frequent patterns. These will be explored in detail in later chapters. Here we list some generic characteristics of interesting patterns.

Stability The stability of a pattern is a measure of the invariance of the pattern throughout the entire data collection. Patterns that hold in just some parts of the data are not so stable. Several objective metrics exist to determine stability.

Novelty A pattern is novel if it was not known previously. Novelty is a subjective measure. Measuring novelty is possible if the users' knowledge of the data is extracted and modelled in some way.

Actionability A pattern is actionable if its existence leads to some profitable decision. This measure is generally application dependent.

1.7.2 Incremental Pattern Discovery

In most organizations, the historical dataset is dynamic in that it is periodically updated with fresh data. For such environments, pattern discovery is not a one-time operation; it is a recurring activity, especially if the dataset has been significantly updated. Pattern discovery may need to be repeated over and over in order to evaluate the effects of the strategies that have been implemented

based on previously discovered patterns. For example, a supermarket owner may decide to package soaps and shampoos together for a discount, based on a discovered pattern that showed that many people purchase soaps and shampoos together. After the strategy is implemented, the owner will want to mine new sales data to see whether or not people are availing this discount. The strategy will need to be revised, reviewed, and/or repeated in light of the conclusions drawn from the data.

In an overall sense, pattern discovery is essentially an exploratory activity and therefore, by its very nature, operates as a *feedback* process where new discoveries are guided by previous discoveries. In this context, it is worthwhile to consider using previously discovered patterns in order to discover new ones instead of processing the updated dataset from scratch. Depending on the context, it may be necessary to mine patterns from the entire updated dataset, or just from the increment, or from both.

In this context, as illustrated in Fig. 1.7, it is worthwhile to consider the possibility of incremental mining—that is, to use the previously mined patterns

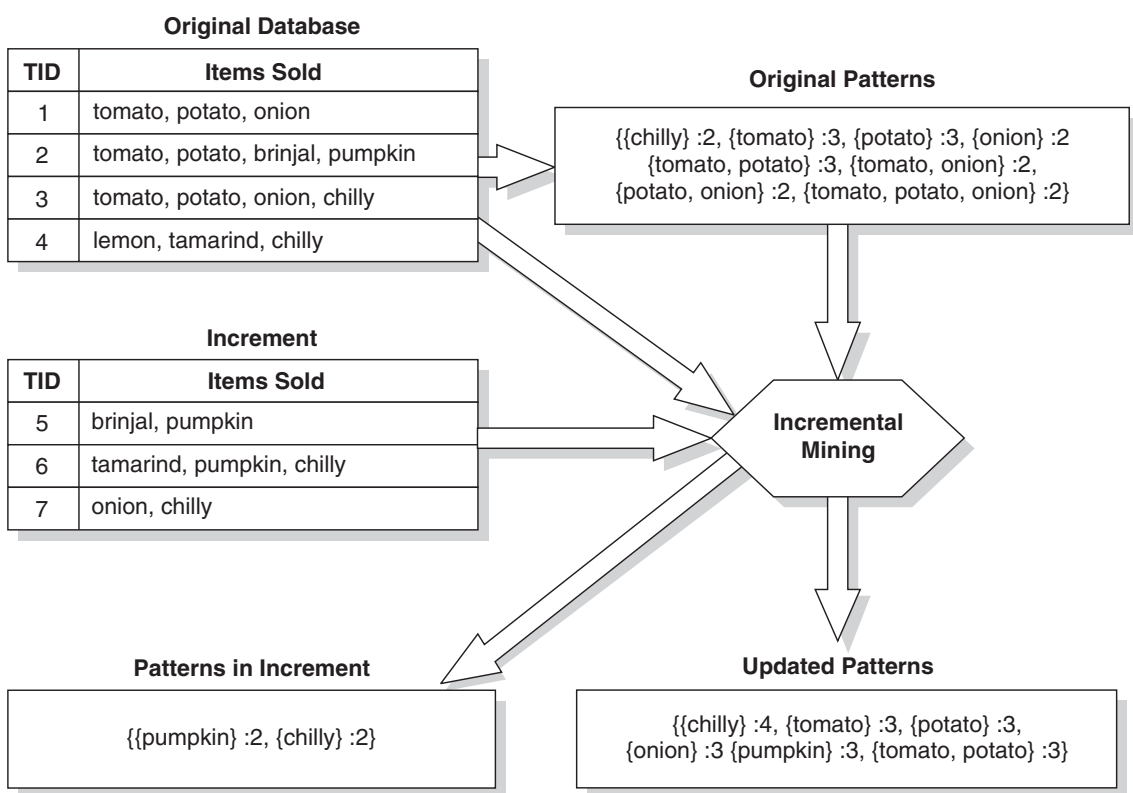

Fig. 1.7 Incremental mining

in order to perform the new mining operation more efficiently. Ideally, it is desirable if we can get away by mining only the freshly added data in order to get the updated patterns. Unfortunately, it is not always possible. Studies have shown that it is sometimes necessary to access the original data in order to mine the updated patterns. However, the incremental mining algorithms are usually able to process the original data very quickly. These techniques can result in substantial benefits, especially as the dataset grows (the original dataset will eventually be many orders of magnitude larger than the fresh data).

1.7.3 Privacy

The prospect of striking gold from data has led many organizations to collect additional information from the people they deal with. For example, a music website may lure you into revealing your music preferences and contact information by promising to deliver a free T-shirt. The intention may be honourable because the purpose is to discover patterns that help make decisions that ultimately benefit everyone concerned. In this case, the data you provide can be used by the website maintainers to deliver a more popular content.

However, some data could be personal (e.g., music preferences) or sensitive (e.g., medical profiles). Sometimes, revealing simple contact information could make a person victim to numerous calls from sales people. This problem is worsened by the fact that some organizations sell their data to third parties–– for instance, hotels in a particular city will be interested in luring potential customers who have recently booked a flight to that city. In a futuristic scenario (perhaps not too far in the future), some web mail servers and search engines may monitor your actions and automatically decide that you are potentially interested in some product, and then alert some vendors with your contact information!

Most people are concerned about protecting their privacy. They do not reveal any data about themselves, unless necessary. Often they provide inaccurate data. Consequently, most data collected by ad hoc sales people or filled out on web forms is inaccurate.

What most people view as a problem, researchers view as an opportunity! They have been studying the privacy problem from different angles as follows:

1. How can patterns be discovered from inaccurate data?
2. How can data be modified to ensure privacy without tempering with the inherent patterns?

3. How can data be modified minimally to hide sensitive patterns?
4. Can discovered patterns be used to retrieve the original data records?

The last question can be answered affirmatively for patterns that are very specific. For example, a pattern such as 'men in ABC company who wear red hats also grow beards' would compromise privacy if there are just a few men who wear red hats.

1.8 Future Trends

Data warehousing and mining technologies have been well researched and deployed in many real-world scenarios. Though research to develop more efficient algorithms for specific components continues, the major focus should be on integrating all warehousing and mining components together with database backends into a *simple unified framework*.

In the context of algorithms for specific components, almost all of them need the user to provide various parameters based on prior experience. Selecting values for these parameters is currently a 'black art', so techniques will be developed to obtain them automatically. Along a similar thread, research is expected to proceed to design algorithms that do not require any parameters at all. A true integration of pattern discovery technologies into database backends should be seamless. In the final scenario, the user will not have to even initiate the process of pattern discovery. Instead, the system will show the patterns as and when they emerge in the data––a technology that can well be called 'proactive data mining'.

Summary

One of the main reasons behind maintaining any database is to enable the users to find interesting patterns and trends in the data. Interesting patterns must be stable, novel, and actionable. Pattern discovery is a multi-disciplinary field, drawing work from artificial intelligence, statistics, information retrieval, database systems, high-performance computing, and data visualization.

Data warehousing is a technology that enables users to explore data in search of patterns. Data warehouse users are analysts who explore the data represented using a multi-dimensional model, in search of useful patterns.

Their exploration involves complex multiple aggregation queries over the data.

Data mining automates the process of pattern discovery. Various kinds of patterns can be discovered such as classification and regression models, clusters, frequent patterns, time-series patterns, and summaries. The algorithms used to discover these patterns typically require several parameters to be input from the user. One future research direction is to automate the process of setting values to these parameters. The integration of data warehousing and mining components with the database backends into a simple unified framework is another major area open for research.

EXERCISES

Test Your Understanding

1. Differentiate between the following:
   (a) Data warehousing and Data mining
   (b) Classification and Clustering
   (c) Classification and Regression
   (d) Data mining and Statistics
2. What is incremental pattern discovery? How is it done?
3. What is frequent pattern mining? What is it useful for?
4. What is the relationship between data compression and data mining?
5. What are the various privacy related issues in pattern discovery technologies?

Improve Your Research Skills

1. In our daily life, we discover patterns in the world around us. We constantly apply classification, regression, clustering, frequent pattern mining, time-series mining, and summarization. Mention one instance in your daily life where you (perhaps subconsciously) apply each of these mining tasks.
2. Explore journals, technical magazines, and the Internet, and write a survey on the applications of pattern discovery technologies. Categorize these applications and mention what you feel is the most successful application in each category.

Improve the Field

1. Several kinds of patterns have been suggested in the research literature. Explore these and come up with the most elementary kinds of patterns from which all other kinds of patterns can be derived?
2. Describe how proactive mining software should look and behave if it is to be used by non-technical users and yet supports the discovery of complex patterns.