1. **Why Data Mining?**
	1. We live in a world where vast amounts of data are collected daily. Analyzing such data is an important need.
		1. **Moving toward the Information Age:** we are living data age but with help of Data mining turns a large collection of data into knowledge we will live in Information Age.
		2. **Data mining as the Evolution of Information Technology:** its result of the natural evolution of information technology. Nowadays numerous database systems offer query and transaction processing as common practice. Advanced data analysis has naturally become the next step. As a result, data collected in large data repositories become “data tombs” data mining try to make it “golden nuggets”.
	2. **What Is Data Mining?** Its “knowledge mining from data,” The knowledge discovery process is: 1- **Data cleaning** (to remove noise and inconsistent data) 2- **Data integration** (where multiple data sources may be combined) 3- **Data selection** (where data relevant to the analysis task are retrieved from the database). 4- **Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations) 5**- Data mining** (an essential process where intelligent methods are applied to extract data patterns). 6- **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on interestingness measures) 7 - **Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users)
	3. **What Kinds of Data Can Be Mined?**
		1. **Database Data:** Relational data can be accessed by **database queries,** when **mining relational databases**, we can go further by *searching for trends* or *data patterns.*
		2. **Data Warehouses**: A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site. By providing multidimensional data views and the precomputation of summarized data, data warehouse systems can provide inherent support for OLAP. Online analytical processing.
		3. **Transactional Data: each record in a transactional database captures a transaction, such as a customer’s purchase.**
		4. **Other Kinds of Data**: time-related or sequence data (e.g., historical records, stock exchange data, and time-series and biological sequence data), data streams (e.g., video surveillance and sensor data, which are continuously transmitted), spatial data (e.g., maps), engineering design data (e.g., the design of buildings, system components, or integrated circuits), hypertext and multimedia data (including text, image, video, and audio data), graph and networked data (e.g., social and information networks), and the Web (a huge, widely distributed information repository made available by the Internet).
	4. **What Kinds of Patterns Can Be Mined?** There are a number of ***data mining*** functionalities **descriptive** and **predictive.**
		1. **Class/Concept Description: Characterization and Discrimination:** Data entries can be associated with classes or concepts **Data characterization** is a summarization of the general characteristics or features of a target class of data.The output of data characterization can be presented in various forms **pie charts**, **bar charts**, **curves**, **multidimensional data cubes. Data discrimination** is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.
		2. **Mining Frequent Patterns, Associations, and Correlations:** are patterns that occur frequently in data. Association rules are discarded as uninteresting if they do not satisfy both a **minimum support threshold** and a **minimum confidence threshold**. Additionalanalysis can be performed to uncover interesting statistical **correlations** between associated attribute**.**
		3. **Classification and Regression for Predictive Analysis: Classification** is the process of finding a **model** (or function) that describes and distinguishes data classes or concepts. The model are derived based on the analysis of a set of **training data** A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.
		4. **Cluster Analysis:** Unlike classification and regression, which analyze class-labeled (training) data sets, **clustering** analyzes data objects without consulting class labels. The objects are clustered or grouped based on the principle of *maximizing the infraclass similarity and minimizing the interclass similarity*.
		5. **Outlier Analysis:** A data set may contain objects that do not comply with the general behavior or model of the data. These data objects are **outliers.** In some applications (e.g., fraud detection) the rare events can be more interesting than the more regularly occurring ones.
		6. **Are All Patterns Interesting?** The answer is no—only a small fraction of the patterns potentially generated would actually be of interest to a given user.
	5. **Which Technologies Are Used?**
		1. **Statistics:** Statistics studies the collection, analysis, interpretation or explanation, and presentation of data. Data mining has an inherent connection with statistics. A statistical model is a set of mathematical functions that describe the behavior of the objects in a target class.
		2. **Machine Learning: Machine learning** investigates how computers can learn (or improve their performance). **Supervised learning** is basically a synonym for classification. **Unsupervised learning** is essentially a synonym for clustering. **Semi-supervised learning** is a class of machine learning techniques that make use of both labeled and unlabeled examples when learning a model **Active learning** is a machine learning approach that lets users play an active role in the learning process.
		3. **Database Systems and Data Warehouses: Database systems research** focuses on the creation, maintenance, and use of databases for organizations and end-users. A **data warehouse** integrates data originating from multiple sources and various timeframes.
		4. **Information Retrieval: Information retrieval** (**IR**) is the science of searching for documents or information in documents. Documents can be text or multimedia, and may reside on the Web. Information retrieval assumes that (1) the data under search are unstructured; and (2) the queries are formed mainly by keywords, which do not have complex structures.
	6. **Which Kinds of Applications Are Targeted?**
		1. **Business Intelligence** (**BI**)**:** technologies provide historical, current, and predictive views of business operations. Examples include reporting, online analytical processing, business performance management,
		2. **Web Search Engines:** A **Web search engine** is a specialized computer server that searches for information on the Web. The search results of a user query are often returned as a list (sometimes called *hits*).
2. **Getting to Know Your Data.**
	1. **Data Objects and Attribute Types:** Data sets are made up of data objects. A **data object** represents an entity, Data objects are typically described by attributes, Data objects can also be referred to as *samples, examples, instances, data points*, or *objects,* the rows of a database correspond to the data objects, and the columns correspond to the attributes.
		1. **What Is an Attribute?** An **attribute** is a data field, representing a characteristic or feature of a data object. The term *(dimension)* is commonly used in data warehousing. Machine learning (*feature*) statisticians prefer the term (*variable*) data mining and database (*attribute*). The **type** of an attribute is determined by the set of possible values.
		2. **Nominal Attributes:** Nominal means “relating to names.” The values of a **nominal attribute** are symbols or *names of things*. Also referred to as **categorical**. The values do not have any meaningful orders.
		3. **Binary Attributes:** A **binary attribute** is a nominal attribute with only two categories or states: 0 or 1. A binary attribute is **symmetric** if both of its states are equally valuable and carry the same weight like gender. A binary attribute is **asymmetric** if the outcomes of the states are not equally important such as the *positive* and *negative* outcomes of a medical test for HIV.
		4. **Ordinal Attributes**: An **ordinal attribute** is an attribute with possible values that have a meaningful order or *ranking* among them, but the magnitude between successive values is not known.
		5. **Numeric Attributes:** A **numeric attribute** is *quantitative*; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be *interval-scaled* or *ratio-scaled*. **Interval-scaled attributes** are measured on a scale of equal-size units. A **ratio-scaled attribute** is a numeric attribute with an inherent zero-point.
		6. **Discrete versus Continuous Attributes:** A **discrete attribute** has a finite or countably infinite set of values, which may or may not be represented as integers. **Continuous**. The terms *numeric attribute* and *continuous attribute* are often used interchangeably in the literature.
	2. **Basic Statistical Descriptions of Data:** Basic statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.
		1. **Measuring the Central Tendency: Mean, Median, and Mode**: The Mean is (algebraic measure) (sample vs. population) the **median** which is the middle value in a set of ordered data values. It is the value that separates the higher half of a data set from the lower half. The *mode* is another measure of central tendency. The **mode** for a set of data is the value that occurs most frequently in the set. The **midrange** can also be used to assess the central tendency of a numeric data set. It is the average of the largest and smallest values in the set.
		2. **Measuring the Dispersion of Data: Range, Quartiles, Variance, Standard Deviation, and Interquartile Range:** We now look at measures to assess the dispersion or spread of numeric data. The **range** of the set is the difference between the largest (max ()) and smallest (min ()) values. **Quantiles** are points taken at regular intervals of a data distribution, dividing it into essentially equal size consecutive sets. The distance between the first and third quartiles is **interquartile range** (**IQR**). The **five-number summary** of a distribution consists of the median (*Q*2), the quartiles *Q*1 and *Q*3. **Boxplots** are a popular way of visualizing a distribution. Variance and standard deviation are measures of data dispersion. They indicate how spread out a data distribution is. A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.
		3. **Graphic Displays of Basic Statistical Descriptions of Data**: A **quantile plot** is a simple and effective way to have a first look at a univariate data distribution. A **quantile–quantile plot**, or **q-q plot**, graphs the quantiles of one univariate distribution against the corresponding quantiles of another. **Histograms** (or **frequency histograms**) are at least a century old and are widely used. A **scatter plot** is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes.
	3. **Data Visualization: Data visualization** aims to communicate data clearly and effectively through graphical representation.
		1. **Pixel-Oriented Visualization Techniques:** A simple way to visualize the value of a dimension is to use a pixel where the color of the pixel reflects the dimension’s value. A drawback of pixel-oriented visualization techniques is that they cannot help us much in understanding the distribution of data in a multidimensional space.
		2. **Geometric Projection Visualization Techniques:** **Geometric projection techniques** help users find interesting projections of multidimensional data sets.
		3. **Icon-Based Visualization Techniques:** use small icons to represent multidimensional data values two popular icon-based techniques: *Chernoff faces* and *stick* *figures*.
		4. **Hierarchical Visualization Techniques:** The visualization techniques discussed so far focus on visualizing multiple dimensions simultaneously.
		5. **Visualizing Complex Data and Relations:** non-numeric data, such as text and social networks, have become available. Visualizing and analyzing such data attracts a lot of interest. Like a **tag cloud** is a visualization of statistics of user-generated tags.
3. **Data Preprocessing**:
	1. **Data Preprocessing: An Overview:**
		1. **Data Quality: Why Preprocess the Data?** There are many factors comprising **data quality**, including **accuracy:** correct or wrong, accurate or not, **completeness:** not recorded, unavailable, **consistency:** some modified but some not, dangling, **timeliness:** timely update, **believability:** how trustable the data are correct, and **interpretability:** how easily the data can be understood.
		2. **Major Tasks in Data Preprocessing: Data cleaning** routines work to “clean” the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. **Data integration** when integrating multiple databases, data cubes, or files given concept may have different names in different databases, causing inconsistencies and redundancies. **Data reduction** obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results. Data reduction strategies include *dimensionality reduction* and *numerosity reduction*. **Dimensionality reduction**, data encoding schemes are applied in **numerosity reduction**, the data are replaced by alternative. Allow data mining at multiple abstraction levels. Normalization, data discretization, and concept hierarchy generation are forms of **data transformation**.
	2. **Data Cleaning**
		1. **Missing Values**: solution methods.1- Ignore the tuple. 2- Fill in the missing value manually. 3- Use a global constant to fill in the missing value. 4- Use a measure of central tendency for the attribute to fill in the missing value 5- Use the attribute mean or median for all samples belonging to the same class as the given tuple. 6- Use the most probable value to fill in the missing value.
		2. **Noisy Data: Noise** is a random error or variance in a measured variable. Data smoothing techniques. **Binning:** Binning methods smooth a sorted data value by consulting its “neighborhood “that is, the values around it. **Regression:** Data smoothing can also be done by regression, a technique that conforms data values to a function. **Outlier analysis**: Outliers may be detected by clustering,
		3. **Data Cleaning as a Process**: 1- Data discrepancy detection: Use metadata, Check field overloading, Check uniqueness rule, consecutive rule and null rule, Use commercial tools, Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections or Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers) 2- Data migration and integration: Data migration tools: allow transformations to be specified, ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface.
4. **Chapter 4:** Data warehouses generalize and consolidate data in multidimensional space. The construction of data warehouses involves data cleaning, data integration, and data transformation, and can be viewed as an important preprocessing step for data mining. Moreover, data warehouses provide online analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities.
	1. **Data Warehouse: Basic Concepts.**
		1. **What Is a Data Warehouse?** Data repository that is maintained separately from an organization’s operational databases. Data warehouse systems allow for integration of a variety of application systems. They support information processing by providing a solid platform of consolidated historic data for analysis. Data warehouse have key features, its **Subject-oriented:** A data warehouse is organized around major subjects not day-to-day operations its focuses on the modeling and analysis of data for decision makers , its simple particular subject issues by excluding data that are not useful in the decision support process. **Integrated**: integrating multiple heterogeneous sources data cleaning and data integration techniques are applied to ensure consistency. **Time-variant**: Data are stored to provide information from an historic perspective. **Nonvolatile**: A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. The traditional database approach to heterogeneous database integration is to build **wrappers** and **integrators** (or **mediators**) on top of multiple, heterogeneous databases.
		2. **Differences between Operational Database Systems and Data Warehouses: online transaction processing (OLTP)** systems. They cover most of the day-to-day operations of an organization such as purchasing **online analytical processing (OLAP)** systems serve users or knowledge workers in the role of data analysis and decision making. **Users and system orientation**: An OLTP system is *customer-oriented* and OLAP system is *market-oriented* **Data contents**: An OLTP system manages current data that, typically, are too detailed and OLAP system manages large amounts of historic data, provides facilities for summarization and aggregation **Database design**: An OLTP system usually adopts an entity-relationship (ER) data OLAP system typically adopts either a *star* or a *snowflake* model and a subject-oriented
		3. **But, Why Have a Separate Data Warehouse?** Because operational databases store huge amounts of data so to help promote the *high performance of both systems* we Separate it
		4. **Data Warehousing: A Multitiered Architecture:** Data warehouses often adopt a three-tier architecture.
		5. **Data Warehouse Models: Enterprise Warehouse, Data Mart, and Virtual Warehouse**: there are three data warehouse models **Enterprise warehouse:** An enterprise warehouse collects all of the information about subjects spanning the entire organization. **Data mart:** A data mart contains a subset of corporate-wide data that is of value to a specific group of users. **Virtual warehouse:** A virtual warehouse is a set of views over operational databases.
		6. **Extraction, Transformation, and Loading:** Data warehouse systems use back-end tools and utilities to populate and refresh their data. **Data extraction**, which typically gathers data from multiple, heterogeneous, and external sources **Data cleaning**, which detects errors in the data and rectifies them when possible. **Data transformation**, which converts data from legacy or host format to warehouse format. **Load**, which sorts, summarizes, consolidates, computes views, checks integrity, and builds indices and partitions. **Refresh**, which propagates the updates from the data sources to the warehouse.
		7. **Metadata Repository: Metadata** are data about data. When used in a data warehouse, metadata are the data that define warehouse objects.
	2. **Data Warehouse Modeling: Data Cube and OLAP**: Data warehouses and OLAP tools are based on a **multidimensional data model**.
		1. **Data Cube: A Multidimensional Data Model: data cube** allows data to be modeled and viewed in multiple dimensions. **Dimensions** are the perspectives or entities with respect to which an organization wants to keep records. **Facts** are numeric measures. The **fact table** contains the names of the *facts*, or measures, as well as keys to each of the related dimension tables. The cuboid that holds the lowest level of summarization is called the **base cuboid**. The 0-D cuboid, which holds the highest level of summarization, is called the **apex cuboid**.
		2. **Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Data Models:** The most popular data model for a data warehouse is a **multidimensional model** which can exist in the form of a **star schema**, a **snowflake schema**, or a **fact constellation schema**. Let’s look at each of these. **Star schema:** The most common modeling paradigm is the star schema. **Snowflake schema:** The snowflake schema is a variant of the star schema model, where some dimension tables are *normalized*, **Fact constellation:** Sophisticated applications may require multiple fact tables to *share* dimension tables.
		3. **Dimensions: The Role of Concept Hierarchies:** A **concept hierarchy** defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts.
		4. **Measures: Their Categorization and Computation: Distributive:** An aggregate function is *distributive* if it can be computed in a distributed manner as follows. **Algebraic:** An aggregate function is *algebraic* if it can be computed by an algebraic function **Holistic:** An aggregate function is *holistic* if there is no constant bound on the storage size needed to describe a subaggregate.
		5. **Typical OLAP Operations**: **Roll-up:** The roll-up operation (also called the *drill-up* operation by some vendors) performs aggregation on a data cube **Drill-down:** Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. **Slice and dice:** The *slice* operation performs a selection on one dimension of the given cube, resulting in a sub cube. **Pivot (rotate):** *Pivot* (also called *rotate*) is a visualization operation that rotates the data axes in view to provide an alternative data presentation.
	3. **Data Warehouse Design and Usage:** Four views regarding the design of a data warehouse **Top-down view:** allows selection of the relevant information necessary for the data warehouse **Data source view:** exposes the information being captured, stored, and managed by operational systems **Data warehouse view:** consists of fact tables and dimension tables **Business query view:** sees the perspectives of data in the warehouse from the view of end-user. There are three kinds of data warehouse applications: *information processing analytical processing*, and *data mining* **Information processing** supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts, or graphs. **Analytical processing** supports basic OLAP operations, including slice-and-dice drill-down, roll-up, and pivoting. **Data mining** supports knowledge discovery by finding hidden patterns and associations.
5. **Chapter** 5:
	1. **Data Cube Computation: Preliminary Concepts:** Data cubes facilitate the online analytical processing of multidimensional data.
		1. **Cube Materialization: Full Cube, Iceberg Cube, Closed Cube, and Cube Shell:** To ensure fast OLAP, it is sometimes desirable to precompute the **full cube** .An Iceberg-Cube contains only those cells of the data cube that meet an aggregate condition. It is called an Iceberg-Cube because it contains only some of the cells of the full cube. A **closed cube** is a data cube consisting of only closed cells. Another strategy for partial materialization is to precompute only the cuboids involving a small number of dimensions such as three to five. These cuboids form a **cube shell** for the corresponding data cube.
		2. **General Strategies for Data Cube Computation:** There are several methods for efficient data cube computation, **Optimization Technique 1: Sorting, hashing, and grouping Optimization Technique 2: Simultaneous aggregation and caching of intermediate results Optimization Technique 3: Aggregation from the smallest child when there exist multiple child cuboids Optimization Technique 4: The Apriori pruning method can be explored to compute iceberg cubes efficiently.**
	2. **Data Cube Computation Methods:** Data cube computation is an essential task in data warehouse implementation. The precomputation of all or part of a data cube can greatly reduce the response time and enhance the performance of online analytical processing. There are some efficient methods for data cube computation.
		1. **Multiway Array Aggregation for Full Cube Computation:** The **multiway array aggregation** (or simply **MultiWay**) method computes a full data cube by using a multidimensional array as its basic data structure. It is a typical MOLAP approach that uses direct array addressing,
		2. **BUC: Computing Iceberg Cubes from the Apex Cuboid Downward: BUC** is an algorithm for the computation of sparse and iceberg cubes. Unlike Multiway BUC constructs the cube from the apex cuboid toward the base cuboid.
		3. **Star-Cubing: Computing Iceberg Cubes Using a Dynamic Star-Tree Structure:** It integrates top-down and bottom-up cube computation and explores both multidimensional aggregation (similar to MultiWay) and Apriori-like pruning (similar to BUC).
		4. **Precomputing Shell Fragments for Fast High-Dimensional OLAP:** Recall the reason that we are interested in precomputing data cubes: Data cubes facilitate fast OLAP in a multidimensional data space.
	3. **Processing Advanced Kinds of Queries by Exploring Cube Technology**: The methods is for develop data cube technology for effective processing of advanced kinds of queries.
		1. **Sampling Cubes: OLAP-Based Mining on Sampling Data:** When collecting data, we often collect only a subset of the data we would ideally like to gather. In statistics, this is known as collecting a sample of the data populationthe resulting data are called sample data. Data are often sampled to save on costs, manpower, time, and materials.
		2. **Ranking Cubes: Efficient Computation of Top-*k* Queries:** The data cube helps not only online analytical processing of multidimensional queries but also search and data mining, Ranking Cube and examine how it contributes to the efficient processing of top-k queries. Instead of returning a large set of indiscriminative answers to a query, a top-k query (or ranking query) returns only the best k results according to a user-specified preference. The results are returned in ranked order so that the best is at the top. The user specified preference generally consists of two components: a selection condition and a ranking function.
	4. **Multidimensional Data Analysis in Cube Space:** Data cubes create a flexible and powerful means to group and aggregate data subsets. They allow data to be explored in multiple dimensional combinations and at varying aggregate granularities. This capability greatly increases the analysis bandwidth and helps effective discovery of interesting patterns and knowledge from data. The use of cube space makes the data space both meaningful and tractable.
		1. **Prediction Cubes: Prediction Mining in Cube Space:** Recently, researchers have turned their attention toward multidimensional data mining to uncover knowledge at varying dimensional combinations and granularities. Such mining is also known as exploratory multidimensional data mining and online analytical data mining (OLAM). Multidimensional data space is huge. Multidimensional data mining in cube space organizes data of interest into intuitive regions at various granularities. It analyzes and mines the data by applying various data mining techniques systematically over these regions. There are four ways in which OLAP-style analysis can be fused with data mining techniques: **1.** *Use cube space to define the data space for mining*. **2.** *Use OLAP queries to generate features and targets for mining*. **3.** *Use data mining models as building blocks in a multistep mining process*. **4.** *Use data cube computation techniques to speed up repeated model construction*.
		2. **Multifeature Cubes: Complex Aggregation at Multiple Granularities:** Multifeature cubes enable more in-depth analysis. They can compute more complex queries of which the measures depend on groupings of multiple aggregates at varying granularity levels. The queries posed can be much more elaborate and task-specific than traditional queries.
		3. **Exception-Based, Discovery-Driven Cube Space Exploration:** **Exception-based,discovery-driven** exploration of cube space displays visual cues to indicate discovered data exceptions at all aggregation levels, thereby guiding the user in the data analysis process.
6. **Chapter 6: Frequent patterns** are patterns (e.g., itemsets, subsequences, or substructures) that appear frequently in a data set.
	1. **Basic Concepts**: Frequent pattern mining searches for recurring relationships in a given data set**.**
		1. **Market Basket Analysis: A Motivating Example:** Frequent itemset mining leads to the discovery of associations and correlations among items in large transactional or relational data sets. A typical example of frequent itemset mining is market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customers place in their “shopping baskets” the patterns can be represented in the form of association rules. customers who purchase computers also tend to buy antivirus software at the same time is represented in the following association rule:

(computer)*antivirus software* [*support* D 2%,*confidence* D 60%]. (6.1).

Rule **support** and **confidence** are two measures of rule interestingness. A support of 2% for Rule (6.1) means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the software. Typically, association rules are considered interesting if they satisfy both a **minimum support threshold** and a **minimum confidence threshold**. These thresholds can be a set by users or domain experts.

* + 1. **Frequent Itemsets, Closed Itemsets, and Association Rules: Association rule mining** consists of first finding **frequent itemsets and** satisfying a *minimum support threshold* in from which **strong** association rules in the form of *A 🡪 B* These rules also satisfy a *minimum confidence threshold* *Associations can be further analyzed to uncover correlation rules.*
	1. **Frequent Itemset Mining Methods:** methods for mining the simplest form of frequent patterns.
		1. **Apriori Algorithm: Finding Frequent Itemsets by Confined Candidate Generation:** Apriori is a seminal algorithm proposed for mining frequent itemsets for Boolean association rules The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties, To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property is used to reduce the search space. *“How is the Apriori property used in the algorithm?”* *consisting of join and prune actions.* *This subset testing can be done quickly by maintaining a hash tree of all frequent itemsets.*
		2. **Generating Association Rules from Frequent Itemsets:** Once the frequent itemsets from transactions in a database it is straightforward to generate strong association rules from them (where strong association rules satisfy both minimum support and minimum confidence).
		3. **Improving the Efficiency of Apriori: Hash-based technique** (hashing itemsets into corresponding buckets): A hash-based technique can be used to reduce the size of the candidate. **Transaction reduction** (reducing the number of transactions scanned in future iterations. **Partitioning** (partitioning the data to find candidate itemsets): A partitioning technique can be used that requires just two database scans to mine the frequent itemsets. **Sampling** (mining on a subset of the given data): The basic idea of the sampling approach is to pick a random sample S of the given data D, and then search for frequent itemsets in S instead of D. **Dynamic itemset counting** (adding candidate itemsets at different points during a scan)
		4. **A Pattern-Growth Approach for Mining Frequent Itemsets:** Apriori can suffer from two nontrivial costs: *It may still need to generate a huge number of candidate sets It may need to repeatedly scan the whole database and check a large set of candidates* by *pattern matching.* An interesting method in this attempt is called **frequent pattern growth,** or simply **FP-growth**, which adopts a *divide-and-conquer* strategy as follows. First, it compresses the database representing frequent items into a **frequent pattern tree,** or **FP-tree**, which retains the itemset association information.
		5. **Mining Frequent Itemsets Using the Vertical Data Format:** Both the Apriori and FP-growth methods mine frequent patterns from a set of transactions in TID-itemset format (i.e., fTID: itemsetg), where TID is a transaction ID and itemset is the set of items bought in transaction TID. This is known as the **horizontal data format**. Alternatively, data can be presented in item-TID set format (i.e., fitem : TID set, where item is an item name, and TID set is the set of transaction identifiers containing the item. This is known as the **vertical data format**
		6. **Mining Closed and Max Patterns**: A recommended methodology is to search for closed frequent itemsets directly during the mining process. This requires us to prune the search space as soon as we can identify the case of closed itemsets during mining. Pruning strategies include the following: **Item merging** and **Sub-itemset pruning** and **Item skipping**.
	2. **Which Patterns Are Interesting?—Pattern Evaluation Methods:** This has been a major bottleneck for successful application of association rule mining.
		1. **Strong Rules Are Not Necessarily Interesting:** Whether or not a rule is interesting can be assessed either subjectively or objectively. Ultimately, only the user can judge if a given rule is interesting, and this judgment, being subjective, may differ from one user to another.
		2. **From Association Analysis to Correlation Analysis:** the support–confidence framework should be augmented with a pattern evaluation measure, which promotes the mining of interesting rules.
		3. **A Comparison of Pattern Evaluation Measure:** A measure is **null-invariant** if its value is free from the influence of **null-transactions** (i.e., the *transactions that do not contain any of the itemsets being examined*). Among many pattern evaluation measures, we examined ***lift***, \_2***, all confidence, max confidence, Kaczynski***, and ***cosine***, and showed.
1. **Chapter** **7:**
	1. **Pattern Mining: A Road Map:** *it is important to lay out a clear road map to help us get an organized picture of the field and to select the best methods for pattern mining applications. Outlines a general road map on pattern mining research. Most studies mainly address three pattern mining aspects: the kinds of patterns mined, mining methodologies, and applications*
	2. **Pattern Mining in Multilevel, Multidimensional Space:**
		1. **Mining Multilevel Associations**: Association rules generated from mining data at multiple abstraction levels are called **multiple-level** or **multilevel association rules**. Multilevel association rules can be mined efficiently using concept hierarchies under a support-confidence framework.

**Using uniform minimum support for all levels** (referred to as **uniform support**): The same minimum support threshold is used when mining at each abstraction level. When a uniform minimum support threshold is used, the search procedure is simplified. The uniform support approach, however, has some drawbacks. It is unlikely that items at lower abstraction levels will occur as frequently as those at higher abstraction levels.

**Using reduced minimum support at lower levels** (referred to as **reduced support**): Each abstraction level has its own minimum support threshold. The deeper the abstraction level, the smaller the corresponding threshold.

**Using item or group-based minimum support (referred to as group-based support):** **Because users or experts often have insight as to which groups are more important than others.**

* + 1. **Mining Multidimensional Associations:**

*(buys*.*X*, “*digital camera*”/)*buys*.*X*, “*HP printer*”). (7.6) we can refer to Rule as a **single dimensional** or **interdimensional association rule** because it contains a single distinct predicate.*(age*.*X*, “20: : :29”/^*occupation*.*X*, “*student*”/)*buys*.*X*, “*laptop*”/). (7.7) Association rules that involve two or more dimensions or predicates can be referred to as **multidimensional association rules**. Has **no repeated predicates**. Multidimensional association rules with no repeated predicates are called **interdimensional association rules**. We can also mine multidimensional association rules with repeated predicates, which contain multiple occurrences of some predicates. These rules are called **hybrid-dimensional association rules**. **Categorical Attributes:** *finite number of possible values, no ordering among values—data cube approach* **Quantitative Attributes:** *Numeric, implicit ordering among Values discretization, clustering, and gradient approaches*

* + 1. **Mining Quantitative Association Rules***:* *three methods that can help overcome this difficulty to discover novel association relationships: (1) a data cube method, (2) a clustering-based method, and (3) a statistical analysis method to uncover exceptional behaviors.*
		2. **Mining Rare Patterns and Negative Patterns:** **Rare patterns:** *Very low support but interesting E.g., buying Rolex watches we can mine by setting individual-based or special group-based support threshold for valuable items.* **Negative patterns:** *Since it is unlikely that one buys Ford Expedition and Toyota Prius together, Ford Expedition and Toyota Prius are likely negatively correlated patterns* *negatively correlated patterns that are infrequent tend to be more interesting than those that are frequent*
	1. **Constraint-Based Frequent Pattern Mining:** good heuristic is to have the users specify such intuition or expectations as *constraints* to confine the search space. This strategy is known as **constraint-based mining**. The constraints can include the following: **Knowledge type constraints:** These specify the type of knowledge to be mined, such as association, correlation, classification, or clustering. **Data constraints:** These specify the set of task-relevant data **Dimension/level constraints:** These specify the desired dimensions (or attributes) of the data, the abstraction levels, or the level of the concept hierarchies to be used in mining. **Interestingness constraints:** These specify thresholds on statistical measures of rule interestingness such as support, confidence, and correlation. **Rule constraints:** These specify the form of, or conditions on, the rules to be mined.
		1. **Metarule-Guided Mining of Association Rules:** *Metarules allow users to specify the syntactic form of rules that they are interested in mining. The rule forms can be used as constraints to help improve the efficiency of the mining process. Metarules may be based on the analyst’s experience, expectations, or intuition regarding the data or may be automatically generated based on the database schema.*
		2. **Constraint-Based Pattern Generation:** *Pruning Pattern Space and Pruning Data Space: rule constraints, how they can be used to make the mining process more efficient, Dimension/level constraints and interestingness constraints can be applied after mining to filter out discovered rules, although it is generally more efficient and less expensive to use them during mining to help prune the search space* *In general, an efficient frequent pattern mining processor can prune its search space during mining in two major ways***: pruning pattern search space and pruning data search space.**
	2. **Mining High-Dimensional Data and Colossal Patterns:** mine high-dimensional data? Researchers have overcome this difficulty in two directions. One direction extends a pattern growth approach by further exploring the vertical data format to handle data sets with a large number of dimensions (also called features or items, e.g., genes) but a small number of rows (also called transactions or tuples, e.g., samples). The second direction. We introduce **Pattern- Fusion**, a new mining methodology that mines colossal patterns (i.e., patterns of very long length). This method takes leaps in the pattern search space, leading to a good approximation of the complete set of colossal frequent patterns.
		1. **Mining Colossal Patterns by Pattern-Fusion:** A new mining strategy called Pattern-Fusion was developed, which fuses a small number of shorter frequent patterns into colossal pattern candidates. It thereby takes leaps in the pattern search space and avoids the pitfalls of both breadth-first and depth first searches. This method finds a good approximation to the complete set of colossal frequent patterns. The Pattern-Fusion method has the following major characteristics. First, it traverses the tree in a bounded-breadth way. Second, Pattern-Fusion has the capability to identify “shortcuts” whenever possible. the Pattern-Fusion method is outlined in the following two phases: **1. Initial Pool:** Pattern-Fusion assumes an initial pool of small frequent patterns is available. **2. Iterative Pattern-Fusion:** Pattern-Fusion takes as input a user-specified parameter *K*, which is the maximum number of patterns to be mined.
	3. **Mining Compressed or Approximate Patterns:** A major challenge in frequent pattern mining is the huge number of discovered patterns. Using a minimum support threshold to control the number of patterns found has limited effect. Too low a value can lead to the generation of an explosive number of output patterns, while too high a value can lead to the discovery of only commonsense patterns. To reduce the huge set of frequent patterns generated in mining while maintaining high-quality patterns, we can instead mine a compressed or approximate set of frequent patterns.
		1. **Mining Compressed Patterns by Pattern Clustering:** Clustering is the automatic process of grouping like objects together, so that objects within a cluster are similar to one another and dissimilar to objects in other clusters. In this case, the objects are frequent patterns. The frequent patterns are clustered using a tightness measure.
		2. **Extracting Redundancy-Aware Top-*k* Patterns:** **Mining the top-k most frequent patterns is a strategy for reducing the number of patterns returned during mining** A small set of representative patterns that have not only high significance but also low redundancy are called **redundancy-aware top-*k* patterns**.
	4. **Pattern Exploration and Application:**
		1. **Semantic Annotation of Frequent Patterns:** Semantic annotations can be generated to help users understand the meaning of the frequent patterns found, such as for textual terms like “frequent, pattern.” These are dictionary-like annotations, providing semantic information relating to the term. This information consists of context indicators (e.g., terms indicating the context of that pattern), the most representative data transactions (e.g., fragments or sentences containing the term), and the most semantically similar patterns (e.g., “maximal, pattern” is semantically similar to “frequent, pattern”). The annotations provide a view of the pattern’s context from different angles, which aids in their understanding.
		2. **Applications of Pattern Mining**: Frequent pattern mining has many diverse applications, ranging from pattern-based data cleaning to pattern-based classification, clustering, and outlier or exception analysis. These methods are discussed in the subsequent chapters in this book.