DATA MINING AND BIG DATA ANALYTICS (18MCA52C)

UNIT I: Introduction

FACULTY

Dr. K. ARTHI MCA, M.Phil., Ph.D.,

Assistant Professor,

Postgraduate Department of Computer Applications,

Government Arts College (Autonomous),

Coimbatore-641018.

DATA MINING AND BIG DATA ANALYTICS (18MCA52C)

SYLLABUS

UNIT I: Introduction - Data Mining - Relational Databases - Data Warehouses - Transactional databases - Data Mining functionalities - Classification of Data Mining systems - Major Issues in Data Mining.

UNIT II: Data Pre-processing - Data cleaning - Missing value, noising data and inconsistent data - Data integration and Transformation - Data reduction - Data cube aggregation - Dimensionality reduction and data compression - Data mining primitives.

UNIT III: Classification and predictions - Issues regarding classification and prediction - Classifications by decision tree induction - Classification by Back propagation - Other classification methods.

UNIT IV: Cluster Analysis - Types of Data in Cluster Analysis - Interval - Scaled variables, Binary variables, Nominal ordinal and ratio - scaled variables - Clustering methods - Partitioning methods - Kmeans, k-medoids and CLARANS - Hierarchical methods - Agglomerative andDivisive, BIRCH, CURE - Outlier analysis - Data Mining applications.

UNIT V: The Big Deal about Big Data: What is Big Data - Why Is Big data important - Big Data. Applying Big Data to Business problems: A sampling of use cases - Big Data use cases - IT for IT - Customer state. Analytics for Big Data at Rest: The Big Data platform for high performance deep analytics- Appliance simplicity - Hardware Acceleration-Balance, massively parallel architecture - Modular design.

TEXT BOOKS: 1. Jinweihan, Micheline Kambler, “Data Mining: Concepts and Techniques”, Morgan Kaufman Publishers, New Delhi. (For Unit I, II, III and IV). 2. Paul C Zikopoulos, Dirk deRoos, Krishnan Parasuraman, Thomas Deutsch, David Corrigan, James Giles, “Harness the Power of Big Data”, The McGraw-Hill Publications, 2013, First Edition. (For Unit V).

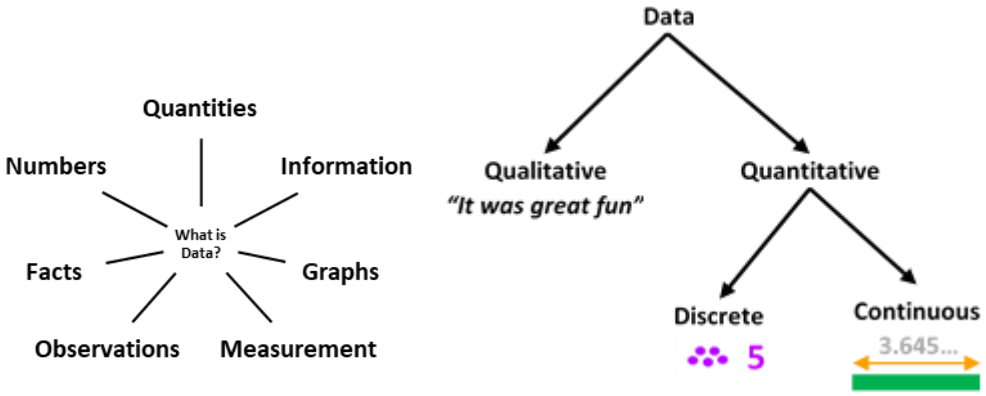
REFERENCE BOOKS: 1. Pieter Adriaans, DolfZantinge, "Data Mining", Addison Wesley, 1998. 2. Sam Anohory, Dennis Murrey, "Dataware housing in the real world", Pearson, 2004.

UNIT - I

DATA

Datum - Data

Data is a collection of facts, such as numbers, words, measurements, observations or just descriptions of things.



1.1. INTRODUCTION

- Searching for knowledge (interesting patterns) in data.
- Knowledge mining from data (does not refer to mining from large data).
- Knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.
- Knowledge discovery from data, or KDD.

1.2. DATA MINING

- Data mining is the process of discovering interesting patterns and knowledge from large amounts of data.
- The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.
- The most basic forms of data for mining applications are database data, data warehouse data and transactional data

- For example, data mining systems can analyze customer data to predict the credit risk of new customers based on their income, age, and previous credit information.

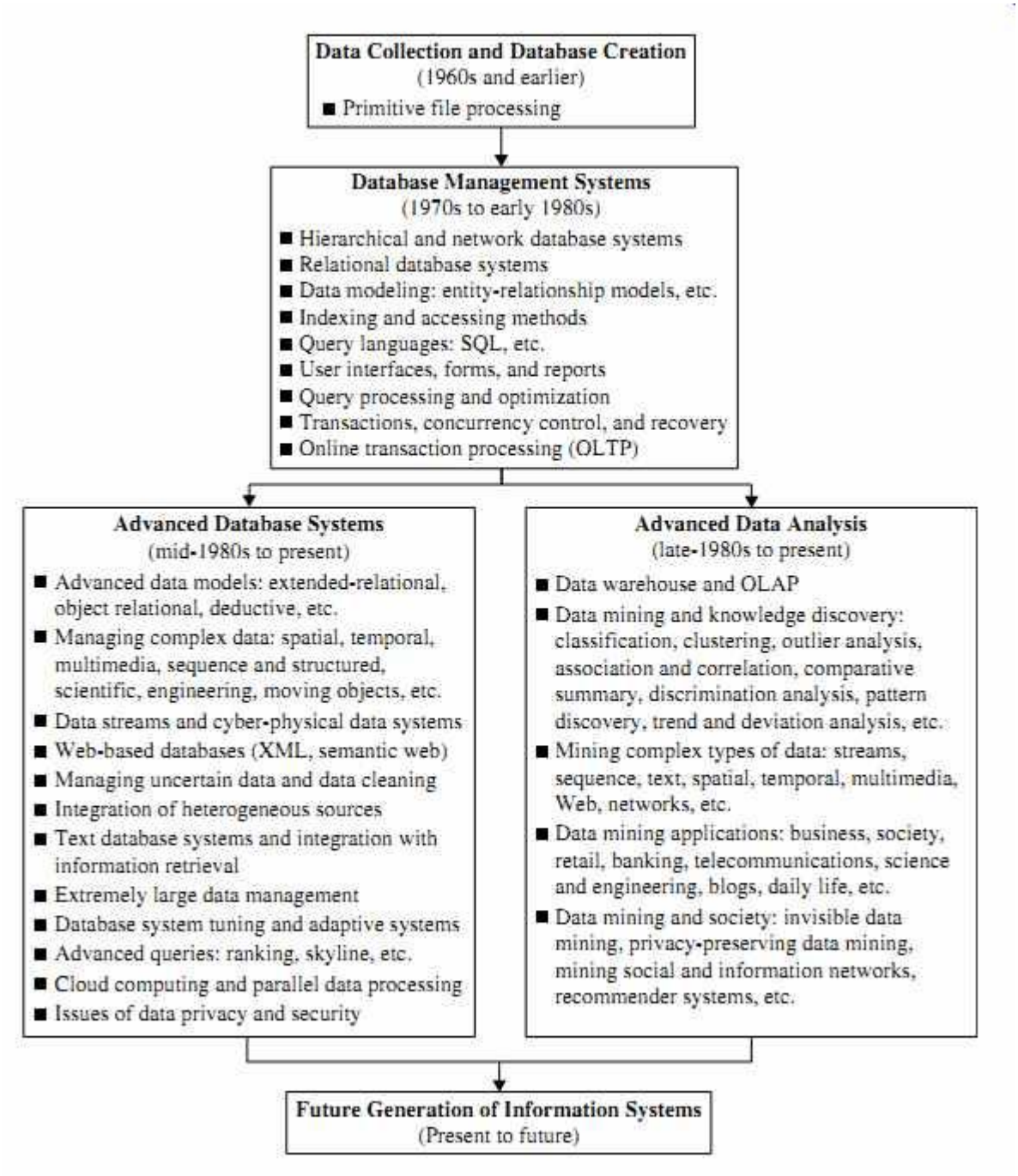
1.3. DATABASE DATA

- A database system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.
- The software programs provide mechanisms for defining database structures and data storage; for specifying and managing concurrent, shared, or distributed data access.
- A relational database is a collection of tables, each of which is assigned a unique name.
- Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows).
- Entity-relationship (ER) data model, is constructed for relational databases. An ER data model represents the database as a set of entities and their relationships.
- Relational databases are one of the most commonly available richest information repositories.
- They are a major data form in the study of data mining.

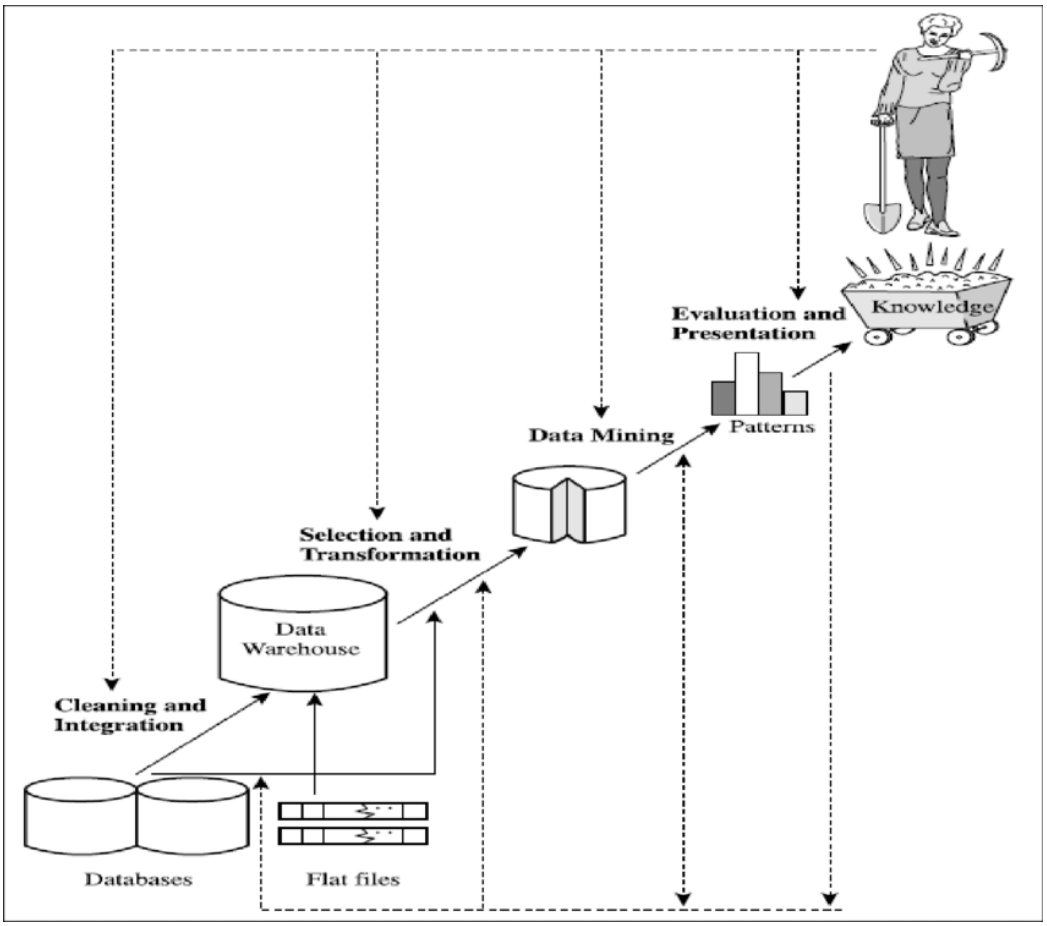
1.4. DATA WAREHOUSES

- Repository of information collected from multiple sources.
- Stored under a unified schema, and residing at a single site.
- Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.
- A data warehouse is a repository for long-term storage of data from multiple sources, organized so as to facilitate management decision making.
- The data are stored under a unified schema and are typically summarized.
- Data warehouse systems provide multidimensional data analysis capabilities, collectively referred to as online analytical processing.
- Data warehouse technology includes data cleaning, data integration, and online analytical processing (OLAP)—that is, analysis techniques with functionalities

such as summarization, consolidation, and aggregation, as well as the ability to view information from different angles.


Data Collection and Database Creation
(1960s and earlier)
■ Primitive file processing
Database Management Systems
(1970s to early 1980s)
■ Hierarchical and network database systems
■ Relational database systems
■ Data modeling: entity-relationship models, etc.
■ Indexing and accessing methods
■ Query languages: SQL, etc.
■ User interfaces, forms, and reports
■ Query processing and optimization
■ Transactions, concurrency control, and recovery
■ Online transaction processing (OLTP)
Advanced Database Systems
(mid-1980s to present)
■ Advanced data models: extended-relational, object relational, deductive, etc.
■ Managing complex data: spatial, temporal, multimedia, sequence and structured, scientific, engineering, moving objects, etc.
■ Data streams and cyber-physical data systems
■ Web-based databases (XML, semantic web)
■ Managing uncertain data and data cleaning
■ Integration of heterogeneous sources
■ Text database systems and integration with information retrieval
■ Extremely large data management
■ Database system tuning and adaptive systems
■ Advanced queries: ranking, skyline, etc.
■ Cloud computing and parallel data processing
■ Issues of data privacy and security
Advanced Data Analysis
(late-1980s to present)
■ Data warehouse and OLAP
■ Data mining and knowledge discovery: classification, clustering, outlier analysis, association and correlation, comparative summary, discrimination analysis, pattern discovery, trend and deviation analysis, etc.
■ Mining complex types of data: streams, sequence, text, spatial, temporal, multimedia, Web, networks, etc.
■ Data mining applications: business, society, retail, banking, telecommunications, science and engineering, blogs, daily life, etc.
■ Data mining and society: invisible data mining, privacy-preserving data mining, mining social and information networks, recommender systems, etc.
Future Generation of Information Systems
(Present to future)

Evolution of database system technology



- Data mining as a step in the process of knowledge discovery
  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 knowledgebased on interestingness measures)
  7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

1.5. TRANSACTIONAL DATABASES

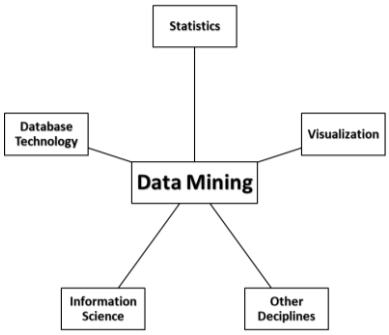
- Transactional database is a DBMS that provides a set of database operation.
- In a database system, a transaction might consist of one or more data-manipulation statements and queries, each reading and/or writing.
- Relational - it means data is organized in tables and recognized like primary key, foreign key, uniqueness, data normalization.
- Transactional database is database able to manipulate data within transactions i.e. information in the database.
- Transactional databases are a collection of data organized by time stamps, date, etc. to represent transaction in databases.

1.6. DATA MINING FUNCTIONALITIES

- Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.
- There are two types of data mining tasks:
  - Descriptive data mining tasks that describe the general properties of the existing data.
  - Predictive data mining tasks that attempt to do predictions based on inference on available data.
    - Predictive mining tasks perform inference on the current data in order to make predictions.

1.7. CLASSIFICATION OF DATA MINING SYSTEMS

- DM is considered as an interdisciplinary field.
- It includes a set of various disciplines such as statistics, database systems, machine learning, visualization and information sciences.
- Classification of the data mining system helps users to understand the system and match their requirements with such systems.



Data mining systems can be categorized as;

- ***Classification according to the application adapted:***
  This involves domain-specific application. For example, for telecommunications, finance, stock markets, e-mails and so on.
- ***Classification according to the type of techniques utilized:***
  This technique involves the degree of user interaction or the technique of data analysis involved. For example, machine learning, visualization, pattern recognition, neural networks, database-oriented or data-warehouse oriented techniques.
- ***Classification according to the types of knowledge mined:***
  This is based on functionalities such as characterization (summarizing data of class), association, discrimination (mapping or classification of a class) and correlation, prediction etc.
- ***Classification according to types of databases mined:***
  A database system can be classified as a 'type of data' or 'use of data' model or 'application of data'.

1.8. MAJOR ISSUES IN DATA MINING

The major issues in data mining research, partitioning them into five groups:

- Mining methodology
- User interaction
- Efficiency and scalability
- Diversity of data types
- Data mining and society

Mining Methodology

- Various aspects of mining methodology
- Mining various and new kinds of knowledge
- Mining knowledge in multidimensional space
- Data mining—an interdisciplinary effort
- Boosting the power of discovery in a networked environment
- Handling uncertainty, noise, or incompleteness of data
- Pattern evaluation and pattern- or constraint-guided mining.

User Interaction

- Interactive mining
- Incorporation of background knowledge
- Ad hoc data mining and data mining query languages
- Presentation and visualization of data mining results

Efficiency and scalability

- Efficiency and scalability of data mining algorithm scalability
- Parallel, distributed, and incremental mining algorithms

Diversity of Database Types

- Handling complex types of data
- Mining dynamic, networked, and global data repositories

Data Mining and Society

- Social impacts of data mining
- Privacy-preserving data mining
- Invisible data mining

1.8.1. Mining Methodology

- Researchers have been vigorously developing new data mining methodologies.
- This involves the investigation of new kinds of knowledge, mining in multidimensional
- space, integrating methods from other disciplines, and the consideration of semantic ties
- among data objects.
- In addition, mining methodologies should consider issues such as data uncertainty, noise, and incompleteness.

Mining various and new kinds of knowledge:

- Data mining covers a wide spectrum of data analysis and knowledge discovery tasks, from data characterization and discrimination to association and correlation analysis, classification, regression, clustering, outlier analysis, sequence analysis, and trend and evolution analysis.
- These tasks may use the same database in different ways and require the development of numerous data mining techniques.

Mining knowledge in multidimensional space:

- When searching for knowledge in large data sets, we can explore the data in multidimensional space. That is, we can search for interesting patterns among combinations of dimensions (attributes) at varying levels of abstraction.
- Such mining is known as (exploratory) multidimensional data mining.
- In many cases, data can be aggregated or viewed as a multidimensional data cube. Mining knowledge in cube space can substantially enhance the power and flexibility of data mining.

Data mining—an interdisciplinary effort:

- The power of data mining can be substantially enhanced by integrating new methods from multiple disciplines.
- For example, to mine data with natural language text, it makes sense to fuse data mining methods with methods of information retrieval and natural language processing.
- As another example, consider the mining of software bugs in large programs. This form of mining, known as bug mining, benefits from the incorporation of software engineeringknowledge into the data mining process.

Boosting the power of discovery in a networked environment:

- Most data objects reside in a linked or interconnected environment, whether it be the Web, database relations, files, or documents.
- Semantic links across multiple data objects can be used to advantage in data mining.
- Knowledge derived in one set of objects can be used to boost the discovery of knowledge in a “related” or semantically linked set of objects.

Handling uncertainty, noise, or incompleteness of data:

- Data often contain noise, errors, exceptions, or uncertainty, or are incomplete.
- Errors and noise may confuse the data mining process, leading to the derivation of erroneous patterns.
- Data cleaning, data preprocessing, outlier detection and removal, and uncertainty reasoning are examples of techniques that need to be integrated with the data mining process.

Pattern evaluation and pattern- or constraint-guided mining:

- Not all the patterns generated by data mining processes are interesting. What makes a pattern interesting may vary from user to user.
- Therefore, techniques are needed to assess the interestingness of discovered patterns based on subjective measures.
- These estimate the value of patterns with respect to a given user class, based on user beliefs or expectations.
- Moreover, by using interestingness measures or user-specified constraints toguide the discovery process, we may generate more interesting patterns and reduce the search space.

1.8.2. User Interaction

- The user plays an important role in the datamining process.
- Interesting areas of research include how to interact with a data mining system, how to incorporate a user's back-ground knowledge in mining, and how to visualize and comprehend data mining results.

Interactive mining:

- The data mining process should be highly interactive.
- Thus, it is important to build flexible user interfaces and an exploratory mining environment, facilitating the user's interaction with the system.
- A user may like to first sample a set of data, explore general characteristics of the data, and estimate potential mining results.
- Interactive mining should allow users to dynamically change the focus of a search, to refine mining requests based on returned results, and to drill, dice, and pivot through the data and knowledge space interactively, dynamically exploring "cube space" while mining.

Incorporation of background knowledge:

- Background knowledge, constraints, rules, and other information regarding the domain under study should be incorporated into the knowledge discovery process.
- Such knowledge can be used for pattern evaluation as well as to guide the search toward interesting patterns.

Ad hoc data mining and data mining query languages:

- Query languages (e.g., SQL) have played an important role in flexible searching because they allow users to pose adhoc queries.
- Similarly, high-level data mining query languages or other high-level flexible user interfaces will give users the freedom to define ad hoc data mining tasks.
- This should facilitate specification of the relevant sets of data for analysis, the domainknowledge, the kinds of knowledge to be mined, and the conditions and constraintsto be enforced on the discovered patterns.
- Optimization of the processing of such flexible mining requests is another promising area of study.

Presentation and visualization of data mining results:

- How can a data mining system present data mining results, vividly and flexibly, so that the discovered knowledge can be easily understood and directly usable by humans?
- This is especially crucial if the data mining process is interactive. It requires the system to adopt expressive knowledge representations, user-friendly interfaces, and visualization techniques.

1.8.3. Efficiency and Scalability

- Efficiency and scalability are always considered when comparing data mining algorithms.
- As data amounts continue to multiply, these two factors are especially critical.

Efficiency and scalability of data mining algorithms:

- Data mining algorithms must be efficient and scalable in order to effectively extract information from huge amounts of data in many data repositories or in dynamic data streams.
- In other words, the running time of a data mining algorithm must be predictable, short, and acceptable by applications.
- Efficiency, scalability, performance, optimization, and the ability to execute in real time are key criteria that drive the development of many new data mining algorithms.

Parallel, distributed, and incremental mining algorithms:

- The humongous size of many data sets, the wide distribution of data, and the computational complexity of some data mining methods are factors that motivate the development of parallel and distributed data intensive mining algorithms.
- Such algorithms first partition the data into “pieces.”
- Each piece is processed, in parallel, by searching for patterns.
- The parallel processes may interact with one another.
- The patterns from each partition are eventually merged.
- Cloud computing and cluster computing, which use computers in a distributed and collaborative way to tackle very large-scale computational tasks.
- In addition, the high cost of some data mining processes and the incremental nature of input promote incremental data mining, which incorporates new data updates without having to mine the entire data “from scratch.”
- Such methods perform knowledge modification incrementally to amend and strengthen what was previously discovered.

1.8.4. Diversity of Database Types

- The wide diversity of database types brings about challenges to data mining. Theseinclude:

Handling complex types of data:

- Diverse applications generate a wide spectrum of new data types, from structured data such as relational and data warehouse data to semi-structured and unstructured data; from stable data repositories to dynamic data streams; from simple data objects to temporal data, biological sequences, sensor data, spatial data, hypertext data, multimedia data, software program code, Web data, and social network data.
- It is unrealistic to expect one data mining system to mine all kinds of data, given thediversity of data types and the different goals of data mining.
- Domain- or application-dedicated data mining systems are being constructed for in-depth mining of specific kinds of data.
- The construction of effective and efficient data mining tools for diverse applications remains a challenging and active area of research.

Mining dynamic, networked, and global data repositories:

- Multiple sources of data are connected by the Internet and various kinds of networks,forming gigantic, distributed, and heterogeneous global information systems and networks.
- The discovery of knowledge from different sources of structured, semi-structured, or unstructured yet interconnected data with diverse data semantics poses great challenges to data mining.
- Mining such gigantic, interconnected information networks may help disclose many more patterns and knowledge in heterogeneous data sets than can be discovered from a small set of isolated data repositories.
- Web mining, multisource data mining, and information network mining have become challenging and fast-evolving data mining fields.

1.8.5. Data Mining and Society

- How does data mining impact society?
- What steps can data mining take to preserve the privacy of individuals?
- Do we use data mining in our daily lives without even knowing that we do?
- These questions raise the following issues:

Social impacts of data mining:

- it is important to study the impact of data mining on society.
- How can we use data mining technology to benefit society?
- How can we guard against its misuse?
- The improper disclosure or use of data and the potential violation of individual privacy and data protection rights are areas of concern that need to be addressed.

Privacy-preserving data mining:

- Data mining will help scientific discovery, business management, economy recovery, and security protection (e.g., the real-time discovery of intruders and cyber-attacks).
- However, it poses the risk of disclosing an individual's personal information.
- Studies on privacy-preserving data publishing and data mining are ongoing.

- The philosophy is to observe data sensitivity and preserve people's privacy while performing successful data mining.

Invisible data mining:

- More and more systems should have data mining functions built within so that people can perform data mining or use data mining results simply by mouse clicking, without any knowledge of data mining algorithms.
- Intelligent search engines and Internet-based stores perform such invisible data mining byincorporating data mining into their components to improve their functionality andperformance.
- For example, when purchasing items online, users may be unaware that the store is likely collecting data on the buying patterns of its customers, which may be used to recommend other items for purchase in the future.

THANK YOU

This content is taken from the text books and reference books prescribed in the syllabus.