Unit 1:Data Mining

Notes by Kapil Shanbhag

What is 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

Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery. The knowledge discovery process is shown in Figure 1.1 as an iterative sequence of the following steps:

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—see Section 1.4.6)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

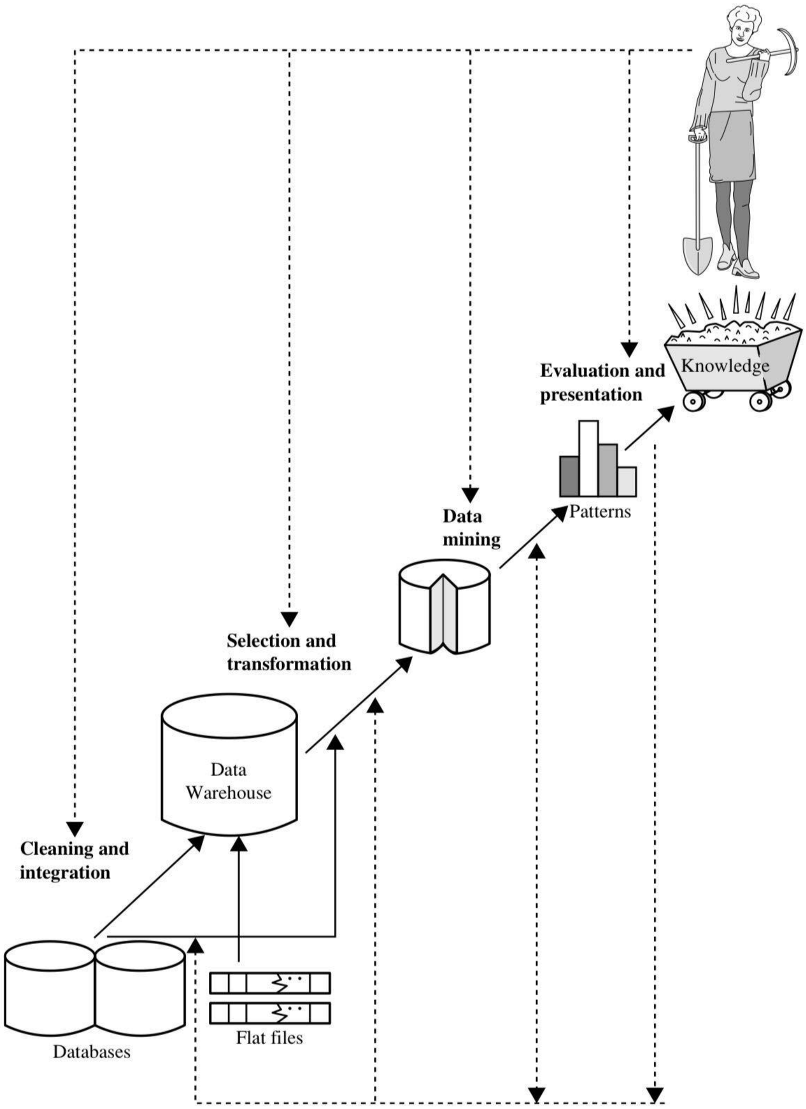


Fig 1.1 Data mining as a step in the process of knowledge discovery.

What kinds of Data can be mined

The most basic forms of data for mining applications are database data , data warehouse data and transactional data. a. Data mining can also be applied to other forms of data (e.g., data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the WWW)

Database Data

- A database system, also called a database management 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; and for ensuring consistency and security of the information stored despite system crashes or attempts at unauthorized 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).
- Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values.
- A semantic data model, such as an entity-relationship (ER) data model, is often constructed for relational databases. An ER data model represents the database as a set of entities and their relationships.
- Relational data can be accessed by database queries written in a relational query language (e.g., SQL) or with the assistance of graphical user interfaces.
- A given query is transformed into a set of relational operations, such as join, selection, and projection, and is then optimized for efficient processing. A query allows retrieval of specified subsets of the data.

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.
- Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.
- To facilitate decision making, the data in a data warehouse are organized around major subjects (e.g., customer, item, supplier, and activity).
- The data are stored to provide information from a historical perspective, such as in the past 6 to 12 months, and are typically summarized.

- A data warehouse is usually modeled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count or sum(sales amount). A data cube provides a multidimensional view of data and allows the precomputation and fast access of summarized data.

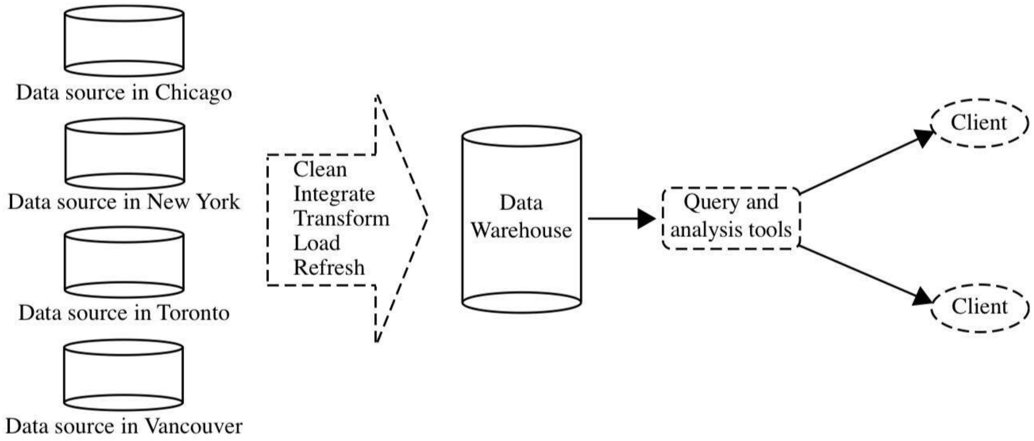
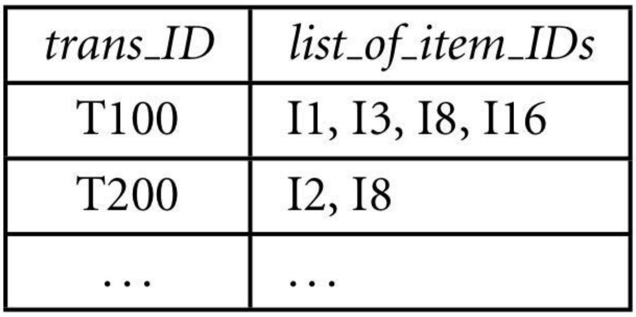


Fig 1.2Typical framework of a data warehouse for All Electronics.

Transactional Data

- In general, each record in a transactional database captures a transaction, such as a customer's purchase, a flight booking, or a user's clicks on a web page.
- A transaction typically includes a unique transaction identity number (trans ID) and a list of the items making up the transaction, such as the items purchased in the transaction
- A transactional database may have additional tables, which contain other information related to the transactions, such as item description, information about the salesperson or the branch, and so on.



trans_ID	*list_of_item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

Fig 1.3 Fragment of a transactional database for sales at All Electronics

Other kinds of Data

Such kinds of data can be seen in many applications: time-related or sequence data data streams , spatial data, engineering design data , hypertext and multimedia data , graph and networked data , and the Web . These applications bring about new challenges, like how to handle data carrying special structures (e.g., sequences, trees, graphs, and networks) and specific semantics (such as ordering, image, audio and video contents, and connectivity), and how to mine patterns that carry rich structures and semantics.

What Kinds of Patterns can be mined

- There are a number of data mining functionalities. These include characterization and discrimination; the mining of frequent patterns, associations, and correlations ; classification and regression ; clustering analysis ; and outlier analysis .
- Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks.
- In general, such tasks can be classified into two categories: descriptive and predictive.
- Descriptive mining tasks characterize properties of the data in a target data set.
- Predictive mining tasks perform induction on the current data in order to make predictions

1) Class/Concept Description: Characterization and Discrimination

- Data entries can be associated with classes or concepts
- It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. Such descriptions of a class or a concept are called class/concept descriptions.
- These descriptions can be derived using (1) data characterization, by summarizing the data of the class under study or (2) data discrimination, by comparison of the target class with one or a set of comparative classes or (3) both data characterization and discrimination
- Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a query.
- There are several methods for effective data summarization and characterization.
- The output of data characterization can be presented in various forms. Examples include pie charts, bar charts, curves, multidimensional data cubes, and multidimensional tables,
- 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.
- The target and contrasting classes can be specified by a user, and the corresponding data objects can be retrieved through database queries
- The forms of output presentation are similar to those for characteristic descriptions, although discrimination descriptions should include comparative measures that help to distinguish between the target and contrasting classes.

Discrimination descriptions expressed in the form of rules are referred to as discriminant rules.

- Read example from text book

2) Mining Frequent Patterns, Associations, and Correlations

- Frequent patterns, as the name suggests, are patterns that occur frequently in data.
- There are many kinds of frequent patterns, including frequent itemsets, frequent subsequences, and frequent substructures. A frequent itemset typically refers to a set of items that often appear together in a transactional data set
- If a substructure occurs frequently, it is called a (frequent) structured pattern. Mining frequent patterns leads to the discovery of interesting associations and correlations within data.
- Typically, association rules are discarded as uninteresting if they do not satisfy both a minimum support threshold and a minimum confidence threshold. Additional analysis can be performed to uncover interesting statistical correlations between associated attribute–value pairs.
- Read Example from text book

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 .The model is used to predict the class label of objects for which the the class label is unknown.
- Whereas classification predicts categorical (discrete, unordered) labels, regression models continuous-valued functions.
- That is, regression is used to predict missing or unavailable numerical data values rather than (discrete) class labels.
- The term prediction refers to both numeric prediction and class label prediction
- Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well
- Regression also encompasses the identification of distribution trends based on the available data.
- Classification and regression may need to be preceded by relevance analysis, which attempts to identify attributes that are significantly relevant to the classification and regression process. Such attributes will be selected for the classification and regression process.
- Read Example from text book

4) Cluster Analysis

- clustering analyses data objects without consulting class labels
- Clustering can be used to generate class labels for a group of data.
- The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity
- Each cluster so formed can be viewed as a class of objects, from which rules can be derived.
- Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.
- Read Example from text book

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.
- The analysis of outlier data is referred to as outlier analysis or anomaly mining.
- Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are remote from any other cluster are considered outliers.
- Read Example from text book

Which are Technologies 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 in terms of random variables and their associated probability distributions.
- Statistical models are widely used to model data and data classes.

2)Machine Learning

- Machine learning investigates how computers can learn (or improve their performance) based on data. A main research area is for computer programs to automatically learn to recognize complex patterns and make intelligent decisions based on data.

3) Database Systems and Data Warehouses

- Database systems research focuses on the creation, maintenance, and use of databases for organizations and end-users.
- Database systems are often well known for their high scalability in processing very large, relatively structured data sets.
- Many data mining tasks need to handle large data sets or even real-time, fast streaming data. Therefore, data mining can make good use of scalable database technologies to achieve high efficiency and scalability on large data sets.

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.
- The typical approaches in information retrieval adopt probabilistic models.
- The document’s language model is the probability density function that generates the bag of words in the document. The similarity between two documents can be measured by the similarity between their corresponding language models
- Furthermore, a topic in a set of text documents can be modeled as a probability distribution over the vocabulary, which is called a topic model.

A classification of data mining systems

Data mining systems can be categorized according to various criteria, as follows.

1) Classification according to the kinds of databases mined.

- A data mining system can be classified according to the kinds of databases mined.
- Database systems themselves can be classified according to different criteria, each of which may require its own data mining technique.
- Data mining systems can therefore be classified accordingly.
- For instance, if classifying according to data models, we may have a relational, transactional, object-oriented, object-relational, or data warehouse mining system.
- If classifying according to the special types of data handled, we may have a spatial, time-series, text, or multimedia data mining system, or a World-Wide Web mining system. Other system types include heterogeneous data mining systems, and legacy data mining systems.

2) Classification according to the kinds of knowledge mined.

- Data mining systems can be categorized according to the kinds of knowledge they mine, i.e., based on data mining functionalities, such as characterization, discrimination, association, classification, clustering, trend and evolution analysis, deviation analysis, similarity analysis, etc.
- A comprehensive data mining system usually provides multiple and/or integrated data mining functionalities.
- Moreover, data mining systems can also be distinguished based on the granularity or levels of abstraction of the knowledge mined, including generalized knowledge , primitive-level knowledge , or knowledge at multiple levels.
- An advanced data mining system should facilitate the discovery of knowledge at multiple levels of abstraction.

3) Classification according to the kinds of techniques utilized.

- Data mining systems can also be categorized according to the underlying data mining techniques employed.
- These techniques can be described according to the degree of user interaction involved or the methods of data analysis employed

- A sophisticated data mining system will often adopt multiple data mining techniques or work out an effective, integrated technique which combines the merits of a few individual approaches.

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, and data mining and society.

1) Mining Methodology

- Mining various and new kinds of knowledge: Data mining covers a wide spectrum of data analysis and knowledge discovery tasks. Due to the diversity of applications, new mining tasks continue to emerge, making data mining a dynamic and fast-growing field
- Mining knowledge in multidimensional space: When searching for knowledge in large data sets, we can explore the data in multidimensional space.
- Data mining—an interdisciplinary effort: The power of data mining can be substantially enhanced by integrating new methods from multiple disciplines.
- Boosting the power of discovery in a networked environment: g. 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
- 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.

2) User Interaction

- 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
- Incorporation of background knowledge: 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: 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. Optimization of the processing of flexible mining requests is another promising area of study
- Presentation and visualization of data mining results: It requires the system to adopt expressive knowledge representations, user-friendly interfaces, and visualization techniques.

3) Efficiency and Scalability

- 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.
- Parallel, distributed, and incremental mining algorithms: Parallel and distributed data-intensive mining 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. 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.

4) Diversity of Database Types

- Handling complex types of data: It is unrealistic to expect one data mining system to mine all kinds of data, given the diversity of data types and the different goals of data mining. 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: Mining 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

5) Data Mining and Society

- Social impacts of data mining: 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. However, it poses the risk of disclosing an individual's personal information

- Invisible data mining: We cannot expect everyone in society to learn and master data mining techniques. 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.