Lecture-2

Data Mining

Introduction

(Architecture & kind of Data used for Data Mining)

Content

- Architecture of a Typical Data Mining System
- Data Mining—On What Kind of Data?
  - Relational databases
  - Data warehouses
  - Transactional databases
  - Advanced database systems
  - World Wide Web.
- Summary

Architecture of a Typical Data Mining System

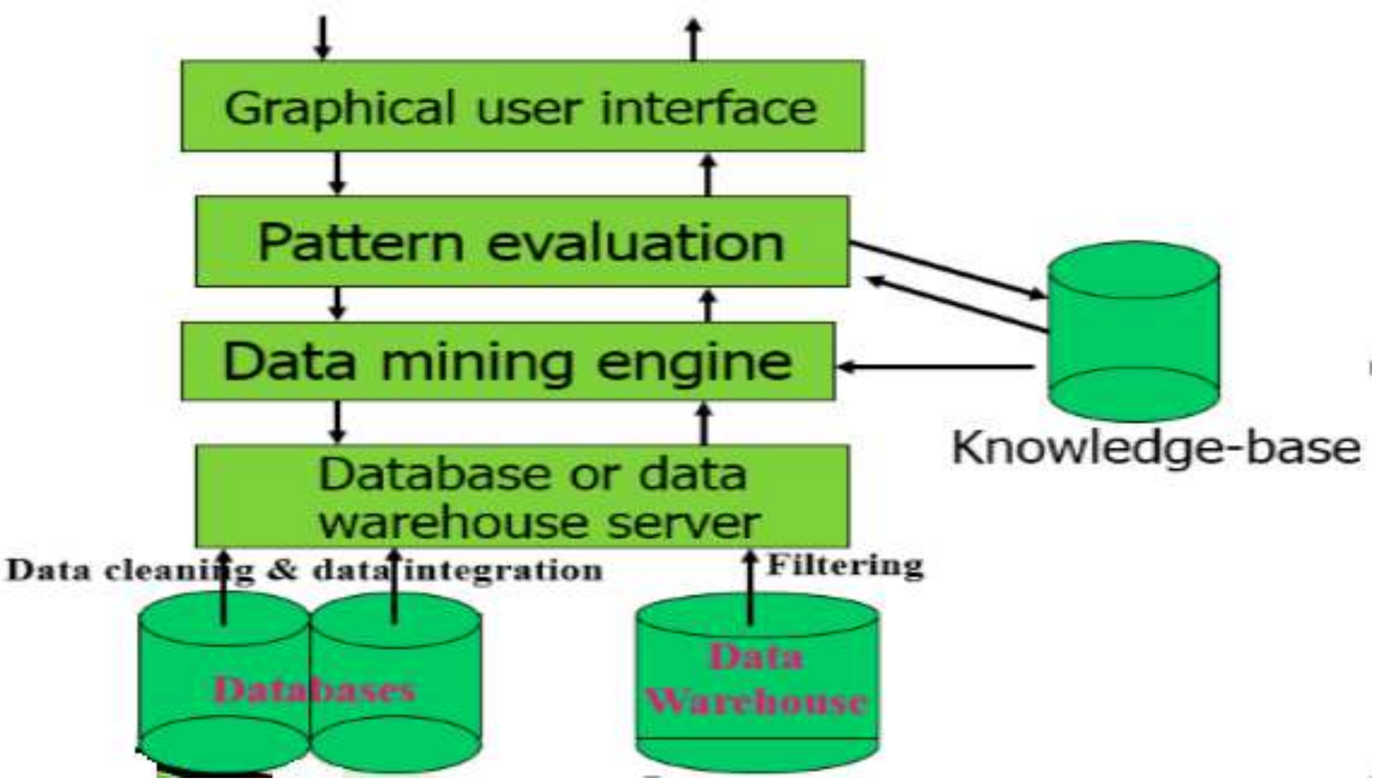


Figure 1. Architecture of Data Mining System

Architecture of a Typical Data Mining System [Cont..]

- The architecture of a typical data mining system may have the following major components:

➢ Database, data warehouse, World Wide Web, or other information repository:

  - Includes databases, data warehouses, spreadsheets, or other kinds of information repositories.
  - Data cleaning and data integration techniques may be performed on the data.

➢ Database or data warehouse server:

  - Responsible for fetching the relevant data, based on the user’s data mining request.

Architecture of a Typical Data Mining System [Cont..]

➢Knowledge base:

- Domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns.
- Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction.
- Knowledge such as user beliefs, which can be used to assess a pattern’s interestingness based on its unexpectedness, may also be included.
- Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata.

Architecture of a Typical Data Mining System [Cont..]

➢Data mining engine:

- This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

Architecture of a Typical Data Mining System [Cont..]

➢Pattern evaluation module:

- Focus the search toward interesting patterns.
- It may use interestingness thresholds to filter out discovered patterns.
- Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used.
- For efficient data mining it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process so as to confine the search to only the interesting patterns.

Architecture of a Typical Data Mining System [Cont..]

➢User interface:

- Communicates between users and the data mining system.
- Allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results.
- Allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

Data Mining—On What Kind of Data?

- Here we will discuss the number of different data repositories on which mining can be performed.
  - The data repositories includes:
    - Relational databases
    - Data warehouses
    - Transactional databases
    - Advanced database systems
    - World Wide Web.

Relational Databases

- A database management system (DBMS), consists of a collection of interrelated data.
- Set of software programs used to manage and access the data from database.
- A relational database is a collection of tables, each of which is assigned a unique name.
- Each table consists of a set of attributes and a large set of tuples.
- Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values.
- The entity-relationship (ER) data model, constructed for relational databases represents the database as a set of entities and their relationships.

DataWarehouses

- A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.
- Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.

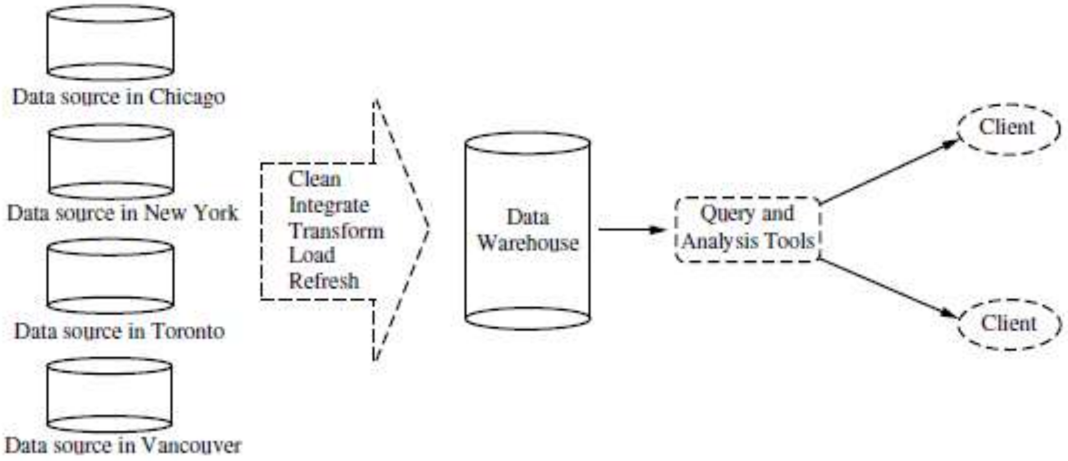


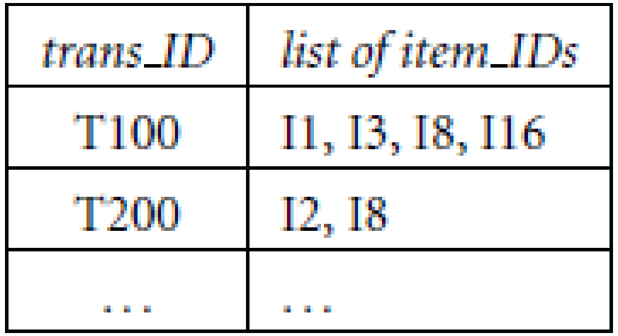
Figure 2. Framework of a data warehouse

Data Warehouses [Cont..]

- To facilitate decision making, the data in a data warehouse are organized around major subjects, such as customer, item, supplier, and activity.
- The data are stored to provide information from a historical perspective and are summarized.
- A data warehouse is usually modeled by a multidimensional database structure, where 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 *sales amount.*
- The actual physical structure of a data warehouse may be a relational data store or a multidimensional data cube.

Transactional Databases

- A transactional database consists of a file where each record represents a transaction.
- A transaction includes a unique transaction identity number (trans ID) and a list of the items making up the transaction.
- The transactional database also stores information regarding the sale, such as the date of the transaction, the customer ID number, the ID number of the salesperson and of the branch at which the sale occurred, and so on.



trans_ID	*list of item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

Advanced Data and Information Systems and Advanced Applications

- It includes:
    - Object-relational databases
    - Spatial Databases and Spatiotemporal Databases
    - Temporal Databases, Sequence Databases, and Time-Series Databases
    - Text databases & Multimedia databases
    - Heterogeneous Databases and Legacy Databases

Object-Relational Databases

- Object-relational databases are constructed based on an object-relational data model.
- Conceptually, the object-relational data model inherits the essential concepts of object-oriented databases, where, in general terms, each entity is considered as an object.
- Each object has associated with it the following:
  - A set of variables
  - A set of messages
  - A set of methods
- Objects that share a common set of properties can be grouped into an object class.

Temporal Databases, Sequence Databases

➢Temporal Databases

- A temporal database typically stores relational data that include time-related attributes.
- These attributes may involve several timestamps, each having different semantics.

➢Sequence Databases

- A sequence database stores sequences of ordered events, with or without a concrete notion of time.
- Examples: customer shopping sequences, Web click streams

Time-Series Databases

- A time-series database stores sequences of values or events obtained over repeated measurements of time (hourly, daily, weekly).
- Example: data collected from the stock exchange, inventory control, and the observation of natural phenomena (like temperature and wind).

Spatial Databases and Spatiotemporal Databases

➢Spatial Databases

- Spatial databases contain spatial-related information.
- Examples include geographic (map) databases, very large-scale integration (VLSI) or computed-aided design databases, and medical and satellite image databases.

➢Spatiotemporal Databases

- A spatial database that stores spatial objects that change with time is called spatiotemporal database.

Text Databases and Multimedia Databases

➢Text Databases

- Text databases are databases that contain word descriptions for objects.
- These word descriptions are usually not simple keywords but rather long sentences or paragraphs, such as product specifications, error or bug reports, warning messages, summary reports, notes, or other documents.

➢Multimedia Databases

- Multimedia databases store image, audio, and video data.
- They are used in applications such as picture content-based retrieval, voice-mail systems, video-on-demand systems, the World Wide Web, and speech-based user interfaces that recognize spoken commands.

Heterogeneous Databases and Legacy Databases

➢Heterogeneous Databases

- A heterogeneous database consists of a set of interconnected, autonomous component databases.
- The components communicate in order to exchange information and answer queries.

➢Legacy Databases

- A legacy database is a group of heterogeneous databases that combines different kinds of data systems, such as relational or object-oriented databases, hierarchical databases, network databases, spreadsheets, multimedia databases, or file systems.

The World Wide Web

- The World Wide Web and its associated distributed information services, such as Yahoo!, Google, America Online, and AltaVista, provide rich, worldwide, on-line information services, where data objects are linked together to facilitate interactive access.
- Users seeking information of interest traverse from one object via links to another.
- Such systems provide the opportunities and challenges for data mining.
- Web community analysis helps identify hidden Web social networks and communities and observe their evolution.
- Web mining is the development of scalable and effective Web data analysis and mining methods.

Summary

- The architecture of a typical data mining system includes a database and/or data warehouse and their appropriate servers, a data mining engine and pattern evaluation module, Knowledge base and a graphical user interface.
- Data patterns can be mined from many different kinds of databases, such as relational databases, data warehouses, and transactional, and object-relational databases.
- Interesting data patterns can also be extracted from other kinds of information repositories, including spatial, time-series, sequence, text, multimedia, and legacy databases and the World Wide Web.

Thank You