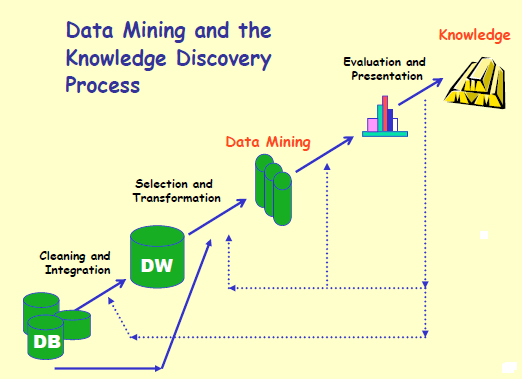
**Unit 1**

**Data Mining Concepts**

Simply stated, data mining refers to *extracting or “mining” knowledge from large amounts of data stored in databases, data warehouses, or other information repositories*. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery. Knowledge discovery consists of an iterative sequence of the following steps:

* **Data cleaning** - It removes noise and inconsistent data
* **Data integration** - This combines data from multiple data sources
* **Data selection** - Data relevant to the analysis task are retrieved from the database
* **Data transformation** - Data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
* **Data mining** - an essential process where intelligent methods are applied in order to extract data patterns
* **Pattern evaluation** - Identifies the truly interesting patterns representing knowledge based on some interestingness measures.
* **Knowledge presentation** - Knowledge representation techniques are used to present the mined knowledge to the user.



**Figure: Knowledge Discovery Process (Stages of KDD)**

According to this view, data mining is only one step in the knowledge discovery process. However, in industry, in media, and in the database research milieu, the term data mining is becoming more popular than the longer term of knowledge discovery from data. Therefore, in this book, we choose to use the term data mining.

Based on this view, the architecture of a typical data mining system may have the following major components.

* **Database, Data Warehouse, World Wide Web, or Other Information Repository**: This is one or a set of 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**: The database or data warehouse server is responsible for fetching the relevant data, based on the user’s data mining request.
* **Knowledge Base**: This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. It is simply stored in the form of set of rules. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction.
* **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.
* **Pattern Evaluation Module**: This component typically employs interestingness measures and interacts with the data mining modules so as to *focus* the search toward interesting patterns. It may use interestingness thresholds to filter out discovered patterns.
* **User interface**: This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task. In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

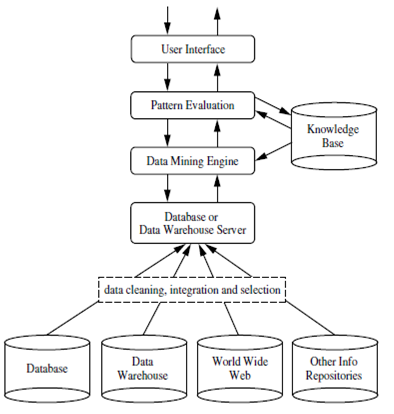


Figure: Architecture of Data Mining System

**Data Warehouse Concepts**

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. To facilitate decision making, the data in a data warehouse are *organized around major subjects*, such as customer, item, supplier, and activity. 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*.

Multidimensional structure is defined as a variation of the relational model that uses multidimensional structures to organize data and express the relationships between data. The structure is broken into cubes and the cubes are able to store and access data within the confines of each cube. "Each cell within a multidimensional structure contains aggregated data related to elements along each of its dimensions.



Figure: Multidimensional Database

By providing multidimensional data views and the pre-computation of summarized data, data warehouse systems are well suited for on-line analytical processing, or OLAP. Examples of OLAP operations include drill-down and roll-up, which allow the user to view the data at differing degrees of summarization. We can drill down on sales data summarized by *quarter* to see the data summarized by *month*. Similarly, we can roll up on sales data summarized by *city* to view the data summarized by *country*.



**Functions of Data Warehousing**

As already mentioned, data Warehousing is a method for gathering and controlling data from different sources making the data easily available for querying and analysis. The following are the functions of data warehouse are discussed below:

* **Data Extraction** - Involves gathering data from multiple heterogeneous sources.
* **Data Cleaning** - Involves finding and correcting the errors in data.
* **Data Transformation** - Involves converting the data from legacy format to warehouse format.
* **Data Loading** - Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.
* **Refreshing** - Involves updating from data sources to warehouse.

**Setting up KDD Environment**

Main goal of KDD is to obtain better understanding of changing organizational environment. It includes a data mining process but using data mining as single generic tool is neither realistic nor desirable. KDD environment needs a suit of data mining tools that needs to be selected and tuned carefully for each organization according to their need. Some of the golden rules that should be followed for setting up KDD environment are discussed below:

* **Support for Extremely Large Data Set**: Data mining deals with billions of records. Thus fast and flexible way of storing and handling such large volume of data is requires along with capacity of storing intermediate results.
* **Support Hybrid Learning:** Leaning can be divided into three categories: Classification, Knowledge engineering, & problem solving. Complex data mining projects needs hybrid learning algorithms that possess capabilities of all three categories.
* **Establish a Data Warehouse**: Data mining mainly depends upon availability and analysis of historic data and hence establishing data warehouse is important for this.
* **Introduce Data Cleaning Facility**: If data stored in databases contains noise, data processing suffers from pollution. Thus data cleaning facility is necessary to avoid such pollution.
* **Facilitate Working with Dynamic Coding**: A KDD Environment should enable the user to experiment with different coding schemes. It must keep the track of genealogy of different samples and tables as well as the semantics and transformations of the different attributes are vital.
* **Beware of False Predictors**: If the results are too good to be true, you probably have found false predictors.
* **Verify Result:** Examine the results carefully and repeat and refine the knowledge discovery process until you are confident.

**Data Mining Functionalities**

In general, data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions. Data mining functionalities, and the kinds of patterns they can discover, are described below.

**Concept/Class Description:** Data Mining can be used to describe individual classes and concepts in data stores. These descriptions can be derived via data characterization or data discrimination. *Data characterization* is a summarization of the general characteristics or features of a target class of data. For example, A data mining system should be able to produce a description summarizing the characteristics of customers who spend more than $1,000 in a year. Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. The target and contrasting classes can be specified by the user. For example, A data mining system should be able to compare two groups of customers, such as those who shop for computer products regularly versus those who rarely shop for such products.

**Mining Frequent Patterns, Associations, and Correlations:** Frequent patterns are patterns that occur frequently in data. Frequent patterns may include frequent itemsets, subsequences, and substructures. A *frequent itemsets* typically refers to a set of items that frequently appear together in a transactional data set, such as milk and bread. A pattern that customers tend to purchase first a PC, followed by a digital camera, and then a memory card, is *pattern* is a frequent subsequence. A different structural form, such as graphs, trees, or lattices, is called a (*frequent*) *structured pattern*. Mining frequent patterns leads to the discovery of interesting associations and correlations within data. For example, a marketing manager of an *Electronics store* would like to determine which items are frequently purchased together within the same transactions. For this mining rule can be

*buys*(*X*; “*computer*”))=>*buys*(*X*; “*software*”) [*support* = 1%; *confidence* = 50%]

*Where X is a variable representing a customer*

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.

**Classification and Prediction**

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts. Mainly it is used to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data. Data object whose class label is known is considered as training data. The derived model may be represented in various forms, such as *classification (IF-THEN) rules*, *decision trees*, *mathematical formulae*, *neural networks* etc.

Whereas classification predicts categorical labels, prediction models continuous-valued functions. That is, it is used to predict missing or unavailable *numerical data values* rather than class labels. Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Prediction also encompasses the identification of distribution *trends* based on the available data. Classification and prediction may need to be preceded by relevance analysis, which attempts to identify attributes that do not contribute to the classification or prediction process. These attributes can then be excluded.

**Cluster Analysis**

Unlike classification and prediction, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of *maximizing the intra-class similarity and minimizing the interclass similarity*. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.



Figure: Three data clusters

**Outlier Analysis**

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions. However, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones. 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 a substantial distance from any other cluster are considered outliers. For example, Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account.

**Evolution Analysis**

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. Distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis. For example, you have the major stock market (time-series) data of the last several years available from the Nepal Stock Exchange and you would like to invest in shares of high-tech industrial companies. A data mining study of stock exchange data may identify stock evolution regularities for overall stocks and for the stocks of particular companies. Such regularities may help predict future trends in stock market prices, contributing to your decision making regarding stock investments.

**Major Issues in Data Mining**

Major issues in data mining are about mining methodology, user interaction, performance, and diverse data types. These issues are introduced below:

**Mining methodology and user interaction issues**

* *Mining different kinds of knowledge in databases:* Because different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, including data characterization, discrimination, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis. These tasks may use the same database in different ways and require the development of numerous data mining techniques.
* *Interactive mining of knowledge at multiple levels of abstraction:* Because it is difficult to know exactly what can be discovered within a database, the data mining process should be interactive. Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results.
* *Incorporation of background knowledge:* Domain knowledge related to databases, such as integrity constraints and deduction rules, can help focus and speed up a data mining process, or judge the interestingness of discovered patterns.
* *Data mining query languages and ad hoc data mining: H*igh-level data mining query languages need to be developed to allow users to describe ad hoc data mining tasks by facilitating the specification of the relevant sets of data for analysis, the domain knowledge, the kinds of knowledge to be mined, and the conditions and constraints to be enforced on the discovered patterns.
* *Presentation and visualization of data mining results:* Discovered knowledge should be expressed in high-level languages, visual representations, or other expressive forms so that the knowledge can be easily understood and directly usable by humans.
* *Handling noisy or incomplete data:* The data stored in a database may reflect noise, exceptional cases, or incomplete data objects. These objects may confuse the data mining process, causing the knowledge model constructed to overfit the data. Thus data cleaning methods and data analysis methods that can handle noise are required, as well as outlier mining methods for the discovery and analysis of exceptional cases.
* *Pattern evaluation—the interestingness problem:* A data mining system can uncover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user, either because they represent common knowledge or lack novelty. Thus the use of interestingness measures or user-specified constraints to guide the discovery process and reduce the search space is another active area of research.

**Performance Issues**

* *Efficiency and scalability of data mining algorithms: R*unning time of a data mining algorithm must be predictable and acceptable in large databases. From a database perspective on knowledge discovery, efficiency and scalability are key issues in the implementation of data mining systems.
* *Parallel, distributed, and incremental mining algorithms:* The huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms. Such algorithms divide the data into partitions, which are processed in parallel. The results from the partitions are then merged. Moreover, the high cost of some data mining processes promotes the need for incremental data mining algorithms that incorporate database updates without having to mine the entire data again “from scratch.” Such algorithms perform knowledge modification incrementally to amend and strengthen what was previously discovered.

**Issues Relating to the Diversity of Database Types**

* *Handling of relational and complex types of data:* Because relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important. However, other databases may contain complex data objects, hypertext and multimedia data, spatial data, temporal data, or transaction data. It is unrealistic to expect one system to mine all kinds of data, given the diversity of data types and different goals of data mining. Specific data mining systems should be constructed for mining specific kinds of data. Therefore, one may expect to have different data mining systems for different kinds of data.
* *Mining information from heterogeneous databases and global information systems:* Local- and wide-area computer networks connect many sources of data, forming huge, distributed, and heterogeneous databases. Data mining may help disclose high-level data regularities in multiple heterogeneous databases that are unlikely to be discovered by simple query systems and may improve information exchange and interoperability in heterogeneous databases. Web mining, which uncovers interesting knowledge about Web contents, Web structures, Web usage, and Web dynamics, becomes a very challenging and fast-evolving field in data mining.

**Major Issues in Data Warehousing**

 Building a data Warehouse is very difficult and a pain. It is challenging, but it is a fabulous project to be involved in, because when data warehouses work properly, they are magnificently useful, huge fun and unbelievably rewarding. Some of the major issues involved in building data warehouse are discussed below:

**General Issues:** It includes but is not limited to following issues:

* What kind of analysis do the business users want to perform?
* Do you currently collect the data required to support that analysis?
* How clean is data?
* Are there multiple sources for similar data?
* What structure is best for the core data warehouse (i.e., dimensional or relational)?

**Technical Issues:** It includes but is not limited to following issues

* How much data are you going to ship around your network, and will it be able to cope?
* How much disk space will be needed?
* How fast does the disk storage need to be?
* Are you going to use SSDs to store “hot” data (i.e., frequently accessed information)?
* What database and data management technology expertise already exists within the company?

**Cultural Issues**: It includes but is not limited to following issues

* How do data definitions differ between your operational systems? Different departments and business units often use their own definitions of terms like “customer,” “sale” and “order” within systems. So you’ll need to standardize the definitions and add prefixes such as “all sales,” “recent sales,” “commercial sales” and so on.
* What’s the process for gathering business requirements? Some people will not want to spend time for you. Instead, they will expect you to use your telepathic powers to divine their warehousing and data analysis needs.

**Applications of Data Warehousing**

Information processing, analytical processing, and data mining are the three types of data warehouse applications that are discussed below:

* **Information Processing** - A data warehouse allows to process the data stored in it. The data can be processed by means of querying, basic statistical analysis, reporting using crosstabs, tables, charts, or graphs.
* **Analytical Processing** - A data warehouse supports analytical processing of the information stored in it. The data can be analyzed by means of basic OLAP operations, including slice-and-dice, drill down, drill up, and pivoting.
* **Data Mining** - Data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction. These mining results can be presented using the visualization tools.

**Application of Data Mining**

* **Market Analysis and Management**: Target marketing, customer relation management, market basket analysis, cross selling, market segmentation, Find clusters of customers who share the same characteristics: interest, income level, spending habits, etc. Determine customer purchasing patterns over time
* **Risk Analysis and Management**: Forecasting, customer retention, improved underwriting, quality control, competitive analysis, credit scoring.
* **Fraud Detection and Management**: Use historical data to build models of fraudulent behavior and use data mining to help identify similar instances. For example, detect suspicious money transactions.
* **Sports:** Data mining can be used to analyze shots & fouls of different athletes, their weaknesses and helps athletes to assist in improving their games.
* **Space Science:** Data mining can be used to automate the analysis image data collected from sky survey with better accuracy.
* **Internet Web Surf-Aid:** Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.
* **Social Web and Networks**: There are a growing number of highly-popular user-centric applications such as blogs, wikis and Web communities that generate a lot of structured and semi-structured information. In these applications data mining can be used to explain and predict the evolution of social networks, personalized search for social interaction, user behavior prediction etc.