Data Mining Systems

There are many data mining systems available or presently being developed. Some are specialized systems dedicated to a given data sources or are confined to limited data mining functionalities, while others are more versatile and comprehensive. This unit examines the various classifications of data mining systems, data mining tasks, the major issues and challenging in data mining.

Classification of Data Mining Systems

The data mining system can be classified according to the following criteria:

a) Statistics
b) Database Technology
c) Machine Learning
d) Information Science
e) Visualization
f) Other Disciplines

Some Other Classification Criteria:

a) Classification according to kind of databases mined
b) Classification according to kind of knowledge mined
c) Classification according to kinds of techniques utilized
d) Classification according to applications adapted

i. **Classification according to kind of databases mined:**

We can classify the data mining system according to kind of databases mined. Database system can be classified according to different criteria such as data models, types of data etc. And the data mining system can be classified accordingly. For example, if we classify the database according to data model then we may have a relational, transactional, object- relational, or data warehouse mining system

ii. **Classification according to kind of knowledge mined:**

We can classify the data mining system according to kind of knowledge mined. It is means data mining system are classified on the basis of functionalities such as:

a) Characterization
b) Discrimination
c) Association and Correlation Analysis
d) Classification
e) Prediction
f) Clustering
g) Outlier Analysis
h) Evolution Analysis

iii. **Classification according to kinds of techniques utilized**

We can classify the data mining system according to kind of techniques used. We can describe these techniques according to degree of user interaction involved or the methods of analysis employed.

iv. **Classification according to applications adapted:**

We can classify the data mining system according to application adapted. These applications are as follows:

a) Finance
b) Telecommunications
c) DNA
d) Stock Markets
e) E-mail

Data Mining Task

The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on inference on available data. The data mining functionalities and the variety of knowledge they discover are briefly presented in the following list:

- **Characterization:** Data characterization is a summarization of general features of objects in a target class, and produces what is called characteristic rules. The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions.

- **Association analysis**: Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules. Association analysis is commonly used for market basket analysis.

- **Classification**: Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects

- **Prediction**: Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values.

- **Clustering**: Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity).

- **Outlier analysis**: Outliers are data elements that cannot be grouped in a given class or cluster. Also known as exceptions or surprises, they are often very important to identify. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable.

- **Evolution and deviation analysis**: Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data. Deviation analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values.

3.3 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 (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semi-structured, or unstructured data with diverse data semantics poses great challenges to data mining. 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.

The above issues are considered major requirements and challenges for the further evolution of data mining technology.

Integration of a Data Mining System with a Database or Data Warehouse System

A critical question in the design of a data mining system is how to integrate or couple the data mining system with a database system and/or a data warehouse system. If a data mining system works as a stand-alone system or is embedded in an application program, there are no database system or Data warehouse systems with which it has to communicate. This simple scheme is called no coupling, where the main focus of the data mining design rests on developing effective and efficient algorithms for mining the available data sets. However, when a data mining system works in an environment that requires it to communicate with other information system components, such as database and Data warehouse systems, possible integration schemes include no coupling, loose coupling, semi tight coupling, and tight coupling. We examine each of these schemes, as follows:

No coupling: No coupling means that a data mining system will not utilize any function of a database or Data warehouse system. It may fetch data from a particular source (such as a file system), process data using some data mining algorithms, and then store the mining results in another file. Such a system, though simple, suffers from several drawbacks. First, a database system provides a great deal of flexibility and efficiency at storing, organizing, accessing, and processing data. Without using a database / Data warehouse, a data mining system may spend a substantial amount of time finding, collecting, cleaning, and transforming data. In database and/or Data warehouse systems, data tend to be well organized, indexed, cleaned, integrated, or consolidated, so that finding the task-relevant, high-quality data becomes an easy task. Second, there are many tested, scalable algorithms and data structures implemented in database
and Data warehouse systems. It is feasible to realize efficient, scalable implementations using such systems. Moreover, most data have been or will be stored in database/Data
warehouse systems. Without any coupling of such systems, a data mining system will need to use other tools to extract data, making it difficult to integrate such a system into an information processing environment. Thus, no coupling represents a poor design.

Loose coupling: Loose coupling means that a data mining system will use some facilities of a database or Data warehouse system, fetching data from a data repository managed by these systems, performing data mining, and then storing the mining results either in a file or in a designated place in a database or data warehouse. Loose coupling is better than no coupling because it can fetch any portion of data stored in databases or data warehouses by using query processing, indexing, and other system facilities. It incurs some advantages of the flexibility, efficiency, and other features provided by such systems. However, many loosely coupled mining systems are main memory-based. Because mining does not explore data structures and query optimization methods provided by database or Data warehouse systems, it is difficult for loose coupling to achieve high scalability and good performance with large data sets.

Semi-tight coupling: Semi-tight coupling means that besides linking a data mining system to a database/Data warehouse system, efficient implementations of a few essential data mining primitives (identified by the analysis of frequently encountered data mining functions) can be provided in the database/Data warehouse system. These primitives can include sorting, indexing, aggregation, histogram analysis, multiway join, and precomputation of some essential statistical measures, such as sum, count, max, min, standard deviation, and so on. Moreover, some frequently used intermediate mining results can be precomputed and stored in the database/Data warehouse system. Because these intermediate mining results are either precomputed or can be computed efficiently, this design will enhance the performance of a data mining system.

Tight coupling: Tight coupling means that a data mining system is smoothly integrated into the database/Data warehouse system. The data mining subsystem is treated as one functional component of an information system. Data mining queries and functions are optimized based on mining query analysis, data structures, indexing schemes, and query processing methods of a database or Data warehouse system. With further technology advances, data mining, database, and Data warehouse systems will evolve and integrate together as one information system with multiple functionalities. This will provide a uniform information processing environment. This approach is highly desirable because it facilitates efficient implementations of data mining functions, high system performance, and an integrated information processing environment.

With this analysis, it is easy to see that a data mining system should be coupled with a database/ Data warehouse system. Loose coupling, though not efficient, is better than no coupling because it uses both data and system facilities of a database/ Data warehouse system. Tight coupling is highly desirable, but its implementation is nontrivial and more research is needed in this area. Semi-tight coupling is a compromise between loose and tight coupling. It is important to identify commonly used data mining primitives and provide efficient implementations of such primitives in database or Data warehouse systems.

Data Mining Issues

a) **Mining different kinds of knowledge in databases:**

The need of different users is not the same. And Different user may be in interested in different kind of knowledge. Therefore, it is necessary for data mining to cover broad range of knowledge discovery task.

b) **Interactive mining of knowledge at multiple levels of abstraction:**

The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on returned results.

c) **Incorporation of background knowledge:**

Background knowledge, or information regarding the domain under study, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction. 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.

d) **Data mining query languages and ad hoc data mining:**
Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.

e) **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. This is especially crucial if the data mining system is to be interactive. This requires the system to adopt expressive knowledge representation techniques, such as trees, tables, rules, graphs, charts, crosstabs, matrices, or curves.

f) **Handling noisy or incomplete data:**
The data cleaning methods are required that can handle the noise, incomplete objects while mining the data regularities. If data cleaning methods are not there then the accuracy of the discovered patterns will be poor.

g) **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. Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns, particularly with regard to subjective measures that estimate the value of patterns with respect to a given user class, based on user beliefs or expectations.

h) **Efficiency and scalability of data mining algorithms:**
To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable. In other words, the running 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.

i) **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.