CHAPTER-4

DATA MINING
CHAPTER-4

DATA MINING

4.1 INTRODUCTION TO DATA MINING

The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed. The proliferation of database management systems has also contributed to recent massive gathering of all sorts of information. Today, we have far more information than we can handle: from business transaction and scientific data, to satellite pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision making. Confronted with huge collections of data, we have now created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the “essence” of information stored, and the discovery of patterns in raw data. We are in an age often referred to as the information age. In this information age, because we believe that information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, etc.,

we have been collecting tremendous amounts of information. Initially, with the advent of computers and means for amass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became over shelling.
4.2 WHAT KIND OF INFORMATION ARE WE COLLECTING?

We have been collecting a myriad of data, from simple numerical measurements and text documents, to more complex information such as spatial data, multimedia channels, and hypertext documents. Here is a non-exclusive list of a variety of information collected in digital form in databases and in flat files.

- **Business transactions**

  Large department stores, for example, thanks to the widespread use of bar codes, store millions of transaction daily representing often terabytes of data. Storage space is not the major problem, as the price of hard disks is continuously dropping, but the effective use of the data in a reasonable time frame for competitive decision making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world (M.S. Chem., et al.) [9]. Every transaction in the business industry is often “memorized” for perpetuity. Such transactions are usually time related and can be inter-business deals such as purchases, exchanges, banking, stock, etc., or intra-business operations such as management of in-house wares and assets.

- **Scientific Data**

  Unfortunately, we can capture and store more new data faster than we can analyze the old data already accumulated. Whether in a Swiss nuclear accelerator laboratory counting particles, in the Canadian forest studying readings from a grizzly bear radio collar, on a South Pole iceberg gathering data about oceanic activity, or in an American university investigating human psychology, our society is amassing colossal amounts of scientific data that need to be analyzed.
• **Medical and personal data**

From government census to personnel and customer files, very large collections of information are continuously gathered about individuals and groups. Governments, companies and organizations such as hospitals, are stockpiling very important quantities of personal data to help them manage human resources, better understand a market, or simply assist clientele. Regardless of the privacy issues this type of data often reveals, this information is collected, used and even shared. When correlated with other data this information can light on customer behavior and the like.

• **Satellite sensing**

There are a countless number of satellites around the globe: Some are geo-stationary above a region, and some are orbiting around the Earth, but all are sending a non-stop stream of data to the surface. NASA, which controls a large number of satellites, received more data every second than what all NASA researchers and engineers can cope with. Many satellite pictures and data are made public as soon as they are received in the hopes that other researchers can analyze them.

• **Games**

Our society is collecting a tremendous amount of data and statistics about games, players and athletes. From hockey scores, basketball passes and car-racing lapses, to swimming times, boxer’s pushes and chess positions, and all the data are stored. Commentators and journalists are using this information for reporting, but trainers and athletes would want to exploit this data to improve performance and better understand opponents.
• **Digital Media**

Many radio stations, television channels and film studio are digitizing their audio and video collections to improve the management of their multimedia assets. Associations such as the NHL and the NBA have already started converting their huge game collection into digital forms. In addition the proliferation of cheap scanners, desktop video cameras and digital cameras is one of the causes of the explosion in digital media repositories.

• **CAD and Software engineering data**

There is a multitude of Computer Assisted Design (CAD) system for architects to design buildings or engineers to conceive system components or circuits (J Han et al) [10], these systems are generating a tremendous amount of data. Moreover, software engineering is a source of considerable similar with code, function libraries, objects, etc., which need powerful tools for management and maintenance.

• **Virtual Worlds**

There is a remarkable amount of virtual reality object and space repositories available. Management of these repositories as well as content-based search and retrieval from these repositories are still research issues, while the size of the collection continues to grow. There are many applications making use of three dimensional virtual spaces. Ideally, these virtual spaces are described in such a way that they can shares objects and places.
Text report and memos (e-mail massage)

Most of the communications within and between companies or research organization or even private people, are based on reports and memos in textual forms often exchanged by e-mail. These messages are regularly stored in digital forms for future use and reference creating formidable digital libraries.

The World Wide Web repositories

Many believe that the World Wide Web will become the compilation of human knowledge. Since the inception of the World Wide Web in 1993, documents of all sorts of formats, contents and description have been collected and inter-connected with hyperlinks making it the largest repository of data ever built. Despite its dynamic and unstructured nature, its heterogeneous characteristic, and it’s very often redundancy and inconsistency, the World Wide Web is the most important data collection regularly used for reference because of the broad variety of topics covered and the infinite contributions of resources and publishers.

4.3 WHAT ARE DATA MINING AND KNOWLEDGE DISCOVERY?

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, of not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or
KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure (figure 1) shows data mining as a step in an iterative knowledge discovery process (G. Piatetsky et al) [11]. The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

**Data cleaning:** also know as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

**Data integration:** at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

**Data Selection:** At this step, the data relevant to the analysis is decided on and retrieved from the data collection.

**Data transformation:** also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

**Data mining:** it is crucial step in which clever techniques are applied to extract patterns potentially useful.

**Pattern evaluation:** in this step, strictly interesting representing knowledge are identified based on given measure.
Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results. It is common to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a preprocessing phase to generate a data warehouse.

Date selection and data transformation can also be combined where the consolidation of the date is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data. The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measure can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more
appropriate results. Data mining derives its name from the similarities between searching for valuable information in a large database and mining. Both imply either sifting through a large amount of material the material to exactly pinpoint where the values reside. Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery.

4.4 WHAT KIND OF DATA CAN BE MINED?

In principle, data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. However, algorithms and approaches may differ when applied to different types of data. Indeed, the challenges presented by different types of data vary significantly. Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object-oriented databases, data warehouse, transactional databases, unstructured and semi structured repositories, such as the World Wide Web, advanced databases such as spatial database, multimedia database, time-series databases and textual databases, and even flat files. Here are some examples in more details.

- **Flat Files**

Flat files are actually the most common data sources for data mining algorithms, especially at the research level. Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied. The data in these files can be transactions, time-series, scientific measurement, etc.
• **Relational Databases**

Briefly, a relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key.

• **Data Warehouse**

A data warehouse as a storehouse is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema. A DW gives the option to analyze data from different sources under the same roof. Let us suppose that Our Video Store becomes a franchise in North America. Many Video stores belonging to Our Video Store Company may have different databases and different structures. If the executive of the company wants to access the data from all stores for strategic decision-making, future direction, marketing, etc., it would be more appropriate to store all the data in one site with a homogeneous structure that allows interactive analysis. In other words, data from the different stores would be loaded, cleaned, transformed and integrated together. To facilitate decision making and multidimensional views, data warehouse are usually modeled by a multi dimensional data structure.

• **Transaction Databases**

A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items. Associated with the transaction files could also be descriptive data for the items. Each record
is a rental contract with a customer identifier, a date, and the list of
times rented (i.e. video tapes, games, VCR, etc.). Since relational
databases do not allow nested tables (i.e. as set as attribute value),
transactions are usually stored in flat files or stored in two normalized
transaction tables, one for the transactions and one for the transaction
items. One typical data mining analysis on such data is the so-called
market.

- **Multimedia Databases**

Multimedia databases include video, images, audio and text media. They
can be stored on extended object-relational or object-oriented databases,
or simply on a file system. Multimedia is characterized by its high
dimensionality, which makes data mining even more challenging. Data
mining from multimedia repositories may require computer vision,
computer graphics, image interpretation, and natural language
processing methodologies.

- **Spatial Databases**

Spatial database are databases that, in addition to usual data, store
geographical information like maps, and global or regional positioning.
Such spatial databases present new challenges to data mining
algorithms.

- **Time-Series Database**

Time-series databases contain time related data such stock market data.
These databases usually have a continuous flow of new data coming in,
which sometimes causes the need for a challenging real time analysis.
Data mining in such databases commonly includes the study of trends
and correlations between evolutions of different variable, as well as the prediction of trends and movements of the variable in time.

World Wide Web

The World Wide Web is the most heterogeneous and dynamic repository available. A very large number of authors and publishers are continuously contributing to its growth and metamorphosis, and a massive number of users are accessing its resources daily. Data in the World Wide Web is organized in inter-connected documents. These documents can be text, audio, video, raw data, and even applications. Conceptually, the World Wide Web is comprised of three major components: The content of the Web, Which encompasses documents available; the structure of the Web, which covers the relationship between documents; and the usage of the web, describing how and when the resources are accessed. A fourth dimension can be added relating the dynamic nature or evolution of the documents. Data mining in the World Wide Web, or web mining, tries to address all these issues and is often divided into web content mining, web structure mining and web usage mining.

4.5 WHAT CAN BE DISCOVERED?

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 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 lists:
• **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. For example, one may want to characterize the Our Video Store customers who regularly rent more than 30 movies a year. With concept hierarchies on the attributes describing the target class, the attribute oriented induction method can be used, for example, to carry out data summarization. Note that with a data cube containing summarization of data, simple OLAP operations fit the purpose of data characterization.

• **Discrimination**

Data discrimination produces what are called discriminate rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class. For example, one may want to compare the general characteristics of the customers who rented more than 30 movies in the last year with those whose rental account is lower than 5. The techniques used for data discrimination are very similar to the techniques used for data characterization with the exception that data discrimination results include comparative measures.

• **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 that 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. For example, it could be useful for the Our Video Store manager to know what movies are often rented together or if there is a relationship between renting a certain type of movies and buying popcorn or pop. The discovered association rules are of the form: P → Q \[s, c\], where P and Q are conjunctions of attribute value-pairs, and s (for support is the probability that P and Q appear together in a transaction and c (for confidence) is the conditional probability that \( Q \) appears in a transaction when \( P \) is present. For example, the hypothetic association rules:

\[
\text{Rent Type (X, “game”) \land \text{Age(X, “13-19”) \land \text{Buys(X, “pop”) \[s=2\%, c=55\%\]}}
\]

would indicate that 2% of the transactions considered are of customers aged between 13 and 19 who are renting a game and buying a pop, and that there is a certainty of 55% that teenage customers who rent a game also buy pop.

### 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. For example, after starting a credit policy, the Our Video Store managers could analyze the customers’ behaviors vis-à-vis their
credit, and label accordingly the customers who received credits with three possible labels “sale”, “risky” and “very risky”.

- **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 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. It is common that users do not have a clear idea of the kind of patterns they can discover or need to discover from the data at hand. It is therefore important to have a versatile and inclusive data mining system that allows the discovery of different kinds of knowledge and at different levels of abstraction. This also makes interactivity an important attribute of a data mining system.

**4.6 IS ALL THAT IS DISCOVERED INTERESTING AND USEFUL?**

Data mining allows the discovery of knowledge potentially useful and unknown. Whether the knowledge discovered is new, useful or interesting, is very subjective and depends upon the application and the user. It is certain that data mining can generate, or discover, a very large number of patterns or rules. In some cases the number of rules can reach the millions. One can even think of a meta-mining phase to mine the oversized data mining results. To reduce the number of patterns or
rules discovered that have a high probability to be non-interesting, one has to put a measurement on the patterns. However, this raises the problem of completeness. The user would want to discover all rules or patterns, but only those that are interesting. The measurement of how interesting a discovery is, often called interestingness, can be based on quantifiable objective elements such as validity of the patterns when tested on new data with some degree of certainly, or on some subjective such as understandability of the patterns, novelty of the patterns, or usefulness.

Discovered patterns can also be found interesting if they confirm or validate a hypothesis sought to be confirmed or unexpectedly contradict a common belief. This brings the issue of describing what is interesting to discover, such as meta-rule guided discovery that describes forms of rules before the discovery process, and interestingness refinement languages that interactively query the results for interesting patterns after the discovery phase. Typically, measurements for interestingness are based on thresholds set by the user. These thresholds define the completeness of patterns discovered. Identifying and measuring the interestingness of patterns and rules discovered, or to be discovered is essential for the evaluation of the mined knowledge and the KDD process as a whole. While some concrete measurements exist, assessing the interestingness of discovered knowledge is still an important research issue.

4.7 HOW DO WE CATEGORIZE DATA MINING SYSTEM?

There are many data mining systems available or being developed. Some are specialized systems dedicated to a given data source or are confined to limited data mining functionalities, other are more versatile and
comprehensive. Data mining systems can be categorized according to various criteria among other classification are the following:

- **Classification according to the type of data source mined**
  This classification categorizes data mining systems according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web etc.

- **Classification according to the data model drawn on**
  This classification categorized data mining systems based on the data model involved such as relational database, object-oriented database, data warehouse, transactional etc.

- **Classification according to the kind of knowledge discovered**
  This classification categorizes data mining systems based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

- **Classification according to mining techniques used.**
  Data mining system employ and provide different techniques. This classification categorized data mining system according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc. The classification can also take into account the degree of user interaction involved in the data mining process such as
query-driven systems, interactive exploratory systems, or autonomous systems.

4.8 WHAT ARE THE ISSUES IN DATA MINING?

Data mining algorithms embody techniques that have sometimes existed for many years, but have only lately been applied as reliable and scalable tools that time and again outperform older classical statistical methods. While data mining is still in its infancy, it is becoming a trend and ubiquitous. Before data mining develops into a conventional, mature and trusted discipline, many still pending issues have to be addressed. Some of these issues are addressed below. Note that these issues are not exclusive and are not ordered in any way.

- Security and social issues

Security is an important issue with any data collection that is shared and/or is intended to be used for strategic decision-making. In addition, when data is collected for customer profiling, user behavior understanding, correlating personal data with other information, etc., large amounts of sensitive and private information about individuals or companies is gathered and stored. This becomes controversial given the confidential nature of some of this data and the potential illegal access to the information. Moreover, data mining could disclose new implicit knowledge about individuals or groups that could be against privacy policies, especially if there is potential dissemination of discovered information. Another issue that arises from this concern is the appropriate use of data mining. Due to the value of data, databases of all sorts of content are regularly sold, and because of the competitive advantage that can be attained from implicit knowledge discovered, some
important information could be withheld, while other information could be widely distributed and used without control.

- **User interface issues.**

  The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the users. Good data visualization eases the interpretation of data mining results, as well as helps users better understand their needs. Many data exploratory tasks significantly facilitated by the ability to see data in an appropriate visual presentation. There are many visualization ideas and proposals for effective data graphical presentation. However, there is still much research to accomplish in order to obtain good visualization tools for large datasets that could be used to display and manipulate mined knowledge. The major issues related to user interfaces and visualization are “screen real-estate”, information rendering, and interaction. Interactivity with the data and data mining tasks, as well as to picture the discovered knowledge from different angles and at different conceptual levels.

- **Mining methodology issues**

  These issues pertain to the data mining approaches applied and their limitations. Topics such as versatility of the mining approaches, the diversity of data available, the dimensionality of the domain, the broad analysis need (when known), the assessment of the knowledge discovered, the exploitation of background knowledge and metadata, the control and handling of noise in data, etc. are all examples that can dictate mining methodology choices. For instance, it is often desirable to have different data mining methods available since different approaches may perform differently depending upon the data at hand. Moreover,
different approaches may suit and solve user’s needs differently. Most algorithms assume the data to be noise-free. This is of course a strong assumption. Most datasets contain exceptions, invalid or incomplete information, etc., which may complicate, if not obscure, the analysis process and in many cases compromise the accuracy of the results. As a consequence, data preprocessing (data cleaning and transformation) becomes vital. It is often seen as lost time, but data cleaning, as time-consuming and frustrating as it may be, is one of the most important phases in the knowledge discovery process. Data mining techniques should be able to handle noise in data or incomplete information. More than the size of data, the size of the search space is even more decisive for data mining techniques. The size of the search space is often depending upon the number of dimensions in the domain space. The search space usually grows exponentially when the number of dimension increases. This is known as the curse of dimensionality. This “curse” affects so badly the performance of some data mining approaches that it is becoming one of the most urgent issues to solve.

- **Performance issues**

Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the very large data sets data mining is dealing with today. Terabyte sizes are common. This raises the issues of scalability and efficiency of the data mining methods when processing considerably large data. Algorithms with exponential and even medium-order polynomial complexity cannot be of practical use for data mining. Linear algorithms are usually the norm. In same theme, sampling can be used for mining instead of the whole dataset. However, concerns such as completeness and choice of samples may arise. Other topics in the issue of
performance are incremental updating, and parallel programming. There is no doubt that parallelism can help solve the size problem if the dataset can be subdivided and the results can be merged later. Incremental updating is important for merging results from parallel mining, or updating data mining results when new data becomes available without having to re-analyze the complete dataset.

- **Date source issues**

There many issues related to the data source, some are practical such as the diversity of data types, while other are philosophical like the data glut problem. We certainly have an excess of data since we already have more data than we can handle and we are still collecting data at an even higher rate. If the spread of database management systems has helped increase the gathering of information, the advent of data mining is certainly encouraging more data harvesting. The current practice is to collect as much data as possible now and process, or try to process it, later. The concern is whether we are collecting the right data at the appropriate amount, whether we known what we want to do with it, and whether we distinguish between what data is important and what data is insignificant. Regarding the practical issues related to data sources, there is the subject of heterogeneous databases and the focus on diverse complex data types. We are storing different types of data in a variety of repositories. It is difficult to expect a data mining system to effectively and efficiently achieve good mining results on all kinds of data and sources. Different kinds of data and sources many require distinct algorithms and methodologies. Currently, there is a focus on relational databases and data warehouse, but other approaches need to be pioneered for other specific complex data types.