# DATA DESCRIPTION FOR DATA MINING

## Types of Information Collected

We collect on daily basis a myriad of data which ranges from simple numerical measurements and text documents to more complex information such as hypertext documents, scientific data, spatial data and multimedia channels. Here is a different kind of information often collected in digital form in databases and flat files, although not exclusive.

### 1. Scientific Data

Our society is seriously gathering great amount of scientific data that needs to be analyzed. Examples are in the Swiss nuclear accelerator laboratory counting particles, South Pole iceberg gathering data about oceanic activity, American university investigating human psychology and Canadian forest studying readings from a grizzly bear radio collar. The unfortunate part of it is we can easily capture and store more new data faster than we can analyze the old data that have been accumulated.

### 2. Personal and Medical Data

From personal data to medical and government, very large amounts of information are continuously collected. Governments, individuals and organizations such as hospitals and schools are on daily basic stockpiling large quantity of very important personal data to help them manage human resources, better understanding of market, or simply assist client. No matter the private issues this type of data reveals, the information is collected used and even shared. And when compared with other data this information can shed more light on customer behavior and likes.

### 3. Games

The rate at which our society gathers data and statistics about games, players and athletes is tremendous. These ranges from car-racing, swimming, hockey scores, footballs, basketball passes, chess positions and boxers' pushes, all these data are stored. Trainers and athletes make use of this data to improve their performances and have a better understanding of their opponents, but the journalists and commentators use this information to report.

### 4. CAD and Software Engineering Data

There are different types of Computer Assisted Design (CAD) systems used by architects and engineers to design buildings and picture system components or circuits. These systems generate a great amount of data. Also, software engineering is a source of data generation with code, function libraries and objects, these needs powerful tools for management and maintenance

### 5. Business Transaction

Every transaction in business is often noted for the sake of continuity. These transactions are usually related and can be inter-business deals such as banking, purchasing, exchanges and stocks or intra-business operations such as management of in-house wares and assets. Large departmental stores for example stores millions of transactions on daily basis with the use barcodes.

The storage space does not pose any problem, as the price of hard disks are dropping, but the effective use of the data within a reasonable time frame for competitive decision-making is certainly the most important problem to be solved for businesses that struggle in competitive world.

### 6. Surveillance Video and Pictures

With the incredible fall in price of video camera prices, video cameras are becoming very common. The video tapes from surveillance cameras are usually recycled, thereby losing its content. But today there is tendency to store the tapes and even digitize them for future use and analysis.

### 7. Satellite Sensing

There are countless numbers of satellites around the globe, some are geo-stationary above a region while some are orbiting round the Earth, but all are sending a non-stop of data to the surface of the earth. NASA which is a body controlling large number of satellites receives more data per second than all NASA engineers and researchers can cope with. Many of the pictures and data captured by the satellite are made public as soon they are received hoping that other researchers can analyze them.

### 8. World Wide Web (WWW) Repositories

Since the advent of World Wide Web in 1993, documents of different formats, contents and description have been collected and inter- connected with hyperlinks making it the largest repository of data ever built, The World Wide Web is the most important data collection regularly used for reference because of the wide variety of topics covered and the infinite contributions of resources and authors. Many even believe that the World Wide Web is a compilation of human knowledge.

## Types of Data Mined

Data mining can be applied to any kind of information in the repository, though algorithms and approaches may differ when applied to different types of data. And the challenges posed by different types of data vary extensively. Data Mining is used and studied for databases including relational databases, object-relational databases and object-oriented data, bases data warehouses, transactional databases, unstructured and semi structured repositories such as the World Wide Web, and advanced databases such as spatial databases, multimedia database, time-series databases and flat files. Some of these are discussed in more details as follows.

### a) Flat Files

These are the commonest data source for data mining algorithms especially at the research level. Flat files are simply data files in text or binary format with a structured known by the data mining algorithms to be applied. The data in these files can be in form of transactions, time-sales data, scientific measurements etc.

### b) Relational Databases

This is the most popular type of database system in use today by computers. It stores data in a series of two-dimensional tables called relation (i.e. tabular form). A relational database consists of a set of tables containing either values of entity attributes, or value of attribute from entity relationships. Tables generally have columns and rows, where columns represent attribute 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. In the table below we present some relations student name, registration number, department and grade in computer representing a fictitious student grade in a class. These relations are just a subset of what could be a database for student score.

### c) Relational Database

| Student Name | Registration | Department | Grade in Data Mining |
|---|---|---|---|
| Ken | BUS/05/MLD/101 | Business | A |
| John | MKT/05/MLD/105 | Marketing | B |
| Nancy | BFN/05/MLD/203 | Banking & Finance | A |
| Mary | ACC/05/MLD/102 | Accountancy | C |
| Victor | BUS/05/MLD/200 | Business | B |

The most commonly used query language for relational database is Structured Query Language (SQL), it allows for retrieval and manipulation of data stored in the tables as well as the calculation of aggregate function such as sum, min, max and count. The data mining algorithm that uses relational databases can be more versatile than data mining algorithm that is specifically designed for flat files because they can always take advantage of the structure inherent in relational databases, while data mining can benefit from Structured Query Language (SQL) for data selection, transformation and consolidation. Also, it goes beyond what SQL can provide like predicting, comparing and detecting deviations.

### d) Data Warehouses

A data warehouse (a storehouse) is a repository of data gathered from multiple data sources (often heterogeneous) and is designed to be used as a whole under the same unified schema. A data warehouse provides an option of analyzing data from different sources under the same roof. The most efficient data warehousing architecture will be able to incorporate or at least reference all management systems using designated technology suitable for corporate database management e.g. Sybase, Ms SQL Server

### e) Transaction Databases

This is a set of records that represent transactions, each with a time stamp, an identifier and set of items. Also, associated with the transaction files is the descriptive data for the items.

Rentals

| Transaction (1) | Data | Time | Customer ID | Item List |
|---|---|---|---|---|
| TI | 14/09/04 | 14.40 | 12 | 10,11,30, 110.. |
| II. | III. | IV. | V. | VI. |
| VII. | VIII. | IX. | X. | XI. |

Fragment of a Transaction Database for Rentals in a Store

Figure above represents a transaction database, each record shows a rental contact with a customer identifier, a date and list of items rented. But relational database do not allow nested tables that is a set as attribute value, transactions are usually stored in flat files or stored in two normalized transaction tables, one for the transactions and the other one for the transaction items. A typical data analysis on such data is the so-called market basket analysis or association rules in which associations occurring together or in sequence are studied.

### f) Spatial Databases

These are databases that in addition to the usual data stores geographical information such as maps, global or regional positioning, and this type of database also present new challenges to data mining algorithms.

### g) Multimedia Databases

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

### h) Time-Series Databases

This type of database contains time related data such as stock market data or logged activities. Time-series database usually contain a continuous flow of new data coming in that sometimes causes the need for a challenging real time analysis. Data mining in these types of databases often include the study of trends and correlations between evolutions of different variables, prediction of trends and movements of the variables in time.

### i) World Wide Web

World Wide Web is the most heterogeneous and dynamic repository available. Large number of authors and publishers are continuously contributing to its growth and metamorphosis, and a massive number of users are assessing its resources daily. The data in the World Wide Web are organized in inter-connected documents, which can be text, audio, video, raw data and even applications. The World Wide Web comprises of three major components: the content of the web, which encompasses document available, the structure of the web, which covers the hyperlinks and the relationships between documents the usage of the web, this describe 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, addresses all these issues and is often divided into web content mining and web usage mining.

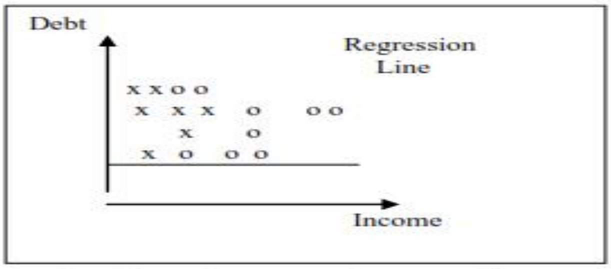
## DATA MINING FUNCTIONALITIES

Data mining functionalities are used to specify the kind of patterns to be found in data mining task. It is a very common phenomenon that many users do not have clear idea of the kind of patterns they can discover or need to discover from the data at hand. It is therefore crucial 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 issue in data mining system.

The data mining functionalities and the variety of knowledge they discover are described in this section as follows:

### 1. Classification

This is also referred to as supervised classification and is a learning function that maps (i.e. classifies) item into several given classes. The classification uses given class labels to order the objects in the data collection. Classification approaches normally make use of 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 which is used to classify new objects. Examples of classification method used in data mining application include the classifying of trends in financial markets and the automated identification of objects of interest in large image database. Figure 2.2 shows a simple partitioning of the loan data into two class regions; this may be done imperfectly using a linear decision boundary. The bank may use the classification regions to automatically decide whether future loan applicants will be given loan or not.



*The shaped region denotes class with no loan*

*A Simple Linear Classification Boundary for the Loan Data Set*
**Source:** *Usama, F. et al. (1996)*

## 2. Characterization

Data characterization is also called summarization and involves methods for finding a compact description (general features) for a subject of data or target class, and produces what is called characteristics rules. The data that is 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. A simple example would be tabulating the mean and standard deviations for all fields. More sophisticated methods involve the deviation of summary rules (Usama et al. 1996; Agrawal et al. 1996), multivariate visualization techniques and the discovery of functional relationships between variables. Summarization techniques are often applied to interactive exploratory data analysis and automated report generation (Usama et al., 1996)

## 3. Clustering

Clustering is similar to classification and is the organization of data in classes. But unlike classification, in clustering class tables are not predefined (unknown) and is up to clustering algorithm to discover acceptable classes. Clustering can also be referred to as unsupervised classification because the classification is not dictated by given class tables. We have so many clustering approaches which are all based on the principle of maximizing the similarity between objects in the same class (that is intra-class similarity) and minimizing the similarity between objects of different classes that is inter-class similarity.

## 4. Prediction (Regression)

This involves learning a function that maps a data item to a real–valued prediction variable. This method has attracted considerable attention given the potential implication of successful forecasting in a business context. Predictions can be classified into two major types namely: one can either try to predict some unavailable data value or pending trends, or predict a class label for some data (this is tied to classification). The moment a classification model is built based on 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 often refers to forecast of missing numerical value, or increase/decrease trends in time related data. Summarily, the main idea of prediction is to use a large number of past values to consider probable future values.

### 5. Discrimination

Data discrimination generates what we call discriminant 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 instance, we may want to characterize the rental customers that regularly rent more than 50 movies last year with those whose rental account is lower than 10. The techniques used for data discrimination are similar to that used for data characterization with the exception that data discrimination results include comparative measures.

### 6. Association Analysis

Association analysis is the discovery of what we commonly refer to as 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 that 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 because it searches for relationship between variable. For example, a supermarket might gather data of what each customer buys. With the use of association rule learning, the supermarket can work out what products are frequently bought together, which is useful for marketing purposes. This is sometimes called market basket analysis.

### 7. Outlier Analysis

This is also referred to as exceptions or surprises. Outliers are data elements that cannot be grouped in a given class or clusters, and often important to identify, though, outliers can be considered noisy and discarded in some applications. They can reveal important knowledge in other domains; this makes them very significant and their analysis valuable.

### 8. Evolution and Deviation Analysis

Evolution and deviation analysis deals with the study of time related data that changes in time. In actual sense evolution analysis models evolutionary trends in data that consent with characterizing, comparing, classifying or clustering of time related data. While deviation analysis is concerned with the differences between measured values and expected values, and attempts to find the cause of the deviations from the expected values.