Chapter 7: Knowledge Discovery Systems: Systems that Create Knowledge

Knowledge Discovery

- refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process.
- Knowledge discovery is an activity that produces knowledge by discovering it or deriving it from existing information.
- Knowledge is then organized by indexing knowledge elements, filtering based on content and establishing linkages and relationship among the elements.
- Subsequently, this knowledge is made available to users for supporting their decision making process.

Knowledge Discovery Systems

- Systems that create knowledge
- support the process of developing new tacit or explicit knowledge from data and information or from the synthesis of prior knowledge. These systems support two KM sub processes associated
- with knowledge discovery: combination, enabling the discovery of new explicit knowledge; and socialization, enabling the discovery of new tacit knowledge.
- rely on mechanisms and technologies that can support the **combination** and the **socialization** processes.
- support the process of developing new tacit or explicit **knowledge** from data and information or from the synthesis of prior **knowledge**.

Mechanisms to Discover Knowledge: using socialization to create new tacit knowledge

- **Socialization** is **a** process of sharing experiences.
- **It** is the synthesis of tacit knowledge across individuals, usually through joint activities rather than written or verbal instructions.
- It creates new tacit knowledge from existing tacit knowledge.
- Typically the new **tacit knowledge** is in a form of shared mental models or technical competences.
- Socialization enables the discovery of tacit knowledge through joint activities between:
    - Masters and apprentices
    - Researchers at an academic conference
- For example by observing **a** colleague the observer can learn **through** imitation or practice.
- Typically the **new tacit knowledge** is **in a** form of shared mental models or technical competences.
- Socialization as a means of knowledge discovery is a common practice at many organizations, pursued either by accident or on purpose.

- Simple discussions over lunch among friends discussing their daily problems often lead to knowledge discovery.
- meetings outside the workplace, meet outside their normal work environment, perhaps at a resort,
- where they are able to discuss their problems in an informal and relaxed environment.
- These meetings serve not only as a medium for creativity to flourish but also to share knowledge and build trust amongst the group members.

Technologies to discover knowledge: Using Data Mining to Create New Explicit Knowledge

- The technologies that enable the discovery of new knowledge uncover the relationships from explicit information. Knowledge discovery technologies can be very powerful for organizations wishing to obtain an advantage over their competition.
- Is discovery by finding interesting patterns in observations, typically embodied in explicit data.
- Another name for Knowledge Discovery in Databases is Data Mining (DM).
- **knowledge discovery in databases (KDD)** is the process of finding and interpreting patterns from data, involving the application of algorithms to interpret the patterns generated by these algorithms.
- Data mining is the process of analyzing data from different perspectives and summarizing it in to useful information.
- Data mining softwares is one of a number of analytical tools for analyzing data.
- A DM software allows users to analyze data from different dimenstions or angles ,categorize it, and summarize the relationships identified.
- Technically, it is the process of finding correlations or patterns among dozens of fields in large relational databases.
- Data mining systems have made a significant contribution in scientific fields for years.
- The recent proliferation of e-commerce applications, providing reams of hard data ready for analysis, presents us with an excellent opportunity to make profitable use of data mining.

Examples of data Mining techniques applications

- **Marketing: Predictive** DM techniques, like artificial neural networks (ANN), have been used for target marketingincluding market segmentation.
- **Direct marketing**: customers are likely to respond to new products based on their previous consumer behavior.
- **Retail:**DM methods have likewise been used for sales forecasting.
- **Market basket analysis:** uncover which products are likely to be purchased together.
- **Banking:** Trading and financial forecasting are used to determine derivative securities pricing, futures price forecasting, and stock performance.
- **Insurance:** DM techniques have been used for segmenting customer groups to determine premium pricing and predict claim frequencies.
- **Telecommunications:** Predictive DM techniques have been used to attempt to reduce churn, and to predict when customers will attrition to a competitor.
- **Operations management:** Neural network techniques have been used for planning and scheduling, project management, and quality control.

Data Mining Techniques

1. Predictive Techniques

- **Classification**: Data mining techniques in this category serve to classify the discrete outcome variable.
- **Prediction or Estimation**: DM techniques in this category predict a continuous outcome (as opposed to classification techniques that predict discrete outcomes).

2. Descriptive Techniques

- **Affinity or association**: Data mining techniques in this category serve to find items closely associated in the data set.
- **Clustering**: DM techniques in this category aim to create clusters of input objects, rather than an outcome variable.

Designing the Knowledge Discovery System: Cross-Industry Process for Data Mining(CRISP DM)

- Discovering knowledge can be different things for different organizations.
- Some organizations have large databases, while others may have small ones. The problems faced by the users of data mining systems may also be quite different.
- Therefore, the developers of DM software face a difficult process when attempting to build tools that are considered generalizable across the entire spectrum of applications and corporate cultures.
- Early efforts to apply data mining in business operations faced the need to learn, primarily via trial and error, how to develop an effective approach to DM.
- In fact, as early adopters of DM observed an exploding interest in the application of techniques, the need to develop a standard process model for KDD became apparent.
- This standard should be well-reasoned, nonproprietary, and freely available to all DM practitioners.

- In 1999, a consortium of vendors and early adopters from Germany, Netherlands and England on DM applications for business operations—developed a set of specifications called **Cross-Industry Standard Process for Data Mining (CRISP-DM).**
- CRISP-DM is an industry consortium that developed an industry-neutral and tool-neutral process for data mining.
- ***CRISP-DM***, is an open standard process model that describes common approaches used by data mining experts.
- is an industry-proven way to guide your *data mining* efforts.
- it is a set of guidelines to help plan, organize, and execute your *data mining* or data analysis project.
- The *CRISP-DM* methodology provides a structured approach to planning a data mining project.
- As a methodology, *CRISP* includes descriptions of the typical phases of a project, the tasks involved with each phase, and an explanation of the relationships between these tasks.

CRISP-DM defines a hierarchical process model that defines the basic steps of data mining for knowledge discovery as follows:

1. **Business Understanding:** To obtain the highest benefit from data mining, there must be a clear statement of the business objectives.

- The first requirement for knowledge discovery is to **understand the business** problem.
- In other words to obtain the highest benefit from data mining, there must be a clear statement of the business objectives. For example, a business goal could be "to increase the response rate of direct mail marketing."
- An economic justification based on the return of investment of a more effective direct mail marketing may be necessary to justify the expense of the data mining study.

2.Data Understanding

- One of the most important tenets in data engineering is "know thy data."
- Knowing the data well can permit the designer to tailor the algorithm or tools used for data mining to his/her specific problem.
- This maximizes the chances for success as well as the efficiency and effectiveness of the knowledge discovery system.
- This step, together with preparation and modeling, consumes most of the resources required for the study.

- In fact, data understanding and preparation may take from 50 percent to 80 percent of the time and effort required for the entire knowledge discovery process.
- Typically, data collection for the data mining project requires the creation of a database, although a spreadsheet may be just as adequate.

The steps required for the data understanding process are:

i. Data Collection

- This step defines the data sources for the study, including the use of external public data (e.g., real estate tax folio) and proprietary databases (e.g., contact information for businesses in a particular zip code).
- The **data collection** report typically includes the following: a description of the data source, data owner, who (organization and person) maintains the data, cost (if purchased), storage format and structure, size (e.g., in records, rows, etc.), physical storage characteristics, security requirements, restrictions on use, and privacy requirements.

ii. Data Description

- This step describes the contents of each file or table. Some of the important items inthis report are number of fields (columns) and percent of records missing.