CHAPTER-1

INTRODUCTION

1.1 Data Mining

Data mining, the extraction of hidden predictive information from large databases, is a powerful technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions.

The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line.

1.2 The Foundations of Data Mining

Data mining techniques are the result of a long process research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time.

Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Steps in evolution of data mining is as below,

Table 1.1 Steps in the Evolution of Data Mining

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What’s likely to be monthly Boston unit sales	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups	Prospective, proactive information delivery

Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature are,

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Commercial databases are growing at unprecedented rates. The accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods.

In the evolution from business data to business information, each new step has built upon the previous one. For example, dynamic data access is critical for drill-through in data navigation applications, and the ability to store large databases is critical to data mining. From the user’s point of view, the four steps listed in Table 1 were revolutionary because they allowed new business questions to be answered accurately and quickly.

The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments.

1.3 Overview Of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- **Automated prediction of trends and behaviours**.

  Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

- **Automated discovery of previously unknown patterns**.

  Data mining tools sweep through databases and identify previously hidden patterns in one step. Pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

  Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

  The most commonly used techniques in data mining are:

  - **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
- **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- **Nearest neighbour method**: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ^{1}). Sometimes called the k-nearest neighbour technique.
- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms.

1.4 Recommender System

Recommender system is now popularly getting noticed by research people. Among them, business recommender system is a new upcoming research areas. Technology’s rapid development shares business based and location based data about a person. Grouping of both the data’s will yield a new data called economy-spatial data. This data can be used in disaster rescue, activity planning, geo-crowd sourcing, spatial task outsourcing, business plans recommendation and travel package recommendation.

Though there were many areas in which economy-spatial data has its roots, business plans recommendation is an area which not only benefits an individual but also the society. The Economy-spatial data can be used for improving business in a location. This in turn, increases the social development.

Business is a competitive market where each follows some strategy to improve their income. All companies have their own department for not only improving their business, but also to analyse their business growth. But, for an individual either by own, thinks about an idea for business and starts collecting information from various resources or gets idea from family, neighbours ,well wishers and starts collecting the information. After collecting the required information, analyzes the feasibility. If it is feasible to carry out, starts planning and works out to make that idea into a realistic one.

Irrespective of whether the idea is own or from social circle, the individual has to do some case study in the idea what they have. Then they have to analyse the constraints, possibilities, permissions to obtain from government, competitors, success of business in the location where they wish to start and this to be factors increases largely by depending on an individual’s thinking. To collect resources and to start analysing feasibility, it consumes more time. There are lot of factors which are to be considered before making a decision to start a business in a particular area.

The system that we propose is to solve this issue. It is to decrease the time consumption for feasibility analysis and to provide required guidelines to obtain necessary information for starting the business.