CHAPTER-30
Additional Themes on Data Mining

30.1 Visual and Audio Data Mining
30.2 Scientific and Statistical Data Mining
30.3 Social Impacts of Data Mining
30.4 Review Questions
30.5 References.

30.Additional Themes on Data Mining

Due to the broad scope of data mining and the large variety of data mining methodologies, not all of the themes on data mining canbe thoroughly covered here.

30.1 Visual and Audio Data Mining

Visual data mining discovers implicit and useful knowledge from large data using data and/or knowledge visualization techniques. The eyes and brain, the latter of which can be thought of as a powerful, highly parallel processing and reasoning engine containing a large knowledge base, control the human visual system. Visual data mining essentially combines the power of these components making it a highly attractive and effective tool for the comprehension of data distributions, patterns, clusters, and outliers in data.

Visual data mining can be viewed as an integration of two disciplines: visualization and data mining. It is also closely related to computer graphical multimedia systems, human computer interfaces, pattern recognition, and high performance computing. In general, data visualization and data mining can be integrated in the following ways:

Data visualization:

Data in a database or data warehouse can be viewed different levels of granularity or abstraction, or as different combination attributes or dimensions. Data can be presented in various visual forms, as box plots, 3-D cubes, data distribution charts, curves, surfaces, link graph and so on. Visual display can help give users' a clear impression and overview of the data characteristics in a database.

Data mining result visualization:

Visualization of data mining results is presentation of the results or knowledge obtained from data mining in via forms. Such forms may include scatter plots and boxplots (obtained in descriptive data mining), as well as decision trees, association rule clusti outliers, generalized rules, and so on. For example, scatter plots are shown in

Data mining process visualization:

This type of visualization presents the various processes of data mining in visual forms so that users can see how the data are extracted and from which database or data warehouse they are extracted, as well as how the selected data are cleaned, integrated, processed, and mined. Moreover, it may also show which method is selected for data mining, where the results are stored, and how they may be viewed.

Interactive visual data mining:

In (interactive) visual data mining, visualization tools can be used in the data mining process to help users make smart data mining decisions. For example, the data distribution in a set of attributes can be displayed using colored sectors or columns (depending on whether the whole space is represented by either a circle or a set of columns). This display may help users determine which sector should first be selected for classification and where a good split point for this sector may be an interesting alternative to visual mining.

Audio data mining:

Uses audio signals to indicate the patterns of data or the features of data mining results. Although visual data mining may disclose interesting patterns using graphical displays, it requires users to concentrate on watching patterns and identifying interesting or novel features within them. This can sometimes be quite tiresome. If patterns can be transformed into sound and music, then instead of watching pictures, we can listen to pitches, rhythms, tune, and melody in order to identify anything interesting or unusual. This may relieve some of the burden of visual concentration and be more relaxing than visual mining in many cases. Therefore, audio data mining can be an interesting alternative to visual mining.

30.2 Scientific and Statistical Data Mining

The data mining techniques described in this book are primarily database oriented, that is, designed for the efficient handling of huge amounts of data tin are typically multidimensional and possibly of various complex types. There are however, many well-established statistical techniques for data analysis, particularly for numeric data. These techniques have been applied extensively to scientific data (e.g., data from experiments in psychology, medicine, electrical engineering and manufacturing), as well as to data from economics and the social sciences.

Some of these techniques, such as principal component analysis, regression, and clustering, have already been addressed in this book. A thorough discussion of major statistical methods for data analysis is beyond the scope of this work; however, several methods are mentioned below for the sake of completeness.

Regression:

In general, these methods are used to predict the value of a response (dependent) variable from one or more predictor (independent) variables where the variables are numeric. There are various forms of regression, such as, linear, multiple, weighted polynomial, non-parametric and robust (where robust methods are useful when errors fail to satisfy normalcy conditions or when the data contain significant outliers).

Generalized linear models:

These models and their generalization (generalized additive models), allow a categorical response variable (or some transformation of it) to be related to a set of predictor variables in a manner similar to the modeling of a numeric response variable using linear regression. Generalized linear models include logistic regression and Poisson regression.

Regression trees:

These can be used for classification and prediction. The trees constructed are binary. A regression tree is similar to a decision tree in the sense that tests are performed at the internal nodes. A major difference is at the leaf level-while in a decision tree a majority voting is performed to assign a class label to the leaf, in a regression tree the mean of the objective attribute is computed and used as the predicted value.

Analysis of variance:

These techniques analyze experimental data for two or more populations described by a numeric response variable and one or more categorical variables (factors). In general, an ANOVA (single-factor analysis of variance) problem involves a comparison of population or treatment means to determine if at least two of the means are different. More complex ANOVA problems also exist.

Mixed-effect models:

These models are for analyzing grouped data-data that can be classified according to one or more grouping variables. They typically describe relationships between a response variable and some covariates in data grouped according to one or more factors. Common areas of application include multilevel data, repeated measures data, block designs, and longitudinal data.

Factor analysis:

This method is use& to determine which variables are combined to generate a given factor. For example, for many psychiatric data it is not possible to measure a certain factor of interest directly (such as intelligence); however, it is often possible to measure other quantities (such as student test scores) that reflect the factor of interest. Here, none of the variables are designated as dependent.

Discriminant analysis:

This technique is used to predict a categorical response variable. Unlike generalized linear models, it assumes that the independent variables follow a multivariate normal distribution. The procedure, attempts to determine several discriminant functions (linear combinations of the independent variables) that discriminate among the groups defined by the response variable. Discriminant analysis is commonly used in the social sciences.

Time series:

These are many statistical techniques for analyzing time-series data, such as autoregression methods, univariate ARIMA (autoregressive integrated moving average) modeling, and long-memory time-series modeling.

Survival analysis:

Several well-established statistical techniques exist for survival analysis, which originally were designed to predict the probability that a patient undergoing a medical treatment would survive at least to time r. Methods for survival analysis, however, are also commonly applied to manufacturing settings to estimate the life span of industrial equipment. Popular methods include Kaplan-Meier estimates of survival, Cox proportional hazards regression models, and their extensions.

Quality control:

Various statistics can be used to prepare charts for quality control, such as Shewhart charts and cusum charts (both of which display group summary statistics). These statistics include the mean, standard deviation, range, count, moving average, moving standard deviation, and moving range.

30.3 Social Impacts of Data Mining

With the fast computerization of society, the social impacts of data mining should not be underestimated. Is data mining a hype, or is it really here to stay? What obstacles must be met in order for data mining to become accepted as a mainstream technology for business, and eventually, for everyone's personal use? What can be done toward protecting data privacy and security? This section addresses each of these questions.

Is Data Mining Hype or a Persistent, Steadily Growing Business?

Data mining has recently become very popular, with many people jumping into data mining research, development, or business, or claiming their software systems to be data mining products. Observing this, you may wonder, "Is data mining a hype, or is it here to stay? How well accepted is it, as a technology?"

Granted, there has been a great deal of hype regarding data mining since its emergence during the late 1980s, especially because many people expect that data mining will become an essential tool for deriving knowledge from data, to help business executives make strategic decisions, to sharpen the competitive edge of a business;, and do many other wonderful things.

Data mining is a technology. Like any other technology, data mining will require time and effort to research, to develop, and to mature, and its adoption will, likely go through a life cycle consisting of the following stages.

- **Innovators:** The new technology starts to take form as researchers begin to realize the need for methods to solve a particular (possibly new) problem. Early adopters: Interest increases as more and more methods for the technology are proposed.

- **Chasm:** This represents the "hurdles" or challenges that must be met before the technology can become widely accepted as mainstream.

- **Early majority:** The technology becomes mature and is generally accepted and used.

- **Late majority:** The technology is well accepted, but interest in it declines as the initial problem either becomes less important or is replaced, by other needs.

- **Laggards:** Use of the technology stars to die out, as it becomes old and outdated.

“So, at what stage is data mining?” Several recent discussions have placed data mining at a chasm. In order for data mining to become fully accepted, as a technology, further research and developments are needed in the many areas mentioned as-efficiency and scalability, increased user interaction, incorporation of background knowledge and visualization techniques, the evolution of a standardized data mining query language, effective methods for finding interesting patterns, improved handling of complex data types, Web mining, and so on.

For data Mining, to "climb out" of the chasm, we also need to focus on the integrated of data mining into existing business technology. Currently, there exists a good variety of generic data mining systems.

However, many of these tend to be designed for specifically trained experts who are familiar with data mining jargon and data analysis techniques, like association, classification, and clustering. This makes such systems difficult to use for business executives and the general public. Moreover, these systems tend to be designed to provide horizontal solutions that are geared to work for all kinds of business but are not specially designed to provide business-specific data mining solutions. Since effective data mining requires the smooth integration of business logic with data mining functions, one cannot expect that generic data mining systems can achieve as great a success in business intelligence as domain-independent relational database systems have done in business transaction and query processing.

Many data mining researchers and developers believe that a promising direction for data mining is to construct data mining systems that provide vertical solutions, that is, the integration of in-depth domain-specific business logic into data mining systems. Business conducted on the Web, or e-commerce, is an obvious venue for data mining, as more companies collect large amounts of data from e-stores set up on the Web (also called Web stores). We will therefore examine how to provide domain-specific data mining solutions for e-commerce applications.

Currently, more tailored systems are required that facilitate marketing campaign management (often called e-marketing). Ideally, such closed-loop systems bring together customer data analysis (with OLAP and mining technologies embedded under a user-friendly interface), customer profiling (or one-to-one segments'), campaign roll-out, and campaign analysis.

These systems increasingly use data mining for customer relationship management (CRM), which helps companies to provide more customized, personal service to their customers in lieu of mass marketing. By studying browsing and purchasing patterns on Web stores (e.g., by analyzing click streams, the information that consumers provide by clicks of the mouse), companies can learn more about individual customers or customer groups. This information can be applied to the benefit of both the company and the customer involved. For example, by having more accurate models of their customers, companies should gain a better understanding of customer needs. Serving these needs can result in greater success regarding cross-selling of related products, up-selling, one-to-one promotions, product affinities, larger baskets, and customer retention. By tailoring advertisements and promotions to customer profiles, customers are less likely to be annoyed with unwanted mass mailings or junk mail. These actions can result in substantial cost savings for companies. The customer further benefits in that she is more likely to be notified of offers that are actually of interest to her, resulting in less waste of personal time and greater satisfaction. Customer-tailored advertisements are not limited to company mail-outs or ads planed on Web stores: In the future, digital television and on-line books and newspapers may also provide advertisements that are designed and selected specifically for the given viewer or viewer group based on customer profiling information and demographics. It is important to note that- data mining is just one piece of the integrated solution. Other components, such as data cleaning and data integration, **OLAP**, user security, inventory and order management, product management, and so on. Must also be in place.

Is Data Mining Merely Managers Business or Everyone's Business?

Data mining will surely help company executives a great deal in understanding the market and their business. However, "is data mining merely managers' business or every ones business?" Since more and more data are being made available on the Web or possibly on your own disks, it is likely you will need data mining to understand the data you can access to benefit your work and daily life. Moreover, in the years to come, it is expected that more and more powerful, user-friendly, diversified, and affordable data mining systems or components will be made available. Therefore, one can expect that everyone will have needs and the means for data mining. In other words, it is unlikely that data mining will remain reserved for today's traditional knowledge workers consisting of managers and business analysts. Instead, data mining will become increasingly available to everyone.

“But, what could I do at home with data mining?" Data mining can have multiple personal uses- For example, you might like to mine your family's medical history, identifying patterns relating to genetically related medical conditions, such as cancer or chromosome abnormalities. Such knowledge may help in making decisions about your lifestyle and health. In the future, you may be able to mine the records of the companies you deal with in order to evaluate their service to you as a customer, or to choose the best companies to deal with, based on customer service. You could apply content-based text mining to search your e-mail messages, or automatically create a classification system to help organize your archived messages you could mine data on stocks and company performance to assist in your financial investments. Other examples include mining Web stores to find the best deal on a particular item or type of vacation. Thus, as data mining crosses the chasm and becomes more affordable, and with the. increased availability of personal computers and data on the Web, it is expected that data mining will become increasingly accessible to the general public and will eventually become a handy tool for everyone.

Is Data Mining a Threat to Privacy and Data Security?

With more and more information accessible in electronic forms and available on the Web, and with increasingly powerful data mining tools being developed and put into use, you may wonder, "is data mining a threat to my privacy and information security?" Like any other technology, data mining can be used for good or bad. Since data mining may disclose patterns and various kinds of knowledge that are difficult to find otherwise, it may pose a threat to privacy and information security if not done or used properly.

Most consumers don't mind providing companies with personal information if they think it will enable the companies to better service their needs. For example, shoppers are usually happy to sign up for loyalty cards at.the local supermarket if it means they can get discounts in return,

Have you ever stopped to think about just, how much information is recorded about you, and what that information says? Profiling information can be collected every time you use your credit card, debit card, supermarket loyalty card, or frequent flyer card, or apply for any of the above. It can be collected when you surf the Web, reply to an Internet newsgroup, subscribe to a magazine, rent a video, join a club, fill out a contest entry form, give information about your new baby (in order to receive coupons, free samples, or gifts), pay for prescription drugs, or present you medical care number when visiting the doctor. Clearly, the information that can easily be collected is not limited to our retail purchasing behavior, but may even reflect our hobbies as well as financial, medical, and insurance data. If you stop to think about this the next time you do any of the above actions, you may get the feeling that “Big Brother" or “Big Banker" or “Big Business” is carefully watching you.

While the collection of our personal data may prove beneficial for companies and consumers, there is also potential for its misuse. What if the data are used for other purposes such as, say, to help insurance companies determine your level of fat consumption based on the food items you purchase? One supermarket recently tried to use loyalty-card data to show that a shopper who slipped and fell was actually a heavy drinker (based on the amount of alcohol purchases). Although the case was dropped, it illustrates how data that are “invisibly” collected on consumers may be used against them.

While pondering the above, you may wonder:

“When I provide a company with information about myself, are these data going to be used in ways I don't expect?"

“Will the data be sold to other companies"?

"Can I end out what is recorded about me?"

"How can I find out which companies have information about me?"

"Do I have the right or the means to refuse companies to use the profiling information they have about me?"

"Are there any means set up by which I can correct any errors in the profile data recorded about me? What if' I Want to erase, complete, amend, or update the data?"

"Will the information about me be `anonymized,’ or will it be traceable to me?"

"How secure are the data?"

“How accountable is the company who collects or stores my data, if these data are stolen or misused?”

There are no easy answers to these questions. International guidelines, known as fair information practices, were established for data privacy protection and cover aspects relating to data collection, use, quality, openness, individual participation, and accountability. They include the following principles:

Purpose specification and use limitation:

The purposes for which personal data are collected should be specified at the time of collection, and the data collected should not exceed the stated purpose. Data mining is typically a secondary purpose of the data collection. It has been argued that attaching a "disclaimer" that the data may also be used for mining is generally not accepted as sufficient disclosure of intent. Due to the exploratory nature of data mining, it is impossible to know what patterns may be discovered; therefore, there is no certainty over how they may be used.

Openness:

Individuals have the right to know what information is collected about them, who have access to the data and how the data are being used.

One social concern of data mining is the issue of privacy and information security. Opt-out policies, which allow consumers to specify limitations on the use of their personal data, are one approach toward data privacy protection, while data security-enhancing techniques can anonymize information for security and privacy.

30.4 Review Questions

1 Explain about Visual and Audio Data Mining

2 Explain about Scientific and Statistical Data Mining

3 Explain about Social Impacts of Data Mining

30.5 References.

[1]. Data Mining Techniques, Arun k pujari 1st Edition

[2] .Data warehousung,Data Mining and OLAP, Alex Berson ,smith.j. Stephen

[3].Data Mining Concepts and Techniques ,Jiawei Han and MichelineKamber

[4]Data Mining Introductory and Advanced topics, Margaret H Dunham PEA

[5] The Data Warehouse lifecycle toolkit , Ralph Kimball Wiley student Edition