5

Emerging Trends and Applications of Data Mining

Objectives:

- Discuss the technology and enterprise – adoption trends associated with business analysis.
- Recent innovations and trends is business analytics – spanning organizations and technical processes, new technologies, user interface design, and system integration – are all driven by business value.
- We discuss the technology and enterprise – adoption trends associated with business analysis.
- In order to achieve the greatest possible business value, analytic solutions have to produce results that are actionable, along with ways to measure the effects of key changes.
- Data mining applications have proved highly effective in addressing many important business problems.
- We outline the related challenges in several emerging domains.
- Biology is in the midst of a revolution with an unprecedented flood of data forcing biologists to rethink their approach to scientific discovery.
- Data mining flourishes in telecommunications due to the availability of vast quantities of high quality data.
- The scope, coverage and volume of digital geographic data seers have grown rapidly in recent years due to the progress in data collection and data processing technologies.
- Earth science data mining consists of two main components: the modeling of ecological data and the design of efficient algorithms for finding spatiotemporal patterns.
- It is needed to develop highly automated, scalable, integrated, reliable data mining systems, and tools.
- It is important to promote information exchange among users, data analysts, system developers, and data mining researchers to facilitate the advances available from data mining research, application development, and technology transfer.

S. Sumathi and S.N. Sivanandam: *Emerging Trends and Applications of Data Mining*, Studies in Computational Intelligence (SCI) **29**, 165–183 (2006)
www.springerlink.com

Abstract. The field of business analytics has improved significantly over the past few years, giving business users insights, particularly from operational data stored in transactional system. An example is e-commerce data analysis, which has recently come to be viewed as a killer appropriate for the field of data mining. The data sets created by integrating click stream records generated by Web site activity with demographic and other behavioral data dwarf, in size and complexity, the largest data warehouses of a few years ago. The result is massive database requiring a mix of automated analysis techniques and human effort to give business users strategic insight about the activity on their sites, as well as about the characteristics of the sites visitors and customers. With many millions of click stream records generated every day, aggregated to customer focused records with hundreds of attributers, there is a clear need for automated techniques for finding patterns in the data. Here, we discuss the technology and enterprise – adoption trends associated with business analysis.

Also in this section, emerging scientific applications in data mining like biomedical engineering, Geospatial data, and Telecommunications are discussed.

5.1 Emerging Trends in Business Analytics

The key consumer is the business user, whose job possibly in merchandising, marketing, or sales, is not directly related to analytical tools to improve the results of some business process along one or more dimensions (such as profit and time to maker). Fortunately, data mining analytic applications, and business intelligence systems are now better integrated with transactional systems than they were once, creating a closed loop between operations and analysis that allows data to be analyzed and the results reflected quickly in business actions. The mined information today is deployed to a broader business audience taking advantage of business analytics in its everyday activities. Analytics are now routinely used in sales, marketing, supply chain optimization, and fraud detection.

5.1.1 Business Users

Even with these advances, business users, while expert in their particular areas, are still unlikely to be expert in data analysis and statistics. To make decisions based on the data collected by and about their organizations, they must either rely on data analysis to extract information from the data or employ analytic applications that blend data analysis technologies with task – specific knowledge. In the former, business users impart domain knowledge to the analyst then wait for the analyst to organize and analyze it and communicate back the results. These results typically raise further questions; hence several iterations are necessary before business users can actually act on the analysis. In the latter, analytic applications incorporate not only a variety of data-mining techniques but provide recommendations to business users as to how to best analyze the data and present the extracted information. Business

users are expected to use it to improve performance along multiple metrics. Unfortunately, the gap between relevant analytics and users strategic business needs is significant. The gap is characterized by several challenges.

Cycle time. The time needed for the overall cycle of collecting analyzing, and acting on enterprise data must be reduced. While business constraints may impose limits on reducing the overall cycle time, business users want to be empowered and rely less on other people to help with these tasks.

Analytic time and expertise. Within the overall cycle, the time and analytic expertise necessary to analyze data must be reduced.

Business goals and metrics. Unrealistic expectations about data mining "magic" often lead to misguided efforts lacking clear goals and metrics.

Goals for data collection and transformations. Once metrics are identified, organizations must collect and transform the appropriate data. Data analysis is often an afterthought, limiting the possible value of any analysis.

Describing analysis results. Most analysis tools are designed for quantitative analysts, not the broader base of business users who need the output translated into language and visualizations appropriate for business needs.

Integrating data from multiple sources. The extract transform load (ETL) process is typically complex and its cost and difficulty are often underestimated.

5.1.2 The Driving Force

The emerging trends and innovations in business analytics embody approaches to these business challenges. Indeed, it is a very healthy sign for the field that regardless of the solution – process, technology, system integration, or user interface, business problems remain the driving force.

"Verticalization"

In order to reduce discovery cycle time, facilitate the definition and achievement of business goals, and deploy analysis results to a wider audience, developers of analytics solutions started vertical zing, their software, or customizing applications within specific industries. The first step toward verticalization is to incorporate task – specific knowledge; examples include: knowledge about how to analyze customer data to determine the effectiveness of a marketing campaign; knowledge of how to analyze click stream data generated by a web site to reduce shopping cart abandonment and improve ad effectiveness; knowledge about how an investment bank consolidates its general ledger and produces various types of forecasts; and how an insurance company analyzes data in order to provide an optimally priced policy to an exiting customer.

In the process of incorporating industry-specific knowledge, companies are also able to optimize the performance of their applications for specific industries. For example, a company that developed an analytic application for budgeting and forecasting targeted at the financial services industry determined that its online analytical processing, or OLAP, engine's execution speed could be optimized by limiting to nine the number of dimensions it had to handle, a number deemed sufficient for the particular application in that industry.

The use of industry – specific knowledge is not limited to the data mining components of analytics applications but also effects how the extracted information is accessed and presented. For example, organizations in the financial services, retail, manufacturing, utilities, and telecommunications industries increasingly want their field personnel to have access to business analytics information through wireless devices. Analytics application vendors are now developing technologies to automatically detect wireless devices and their form factors, automatically tailoring analysis results to fit the capabilities of a particular device. For example, if the information is to be displayed on a phone supporting the Wireless Access Protocol (implying small screen size), it may be necessary to automatically summarize text, abbreviate words, and limit the use of graphics by automatically selecting only the most relevant figures.

Comprehensible Models and Transformations

In light of the need to let the business users analyze the data and quickly gain insight, and aiming for the goal of reducing reliance on data mining experts, comprehensible models are more but popular than the opaque models. For example, in KDD-cup 2000 a data mining competition in which insight was important, the use of decision trees, generally accepted as relatively easy to understand, outnumbered the other methods more than two to one. Business users do not want to deal with advanced statistical concepts; they want straightforward visualizations and task relevant outputs.

Part of the Larger System

The needs of data analysis are being designed into systems, instead of being an afterthought, typically addressing the following area:

Data Collection: We cannot analyze what we do not collect, so collection of rich data is critical. For example, e-commerce systems can collect attributes ranging from the users local time, screen resolution (useful for determining the quality of images to send), and network bandwidth.

Generation (and Storage) of Unique Identifiers: In order to help merge information from several records and remove duplicate records, systems must generate unique keys to join data and store them. For example, all click stream records in the same session should store the session IDs so they can be joined later to session records stored in other tables.

Integration with Multiple Data Sources: Analysis is more effective when data is available from multiple sources. For example, in customer analytics, data should be merged from multiple touch points, including the web, call centers, physical store, wireless access, and ad campaigns (both direct and online). Behavioral data can be more powerful when overlaid with demographic and socioeconomic data from other sources.

Hardware Sizing: Analysis requires hardware capable of dealing with large amounts of data. Some organizations have traditionally underestimated the need for sophisticated IT infrastructure and the hardware needed to make timely analysis feasible.

In New Areas

During the past few years, recognition of the strategic value of business analytics has led to significant development in business application that analyzes the customer data. They have been used to reduce customer attrition, improve customer profitability, increase the value of e-commerce purchases, and increase the response of direct mail and email marketing campaigns.

This success has paved the way for the new applications; three are particularly promising; supply chain visibility, price optimization, and work force analysis. Organizations have automated portions of their supply chains; enabling collection of significant data about inventory, supply performance and logistics of materials, and finished goods. Newer applications analyze this data to provide insights about the performance of the suppliers and partners, material expenditures, accuracy of sales forecasts for controlling materials inventory, accuracy of production plans, and accuracy of plans for order delivery.

The wide adoption of customer relationship management or CRM, and supply chain management software has allowed enterprises to fully interface and integrate their demand and supply chains. Based on this integration they are better able to capture up to the minute data about demand for a particular product as well as data of similar granularity about the supply of corresponding data. Analyzing these two data streams, organizations optimize the price of a particular product along several dimensions so demand meets available supply; for example, the price of a product may be different through one channel (such as the web) than through another (such as retail store). Price optimization allows any type of organization to maximize the profit margins for each item sold while reducing inventory.

Once organizations are able to analyze data about their customers and their suppliers, they begin analyzing data about their employees too. A new generation of analytic applications allows enterprises to identify work force trends (such as attrition rates) and perform HR management tasks (such as compensation and benefits analyses). Companies whose cost or revenue model is dependent on hourly models (such as contact centers and systems integrators) use it to optimize staffing levels and skill requirements while minimizing the number of employees who are not able to bill.

Integration with action and measurement. With increased understanding of and experience in analytics, business users become more demanding and discerning, particularly when it comes to action based on insight and return on investment (ROI). Increasingly, analytics users ask two key questions: How do we turn discovered information into action? And how can we determine the effect of each action on the organization's business performance? Tales of data mining applications used to end with some novel analytical result; today, however, it is increasingly necessary that solutions use analytic results as a starting point toward the critical next steps of action and measurement. It is no longer enough for, say, cluster discovery algorithms to uncover interesting groups of customers. The successful analytic solution must make it easier for the user to grasp the significance of these clusters in the context of a business action plan; for example, these people have a propensity for purchasing new fashions. Achieving these results requires nontrivial transformations from the base statistical model. Traditionally, achieving these results necessitated the participation of expert human analysts.

Integration analytics with existing systems is a key to both action and measurement. For example, if the analytic application identifies customers likely to respond to promotion, but it takes a cadre of IT specialists to incorporate the relevant data into the advertising system to run the promotion, the results are unlikely to be used, as IT specialists are likely to be in short supply. Similarly, if promotion-targeting solutions enable distribution of catalogs with optimized promotions, but the order submission system is not closely tied back into the customer analytics, the resulting lag in ROI reports inhibits timely adjustment in the next catalog mailing. Efforts to integrate operations and analytic systems have seen major initiatives over the past five years including entire product lines whose value proposition is the optimization of the collect–analyze–act–measurement cycle.

Broadening the effects of analytics in the business process, solutions go beyond customer-centric applications to support sales, marketing, supply chain visibility, price optimization, and work force analysis. Finally, in order to achieve the greatest possible business value, analytic solutions have to produce results that are actionable, along with ways to measure the effects of key changes.

5.2 Business Applications of Data Mining

The traditional approach to data analysis for decision support has been to couple domain expertise with statistical modeling techniques to develop handcrafted solutions for specific problems. More recently, several trends have emerged to challenge this approach. One is the increasing availability of large volumes of high-dimensional data occupying database tables with millions of rows and thousands of columns. Another is the competitive demand for the rapid construction and deployment of data-driven analytics. Third is the

need to give end users analysis results in a form they readily understand and assimilate, helping them gain the insights they need to make critical business decisions. Moreover, knowledge discovery in databases (KDD) techniques emphasizing scalable, reliable, fully automated, explanatory structures have shown that in data analysis, such structures supplement, and sometimes supplant, existing human-expert-intensive analytic techniques for improving decision quality.

Measurable Benefits

KDD applications deliver measurable benefits, including reduced cost of doing business, improved profitability, and enhanced quality of service. Industries, in which such benefits have been demonstrated, include insurance, direct-mail marketing, telecommunications, retail, and health care.

Risk management and targeted marketing. Insurance and direct mail are two industries that rely on data analysis to make profitable business decisions. For example, insurers must be able to accurately assess the risks posed by their policyholders to set insurance premiums at competitive levels. For example, overcharging low-risk policyholders would motivate them to seek premiums elsewhere; undercharging high-risk policyholders would attract more of them due to lower premiums. In either case, costs would increase and profits inevitably decrease. Effective data analysis leading to the creation of accurate predictive models is essential for addressing these issues.

In direct-mail targeted marketing, retailers must be able to identify subsets of the population likely to respond to promotions in order to offset mailing and printing costs. Profits are maximized by mailing only to those potential customers most likely to generate net income to a retailer in excess of the retailer's mailing and printing costs.

Businesses relying on data-driven analysis for decision making typically construct data warehouses to capture as much information as possible about their customers. Examples of such information include details of past customer transactions, as well as additional information obtained from third-party data providers, including credit scores and demographics, for targeted marketing purposes and motor vehicle records for insurance purposes.

To aid decision-making, analysts construct predictive models using warehouse data to predict the outcomes of variety of decision alternatives. For example, in order to set policy premiums, insurers need to predict the cost of claims filed by policyholders annually, what is known about each policyholder. In order to select customers for a targeted marketing campaign, retailers need to predict revenues or gross profits that would be generated the customers receiving the mailings.

A popular approach to predictive modeling used by many data analysts and applied statisticians involve partitioning the data records for a population of customers (or other entities) into segments, then developing separate predictive models for each segment. Typically, data is partitioned through

a combination of domain knowledge, simple heuristics, and clustering algorithms. Predictive models are constructed once segments are identified. The drawback is that this sequential approach ignores the strong influence segmentation exerts on the predictive accuracies of the models within each segment. Good segmentations tend to be obtained only through trial and error by varying the segmentation criteria.

A better approach is to simultaneously perform segmentation and predictive modeling within each segment, optimizing the segmentation so as to maximize the overall predictive accuracy of the resulting model. This approach is built into the IBM Probabilistic Estimation (ProbE) data mining server, making it possible to automatically construct high-quality segmentation-based predictive models from very large high-dimensional data sets. A top down tree-based algorithm is used to construct the segmentations. A collection of other algorithms is incorporated for constructing segment models, including stepwise linear regression and stepwise native Bays algorithms for general-purpose modeling and a joint Poisson/log-normal algorithm for insurance risk modeling. A key feature of the ProbE server is it is readily extended to incorporate different types of predictive modeling algorithms for the segments, as well as different types of segmentation algorithms.

Two different client applications have been developed by IBM's Data Abstraction Research Group that utilizes the ProbE data-mining server. One is called IBM Advanced Targeted Marketing for Single Events (ATM-SE), built jointly with the Business Intelligence group at Fingerhut, Inc., a large US catalog and Internet retailer based in Minnetonka, MN, for constructing customer-profitability and response likelihood models for targeted marketing in the retail industry. The other is the IBM Underwriting Profitability Analysis (UPA) application, co-developed with Farmers Insurance Group, a large automobile and home insurance company based in Los Angles, for discovering homogeneous insurance risk groups.

Fingerhut's 2000 evaluation of the ATM-SE application for direct-mail response modeling demonstrated the application produced segmentation-based response models that either equaled or slightly outperformed Fingerhut's own proprietary models. This evaluation was significant because numerous vendors and consultants had previously failed to beat Fingerhut's in-house modeling capability. If these results ultimately hold across all of Fingerhut's models, the ATM-SE models would yield an estimated increase in annual profits directly to Fingerhut of more than $1 million. Moreover, the ProbE server achieved its result in a fully automated mode of operation, with no manual intervention.

The UPA application configures the ProbE server so as to use a point Poisson/log-normal statistical model within each segment to simultaneously model both frequency with which insurance claims are filed by policyholders and the amounts, or severities, of these claims for each segment. Using this class of segment model, the identified segments correspond to distinct risk groups whose loss characteristics, such as claim frequency and severity, are estimated in accordance with standard actuarial practices.

The Farmers Group's 1997 evaluation of the application's ability to analyze insurance policy and claims data for all policyholders in one state involved mining runs for 18 unique combinations of customers with specific insurance products and coverage, including explanatory variables. Each run generated about 40 rules, from which 43 combinations were identified as "nuggets," or previously unknown rules with significant potential value. Six nuggets were selected for a detailed benefits assessment which indicated that implementing just these six is a single state could potentially yield a net profit gained several million dollars in the first year alone.

Although insurers know that drivers of high-performance sports cars are more likely to have accidents than drivers of other types of cars, the UPA fours that if a sports car was not the only vehicle in the household, the accident rate is not much greater than that of a regular car. One estimate determined that "just letting Corvettes and Porsches into [the insurer's] 'preferred premium' plan could bring in an additional $4.5 million in premium revenue over the next two years without a significant rise in claims. Another publicly disclosed nugget related to experienced drivers, who tend to have relatively low case frequencies. However, the UPA also turned up the particular segment of experienced drivers who are unusually accident prone.

ProbE's segmentation-based predictive modeling capability permits construction of mining applications optimized for specific problems. Indications are that the ProbE server can consistently produce high quality on a fully automated basis without requiring costly manual adjustments of the models of the mining parameters by data mining experts. These characteristics will make data mining increasingly attractive to mid-size business as well as to be much lager counterparts.

Customer profiles and feature construction. An important ingredient for obtaining highly predictive models is to use highly predictive features, or attributes and variables, as model input. Although a database might contain sufficient information to construct highly predictive models, it is not always stored in a form that permits the data to be uses directly as input to a model. In such cases, the data must be transformed to obtain accurate models.

Transaction data is notorious for requiring transformation before it can be used for data mining applications. Such data consists of records of pairs of individuals and events. An example is a set of overall items purchased by a customer and grouped into a "market basket." Another is a set of Web pages requested from a Web site by a particular suffers and grouped by session. The ability of companies' worldwide to collect vast amounts of such transaction data has far outpaced their ability to analyze it. Transaction data is especially challenging from data mining respective due to several factors:

Massive numbers of records. Large retail chains generate millions of transaction per day.

Sparseness. A typical basket contains only a small fraction of the total possible number of items; individual customers may have few baskets, perhaps only one.

Homogeneity. Purchasing behavior varies consider holy depending on individual tastes and means, along with individual purchasing patterns over time.

These factors combine to make transaction data highly nontrivial when using traditional data analysis techniques. The related challenges, along with the transaction data itself, motivated much of the early work in data mining, including development of association-rule algorithms for efficiently searching for correlations among items in retail transaction data. While the association-rule approach can be useful for exploratory analysis of transaction data, such as discovering combinations of products purchased together, it is less well suited for predicting individual customer behavior.

A recently developed framework called predictive profiting handles transaction detain predictive modeling. A predictive profile is a model that predicts future purchasing behavior of an individual customer, given historical transaction data for both the individual and for the larger population of all of a particular company's customers. The predictive profiling approach is based on a flexible probabilistic model that works in the following way: Let y be a randomly chosen market, where y is a d-dimensional vector describing how many of each of the d items were purchased in the basket. The high-dimensional joint distribution on baskets $p(y)$, is approximated by a linear combination of K simpler models. Each of the K simpler models in effect captures" prototype combinations" of products in baskets.

In the first phase of modeling, the K prototype combinations are learned from the data through a well-known expectation-maximization procedure for statistical estimation. In the second phase, each customer is "mapped" onto the product space represented by the K prototypes, where the mapping is based on individual past purchasing patterns. The mapping effectively transforms the transaction data for each customer into a set of feature values that are then used to make prediction about future purchasing behavior. The transformation is not defined prior to data mining but is inferred by the mining algorithm.

This model is not designed to capture all aspects of individual customer behavior but to extract useful first-order characteristics in terms of how customers shop. Figure 5.1 outlines how the prototypes are used to support exploratory visualizations of the data, providing an interpretable description of the heterogeneity of customer behavior as reflected by different basket prototypes. The predictive profiling method was tested by the university of California, Irvine, research team on two large real-world transaction data sets collected over several years.

The data sets involved several million baskets and about 500,000 customers. Models were trained using historical data from the early years of each data set, then tested on data from later years, typically using from K = 20

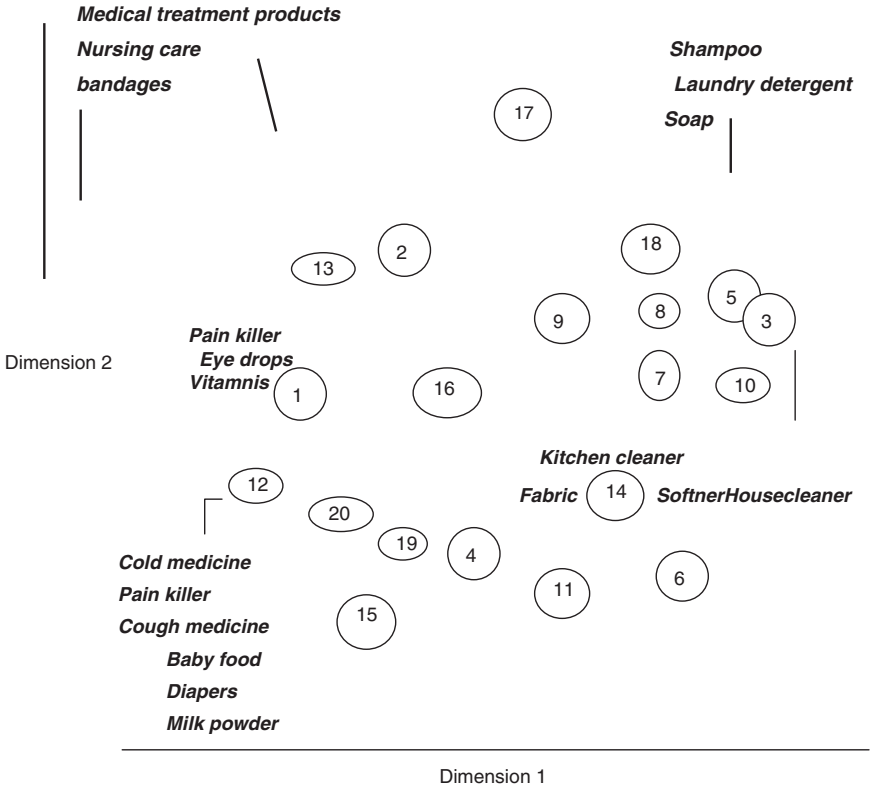

Fig. 5.1. A set of K = 20 prototypes represented here in a 2D space by using multidimensional scaling. The prototype baskets were learned from a set of about six million baskets from a chain of drugstores in Japan. The numbers in each circle refers to different prototypes; the area of each circle represents how likely a randomly chosen basket belongs to that prototype. The names of three items with the greatest lift (defined as p{item\prototype})/p(item)) are also displayed for some of the prototypes. Prototypes close together are also statistically close in the data.

to K = 100 prototypes. The models demonstrated systematic improvement in out of sample predictive performance compared to more standard alternatives. Empirically, the time taken to fit the models was found to scale linearly with both number of baskets and number of fitted prototypes K. The wall clock time to learn all the prototypes and customers profiles took only a few hours on a standard PC.

Such methods for handling transaction data are likely to prove useful across a variety of business applications, including customer segmentation, personalization, forecasting, and change detection, especially in e-commerce environments, where real time modeling of an individual customer and personalized

feedback is valuable. Scalable, robust, and accurate solutions to these problems promise significant economic payoff in the business world.

Medical applications (diabetic screening). Preprocessing and postprocessing steps are often the most critical elements determining the effectiveness of real life data mining applications, as illustrated by the following recent medical application in diabetic patient screening. In 1990s in Singapore, about 10% of the population were diabetic, a disease with many side effects, including increased rescue of eye disease, kidney failure, other complications. However, early detection and proper care management can make a difference in health and longevity of individual sufferers. For example, to combat the disease, the Government of Singapore introduced a regular screening program for diabetic patients in its public hospitals in 1992. Patient information, clinical symptoms, eye disease diagnosis, treatment, and other details were captured in a database maintained by Government medical authorities. Today, after almost 10 years of collecting data, a wealth of medical information is available. This vast store of historical data leads naturally to the application of data mining techniques to discover interesting patterns. The objective is to find rules physicians can use to understand more about diabetic and how it might be associated with different segments of the population.

However, the data miners encountered two major problems. First, the data captured by health clinics turned out to be very noisy; for example, patient's records in the database contained typographical errors, missing values, and incorrect information, including street names and date of birth. Worse, many records contained duplicate data. Cleaning data takes a great deal of effort and time. In addition, many of these records were not in a form suitable for data mining; they had to be transformed to more meaningful attributes before mining could proceed. The second problem was that some state of the art association rule algorithms generate too many rules from the data, no matter how clean it is. Because physicians are busy seeing patients, they cannot take the time to shift through large numbers of rules. It was therefore important to present the discovered rules in some easy to understand form.

To overcome the problem noisy data, a data mining team at National University of Singapore developed a semiautomatic data cleaning system to reconcile database format differences by allowing physicians to specify the mapping between attributes in different format styles and/or different encoding schemes. Reconciling the format differences addressed the problem of identifying and removing duplicate records.

To resolve the problem of too many rules generated by the mining algorithms, the same team developed a user-oriented approach providing step-by-step exploration of both the data and the discovered patterns. Data visualization is used as an integral part of the process to give users a general view of findings. During rule mining, the mining algorithm employs a pruning method to remove insignificant rules. The final rules were also organized into general rules and exceptions to facilitate browsing and analysis. This rule mining approach to organizing mining results is useful to Singapore's medical

authorities because it allows them to view the general patterns that are discovered as well as the detailed patterns. Because it is also a common strategy people employ in everyday learning, the mining results are easy to interpret.

The physicians confirmed that many of the rules and causal relationships the data mining algorithms discovered conformed to the trends they observed in their practices. However, they were surprised by many of the exceptions they did not know before. As a result of data mining they gained a much better understanding of how diabetes progresses over time and how various treatments affect its progress.

Data mining applications have proved highly effective in addressing many important business problems. We expect to see the continued construction and deployment of KDD applications for crucial business decision support systems. Exemplary applications employing data mining analytical techniques will require the KDD technical community to keep improving the underlying techniques for model building and model understanding. The emphasis in model building will be on developing mining techniques that are automated, scalable, and reliable. For domain understanding, the challenge is to keep developing sophisticated techniques that assist users in analyzing discovered knowledge easily and quickly.

5.3 Emerging Scientific Applications in Data Mining

Recent progress in scientific and engineering applications has accumulated huge volumes of high dimensional data, stream data, unstructured and semi-structured data, and spatial and temporal data. Highly scalable and sophisticated data mining tools for such applications represent one of the most active research frontiers in data mining. Here we outline the related challenges in several emerging domains.

5.3.1 Biomedical Engineering

Biology is in the midst of a revolution, with an unprecedented flood of data forcing biologists to rethink their approach to scientific discovery. First, large-scale data-collection techniques have emerged for a number of data sources limited by throughput, of the amount of available data. Examples of the data glut include: systematic genome DNA sequencing of organisms; high throughput determination of small molecule structures, as well as large macromolecular structures (such as proteins, RNA, and DNA); large scale measurements of molecular interactions; and simultaneous measurement expression level of all genes (thousands to tens of thousands) in a population of cells. Second, the availability of this data requires the biologists to create system for organizing, storing, and disseminating it, thus creating a need for standard terminologies and development of standards of interchange and annotations. Third, because of the apparent opportunities for automated learning from the data sets, a

market for robust machine learning and data mining algorithms has emerged to take advantage of previous knowledge without being overly biased in the search for new knowledge. As a result, biology has changed from a field dominated by an attitude of "formulate hypothesis, conduct experiment, evaluate results" to more of a big science attitude of "collect and store data, mine for new hypothesis, confirm with data or supplemental experiment." the long-term significance of new data of molecular biology is that it can be combined with clinical medical data to achieve a higher resolution understanding of the causes for and treatment of disease. A major challenge for data mining in biomedicine is therefore the organization of the molecular data, cellular data, and clinical data in ways allowing them to integrate for the sake of knowledge extraction.

A major additional source of information is the published medical literature, increasingly available online in full text form or as useful (but unstructured) summaries of the main data and biomedical hypothesis.

5.3.2 Telecommunications

Data mining flourishes in telecommunications due to the availability of vast quantities of high quality data. A significant stream of it consists of call records collected at network switches used primarily for billing; it enables data mining applications in toll-fraud detection and consumer marketing.

Perhaps the best-known marketing application of data mining, albeit via unconfirmed anecdote, concerns MCI's "Friends & Family" promotion launched in the domestic US market in 1991. As the anecdote goes, market researchers observed relatively small subgraphs in this long-distance phone company's large call-graph of network activity, revealing the promising strategy of adding entire calling circles to the company's subscriber base, rather than the traditional and costly approach of seeking individual customers one at a rime. Indeed, MCI increased its domestic US market share in the succeeding years by exploiting the "viral" capabilities of calling circles; one infected member causes others to become infected. Interestingly, the plan was abandoned some years later (not available since 1997), possibly because the virus had run its course but more likely due to other competitive forces.

In toll-fraud detection, data mining has been instrumental in completely changing the landscape for how anomalous behaviors are detected. Nearly all fraud detection systems in the telecommunications industry 10 years ago were based on global threshold models; they can be expressed as rule sets of the form "If a customer makes more than X calls per hour to country Y; then apply treatment Z." The placeholders X, Y, and Z are parameters of these rule sets applied to all customers.

Given the range of telecommunication customers, blanket application of these rules produces many false positives. Data mining methods for customized monitoring of land and mobile phone lines were subsequently developed by leading service providers, including AT&T, MCI, and Verizon,

whereby each customer's historic calling patterns are used as a baseline against which all new calls are compared. So, for customers routinely calling country Y more than X times a day, such alerts would be suppressed, but if they ventured to call a different country Y', an alert might be generated.

Methods of this type were presumably in place for the credit card industry a few years before emerging in telecom. But the size of the transaction streams is far greater in telecom, necessitating new approaches to the problem.

It is expected that algorithms based on call-graph analysis and customized monitoring will become more prevalent in both toll-fraud detection and marketing of telecommunications services. The emphasis on so-called "relational data" is an emerging area for data mining research, and telecom provides relational data of unprecedented size and scope.

These applications are enabled by data from the billing stream. As the industry transforms itself from a circuit-switched to a packet-switched paradigm, the data mining community could well experience a dearth of data, since billing is likely to be increasingly insensitive to usage. Moreover, the number of records that could potentially be recorded in a packet-switched network (such as packet headers) is orders of magnitude greater than today's circuit-switched networks are. Thus, unless a compelling business need is identified, the cost of collecting, transmitting, parsing, and storing this data will be too great for the industry to willingly accept. A dearth of data could well spell the end to future significant data mining innovations in telecommunications.

However, this view might yet be altered by the following scenarios:

New network architectures. New-generation network infrastructure will have to adapt to changes in demand yet be more reliable and secure; for example, capacity in mobile networks will have to be assigned dynamically, necessitating development of new data mining techniques for understanding and predicting network load. Similarly, network-intrusion detection will continue to be important to data mining, helping ensure that artificially induced traffic cannot cripple a network.

Mobility and microbilling. In Europe and Japan, merchants use mobile handsets for selling (and billing) a variety of consumer goods, including vending machine purchases and parking tees. Since these consumer activities correspond to "billable events," data will certainly be collected and maintained for such services.

Mobile services. Ease of use is crucial for enticing customers to adopt new mobile services. Data mining will probably play a major role in the design of adaptive solutions enabling users to obtain useful information with relatively few keystrokes.

Homeland security. Federal regulations require US telecommunications companies to maintain call records for two years. With the recent emphasis on homeland security, along with the role telecom data can play identifying and tracking terrorist cells and activities, data will continue to be collected and maintained, even though the records may not be used for billing.

5.3.3 Geospatial Data

The scope, coverage and volume of digital geographic data seers have grown rapidly in recent years due to the progress in data collection and data processing technologies. These data sets include digital data of all sorts, created, processed, and disseminated by government- and private-sector agencies on land use and socioeconomic infrastructure; vast amounts of georeferenced digital imagery and video data acquired through high-resolution remote sensing systems and other monitoring devices; geographic and spatiotemporal data collected by global positioning systems, as well as other position-aware devices, including cellular phones, in-vehicle navigation systems, and wireless Internet clients; and digital geographic data repositories on the Web. Moreover, information infrastructure initiatives, including the US National Spatial Data Infrastructure, facilitate data sharing and interoperability, making enormous amounts of space-related data sharable and analyzable worldwide.

The increasing volume and diversity of digital geographic data easily overwhelm traditional spatial analysis techniques that handle only limited and homogeneous data sets with high-computational burden. To discover new and unexpected patterns, trends, and relationships embedded within large and diverse geographic data sets, several recent studies of geospatial data mining have developed a number of sophisticated and scalable spatial clustering algorithms, outlier analysis techniques, spatial classification and association analysis methods, and spatial data-cleaning and integration tools.

Nevertheless, considering the challenges posed by the already enormous and increasing amount of spatial data, geospatial data mining is in its infancy. Lots of research needs to be done, especially concerning the following pressing issues.

Developing and supporting geographic data warehouses. Although data warehouses are central to the knowledge discovery process, no true geospatial data warehouse exists today. Creating one requires solutions to problems in geographic and temporal data compatibility, including reconciling semantics, referencing systems, geometry, accuracy, and precision. Creating a warehouse might also need to solve the problems of efficient computation of sophisticated spatial aggregations, as well as how to handle spatial-related data streams. However, spatial data warehouses are likely to eventually play an essential role in geospatial information exchanges and data mining, so it is critical that we develop and support such an infrastructure today.

Exploring and mining richer geographic data types. Geographic data sets are moving beyond the well-structured vector and raster formats to include semistructured and unstructured data, especially georeferenced stream data and multimedia data. Techniques have to be developed to handle spatiotemporal data, robust geographic concept hierarchies and granularities, and sophisticated geographic relationships, including non-Euclidean distances, direction, connectivity, attributed geographic space (such as terrain), and constrained interaction structures (such as networks).

Reaching a broader user community. Geospatial data mining needs to go beyond researchers to also deliver its benefits to general users. This requires high-level user interfaces and visualization tools that aid diverse users in geospatial data mining. Moreover, these interfaces and tools have to be integrated with existing geographical information systems and database systems to guide users searching for geographic knowledge, interpreting and visualizing discovered knowledge, and using the discovered geographic knowledge in their decision-making.

5.3.4 Climate Data and the Earth's Ecosystems

The large amount of climate data acquired through NASA's Earth observation satellites, terrestrial observations, and ecosystem models offers an unprecedented opportunity for predicting and preventing future ecological problems by managing the ecology and health of the planet. Such data consists of a sequence of global snapshots of the Earth, typically available at monthly intervals, including various atmospheric, land, and ocean variables (such as sea surface temperature, precipitation, and net primary production, or the net photosynthetic accumulation of carbon by plants). Due to the nature and scale of this data, data mining techniques can play a major role in the automatic extraction and analysis of interesting patterns, thus complementing, existing statistical techniques.

Earth science data mining consists of two main components: the modeling of ecological data and the design of efficient algorithms for finding spatiotemporal patterns. An important goal is the discovery of teleconnection patterns, or recurring and persistent climate patterns spanning vast geographical regions. They manifest themselves as spatiotemporal relationships among ecological variables observed at various locations on the Earth and are critical for understanding how the ecosystem's various elements interact with one another. Clustering techniques, which divide data into meaningful or useful groups, help automate the discovery of teleconnections. Specifically, clustering identifies regions of the Earth whose constituent points have similar short- and long-term climate characteristics. By analyzing correlations among climate variables across these regions, it is possible to rediscover existing patterns (such as the El Nino periodic ocean–atmosphere disruption in the tropical Pacific Ocean), as well as new, previously unrecognized teleconnections. An alternative approach is to convert the time series into sequences of events, then apply existing association-rule techniques to discover interesting patterns in the sequences.

The difficulty of mining Earth science data is illustrated by the following examples of issues arising during the various stages of data mining analysis:

Preprocessing. It is often beneficial to aggregate data into a smaller number of points, easing computational requirements and (typically) reducing the amount of noise. However, it can be difficult for researchers to choose the

proper level of aggregation, since too much limits the patterns that can be directed, while too little results in noisy data in which only the strongest patterns can be discovered. Event definition is another necessary but ill-defined task. In the spatial domain, the problem is too many events, and in the temporal domain, events are rare; for example, El Nino events occur only every four to seven years. Yet another concern is integrating data from heterogeneous sources (such as data covering different time periods). Earlier data may come from manual, earth-based observations, while later data may originate from satellites.

Similarity of time series. The "proper" measure of similarity between time series is fraught with challenges. Linear correlation works well with standard clustering algorithms and lends itself to statistical tests. Nevertheless, alternate measures of time series similarity would be beneficial if they allowed the detection of patterns that could not be detected via linear correlation, and might, for example, be based on either dynamic time warping or cepstral coefficients representing the frequency spectrum of a time series. An "ideal" measure of similarity would account for time lag and the fact that only extreme events are usually correlated.

Identifying interesting patterns. Once patterns are discovered, it is difficult to distinguish the spurious ones from the significant ones. For example, given 40,000 time series recording the sea surface temperature at various points on the ocean's surface and 60,000 time series representing precipitation on land, some of these series might, just by chance, have strong correlations. While a number of statistical approaches estimate significance levels, it is not possible to apply such approaches directly due to spatial and temporal autocorrelation. When genuine patterns are identified, domain-specific knowledge is inevitably still needed to identify patterns of interest to Earth scientists.

5.4 Summary

Recent innovations and trends in business analytics – spanning organizations and technical processes, new technologies, user interface design, and system integration – are all driven by business value. Business value is measured in terms of progress toward bridging the gap between the needs of the business user and the accessibility and usability of analytic tools. In order to make analytics more relevant and tangible for business users, solutions increasingly focus on specific vertical applications tailoring results and interfaces for these users, yielding human-level insight. For ease of use, simpler and more effective deployment, and optimal value, analytics are also increasingly embedded in larger systems. Consequently, data collection, storage, processing, and other issues specific to analytics are incorporated into overall system design. Thus the emerging applications involve great data management challenges.

These emerging applications involve great data-management challenges that also represent new opportunities for data mining research, Methods for mining biomedical, telecommunication, geospatial, and climate data are under active development. However, in light of the tremendous amount of fast-growing and sophisticated types of data and comprehensive data analysis tasks, data mining technology may be only in its infancy, as the technology is still far from adequate for handling the large-scale and complex emerging application problems. Research is needed to develop highly automated, scalable, integrated, reliable data mining systems, and tools, Moreover, it is important to promote information exchange among users, data analysts, system developers, and data mining researchers to facilitate the advances available from data mining research, application development, and technology transfer.

5.5 Review Questions

1. Define cycle time
2. Explain in detail on the driving forces used in the business strategies
3. Justify – KDD applications deliver measurable benefits
4. State in detail on the scientific applications in data mining