UNIT –VI APPLICATION AND TRENDS IN DATA MINING

Spatial Data Mining

A spatial database stores a large amount of space-related data, such as maps, preprocessedremote sensing or medical imaging data, and VLSI chip layout data. Spatialdatabases have many features distinguishing them from relational databases. Theycarry topological and/or distance information, usually organized by sophisticated,multidimensional spatial indexing structures that are accessed by spatial data accessmethods and often require spatial reasoning, geometric computation, and spatialknowledge representation techniques.

Spatial data mining refers to the extraction of knowledge, spatial relationships, orother interesting patterns not explicitly stored in spatial databases. Such mining demandsan integration of data miningwith spatial database technologies. It can be used for understandingspatial data, discovering spatial relationships and relationships between spatialand nonspatial data, constructing spatial knowledge bases, reorganizing spatial databases,and optimizing spatial queries. It is expected to have wide applications in geographicinformation systems, geomarketing, remote sensing, image database exploration, medicalimaging, navigation, traffic control, environmental studies, and many other areaswhere spatial data are used. A crucial challenge to spatial data mining is the explorationof *efficient* spatial data mining techniques due to the huge amount of spatial data and thecomplexity of spatial data types and spatial access methods.

Multimedia Data Mining

"What is a multimedia database?" A multimedia database system stores and manages a large collection of *multimedia data*, such as audio, video, image, graphics, speech, text, document, and hypertext data, which contain text, text markups, and linkages. Multimedia database systems are increasingly common owing to the popular use of audiovideo equipment, digital cameras, CD-ROMs, and the Internet. Typical multimedia database systems include NASA's EOS (Earth Observation System), various kinds of image and audio-video databases, and Internet databases.

Similarity Search in Multimedia Data

"When searching for similarities in multimedia data, can we search on either the data description or the data content?" That is correct. For similarity searching in multimedia data, we consider two main families of multimedia indexing and retrieval systems: (1) description-based retrieval systems, which build indices and perform object retrieval based on image descriptions, such as keywords, captions, size, and time of creation; and (2) content-based retrieval systems, which support retrieval based on the image content, such as color histogram, texture, pattern, image topology, and the shape of objects and their layouts and locations within the image. Description-based retrieval islabor-intensive if performed manually. If automated, the results are typically of poor quality. For example, the assignment of keywords to images can be a tricky and arbitrary task. Recent development of Web-based image clustering and classification methods has improved the quality of description-based Web image retrieval, because imagesurrounded text information as well as Web linkage information can be used to extract proper description and group images describing a similar theme together. Content-based retrieval uses visual features to index images and promotes object retrieval based on feature similarity, which is highly desirable in many applications.
In a content-based image retrieval system, there are often two kinds of queries: *imagesample-based queries* and *image feature specification queries*. Image-sample-based queries find all of the images that are similar to the given image sample. This search compares thefeature vector (or signature) extracted from the sample with the feature vectors of

images that have already been extracted and indexed in the image database. Based on this comparison, images that are close to the sample image are returned. Image feature specification queries specify or sketch image features like color, texture, or shape, which are translated into a feature vector to be matched with the feature vectors of the images in thedatabase.Content-based retrieval haswide applications, includingmedical diagnosis, weather prediction, TV production, Web search engines for images, and e-commerce. Some systems, such as *QBIC* (*Query By Image Content*), support both sample-based and image feature specification queries. There are also systems that support both contentbased and description-based retrieval.

To facilitate themultidimensionalanalysis of largemultimedia databases,multimedia data cubes can be designed andconstructed in a manner similar to that for traditional data cubes from relational data.A multimedia data cube can contain additional dimensions and measures for multimediainformation, such as color, texture, and shape.

Text Mining

Most previous studies of data mining have focused on structured data, such as relational, transactional, and data warehouse data. However, in reality, a substantial portion ofthe available information is stored in text databases (or document databases), whichconsist of large collections of documents from various sources, such as news articles,research papers, books, digital libraries, e-mail messages, andWeb pages. Text databasesare rapidly growing due to the increasing amount of information available in electronicform, such as electronic publications, various kinds of electronic documents, e-mail, andtheWorldWideWeb (which can also be viewed as a huge, interconnected, dynamic textdatabase). Nowadays most of the information in government, industry, business, andother institutions are stored electronically, in the form of text databases.Data stored in most text databases are *semistructured data* in that they are neithercompletely unstructured nor completely structured. For example, a document maycontain a few structured fields, such as *title*, *authors*, *publication date*, *category*, andso on, but also contain some largely unstructured text components, such as *abstract*and *contents*. There have been a great deal of studies on the modeling and implementationof semistructured data in recent database research. Moreover, informationretrieval techniques, such as text indexing methods, have been developed to handleunstructured documents.

Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data. Typically, only a small fraction of the many available documents will be relevant to a given individual user.Without knowing what could be in thedocuments, it is difficult to formulate effective queries for analyzing and extracting usefulinformation from the data. Users need tools to compare different documents, rank theimportance and relevance of the documents, or find patterns and trends across multipledocuments. Thus, text mining has become an increasingly popular and essential themein data mining.

There are many approaches to text mining, which can be classified from different perspectives, based on the inputs taken in the text mining system and the data mining tasks to be performed. In general, the major approaches, based on the kinds of data they take as input, are:

(1) the keyword-based approach, where the input isa set of keywords or terms in the documents,

(2) the tagging approach, where theinput is a set of tags, and

(3) the information-extraction approach, which inputssemantic information, such as events, facts, or entities uncovered by informationextraction.

A simple keyword-based approach may only discover relationships at arelatively shallow level, such as rediscovery of compound nouns (e.g., “database”and “systems”) or co-occurring patterns with less significance (e.g., “terrorist” and“explosion”). It may not bring much deep understanding to the text. The taggingapproach may rely on tags obtained by manual tagging (which is costly and is unfeasiblefor large collections of documents) or by some automated categorization algorithm(which may process a relatively small set of tags and require defining the categoriesbeforehand). The information-extraction approach is more advanced and may leadto the discovery of some deep knowledge, but it requires semantic analysis of text bynatural language understanding and machine learning methods. This is a challengingknowledge discovery task.

Various text mining tasks can be performed on the extracted keywords, tags, or semantic information. These include document clustering, classification, information extraction,association analysis, and trend analysis

Mining theWorldWideWeb

TheWorldWideWeb serves as a huge,widely distributed, global information service centerfor news, advertisements, consumer information, financial management, education, government, e-commerce, and many other information services. TheWeb also containsa rich and dynamic collection of hyperlink information and Web page access and usageinformation, providing rich sources for data mining. However, based on the followingobservations, the Web also poses great challenges for effective resource and knowledgediscovery.

The Web seems to be too huge for effective data warehousing and data mining. The size of theWeb is in the order of hundreds of terabytes and is still growing rapidly. Manyorganizations and societies place most of their public-accessible information on theWeb. It is barely possible to set up a data warehouse to replicate, store, or integrate allof the data on theWeb.

How can the portion oftheWeb that is truly relevant to your interest be determined? How can we find highqualityWebpages on a specified topic?

Webmining is a more challenging task that searches forWeb structures, ranks the importanceofWeb contents, discovers the regularity and dynamics ofWeb contents, and minesWebaccess patterns. However,Web mining can be used to substantially enhance the power of

AWeb search engine sinceWeb miningmay identify authoritativeWeb pages, classifyWebdocuments, and resolve many ambiguities and subtleties raised in keyword-based Websearch. In general, Web mining tasks can be classified into three categories: Web contentmining, Web structure mining, and Web usage mining. Alternatively, Web structures canbe treated as a part ofWeb contents so thatWeb mining can instead be simply classifiedinto Web content mining and Web usage mining.

Data Mining Applications

1.Data Mining for Financial Data Analysis

Most banks and financial institutions offer a wide variety of banking services, credit and investment services. Some also offer insurance services and stock investment services.

Financial data collected in the banking and financial industry are often relatively complete, reliable, and of high quality, which facilitates systematic data analysis and data mining. Here we present a few typical cases:

Design and construction of data warehouses for multidimensional data analysis and data mining:

Data warehouses need to be constructedfor banking and financial data. Multidimensional data analysis methods should beused to analyze the general properties of such data. For example, one may like to viewthe debt and revenue changes by month, by region, by sector, and by other factors,along with maximum, minimum, total, average, trend, and other statistical information.

Data warehouses, data cubes, multifeature and discovery-driven data cubes,characterization and class comparisons, and outlier analysis all play important rolesin financial data analysis and mining.

Loan payment prediction and customer credit policy analysis:

Loan payment predictionand customer credit analysis are critical to the business of a bank. Many factorscan strongly or weakly influence loan payment performance and customer credit rating.Data mining methods, such as attribute selection and attribute relevance ranking, may help identify important factors and eliminate irrelevant ones.

Analysis of the customer payment history may find that, say, payment-toincomeratio is a dominant factor, while education level and debt ratio are not. Thebank may then decide to adjust its loan-granting policy so as to grant loans to those customers whose applications were previously denied but whose profiles show relativelylow risks according to the critical factor analysis.

Classification and clustering of customers for targeted marketing:

Classification andclustering methods can be used for customer group identification and targeted marketing.For example, we can use classification to identify the most crucial factors that may influence a customer's decision regarding banking. Customers with similarbehaviors regarding loan payments may be identified bymultidimensional clusteringtechniques. These can help identify customer groups, associate a new customer withan appropriate customer group, and facilitate targeted marketing.

Detection of money laundering and other financial crimes:

To detect money launderingand other financial crimes, it is important to integrate information frommultipledatabases, as long as they are potentially related to the study. Multiple data analysistools can then be used to detect unusual patterns, such as large amounts of cash flowat certain periods, by certain groups of customers. Useful tools include data visualizationtools, linkage analysis tools, classification tools, clustering tools, outlier analysis tools and sequential patternanalysis tools.

2 Data Mining for the Retail Industry

The retail industry is a major application area for data mining, since it collects huge amounts of data on sales, customer shopping history, goods transportation, consumption,and service. The quantity of data collected continues to expand rapidly, especiallydue to the increasing ease, availability, and popularity of business conducted on theWeb,or e-commerce. Today, many stores also have websites where customers can make purchaseson-line. Some businesses, such as Amazon.com (*www.amazon.com*), exist solely on-line, without any physical store locations. Retail data providea rich source for data mining.

Retail data mining can help identify customer buying behaviors, discover customershopping patterns and trends, improve the quality of customer service, achieve bettercustomer retention and satisfaction, enhance goods consumption ratios, designmore effective goods transportation and distribution policies, and reduce the cost ofbusiness.Multidimensional analysis of sales, customers, products, time, and region: The retailindustry requires timely information regarding customer needs, product sales, trends,and fashions, as well as the quality, cost, profit, and service of commodities. It is thereforeimportant to provide powerfulmultidimensional analysis and visualization tools.

Analysis of the effectiveness of sales campaigns:

The retail industry conducts salescampaigns using advertisements, coupons,andvarious kindsof discountsandbonusesto promote products and attract customers. Careful analysis of the effectivenessof sales campaigns can help improve company profits.Multidimensional analysis canbe used for this purpose by comparing the amount of sales and the number of transactionscontaining the sales items during the sales period versus those containing thesame items before or after the sales campaign.

Customer retention—analysis of customer loyalty:

With customer loyalty card information,one can register sequences of purchases of particular customers. Customerloyalty and purchase trends can be analyzed systematically. Goods purchased at differentperiods by the same customers can be grouped into sequences. Sequential patternmining can be used to investigate changes in customer consumptionor loyalty and suggest adjustments on the pricing and variety of goods in orderto help retain customers and attract new ones.

Product recommendation and cross-referencing of items:

By mining associationsfrom sales records, one may discover that a customer who buys a digital camera islikely to buy another set of items. Such information can be used to form productrecommendations. *Collaborative recommender systems* use data mining techniquesto make personalized product recommendations during live customer transactions,based on the opinions of other customers.Product recommendationscan also be advertised on sales receipts, in weekly flyers, or on the Web to helpimprove customer service, aid customers in selecting items, and increase sales.

3Data Mining for the Telecommunication Industry

With the deregulation of the telecommunicationindustry in many countries and the development of new computer and communicationtechnologies, the telecommunication market is rapidly expanding and highly competitive.This creates a great demand for data mining in order to help understand the businessinvolved, identify telecommunication patterns, catch fraudulent activities, make betteruse of resources, and improve the quality of service.

The following are a few scenarios for which data mining may improve telecommunication services:

Multidimensional analysis of telecommunication data:

Telecommunication data areintrinsically multidimensional, with dimensions such as calling-time, duration,location of caller, location of callee, and type of call. The multidimensional analysisof such data can be used to identify and compare the data traffic, system workload,resource usage, user group behavior, and profit. Therefore, it is often useful to consolidatetelecommunication data into large data warehouses and routinely perform multidimensionalanalysis using OLAP and visualization tools.

Fraudulent pattern analysis and the identification of unusual patterns:

Fraudulentactivity costs the telecommunication industry millions of dollars per year. It is important to (1) identify potentially fraudulent users and their atypical usagepatterns; (2) detect attempts to gain fraudulent entry to customer accounts; and(3) discover unusual patterns that may need special attention, such as busy-hourfrustrated call attempts, switch and route congestion patterns, and periodic callsfrom automatic dial-out equipment (like fax machines) that have been improperlyprogrammed.Many of these patterns can be discovered bymultidimensional analysis,cluster analysis, and outlier analysis.

Multidimensional association and sequential pattern analysis:

The discovery of associationand sequential patterns in multidimensional analysis can be used to promotetelecommunication services. This can help promote the sales of specific long-distance and cellular phonecombinations and improve the availability of particular services in the region.

Mobile telecommunication services:

Mobile telecommunication, Web and informationservices, and mobile computing are becoming increasingly integrated andcommon in our work and life. One important feature of mobile telecommunicationdata is its association with spatiotemporal information. Spatiotemporal datamining may become essential for finding certain patterns. Data mining will likely play a major role inthe design of adaptive solutions enabling users to obtain useful information with relatively few keystrokes.

Use of visualization tools in telecommunication data analysis:

Tools for OLAPvisualization, linkage visualization, association visualization, clustering, and outliervisualization have been shown to be very useful for telecommunication data analysis.

4.Data Mining for Biological Data Analysis

Data mining may contribute to biological data analysis in the following aspects:

Semantic integration of heterogeneous, distributed genomic and proteomic databases:

Genomic and proteomic data sets are often generated at different labs and bydifferent methods. They are distributed, heterogenous, and of a wide variety. Thesemantic integration of such data is essential to the cross-site analysis of biologicaldata. Moreover, it is important to find correct linkages between research literatureand their associated biological entities. Such integration and linkage ananalysis wouldfacilitate the systematic and coordinated analysis of genome and biological data.

This has promoted the development of integrated data warehouses and distributedfederated databases to store and manage the primary and derived biological data.Data cleaning, data integration, reference reconciliation, classification, and clusteringmethods will facilitate the integration of biological data and the construction of datawarehouses for biological data analysis.

Alignment, indexing, similarity search, and comparative analysis ofmultiple nucleotide/ protein sequences:

Various biological sequence alignment methods have beendeveloped in the past two decades. BLAST and FASTA, in particular, are tools for thesystematic analysis of genomic and proteomic data.Biological sequence analysismethodsdiffer frommany sequential pattern analysis algorithms proposed in data miningresearch. They should allow for gaps and mismatches between a query sequence andthe sequence data to be searched in order to deal with insertions, deletions, andmutations.

From the point of view of medical sciences, genomic and proteomic sequences isolated from diseased and healthy tissues can be compared to identify critical differences between them. Sequences occurring more frequently in the diseased samplesmay indicate the genetic factors of the disease. Those occurring more frequentlyonly in the healthy samples may indicate mechanisms that protect the body from thedisease.

Discovery of structural patterns and analysis of genetic networks and protein pathways:

In biology, protein sequences are folded into three-dimensional structures,and such structures interact with each other based on their relative positions and thedistances between them. Such complex interactions form the basis of sophisticatedgenetic networks and protein pathways. It is crucial to discover structural patternsand regularities among such huge but complex biological networks. To this extent,it is important to develop powerful and scalable data mining methods to discoverapproximate and frequent structural patterns and to study the regularities and irregularitiesamong such interconnected biological networks.

Association and path analysis:

Identifying co-occurring gene sequences and linkinggenes to different stages of disease development: Currently, many studies havefocused on the comparison of one gene to another. However, most diseases are nottriggered by a single gene but by a combination of genes acting together. *Associationanalysis* methods can be used to help determine the kinds of genes that are likely toco-occur in target samples. Such analysis would facilitate the discovery of groups ofgenes and the study of interactions and relationships between them.

While a group of genes may contribute to a disease process, different genes maybecome active at different stages of the disease. If the sequence of genetic activitiesacross the different stages of disease development can be identified, it may be possibleto develop pharmaceutical interventions that target the different stages separately,therefore achieving more effective treatment of the disease. Such *path analysis* isexpected to play an important role in genetic studies.

Visualization tools in genetic data analysis:

Alignments among genomic orproteomic sequences and the interactions among complex biological structures aremost effectively presented in graphic forms, transformed into various kinds ofeasy-to-understand visual displays. Such visually appealing structures and patternsfacilitate pattern understanding, knowledge discovery, and interactive data exploration.Visualization and visual data mining therefore play an important role inbiological data analysis.

5Data Mining in Other Scientific Applications

Data warehouses and data preprocessing:

Data warehouses are critical for informationexchange and data mining. In the area of geospatial data, however, no truegeospatial data warehouse exists today. Creating such a warehouse requires findingmeans for resolving geographic and temporal data incompatibilities, such as reconcilingsemantics, referencing systems, geometry, accuracy, and precision. For scientificapplications in general, methods are needed for integrating data from heterogeneoussources (such as data covering different time periods) and for identifying events.

Mining complex data types:

Scientific data sets are heterogeneous in nature, typicallyinvolving semi-structured and unstructured data, such as multimedia data andgeoreferenced stream data. Robust methods are needed for handling spatiotemporaldata, related concept hierarchies, and complex geographic relationships.

Graph-based mining:

It is often difficult or impossible to model several physicalphenomena and processes due to limitations of existing modeling approaches. Alternatively,labeled graphs may be used to capture many of the spatial, topological,geometric, and other relational characteristics present in scientific data sets. In graphmodeling,each object to be mined is represented by a vertex in a graph, and edgesbetween vertices represent relationships between objects. For example, graphs canbe used to model chemical structures and data generated by numerical simulations,such as fluid-flow simulations. The success of graph-modeling, however, depends onimprovements in the scalability and efficiency of many classical data mining tasks,such as classification, frequent pattern mining, and clustering.

Visualization tools and domain-specific knowledge:

High-level graphical user interfacesand visualization tools are required for scientific data mining systems. Theseshould be integrated with existing domain-specific information systems and databasesystems to guide researchers and general users in searching for patterns, interpretingand visualizing discovered patterns, and using discovered knowledge in their decisionmaking.

6Data Mining for Intrusion Detection

Current traditional intrusion detection systemsface many limitations. This has led to an increased interest in data mining for intrusiondetection. The following are areas in which data mining technology may be applied orfurther developed for intrusion detection:

Development of data mining algorithms for intrusion detection:

Data miningalgorithms can be used for misuse detection and anomaly detection. In misuse detection,training data are labeled as either “normal” or “intrusion.” A classifier canthen be derived to detect known intrusions. Research in this area has included theapplication of classification algorithms, association rule mining, and cost-sensitivemodeling. Anomaly detection builds models of normal behavior and automaticallydetects significant deviations from it. Supervised or unsupervised learning can beused. In a supervised approach, the model is developed based on training data thatare known to be “normal.” In an unsupervised approach, no information is givenabout the training data. Anomaly detection research has included the application ofclassification algorithms, statistical approaches, clustering, and outlier analysis. Thetechniques used must be efficient and scalable, and capable of handling network dataof high volume, dimensionality, and heterogeneity.

Association and correlation analysis, and aggregation to help select and build discriminatingattributes:

Association and correlation mining can be applied to findrelationships between system attributes describing the network data. Such informationcan provide insight regarding the selection of useful attributes for intrusiondetection. New attributes derived from aggregated data may also be helpful, such assummary counts of traffic matching a particular pattern.

Analysis of stream data:

Due to the transient and dynamic nature of intrusions andmalicious attacks, it is crucial to perform intrusion detection in the data streamenvironment. Moreover, an event may be normal on its own, but considered maliciousif viewed as part of a sequence of events. Thus it is necessary to study whatsequences of events are frequently encountered together, find sequential patterns,and identify outliers. Other data mining methods for finding evolving clusters andbuilding dynamic classificationmodels in data streams are also necessary for real-timeintrusion detection.

Distributed data mining:

Intrusions can be launched fromseveral different locationsand targeted to many different destinations. Distributed data mining methods maybe used to analyze network data from several network locations in order to detectthese distributed attacks.

Visualization and querying tools:

Visualization tools should be available for viewingany anomalous patterns detected. Such tools may include features for viewing associations,clusters, and outliers. Intrusion detection systems should also have a graphicaluser interface that allows security analysts to pose queries regarding the network dataor intrusion detection results.

Trends in Data Mining

Following are some of the trends in data mining:

Application exploration:

Early data mining applications focused mainly on helpingbusinesses gain a competitive edge. The exploration of data mining for businessescontinues to expand as e-commerce and e-marketing have become mainstream elementsof the retail industry. Data mining is increasingly used for the explorationof applications in other areas, such as financial analysis, telecommunications,biomedicine, and science. Emerging application areas include data mining for counterterrorism and mobile datamining.

Scalable and interactive data mining methods:

In contrast with traditional data analysismethods, data mining must be able to handle huge amounts of data efficientlyand interactively. One important direction toward improving the overallefficiency of the mining process while increasing user interaction is constraint basedmining. This provides users with added control by allowing the specification and useof constraints to guide data mining systems in their search for interesting patterns.

Integration of data mining with database systems, data warehouse systems, and Web database systems:

Database systems, data warehouse systems, and theWeb havebecome mainstream information processing systems. It is important to ensure thatdata mining serves as an essential data analysis component that can be smoothlyintegrated into such an information processing environment. Transaction management, query processing, on-line analytical processing,and on-line analytical mining should be integrated into one unified framework. Thiswill ensure data availability, data mining portability, scalability, high performance,and an integrated information processing environment for multidimensional dataanalysis and exploration.

Standardization of data mining language:

A standard data mining language or otherstandardization efforts will facilitate the systematic development of data mining solutions,improve interoperability among multiple data mining systems and functions,and promote the education and use of data mining systems in industry and society.Recent efforts in this direction include Microsoft's OLE DB for Data Mining, PMML, and CRISP-DM.

Visual data mining:

Visual data mining is an effective way to discover knowledgefrom huge amounts of data. The systematic study and development of visual datamining techniques will facilitate the promotion and use of data mining as a tool fordata analysis.

New methods for mining complex types of data:

Mining complex types of data is an important research frontier in data mining.Although progress has been made in mining stream, time-series, sequence, graph,spatiotemporal,multimedia, and text data, there is still a huge gap between the needsfor these applications and the available technology. More research is required, especiallytoward the integration of data mining methods with existing data analysistechniques for these types of data.

Biological data mining:

Although biological data mining can be considered under"application exploration" or "mining complex types of data," the unique combinationof complexity, richness, size, and importance of biological data warrantsspecial attention in data mining. Mining DNA and protein sequences, mining highdimensionalmicroarray data, biological pathway and network analysis, link analysisacross heterogeneous biological data, and information integration of biological databy data mining are interesting topics for biological data mining research.

Data mining and software engineering:

The analysis of theexecutions of a buggy software program is essentially a data mining process tracing the data generated during program executions may disclose importantpatterns and outliers that may lead to the eventual automated discovery of softwarebugs.We expect that the further development of data mining methodologies for softwaredebugging will enhance software robustness and bring new vigor to softwareengineering.

Web mining:

Giventhe huge amount of information available on theWeb and the increasingly importantrole that the Web plays in today's society, Web content mining, Weblog mining, and data mining services on the Internet will become one of the most important andflourishing subfields in data mining.

Distributed data mining:

Traditional data mining methods, designed to work at acentralized location, do not work well in many of the distributed computing environmentspresent today (e.g., the Internet, intranets, local area networks, high-speedwireless networks, and sensor networks). Advances in distributed data mining methodsare expected.

Real-time or time-critical data mining:

Many applications involving stream data(such as e-commerce, Web mining, stock analysis, intrusion detection, mobile datamining, and data mining for counterterrorism) require dynamic data mining modelsto be built in real time. Additional development is needed in this area.Graph mining, link analysis, and social network analysis: Graph mining, link analysis,and social network analysis are useful for capturing sequential, topological, geometric,and other relational characteristics of many scientific data sets (such as forchemical compounds and biological networks) and social data sets (such as for theanalysis of hidden criminal networks). Suchmodeling is also useful for analyzing linksin Web structure mining. The development of efficient graph and linkage models isa challenge for data mining.

Multirelational andmultidatabase data mining:

Most data mining approaches searchfor patterns in a single relational table or in a single database. However, most realworlddata and information are spread acrossmultiple tables and databases.Multirelationaldata mining methods search for patterns involvingmultiple tables (relations)from a relational database. Multidatabase mining searches for patterns across multipledatabases. Further research is expected in effective and efficient data miningacross multiple relations and multiple databases.

Privacy protection and information security in data mining:

An abundance ofrecorded personal information available in electronic forms and on the Web, coupledwith increasingly powerful data mining tools, poses a threat to our privacyand data security. Growing interest in data mining for counterterrorism also addsto the threat. Further development of privacy-preserving data mining methods is foreseen. The collaboration of technologists, social scientists, law experts, andcompanies is needed to produce a rigorous definition of privacy and a formalismto prove privacy-preservation in data mining.