5.1 Applications of Data Mining

A wide range of companies have deployed successful applications of data mining. While early adopters of this technology have tended to be in information-intensive industries such as financial services and direct mail marketing, the technology is applicable to any company looking to leverage a large data warehouse to better manage their customer relationships. Two critical factors for success with data mining are: a large, well-integrated data warehouse and a well-defined understanding of the business process within which data mining is to be applied (such as customer prospecting, retention, campaign management, and so on).

Some successful application areas include:

- A pharmaceutical company can analyze its recent sales force activity and their results to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The data needs to include competitor market activity as well as information about the local health care systems. The results can be distributed to the sales force via a wide-area network that enables the representatives to review the recommendations from the perspective of the key attributes in the decision process. The ongoing, dynamic analysis of the data warehouse allows best practices from throughout the organization to be applied in specific sales situations.
- A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product. Using a small test mailing, the attributes of customers with an affinity for the product can be identified. Recent projects have indicated more than a 20-fold decrease in costs for targeted mailing campaigns over conventional approaches.
- A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. Using data mining to analyze its own customer experience, this company can build a unique segmentation identifying the attributes of high-value prospects. Applying this segmentation to a general business database such as those provided by Dun & Bradstreet can yield a prioritized list of prospects by region.
- A large consumer package goods company can apply data mining to improve its sales process to retailers. Data from consumer panels, shipments, and competitor activity can be applied to understand the reasons for brand and store switching. Through this analysis, the manufacturer can select promotional strategies that best reach their target customer segments.

Each of these examples have a clear common ground. They leverage the knowledge about customers implicit in a data warehouse to reduce costs and improve the value of customer relationships. These organizations can now focus their efforts on the most important (profitable) customers and prospects, and design targeted marketing strategies to best reach them.

There are a number of applications that data mining has. The first is called market segmentation. With market segmentation, you will be able to find behaviors that are common among your customers. You can look for patterns among customers that seem to purchase the same products at the same time. Another application of data mining is called customer churn. Customer churn will allow you to estimate which customers are the most likely to stop purchasing your products or services and go to one of your competitors. In addition to this, a company can use data mining to find out which purchases are the most likely to be fraudulent.

For example, by using data mining a retail store may be able to determine which products are stolen the most. By finding out which products are stolen the most, steps can be taken to protect those products and detect those who are stealing them. While direct mail marketing is an older technique that has been used for many years, companies who combine it with data mining can experience fantastic results. For example, you can use data mining to find out which customers will respond favorably to a direct mail marketing strategy. You can also use data mining to determine the effectiveness of interactive marketing. Some of your customers will be more likely to purchase your products online than offline, and you must identify them.

While many businesses use data mining to help increase their profits, many of them don't realize that it can be used to create new businesses and industries. One industry that can be created by data mining is the automatic prediction of both behaviors and trends. Imagine for a moment that you were the owner of a fashion company, and you were able to precisely predict the next big fashion trend based on the behavior and shopping patterns of your customers? It is easy to see that you could become very wealthy within a short period of time. You would have an advantage over your competitors. Instead of simply guessing what the next big trend will be, you will determine it based on statistics, patterns, and logic.

Another example of automatic prediction is to use data mining to look at your past marketing strategies. Which one worked the best? Why did it work the best? Who were the customers that responded most favorably to it? Data mining will allow you to answer these questions, and once you have the answers, you will be able to avoid making any mistakes that you made in your previous marketing campaign. Data mining can allow you to become better at what you do. It is also a powerful tool for those who deal with finances. A financial institution such as a bank can predict the number of defaults that will occur among their customers within a given period of time, and they can also predict the amount of fraud that will occur as well.

Another potential application of data mining is the automatic recognition of patterns that were not previously known. Imagine if you had a tool that could automatically search your database to look for patterns which are hidden. If you had access to this technology, you would be able to find relationships that could allow you to make strategic decisions.

Data mining is becoming a pervasive technology in activities as diverse as using historical data to predict the success of a marketing campaign, looking for patterns in financial transactions to discover illegal activities or analyzing genome sequences. From this perspective, it was just a matter of time for the discipline to reach the important area of computer security.

Applications of Data Mining in Computer Security presents a collection of research efforts on the use of data mining in computer security.
Data mining has been loosely defined as the process of extracting information from large amounts of data. In the context of security, the information we are seeking is the knowledge of whether a security breach has been experienced, and if the answer is yes, who is the perpetrator. This information could be collected in the context of discovering intrusions that

aim to breach the privacy of services, data in a computer system or alternatively, in the context of discovering evidence left in a computer system as part of criminal activity.

Applications of Data Mining in Computer Security concentrates heavily on the use of data mining in the area of intrusion detection. The reason for this is twofold. First, the volume of data dealing with both network and host activity is so large that it makes it an ideal candidate for using data mining techniques. Second, intrusion detection is an extremely critical activity. This book also addresses the application of data mining to computer forensics. This is a crucial area that seeks to address the needs of law enforcement in analyzing the digital evidence.

Applications of Data Mining in Computer Security is designed to meet the needs of a professional audience composed of researchers and practitioners in industry and graduate level students in computer science.

5.2 Social Impacts of Data Mining

Data Mining can offer the individual many benefits by improving customer service and satisfaction, and lifestyle in general. However, it also has serious implications regarding one's right to privacy and data security.

<u>Is Data Mining a Hype or a persistent, steadily growing business?</u>
Data Mining has recently become very popular area for research, development and business as it becomes an essential tool for deriving knowledge from data to help business person in decision making process.

<u>Several phases of Data Mining technology is as follows:</u>

- Innovators
- Early Adopters
- Chasm
- Early Majority
- Late Majority
- Laggards

<u>Is Data Mining Merely Managers Business or Everyone's Business?</u>
Data Mining will surely help company executives a great deal in understanding the market and their business. However, one can expect that everyone will have needs and means of data mining as it is expected that more and more powerful, user friendly, diversified and affordable data mining systems or components are made available.

<u>Data Mining can also have multiple personal uses such as:</u>
Identifying patterns in medical applications
To choose best companies based on customer service.
To classify email messages etc.

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, there are increasing concern that data mining may pose a threat to our privacy and data security.

Data Privacy:

In 1980, the organization for Economic co-operation and development (OECD) established as set of international guidelines, referred to as fair information practices. These guidelines aim to protect privacy and data accuracy.

They include the following principles:

- Purpose specification and use limitation.
- Openness
- Security Safeguards
- Individual Participation

Data Security:

Many data security enhancing techniques have been developed to help protect data. Databases can employ a multilevel security model to classify and restrict data according to various security levels with users permitted access to only their authorized level.

Some of the data security techniques are:

Encryption Technique

Intrusion Detection

In secure multiparty computation

In data obscuration

5.3 Tools

Data Mining Tools:

1. Auto Class III:
   Auto Class is an unsupervised Bayesian Classification System for independent data.
2. Business Miner:
   Business Miner is a single strategy easy to use tool based on decision trees.
3. CART:
   CART is a robust data mining tool that automatically searches for important patterns and relationships in large data sets.
4. Clementine:
   It finds sequence association and clustering for web data analysis.
5. Data Engine:
   Data Engine is a multiple strategy data mining tool for data modeling, combining conventional data analysis methods with fuzzy technology.
6. DB Miner:
   DB Miner is a publicly available tool for data mining. It is multiple strategy tool and it supports clustering and Association Rules.

7. Delta Miner:
   Delta Miner is a multiple strategy tool for supporting clustering, summarization, and deviation detection and visualization process.
8. IBM Intelligent Miner:
   Intelligent Miner is a integrated and comprehensive set of data mining tools. It uses decision trees, neural networks and clustering.
9. Mine Set:
   Mine Set is comprehensive tool for data mining. Its features include extensive data manipulation and transformation.
10. SPIRIT:
    SPIRIT is a tool for exploration and modeling using Bayesian techniques.
11. WEKA:
    WEKA is a S/W environment that integrates several machine learning tools within a common framework and Uniform GUI.

5.4 An Introduction to DB Miner

A Data Mining system, DB Miner has been developed for interactive mining of multiple-level knowledge in large relational databases. The system implements wide spectrum of data mining functions, including generalization, characterization, association, classification and prediction.

Introduction:
With the upsurge of research and development activities on knowledge discovery in databases, a data mining system, db miner, has been developed based on our studies of data mining techniques and our experience in the development of an early system prototype, DBlearn.

The system has the following distinct features:

1. It incorporates several interesting data mining techniques, including attribute-oriented induction, statistical analysis, progressive deepening for mining multiple level rules and meta-rule guided knowledge mining.
2. It performs interactive data mining and multiple concept levels on any user-specified set of data in a database using an SQL-like Data mining Query Language, DMQL or a GUI.
3. Efficient implementation techniques have been explored using different data structures, including generalized relations and multiple-dimensional data cubes.
4. The data mining process may utilize user or expert defined set-grouping or schema level concept hierarchies which can be specified flexibly, adjusted dynamically based on data distribution and generated automatically for numerical attributes.
5. Both UNIX and PC (Windows / NT) versions of the system adopt client / server architecture. The later may communicate with various commercial database systems for data mining using the ODBC technology.

Architecture and Functionalities:
The general architecture of DB Miner is shown in the figure A1, tightly integrates a relational database system, such as Sybase SQL Server, with a concept hierarchy module, and a set of knowledge discovery modules of DB Miner.

Graphical User Interface:
Knowledge discovery modules of DB Miner includes characterizer, discriminator, classifier, association rule finder, meta rule guided miner, predictor, evolution evaluator, deviation evaluator and some planned future modules.

The functionalities of the knowledge discovery modules are brief described as follows:
The characterizer generalizes a set of task-relevant data into a generalized relation which can then be used for extraction of different kinds of rules to be viewed at multiple concept levels from different angles.

A discriminator discovers a set of discriminator rules which summarize the features that distinguish the class being examined from other classes.

An Association Rule Finder discovers a set of association rules at the multiple concept levels from the relevant sets of data in a database.

A meta-rule guided miner is a data mining mechanism which takes a user specified meta-rule form as a pattern to confine the search for desired rule.

A predictor predicts the possible values of some mining data or the value distribution of certain attributes in a set of objects.

A data evolution evaluator evaluates the data evolution regularities for certain objects where behavior changes over time.

A deviation evaluator evaluates the deviation patterns for a set of task relevant data in a database.

Another important function module of DB Miner is concept hierarchy which provides essential background knowledge for data generalization and multiple-level data mining.

5.5 Case Studies

Data mining is the process of discovering previously unknown, actionable and profitable information from large consolidated databases and using it to support tactical and strategic business decisions.

The statistical techniques of data mining are familiar. They include linear and logistic regression, multivariate analysis, principal components analysis, decision trees and neural networks. Traditional approaches to statistical inference fail with large databases, however, because with thousands or millions of cases and hundreds or thousands of variables there will be a high level of redundancy among the variables, there will be spurious relationships, and even the weakest relationships will be highly significant by any statistical test. The objective is to build a model with significant predictive power. It is not enough just to find which relationships are statistically significant.

Consider a campaign offering a product or service for sale, directed at a given customer base. Typically, about 1% of the customer base will be "responders," customers who will purchase the product or service if it is offered to them. A mailing to 100,000 randomly-chosen customers will therefore generate about 1000 sales. Data mining techniques enable customer relationship marketing, by identifying which customers are most likely to respond to the campaign. If the response can be raised from 1% to, say, 1.5% of the customers contacted (the "lift value"), then 1000 sales could by achieved with only 66,666 mailings, reducing the cost of mailing by one-third.

Case Study: Data Mining the Northridge Earthquake

The data collected during the Northridge, California earthquake occupied several warehouses, and ranged from magnetic media to bound copies of printed reports. Nautilus Systems personnel sorted, organized, and cataloged the materials. Document were scanned and converted to text. Data were organized chronologically and according to situation reports, raw data, agency data, and agency reports. For example, the Department of Transportation had information on highways, street structures, airport structures, and related damage assessments.

Nautilus Systems applied its proprietary data mining techniques to extract and refine data. Geography was used to link related information, and text searches were used to group information tagged with specific names (e.g., Oakland Bay Bridge, San Mateo, Marina). The refined data were further analyzed to detect patterns, trends, associations and factors not readily apparent. At that time, there was not a seismographic timeline, but it was possible to map the disaster track to analyze the migration of damage based upon geographic location. Many types of analyses were done. For example, the severity of damage was analyzed according to type of physical structure, pre- versus post- 1970 earthquake building codes, and off track versus on track damage. It was clear that the earthquake building codes limited the degree of damage.

Nautilus Systems also looked at the data coming into the command and control center. The volume of data was so great that a lot was filtered out before it got to the decision support level. This demonstrated the need for a management system to build intermediate decision blocks and communicate the information where it was needed. Much of the information needed was also geographic in nature. There was no ability to generate accurate maps for response personnel, both route maps including blocked streets and maps defining disaster boundaries. There were no interoperable communications between local police, the fire department, utility companies, and the disaster field office. There were also no predefined rules of engagement between FEMA and local resources, resulting in delayed response (including such critical areas as firefighting)

Benefits

Nautilus Systems identified recurring data elements, data relationships and metadata, and assisted in the construction of the Emergency Information Management System (EIMS). The EIMS facilitates rapid building and maintenance of disaster operations plans, and provides consistent, integrated command (decision support), control (logistics management), and communication (information dissemination) throughout all phases of disaster management. Its remote GIS capability provides the ability to support multiple disasters with a central GIS team, conserving scarce resources.

5.6 Mining WWW

Mining the World Wide Web:

WWW is huge, widely distributed, global information source for:

Information Services

- News, Advertisements, consumer information, financial management, education, government, e-commerce etc.

Hyper-link Information

Access and usage information

Web-site contents and Organization

Growing and changing very rapidly

- Broad diversity of user communities.

Only a small portion of the information on the web is truly relevant or useful to web users.

- How to find high-quality Web pages on a specified topic?

WWW provides rich sources of data mining.

Challenges on WWW Interactions:

Creating knowledge from Information available

Personalization of the information

Learning about customers / individual users

Finding relevant information

Searches for:

Web access patterns

Web Structures

Regularity and dynamics of Web contents

Problems:

The “abundance” problem

Limited coverage of the Web

- Hidden Web sources, majority of data in DBMS.

Limited query interface based on keyword-oriented search.

Limited customization to individual users

Dynamic and semi-structured

Web Search Engines:

Index-based search the web, index web pages and build and store huge keyword based indices.

Helps locate sets of Web pages containing certain keywords.

Deficiencies

- A topic of any breadth may easily contain hundreds of thousands of documents
- Many documents that are highly relevant to a topic may not contain keywords defining them.

Web Mining Subtasks:

Resource Finding

- Task of retrieving intended web-documents.

Information Selection and Pre-Processing

- Automatic Selection and pre-processing specific information from retrieved web resources.

Generalization

- Automatic discovery of patterns in Web Sites.

Analysis

- Validation and / or interpretation of mined patterns

Web Content Mining:

Discovery of useful information from web contents / data documents

- Web data contents: text, image, audio, video, metadata and hyperlinks

Information Retrieval View

- Assist / Improve information finding
- Filtering information to users on user profiles

Database View

- Model data on the web integrate them for more sophisticated queries.

Web Structure Mining:

To discover the link structure of the hyperlinks at the inter-document level to generate structural summary about the website and web page.

Direction 1: based on the hyperlinks, categorizing the web pages and generated information.
Direction 2: discovering the structure of Web document itself.
Direction 3: Discovering the nature of the Web site.

Finding authoritative web pages

- Retrieving pages that are not only relevant, but also of high quality or authoritative on the topic.

Hyperlinks can infer the notion of authority.

- Web consists not only of pages, but also of hyperlinks pointing from one page to another
- These hyperlinks contain an enormous amount of latent human annotation
- A hyperlink pointing to another web page, this can be considered as the authors endorsement of the other page.

Web Usage Mining:

Web Usage Mining also known as Web log Mining
Mining Techniques to discover interesting usage patterns from the secondary data derived from the interactions of the users while surfing the web.

Techniques for Web Usage Mining:

Construct multidimensional view on the Web log database.
Perform data mining on Web log records
Conduct Studies to analyze system performance

Design of a Web log Miner:

Web log is filtered to generate a relational database
A data cube is generated from database
OLAP is used to drill-down and roll-up in the cube
OLAM is used for mining interesting knowledge

Mining the Web's link structures to identify authoritative web pages:

Identify Authoritative Web Pages:

Hub: Web page links to a collection of prominent sites on a common topic
Authority: Pages that link to a collection of authoritative pages on a broad topic; web page pointed to by hubs.

Mutual Reinforcing Relationship:

A good authority is a page that is pointed to by many good hubs, while a good hub is a page that points to many good authorities.

Finding Authoritative Web Pages:

Retrieving pages that are not only relevant, but also of high quality or authoritative on the topic.

Hyperlinks can infer the notion of authority:

- The Web consists not only of pages, but also of hyperlinks pointing from one page to another.
- These hyperlinks contain an enormous amount of latent human annotation
- A hyperlink pointing to another web page, this can be considered as the author's endorsement of the other page.

Problems with the web linkage structure:

- Not every hyperlink represents an endorsement
- One authority will seldom have its web page point to its rival authorities in the same field
- Authoritative pages are seldom particularly descriptive.

HITS (Hyperlink Inducted Topic Search):

Explore interactions between hubs and authoritative pages.

Use an index-based search engine to form the root set.

Expand the root set into a base set

- Include all of the pages that the root set pages link to, and all of the pages that link to a page in the root set, up to a designated size cut off.

Apply weight-propagation

- An iterative process that determines numerical estimates of hub and authority weights

System based on the HITS Algorithm:

- Clever Google: Achieve better quality search results than those generated by term index engines such as Alta Vista.

Difficulties from ignoring textual contexts

- Drifting
- Topic Hijaking

Automatic Classification of Web Documents:

Assign a class label to each document from a set of predefined topic categories.

Based on a set of examples of pre-classified documents

Keyword-based document classification methods

Statistical Models

Multilayered Web Information Base:

Layer 0: the Web itself

Layer 1: the Web page descriptor layer

Layer 2 and up: various web directory services constructed on the top of layer 1

Applications of Web Mining:

Target potential customers for e-commerce

Improve web server system performance

Identify potential prime advertisement locations

Facilitates adaptive / personalized sites

Improve site design

Fraud / Intrusion detection

Predicts user's actions

5.7 Mining Text Database

Mining Text Databases:

Text Databases and Information Retrieval:

Text Databases (Document Databases):

- Large collections of documents from various sources, new articles, research papers, books, digital libraries, email messages and web pages, library databases etc.
- Data stored is usually semi-structured
- Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data.

Information Retrieval:

- A field developed in parallel with database systems.
- Information is organized into a large number of documents
- Information retrieval problem: locating relevant documents based on user input, such as keywords or example documents.

Information Retrieval:

Typical IR Systems:

- Online Library Catalogs
- Online document management systems

Information Retrieval Vs Database Systems:

- Some DB problems are not present in IR, eg., update, transaction management, complex objects.
- Some IR problems are not addressed well in DBMS, eg., unstructured documents, approximate search using keywords and relevance.

Precision: the percentage of retrieved documents that are in fact relevant to the query.

Precision = |{Relevant} ^ {Retrieved}|

|{Retrieved}|

Recall: the percentage of documents that are relevant to the query and were in fact retrieved.

Recall = |{Relevant} ^ {Retrieved}|

|{Relevant}|

Keyword – Based Retrieval:

A document is represented by a string, which can be identified by a set of keywords. Queries may use expressions of keywords.

- Eg. Car amd Repair shop, tea, coffee, DBMS but not Oracle
- Queries and retrieval should consider synonyms, eg. Repair and maintenance.

Major difficulties of the Model:

Synonymy: A keyword T does not appear anywhere in the document, even though the document is closely related to T, eg. Data mining

Polysemy: The same keyword may mean different things in different contexts eg. Mining.

Similarity – Based Retrieval in Text DBs:

Finds similar documents based on a set of common keywords.

Answer should be based on the degree of relevance based on the nearness of the keywords, relative frequency of the keywords etc.

Basic Techniques:

Stop List:

Set of words that is deemed "irrelevant" even though they may appear frequently.

Eg. A, the, of, for, with etc.

Stop lists may vary when document set varies.

Word Stem:

Several words are small syntactic variants of each other since they share a common word stem.

A term frequency table:

Each entry frequent table (i,j)

- No. of occurrences of the word ti in document di
- Usually, the ratio instead of the absolute number of occurrences is used.

Similarity Metrics:

Measure the closeness of the document to a query (a set of keywords)

Relative term occurrences

Cosine Distance

Latent Semantic Indexing:

Basic Idea:

- Similar documents have similar word frequencies.
- Difficulty: the size of the term frequency matrix is very large.
- Use a singular value decomposition (SVD) techniques to reduce the size of the frequency table.
- Retain the K most significant rows of the frequency table.

Method:

- Create a term frequency matrix, freq-matrix.
- SVD Construction: Compute the singular valued decomposition of the freq-matrix by splitting it into 3 matrices, U, S, V.

Vector Identification:

- For each document d, replace its original document vector by a new excluding the eliminated terms.

Index Creation:

- Store the set of all vectors, indexed by one of a number of techniques (such as TV-tree)

Other Text Retrieval Indexing Techniques:

Inverted Index:

- Maintains two hash or B +tree indexed tables.

Document Table:

- a set of documents records < doc_id, postings_list>

Term-table: a set of term records, < term, postings_list>

Answer Query: Find all docs associated with one or a set of terms.

Advantage: Easy to implement

Disadvantage: Do not handle well synonymy and polysely and posting lists could be too long (storage could be very large)

Signature File:

- Associate a signature with each document.

- A signature is a representation of an ordered list of terms that describe the document.
- Order is obtained by frequency analysis, stemming and stop lists.

Types of Text Data Mining:

Keyword – based association analysis.

Automatic document classification

Similarity detection

Cluster documents by a common author

Cluster documents containing information from a common source

Link analysis: Unusual Correlation between entities.

Sequence Analysis: Predicting a recurring event.

Anomaly Detection: Find information that violates usual patterns.

Hypertext Analysis:

Patterns in anchors / links

Anchor text correlations with linked objects.

Keyword based Association Analysis:

Collect sets of keywords or terms that occur frequently together and then find the association or correlation relationships among them.

First preprocess the text data by parsing, stemming, removing stop words etc.

Then evoke association mining algorithms.

Consider each document as a transaction

View a set of keywords in the document as a set of items in the transaction.

Term level Association Mining:

No need for human effort in tagging documents

The number of meaningless results and the execution time is greatly reduced.

Automatic Document Classification:

Motivation

Automatic Classification for the tremendous number of on-line text documents.

A Classification Problem:

Training set: Human experts generate a training data set.

Classification: The computer system discovers the classification rules.

Application: The discovered rules can be applied to classify new / unknown documents.

Text Document Classification differs from the classification of relational data

Document databases are not structured according to attribute-value pairs.

Association-Based Document Classification:

Extract keywords and terms by information retrieval and simple association analysis techniques.

Obtain concept hierarchies of keywords and terms using:

Available term classes, such as Word Net

Expert Knowledge

Some keyword classification systems.

Classify documents in the training set into class hierarchies

Apply term association mining method to discover sets of associated terms.

Use the terms to maximally distinguish one class of documents from others.

Derive a set of association rules associated with each document class

Order the classification rules based on their occurrence frequency and discriminative power.

Used the rules to classify new documents.

Document Clustering:

Automatically group related documents based on their contents.

Require no training sets or predetermined taxonomies, generate a taxonomy at runtime.

Major Steps:

Preprocessing

Remove stop words, stem, feature extraction, lexical analysis,…

Hierarchical Clustering

Compute similarities applying clustering algorithms,…

Slicing

Fan out controls, flatten the tree to configurable number of levels,…

5.8 Mining Spatial Databases

Spatial Data Mining refers to the extraction of knowledge, spatial relationships or other interesting patterns not explicitly stored in spatial databases.

A spatial database stores a large amount of space-related data, such as maps, preprocessed remote sensing or medical imaging data, and VLSI chip layout data.

Statistical spatial data analysis has been a popular approach to analyzing spatial data and exploring geographic information.

The term ‘geostatistics’ is often associated with continuous geographic space, whereas the term ‘Spatial statistics’ is often associated with discrete space.

Spatial Data Mining Applications:

Geographic information systems

Geo marketing

Remote sensing

Image database exploration

Medical Imaging

Navigation

Traffic Control

Environmental Studies

Spatial Data Cube Construction and Spatial OLAP:

Spatial data warehouse is a subject-oriented integrated, time-variant and non-volatile collection of both spatial and non-spatial data in support of spatial data mining and spatial data related decision-making process.

There are three types of dimensions in a Spatial Data Cube:

A non-spatial dimension contains only non-spatial data, each contains nonspatial data whose generalizations are non-spatial.

A Spatial-to-nonspatial dimension is a dimension whose primitive-level data are spatial but whose generalization, starting at a certain high level, becomes non-spatial.

A Spatial-to-Spatial dimension is a dimension whose primitive level and all of its high level generalized data are spatial.

Measures of Spatial Data Cube:

A numerical measure contains only numeric data

A Spatial measure contains a collection of pointers to spatial objects.

Computation of Spatial Measures in Spatial Data Cube Construction:

Collect and store the corresponding spatial object pointers but do not perform precomputation of spatial measures in the spatial data cube.

Precompute and store a rough approximation of the spatial measures in the spatial data cube.

Selectively pre-compute some spatial measures in the spatial data cube.

Mining Spatial Association and Co-Location Pattern:

Spatial Association rules can be mined in spatial databases.

A Spatial association rule is of the form A ➔ B [s%, c%] where A & B are sets of spatial or non-spatial predicates.

S% is the support of the rule; c% is the confidence of the rule

Spatial Association mining needs to evaluate multiple spatial relationships among a large number of spatial objects, the process could be quite costly.

An interesting mining optimization called ‘progressive refinement’ can be adopted in spatial association analysis.

The method first mines large data sets roughly using a fast algorithm and then improves the quality of mining in a pruned data set using a more expensive algorithm.

Superset Coverage Property:

It should allow a false-positive test, which might include some data sets that do not belong to the answer sets, but it should not allow a ‘false-negative test’, which might exclude some potential answers.

For mining spatial associations related to the spatial predicate close to and collect the candidates that pass the minimum support threshold by

Applying certain rough spatial evaluation algorithms.

Evaluating the relaxed spatial predicate, ‘g close to’, which is generalized close to covering a broader context that includes ‘close to’, ‘touch’ and intersect’

Spatial Clustering methods:

Spatial data clustering identifies clusters, or densely populated regions, according to some distance measurement in a large, multi dimensional data set.

Spatial Classification and Spatial Trend Analysis:

Spatial Classification analyzes spatial objects to derive classification schemes in relevance to certain spatial properties.

Example: Classify regions in a province into rich Vs poor according to the average family income.

Trend analysis detects changes with time, such as the changes of temporal patterns in time-series data.

Spatial trend analysis replaces time with space and studies the trend of non-spatial or spatial data changing with space.

Example: Observe the trend of changes of the climate or vegetation with the increasing distance from an ocean.

Regression and correlation analysis methods are often applied by utilization of spatial data structures and spatial access methods.

Mining Raster Databases:

Spatial database systems usually handle vector data that consists of points, lines, polygons (regions) and their compositions, such as networks or partitions.

Huge amounts of space-related data are in digital raster forms such as satellite images, remote sensing data and computer tomography.

Review Questions

Two Marks:

1. List out some of the application areas of Data mining systems.
2. Is data mining a hype or a persistent?
3. Write short notes on text mining.
4. What are the applications of spatial data bases?
5. Define Spatial Data Mining.
6. List out any five various commercial data mining tools.
7. What are the different Data security techniques used in data mining?
8. What is information retrieval?
9. What is keyword-based association analysis?
10. What is HITS algorithm?
11. List out some of the challenges of WWW.
12. What is web usage mining?
13. What are the three types of dimensions in Spatial data cube?

Sixteen Marks:

1. Discuss in detail the application of Data Mining for financial data analysis?
2. Discuss the application of data mining in business.
3. Discuss in detail of applications of data mining for biomedical and DNA data analysis and telecommunication industry.
4. Discuss the Social impacts of Data Mining Systems.
5. Discuss about the various data mining tools.
6. Explain the Mining of Spatial databases.
7. Discuss the Mining of Text Databases,
8. What is web mining? Discuss the various web mining techniques.

Assignment Topic:

1. Explain in detail about the data mining tool DB-Miner.