UNIT-V

Advanced Concepts

Basic concepts in Mining data streams–Mining Time–series data—Mining sequence patterns in transactional databases– Mining Object– Spatial– Multimedia–Text and Web data – Spatial Data mining– Multimedia Data mining–Text Mining– Mining the World Wide Web.

Mining Data Streams

Tremendous and Potentially infinite volumes of data streams are often generated by real time surveillance systems, communication networks, Internet traffic, on-line transactions in the financial market or retail industry electric power grids, industry production processes, remote sensors and other dynamic environments.

Unlike traditional data sets, stream data flow in and out of a computer system continuously and with varying update rates. They are temporally ordered, fast changing, massive, and potentially infinite. It may be impossible to store an entire data stream or to scan through it multiple times due to its tremendous volume. More-over, stream data tend to be of a rather low level of abstraction, whereas most analysts are interested in relatively high-level dynamic changes, such as trends and deviations. **Data Streams**

- Data streams—continuous, ordered, changing, fast, huge amount
- Traditional DBMS—data stored in finite, persistent data sets

Characteristics

- Huge volumes of continuous data, possibly infinitev Fast changing and requires fast, real-time response
- Data stream captures nicely our data processing needs of today
- Random access is expensive—single scan algorithm (can only havev one look)
- Store only the summary of the data seen thus far

Most stream data are at pretty low-level or multi-dimensional in nature, needs multilevel and multi-dimensional processing

Stream Data Applications

- Telecommunication calling records
- Business: credit card transaction flowsv Network monitoring and traffic engineering
- Financial market: stock exchange
- Engineering & industrial processes: power supply & manufacturing
- Sensor, monitoring & surveillance: video streams, RFIDs Security monitoring
- Web logs and Web page click streams

Mining Time-Series Data

A time-series database consists of sequences of values or events obtained over repeated measurements of time. The values are typically measured at equal time intervals (e.g., hourly, daily, weekly). Time-series databases are popular in many applications, such as stock market analysis, economic and sales fore-casting, budgetary analysis, utility studies, inventory studies, yield projections, work-load projections, process and quality control, observation of natural phenomena (such as atmosphere, temperature, wind, earthquake), scientific and engineering experiments, and medical treatments. A time-series database is also a sequence database. However, a sequence database is any database that consists of sequences of ordered events, with or without concrete notions of time. For example, Web page traversal sequences and cus-tomer shopping transaction sequences are sequence data, but they may not be time series data. Time-series database

Consists of sequences of values or events changing with time

Data is recorded at regular intervals

Characteristic time-series components

Trend, cycle, seasonal, irregular

Applications Financial: stock price, inflation

Industry: power consumptionv Scientific: experiment results

Meteorological: precipitation

Trend Analysis

A time series involving a variable Y, representing, say, the daily closing price of a share in a stock market, can be viewed as a function of time t, that is, Y = F(t). Such a function can be illustrated as a time-series graph, as shown in Figure 8.4, which describes a point moving with the passage of time.

In general there are two goals in time-series analysis: (1) modeling time series (i.e., to gain insight into the mechanisms or underlying forces that generate the time series), and (2) forecasting time series (i.e., to predict the future values of the time-series variables).Trend analysis consists of the following four major components or movements for characterizing time-series data:

1) Trend or long-term movements: These indicate the general direction in which a time-series graph is moving over a long interval of time. This movement is displayed by a trend curve, or a trend line. For example , the trend curve of Figure 8.4 is indicated by a dashed curve. Typical methods for determining a trend curve or trend line include the weighted moving average method and the least squares method, discussed later.

2) Cyclic movements or cyclic variations: These refer to the cycles, that is, the long-term oscillations about a trend line or curve, which may or may not be periodic. That is, the cycles need not necessarily follow exactly similar patterns after equal intervals of time

3) Seasonal movements or seasonal variations: These are systematic or calendar related. Examples include events that recur annually, such as the sudden increase in sales of chocolates and flowers before Valentine's Day or of department store items before Christmas. The observed increase in water consumption in summer due to warm weather is another example. In these examples, seasonal movements are the identical or nearly identical patterns that a time series appears to follow during corresponding months of successive years.

4) Irregular or random movements: These characterize the sporadic motion of time series due to random or chance events, such as labor disputes, floods, or announced personnel changes within companies.

Similarity Search in Time-Series Analysis

1) Normal database query finds exact match

2) Similarity search finds data sequences that differ only slightly from the given query sequence 3) Two categories of similarity queries

- Whole matching: find a sequence that is similar to the query sequence
- Subsequence matching: find all pairs of similar sequences

4) Typical Applications

- Financial market
- Market basket data analysis
- Scientific databases

Mining Sequence Patterns in Transactional Databases

A sequence database consists of sequences of ordered elements or events, recorded with or without a concrete notion of time. There are many applications involving sequence data. Typical examples include customer shopping sequences, Web click streams, bio-logical sequences, sequences of events in science and engineering, and in natural and social developments.

Sequential Pattern Mining: Concepts and Primitives

Sequential pattern mining is the mining of frequently occurring ordered events or subsequences as patterns. An example of a sequential pattern is "Customers who buy a Canon digital camera are likely to buy an HP color printer within a month." For retail data, sequential patterns are useful for shelf placement and promotions. This industry, as well as telecommunications and other businesses, may also use sequential patterns for targeted marketing, customer retention, and many other tasks. Other areas in which sequential patterns can be applied include Web access pat-tern analysis, weather prediction, production processes, and network intrusion detection.

Scalable Methods for Mining Sequential Patterns

Sequential pattern mining is computationally challenging because such mining may generate and/or test a combinatorially explosive number of intermediate subsequences.

"How can we develop efficient and scalable methods for sequential pattern mining?" Recent developments have made progress in two directions: (1) efficient methods for mining the full set of sequential patterns, and (2) efficient methods for mining only the set of closed sequential patterns, where a sequential pattern s is closed if there exists no sequential pattern s 0 where s 0 is a proper super sequence of s, and s 0 has the same (frequency) support as s.

Three such approaches for sequential pattern mining, represented by the algorithms GSP, SPADE, and Prefix Span, respectively. GSP adopts a candidate generate-and-test approach using horizonal data format (where the data are represented as h sequence ID : sequence of itemsets i, as usual, where each itemset is an event). SPADE adopts a candidate generate-and-test approach using vertical data format (where the data are represented as hitemset: (sequence ID, event ID)i). The vertical data format can be obtained by transforming from a horizontally formatted sequence database in just one scan. Prefix Span is a pat-tern growth method, which does not require candidate generation.

All three approaches either directly or indirectly explore the Apriori property, stated as follows: every nonempty subsequence of a sequential pattern is a sequential pattern. (Recall that for a pattern to be called sequential, it must be frequent. That is, it must sat-isfy minimum support.) The Apriori property is antimonotonic (or downward-closed) in that, if a sequence cannot pass a test (e.g., regarding minimum support), all of its super sequences will also fail the test. Use of this property to prune the search space can help make the discovery of sequential patterns more efficient.

Mining Object, Spatial, Multimedia, Text, and Web Data

Multidimensional Analysis and Descriptive Mining of Complex Data Objects

Many advanced, data-intensive applications, such as scientific research and engineering design, need to store, access, and analyze complex but relatively structured data objects. These objects cannot be represented as simple and uniformly structured records (i.e., tuples) in data relations. Such application requirements have motivated the design and development of object-relational and object-oriented database systems. Both kinds of systems deal with the efficient storage and access of vast amounts of disk-based complex structured data objects. These systems organize a large set of complex data objects into classes, which are in turn organized into class/subclass hierarchies. Each object in a class is associated with (1) an object-identifier, (2) a set of attributes that may contain sophisticated data structures, set- or list-valued data, class composition hierarchies, multimedia data, and (3) a set of methods that specify the computational routines or rules associated with the object class. There has been extensive research in the field of database systems on how to efficiently index, store, access, and manipulate complex objects in object-relational and object-oriented database systems. Technologies handling these issues are discussed in many books on database systems, especially on object-oriented and object-relational database systems.

One step beyond the storage and access of massive-scaled, complex object data is the systematic analysis and mining of such data. This includes two major tasks: (1) construct multidimensional data warehouses for complex object data and perform online analytical processing (OLAP) in such data warehouses, and (2) develop effective and scalable methods for mining knowledge from object databases and/or data warehouses. The second task is largely covered by the mining of specific kinds of data (such as spatial, temporal, sequence, graph- or tree-structured, text, and multimedia data), since these data form the

major new kinds of complex data objects. As in Chapters 8 and 9, in this chapter we continue to study methods for mining complex data. Thus, our focus in this section will be mainly on how to construct object data warehouses and perform OLAP analysis on data warehouses for such data.

A major limitation of many commercial data warehouse and OLAP tools for multidimensional database analysis is their restriction on the allowable data types for dimensions and measures. Most data cube implementations confine dimensions to nonnumeric data, and measures to simple, aggregated values. To introduce data mining and multidimensional data analysis for complex objects, this section examines how to perform generalization on complex structured objects and construct object cubes for OLAP and mining in object databases.

To facilitate generalization and induction in object-relational and object-oriented databases, it is important to study how each component of such databases can be generalized, and how the generalized data can be used for multidimensional data analysis and data mining.

Spatial Data Mining

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. Spatial databases have many features distinguishing them from relational databases.

They carry topological and/or distance information, usually organized by sophisticated, multidimensional spatial indexing structures that are accessed by spatial data access methods and often require spatial reasoning, geometric computation, and spatial knowledge representation techniques. Spatial data mining refers to the extraction of knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases. Such mining demands an integration of data mining with spatial database technologies.

It can be used for understanding spatial data, discovering spatial relationships and relationships between spatial and nonspatial data, constructing spatial knowledge bases, reorganizing spatial databases, and optimizing spatial queries. It is expected to have wide applications in geographic information systems, geomarketing, remote sensing, image database exploration, medical imaging, navigation, traffic control, environmental studies, and many other areas where spatial data are used.

A crucial challenge to spatial data mining is the exploration of efficient spatial data mining techniques due to the huge amount of spatial data and the complexity of spatial data types and spatial access methods. "What about using statistical techniques for spatial data mining?" 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. In a statistical model that handles nonspatial data, one usually assumes statistical independence among different portions of data. However, different from traditional data sets, there is no such independence among spatially distributed data because in reality, spatial objects are often interrelated, or more exactly spatially co-located, in the sense that the closer the two objects are located, the more likely they share similar properties.

Such a property of close interdependency across nearby space leads to the notion of spatial autocorrelation. Based on this notion, spatial statistical modeling methods have been developed with good success. Spatial data mining will further develop spatial statistical analysis methods and extend them for huge amounts of spatial data, with more emphasis on efficiency, scalability, cooperation with database and data warehouse systems, improved user interaction, and the discovery of new types of knowledge.

Spatial Data Cube Construction and Spatial OLAP

"Can we construct a spatial data warehouse?" Yes, as with relational data, we can integrate spatial data to construct a data warehouse that facilitates spatial data mining. A spatial data warehouse is a subject-

oriented, integrated, time-variant, and nonvolatile collection of both spatial and nonspatial data in support of spatial data mining and spatial-datarelated decision-making processes

Spatial Clustering Methods

Spatial data clustering identifies clusters, or densely populated regions, according to some distance measurement in a large, multidimensional data set. since cluster analysis usually considers spatial data clustering in examples and applications.

Spatial Classification and Spatial Trend Analysis

Spatial classification analyzes spatial objects to derive classification schemes in relevance to certain spatial properties, such as the neighborhood of a district, highway, or river.

Multimedia Data mining

"What kinds of associations can be mined in multimedia data?" Association rules involving multimedia objects can be mined in image and video databases. At least three categories can be observed:

- Associations between image content and nonimage content features: A rule like "If at least 50% of the upper part of the picture is blue, then it is likely to represent sky" belongs to this category since it links the image content to the keyword sky
- . Associations among image contents that are not related to spatial relationships: A rule like "If a picture contains two blue squares, then it is likely to contain one red circle as well" belongs to this category since the associations are all regarding image contents.
- Associations among image contents related to spatial relationships: A rule like "If a red triangle is between two yellow squares, then it is likely a big oval-shaped object is underneath" belongs to this category since it associates objects in the image with spatial relationships.

To mine associations among multimedia objects, we can treat each image as a transaction and find frequently occurring patterns among different images.

"What are the differences between mining association rules in multimedia databases versus in transaction databases?" There are some subtle differences. First, an image may contain multiple objects, each with many features such as color, shape, texture, keyword, and spatial location, so there could be many possible associations. In many cases, a feature may be considered as the same in two images at a certain level of resolution, but different at a finer resolution level. Therefore, it is essential to promote a progressive resolution refinement approach. That is, we can first mine frequently occurring patterns at a relatively rough resolution level, and then focus only on those that have passed the minimum support threshold when mining at a finer resolution level. This is because the patterns that are not frequent at a rough level cannot be frequent at finer resolution levels. Such a multiresolution mining strategy substantially reduces the overall data mining cost without loss of the quality and completeness of data mining results. This leads to an efficient methodology for mining frequent itemsets and associations in large multimedia databases.

Second, because a picture containing multiple recurrent objects is an important feature in image analysis, recurrence of the same objects should not be ignored in association analysis. For example, a picture containing two golden circles is treated quite differently from that containing only one. This is quite different from that in a transaction database, where the fact that a person buys one gallon of milk or two may often be treated the same as "buys milk." Therefore, the definition of multimedia association and its measurements, such as support and confidence, should be adjusted accordingly.

Third, there often exist important spatial relationships among multimedia objects, such as above, beneath, between, nearby, left-of, and so on. These features are very useful for exploring object associations and correlations. Spatial relationships together with other content-based multimedia features, such as color,

shape, texture, and keywords, may form interesting associations. Thus, spatial data mining methods and properties of topological spatial relationships become important for multimedia mining.

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 of the available information is stored in text databases (or document databases), which consist of large collections of documents from various sources, such as news articles, research papers, books, digital libraries, e-mail messages, and Web pages. Text databases are rapidly growing due to the increasing amount of information available in electronic form, such as electronic publications, various kinds of electronic documents, e-mail, and the World Wide Web (which can also be viewed as a huge, interconnected, dynamic text database). Nowadays most of the information in government, industry, business, and other institutions are stored electronically, in the form of text databases.

Data stored in most text databases are semistructured data in that they are neither completely unstructured nor completely structured. For example, a document may contain a few structured fields, such as title, authors, publication date, category, and so 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 implementation of semistructured data in recent database research. Moreover, information retrieval techniques, such as text indexing methods, have been developed to handle unstructured 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 the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users need tools to compare different documents, rank the importance and relevance of the documents, or find patterns and trends across multiple documents. Thus, text mining has become an increasingly popular and essential theme in data mining.

Text Data Analysis and Information Retrieval

"What is information retrieval?" Information retrieval (IR) is a field that has been developing in parallel with database systems for many years. Unlike the field of database systems, which has focused on query and transaction processing of structured data, information retrieval is concerned with the organization and retrieval of information from a large number of text-based documents. Since information retrieval and database systems each handle different kinds of data, some database system problems are usually not present in information retrieval systems, such as concurrency control, recovery, transaction management, and update.

Also, some common information retrieval problems are usually not encountered in traditional database systems, such as unstructured documents, approximate search based on keywords, and the notion of relevance. Due to the abundance of text information, information retrieval has found many applications. There exist many information retrieval systems, such as on-line library catalog systems, on-line document management systems, and the more recently developed Web search engines.

A typical information retrieval problem is to locate relevant documents in a document collection based on a user's query, which is often some keywords describing an information need, although it could also be an example relevant document.

In such a search problem, a user takes the initiative to "pull" the relevant information out from the collection; this is most appropriate when a user has some ad hoc (i.e., short-term) information need, such as finding information to buy a used car. When a user has a long-term information need (e.g., a researcher's interests), a retrieval system may also take the initiative to "push" any newly arrived information item to a user if the item is judged as being relevant to the user's information need. Such an information access process

is called information filtering, and the corresponding systems are often called filtering systems or recommender systems.

From a technical viewpoint, however, search and filtering share many common techniques. Below we briefly discuss the major techniques in information retrieval with a focus on search techniques.

Mining the World Wide Web

The World Wide Web serves as a huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other information services. The Web also contains a rich and dynamic collection of hyperlink information and Web page access and usage information, providing rich sources for data mining. However, based on the following observations, the Web also poses great challenges for effective resource and knowledge discovery.

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

- The complexity of Web pages is far greater than that of any traditional text document collection. Web pages lack a unifying structure. They contain far more authoring style and content variations than any set of books or other traditional text-based documents. The Web is considered a huge digital library; however, the tremendous number of documents in this library are not arranged according to any particular sorted order. There is no index by category, nor by title, author, cover page, table of contents, and so on. It can be very challenging to search for the information you desire in such a library!
- The Web is a highly dynamic information source. Not only does the Web grow rapidly, but its information is also constantly updated. News, stock markets, weather, sports, shopping, company advertisements, and numerous other Web pages are updated regularly on the Web. Linkage information and access records are also updated frequently.
- The Web serves a broad diversity of user communities. The Internet currently connects more than 100 million workstations, and its user community is still rapidly expanding. Users may have very different backgrounds, interests, and usage purposes. Most users may not have good knowledge of the structure of the information network and may not be aware of the heavy cost of a particular search. They can easily get lost by groping in the "darkness" of the network, or become bored by taking many access "hops" and waiting impatiently for a piece of information.
- Only a small portion of the information on the Web is truly relevant or useful. It is said that 99% of the Web information is useless to 99% of Web users. Although this may not seem obvious, it is true that a particular person is generally interested in only a tiny portion of the Web, while the rest of the Web contains information that is uninteresting to the user and may swamp desired search results. How can the portion of the Web that is truly relevant to your interest be determined? How can we find highquality Web pages on a specified topic?
- These challenges have promoted research into efficient and effective discovery and use of resources on the Internet.

There are many index-based Web search engines. These search the Web, index Web pages, and build and store huge keyword-based indices that help locate sets of Web pages containing certain keywords. With such search engines, an experienced user may be able to quickly locate documents by providing a set of

tightly constrained keywords and phrases. However, a simple keyword-based search engine suffers from several deficiencies. First, a topic of any breadth can easily contain hundreds of thousands of documents. This can lead to a huge number of document entries returned by a search engine, many of which are only marginally relevant to the topic or may contain materials of poor quality. Second, many documents that are highly relevant to a topic may not contain keywords defining them.

For example, the keyword Java may refer to the Java programming language, or an island in Indonesia, or brewed coffee. As another example, a search based on the keyword search engine may not find even the most popular Web search engines like Google, Yahoo!, AltaVista, or America Online if these services do not claim to be search engines on their Web pages. This indicates that a simple keywordbased Web search engine is not sufficient for Web resource discovery. "If a keyword-based Web search engine is not sufficient for Web resource discovery, how can we even think of doing Web mining?" Compared with keyword-based Web search, Web mining is a more challenging task that searches for Web structures, ranks the importance of Web contents, discovers the regularity and dynamics of Web contents, and mines Web access patterns. However, Web mining can be used to substantially enhance the power of a Web search engine since Web mining may identify authoritative Web pages, classify Web documents, and resolve many ambiguities and subtleties raised in keyword-based Web search.