B.C.A. (Sem – VI)

B.C.A. - 603

Data Warehousing & Data Mining

Purushottam Singh

Unit:- 4

Unit - 4

Spatial Mining:-

- Spatial data mining is the application of data mining methods to spatial data.
- The end objective of spatial data mining is to find patterns in data with respect to geography.
- Spatial Data mining is the process of discovering interesting and previously unknown but potentially useful patterns from large spatial datasets.
- Spatial Data mining are more complex than the inputs of classical Data mining because they include extended object such as points, lines and polygons.
- The data inputs of spatial data mining have two distinct types of attributes.
  1. Non Spatial Attributes
  2. Spatial Attributes
- **Non Spatial Attributes** are use to characterize non spatial features of object, such as name population and unemployment rate for a city.
- **Spatial Attributes** are use to define the spatial location and extent of spatial objects.

Temporal Mining:-

- Temporal data mining refers to the extraction of implicit, non-trivial, and potentially useful abstract information from large collections of temporal data.
- Temporal data are sequences of a primary data type, most commonly numerical or categorical values and sometimes multivariate or composite information.
- Temporal data are regular time series (e.g., stock ticks), event sequences (e.g., sensor readings, medical records,), and temporal databases (e.g., relations with timestamped tuples, databases).
- Temporal data mining (TDM) addresses tasks such as segmentation, classification, clustering, forecasting, and indexing of time series, event sequences, or sections of time series or sequences.
- The field of temporal data mining is concerned with such analysis in the case of ordered data streams with temporal interdependencies.
- Temporal data mining were proposed and shown to be useful in many applications.
- Since temporal data mining brings together techniques from different fields such as statistics, machine learning and databases, the literature is scattered among many different sources.

Multimedia Database:-

- A multimedia database is a database that hosts one or more primary media file Types such as .txt (documents), .jpg (images), .mp3 (audio), etc.
- Multimedia database is used to provide information about science, engineering, Medicine, entertainment, Modern biology, social sciences.
- Multimedia database systems will improve the quantity and quality of information.
- A multimedia database system stores and manages a large collection of multimedia objects.
- Such as audio data, image data, video data, sequence data, and hypertext data, which contain text, text mark-ups and linkages.
- Multimedia database systems include NASA's EOS (Earth Observation System), various kinds of image and audio-video databases, human genome databases, and Internet databases.
- **Multimedia Database system have to be capable:-**

  1. Support of multimedia data types.Ex:-data types as data structure, including type of data and operation.
  2. Capability of manage very numerous multimedia objects, store them and search for them
  3. To include a suitable memory management system, to improve performance and high capacity.

Web content mining:-

- Web content mining is related to data mining and text mining.
- It is related to **data mining** because many data mining techniques can be applied in web content mining.
- It is related to **text mining** because much of the web contents or texts.
- Web content consists of several types of data such as textual, image, audio, video, metadata as well as hyperlinks.
- Examine the contents of web pages as well as result of web searching.
- Web Content mining is the process of extracting knowledge from web contents.
- The Web is perhaps the single largest data source in the world.
- Web mining aims to extract and mine useful knowledge from the Web.
- . Web mining generally consists of:
  - ✓ **Web usage mining:-** the discovery of user access patterns from web usage logs.
  - ✓ **Web structure mining:-** the discovery of useful knowledge from the structure of hyperlinks.
  - ✓ **Web content mining:-** mining, extraction and Integration of useful data, information and knowledge from Web page contents.

Web Structure and usage mining:-

Web Structure:- The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages.

- Web Structure Mining can be is the process of discovering structure information from the Web.
- This type of mining can be performed either at the document level or at the hyperlink level.
- The research at the hyperlink level is also called *Hyperlink Analysis*
- **Web Structure Terminology:-**
  - ➢ ***Web-graph:-*** A directed graph that represents the Web.
  - ➢ ***Node:-*** Each Web page is a node of the Web-graph.
  - ➢ ***Link:-*** Each Hyperlink on the web is a directed edge of the web-graph.

Web Usage Mining:-

- A Web is a collection of inter-related files on one or more Web servers.
- Web usage mining is Discovery of meaningful patterns from data generated by client server transaction on one or more web localities.
- Web usage mining is the process of extracting useful information from server logs.
- Web usage mining is the process of finding out what users are looking for on the Internet.

- Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications.
- Web usage Mining is the Following types of data.
  - **Web Server Data:-** The user logs are collected by the Web server. Typical data includes IP address, page reference and access time.
  - **Application Server Data:-** Commercial application servers have significant features to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.
  - **Application Level Data:-** New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the categories above.

Data Generalization and summarization:-

Data Generalization

- Data Generalization is the process of creating successive layers of summary data in an evaluation database.
- It is a process of zooming out to get a broader view of a problem, trend or situation.
- It is also known as rolling-up data.
- Data generalization can provide a great help in Online Analytical Processing (OLAP) technology.
- OLAP is used for providing quick answers to analytical queries which are by nature multidimensional.
- Data generalization is also especially beneficial in the implementation of an online transaction processing (OLTP).
- OLTP refers to a class systems designed for managing and facilitating transaction oriented applications especially those involved with data entry and retrieval transaction processing.
- A process which abstracts a large set of task-relevant data in a database from a low conceptual levels to higher ones.

Data Summarization:-

- Data Summarization summarizes evolutional data included both primitive and derived data, in order to create a derived evaluation data that is general in nature.
- Since the data in the data warehouse is of very high volume, there needs to be a mechanism in order to get only the relevant and meaningful information in a less messy format.
- Data summarization provides the capacity to give data consumers generalize view of disparate bulks of data.
- Data summarization in very large multi-dimensional datasets as in the case of data warehouses is a very challenging work.
- Data summarization is quite a common thing but may require a very powerful and time consuming approach in order to analyze ultra large datasets.
- Data summarization can also be done with a simple spreadsheet application such as Microsoft Excel.

Mining Class Comparison:-

Comparison
- Comparing two or more classes.

Method
- Partition the set of relevant data into the target class and the contrasting classes.
- Generalize both classes to the same high level concepts.
- Compare tuples with the same high level descriptions.
- Present for every tuple its description and two measures:
  - support - distribution within single class
  - comparison - distribution between classes
- Highlight the tuples with strong discriminate features
- **Relevance Analysis:-**
- Find attributes (features) which best distinguish different classes.

Task
- Compare graduate and undergraduate students using discriminate rule.
- DMQL query.

Data Characterization:-

Data characterization is a summarization of the general characteristics or features of a target Class of data. The data corresponding to the user-specified class are typically collected by a Query.
The output of data characterization can be presented in pie charts, bar charts, multidimensional Data cubes, and multidimensional tables. They can also be presented in rule form.

Application:-

- Identify potential customers for e-commerce.
  - Recommendation list of Amazon.com
- Enhance the quality and delivery of Internet information services to the end user.
  - Yahoo!
- Improve Web server system performance.
  - Google
- Includes URL requested, IP address and timestamp
- Clean, condense and transform the raw data of Weblog before performing data mining.
- Multidimensional OLAP analysis
  - Top N users
  - Top N accessed Web pages
  - Most frequently accesses time periods.
- Association patterns, sequential patterns and trends of web accessing.

Additional Themes on mining:-

1. Visual and Audio Data Mining
2. Scientific and Statistical Data Mining
3. Social Impacts of Data Mining

- **Visual and Audio Data Mining:-**

• Visual data mining discovers implicit and useful knowledge from large data using data and/or knowledge visualization techniques.

• The eyes and brain, the latter of which can be thought of as a powerful, highly parallel processing and reasoning engine containing a large knowledge base, control the human Visual system.

• Visual data mining essentially combines the power of these components making it a highly attractive and effective tool for the comprehension of data distributions, patterns, clusters, and outliers in data.

• data visualization and data mining can be integrated in the following ways:

➢ **Data visualization:-**

• Data in a database or data warehouse can be viewed different levels of granularity or Abstraction, or as different combination attributes or dimensions.

• Data can be presented in various visual forms, as box plots, 3-D cubes, data distribution charts, curves, surfaces, link graph and so on.

➢ **Data mining result and visualization:-**

• Visualization of data mining results is presentation of the results or knowledge obtained from data mining in via forms.

• Such forms may include scatter plots and box plots (obtained in descriptive data mining), as well as decision trees, association rule, outliers, generalized rules, and so on.

➢ **Data mining process visualization:-**

• This type of visualization presents the various processes of data mining in visual forms so that users can see how the data are extracted and from which database or data warehouse they are extracted, as well as how the selected data are cleaned, integrated, processed, and mined.

• **Audio Data mining:-**

• Uses audio signals to indicate the patterns of data or the features of data mining results.

• Although visual data mining may disclose interesting patterns using graphical displays, it requires users to concentrate on watching patterns and identifying interesting or novel features within them.

❖ **Scientific and Statistical Data Mining**

• These techniques have been applied extensively to scientific data (e.g., data from experiments in psychology, medicine, electrical engineering and manufacturing), as well as to data from economics and the social sciences.

o **Regression:-**

▪ These methods are used to predict the value of a response (dependent) variable from one or more predictor (independent) variables where the variables are numeric.

▪ There are various forms of regression, such as, linear, multiple, weighted polynomial, non-parametric and robust.

o **Regression trees :-**

▪ These can be used for classification and prediction. The trees constructed are binary.

▪ A regression tree is similar to a decision tree in the sense that tests are performed at the internal nodes.

o **Time series :-**

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

o **Quality control :-**

▪ Various statistics can be used to prepare charts for quality control.

▪ These statistics include the mean, standard deviation, range, count, moving average, moving standard deviation, and moving range.

❖ **Social Impacts of Data Mining :-**

▪ With the fast computerization of society, the social impacts of data mining should not be Underestimated.

- Data mining has recently become very popular, with many people jumping into data mining research, development, or business, or claiming their software systems to be data mining products.
- Data mining will surely help company executives a great deal in understanding the market and their business.

Trends in mining:-

❖**Application Exploration:**

o Early data mining applications put a lot of effort into helping businesses gain a competitive edge.

o The Exploration of data mining for businesses continues to expand as ecommerce and e-marketing have become mainstream in the retail industry.

o Data Mining is increasingly used for the exploration of applications in other areas Such as web and text analysis, financial analysis, industry government, Biomedicine and science.

❖**Scalable and interactive data mining methods:**

o In contrast with traditional data analysis methods, data mining must be able to handle huge amounts of data efficiently and if possible, interactively.

o Because the amount of data being collected continues to increase rapidly, scalable algorithms for individual and integrated data mining functions become essential.

❖**Mining social and information networks:**

o Mining social and information networks and link analysis are critical tasks because such networks are complex.

o The Development of scalable and effective knowledge discovery methods and application for large numbers of network data is essential.

❖**Mining multimedia, text, and web data:**

o Such kind of data is a recent focus in data mining research.

o Great progress has been made, yet there are still many open issues to be solved.

❖**Mining biological and biomedical data:**

o The unique combination of complexity, richness, size and importance of biological and biomedical data warrants special attention in data mining.

o Biological data mining research include mining biomedical literature, link analysis and information integration of biological data by data mining.

❖**Visual and Audio data mining:**

o Visual and audio data mining is an effective way to integrate with humans' visual and audio system and discover knowledge from huge amounts of data.

❖**Distributed data mining and real-time data stream mining:**

o Traditional data mining methods, designed to work at a centralized location, do not work well in many of the distributed computing environments present today.

o Many Application involving stream data ex: - web mining.e-commerce, intrusion Detection, mobile data mining.