OLAP and Data Mining

Data Warehousing and End-User Access Tools

- Accompanying growth in data warehouses is increasing demands for more powerful access tools providing advanced analytical capabilities.

- Key developments include:
  - Online analytical processing (OLAP).
  - SQL extensions for complex data analysis.
  - Data mining tools.

Introducing OLAP

- The dynamic synthesis, analysis, and consolidation of large volumes of multi-dimensional data, Codd (1993).

- Describes a technology that uses a multi-dimensional view of aggregate data to provide quick access to strategic information for purposes of advanced analysis.

Introducing OLAP

- Enables users to gain a deeper understanding and knowledge about various aspects of their corporate data through fast, consistent, interactive access to a wide variety of possible views of the data.

- Allows users to view corporate data in such a way that it is a better model of the true dimensionality of the enterprise.
Introducing OLAP

◆ Can easily answer ‘who?’ and ‘what?’ questions, however, ability to answer ‘what if?’ and ‘why?’ type questions distinguishes OLAP from general-purpose query tools.

◆ Types of analysis ranges from basic navigation and browsing (slicing and dicing) to calculations, to more complex analyses such as time series and complex modeling.

Examples of OLAP Applications in Various Functional Areas

<table>
<thead>
<tr>
<th>Functional area</th>
<th>Examples of OLAP applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finance</td>
<td>Budgeting, activity-based costing, financial performance analysis, and financial modeling</td>
</tr>
<tr>
<td>Sales</td>
<td>Sales analysis and sales forecasting</td>
</tr>
<tr>
<td>Marketing</td>
<td>Market research analysis, sales forecasting, promotions analysis, customer analysis, and market/customer segmentation</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>Production planning and defect analysis</td>
</tr>
</tbody>
</table>

OLAP Applications

◆ Although OLAP applications are found in widely divergent functional areas, all have following key features:
  – multi-dimensional views of data;
  – support for complex calculations;
  – time intelligence.

OLAP Applications - Multi-Dimensional Views of Data

◆ Core requirement of building a ‘realistic’ business model.

◆ Provides basis for analytical processing through flexible access to corporate data.

◆ The underlying database design that provides the multi-dimensional view of data should treat all dimensions equally.
OLAP Applications - Support for Complex Calculations

- Must provide a range of powerful computational methods such as that required by sales forecasting, which uses trend algorithms such as moving averages and percentage growth.

- Mechanisms for implementing computational methods should be clear and non-procedural.

OLAP Applications – Time Intelligence

- Key feature of almost any analytical application as performance is almost always judged over time.

- Time hierarchy is not always used in same manner as other hierarchies.

- Concepts such as year-to-date and period-over-period comparisons should be easily defined.

OLAP Benefits

- Increased productivity of end-users.
- Reduced backlog of applications development for IT staff.
- Retention of organizational control over the integrity of corporate data.
- Reduced query drag and network traffic on OLTP systems or on the data warehouse.
- Improved potential revenue and profitability.

Representing Multi-Dimensional Data

- Example of two-dimensional query.
  - What is the total revenue generated by property sales in each city, in each quarter of 1997?

- Choice of representation is based on types of queries end-user may ask.

- Compare representation - three-field relational table versus two-dimensional matrix.
Multi-Dimensional Data as Three-Field Table versus Two-Dimensional Matrix

Representing Multi-Dimensional Data

◆ Example of three-dimensional query.
  – ‘What is the total revenue generated by property sales for each type of property (Flat or House) in each city, in each quarter of 1997?’

◆ Compare representation - four-field relational table versus three-dimensional cube.

Multi-Dimensional Data as Four-Field Table versus Three-Dimensional Cube

Representing Multi-Dimensional Data

◆ Cube represents data as cells in an array.

◆ Relational table only represents multi-dimensional data in two dimensions.
Multi-Dimensional OLAP Servers

- Use multi-dimensional structures to store data and relationships between data.
- Multi-dimensional structures are best visualized as cubes of data, and cubes within cubes of data. Each side of cube is a dimension.
- A cube can be expanded to include other dimensions.

A cube supports matrix arithmetic.

- Multi-dimensional query response time depends on how many cells have to be added ‘on the fly’.
- As number of dimensions increases, number of the cube’s cells increases exponentially.

Multi-Dimensional OLAP Servers

- However, majority of multi-dimensional queries use summarized, high-level data.
- Solution is to pre-aggregate (consolidate) all logical subtotals and totals along all dimensions.
- Pre-aggregation is valuable, as typical dimensions are hierarchical in nature.
  - (e.g. Time dimension hierarchy - years, quarters, months, weeks, and days)

Predefined hierarchy allows logical pre-aggregation and, conversely, allows for a logical ‘drill-down’.

- Supports common analytical operations
  - Consolidation.
  - Drill-down.
  - Slicing and dicing.
Multi-Dimensional OLAP Servers

- **Consolidation** - aggregation of data such as simple ‘roll-ups’ or complex expressions involving inter-related data.
- **Drill-Down** - is reverse of consolidation and involves displaying the detailed data that comprises the consolidated data.
- **Slicing and Dicing** - (also called pivoting) refers to the ability to look at the data from different viewpoints.

Multi-Dimensional OLAP servers

- Can store data in a compressed form by dynamically selecting physical storage organizations and compression techniques that maximize space utilization.
- Dense data (i.e., data that exists for high percentage of cells) can be stored separately from sparse data (i.e., significant percentage of cells are empty).

Multi-Dimensional OLAP Servers

- Ability to omit empty or repetitive cells can greatly reduce the size of the cube and the amount of processing.
- Allows analysis of exceptionally large amounts of data.

Multi-Dimensional OLAP Servers

- In summary, pre-aggregation, dimensional hierarchy, and sparse data management can significantly reduce the size of the cube and the need to calculate values ‘on-the-fly’.
- Removes need for multi-table joins and provides quick and direct access to arrays of data, thus significantly speeding up execution of multi-dimensional queries.
Codd’s Rules for OLAP Systems

- In 1993, E.F. Codd formulated twelve rules as the basis for selecting OLAP tools.
  - Multi-dimensional conceptual view
  - Transparency
  - Accessibility
  - Consistent reporting performance
  - Client-server architecture
  - Generic dimensionality

Codd’s Rules for OLAP

- Dynamic sparse matrix handling
- Multi-user support
- Unrestricted cross-dimensional operations
- Intuitive data manipulation
- Flexible reporting
- Unlimited dimensions and aggregation levels.

Categories of OLAP Tools

- OLAP tools are categorized according to the architecture of the underlying database.

- Three main categories of OLAP tools include
  - Multi-dimensional OLAP (MOLAP or MD-OLAP)
  - Relational OLAP (ROLAP), also called multi-relational OLAP
  - Managed query environment (MQE)
Multi-Dimensional OLAP (MOLAP)

- Uses specialized data structures and multi-dimensional Database Management Systems (MDDBMSs) to organize, navigate, and analyze data.

- Data is typically aggregated and stored according to predicted usage to enhance query performance.

Multi-Dimensional OLAP (MOLAP)

- Use array technology and efficient storage techniques that minimize the disk space requirements through sparse data management.

- Provides excellent performance when data is used as designed, and the focus is on data for a specific decision-support application.

Multi-Dimensional OLAP (MOLAP)

- Traditionally, require a tight coupling with the application layer and presentation layer.

- Recent trends segregate the OLAP from the data structures through the use of published application programming interfaces (APIs).

Typical Architecture for MOLAP Tools

[Diagram of typical architecture for MOLAP tools]

- Relational database server and/or legacy systems
- Load
- Database and application logic layer
- MOLAP server
- Data request
- Result set
- End-user access tools
- Presentation layer
MOLAP Tools - Development Issues

- Underlying data structures are limited in their ability to support multiple subject areas and to provide access to detailed data.

- Navigation and analysis of data is limited because the data is designed according to previously determined requirements.

MOLAP Tools - Development Issues

- MOLAP products require a different set of skills and tools to build and maintain the database, thus increasing the cost and complexity of support.

Relational OLAP (ROLAP)

- Fastest growing style of OLAP technology.

- Supports RDBMS products using a metadata layer - avoids need to create a static multi-dimensional data structure - facilitates the creation of multiple multi-dimensional views of the two-dimensional relation.

Relational OLAP (ROLAP)

- To improve performance, some products use SQL engines to support complexity of multi-dimensional analysis, while others recommend, or require, the use of highly denormalized database designs such as the star schema.
Typical Architecture for ROLAP Tools

- Middleware to facilitate the development of multi-dimensional applications. (Software that converts the two-dimensional relation into a multi-dimensional structure).
- Development of an option to create persistent, multi-dimensional structures with facilities to assist in the administration of these structures.

ROLAP Tools - Development Issues

Managed Query Environment (MQE)
- Relatively new development.
- Provide limited analysis capability, either directly against RDBMS products, or by using an intermediate MOLAP server.

Managed Query Environment (MQE)
- Deliver selected data directly from DBMS or via a MOLAP server to desktop (or local server) in form of a datacube, where it is stored, analyzed, and maintained locally.
- Promoted as being relatively simple to install and administer with reduced cost and maintenance.
Typical Architecture for MQE Tools

MQE Tools - Development Issues

- Architecture results in significant data redundancy and may cause problems for networks that support many users.
- Ability of each user to build a custom datacube may cause a lack of data consistency among users.
- Only a limited amount of data can be efficiently maintained.

Data Mining

- The process of extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions (Simoudis, 1996).
- Involves analysis of data and use of software techniques for finding hidden and unexpected patterns and relationships in sets of data.

Data Mining

- Reveals information that is hidden and unexpected, as little value in finding patterns and relationships that are already intuitive.
- Patterns and relationships are identified by examining the underlying rules and features in the data.
- Tends to work from the data up and most accurate results normally require large volumes of data to deliver reliable conclusions.
Data Mining

- Starts by developing an optimal representation of structure of sample data, during which time knowledge is acquired and extended to larger sets of data.

- Data mining can provide huge paybacks for companies who have made a significant investment in data warehousing.

- Relatively new technology, however already used in a number of industries.

Examples of Applications of Data Mining

- Retail / Marketing
  - Identifying buying patterns of customers.
  - Finding associations among customer demographic characteristics.
  - Predicting response to mailing campaigns.
  - Market basket analysis.

- Banking
  - Detecting patterns of fraudulent credit card use.
  - Identifying loyal customers.
  - Predicting customers likely to change their credit card affiliation.
  - Determining credit card spending by customer groups.

- Insurance
  - Claims analysis.
  - Predicting which customers will buy new policies.

- Medicine
  - Characterizing patient behavior to predict surgery visits.
  - Identifying successful medical therapies for different illnesses.
Data Mining Operations

- Four main operations include:
  - Predictive modeling.
  - Database segmentation.
  - Link analysis.
  - Deviation detection.

- There are recognized associations between the applications and the corresponding operations.
  - e.g. Direct marketing strategies use database segmentation.

Data Mining Techniques

- Techniques are specific implementations of the data mining operations.

- Each operation has its own strengths and weaknesses.

- Data mining tools sometimes offer a choice of operations to implement a technique.

Data Mining Operations and Associated Techniques

<table>
<thead>
<tr>
<th>Operations</th>
<th>Data mining techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive modeling</td>
<td>Classification</td>
</tr>
<tr>
<td></td>
<td>Value prediction</td>
</tr>
<tr>
<td>Database segmentation</td>
<td>Demographic clustering</td>
</tr>
<tr>
<td></td>
<td>Neural clustering</td>
</tr>
<tr>
<td>Link analysis</td>
<td>Association discovery</td>
</tr>
<tr>
<td></td>
<td>Sequential pattern discovery</td>
</tr>
<tr>
<td>Deviation detection</td>
<td>Similar time sequence discovery</td>
</tr>
<tr>
<td></td>
<td>Statistics</td>
</tr>
<tr>
<td></td>
<td>Visualization</td>
</tr>
</tbody>
</table>
Predictive Modeling

- Similar to the human learning experience – uses observations to form a model of the important characteristics of some phenomenon.

- Uses generalizations of ‘real world’ and ability to fit new data into a general framework.

- Can analyze a database to determine essential characteristics (model) about the data set.

Predictive Modeling

- Model is developed using a supervised learning approach, which has two phases: training and testing.
  - Training builds a model using a large sample of historical data called a training set.
  - Testing involves trying out the model on new, previously unseen data to determine its accuracy and physical performance characteristics.

Predictive Modeling

- Applications of predictive modeling include customer retention management, credit approval, cross selling, and direct marketing.

- Two techniques associated with predictive modeling: classification and value prediction, distinguished by nature of the variable being predicted.

Predictive Modeling - Classification

- Used to establish a specific predetermined class for each record in a database from a finite set of possible class values.

- Two specializations of classification: tree induction and neural induction.
Example of Classification using Tree Induction

Example of Classification using Neural Induction

Predictive Modeling - Value Prediction

- Used to estimate a continuous numeric value that is associated with a database record.

- Uses the traditional statistical techniques of linear regression and nonlinear regression.

- Relatively easy to use and understand.

Predictive Modeling - Value Prediction

- Data mining requires statistical methods that can accommodate non-linearity, outliers, and non-numeric data.

- Applications of value prediction include credit card fraud detection or target mailing list identification.
Database Segmentation

- Aim is to partition a database into an unknown number of segments, or clusters, of similar records.

- Uses unsupervised learning to discover homogeneous sub-populations in a database to improve the accuracy of the profiles.

Database Segmentation

- Less precise than other operations thus less sensitive to redundant and irrelevant features.

- Sensitivity can be reduced by ignoring a subset of the attributes that describe each instance or by assigning a weighting factor to each variable.

- Applications of database segmentation include customer profiling, direct marketing, and cross selling.

Example of Database Segmentation using a Scatterplot

- Associated with demographic or neural clustering techniques, distinguished by:
  - Allowable data inputs.
  - Methods used to calculate the distance between records.
  - Presentation of the resulting segments for analysis.
Link Analysis

- Aims to establish links (associations) between records, or sets of records, in a database.

- There are three specializations
  - Associations discovery.
  - Sequential pattern discovery.
  - Similar time sequence discovery.

- Applications include product affinity analysis, direct marketing, and stock price movement.

Link Analysis - Associations Discovery

- Finds items that imply the presence of other items in the same event.

- Affinities between items are represented by association rules.
  - e.g. ‘When customer rents property for more than 2 years and is more than 25 years old, in 40% of cases, customer will buy a property. Association happens in 35% of all customers who rent properties’.

Link Analysis - Sequential Pattern Discovery

- Finds patterns between events such that the presence of one set of items is followed by another set of items in a database of events over a period of time.
  - e.g. Used to understand long-term customer buying behavior.

Link Analysis - Similar Time Sequence Discovery

- Finds links between two sets of data that are time-dependent, and is based on the degree of similarity between the patterns that both time series demonstrate.
  - e.g. Within three months of buying property, new home owners will purchase goods such as cookers, freezers, and washing machines.
Deviation Detection

- Relatively new operation in terms of commercially available data mining tools.
- Often a source of true discovery because it identifies outliers, which express deviation from some previously known expectation and norm.

Deviation Detection

- Can be performed using statistics and visualization techniques or as a by-product of data mining.
- Applications include fraud detection in the use of credit cards and insurance claims, quality control, and defects tracing.

Example of Database Segmentation using a Visualization

Data Mining Tools

- There are a growing number of commercial data mining tools on the marketplace.
- Important characteristics of data mining tools include:
  - Data preparation facilities.
  - Selection of data mining operations.
  - Product scalability and performance.
  - Facilities for visualization of results.
Data Mining and Data Warehousing

- Major challenge to exploit data mining is identifying suitable data to mine.
- Data mining requires single, separate, clean, integrated, and self-consistent source of data.

Data Mining and Data Warehousing

- A data warehouse is well equipped for providing data for mining.
- Data quality and consistency is a prerequisite for mining to ensure the accuracy of the predictive models. Data warehouses are populated with clean, consistent data.

Data Mining and Data Warehousing

- Advantageous to mine data from multiple sources to discover as many interrelationships as possible. Data warehouses contain data from a number of sources.
- Selecting relevant subsets of records and fields for data mining requires query capabilities of the data warehouse.

Data Mining and Data Warehousing

- Results of a data mining study are useful if there is some way to further investigate the uncovered patterns. Data warehouses provide capability to go back to the data source.