**Document: Unit-2 DM himani (1).pdf**

**=== Page 1 ===**

**1**

5. Representation for visualizing the discovered patterns. 4. Interestingness measures and thresholds for pattern evaluation. 3. Background knowledge to be used in the discovery proce ss. 2. Kind of knowledge to be mined. 1. Set of task - relevant data to be mined.

** The key primitives are:**

users to interact with data mining systems flexibly.  A data mining query language can be designed to incorporate these primitives, allowing primitives. the data mining system. A data mining query is defined in terms of data mining task  A data mining task can be specified in the form of a data mining query, which is input to how to evaluate the results . They allow users to specify what data to use , what kind of knowledge to discover , and  Data mining task primitives are the basic building blocks that define a data mining task.

**DATA MINING TASK PRIMITIVES:**

**UNIT - 2 DATA MINING PRIMITIVES, LANGUAGE, ARCHITECTURE**

**=== Page 2 ===**

**2**

hierarchy for a given attribute or dimensi on. concepts. Based on different user viewpoints, there may be more than one concept  Drilling Down - Specialization of data: Concept values replaced by lower - level would require fewer input/output operations. explicit abstractions and makes it easier to understand. It compresses the data , and it  Rolling Up - Generalization of data: Allow to view data at more meaningful and sequence of mappings from low - level concepts to higher - level. Such as:

allows data to be mined at multiple levels of abstraction. C oncept hierarchy defines a  For example , Concept hierarchies are a popular form of background knowledge, which the insights obtaine d from the data mining process.  The use of background knowledge can help to improve the accuracy and relevance of trends as well as knowledge about the data itself.  This can include domain - specific knowledge, such as industr y - specific terminology, process.  It refers to any prior information or understanding that is used to guide the data mining

3. Background knowledge to be used in the discovery process: outlier detection, and correlation analysis. mine useful information such as classification, clustering, prediction, discrimination,  For example , it determines the task to be performed on the relevant data in order to characterization, assoc iation, and evolution analysis.  It includes various tasks such as classification, clustering, discrimination, techniques. This describes the data mining tasks or functions to be performed .

 It refers to the type of information that are being required through the use of data minin g

**2. Kind of knowledge to be mined:**

attributes from the dataset from the provided input database.  For example : Extracting the database name, database tables, and relevant required data may be necessary or relevant for the task.  The data selected for mining is typically a subset of the overall data available, as not all to the task at hand, such as customer d ata , sales data, or website usage statistics.  This data may include specific attributes, variables, or characteristics that are relevant being conducted using data mining techniques. This specifies the data to be mined.

 It refers to the specific data that is relevant and necessary for a particular task or analysis

**1. Set of task - relevant data to be mined:**

**=== Page 3 ===**

**3**

visualization techniques such as bar graphs , charts, graphs, tables, etc.  For example : Presentation and visualization of discovered pattern of data using various  Decision trees: Representing classific ation models in a tree - like structure.  Charts and graphs: Visualizing patterns using different types of diagrams.  Tables: Presenting patterns in a structured format.  Rules: Expressing patterns in the form of "if - then" statements.

**visualiz ation techniques include:**

 This specifies how the discovered patterns should be presented to the user. Common technical stakeholders. mining process more accessible and understandable to a wider audience, including non -  Visualizing the discovered pat tern helps to make the patterns obtained from the data within the data. represent the data and can help to highlight important trends, patterns, or relationships  Visualization techniques such as charts, graphs, and maps are commonly used to a way that is easy to understand and interpret.

 It refers to the methods used to represent the patterns discovered through data mining in

**5. Representation for visualizing the discovered patterns:**

novelty for the data and setting an appropriate threshold value for the pattern evaluation.  For example : Evaluating the interestingness measures such as utility, certainty, and  Confidence: How likely it is that a pattern is true.  Support: How often a pattern occurs in the data. help to ide ntify patterns that are truly meaningful and useful. Examples include:  These are used to evaluate the quality and relevance of the discovered patterns. They that pattern must meet in order to be considered for further analysis or action.

 Thr eshol ds for pattern evaluation are used to set a minimum level of i nterestingness  These measures are used to identify patterns that are mean ingful or relevant to the task. confidence . considered to be interesting or relevant based on certain criteria, such as its support or  Interestingness measures are used to quantify the degree to which a pattern is patterns or insights discovered through data mining.  It refers to the methods and criteria used to evaluate the quality and relevance of the

4. Interestingness measures and thresholds for pattern evaluation:

**=== Page 4 ===**

**4**

over time. Finding anomalies or changes in patterns Detecting unusual drops in sales for a product

**Deviation Analysis: Example:**

sequences over time. data. Identifying patterns that occur in Predicting stock price trends based on historical

**Sequential Pattern Mining: Example:**

based on input data. location, size, and number of rooms Predicting continuous numerical values Predicting house prices based on features lik e

**Regression Analysis: Example:**

differ from others . Identifying data points that significantly Detecting fraudulent transactions in banking.

**Outlier Detection: Example:**

similarities behaviour. Grouping data into clusters based on Segmenting customers based on purchasing

**Clustering: Example:**

categories or classes. Spam." Assigning dat a objects to predefined Email classification into "Spam" and "Not

**Classification: Example:**

attributes in a dataset. Identifying relationships between different "Customers who buy bread often buy butter."

**Association Rule Mining: Example:**

differences. performing and low - performing students. Comparing two or more datasets to find C omparing the characteristics of high -

**Data Discrimination: Example:**

data in a target class. employees in an organization. Summarizing general characteristics of Finding the average salary and job roles of

**Data Characterization: Example:**

**DATA MINING TASK S with examples :**

**=== Page 5 ===**

**5**

7. { with [ (type\_of) ] threshold = (threshold\_value) [ for(attribute(s)) ] } 6. [ order by(order\_list) ] 5. from (relation(s)) [ where(condition) ] 4. related to (attribute\_or\_aggreagat e\_list) 3. (rule\_specified) 2. { use hierarchy (hierarchy\_name) for (attribute) } 1. use database (database\_name)

**Syntax -- To retrieve relevant dataset:**

Basic syntax in DMQL: DMQL acquires syntax like the relational query language, SQL . 5. Presentation of Results: Specifies how results should be visualized 4. Interestingness Measures: Defines criteria for meaningful patterns 3. Background Knowledge: Includes additional rules, constraints, or hierarchies. Discovered : classification, prediction, clustering) 2. Type of Knowledge to be Defines the type of pattern (e.g., association, 1. Data to be Mined : Specifies the dataset.

Basic Structure of DMQL : A DMQL query consists of the following main components: and integrates with the overall information processi ng environment.  This facilitates a data mining system's communication with other information systems power, limitation, and fundamental mechanisms of various kinds of data mining tasks.  The design of an effective data mining query languag e requires deep understanding of task has different requirements. covers a wide spectrum of tasks, from data characterization to evolution analysis. Each

 Designing a comprehensive data mining language is challenging because data mining  How results should be presented  Which patterns to discover  What data to mine

** It helps users specify:**

similar to how SQL is used for querying databases. data mining tasks in dat abases. DMQL provides a structured way to extract patterns,  Data Mining Query Language (DMQL) is a specialized query language used to define

**DATA MINING QUERY LANGUAGE (DMQL):**

**=== Page 6 ===**

**6**

3. High - income customers 2. Middle - income customers 1. Low - income customers

**The system clusters customers into three groups:**

**Output:**

with k = 3;  with k = 3: Creates 3 clusters. using k - means  using k - means: Uses the k - means clustering algorithm. from CustomerData  from CustomerData: Uses data from CustomerData table. mine clusters  mine clusters: Specifies that clustering should be performed.

**DMQL Query: Explanation:**

Clustering groups data into meaningful clusters based on similarities.

**Example - 1 Clustering Analysis:**

Characterization: find characteristic rules [as (rule\_name)] Classification: find classification rules [as (rule\_name) ] according to [(attribute)] Association: find association rules [as ( rule\_name)] Generalization: generalize data [into (relation\_name)]

**RULES SYNTAX**

For the rule \_specified in DMQL , The syntax is given below: with respect to attributes.  In sixth line, Then they are or dered using “order by” for a designated threshold value  In fifth line, FROM and WHERE clauses make s sure of given condition being satisfied. on the attribute or aggregation which helps in generalization.  In fourth line, To find out the various specified rules, one must find the related set based  In third line, ( rule\_specified ) denotes the types of rules to be specified.

 The second line uses hierarchy , one has chosen(hierarchy\_name) with given attribute.  T he first line retrieves the required database (database\_name). In the above syntax of Data Mining Q uery,

**=== Page 7 ===**

**7**

(Country) to the lowest (City).  Defines the levels of hierarchy from the highest as (Country, State, City); attribute in the Customer table. on Customer(Location)  Specifies that the hierarchy applies to the Location define hierarchy location\_hierarchy  Defines a new hierarchy named location\_hierarchy.

**D MQL Query: Explanation:**

we can define a hierarchy for location data (Country → State → City). A concept hierarchy allows data to be viewed at different levels of abstraction. For example, Example - 3 : Concept Hierarchy Definition chance they also buy butter"). 70%; least 70% (e.g., "if a customer buys bread, there is a 70% and confidence threshold = Specifies th at the confidence level for a rule must be at 30% transactions to be considered frequent. with support threshold = Means that an itemset must appear in at least 30% of the

frequently. extractin g frequent patterns Indicates that we want to find itemsets that appear from TransactionData Specifies table (or dataset) that contains transaction data. mine association rules Instructs the system to perform association rule mining. use database RetailDB; Specifies the database that contains the transaction data.

**Explanation of the Query:**

and confidence threshold = 70%; with support threshold = 30% extracting frequent pa tterns from TransactionData Butter too. customers who buy Milk and Bread often buy mine association rules {Milk, Bread} → {Butter}, Which means use database RetailDB; The system will return frequent itemsets such as

**DMQL Query: Output:**

find frequent itemsets that customers often purchase together. Problem: Suppose we have a retail dataset containing customer transactions, and we want to Example - 2: Association Rule Mining in DMQL

**=== Page 8 ===**

**8**

**DATA MINING ARCHITECTURE:**

data mining techniques to write effective DMQL queries. 3. Requires Strong Understanding of Data Mining Concepts : Users must be familiar with can be slow if not optimized properly. 2. Performance Issues with Large Datasets : Executing DMQL queries on massive d atasets challenging for users without experience in query languages. 1. Complex Syntax for Beginners : While it simplifies some tasks, DMQL can still be

**DISADVANTAGES OF DMQL:**

complex mining operations compared to traditional programming or SQL - based queries. 4. User - Friendly & Simplifies Complex Queries : It provides easier way for users to specify association rule mining, and other tasks. 3. Supports Multiple Data Mi ning Functions : DMQL supports classification, clustering, warehouses, allowing seamless data retrieval and mining. 2. Integration with Databases : It can be integrated with traditional databases and data data mining, making it easier to define and execute mining tasks.

1. Standardized Query Language: DMQL provides a structured and systematic approach to

**ADVANTAGES OF DMQL:**

**=== Page 9 ===**

**9**

The list of Integration Schemes is as follows - effective algorithms for mining the available data sets.  In this scheme, the main focus is on data mining design and on developing efficient and interact w ith databases , how they process data , and their level of integration .  Data mining systems can be categorized into different architectures based on how they

**VARIOUS ARCHITECTURES OF DATA MINING SYSTEMS:**

4. The objective of the knowledge base is to make the result more accurate and reliable. 3. This knowledge base may contain da ta from user experiences. 2. Data mining engines may also sometimes get inputs from the knowledge base. beneficial in guiding the search for the result patterns. 1. Knowledge Base is an important part of the data mining engine that is quite

**6. Knowle dge Base:**

mining system. graphical user interface helps the user to communicate effectively with the data 1. Since the user cannot fully understand the complexity of the data mining process so

**5. Graphic User Interface:**

also interact with the database servers fo r producing the result of the user requests. 1. They are responsible for finding interesting patterns in the data and sometimes they

**4. Pattern Evaluation Modules:**

kinds of data mining techniques like association, classification, clustering etc. 1. It is one of the core components of the data mining architecture that performs all

**3. Data Mining Engine:**

2. It performs the task of handling data retrieval as per the request of the user. 1. The database server contains the actual data ready to be processed.

**2. Database Server:**

3. WWW is one of the biggest sources of data . forms of media like photos or videos. 2. The data in these sources may be in the form of plain text, spreadsheets, or other 1. Database, World Wide Web(WWW), and data warehouse are parts of data sources.

**1. Data Sources:**

**Components of Data Mining Architecture:**

**=== Page 10 ===**

**10**

 Better scalability for big data applications. database systems.  Supports real - time mining.  Requires modifications to existing  High efficiency and performance .  Complex implementation.

**Pros: Cons:**

built - in mining functions. No need for separate data extraction. mining is performed directly inside the database system. It u ses SQL - based queries or  In this, Fully integrated with database systems or data warehouses. In this, Data

**4. Tight Coupling Architecture :**

 Takes advantage of database capabilities. inefficiencies.  Faster processing than loose coupling.  Still not fully integrated, leading to

**Pros: Cons:**

performed within the database. The min ing system still works as a separate module. indexing, query optimization). Here, s ome pre - processing and mining tasks are warehouses. In this, Data mining components use database functionalities (like  In this, Partial integration of the data mining system with databases or data

**3. Semi - Tight Coupling Architecture :**

extraction and storage.  Can work with multiple data sources.  Additional steps are required for data  More flexible than no - coupling architecture.  Not optimized for real - time mining.

**Pros: Cons:**

separately. Results are stored back in the system or displayed to the user. from databases or data warehouses. The mining system processes the extracted data  In this, Data mining system accesses databases indirectly. In this, Data is extracted

**2. Loose Coupling Architecture :**

research support.  Useful for small datasets or experimental  No real - time or dynamic querying  Simple to implement.  Inefficient for large datasets.

**Pros: Cons:**

before running mining algorithms. independently without direct access to databases. Users must manually extract data warehouse. In this, Data is stored separately in local files. Data mining is performed  It is a Standalone data mining system that does not interact with a database or data

**1. No - Coupling Architecture :**

**=== Page 11 ===**

**11**

certain disease based on their medical history and lifestyle factors. 3. Identifying risk factors: Characterizing individuals with a high risk of developing a trends and opportunities. 2. Analysing market trends: Comparing product sales across different regions to identify 1. Understanding customer behaviour: It means customer segmentation in marketing.

**Applications of Concept Description :**

4. Visualization 3. Rule - Based Description 2. Data summariza tion/ Statistical Summarization 1. Attribute - Oriented Induction (AOI)

**T echniques of Concept Description :**

incomes compared to those who buy budget items."  For e xample: "Customers who purchase luxury items tend to have higher annual It helps in distinguishing one category from another based on k ey attributes.  It c ompares two or more classes / datasets to identify their differences and similarities.

**2. Discrimination / Comparison:**

online shopping."  For example: "The majority of high - spending customers are aged 25 - 40 and prefer purchase a product).  It provides a general summary of a specific set of data (e.g., customers who frequently measures such as mean, median, mode, and standard deviation. identifying common features, patterns, and trends within the data. It uses statistical  It s ummarizes the general characteristics of a target class or concept. This involves

**1. Characterization :**

**Types of Concept Description:**

comparing different concepts or classes within a dataset.  It a descriptive task that aim s to provide summaries of data. It involves characterizing and summarized information about a dataset . to extract meaningful patterns, trends, and relationships. It involves generating  C oncept D escription refers to the process of summarizing and characterizing a set of data

**CONCEPT DESCRIPTION: CHARACTERIZATION AND COMPARISIONS**

**=== Page 12 ===**

**12**

A significant percentage have unstable employment histories . Findings d efaulters often have low credit scores . Many have multiple high - interest loans . 3. Banking and Finance: It c haracterizes individuals likely to default on loans. Example: history of diabetes . A majority have sitting lifestyles with high carbohydrate intake . Example: Findings Most diabetic patients are over 40 year s old . Many have a family 2. Healthcare: Characterize diabetic patients based on lifestyle and medical history.

over in - store purchases. They purchase electronics and fashion items frequently. Example: Finding most high - value customers are aged 25 - 40 . They prefer online shopping 1. Retail and Marketing: Identify ing the characteristics of high - value d customers.

**Applications of Characterization :**

3. It helps in personalized marketing and targeted campaigns. 2. Summarized data helps businesses and researchers to simplify decision making. 1. It h elps in understanding key t rends and patterns in the data.

**Advantages of Characterization :**

6. Visualization & Presentation: Represent data using charts, graphs, and reports. classification rules. 5. Pattern Extraction and Rule Generation: Identify trends, associations, and 4. Data Summarization: Find statistics like mean, median, and frequency. 3. Attribute - Oriented Induction (AOI): Generalize raw data into higher - level concepts. 2. Data Pre - processing: Clean, normalize, and prepare data for analysis. 1. Data Selection: Choose relevant data for characterization.

Steps in Characterization: (Data Characterization Process) base d on their purchasing patterns and shopping behaviour .  For example , in a retail business, characterization might help describe frequent customers behaviours, and patterns efficiently. summarization , rule - based descriptions , and visualization can identify trends,  By using techniques like attribute - oriented induction (AOI) , statistical makers understand key properties of data and make data - driv en decisions.

the general features of a specific dataset. It helps businesses, researchers, and decision -  Characterization is a fundamental concept in data mining used to describe and summarize

**CHARACTERIZATION:**

**=== Page 13 ===**

**13**

customers from different age groups in a store. are used to display summarized data. Example: A bar chart showing the percentage of communicate the characteristics of a class, making it easier to understand the dataset . They 4. Visualization Techniques : Charts, graphs, and other visual representations can effectively businesses target high - income individuals for luxury car promotions. Age > 30 and Income > $50,000 then Likely to Buy a Luxury Car. This rule helps 3. Rule - Based Description : It u ses associatio n rules or classification rules. For e xample: If

o Standard Deviation : Meas ures data spread or variability. o Mode : The most frequently occurring value. o Median : The middle value when data is sorted. o Mean (Average): The central value of a numerical attribute. department is $80,000, with a standard deviation of $10,000." distribution and key features. Example: "The average salary of employees in the IT generating box plots provide s a visual and quantitative overview of the data's mode. Techniques like calculating summary statistics, creating histograms, and

 It p rovides an overview of data using statistical measures, such as mean, median, and

**2. Statistical Summarization :**

segments based on its purchasing behaviour. example , instead of describing individual customers, we might describe customer  Data Generalization involves summariz ing data to a higher level of abstraction. For Example: If dataset contains "Age = 21, 22, 23," it can be generalized to "young adult." data. It generalizes data by replacing specific values with higher - level concepts .  A technique that uses data generalization to extract characteristic rules and summarize

**1. Attribute - Oriented Induction (AOI) :**

**Techniques of Characterization :**

**=== Page 14 ===**

**14**

performing stocks show low demand and declining revenue. Findings: Performing stocks have high trading volume and positive earnings reports. Non - 4. Stock Market Analysis : Compare performing and non - p erforming stocks. patients have normal glucose levels and exercise regularly. Findings: Diabetic patients have higher glucose levels and sitting lifestyles. Non - diabetic 3. Medical Diagnosis : Compare patients with and without a disease. Valid transactions match custom er spending history and location.

Findings: Fraud transactions are often international, high - value, and made at unusual times. 2. Fraud Detection in Banking : Compare fraud and valid transactions. customers shop occasionally and buy discount products. Findings: High - value customers shop frequently and prefer premium brands. Low - value 1. Customer Segmentation in Marketing : Compare high - value and low - value customers.

**Applications of Comparison :**

5. Present Findings : Use tables, charts, graphs, or rules to explain the differences. 4. Analyse Differences : Identify key distinguishing features and trends. visualization techniques. 3. Choose Comparison Method : Statistical, machine learning, rule - based, or 2. Pre - process Data : Clean, filter, and transform data to ensure accuracy. 1. Select Data Groups : Define the two or more datasets to compare.

**Steps in the Comparison Process :**

spending customers based on their purchasing behaviour .  For example : I n marketing, comparison can help differentiate high - spending and low - and improve healthcare outcomes. organizations can make data - driven decisions, optimize marketing strategies, detect fraud,  Usin g statistical techniques, rule - based methods, and machine learning models, data segments. product performance, or market trends. It support s decision - making by analysing different  It helps businesses, researchers, and analysts to find key differences in customer behaviour,

two or more groups of data to identify differences, distinguishing features, and patterns .  Comparison is also called Discrimination . It is the process of analysing and contrasting

**COMPARISON / DISCRIMINATION:**

**=== Page 15 ===**

**15**

"FREE," "WINNER," and "CASH PRIZE." greetings and business language Spam emails often contain words like Non - spam emails have more personalized Group - 1: Spam Emails Group - 2: Non - Spam Emails Forest, and Neural Networks to find key distinguishing factors. Example: 5. Machine Learning - Based Comparison : Uses models like Decision Trees, Random to prefer digital banking. active social media presence, they are likely likely to visit physical bank branches. If the customer is under 30 and has an If the customer is over 50, they are more

**Rule 1: Rule 2:**

**differentiating datasets. Example:**

4. Rule - Based Comparison : Uses association rules and decision trees to extract rules unstable employment stable jobs. Defaulters have low credit scores, high and Non - defaulters have high credit scores and Group - 1: Loan Defaulters Group - 2: Non - Defaulters

**differentiate between two groups. Example:**

3. Discriminant Analysis : A classification technique that finds attributes that best Finding : Online shoppers spend more per transaction than in - store shoppers . purchase = $150 purchase = $100 Group - 1 (Online Shoppers): Average Group - 2 (In - store Shoppers): Average median, mode to f ind the central tendency of different groups. Example: 2. Statistical Comparison : Uses statistic al measures to compare datasets. Such as, mean, customers prefer budget - friendly products.

Finding: High - income customers frequently buy luxury items, whereas low - income G roup - 1: High - income customers G roup - 2: Low - income customers

**two datasets. Example:**

of summarizing, it finds key differences. Attributes are generalized and compared between 1. Att ribute - Oriented Induction (AOI): Similar to AOI used in characterization, but instead

**Techniques for Comparison :**

**=== Page 16 ===**

**16**

diabetic vs. non - di abetic patients)  Disease classification (e.g., dataset valid transactions)  Identifying key attributes in a  Fraud detection (e.g., fraud vs.  Summarizing medical records high vs. low spenders) Use Cases  Understanding customer behaviour  Customer segmentation ( e.g., Type trends. and contrasting patterns. Output Generalized patterns, summaries, and Differences, distinguishing features, 40." budget - friendly products." purchases come from people aged 25 - those with lower income prefer

this dataset is $70,000, and most $70,000 buy luxury items, while Example "The average income of customers in "Customers with an income above  Visualization  Rule - Based Comparison Used  Statistical Summarization  Discriminant Analysis Techniques  Attrib ute - Oriente d Induction (AOI)  Statistical Analysis features of a dataset. groups or datasets. Purpose Provides an overview of the key Highlights differences between characteristics of a dataset. identify differences.

Definition Summarizes and describes the general Compares two or more datasets to Feature Data Characterization Data Comparison Difference between Data Characterization and Data Comparison : 3. It finds anomalies that indicate fraud activity. 2. It d istinguishes different consumer gro ups for personalized marketing. 1. It h elps in busines s strategy and decision - making.

**Advantages of Comparison :**

**=== Page 17 ===**

**17**

and many other appl ications.  These materialized views can then be used for decision support, knowledge discovery, average(), and max().  These operations typically involve aggregate functions, such as count(), sum(),  It uses Roll - up and Drill - down operations on a data cube.  In this approach, computation and results are stored in the Data cube.  It is an efficient approach as it is helpful to make the past selling graph. 1. Data cube approach (OLAP - Online Analytical Processing approach ) :

There are two basic approaches of data generalization: OLAP approach & AOI approach Salary: $55,000 Salary: $50K - $60K City: New York City: USA Age: 27 Age: 20 - 30 Age: 22 Age: 20 - 30 Original Data Generalized Data For example: Instead of listing every single value, data is grouped into ranges or categories . privacy by masking specific details in sensitive data. in raw data. It r educes data storage needs by eliminating unnecessary details. It e nhances  It s implifies lar ge datasets by summarizing data and i dentifies trends that may not be visible

values with higher level concepts. It is a form of descriptive data mining.  Data Generalization is the process of summarizing data by replacing relatively low level

**DATA GENERALIZATION:**

**=== Page 18 ===**

**18**

35 Chicago $65K 28 New York $50K Sarah 35 Chicago $65K 987 - 654 - 3210 Age City Salary John 28 New York $50K 123 - 456 - 7890

**Number) :**

Name Age City Salary Phone Number (Removing Name & Phone Before Attribute removal: After Attribute Removal  Efficiency → Fewer columns mean faster processing.  Relevance → Name doesn’t affect salary analysis.  Privacy → Phone numbers & names are sensitive. Example: In Customer Data , removing name and Phone number. We have removed because: performance . similar information). computations and im proves removing one column if another provides  Less data to process means faster  When data redundancy is high ( e.g.,

user identities . personal information like A/c no. ).  Sensitive data is removed, protecting  When privacy concerns exist ( e.g., removing analyze. customer names in sales trend analysis).  Few attribut es make data easier to  When certain attributes are irrelevant ( e.g., Why Use Attribute Removal? When to Use Attribute Removal? analysis more efficient.  It involves removing less impor tant or sensitive attributes from a dataset to make improving privacy, and highlighting only the most relevant features.

in a dataset are eliminated to simplify the data. This helps in reducing complexity,  Attribute removal is a technique used in data generalization where specific attributes

**a) Attribute Removal :**

b) Attribute generalization. a) Attribute removal.

** Attribute oriented induction approach uses two method :**

respective counts are accumulated in order to perform aggregation. attributes within the relevant data set. Af ter that same tuple are merged and their  In this approach, we perform generalization on basis of different values of each  It is an online data analysis, query oriented and generalization based approach.

**2. Attribute O riented I nduction (AOI) :**

**=== Page 19 ===**

**19**

**$98,000 $90K - $100K**

**$52,000 $50K - $60K**

Exact Salary Generalized Salary Example: Income Generalization focusing on individual salaries. individual values. This is useful when analysing salary distributions rather than 2. Range - Based Generalization: Numerical data is grouped into ranges instead of Paris Île - de - France France Los Angeles California USA New York New York (State) USA City Generalized Level 1 Generalized Level 2 Example: Location Generalization for abstraction. The h igher levels provide more generalization.

1. Concept Hierarchy - Based Generalization : Data is grouped into hierarchical levels

**Techniques of Attribute Generalization :**

65 60 - 70 Senior 37 30 - 40 Middle - aged 23 20 - 30 Young Original Age Generalized Age (Level 1) Generalized Age (Level 2) Example: Instead of dealing with exact ages, we group them into broader categories . summarizing attributes. hierarchies. It simplifies large datasets by reducing complexity. It u ses less space by  It involves replacing detailed va lues with generalized categories based on pre - defined This helps in summarizing data and identifying patterns efficiently.

values are replaced with higher - level, more abstract concepts using concept hierarchies.  Attribute Generalization is a technique in data generalization where specific attribute

**b) Attribute Generalization :**

**=== Page 20 ===**

**20**

d) Social Media: Aggregating trends from large - scale text data. c) Finance: Summarizing stock market data for investment decisions. b) Healthcare: Summarizing patient records for better diagnosis. a) Business: Understanding customer behaviour , sales trends, and market performance.

**Applications of Data Summarization :**

 Used for multidimensional data analysis (e.g., sales by region, time, and product).  Online Analytical Processing (OLAP) cubes store pre - aggregated data for fast retri eval.

**5. Data Cube (OLAP Summarization) :**

 Heat - maps: Display correlations between variables.  Box Plots: Indicate data spread and outliers.  Pie Charts: Represent proportions of categories.  Histograms & Bar Charts: Show frequency distributions.

**4. Visualization Techniques :**

 Segmentation: Dividing data into meaningful groups based on predefined attributes.  Clustering: Grouping similar data points together to form patterns.

**3. Clustering and Segmentation :**

 Averaging customer ratings for a product.  Summing up monthly sales data to get yearly totals.

**summary. For example:**

2. Aggregation : Combining multiple data points into a single value to provide a higher - level  Variance & Standard Deviation: Me asures of data dispersion.  Mode: The most frequently occurring value.  Median: The middle value when data is sorted.  Mean (Average): The central value of a dataset. 1. Statistical Measures : Summarization begins with computing basic statistical measures:

**Techniques of Data Summarization :**

visual or textual summaries that highlight key patterns, trends, and distributions. decision - making. It involves computing statistical measures, aggregating data, and creating  It enables to understand large datasets quickly , reduces storage requirements, and helps in information from large datasets by providing a compact representation of the data.  Data summarization is a key technique in data mining that helps in extracting useful

**DATA SUMMARIZATION:**

**=== Page 21 ===**

**21**

5. Pivot (Rotate) : Re - orients the data view for better analysis. 4. Dice : Filters data based on multiple dimensions. 3. Slice : Filters data along a single dimension. 2. Drill - down : Breaks down data to a more detailed level. 1. Roll - up : Aggregates data to a higher level of hierarchy.

**The OLAP operations are:**

the target and contrasting classes. includes a "contrasting" measure (such as count %) that reflects the comparison between can be visualized in the form of tables, charts, and rules. This presentation usually 5. Presentation of the derived comparison: The resulting class comparison descriptio n pivoting can be performed on target an d contrasting classes based on user instructions. asynchronous drill - down, roll - up & other OLAP operations such as, slicing, dicing and 4. Drilling Down, Rolling Up and other OLAP adjustments: Synchronous and

in the prime target class relation, forming the prime contrasting class relation or cuboid. relation. The concepts in the contrasting class are generalized to the same level as those target class to the level controlled by the user, which results in a prime target cl ass 3. Synchronous Generalization: The process of generalization is performed upon the these classes and o nly highly relevant dimensions are included in the further analysis. comparisons are desired, then dimension relevance analysis should be performed on

2. Dimension relevance analysis: If there are many dimensions and analytical contrasting classes. collected by query Processing and partitioned into a target class and one or a set of 1. Data Collection: The set of relevant data in the database and data warehouse is

**The general procedure for class comparison is as follows:**

science candidates versus physics candidates. sales in the last thr ee years are comparable classes. Also, we can compare computer  For example, t he three classes - person, address, and item are not comparable. But the because they share similar dimensions and attributes. from its contrasting classes. The target and contrasting classes must be comparable  Class discrimination or class comparison mines descriptions that distinguish a target cl ass that distinguish one class from another.

or groups of data based on their attributes. The goal is to identify discriminative features  Mining class comparisons is a technique used to anal yse and compare different classes

**MINING CLASS COMPARISIONS:**

**=== Page 22 ===**

**22**

**…… …… …… …… …… …… ……**

Priya F Monitor Jaipur 22/06/2002 Jain colony 9 460768333 Harish M Mouse Jodhpur 14/07/1993 11/B - road 99 41852791 N ame Gender Product Birthplace Birthdate Residence Phone Contrasting class: new customer

**…… …… …… …… …… …… ……**

Minal F Keyboard Jodhpur 11/08/1992 Bank colony 9928509928 Sumit M Scanner Jaipur 22/06/2002 A.G. colony 9875894102 Rakesh M Printer Jodhpur 14/07/1993 3/A C.H.B. 9925852890 N ame Gender Product Birthplace Birthdate Residence Phone Target class: old customer attributes name, gender, product, b irthplace, birthdate, residence and phone . customer of Royal Electronics which deals in computer’s electronic products, given the Suppose we would like to compare the general properties between the old customer and new

**Example of Mining Class Comparison:**

**=== Page 23 ===**

**23**

presented in form of tables, graphs or rules. 5. Presentation of the derived comparison: Finally, the resulting class comparison is the level of abstraction. performed on the target and contrasting class, based on the user’s instruction to adjust 4. Drilling Down, Rolling Up and other OLAP adjustments: The OLAP operations are Others Over 30 Others Over 30 Jaipur 20 - 25 J odhpur 18 - 25 Jodhpur 25 - 30 Jodhpur 18 - 25 Birthplace Age - Range Birthplace Age - Range

t arget class: old customer contrasting class: new customer Prime generalized relation for the Prime generalized relation for the the prime contrasting class relation. class is also generalized to the same level s as those in the prime target class, forming t o the levels controlled by user, forming the prime target cla ss relation. The contrasting 3. Synchronous Generalization: Here, the generalization is performed on the target class attributes are included in the subsequent analysis.

product, residence and phone are removed from the resulting class. Only highly relevant data. After this analysis, irrelevant or weakly relevant dimensions such as name, gender, 2. Dimension relevance analysis: Now, this analysis is performed on the two classes of class working relation and the other for the initial contrasting class working relation. 1. Data Collection: In this, we select two set of task relevant data. One for the initial target Initial working relations: the target class vs. the contrasting class.