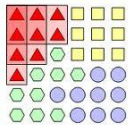
Data Warehousing and Data Mining

Unit 5-6

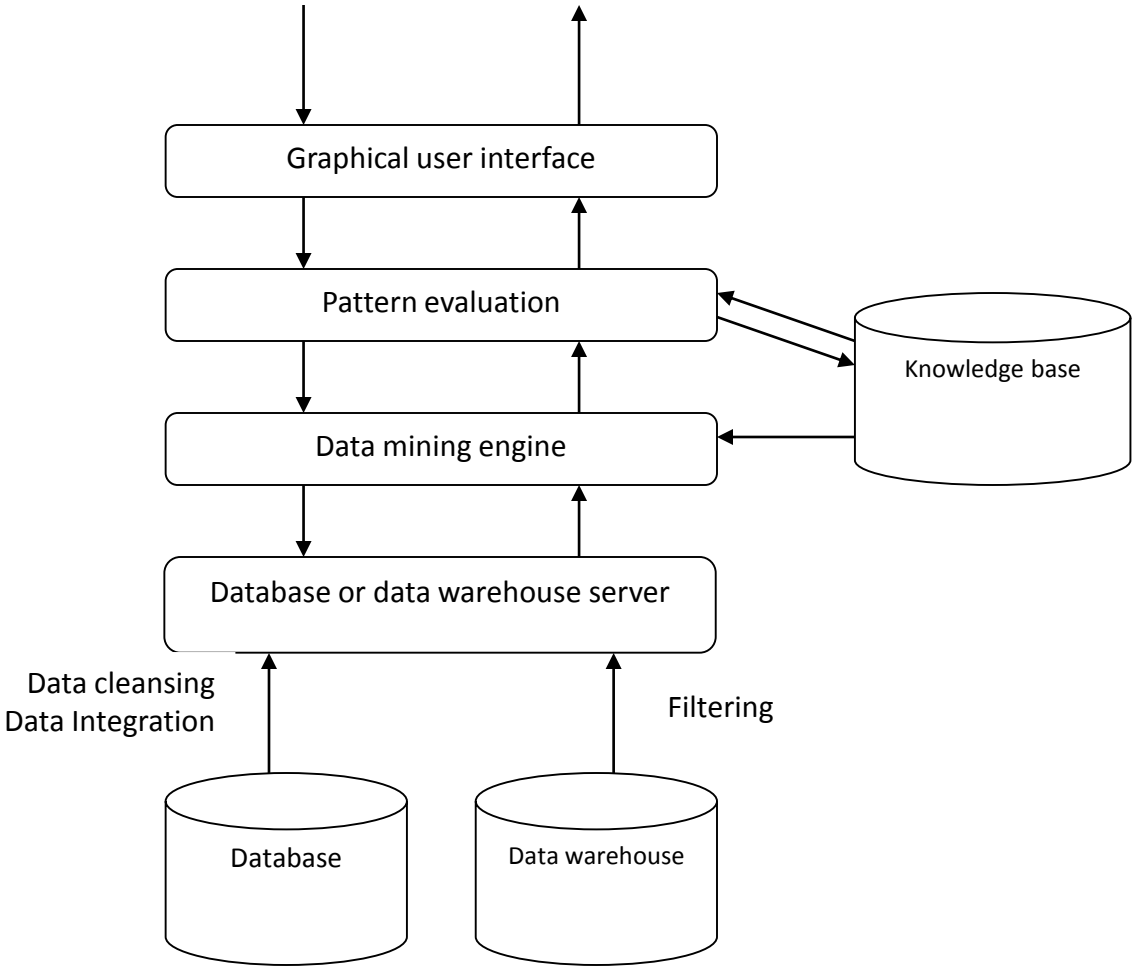
Suresh Pokharel

Content

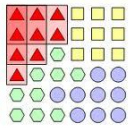


- Data Mining Definition and Task,
- KDD versus Data Mining Techniques,
- Tools and application
- Data mining query language
- Data specification, specifying knowledge
- Hierarchy specification,
- Pattern presentation & visualization specification,
- Data mining language and
- Standardization of data mining

Architecture of a typical data mining system

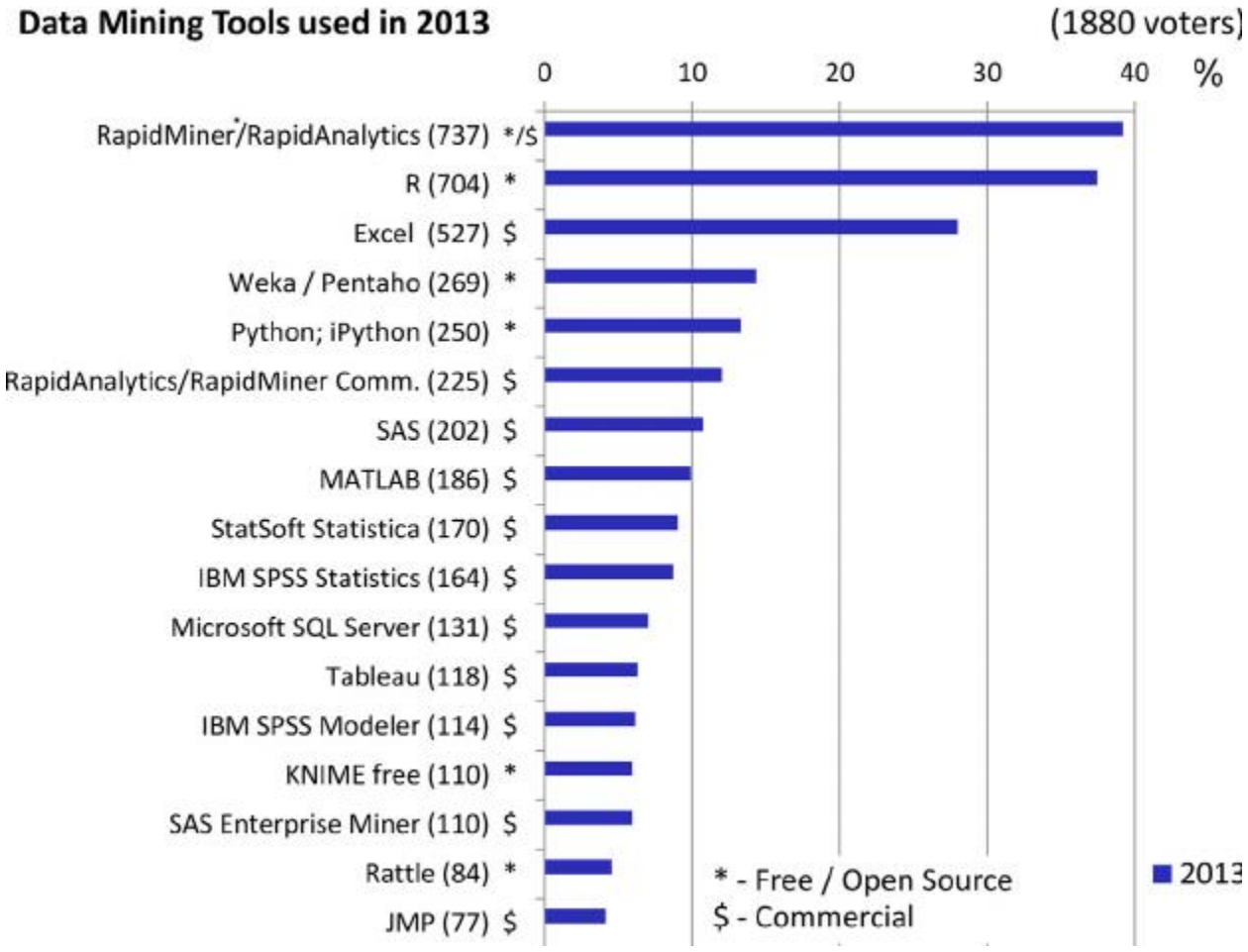


Data Mining Tasks

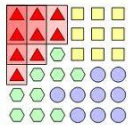


1. **Classification:** learning a function that maps an item into one of a set of predefined classes
2. **Regression:** learning a function that maps an item to a real value
3. **Clustering:** identify a set of groups of similar items
4. **Dependencies and associations:** identify significant dependencies between data attributes
5. **Summarization:** find a compact description of the dataset or a subset of the dataset

Data Mining Tools

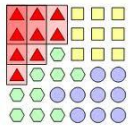


Data Mining Tools



Characteristic	RapidMiner	R	Weka	Orange	KNIME	scikit-learn
Developer:	RapidMiner, Germany	worldwide development	Univ. of Waikato, New Zealand	Univ. of Ljubljana, Slovenia	KNIME.com AG,Switzerland	multiple; support: INRIA, Google
Programming language:	Java	C, Fortran, R	Java	C++, Python, Qt framew.	Java	Python+NumPy+ SciPy+matplotlib
License:	open s. (v.5 or lower); closed s., free Starter ed. (v.6)	free software, GNU GPL 2+	open source, GNU GPL 3	open source, GNU GPL 3	open source, GNU GPL 3	FreeBSD
Current version:	6	3.02	3.6.10	2.7	2.9.1	0.14.1
GUI / command line:	GUI	both; (GUI for DM = Rattle)	both	both	GUI	command line
Main purpose:	general data mining	sci. computation and statistics	general data mining	general data mining	general data mining	machine learning package add-on
Community support (est.):	large (~200 000 users)	very large (~ 2 M users)	large	moderate	moderate (~ 15 000 users)	moderate

Data Mining Application



Telecom Industry

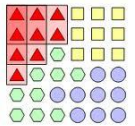
- Fraudulent pattern analysis and the identification of unusual patterns
- Multidimensional association and sequential pattern analysis
- Mobile telecommunication services
- Use of visualization tools in telecommunication data analysis

Biomedical Data Analysis

Social Network

- And many more…..

Data Mining Application



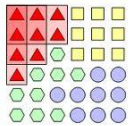
Telecom Industry

- Fraudulent pattern analysis and the identification of unusual patterns
- Multidimensional association and sequential pattern analysis
- Mobile telecommunication services
- Use of visualization tools in telecommunication data analysis

Biomedical Data Analysis

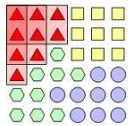
Social Network

- And many more…..



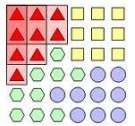
Unit 6

Why Data Mining Primitives and Languages?



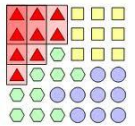
- Finding all the patterns autonomously in a database?
  - unrealistic because the patterns could be too many but uninteresting
- Data mining should be an interactive process
  - User directs what to be mined
- Users must be provided with a set of primitives to be used to communicate with the data mining system
- Incorporating these primitives in a data mining query language
  - More flexible user interaction
  - Foundation for design of graphical user interface
  - Standardization of data mining industry and practice

What Defines a Data Mining Task ?



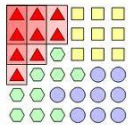
- Task-relevant data
  - Typically interested in only a subset of the entire database
  - Specify
    - the name of database/data warehouse (AllElectronics_db)
    - names of tables/data cubes containing relevant data (item, customer, purchases, items_sold)
    - conditions for selecting the relevant data (purchases made in Nepal for relevant year)
    - relevant attributes or dimensions (name and price from item, income and age from customer)

What Defines a Data Mining Task ? (continued)



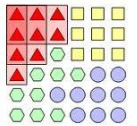
- Type of knowledge to be mined
  - Concept description, association, classification, prediction, clustering
    - Studying buying habits of customers, mine associations between customer profile and the items they like to buy
      - Use this info to recommend items to put on sale to increase revenue
    - Studying real estate transactions, mine clusters to determine house characteristics that make for fast sales
      - Use this info to make recommendations to house sellers who want/need to sell their house quickly
    - Study relationship between individual's sport statistics and salary
      - Use this info to help sports agents and sports team owners negotiate an individual's salary

What Defines a Data Mining Task ? (continued)



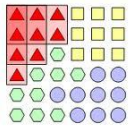
- Type of knowledge to be mined
  - Search for association rules is confined to those matching some set of rules, such as:
    - Age(X, “30..39”) & income (X, “40K..49K”) => buys (X, “VCR”) [2.2%, 60%]
    - Customers in their thirties, with an annual income of 40-49K, are likely (with 60% confidence) to purchase a VCR, and such cases represent about 2.2% of the total number of transactions

What Defines a Data Mining Task ?



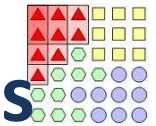
- Task-relevant data
- Type of knowledge to be mined
- Background knowledge
- Pattern interestingness measurements
- Visualization of discovered patterns

Task-Relevant Data (Minable View)



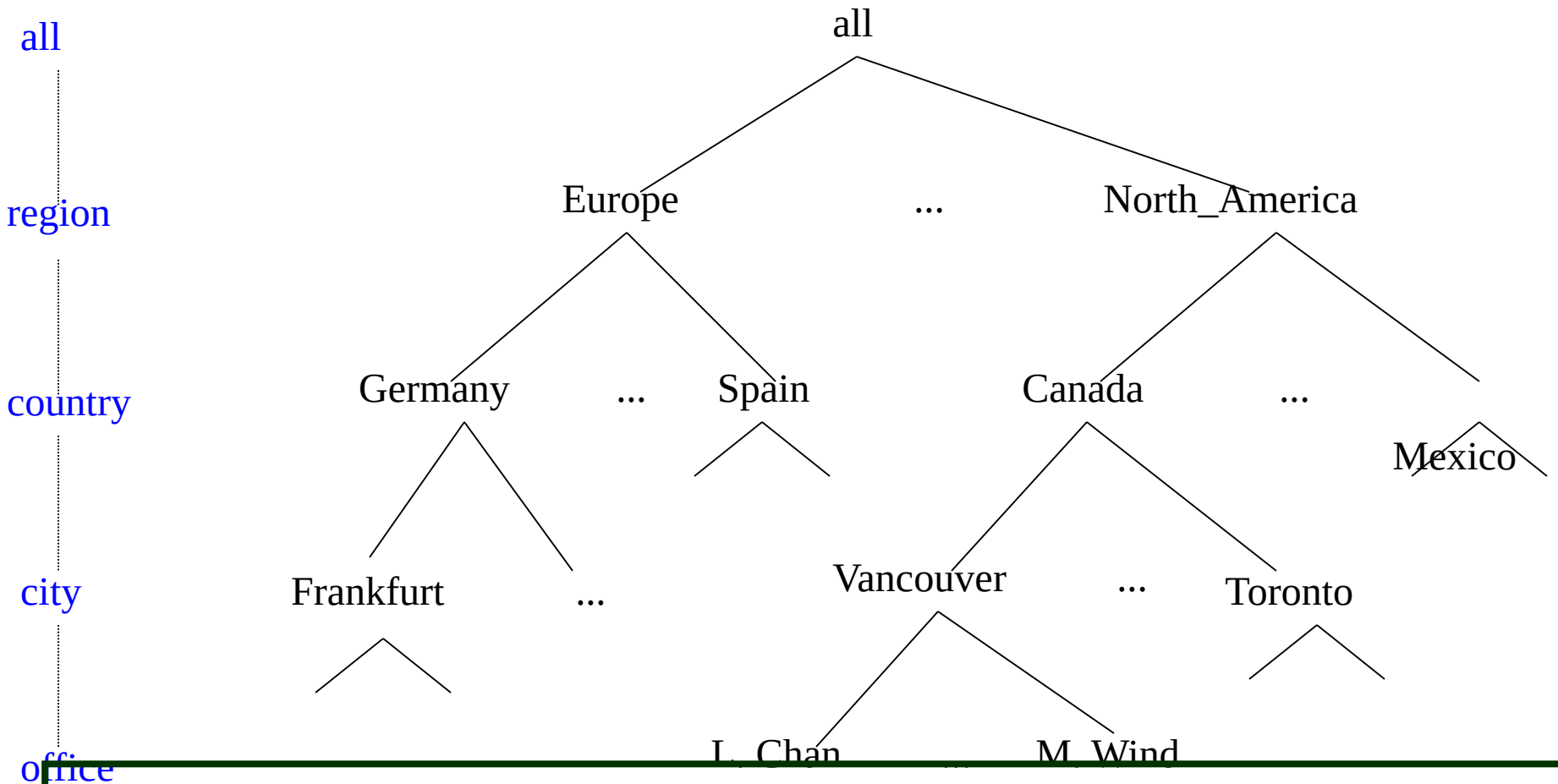
- Database or data warehouse name
- Database tables or data warehouse cubes
- Condition for data selection
- Relevant attributes or dimensions
- Data grouping criteria

Background Knowledge: Concept Hierarchies



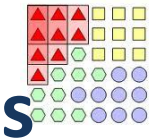
- Allow discovery of knowledge at multiple levels of abstraction
- Represented as a set of nodes organized in a tree
  - Each node represents a concept
- Concept hierarchies allow raw data to be handled at a higher, more generalized level of abstraction
- Four major types of concept hierarchies, schema, set-grouping, operation derived, rule based

A Concept Hierarchy: Dimension (location)



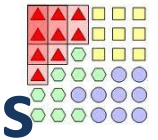
Define a sequence of mappings from a set of low level concepts to higher-level, more general concepts

Background Knowledge: Concept Hierarchies



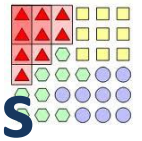
- Schema hierarchy – total or partial order among attributes in the database schema, formally expresses existing semantic relationships between attributes
  - Table address
    - create table address (street char (50), city char (30), province_or_state char (30), country char (40));
  - Concept hierarchy location
    - street < city < province_or_state < country
- Set-grouping hierarchy – organizes values for a given attribute or dimension into groups or constant range values
  - {young, middle_aged, senior} subset of all(age)
    - {20-39} = young
    - {40-59} = middle_aged
    - {60-89} = senior

Background Knowledge: Concept Hierarchies



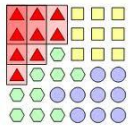
- Operation-derived hierarchy – based on operations specified by users, experts, or the data mining system
  - email address or a URL contains hierarchy info relating departments, universities (or companies) and countries
  - E-mail address
    - dmbook@cs.sfu.ca
  - Partial concept hierarchy
    - login-name < department < university < country

Background Knowledge: Concept Hierarchies



- Rule-based hierarchy – either a whole concept hierarchy or a portion of it is defined by a set of rules and is evaluated dynamically based on the current data and rule definition
  - Following rules used to categorize items as low profit margin, medium profit margin and high profit margin
    - Low profit margin - < $50
    - Medium profit margin – between $50 & $250
    - High profit margin - > $250
  - Rule based concept hierarchy
    - low_profit_margin (X) <= price(X, P1) and cost (X, P2) and (P1 - P2) < $50
    - medium_profit_margin (X) <= price(X, P1) and cost (X, P2) and (P1 - P2) >= $50 and (P1 – P2) <= $250
    - high_profit_margin (X) <= price(X, P1) and cost (X, P2) and (P1 - P2) > $250

Measurements of Pattern Interestingness



- After specification of task relevant data and kind of knowledge to be mined, data mining process may still generate a large number of patterns
- Typically, only a small portion of these patterns will actually be of interest to a user
- The user needs to further confine the number of uninteresting patterns returned by the data mining process
  - Utilize interesting measures
- Four types: simplicity, certainty, utility, novelty

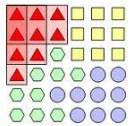
Measurements of Pattern Interestingness (continued)

- Simplicity – A factor contributing to interestingness of pattern is overall simplicity for comprehension
  - Objective measures viewed as functions of the pattern structure or number of attributes or operators
  - More complex a rule, more difficult it is to interpret, thus less interesting
  - Example measures: rule length or number of leaves in a decision tree
- Certainty – Measure of certainty associated with pattern that assesses validity or trustworthiness
  - Confidence (A=>B) = # tuples containing both A & B/ #tuples containing A
  - Confidence of 85% for association rule buys (X, computer) => buys (X, software) means 85% of all customers who bought a computer bought software also

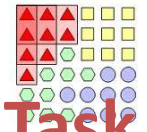
Measurements of Pattern Interestingness (continued)

- Utility – potential usefulness of a pattern is a factor determining its interestingness
  - Estimated by a utility function such as support – percentage of task relevant data tuples for which pattern is true
    - Support (A=>B) = # tuples containing both A & B/ total # of tuples
- Novelty – those patterns that contribute new information or increased performance to the pattern set
  - not previously known, surprising

Visualization of Discovered Patterns



- Different backgrounds/usages may require different forms of representation
  - E.g., rules, tables, crosstabs, pie/bar chart etc.
- Concept hierarchy is also important
  - Discovered knowledge might be more understandable when represented at high level of abstraction
  - Interactive drill up/down, pivoting, slicing and dicing provide different perspective to data
- Different kinds of knowledge require different representation: association, classification, clustering, etc.

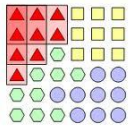


Summary: Five Primitives for Specifying a Data Mining Task

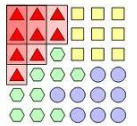
- task-relevant data
  - database/date warehouse, relation/cube, selection criteria, relevant dimension, data grouping
- kind of knowledge to be mined
  - characterization, discrimination, association...
- background knowledge
  - concept hierarchies,..
- interestingness measures
  - simplicity, certainty, utility, novelty
- knowledge presentation and visualization techniques to be used for displaying the discovered patterns
  - rules, table, reports, chart, graph, decision trees, cubes ...
  - drill-down, roll-up,....

Source:JH

A Data Mining Query Language (DMQL)



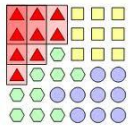
- Motivation
  - A DMQL can provide the ability to support ad-hoc and interactive data mining
  - By providing a standardized language like SQL
    - Hope to achieve a similar effect like that SQL has on relational database
    - Foundation for system development and evolution
    - Facilitate information exchange, technology transfer, commercialization and wide acceptance
- Design
  - DMQL is designed with the primitives described earlier



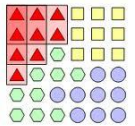
Syntax for DMQL

- Syntax for specification of
  - task-relevant data
  - the kind of knowledge to be mined
  - concept hierarchy specification
  - interestingness measure
  - pattern presentation and visualization
- Putting it all together — a DMQL query

Syntax for task-relevant data specification



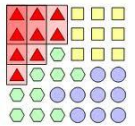
- *use database* database_name, or *use data warehouse* data_warehouse_name
  - directs the data mining task to the database or data warehouse specified
- *from relation*(s)/cube(s) [*where* condition]
  - specify the database tables or data cubes involved and the conditions defining the data to be retrieved
- *in relevance to* att_or_dim_list
  - Lists attributes or dimensions for exploration



Syntax for task-relevant data specification

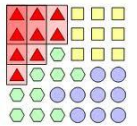
- *order by* order_list
  - Specifies the sorting order of the task relevant data
- *group by* grouping_list
  - Specifies criteria for grouping the data
- *having* condition
  - Specifies the condition by which groups of data are considered relevant

Top Level Syntax of DMQL



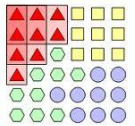
- *(DMQL) ::= (DMQL_Statement);{(DMQL_Statement)*
- *(DMQL_Statement) ::= (Data_Mining_Statement) | (Concept_Hierarchy_Definition_Statement) | (Visualization_and_Presentation)*

Top Level Syntax of DMQL (continued)



- *(Data_Mining_Statement) ::=* ***use database*** *(database_name)* | ***use data warehouse*** *(data_warehouse_name)* {***use hierarchy*** *(hierarchy_name)* ***for*** *(attribute_or_dimension)}* *(Mine_Knowledge_Specification)* ***in relevance to*** *(attribute_or_dimension_list)* ***from*** *(relation(s)/cube(s))* [***where*** *(condition)]* [***order by*** *(order_list)]* [***group by*** *(grouping_list)]* [***having*** *(condition)]* {***with*** *[(interest_measure_name)]* ***threshold*** *= (threshold_value)* [***for*** *(attribute(s))]}

Specification of task-relevant data



Example 4.11 This example shows how to use DMQL to specify the task-relevant data described in Example 4.1 for the mining of associations between items frequently purchased at *AllElectronics* by Canadian customers, with respect to customer *income* and *age*. In addition, the user specifies that she would like the data to be grouped by date. The data are retrieved from a relational database.

```
use database AllElectronics_db
in relevance to I.name, I.price, C.income, C.age
from customer C, item I, purchases P, items_sold S
where I.item_ID = S.item_ID and S.trans_ID = P.trans_ID and P.cust_ID = C.cust_ID
        and C.address = "Canada"
group by P.date
```

Putting It All Together: the Full Specification of a DMQL Query

use database **OurVideoStore_db**

use hierarchy **location_hierarchy for B.address**

mine characteristics as **customerRenting**

analyze **count%**

in relevance to **C.age, I.type, I.place_made**

from **customer C, item I, rentals R, items_rent S, works_at W, branch**

where **I.item_ID = S.item_ID and S.trans_ID = R.trans_ID**

and R.cust_ID = C.cust_ID and R.method_paid = ``Visa''

and R.empl_ID = W.empl_ID and W.branch_ID = B.branch_ID and B.address = ``Alberta'' and I.price >= 100

with **noise** threshold **= 0.05**

display as **table**