Principles of Knowledge Discovery in Data

Fall 2004

<u>Chapter 5</u>: Data Summarization

Dr. Osmar R. Zaïane

University of Alberta

Source:
Dr. Jiawei Han

Summary of Last Chapter

- What is the motivation for ad-hoc mining process?
- What defines a data mining task?
- Can we define an ad-hoc mining language?

Course Content

- Introduction to Data Mining
- Data warehousing and OLAP
- Data cleaning
- Data mining operations
- **Data summarization**
- Association analysis
- Classification and prediction
- Clustering
- Web Mining
- Spatial and Multimedia Data Mining
- *Other topics if time permits*

Chapter 4 Objectives

Understand Characterization and Discrimination of data.

See some examples of data summarization.

Data Summarization Outline

- What are summarization and generalization?
- What are the methods for descriptive data mining?
- What is the difference with OLAP?
- Can we discriminate between data classes?

Descriptive vs. Predictive Data Mining

- Descriptive mining: describe concepts or task-relevant data sets in concise, informative, discriminative forms.
- Predictive mining: Based on data and analysis, construct models for the database, and predict the trend and properties of unknown data.

Concept description:

- Characterization: provides a concise and succinct summarization of the given collection of data.
- Comparison: provides descriptions comparing two or more collections of data.

Need for Hierarchies in Descriptive Mining

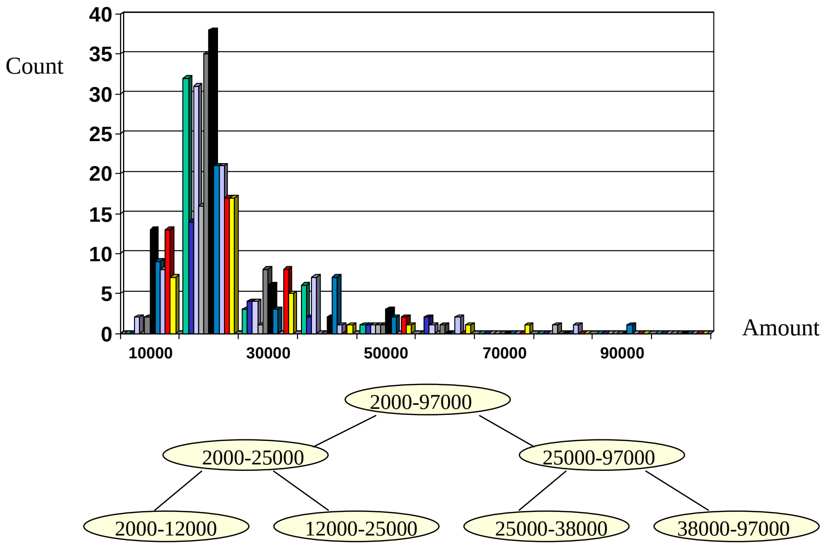
- Schema hierarchy
  - Ex: *house_number* < *street* < *city* < *province* < *country*
    - define hierarchy as [*house_number, street, city, province, country*]
- Instance-based (Set-Grouping Hierarchy):
  - Ex: {*freshman, ..., senior*} ⊂ *undergraduate.*
    - define hierarchy statusHier as

      level2: {freshman, sophomore, junior, senior} < level1:undergraduate;
      level2: {M.Sc, Ph.D} < level1:graduate;
      level1: {undergraduate, graduate} < level0: allStatus
- Rule-based:
  - *undergraduate(x)* $\wedge$ *gpa(x)* > *3.5* $\Rightarrow$ *good(x).*
- Operation-based:
  - aggregation, approximation, clustering, etc.

Creating Hierarchies

- Defined by database schema:
  - Some attributes naturally form a hierarchy:
    - *Address (street, city, province, country, continent)*
  - Some hierarchies are formed with different attribute combinations:
    - *food*(*category, brand, content _spec, package _size, price*).
- Defined by set-grouping operations (by users/experts).
    - {*chemistry, math, physics*} ⊂ *science.*
- Generated automatically by data distribution analysis.
- Adjusted automatically based on the existing hierarchy**.**

Automatic Generation of Numeric Hierarchies



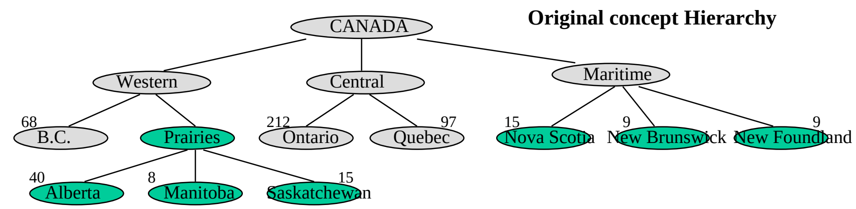
Methods for Automatic Generation of Hierarchies

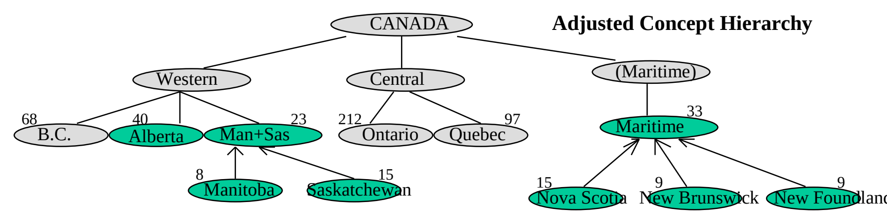
- Categorical hierarchies: (Cardinality heuristics)
  - Observation: the higher hierarchy, the smaller cardinality.
    - card(city) < card(state) < card (country).
  - There are exceptions, e.g., {day, month, quarter, year}.
  - Automatic generation of categorical hierarchies based on cardinality heuristic:
    - location: {country, street, city, region, big-region, province}.
- Numerical hierarchies:
  - Many algorithms are applicable for generation of hierarchies based on data distribution.
  - Range-based vs. distribution-based (different binning methods)

Automatic Hierarchy Adjustment

- Why adjusting hierarchies dynamically?
  - Different applications may view data differently.
  - Example: Geography in the eyes of politicians, researchers, and merchants.
- How to adjust the hierarchy?
  - Maximally preserve the given hierarchy shape.
  - Node merge and split based on certain weighted measure (such as count, sum, etc.)
    - E.g., small nodes (such as small provinces) should be merged and big nodes should be split.

Dynamic Adjustment of Concept Hierarchies



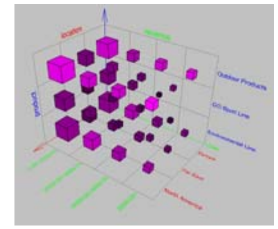


Data Summarization Outline

- What are summarization and generalization?
- What are the methods for descriptive data mining?
- What is the difference with OLAP?
- Can we discriminate between data classes?

Methods of Descriptive Data Mining

- Data cube-based approach:
  - Dimensions: Attributes form concept hierarchies
  - Measures: sum, count, avg, max, standard-deviation, etc.
  - Drilling: generalization and specialization.
  - Limitations: dimension/measure types, intelligent analysis.



- Attribute-oriented induction:
  - Proposed in 1989 (KDD'89 workshop).
  - Not confined to categorical data nor particular measures.
  - Can be presented in both table and rule forms.

Basic Principles of Attribute-Oriented Induction

- <u>Data focusing</u>: task-relevant data, including dimensions, and the result is the *initial relation*.
- <u>Attribute-removal</u>: remove attribute *A* if there is a large set of distinct values for *A* but (1) there is no generalization operator on *A*, or (2)*A*'s higher level concepts are expressed in terms of other attributes.
- <u>Attribute-generalization</u>: If there is a large set of distinct values for *A*, and there exists a set of generalization operators on *A*, then select an operator and generalize *A*.
- <u>Attribute-threshold control</u>: typical 2-8, specified/default.
- <u>Generalized relation threshold control</u>: control the final relation/rule size.

Basic Algorithm for Attribute-Oriented Induction

- <u>InitialRel</u>: Query processing of task-relevant data, deriving the *initial relation*.
- <u>PreGen:</u> Based on the analysis of the number of distinct values in each attribute, determine generalization plan for each attribute: removal? or how high to generalize?
- <u>PrimeGen</u>: Based on the PreGen plan, perform generalization to the right level to derive a "prime generalized relation".
- <u>Presentation</u>: User interaction: (1) adjust levels by drilling, (2) pivoting, (3) mapping into rules, cross tabs, visualization presentations.

Class Characterization: An Example

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC,Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 !st Ave., Vancouver	253-9106	3.70
Laura Lee	F	physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
…	..	…	…	…	…	…	…

Gender	Major	Birth_region	Age_range	Residence	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
…	…	…	…	…	…	…

Gender \ Birth_Region	Canada	Foreign	Total
M	16	14	30
F	10	22	32
Total	26	36	62

Presentation of Generalized Results

- Generalized relation:
  - Relations where some or all attributes are generalized, with counts or other aggregation values accumulated.
- Cross tabulation:
  - Mapping results into cross tabulation form (similar to contingency tables).
- Visualization techniques:
  - Pie charts, bar charts, curves, cubes, and other visual forms.
- Quantitative characteristic rules:
  - Mapping generalized result into characteristic rules with quantitative information associated with it, e.g.,

$$grad(x) \wedge male(x) \Rightarrow$$
$$birth\_region(x) = "Canada"[53\%] \vee birth\_region(x) = "foreign"[47\%].$$

Example: Grant Distribution in Canadian CS Departments

org_name	count%	amount%
Toronto	7.92%	12.60%
Waterloo	8.87%	10.45%
British Columbia	5.85%	7.15%
Simon Fraser	4.34%	4.97%
Concordia	4.91%	4.81%
Alberta	4.15%	4.26%
Calgary	3.77%	4.21%
McGill	3.02%	4.12%
Victoria	3.96%	3.91%
Queen's	4.34%	3.90%
Carleton	3.40%	3.54%
Western Ontario	3.77%	3.25%
Ottawa	3.40%	2.87%
York	2.45%	2.41%
Saskatchewan	2.45%	2.36%
McMaster	2.26%	2.18%
Manitoba	2.64%	2.15%
Regina	2.26%	1.76%
New Brunswick	1.89%	1.24%

DBMiner Query:

Find NSERC operating research grant **distributions according to** Canadian universities.

use nserc96
mine characteristic rule
for "CS.Organization_Grants"
from award A, organization O, grant_type G
where A.grant_code = G.grant_code and
O.org_code = A.org_code and
A.disc_code = 'Computer" and
G.grant_order = "Operation Grant"
in relevance to amount, org_name, count(*)%, amount(*)%
set attribute threshold 1 **for** amount
unset attribute threshold for org_name

Data Summarization Outline

- What are summarization and generalization?
- What are the methods for descriptive data mining?
- What is the difference with OLAP?
- Can we discriminate between data classes?

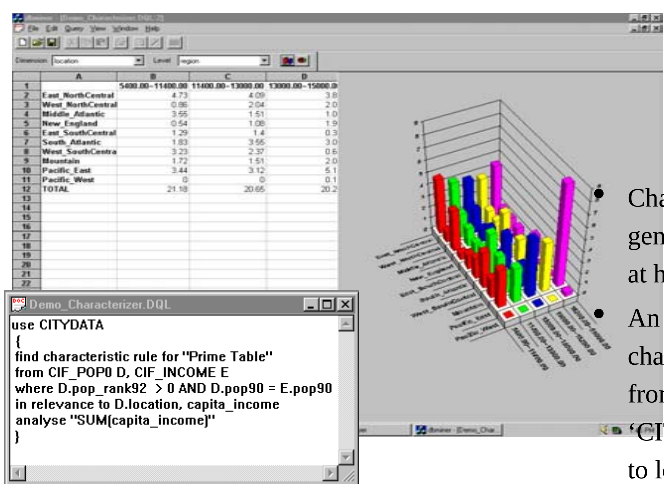
Characterization vs. OLAP

- Similarity:
  - Presentation of data summarization at multiple levels of abstraction.
  - Interactive drilling, pivoting, slicing and dicing.
- Differences:
  - Automated desired level allocation.
  - Dimension relevance analysis and ranking when there are many relevant dimensions.
  - Sophisticated typing on dimensions and measures.
  - Analytical characterization: data dispersion analysis.

Attribute/Dimension Relevance Analysis

- Why attribute-relevance analysis?
  - There are often a large number of dimensions, and only some are closely relevant to a particular analysis task.
  - The relevance is related to both dimensions and levels.
- How to perform relevance analysis?
  - Identify class to be analyzed and its comparative classes.
  - Use information gain analysis (e.g., entropy or other measures) to identify highly relevant dimensions and levels.
  - Sort and select the most relevant dimensions and levels.
  - Use the selected dimension/level for induction.
  - Drilling and slicing follow the relevance rules.

Mining Characteristic Rules



- Characterization: Data generalization/summarization at high abstraction levels.
- An example query: Find a characteristic rule for Cities from the database 'CITYDATA' in relevance to location, capita_income, and the distribution of count% and amount%.

Specification of Characterization by DMQL

- A summarization data mining query:

  MINE Summary

  ANALYZE cost, order_qty, revenue

  WITH RESPECT TO cost, location, order_qty, product, revenue

  FROM CUBE sales_cube
- Analytical characterization.

  If user writes,

  WITH RESPECT TO *

  relevance analysis is often required.

Results of Summarization

Summary #1

Dim: location — Level: country

Meas: revenue — % #

	A	B	C	D	E
1	PRODUCT	LOCATION			
2		Canada	Mexico	United States	North America
3	Back Packs	3294.76	1884	9111.01	14289.77
4	Cooking Equipment	27289.12	2106.9	49630.33	79026.35
5	Sleeping Bags	14820.45	600	20425.8	35846.25
6	Tents	43821.75	21540	224225.3	289587.05
7	Outdoor Products	89226.08	26130.9	303392.44	418749.42
8					
9					
10					

Sheet1

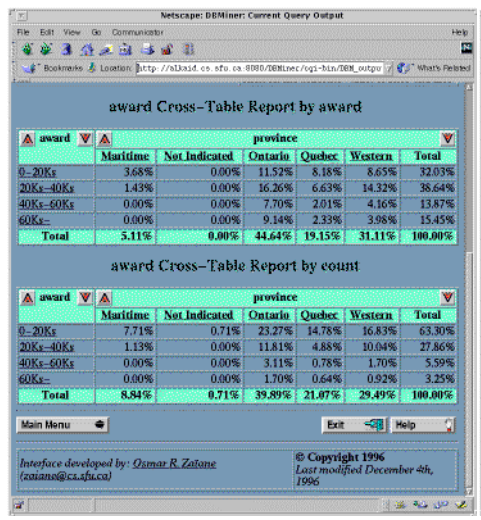
Data Summarization Outline

- What are summarization and generalization?
- What are the methods for descriptive data mining?
- What is the difference with OLAP?
- Can we discriminate between data classes?

Mining Discriminant Rules

- Discrimination: Comparing two or more classes.
- Method:
  - Partition the set of relevant data into the target class and the contrasting class(es)
  - 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 discriminant features
- Relevance Analysis:
  - Find attributes (features) which best distinguish different classes.

Visualization of Characteristic Rules Using Tables and Graphs (DBMiner Web version)



Netscape: DBMiner: Current Query Output

award Cross–Table Report by award

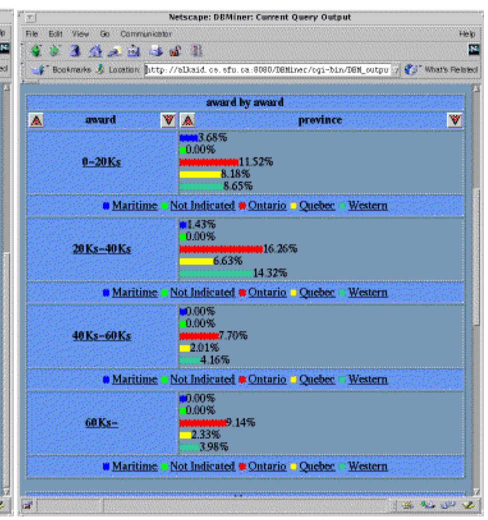
award	Maritime	Not Indicated	Ontario	Quebec	Western	Total
0–20Ks	3.68%	0.00%	11.52%	8.18%	8.65%	32.03%
20Ks–40Ks	1.43%	0.00%	16.26%	6.63%	14.32%	38.64%
40Ks–60Ks	0.00%	0.00%	7.70%	2.01%	4.16%	13.87%
60Ks–	0.00%	0.00%	9.14%	2.33%	3.98%	15.45%
Total	5.11%	0.00%	44.64%	19.15%	31.11%	100.00%

award Cross–Table Report by count

award	Maritime	Not Indicated	Ontario	Quebec	Western	Total
0–20Ks	7.71%	0.71%	23.27%	14.78%	16.83%	63.30%
20Ks–40Ks	1.13%	0.00%	11.81%	4.88%	10.04%	27.86%
40Ks–60Ks	0.00%	0.00%	3.11%	0.78%	1.70%	5.59%
60Ks–	0.00%	0.00%	1.70%	0.64%	0.92%	3.25%
Total	8.84%	0.71%	39.89%	21.07%	29.49%	100.00%

Main Menu Exit Help

Interface developed by: Osmar R. Zaïane (zaiane@cs.sfu.ca) © Copyright 1996 Last modified December 4th, 1996



Visualization of Discriminant Rules Using Graphs (DBMiner Web version)