Chapter 4: Data Generalization

- Attribute-Oriented Induction — An Alternative Data Generalization Method

What is Concept Description?

- Descriptive vs. predictive data mining
  - Descriptive mining: describes concepts or task-relevant data sets in *concise, summarative, informative, discriminative* forms
  - Predictive mining: Based on data analysis, constructs a model for the data, and *predicts the trend and properties* of unknown data based on the model
- Concept description:
  - Characterization: provides a concise and succinct summarization of a *given* collection of data
  - Comparison: provides descriptions *comparing* two or more collections of data

Class Characterization: An Example

Initial Relation

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 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
…	…	…	…	…	…	…	…
Removed	Retained	Sci, Eng, Bus	Country	Age range	City	Removed	Excl, VG,..

Prime Generalized Relation

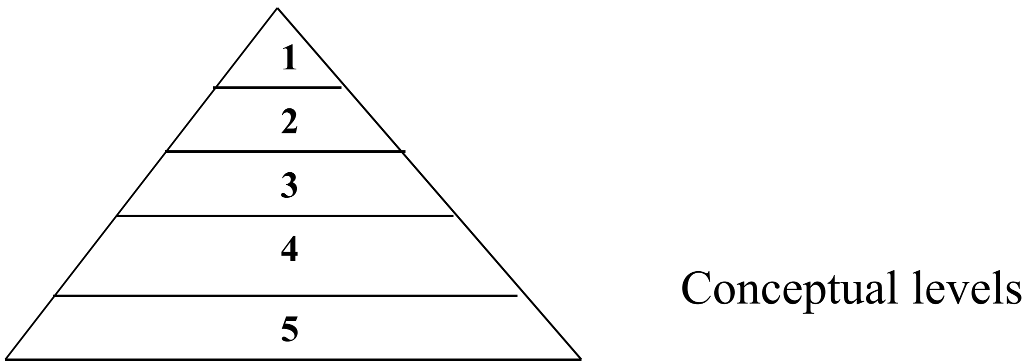
Gender	Major	Birth_region	Age_range	Residence	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
…	…	…	…	…	…	…

Concept Description

- Data generalization
  - Has close ties with concept description
  - Allows data sets to be generalized at multiple levels of abstraction
  - Example
    - nation => province => city => address

Data Generalization and Summarization-based Characterization

- Data generalization
  - A process which abstracts a large set of task-relevant data in a database *from a low conceptual levels to higher ones*



  - Approach:
    - *Attribute-oriented induction approach*

Attribute-Oriented Induction

- Proposed in 1989 (KDD '89 workshop)
- Not confined to categorical data
- How it is done?
  - Collect the task-relevant data (*initial relation*) using a relational database query
  - Perform generalization by *attribute removal* or *attribute generalization*
  - Apply aggregation by merging identical, generalized tuples and accumulating their respective counts
  - Interactive presentation with users

Basic Principles of Attribute-Oriented Induction

- Data focusing: task-relevant data including dimensions, and the result is the *initial relation*
- Attribute-removal: remove attribute *A* if there is a large set of distinct values for *A* but there is no generalization operator on *A*
- Attribute-generalization: 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*
- Attribute-threshold control:
  compare the number of distinct *attribute values* & threshold, typical 2-8
  Increase the threshold -> drilling down
  reduce the threshold -> rolling up
- Generalized relation threshold control:
  compare the number of (distinct) *tuples* & threshold
  Increase the threshold -> drilling down
  reduce the threshold -> rolling up

Attribute-Oriented Induction: Basic Algorithm

- InitialRel: Query processing of task-relevant data, deriving the *initial relation*
- PreGen: 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?*
- PrimeGen: Based on the PreGen plan, perform generalization to the right level to derive a "*prime generalized relation*", accumulating the counts
- Presentation: User interaction: (1) adjust levels by drilling and then (2) mapping into rules or cross tabs for visualization

Example

- DMQL: Describe general characteristics of graduate students in the Big-University database

  use Big_University_DB

  mine characteristics as "Science_Students"

  in relevance to name, gender, major, birth_place, birth_date, residence, phone#, gpa

  from student

  where status in "graduate"

- Corresponding SQL statement:

  Select name, gender, major, birth_place, birth_date, residence, phone#, gpa

  from student

  where status in {"Msc", "MBA", "PhD" }

Class Characterization: An Example

Initial Relation

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 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
…	…	…	…	…	…	…	…
Removed	Retained	Sci,Eng, Bus	Country	Age range	City	Removed	Excl, VG,..

Prime Generalized Relation

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
…	…	…	…	…	…	…

Presentation of Generalized Results

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

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

Presentation—Generalized Relation

location	item	sales (in million dollars)	count (in thousands)
Asia	TV	15	300
Europe	TV	12	250
North_America	TV	28	450
Asia	computer	120	1000
Europe	computer	150	1200
North_America	computer	200	1800

Table 5.3: A generalized relation for the sales in 1997.

location \ item	TV		computer		*both_items*	
	sales	count	sales	count	sales	count
Asia	15	300	120	1000	135	1300
Europe	12	250	150	1200	162	1450
North_America	28	450	200	1800	228	2250
all_regions	45	1000	470	4000	525	5000

Table 5.4: A crosstab for the sales in 1997.

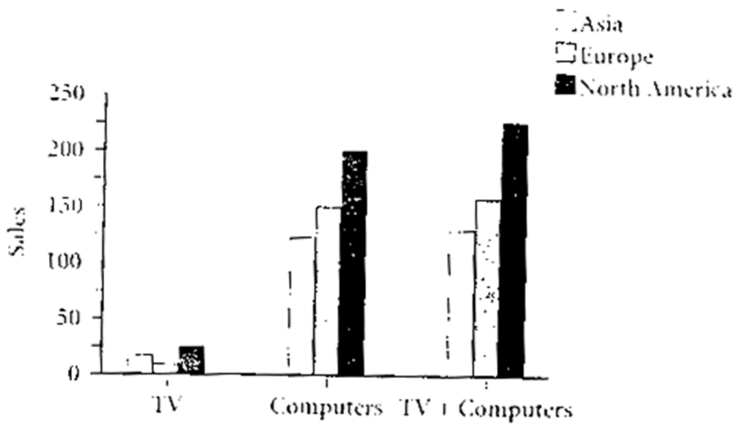


Figure 5.2 Bar chart representation of the sales in 1999

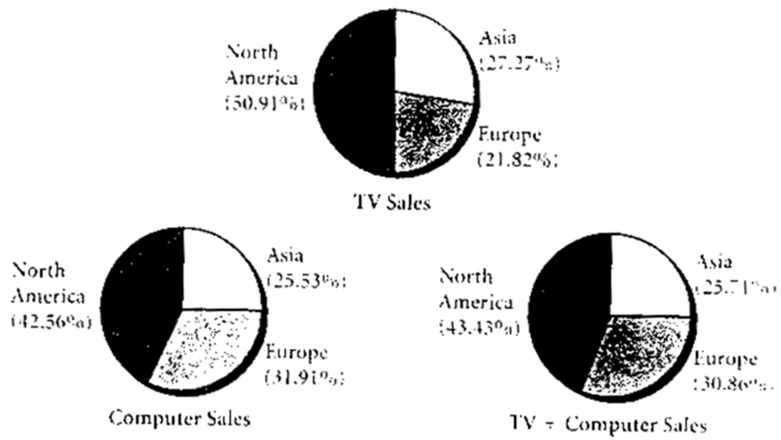
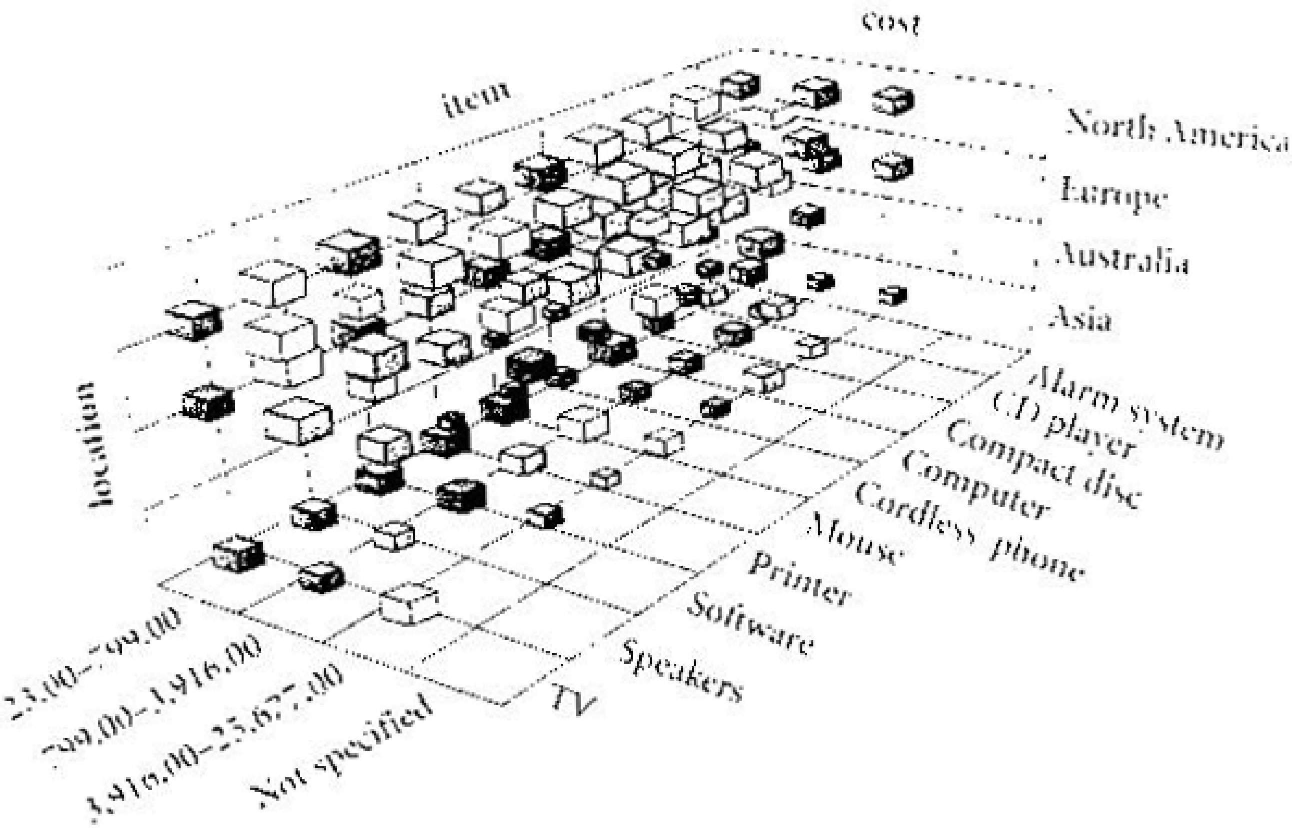


Figure 5.3 pie chart representation of the sales in 1999



3-D cube view representation of the sales in 1999

location \ item	TV		computer		*both_items*	
	sales	count	sales	count	sales	count
Asia	15	300	120	1000	135	1300
Europe	12	250	150	1200	162	1450
North_America	28	450	200	1800	228	2250
all_regions	45	1000	470	4000	525	5000

Table 5.4: A crosstab for the sales in 1997.

$$\forall X, target\_class(X) \Rightarrow condition_1(X)[t:w_1] \vee \ldots \vee condition_m[t:w_m]$$

$$\forall X, item(X) = "computer" \Rightarrow$$
$$(location(X) = "Asia")[t:25.00\%] \vee (location(X) = "Europe")[t:30.00\%] \vee$$
$$(location(X) = "North\_America")[t;45.00\%]$$

Mining Class Comparisons

- Comparison: Comparing two or more classes
- Method:
  - Partition a 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

Quantitative Discriminant Rules

- Cj = target class
- q_a = a generalized tuple covers some tuples of a class
  - but can also cover some tuples of a contrasting class
- d-weight
  - range: [0, 1]

$$d-weight = \frac{count(q_a \in C_j)}{\sum_{i=1}^{m} count(q_a \in C_i)}$$

- quantitative discriminant rule form

$$\forall X,\ target\_class(X) \Leftarrow condition(X) \quad [d : d\_weight]$$

Example: Quantitative Discriminant Rule

Status	Birth_country	Age_range	Gpa	Count
Graduate	Canada	25-30	Good	90
Undergraduate	Canada	25-30	Good	210

Count distribution between graduate and undergraduate students for a generalized tuple

- Quantitative discriminant rule

$$\forall X,\ graduate\_student(X) \Leftarrow$$
$$birth\_country(X) = \text{"}Canada\text{"} \wedge age\_range(X) = \text{"}25-30\text{"} \wedge gpa(X) = \text{"}good\text{"} \quad [d: 30\%]$$

  - where 90/(90 + 210) = 30%

Class Description

- Quantitative characteristic rule

$$\forall X,\ target\_class(X) \Rightarrow condition(X) \quad [t : t\_weight]$$

  - necessary
- Quantitative discriminant rule

$$\forall X,\ target\_class(X) \Leftarrow condition(X) \quad [d : d\_weight]$$

  - sufficient
- Quantitative description rule

$$\forall X,\ target\_class(X) \Leftrightarrow$$
$$condition_1(X)[\mathrm{t} : \mathrm{w}_1, \mathrm{d} : \mathrm{w}'_1] \vee \ldots \vee condition_n(X)[\mathrm{t} : \mathrm{w}_n, \mathrm{d} : \mathrm{w}'_n]$$

  - necessary and sufficient

Example: Quantitative Description Rule

Location/item		TV			Computer			Both_items	
	Count	*t-wt*	*d-wt*	*Count*	*t-wt*	*d-wt*	*Count*	*t-wt*	*d-wt*
Europe	80	25%	40%	240	75%	30%	320	100%	32%
N_Am	120	17.65%	60%	560	82.35%	70%	680	100%	68%
Both_ regions	200	20%	100%	800	80%	100%	1000	100%	100%

Crosstab showing associated t-weight, d-weight values and total number (in thousands) of TVs and computers sold at AllElectronics in 1998

- Quantitative description rule for target class *Europe*

$$\forall X, Europe(X) \Leftrightarrow$$

$$(item(X)="TV")[\mathrm{t}:25\%,\mathrm{d}:40\%] \vee (item(X)="computer")[\mathrm{t}:75\%,\mathrm{d}:30\%]$$

Summary

- Generalization Approaches
  - Data-cube approach
  - Attribute-oriented induction