COURSE NAME:
DATA WAREHOUSING & DATA MINING

LECTURE 20
TOPICS TO BE COVERED:

- Mining complex data objects
- Spatial databases

MINING COMPLEX DATA OBJECTS: GENERALIZATION OF STRUCTURED DATA

- Set-valued attribute
  - Generalization of each value in the set into its corresponding higher-level concepts
  - Derivation of the general behavior of the set, such as the number of elements in the set, the types or value ranges in the set, or the weighted average for numerical data
  - E.g., *hobby* = {*tennis, hockey, chess, violin, nintendo_games*} generalizes to {*sports, music, video_games*}
- List-valued or a sequence-valued attribute
  - Same as set-valued attributes except that the order of the elements in the sequence should be observed in the generalization

GENERALIZING SPATIAL AND MULTIMEDIA DATA

- Spatial data:
  - Generalize detailed geographic points into clustered regions, such as business, residential, industrial, or agricultural areas, according to land usage
  - Require the merge of a set of geographic areas by spatial operations
- Image data:
  - Extracted by aggregation and/or approximation
  - Size, color, shape, texture, orientation, and relative positions and structures of the contained objects or regions in the image
- Music data:
  - Summarize its melody: based on the approximate patterns that repeatedly occur in the segment
  - Summarized its style: based on its tone, tempo, or the major musical instruments played

GENERALIZING OBJECT DATA

- Object identifier: generalize to the lowest level of class in the class/subclass hierarchies
- Class composition hierarchies
  - generalize nested structured data
  - generalize only objects closely related in semantics to the current one
- Construction and mining of object cubes
  - Extend the attribute-oriented induction method
    - Apply a sequence of class-based generalization operators on different attributes
    - Continue until getting a small number of generalized objects that can be summarized as a concise in high-level terms
  - For efficient implementation
    - Examine each attribute, generalize it to simple-valued data
    - Construct a multidimensional data cube (object cube)
    - Problem: it is not always desirable to generalize a set of values to single-valued data

AN EXAMPLE: PLAN MINING BY DIVIDE AND CONQUER

- Plan: a variable sequence of actions
  - E.g., Travel (flight): <traveler, departure, arrival, d-time, a-time, airline, price, seat>
- Plan mining: extraction of important or significant generalized (sequential) patterns from a planbase (a large collection of plans)
  - E.g., Discover travel patterns in an air flight database, or
  - find significant patterns from the sequences of actions in the repair of automobiles
- Method
  - Attribute-oriented induction on sequence data
    - A generalized travel plan: <small-big-small>
  - Divide & conquer:Mine characteristics for each subsequence
    - E.g., big: same airline, small-big: nearby region

A TRAVEL DATABASE FOR PLAN MINING

- Example: Mining a travel planbase

Travel plans table

plan#	action#	departure	depart_time	arrival	arrival_time	airline	…
1	1	ALB	800	JFK	900	TWA	…
1	2	JFK	1000	ORD	1230	UA	…
1	3	ORD	1300	LAX	1600	UA	…
1	4	LAX	1710	SAN	1800	DAL	…
2	1	SPI	900	ORD	950	AA	…
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

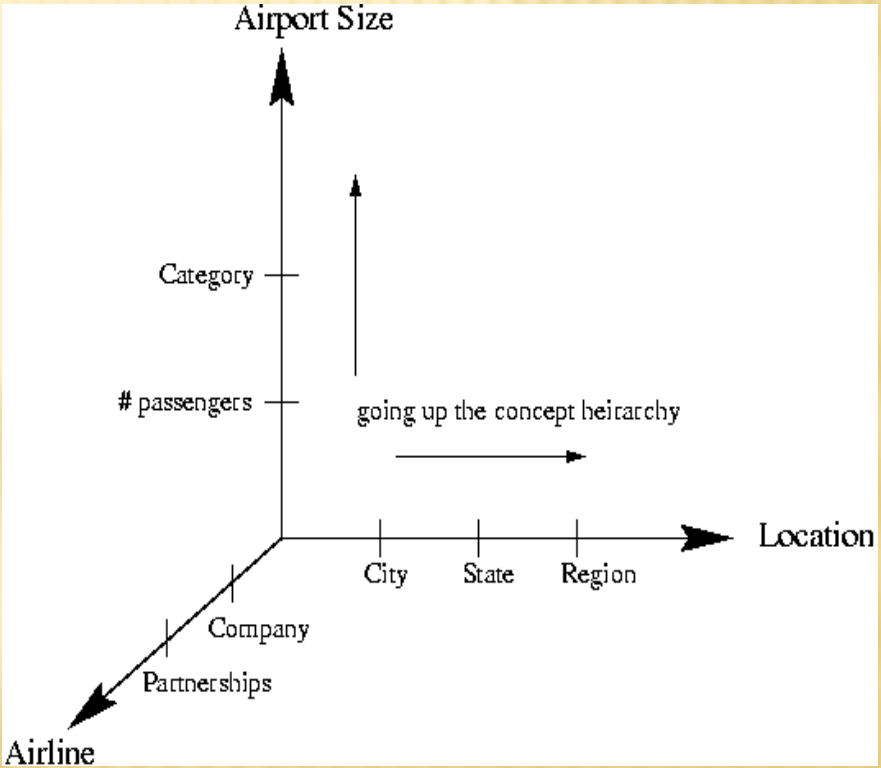
Airport info table

airport_code	city	state	region	airport_size	…
1	1	ALB		800	…
1	2	JFK		1000	…
1	3	ORD		1300	…
1	4	LAX		1710	…
2	1	SPI		900	…
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.

MULTIDIMENSIONAL ANALYSIS

- Strategy
  - Generalize the planbase in different directions
  - Look for sequential patterns in the generalized plans
  - Derive high-level plans

A multi-D model for the planbase



MULTIDIMENSIONAL GENERALIZATION

Multi-D generalization of the planbase

Plan#	Loc_Seq	Size_Seq	State_Seq
1	ALB - JFK - ORD - LAX - SAN	S - L - L - L - S	N - N - I - C - C
2	SPI - ORD - JFK - SYR	S - L - L - S	I - I - N - N
.	.		.
.	.		.
.	.		.

Merging consecutive, identical actions in plans

Plan#	Size_Seq	State_Seq	Region_Seq	…
1	S - L+ - S	N+ - I - C+	E+ - M - P+	…
2	S - L+ - S	I+ - N+	M+ - E+	…
.		.		.
.		.		.
.		.		.

$$flight(x, y,) \wedge airport\_size(x, S) \wedge airport\_size(y, L)$$
$$\Rightarrow region(x) = region(y) \quad [75\%]$$

GENERALIZATION-BASED SEQUENCE MINING

- Generalize planbase in multidimensional way using dimension tables
- Use # of distinct values (cardinality) at each level to determine the right level of generalization (level-“planning”)
- Use operators *merge* “+”, *option* “[]” to further generalize patterns
- Retain patterns with significant support

GENERALIZED SEQUENCE PATTERNS

- AirportSize-sequence survives the min threshold (after applying *merge* operator):

  S-*L*$^{+}$-**S** [35%], *L*$^{+}$-**S** [30%], **S**-*L*$^{+}$ [24.5%], *L*$^{+}$ [9%]

- After applying *option* operator:

  [S]-*L*$^{+}$-[**S]** [98.5%]

  - Most of the time, people fly via large airports to get to final destination

- Other plans: 1.5% of chances, there are other patterns: **S**-**S**, **L**-**S**-**L**

SPATIAL DATA WAREHOUSING

- Spatial data warehouse: Integrated, subject-oriented, time-variant, and nonvolatile spatial data repository for data analysis and decision making
- Spatial data integration: a big issue
  - Structure-specific formats (raster- vs. vector-based, OO vs. relational models, different storage and indexing, etc.)
  - Vendor-specific formats (ESRI, MapInfo, Integraph, etc.)
- Spatial data cube: multidimensional spatial database
  - Both dimensions and measures may contain spatial components

DIMENSIONS AND MEASURES IN SPATIAL DATA WAREHOUSE

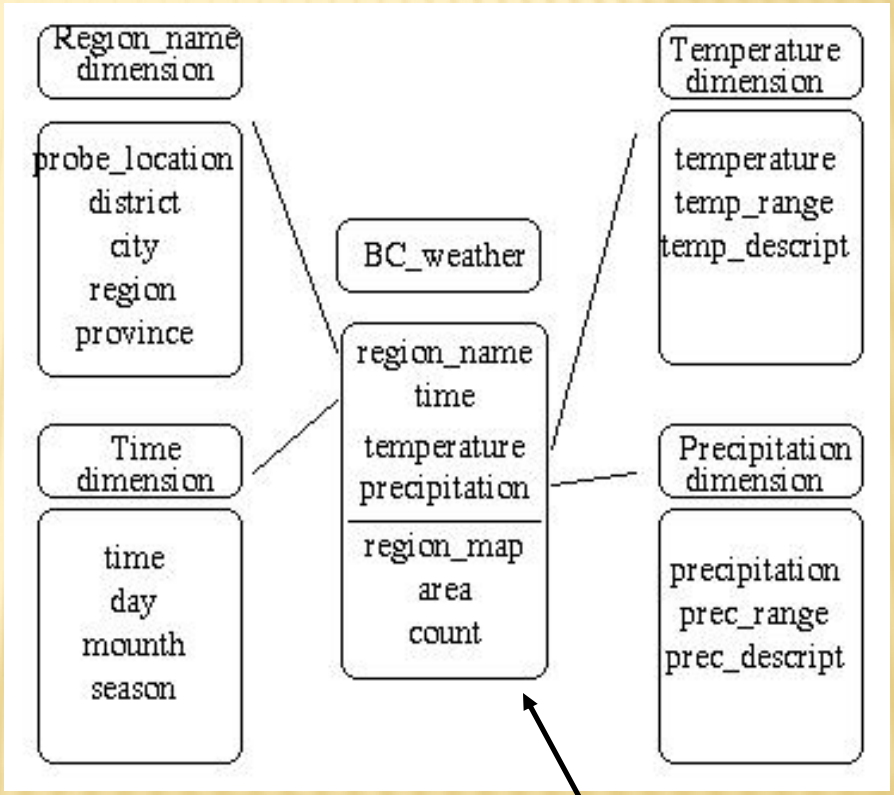
- Dimension modeling
  - Non-spatial
    - e.g. temperature: 25-30 degrees generalizes to *hot*
  - Spatial to non-spatial
    - e.g. region "B.C." generalizes to description "*western provinces*"
  - spatial-to-spatial
    - e.g. region "Burnaby" generalizes to region "Lower Mainland"
- Measures
  - numerical
    - distributive (e.g. count, sum)
    - algebraic (e.g. average)
    - holistic (e.g. median, rank)
  - spatial
    - collection of spatial pointers (e.g. pointers to all regions with 25-30 degrees in July)

EXAMPLE: BC WEATHER PATTERN ANALYSIS

- Input
  - A map with about 3,000 weather probes scattered in B.C.
  - Daily data for temperature, precipitation, wind velocity, etc.
  - Concept hierarchies for all attributes
- Output
  - A map that reveals patterns: merged (similar) regions
- Goals
  - Interactive analysis (drill-down, slice, dice, pivot, roll-up)
  - Fast response time
  - Minimizing storage space used
- Challenge
  - A merged region may contain hundreds of "primitive" regions (polygons)

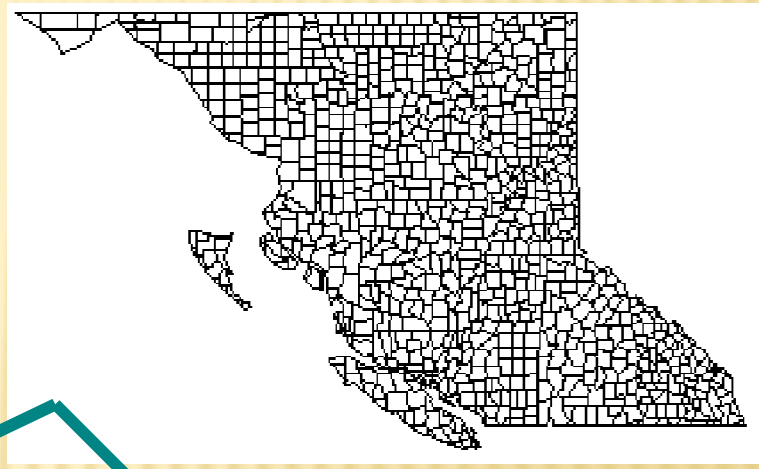
STAR SCHEMA OF THE BC WEATHER WAREHOUSE

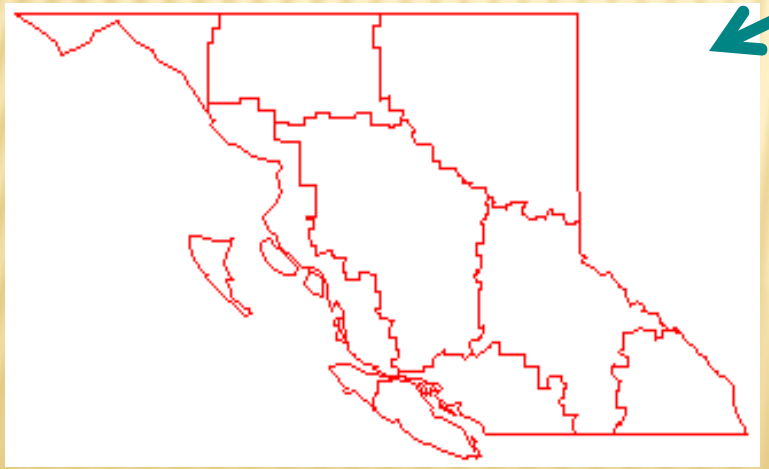
- Spatial data warehouse
  - Dimensions
    - region_name
    - time
    - temperature
    - precipitation
  - Measurements
    - region_map
    - area
    - count

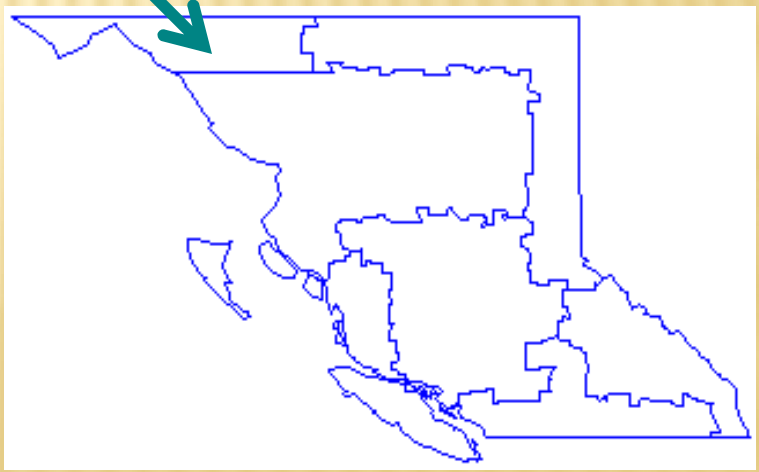


SPATIAL MERGE

- **Precomputing all: too much storage space**
- **On-line merge: very expensive**







METHODS FOR COMPUTATION OF SPATIAL DATA CUBE

- On-line aggregation: collect and store pointers to spatial objects in a spatial data cube
  - expensive and slow, need efficient aggregation techniques
- Precompute and store all the possible combinations
  - huge space overhead
- Precompute and store rough approximations in a spatial data cube
  - accuracy trade-off
- Selective computation: only materialize those which will be accessed frequently
  - a reasonable choice

SPATIAL ASSOCIATION ANALYSIS

- Spatial association rule: $A \Rightarrow B$ [s%, c%]
  - A and B are sets of spatial or nonspatial predicates
    - Topological relations: *intersects, overlaps, disjoint,* etc.
    - Spatial orientations: *left_of, west_of, under,* etc.
    - Distance information: *close_to, within_distance,* etc.
  - s% is the support and c% is the confidence of the rule
- Examples

is_a(x, large_town) ^ intersect(x, highway) → adjacent_to(x, water)

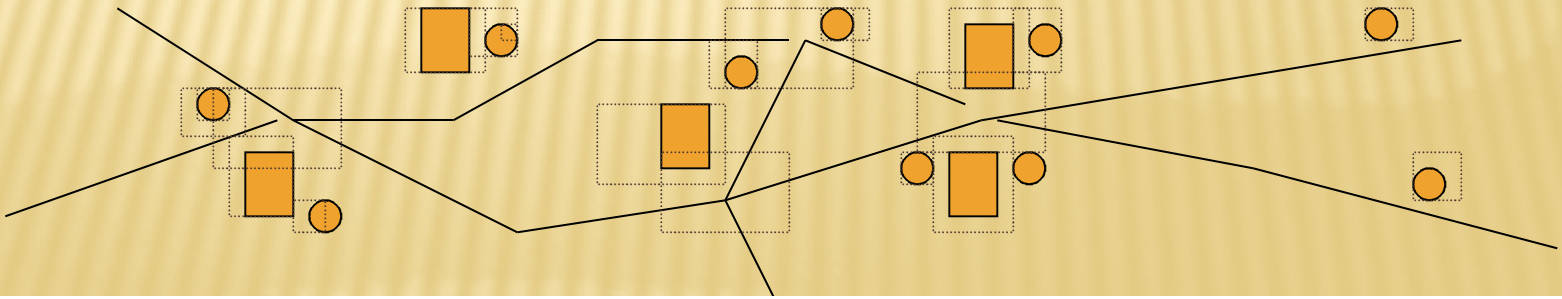
[7%, 85%]

is_a(x, large_town) ^adjacent_to(x, georgia_strait) → close_to(x, u.s.a.)

[1%, 78%]

PROGRESSIVE REFINEMENT MINING OF SPATIAL ASSOCIATION RULES

- Hierarchy of spatial relationship:
  - *g_close_to*: *near_by*, *touch*, *intersect*, *contain*, etc.
  - First search for rough relationship and then refine it
- Two-step mining of spatial association:
  - Step 1: Rough spatial computation (as a filter)
  - Step2: Detailed spatial algorithm (as refinement)
    - Apply only to those objects which have passed the rough spatial association test (no less than *min_support*)



SPATIAL CLASSIFICATION AND SPATIAL TREND ANALYSIS

- Spatial classification
  - Analyze spatial objects to derive classification schemes, such as decision trees in relevance to certain spatial properties (district, highway, river, etc.)
  - Example: Classify regions in a province into *rich* vs. *poor* according to the average family income
- Spatial trend analysis
  - Detect changes and trends along a spatial dimension
  - Study the trend of nonspatial or spatial data changing with space
  - Example: Observe the trend of changes of the climate or vegetation with the increasing distance from an ocean