Complex Types of Data

Learning Outcomes

- Multidimensional analysis and descriptive mining of complex data objects
- Mining spatial databases
- Mining time-series and sequence data

Generalising Spatial and Multimedia Data

- Spatial data
  - Generalise 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, colour, shape, texture, orientation, and relative positions and structures of the contained objects or regions in the image
- Music data
  - Summarise its melody: based on the approximate patterns that repeatedly occur in the segment
  - Summarise its style: based on its tone, tempo, or the major musical instruments played
- Spatio-Temporal data
  - Hurricane data, environmental data, global warming data

Generalising Object Data

- Object identifier
  - Generalise to the lowest level of class in the class/subclass hierarchies
- Class composition hierarchies
  - generalise nested structured data
  - generalise 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 generalisation operators on different attributes
    - Continue until getting a small number of generalised objects that can be summarized as a concise in high-level terms
  - For efficient implementation
    - Examine each attribute, generalise it to simple-valued data
    - Construct a multidimensional data cube (object cube)

Example: Plan Mining by Divide and Conquer

- Plan: a variable sequence of actions
  - E.g., Travel (flight): <traveller, departure, arrival, d-time, a-time, airline, price, seat>
- Plan mining
  - extract significant generalised (sequential) patterns from a plan-base (a large collection of plans)
  - E.g., Discover travel patterns in an air flight database
- Method
  - Attribute-oriented induction on sequence data
    - A generalised 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 plan-base

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	...
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

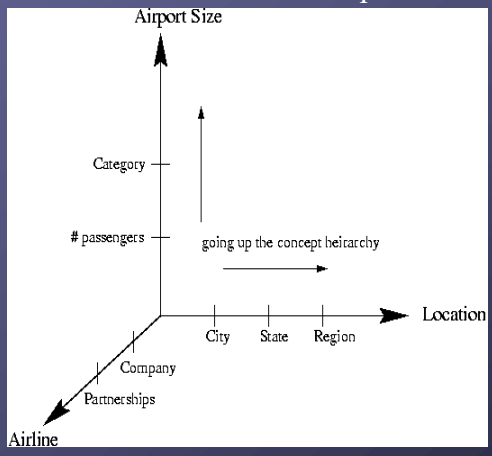
Airport info table

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
  - Generalise the plan-base in different directions
  - Look for sequential patterns in the generalised plans
  - Derive high-level plans

A multi-D model for the plan-base



Multidimensional generalisation

Multi-D generalisation 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+	…
.		.		.
.		.		.
.		.		.

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

Generalisation-Based Sequence Mining

- Dimension Tables
  - Generalise plan-base in multidimensional way using dimension tables
- Cardinality
  - Use the number of distinct values at each level to determine the right level of generalisation (level-"planning")
- Operators
  - Use operators *merge* "+", *option* "[]" to further generalise patterns
- Retain patterns with significant support

Generalised Sequence Patterns

- Airport Size-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 non-volatile 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, Inter-graph, etc.)
- Spatial data cube
  - multidimensional spatial database
  - Both dimensions and measures may contain spatial components

Dimensions and Measures in SDW

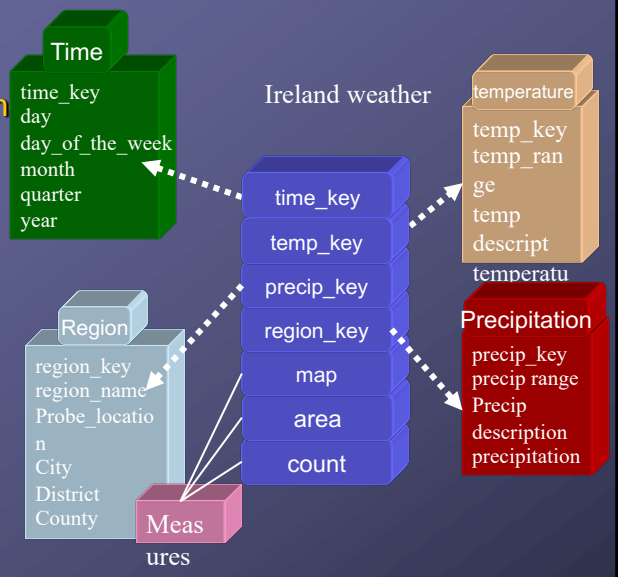
- Dimension modelling
  - Non-spatial
    - e.g. temperature: 25-30 degrees generalises to *hot*
  - Spatial to non-spatial
    - e.g. city "Limerick" generalises to description "*Munster*"
  - Spatial to spatial
    - e.g. region "Athlone" generalises to region "Lower Mainland"
- Measures
  - Numerical
    - 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 of weather pattern analysis

- Input
  - A map with about 3,000 weather probes scattered in an area
  - 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
  - Minimising storage space used
- Challenge
  - A merged region may contain hundreds of "primitive" regions (polygons)

Star Schema of a Weather Warehouse

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



Computation Methods for Spatial Data Cube

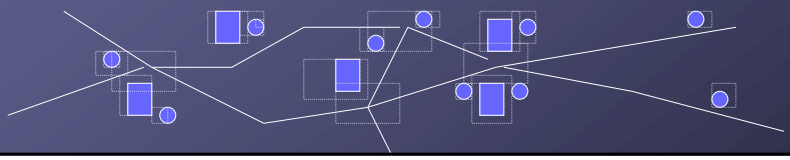
- On-line aggregation
  - Collect and store pointers to spatial objects in a spatial data cube
  - Expensive and slow, need efficient aggregation techniques
- Pre-processing
  - Pre-compute and store all the possible combinations: huge space overhead
  - Pre-compute and store rough approximations in a spatial data cube: accuracy trade-off
- Selective computation
  - Only materialise those which will be accessed frequently: a reasonable choice

Spatial Association Analysis

- Spatial association rule: *A -> B* [*s*%, *c*%]
  - A and B are sets of spatial or non-spatial 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, school) ^close_to(x, sport_centre) –> close_to(x,park) [1%, 78%]*

Progressive Refinement Mining of Spatial Association Rules

- Hierarchy of spatial relationship
  - *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)
    - Use MBR (minimum bounding rectangle) for rough estimation
  - 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

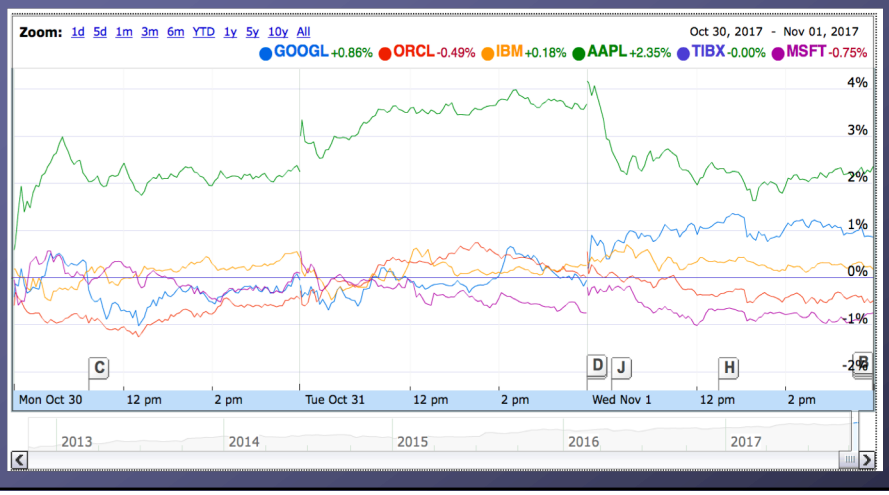
- Spatial classification
  - Analyse spatial objects to derive classification schemes, such as decision trees in relevance to certain spatial properties (district, highway, river, etc.)
  - Example: Classify regions in a province into *rich* vs. *poor* according to the average family income
- Spatial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

Mining Time-Series and Sequence Data

- Time-series database
  - Consists of sequences of values or events changing with time
  - Data is recorded at regular intervals
  - Characteristic time-series components
    - Trend, cycle, seasonal, irregular
- Applications
  - Financial: stock price, inflation
  - Biomedical: blood pressure
  - Meteorological: precipitation

Mining Time-Series and Sequence Data

Time-series plot



Mining Time-Series and Sequence Data: Trend analysis

- A time series Data
  - can be illustrated as a time-series graph which describes a point moving over time
- Categories of Time-Series Movements
  - Long-term or trend movements (trend curve)
  - Cyclic movements or cycle variations, e.g., business cycles
  - Seasonal movements or seasonal variations
    - i.e, almost identical patterns that a time series appears to follow during corresponding months of successive years
  - Irregular or random movements

Estimation of Trend Curve

- The freehand method
  - Fit the curve by looking at the graph
  - Costly and barely reliable for large-scaled data mining
- The least-square method
  - Find the curve minimising the sum of the squares of the deviation of points on the curve from the corresponding data points
- The moving-average method
  - Eliminate cyclic, seasonal and irregular patterns
  - Loss of end data
  - Sensitive to outliers

Discovery of Trend in Time-Series

- Estimation of seasonal variations
  - Seasonal index
    - Set of numbers showing the relative values of a variable during the months of the year
    - E.g., if the sales during October, November, and December are 80%, 120%, and 140% of the average monthly sales for the whole year, respectively, then 80, 120, and 140 are seasonal index numbers for these months
  - Deseasonalised data
    - Data adjusted for seasonal variations
    - E.g., divide the original monthly data by the seasonal index numbers for the corresponding months

Discovery of Trend in Time-Series

- Estimation of cyclic variations
  - If (approximate) periodicity of cycles occurs, cyclic index can be constructed in the same manner as seasonal indexes
- Estimation of irregular variations
  - By adjusting the data for trend, seasonal and cyclic variations
- Prediction
  - With systematic analysis of the trend, cyclic, seasonal, and irregular components, it is possible to make long- or short-term predictions with reasonable quality

Similarity Search in Time-Series Analysis

- Similarity search
  - finds data sequences that differ only slightly from the given query sequence
- Two categories of similarity queries
  - Whole matching: find a set of sequences that are similar to each other (as a whole)
  - Subsequence matching: find sequences that contain sub-sequences that are similar to a given query sequence x
- Typical Applications
  - Financial market & Market basket data analysis
  - Scientific databases
  - Medical diagnosis

Data Transformation

- Domain
  - Many techniques for signal analysis require the data to be in the frequency domain
- Usually data-independent transformations are used
  - The transformation matrix is determined a priori
    - E.g., discrete Fourier transform (DFT), discrete wavelet transform (DWT)
  - The distance between two signals in the time domain is the same as their Euclidean distance in the frequency domain
  - DFT does a good job of concentrating energy in the first few coefficients
  - If we keep only first a few coefficients in DFT, we can compute the lower bounds of the actual distance

Subsequence Matching

- Breaking sequences
  - into a set of pieces of window with length *w*
- Extracting the features
  - For each subsequence inside the window
- Mapping
  - each sequence to a “trail” in the feature space
- Dividing the trail
  - of each sequence into “sub-trails” and represent each of them with minimum bounding rectangle
- Multi-piece assembly algorithm
  - to search for longer sequence matches

Sequential Pattern Mining

- Frequency
  - Mining of frequently occurring patterns related to time or other sequences
- Patterns
  - Sequential pattern mining usually concentrates on symbolic patterns
- Examples
  - Renting “Star Wars”, then “Empire Strikes Back”, then “Return of the Jedi” in that order
  - Collection of ordered events within an interval
- Applications
  - Targeted marketing & Customer retention
  - Weather prediction

Mining Sequences (cont.)

Customer-sequence

CustId	Video sequence
1	{(C), (H)}
2	{(AB), (C), (DFG)}
3	{(CEG)}
4	{(C), (DG), (H)}
5	{(H)}

Map Large Itemsets

Large Itemsets	MappedID
(C)	1
(D)	2
(G)	3
(DG)	4
(H)	5

Sequential patterns with support ≥ 0.25

{(C), (H)}

{(C), (DG)}

Sequential pattern mining: Parameters

- Duration of a time sequence *T*
  - Sequential pattern mining can be confined to the data within a specified duration
  - E.g., Subsequence corresponding to the year of 2009
  - E.g., Partitioned sequences, such as every year, or every week after stock crashes, or every two weeks before and after a volcano eruption
- Event folding window *w*
  - If $w = T$, time-insensitive frequent patterns are found
  - If $w = 0$ (no event sequence folding), sequential patterns are found where each event occurs at a distinct time instant
  - If $0 < w < T$, sequences occurring within the same period *w* are folded in the analysis

Sequential pattern mining: Parameters

- Time interval, *int*, between events in the discovered pattern
  - *int* = 0: no interval gap is allowed, i.e., only strictly consecutive sequences are found
    - Ex. "Find frequent patterns occurring in consecutive weeks"
  - $min\_int \leq int \leq max\_int$: find patterns that are separated by at least *min_int* but at most *max_int*
    - Ex. "If a person rents movie A, it is likely s/he will rent movie B within 30 days" ($int \leq 30$)
  - $int = c \neq 0$: find patterns carrying an exact interval
    - Ex. "Every time when Dow Jones drops more than 5%, what will happen exactly two days later?" ($int = 2$)

Sequential pattern mining

- Other methods for specifying the kinds of patterns
  - Serial episodes: $A \rightarrow B$
  - Parallel episodes: A & B
  - Regular expressions: $(A \mid B)C^*(D \rightarrow E)$
- Methods for sequential pattern mining
  - Variations of Apriori-like algorithms, e.g., GSP (Generalised Sequence Patterns)
  - Database projection-based pattern growth
    - Similar to the frequent pattern growth without candidate generation

Periodicity Analysis

- Periodicity is everywhere
  - tides, seasons, daily power consumption, etc.
- Full periodicity
  - Every point in time contributes (precisely or approximately) to the periodicity
- Partial periodicity: A more general notion
  - Only some segments contribute to the periodicity
    - Jim reads NY Times 7:00-7:30 am every week day
- Cyclic association rules
  - Associations which form cycles
- Methods
  - Full periodicity: FFT, other statistical analysis methods
  - Partial and cyclic periodicity: Variations of Apriori-like mining methods