Principles of Knowledge Discovery in Data

Fall 2002

Chapter 10: Multimedia and Spatial Data Mining

Dr. Osmar R. Zaïane

University of Alberta

Summary of Previous Chapter

- Introduction to Web Mining
  - What are the incentives of web mining?
  - What is the taxonomy of web mining?
- Web Content Mining: Getting the Essence From Within Web Pages.
- Web Structure Mining: Are Hyperlinks Information?
- Web Usage Mining: Exploiting Web Access Logs.
- Warehousing the Web

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
- **Multimedia and Spatial Data Mining**
- *Other topics if time permits*

Chapter 10 Objectives

•Present some applications of DM and KDD in Multimedia Data.

• Present some DM applications and solutions in Spatial data.

Multimedia

Outline

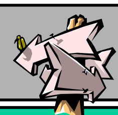
- Knowledge Discovery and Data Mining
- Confusion with MDM
- Mining from Sound
- Mining from Video
- Mining from Images
- Spatial Data Mining

Many Steps in KD Process

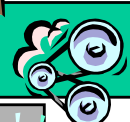
- Gathering the data together

- Cleanse the data and fit it in together

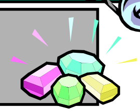
- Select the necessary data



- **Crunch and squeeze the data to extract the *essence* of it**



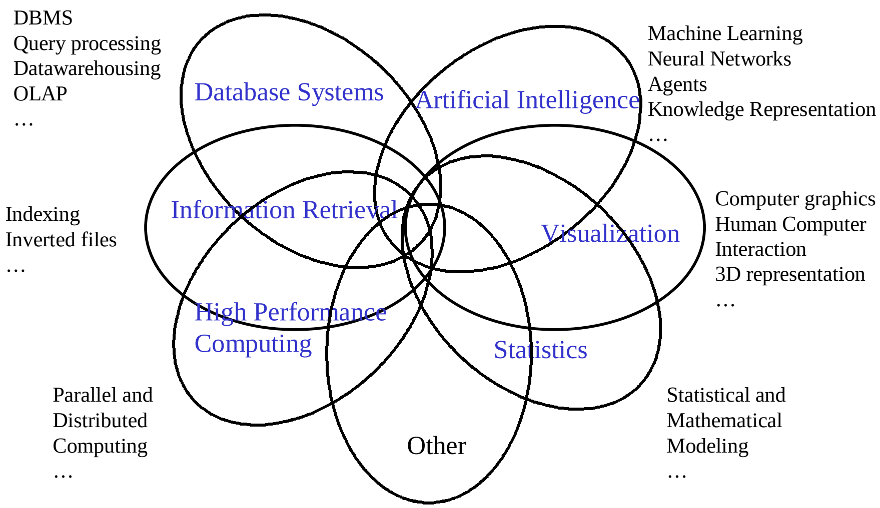
- Evaluate the output and use it



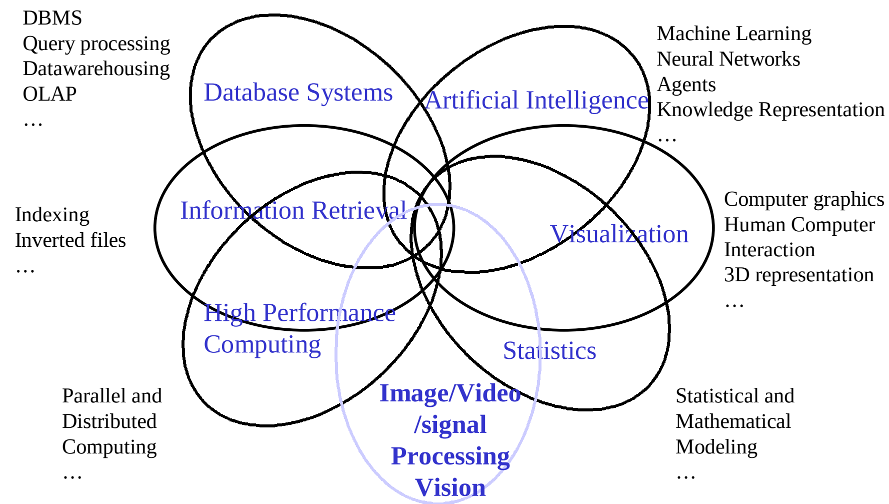
So What Is Data Mining?

In theory, *Data Mining* is <u>a step</u> in the knowledge discovery process. It is the <u>extraction of implicit information</u> from <u>a large dataset</u>.

KDD at the Confluence of Many Disciplines



KDD at the Confluence of Many Disciplines



Outline

- Knowledge Discovery and Data Mining
- **Confusion with MDM**
- Mining from Sound
- Mining from Video
- Mining from Images
- Spatial Data Mining

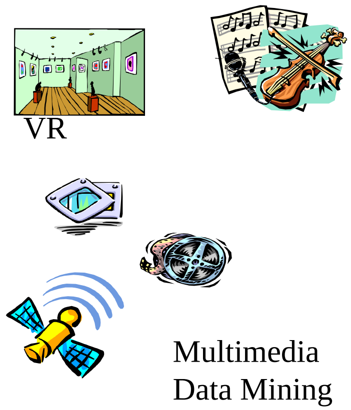
Confusion

If multimedia subsumes everything, is every data mining multimedia mining?

Are Web mining and multimedia mining the same thing?

No!
Multimedia mining is not mining FROM the Web.
It is better to define or restrict the type of media we consider.

Classes of Mining Problems

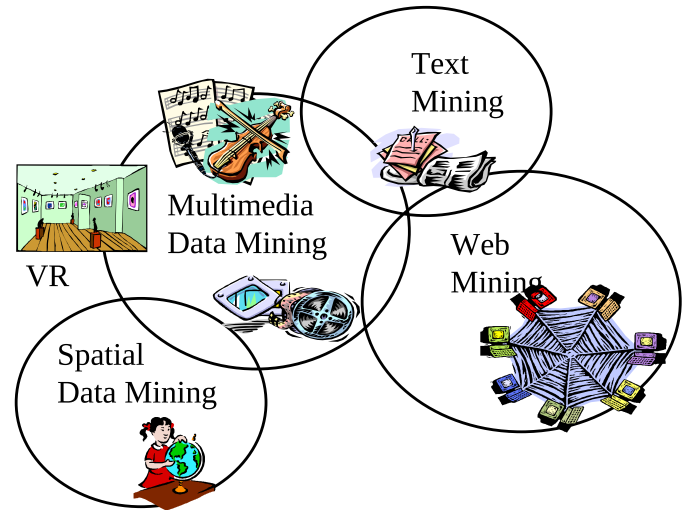


Web Mining
- Content
- Structure
- Usage

Text Mining
- Unstructured
- Semi-structured

Geo-Spatial Data Mining
- Spatial
- Non-spatial

The Big Picture



Content-Based Image Retrieval

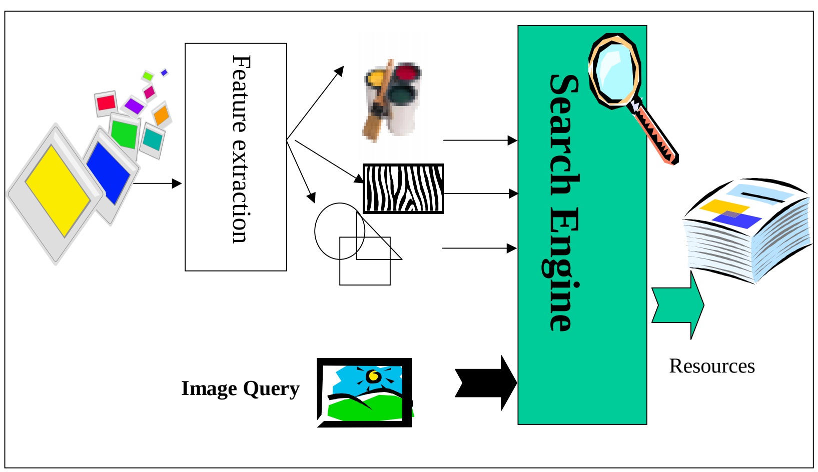
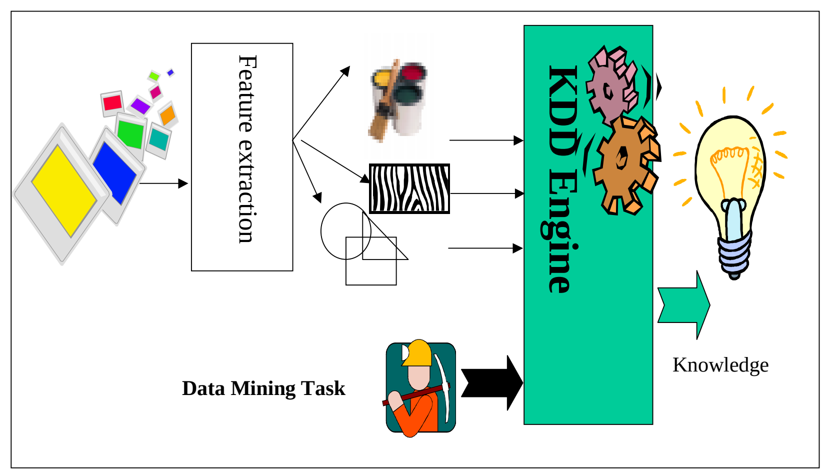
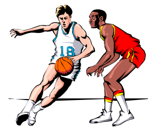


Image Mining





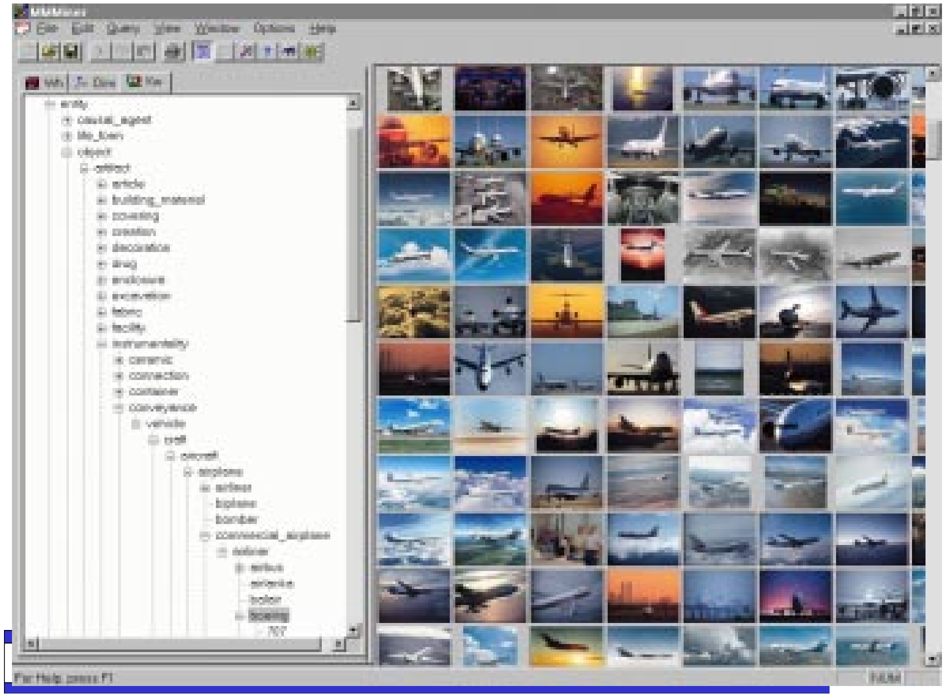
Types of MDM

- Use of multimedia in KDD
- Mining multimedia descriptors (metadata)
- Extraction of features from multimedia for a higher level application
- Pure multimedia mining

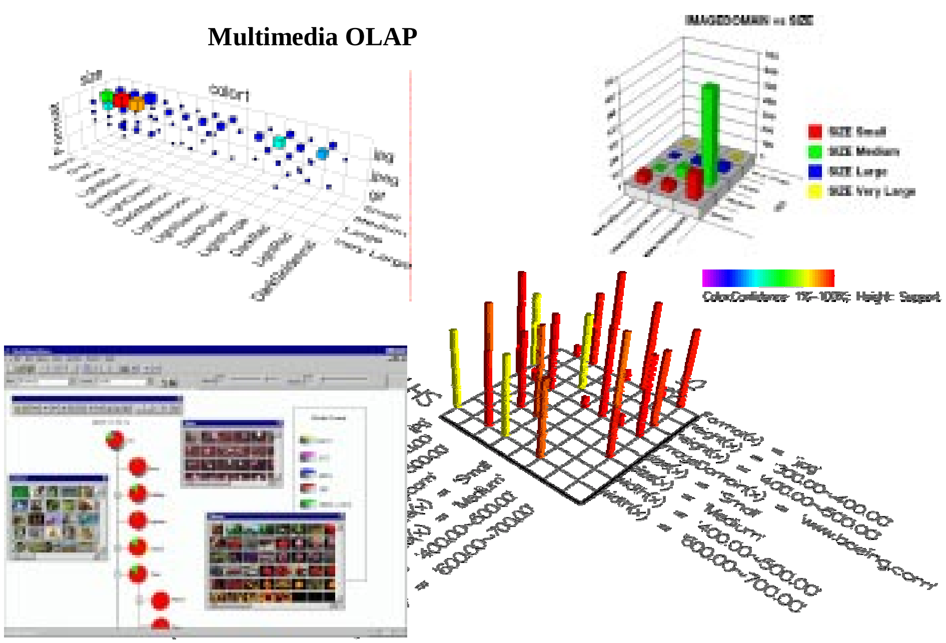
Not real MDM

MDM

Content-Based Multimedia Mining vs.
Non Content-Based Multimedia Mining



Multimedia OLAP



Outline

- Knowledge Discovery and Data Mining
- Confusion with MDM
- **Mining from Sound**
- Mining from Video
- Mining from Images
- Spatial Data Mining

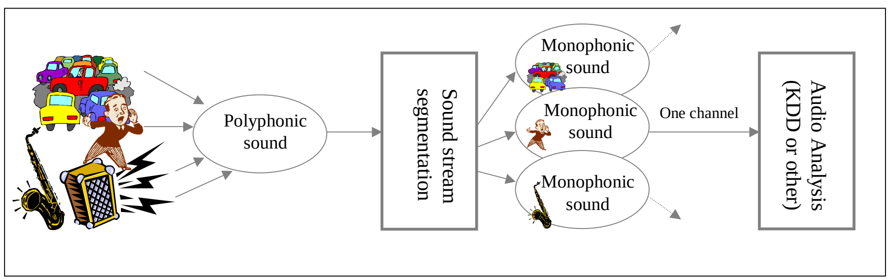
State of the Art

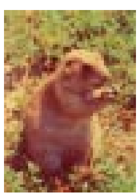
- Not much substantial has been done on knowledge discovery from sound
- Classification of high level sound segments using neural networks.
- Mining speech is usually done after conversion speech-to-text or close caption.

Objectives

- Sound Recognition
- Sound Indexing (improve audio retrieval)
- Sound Segment Identification (recognize themes and melodies in music))
- Noise Filtering (reduction)
- Compression
- Categorization (music/speech, gender, accents, …)

Analysis of Auditory Scenes





Prairie dogs have different alarm calls for different predators ➔ communication language based on pitch (Northern Arizona University study)

Analysis Procedure

- Digitize sound sample
- Segment sound according to some heuristic
- Convert sound segments into properties
- Physical features (frequency, duration, energy, spectrum, harmonics, zcr, formant, prosody, etc.)
- Perceptual features (pitch, timbre, rhythm)

Some Examples

- Enhancing old audio recordings using neural networks (Czyzewski 1996)
- Discriminating between speech and music using nearest neighbour classifier with modulation of energy etc.(Sheiner et al. 1993)
- Learn prosodic patterns in Chinese using decision trees after isolating syllables and pitch (Chen et al. 2000)

Case of clustering sound segments

- Digitize sound sample
- Segment sound according to some heuristic
  - speech: 1/10th sec. intervals
- Convert sound segments to frequency domain using the FFT
- Classify segments using feature identification derived from FFT analysis

Unsupervised Classification

- Use of clustering can automate the analysis of sound segment classification
- Needs a robust technique that allows complex similarity metrics to be used
- The ROCK algorithm works well

Clustering Sound with Rock

- We examined:
  - frequency composition
  - harmonic composition
- Devised
  - frequency composition comparison metric
  - harmonic composition comparison metric

Results

- Only speech sounds were analyzed
- ROCK works well but is sensitive during the initial clustering phase to θ threshold settings
- Used modifications to address sensitivity issues
  – Trimming (to reduce “gravity” effect)
  – Threshold Searching

Outline

- Knowledge Discovery and Data Mining
- Confusion with MDM
- Mining from Sound
- **Mining from Video**
- Mining from Images
- Spatial Data Mining

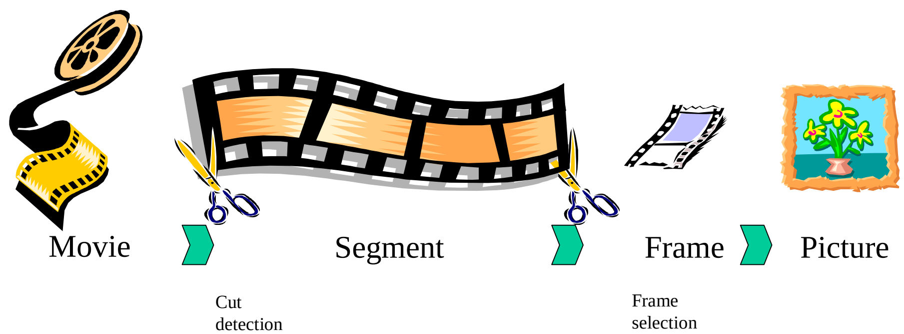
State of the Art

- Very complex problem
- Multiple channels (motion pictures, polyphonic sound, text…)
- Most application concentrate on the motion images and ignore other channels
- Manual annotation and segmentation
- Use of closed captioning (text)

Objectives

- Automatic annotation
- Segmentation
- Identification of objects
- Spatio-temporal patterns (path traversals, tracking, …)
- Understand (discover) general trends in movies

Converting Video to Still Images



Examples of Video Mining

- Mining surveillance camera data
  - Detecting trajectories with stereo cameras
  - Discovering outliers in trajectories
  - Raymond Ng et al. 1998

Examples of Video Mining

- Detecting Narrative structure of news broadcasts
  - Uses dedicated tools for video segmentation
  - Use closed captioning
  - Classifies segments into: anchor shot, footage with voice over, or sound bite
  - Used for retrieval or browsing
  - Shearer et al. 2000

Examples of Video Mining

- Tracking Pedestrians (Papageorgio et al. 1998)
  - Wavelet transf. In gray scale of F S R sides
  - Learning with Support Vector Machine classifier

Examples of Video Mining

- Mining commercial movies
  - Extracting content features such as class of events (violence, happiness…), explosion, rudeness, etc.
  - Expressing sequences of events
  - Correlating with metadata (actors, box office, budget, etc.) to discover rules
  - If ending=s(explosion,violence)➔ W1 > $5M
  - Daniel Barbara et al. 2000

Outline

- Knowledge Discovery and Data Mining
- Confusion with MDM
- Mining from Sound
- Mining from Video
- **Mining from Images**
- Spatial Data Mining

State of the Art

- Identification of objects is difficult
- Diverse image modeling approaches
- Manual annotation is common (semantic descriptors)
- Classification of images/objects
- Clustering of images/objects
- Association rule mining for image content

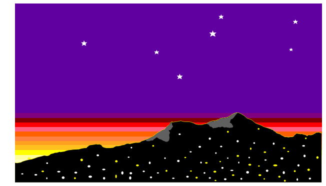
Objectives

- Discrimination of images (small number of classes)
- Grouping of images (clustering)
- Recognize (compound) objects
- Enumerate (estimate) objects
- Determine good image models for image interpretation and indexing

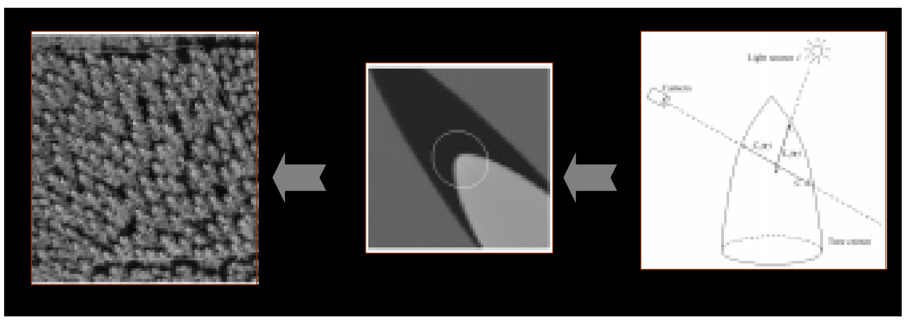
Analysis Procedure

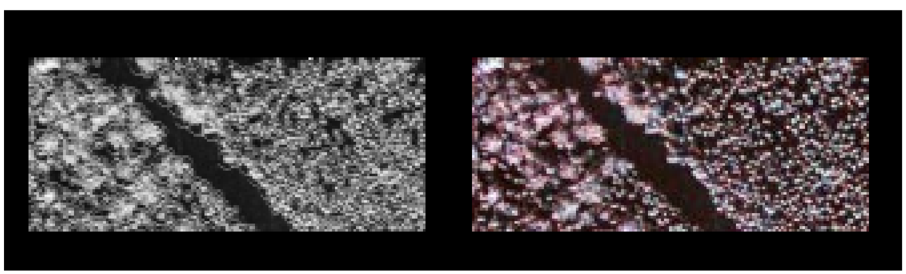
- **Collection and retrieval phase** (Collecting images)
- **Image selection** (Choosing relevant images for the task at hand)
- **Image pre-processing** (Extracting visual features)
- **Mining** (Discovering patterns at individual image level of image group level)
- **Analysis** (validating and interpreting the discovered patterns)

Mining in Image & Raster Databases

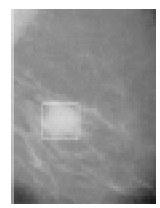


- Magellan project (Fayyad et al.'96, JPL).
  - identify volcanos on Venus surface
  - over 30,000 high resolution images
  - Resolution accuracy: 80%
  - 3 steps: data focusing, feature extraction, and classification learning
- POSSII project (Palomar Obervatory Sky Survey II,)
  - 2×10^9 stellar objects (galaxies, stars, etc.) classified
  - Resolution:one magnitude better than in previous studies
  - Classification accuracy: no normalization 75%, with normalization 94%, and compared with neural networks.
- QuakeFinder (Stolorz et al'96): Find earth quakes from space.
  - using statistics, massive parallelism, and global optimization

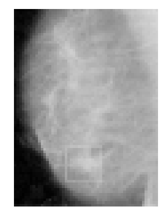




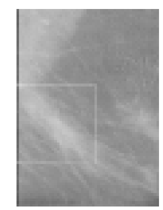
Mammography Database



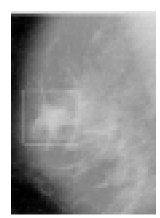
Circumscribed Mass



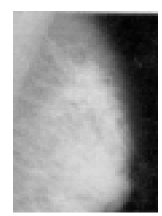
Spiculated Mass



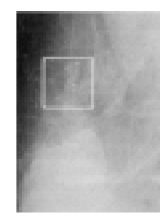
Ill-defined Mass



Architectural Distortion



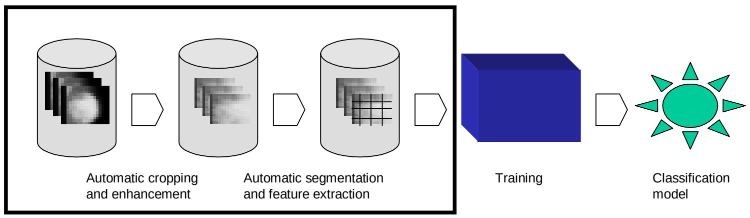
Asymmetry



Clustered Microcalcifications

Digital Mammograms

- Mammograms are difficult to read even by specialists due to low contrast and different types of tissue.
- In order to extract visual features Image enhancement is necessary

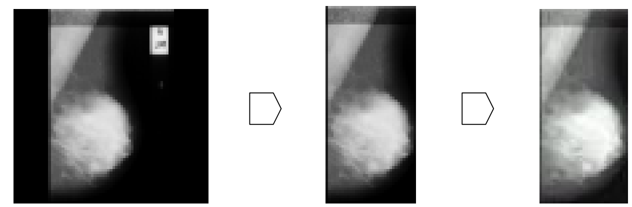


Improving the Quality of Images

- Digitization introduces noise
- Inconsistent illumination conditions
- Inconsistent sizes and distributions

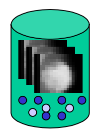
Automatic Cropping: Removes unwanted parts and artifacts.

Enhancement: Diminishes the effect of over brightness and over darkness. Histogram equalization to increase contrast range.



Original mammogram Cropping Histogram Equalization

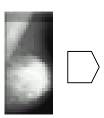
Feature Extraction



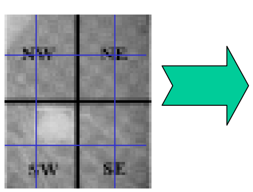
In original DB images are associated with many attributes (age position, tissue…)

We opted for:

- Position of breast (left/right)
- Type of tissue (dense/fatty/fatty-glandular)

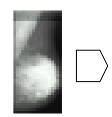


Transaction (ImageID, $F_1, F_2, F_3, \ldots F_f$)



- Mean $\bar{x} = \frac{\sum x}{N}$
- Variance $\sigma^2 = \frac{\sum (x-\bar{x})^2}{N}$
- Skewness $Sk = \frac{1}{N}\left(\frac{(x-\bar{x})}{\sigma}\right)^3$
- Kurtosis $Kurt = \frac{1}{N}\left(\frac{(x-\bar{x})}{\sigma}\right)^4 - 3$

Neural Networks



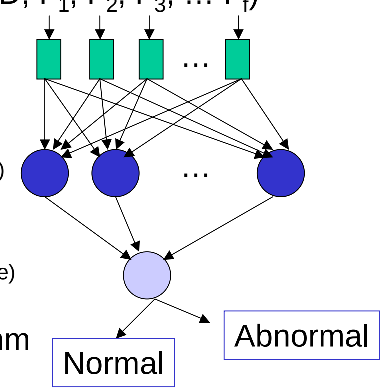
Transaction (IID, F_1, F_2, F_3, … F_f)

Input Layer (69 nodes)

Hidden Layer (10 nodes)

Output Layer (1 node)

Back-propagation algorithm Adjusts internal weights



Association Rules

- Association rule mining aims at discovering associations between items in a transactional database.
- Given D={T_1…T_n} a set of transactions and I={i_1…i_n} a set of items such that any T_i in D is a set of items in I.
- An association rule is an implication A➔B where A and B are subsets of T_i given some support and confidence thresholds.
- In our case T is an image and the items are the extracted features in addition to the known class label

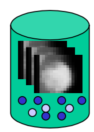
Transaction (IID, **class**, F_1, F_2, F_3, … F_f)

Association Rules With Constraints

- We want to find associations between extracted features and class labels
- Constrain the association rule mining such that the interesting rules A➔B are such that the consequent B is always a class label and the antecedent A is always a conjunction of extracted features.
- We used a constrained version of apriori algorithm to find frequent itemsets.

$F_\alpha \wedge F_\beta \wedge F_\gamma \wedge \ldots \wedge F_\delta$ ➔ **class**

Data Set



- We used a dataset from the mammographic Image Analysis Society (MIAS).
- It was used in other published research.
- The corpus consists of 322 images with 208 normal, 63 benign and 51 malign.
- Location of abnormality, radius, breast position, type of breast tissue, tumor type…
- Notice that these attributes are available for the training set but are not available with images to classify ➔ Shall not be used in testing phase.

Some Results

- Use 90% of images for training and 10% for testing.
- We considered 10 splits of the image collection.

Neural Network-based

Database split	Success ratio (percentage)
1	96.87
2	90.62
3	90.62
4	78.125
5	81.25
6	84.375
7	65.625
8	75
9	56.25
10	93.75
	Average: 81.25

- Average high but inconsistent accuracy
- Performed extremely well compared to other methods in literature

Association Rule-based

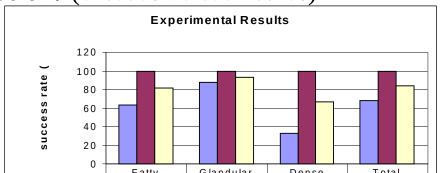
Database split	Success ratio (percentage)
1	67.647
2	79.412
3	67.647
4	61.765
5	64.706
6	64.706
7	64.706
8	64.706
9	67.647
10	88.235
	Average: 69.11

- Average not as high but more consistent accuracy
- support at 10% and no confidence
- close to other results in literature

Observations

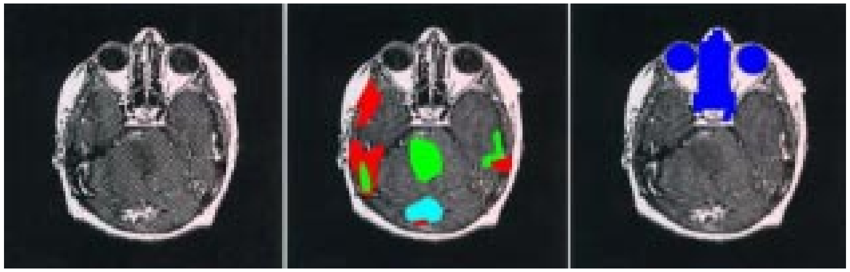
- Pre-processing and feature extraction is very important and influences the accuracy rates.
- Association rule-based classifier is sensitive to the unbalance and size of dataset.
- We made experiments with AR on equilibrated distributions of normal and abnormal with split used in Christoyianni et al. (84.09% versus 75.2%)

We obtained a lower recognition rate for fatty abnormal but higher rates for all normal cases.



	Fatty	Glandular	Dense	Total
Abnormal	63.6	87.5	33.3	68.1
Normal	100	100	100	100
Total	81.8	93.7	66.6	84.09

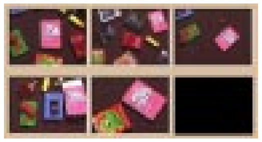
Associations in Medical Images



- Associations: presence of lesions, relative positions and spatial relationships.
- Associations with diagnoses and attributes in patient records.

Locales: Colour Localization

➔ Search by **Object** in images



- Locales can have ***any shape;***
- Locales are ***not*** necessarily ***disjoint;***
- Locales can be ***disconnected;***
- The set of locales is ***not*** necessarily ***complete.***

Locales and Their Features

Visual

Colour(X, *colour*)
Size(X, *size*)
Texture(X, *texture*)
Shape(X, *shape*)

Topology

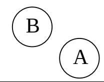
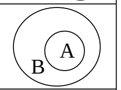
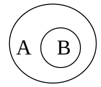
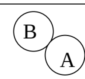
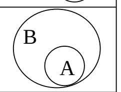
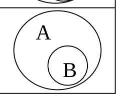
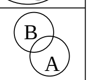
H-next-to(X,Y)
V-next-to(X,Y)
Overlap(X,Y)
Include(X,Y)

Location

Vertical(X, *v*)
Horizontal(X, *h*)

Movement

Motion(X, *motion*)
Speed(X, *speed*)

$A° \cap B°$	$\delta A \cap \delta B$	$\delta A \cap B°$	$A° \cap \delta B$		
$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	A disjoint B	
$\neg\varnothing$	$\varnothing$	$\neg\varnothing$	$\varnothing$	A inside B	
$\neg\varnothing$	$\varnothing$	$\varnothing$	$\neg\varnothing$	A contains B	
$\neg\varnothing$	$\neg\varnothing$	$\varnothing$	$\varnothing$	A equals B	
$\varnothing$	$\neg\varnothing$	$\varnothing$	$\varnothing$	A meets B	
$\neg\varnothing$	$\neg\varnothing$	$\neg\varnothing$	$\varnothing$	A covered by B	
$\neg\varnothing$	$\neg\varnothing$	$\varnothing$	$\neg\varnothing$	A covers B	
$\neg\varnothing$	$\neg\varnothing$	$\neg\varnothing$	$\neg\varnothing$	A overlaps B	

Association Rules

- **Multimedia Association Rule with Recurrent Items:** associate visual object features in images or video frames**:**

$$\alpha_1 P_1 \wedge \alpha_2 P_2 \wedge \ldots \wedge \alpha_n P_n \rightarrow \lambda_1 Q_1 \wedge \lambda_2 Q_2 \wedge \ldots \wedge \lambda_m Q_m \ (c\%)$$

P_i, $i \in [1..n]$, Q_j, $j \in [1..m]$ are predicates instantiated to topological, visual and kinematics descriptors, α_i and λ_j are integers, αP is true iff P has α occurrences, and c is the confidence.

How to Find Associations

- Find all frequent items (more frequent than minimum support);
- Combine frequent items into itemsets;
- Find frequent itemsets;
- Use frequent itemsets for produce association rules.

➔The problem is to find frequent itemsets
Most famous algorithm is Apriori (R. Agrawal 1994).
There are many other variations and improvements.
However, Apriori misses all item-sets with recurrent items. The re-occurrence of an object can be more significant that its existence in the image.

Transaction-Based vs. Object-Based

Transaction-based support:
Support of an object is the percentage of transactions containing the object. (Number of T_{ob}/ Number of Transactions)

Object-based support:
Support of an object is the percentage of objects equal to the object. (Number of Object occurrences/ Number of Objects)

Finding Itemsets with Reoccurring Items

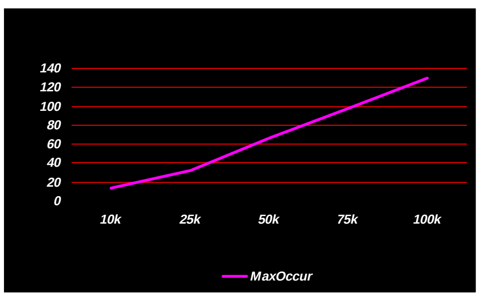
- Devised a new algorithm MaxOccur (zaiane ICDE'2000) (variation of Apriori)
- Finds the maximum time an object can reoccur in an image
- Us max occurrence to increase candidate itemsets during the join

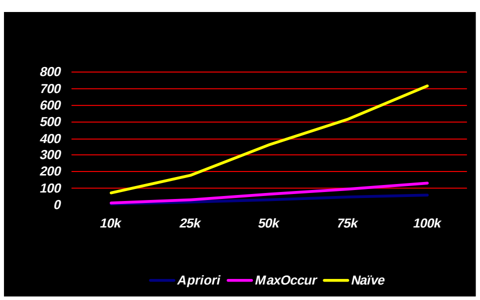
Multimedia Association Rules With Recurrent Spatial Relationships

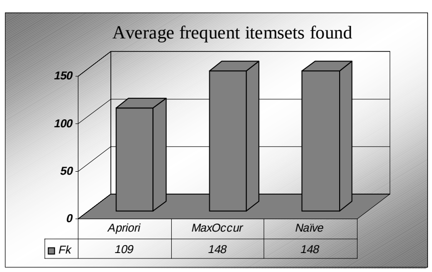
- A spatial relationship is a relationship between 2 objects.
- A spatial relationship is frequent only if the pair of objects is frequent.
- A pair of objects is frequent only if each object is frequent.

Overlap (O_1,O_2) frequent ➔ (O_1, O_2) frequent ➔ O_1 and O_2 frequent

1. Calculate Frequent Atomic Items;
2. Calculate Frequent Pairs of Atomic Items;
3. Calculate Frequent combinations of Spatial Predicates and Pairs of Atomic Items;
4. Use the set in (3) as Frequent 1-item-sets and Call MaxOccur.







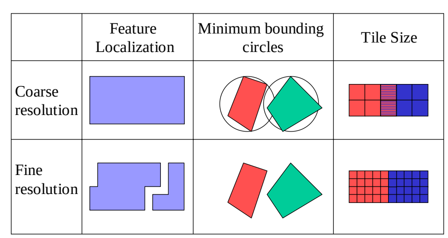
Progressive Resolution Refinement

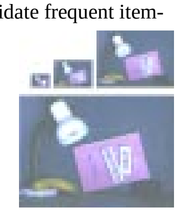
Progressive Resolution Refinement

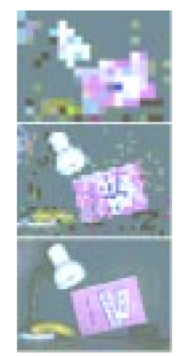
i=0; D_0 =D;
while (i<maxResLevel) do {
  R_i ={sufficiently frequent item-sets at resolution i}
  i=i+1; D_i=Filter(D_{i-1}, R_{i-1});
}

From Coarse to Fine Resolution Mining

Progressively mine finer resolutions only on candidate frequent item-sets

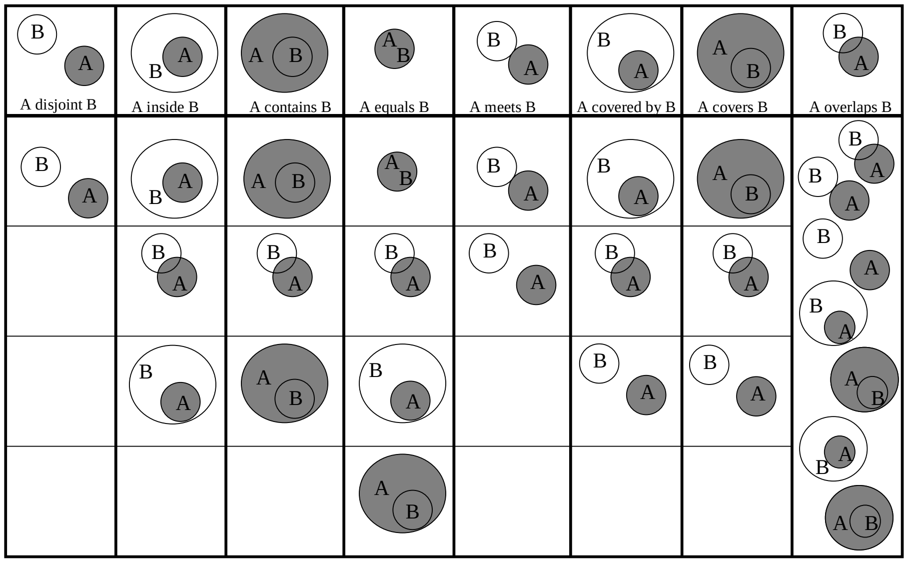




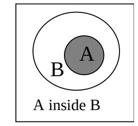


$A^\circ \cap B^\circ$, $\delta A \cap B^\circ$
$A^\circ \cap \delta B$, $\delta A \cap \delta B$

$\emptyset, \emptyset$ $\emptyset, \emptyset$	$\neg\emptyset, \neg\emptyset$ $\emptyset, \emptyset$	$\neg\emptyset, \emptyset$ $\neg\emptyset, \emptyset$	$\neg\emptyset, \emptyset$ $\emptyset, \neg\emptyset$	$\neg\emptyset, \emptyset$ $\emptyset, \emptyset$	$\neg\emptyset, \neg\emptyset$ $\emptyset, \neg\emptyset$	$\neg\emptyset, \emptyset$ $\neg\emptyset, \neg\emptyset$	$\neg\emptyset, \neg\emptyset$ $\neg\emptyset, \neg\emptyset$



The topological relation between two areas A and B at any resolution level is defined by a matrix $\mathcal{R}$:



$$\mathcal{R}(A,B)=\begin{bmatrix} A^\circ \cap B^\circ & \delta A \cap B^\circ \\ A^\circ \cap \delta B & \delta A \cap \delta B \end{bmatrix}$$

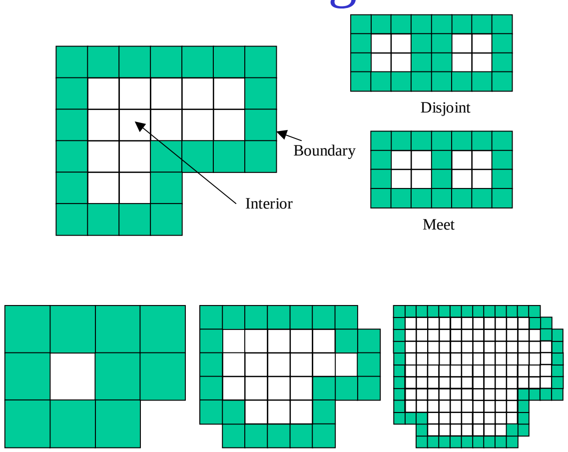
A inside B ➔ $A^\circ \cap B^\circ = \neg\emptyset$ and $\delta A \cap B^\circ = \neg\emptyset$ and
$A^\circ \cap \delta B = \emptyset$ and $\delta A \cap \delta B = \emptyset$

Given ***a*** and ***b*** higher resolutions of **A** and **B**.

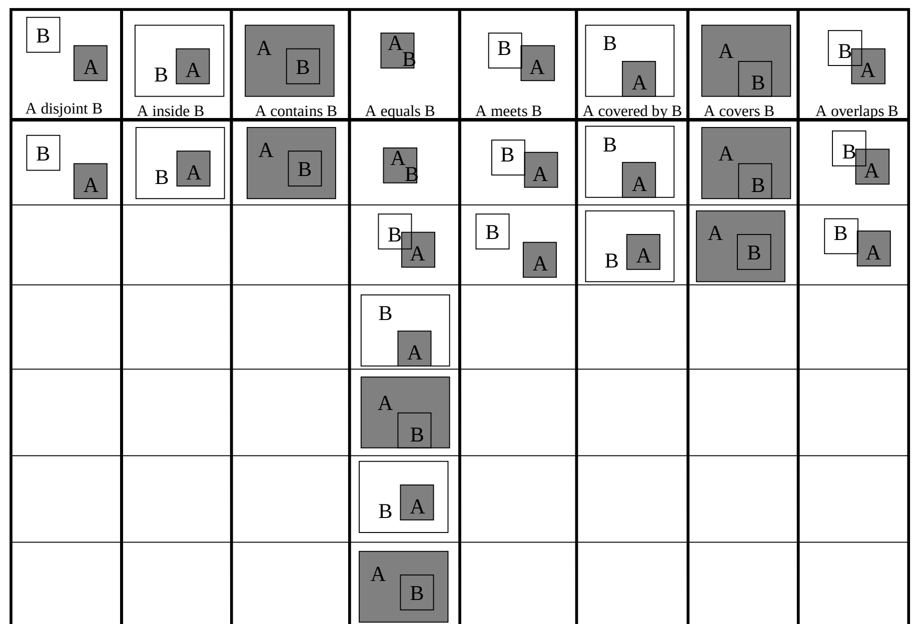
$A^\circ \cap B^\circ = \neg\emptyset$ ➔ $a^\circ \cap b^\circ = \neg\emptyset$
$\delta A \cap B^\circ = \neg\emptyset$ ➔ $\delta a \cap b^\circ = \neg\emptyset$
$A^\circ \cap \delta B = \emptyset$ ➔ $a^\circ \cap \delta b = \emptyset$ or $a^\circ \cap \delta b = \neg\emptyset$
$\delta A \cap \delta B = \emptyset$ ➔ $\delta a \cap \delta b = \emptyset$ or $\delta a \cap \delta b = \neg\emptyset$

$$\mathcal{R}(a,b)=\begin{bmatrix} \neg\emptyset & \neg\emptyset \\ \emptyset & \emptyset \end{bmatrix} \text{ or } \begin{bmatrix} \neg\emptyset & \neg\emptyset \\ \neg\emptyset & \neg\emptyset \end{bmatrix} \text{ or } \begin{bmatrix} \neg\emptyset & \neg\emptyset \\ \emptyset & \neg\emptyset \end{bmatrix}$$

Resolution Refinement With Tile Resizing



$\emptyset, \emptyset$ $\emptyset, \emptyset$	$\neg\emptyset, \neg\emptyset$ $\emptyset, \emptyset$	$\neg\emptyset, \emptyset$ $\neg\emptyset, \emptyset$	$\neg\emptyset, \emptyset$ $\emptyset, \neg\emptyset$	$\neg\emptyset, \emptyset$ $\emptyset, \emptyset$	$\neg\emptyset, \neg\emptyset$ $\emptyset, \neg\emptyset$	$\neg\emptyset, \emptyset$ $\neg\emptyset, \neg\emptyset$	$\neg\emptyset, \neg\emptyset$ $\neg\emptyset, \neg\emptyset$



Summary

- Multimedia data mining one phase, in the knowledge discovery process, that extract implicit patterns from large multimedia content.
- Multimedia mining vs. text mining, web mining, spatial mining.
- Usually: video, images, sound.

Outline

- Knowledge Discovery and Data Mining
- Confusion with MDM
- Mining from Sound
- Mining from Video
- Mining from Images
- **Spatial Data Mining**

Section Objectives

- Introduce two algorithms for KD in large Spatial DB
  - *Nonspatial-Data-Dominated Generalization*
  - *Spatial-Data-Dominated Generalization*
- Show KD has wide applications in Spatial DB.

Spatial Data Mining

- Data in Spatial Database:
  - Non-Spatial component - *usual data, stored in relational DB.*
  - Spatial component - *multi-dimensional, stored in spatial data structures.*

    Spatial data: maps, images from satellites, video cameras, and medical equipment, etc.
- Knowledge Discovery in Spatial BD: is the extraction of
  - *Interesting spatial patterns and features,*
  - *General relationships between spatial and non-spatial data, and*
  - *Other implicit general data characteristics.*

Motivation

- Spatial data availability
- Human limitation
- Needs:

  Knowledge from spatial data is crucial in development of

  Geographical information system

  Medical imaging and robotics systems

Primitives and assumptions

Assumption 1:

- The spatial DB store a large amount, info-rich, relatively reliable and stable data.

Assumption 2:

- A knowledge discovery process is initialized by user's learning request - command-driven discovery.

Assumption 3:

- Strong background knowledge support - conceptual hierarchy information available.

Primitives and Assumptions (con't)

To confine the research to a well-defined domain:

- Rules to be extracted are general data characteristics and/or relationships - generalization rules.
- The spatial database consists of both spatial and nonspatial data. Spatial objects and their associated nonspatial info. are linked to each other.

Spatial Object Representations

In spatial database:

- Spatial data is stored in thematic maps.
- Each thematic map contains a set of spatially ordered objects.
- Each spatial object has a spatial component and nonspatial component,
- Spatial object can be denoted as < geometry, attribute>

APPROACH

- Task:

  Discover generalization rules from data in Spatial DB.

  Discovered knowledge should be represented by high level concepts with a small number of disjuncts.

- Underlining notion:

  To extract general knowledge from spatial DB, generalization needs to be performed on both spatial and non-spatical data. When one of the components is generalized, the other component will be adjusted accordingly.

- Algorithms

  Nonspatial-Data-Dominated Generalization

  Spatial-Data-Dominated Generalization

Algorithm I:
Nonspatial-Data-Dominated Generalization

- Input:

  i) A spatial database consisting of a set of nonspatial data and a spatial map.

  ii) A learning request

  iii) a set of concept hierarchies

- Output:

  A rule that characterizes the general properties/relationships of spatial objects.

Algorithm I:
Nonspatial-Data-Dominated Generalization

- Method:

  1. Collect the set of task-relevant nonspatial data by SQL query
  2. Perform attribute-oriented induction repeatedly on the collected nonspatial data by
     ascending the concept hierarchy,
     merging identical tuples until desired concept level reached,
     collecting spatial object pointers.
  3. Generalize the spatial data
     Retrieve the spatial objects for each generalized nonspatial tuple,
     Perform spatial merge.
  4. Output the generalized rule or the relationship between the generalized nonspatial and spatial data

Algorithm I:
Nonspatial-Data-Dominated Generalization

- Algorithm Summary:

  The learning process generalizes nonspatial data attributes by concept hierarchy ascension and consolidation of adjacent spatial objects with similar attribute values.

  Genaralization terminates when the generalized concept level reaches the desired concept level.

Example:
Nonspatial-Data-Dominated Generalization

- Task:

  To report general precipitation pattern zones of British Columbia in spring 1990.
- Input:

  The spatial DB that stores a map of British Columbia with a set of weather stations (8) scattered around the province.

  Climate data that contains average monthly precipitation for each of the eight regional stations.
- Learning Process:

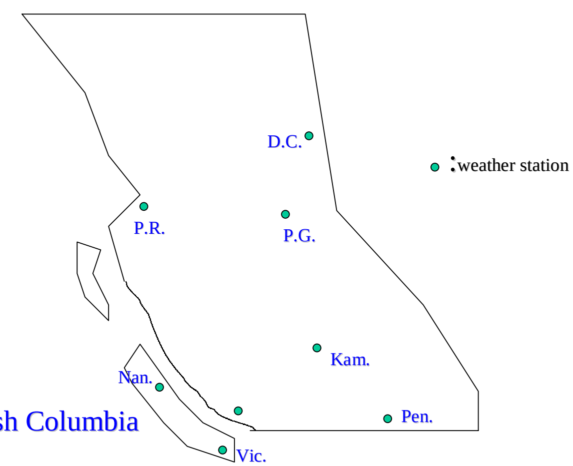
  Perform generalization on nonspatial attribute *precipitation* first using concept hierarchy, then merge crresponding spatial objects accordingly.

Example:
Nonspatial-Data-Dominated Generalization

- Precipitation data (in inch) of 1990

Region	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec	yr total
Nanaimo	6.37	4.36	3.99	2.05	1.47	1.55	0.91	1.01	1.73	4.19	6.06	7.11	41.25
Van.	8.6	6.1	5.3	3.3	3.0	2.7	1.3	1.7	4.1	5.9	10.0	7.8	59.8
Victoria	11.12	9.74	5.15	2.68	2.51	1.07	0.42	2.42	0.95	2.69	2.64	4.36	45.75
P. Rupert	9.8	7.6	8.4	6.7	5.3	4.1	4.7	5.2	7.7	12.2	12.3	11.3	95.16
D. Creek	...	...	...	...	...	...	...	...	...	...	...	...	...
P. George	...	...	...	...	...	...	...	...	...	...	...	...	...
Kamloops	...	...	...	...	...	...	...	...	...	...	...	...	...
Penciton	...	...	...	...	...	...	...	...	...	...	...	...	...

Example:
Nonspatial-Data-Dominated Generalization



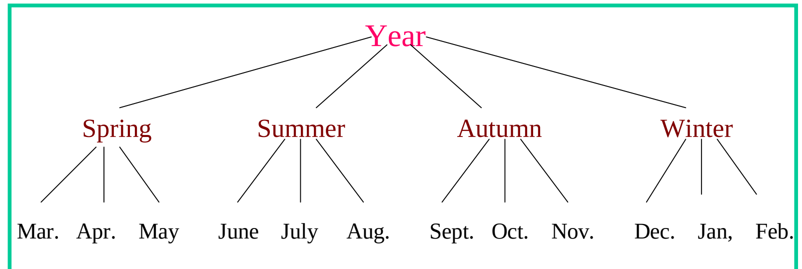
- A map of British Columbia

Example:
Nonspatial-Data-Dominated Generalization

- Two Concept Hierarchies:

  Year-season-month time hierarchy

  Precipitation concept hierarchy



A year-season-month hierarchy

Example:
Nonspatial-Data-Dominated Generalization

- Two Concept Hierarchies:
  - *Year-season-month time hierarchy*
  - *Precipitation concept hierarchy*

High-level precipitation concept hierarchy

very dry (v.d.)	dry (d.)	moderately dry (m.d.)	fair (f.)	moderately wet (m.w.)	wet (w.)	very wet (v.w.)
[0-0.1]*	[0.1-0.3]	[0.3-1.0]	[1.0-1.2]	[1.2-2.0]	[2.0-5.0]	[5.0 & up]

* : Unit: inch

Example:
Nonspatial-Data-Dominated Generalization

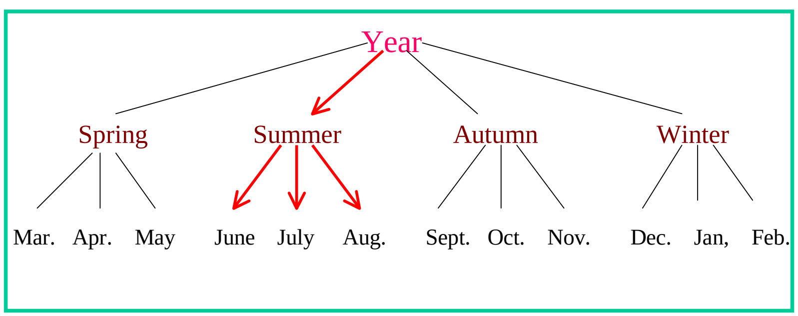
- User's Query:

Report general precipitation pattern zones of B.C. in spring 1990.

extract region
from precipitation-map
where province = "B.C." **and** period = "spring" **and** year = 1990
in relevance to precipitation **and** region

Example:
Nonspatial-Data-Dominated Generalization

- Year-season-month time concept hierarchy



Example:
Nonspatial-Data-Dominated Generalization

- The relevant precipitation data of the regions and its generalization:

Region	Mar.	Apr.	May
Nanaimo	3.99	2.50	1.47
Vancouver	5.3	3.3	3.0
Victoria	5.15	2.68	3.51
…	…	…	…

Average

Region	Mar.	Apr.	May	Summer
Nanaimo	3.99	2.50	1.47	2.85
Vancouver	5.3	3.3	3.0	4.1
Victoria	5.15	2.68	3.51	3.43
…	…	…	…	…

Example:
Nonspatial-Data-Dominated Generalization

- The relevant precipitation data of the regions and its generalization:

Region	Mar.	Apr.	May	Summer	High-level Concept
Nanaimo	3.99	2.50	1.47	2.85	wet
Vancouver	5.3	3.3	3.0	4.1	wet
Victoria	5.15	2.68	3.51	3.43	wet
…	…	…	…	…	…

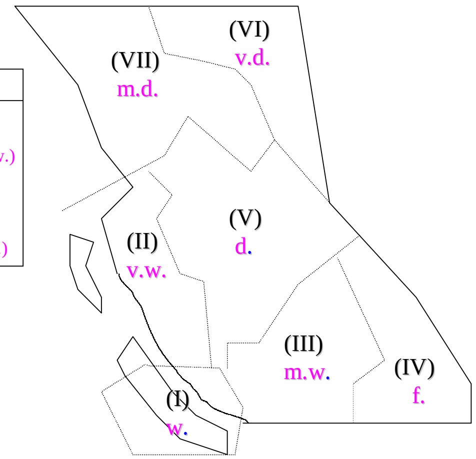
Precipitation concept hierarchy

very dry (v.d.)	dry (d.)	moderately dry (m.d.)	fair (f.)	moderately wet (m.w.)	wet (w.)	very wet (v.w.)
[0-0.1]	[0.1-0.3]	[0.3-1.0]	[1.0-1.2]	[1.2-2.0]	[2.0-5.0]	[5.0 & up]

Example:
Nonspatial-Data-Dominated Generalization

- *Learning result* of precipitation in spring 1990 for B.C.

Region	Rainfall
Region I	wet (w.)
Region II	very wet(v.w.)
Region III	moderately wet(m.w.)
Region IV	fair(f.)
Region V	dry(d.)
Region VI	very dry(v.d.)
Region VII	moderately dry(m.d.)



Algorithm II:
Spatial-Data-Dominated Generalization

- Input:

  i) A spatial database consisting of a set of nonspatial data and a spatial map,

  ii) a learning request,

  iii) a set of concept hierarchies,

  iv) a spatial hierarchy,

- Output:

  A rule that characterizes the general properties/relationships of spatial objects.

Algorithm II:
Spatial-Data-Dominated Generalization

- Method:

  1. Collect the set of task-relevant <u>spatial data</u> by SQL query,
  2. Perform spatial-oriented induction on the collected spatial data by spatial hierarchy ascension
     cluster spatial data objects according to their regions,
     merge the corresponding nonspatial pointers,
     repeat the two steps above until the number of generalized spatial objects is within the threshhold.
  3. Retrieve nonspatial data, generalize nonspatial data for each spatial object,
  4. Output the generalized rule or the relationship between the generalized nonspatial and spatial data.

Algorithm II:
Spatial-Data-Dominated Generalization

- Algorithm Summary:

  The learning process utilizes the spatial hierarchy to obtain generalized objects. The generalized attribute value of the new object is obtained by climbing up the attribute concept hierarchy to find a minimal concept which subsumes the attribute values of the corresponding sub-objects.

Example:
Spatial-Data-Dominated Generalization

- Task:

  To report general temperature pattern in pre-specified regions of British Columbia for summer 1990.

- Input:

  The spatial DB that stores a map of British Columbia with a set of weather stations scattered around the province.

  Climate data that contains min., max., and average monthly temperature for each of the regional stations.

Example:
Spatial-Data-Dominated Generalization

- User's Query:

  Report general temperature pattern in Spring 1990 for B.C.

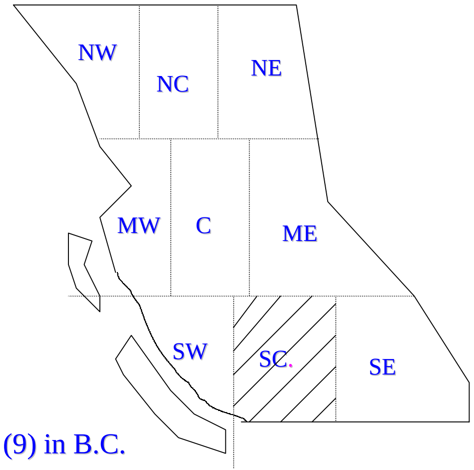
extract characteristic rule
from temperature-map
where province = "B.C." **and** period = "summer" **and** year = 1990
in relevance to temperature **and** region

Example:
Spatial-Data-Dominated Generalization

- High-level temperature concepts:

very cold (v.c.)	cold (c.)	moderately cold (m.c.)	mild (m.)	moderately hot (m.h.)	hot (h.)	very hot (v.h.)
-5 & below	[-5-10]	[10-32]	[32-50]	[50-70]	[70-90]	90 & up

Example: Spatial-Data-Dominated Generalization



- Pre-specified regions (9) in B.C.

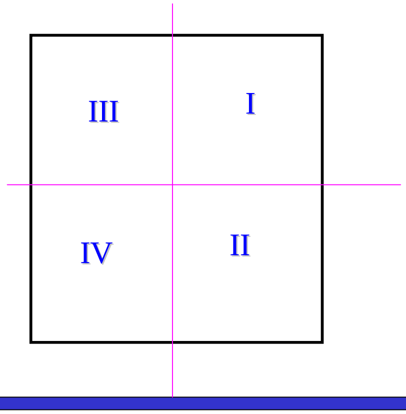
Example: Spatial-Data-Dominated Generalization

South-Central region of B.C.: with 16 cities marked as R1 - R16

R1	R2	R5	R6
R3	R1	R7	R8
R9	R10	R13	R14
R11	R12	R15	R16

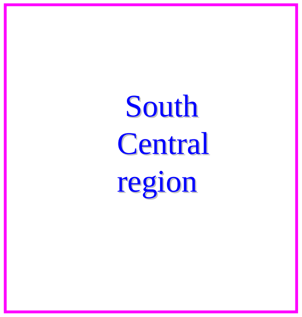
Example: Spatial-Data-Dominated Generalization

- Clustered South-Central region of B.C.



Example: Spatial-Data-Dominated Generalization

- Further clustered South-Central region of B.C.



Example:
Spatial-Data-Dominated Generalization

- Temperature data for south-central region:

Region	June	July	Aug.	summer
R1	62	67	64	64.3
R2	68	65	64	65.7
R3	60	64	63	62.3
R4	…	…	…	63.3
R5	…	…	…	67.5
R6	…	…	…	59.7
R7	…	…	…	59.3
R8	…	…	…	60.2
R9	…	…	…	58.7
R10	…	…	…	68
R11	…	…	…	61.8
R12	…	…	…	59.4
R13	…	…	…	67.3
R14	…	…	…	67.4
R15	…	…	…	63.5
R16	…	…	…	61.3

Avg.: 62.8

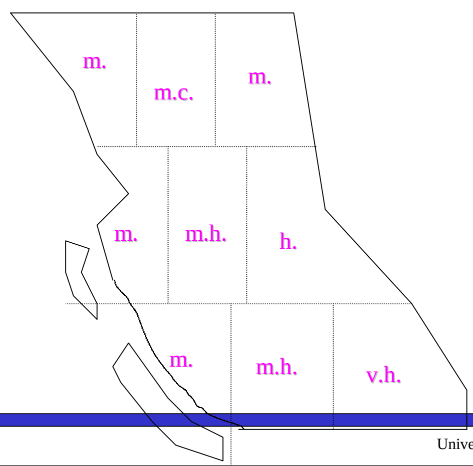
Example:
Spatial-Data-Dominated Generalization

- Genaralized temperature info.:

Region	Temperature
North-West	mild
North-Central	moderately cold
North-East	mild
Mid-West	mild
Central	moderately hot
Mid-East	hot
South-West	mild
South-Central	moderately hot
South-East	very hot

Example:
Spatial-Data-Dominated Generalization

- Generalized temperature pattern of B.C., summer, 1990.



EXTENSIONS

- Interleaved Generalization:

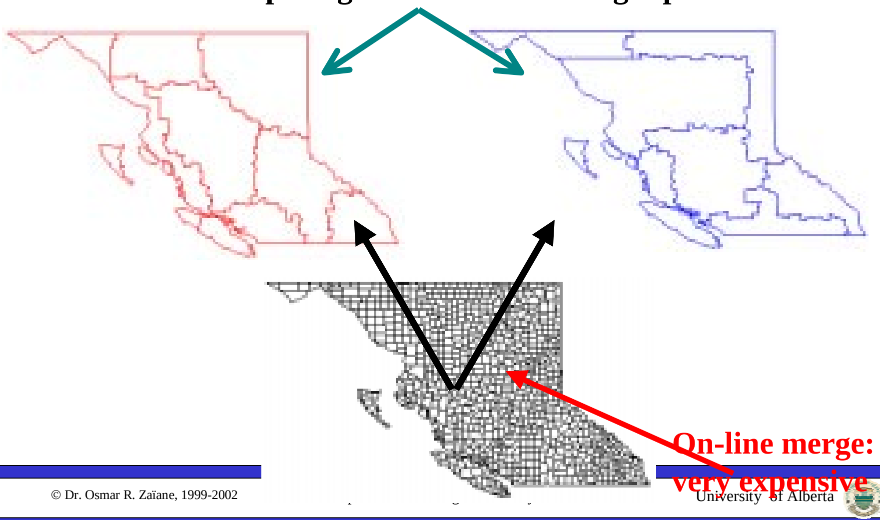
To achieve satisfactory results with reasonable performance.
1. perform nonspatial generalization to certain level,
2. perform high-level spatial merge,
3. repeat the two steps above until satisfactory results reached.

- Generalization on Multiple Thematic Maps:

To handle generalization on more than one thematic maps.
1. generalize each map based on the generalized properties,
2. apply spatial merge on the overlap of the maps to find the regions with generalized properties.

Spatial Merge: Pre- vs On-line Computation

Precomputing all: too much storage space



Result of a roll-up operation

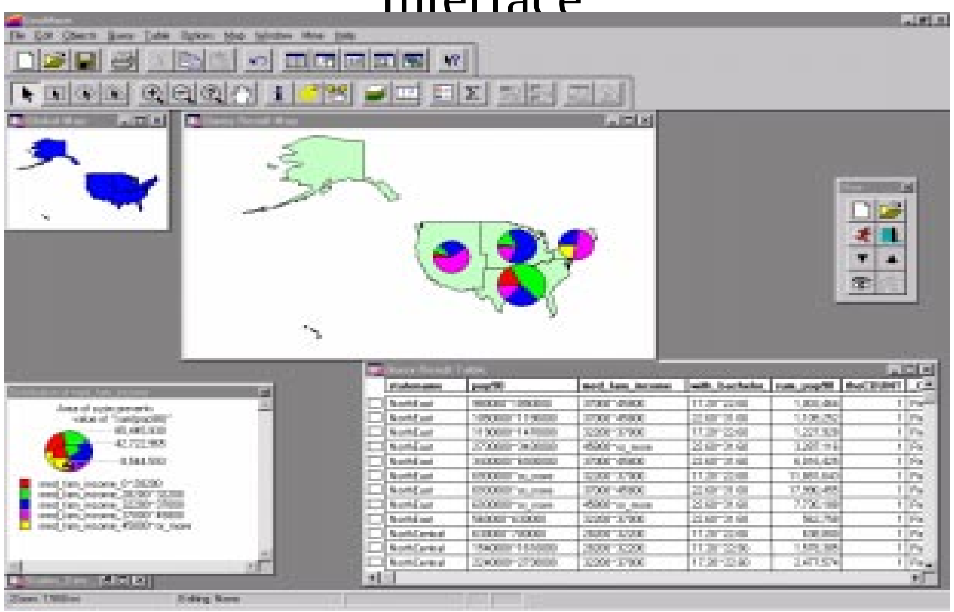
Time	Temperature	Precipitation	Region_map
January	below –20	dry	{AK04, AK07, … VS67}
January	below –20	fair	{AG05, AG10, … TP90}
…	…	…	…

Spatial measurements

- Can we merge {AK04, AK07, …VS67} into a single spatial object?

  Only the subsets that are neighbors on the map can be merged (transitivity applies).

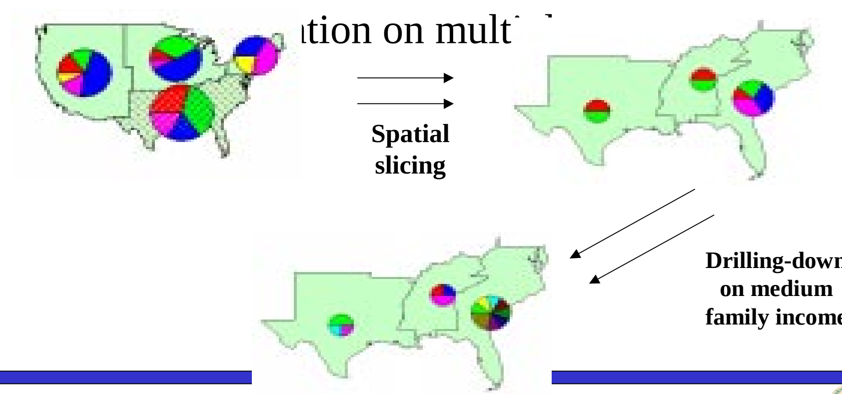
GeoMiner: Graphical User Interface



Spatial OLAP (Characterization)

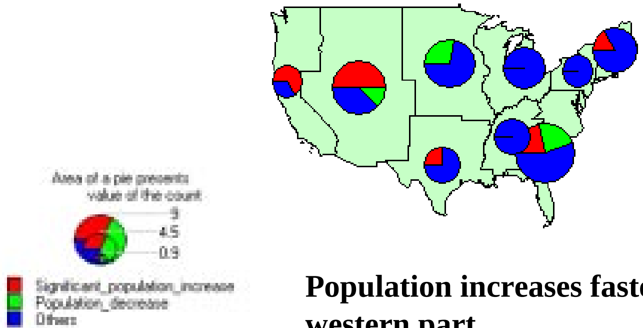
- Viewing data from different angles

tion on mult



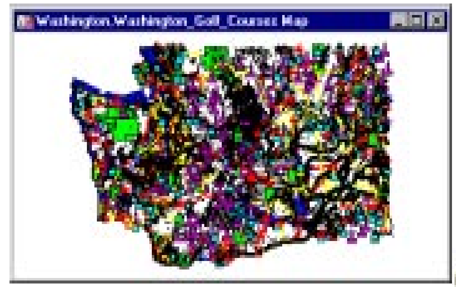
Spatial OLAP (Comparison)

- Comparing different classes of data

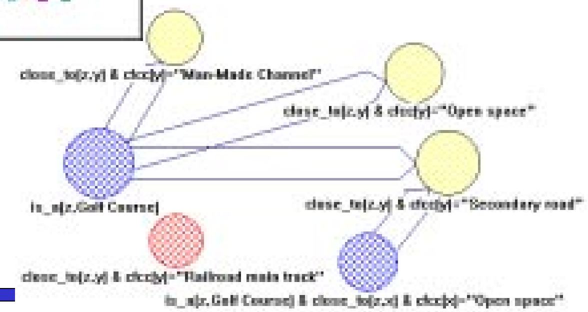


Population increases faster in the western part.
Drill down, and look at different dimensions to get explanation!!

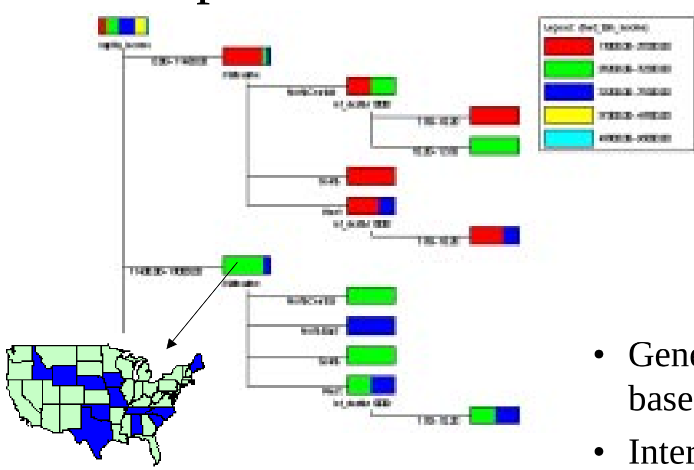
Spatial Association



FIND SPATIAL ASSOCIATION RULE
DESCRIBING "Golf Course"
FROM Washington_Golf_courses, Washington
WHERE CLOSE_TO(Washington_Golf_courses.Obj, Washington.Obj, "3 km")
AND Washington.CFCC <> "D81"
IN RELEVANCE TO Washington_Golf_courses.Obj, Washington.Obj, CFCC
SET SUPPORT THRESHOLD 0.5



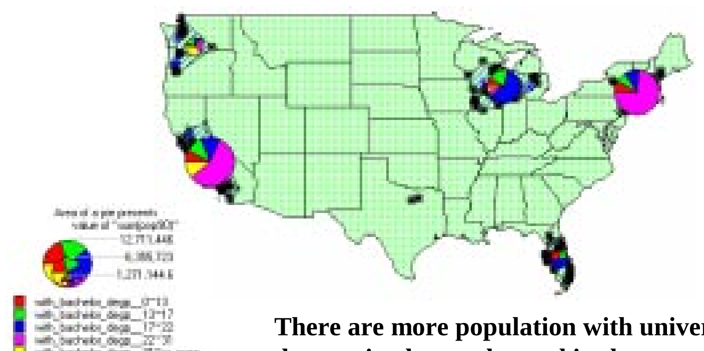
Spatial Classification



- Generalization-based induction
- Interactive classification

Spatial Clustering

- How can we cluster points?
- What are the distinct features of the clusters?



There are more population with university degrees in clusters located in the west, probably because of the distribution of high tech industry

Future Research

- Foundation of spatial data warehousing and data mining.
- Implementation methods:
  - Efficient construction of spatial data cubes.
  - A set of well-tuned spatial data mining operators.
  - Spatial data and knowledge visualization tools.
  - Integration of multiple mining tasks with OLAP functions.
- New spatial indexing techniques for spatial data warehousing and spatial mining.
- New spatial data mining methodologies: Statistical tools, neural nets, and ad-hoc query-based mining, etc.
- Mining spatiotemporal data, raster data, and integration with existing spatial analysis techniques.

References

- R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. SIGMOD'98, 94-105, Seattle, Washington, June 1998.
- M. Egenhofer. Spatial SQL : A query and presentation language. IEEE Trans. Knowledge and Data Engineering, 6:86-95, 1994.
- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. KDD'96, 226-231, Portland, Oregon, August 1996.
- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. Density-connected sets and their application for trend detection in spatial databases. KDD'97, 10-15, Newport Beach, California, August 1997.
- M. Ester, H.-P. Kriegel, and X. Xu. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. SSD'95, 67-82, Portland, Maine, August 1995.
- S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. SIGMOD'98, 73-84, Seattle, Washington, June 1998.
- R. H. Gueting. An introduction to spatial database systems. The VLDB Journal, 3:357-400, 1994.
- J. Han, K. Koperski, and N. Stefanovic. GeoMiner: A system prototype for spatial data mining. SIGMOD'97, 553-556, Tucson, Arizona, May 1997.
- J. Han, N. Stefanovic, and K. Koperski. Selective materialization: An efficient method for spatial data cube construction. PAKDD'98, Melbourne, Australia, April 1998.

References (cont.)

- E. Knorr and R. Ng. Finding aggregate proximity relationships and commonalities in spatial data mining. IEEE Trans. Knowledge and Data Engineering, 8:884-897, Dec. 1996.
- E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. VLDB'98.
- K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. SSD'95, 47-66, Portland, Maine, Aug. 1995.
- K. Koperski, J. Han, and J. Adhikary. Mining knowledge in geographic data. In Comm. ACM, 1998.
- R. Laurini and D. Thompson. Fundamentals of Spatial Information Systems. Academic, 1992.
- W. Lu, J. Han, and B. C. Ooi. Knowledge discovery in large spatial databases. In Proc. Far East Workshop Geographic Information Systems, 275-289, Singapore, June 1993.
- R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. VLDB'94, 144-155, Santiago, Chile, September 1994.
- T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. SIGMOD'96, 103-114, Montreal, Canada, June 1996.
- X. Zhou, D. Truffet, and J. Han. Efficient polygon amalgamation methods for spatial olap and spatial data mining. SSD'99, Hong Kong, Aug. 1999.