Geographic Data Mining

Studies represent research papers

By: Mohammed Alsolami
Adviser: Dr. Yang

Data Mining - Summer I - 2012 - WMU CS
Outline

- Introduction
- Why Geographic Knowledge Discovery?
  - Why Geographic Knowledge Discovery?
  - Where we need it?
- Geographic Knowledge Discovery GKD
  - Geographic Information Science GIS
  - Geographic Research
  - Geographic applications
- Geographic Data Warehousing
  - Spatial Data Warehousing
  - Differences between SDW & DW
- Types of Spatial Data Mining
  A) - (SDM) Spatial Classification
  B) - Spatial association
  C) - Discovering Co-location Patterns from Spatial
  D) - Spatial Classification and prediction
  E) - Spatial Clustering
  F) - Spatial outlier analysis
  G) - Geovisualization
- Discovering Co-location Patterns from Spatial
- Georeferenced multimedia
- Parallel geographic data mining
- Developing Structure of GDM
- Reference and books
- Conclusion
- Why Geographic Knowledge Discovery?

GKD is based on a belief that there is novel and useful geographic knowledge hidden in the unprecedented amounts of data.

Large databases contain interesting patterns: non-random properties and relationships that are:
- Valid data (general enough to apply to new data)
- Novel data (non-trivial and unexpected)
- Useful data (leads to effective action: decision making or investigation)
- Ultimately understandable data (simple, and interpretable by humans)
Outline

- Introduction
- Why Geographic Knowledge Discovery?
- Where we need it?
- Geographic Knowledge Discovery GKD
  - Geographic Information Science GIS
  - Geographic Research
  - Geographic applications
- Geographic Data Warehousing
  - Spatial Data Warehousing
  - Differences between SDW & DW
- Types of Spatial Data Mining
  A) (SDM) Spatial Classification
  B) Spatial association
- Discovering Co-location Patterns from Spatial
  C) Spatial Classification and prediction
  D) Spatial Clustering
  E) Spatial outlier analysis
  F) Geovisualization
- Resource Geographic Knowledge
- Important researches in GDM
  - Georeferenced multimedia
  - Parallel geographic data mining
  - Developing Structure of GDM
- References and books
- Conclusion
Where we need GKD

Meteorology

Climate
Where we need GKD

Location

Access
Where we need GKD

Navigation  Visualisation  Mapping

Mapping: that complies with rules of colors, symbols, generalisation, and so on
Where we need GKD

Online and large geographic databases

Data Mining methods are increasingly important for making sense of geographic data
Where we need GKD

Augmented geographic reality, Street View
Outline

- Introduction
- Why Geographic Knowledge Discovery?
- Where we need it?
- Geographic Knowledge Discovery GKD
  - Geographic Information Science GIS
  - Geographic Research
  - Geographic applications
- Geographic Data Warehousing
  - Spatial Data Warehousing
  - Differences between SDW & DW
- Types of Spatial Data Mining
  A) - (SDM) Spatial Classification
  B) - Spatial association
  C) - Discovering Co-location Patterns from Spatial
  D) - Spatial Classification and prediction
  E) - Spatial outlier analysis
  F) - Geovisualization
- Important researches in GDM
  - Georeferenced multimedia
  - Parallel geographic data mining
  - Developing Structure of GDM
- References and books
- Conclusion
Data mining is only one step of the knowledge discovery from databases (KDD) process. Data mining involves the application of techniques for distilling data into information or facts implied by the data. KDD is the higher level process of obtaining facts through data mining and distilling this information into knowledge or ideas and beliefs about the mini world described by the data. This generally requires a human-level intelligence to guide the process and interpret the results based on pre-existing knowledge (Miller and Han 2001).
Data mining is only one step of the knowledge discovery from databases (KDD) process. Data mining involves the application of techniques for distilling data into information or facts implied by the data. KDD is the higher level process of obtaining facts through data mining and distilling this information into knowledge or ideas and beliefs about the mini world described by the data. This generally requires a human-level intelligence to guide the process and interpret the results based on pre-existing knowledge (Miller and Han 2001).

The geographic data mining is the critical interface to capture, store, manipulate, analyze, manage, and present all types of geographical data by machines and the semantic knowledge required by humans for reasoning about the real world (Gahegan et al 2001).
Geographic knowledge discovery (GKD) is the process of extracting information and knowledge from massive georeferenced databases. (Shekhar et al. 2003).

Why GKD has always huge data:
- Because of the nature of geographic space, the complexity of spatial objects and relationships between objects.
- Also, because of the heterogeneous and sometimes ill-structured nature of georeferenced data, and the nature of geographic knowledge (SMB'05).
Outline

- Introduction
- Why Geographic Knowledge Discovery?
- Where we need it?
- Geographic Knowledge Discovery GKD
  - Geographic Information Science GIS
  - Geographic Research
  - Geographic applications
- Geographic Data Warehousing
  - Spatial Data Warehousing
  - Differences between SDW & DW
- Types of Spatial Data Mining
  A) - (SDM) Spatial Classification
  B) - Spatial association
- Discovering Co-location Patterns from Spatial
  C) - Spatial Classification and prediction
  D) - Spatial Clustering
  E) - Spatial outlier analysis
  F) - Geovisualization
- Resource Geographic Knowledge
- Important researches in GDM
  - Georeferenced multimedia
  - Parallel geographic data mining
  - Developing Structure of GDM
- References and books
- Conclusion
Geographic Information Systems (GIS)

A geographic information system is a system designed to capture, store, manipulate, analyze, manage, and present all types of geographical data. (Sun 2010)

It is important to know that exploring geo-spaces or levels for representing geographic data, that is a form of data preprocessing that could substantially enhance the GKD process. (Miller and Wentz 2003).

*GKD: Geographic knowledge discovery
- Introduction
- Why Geographic Knowledge Discovery?
- Where we need it?
- Geographic Knowledge Discovery GKD
  - Geographic Information Science GIS
- Geographic Research
- Geographic applications
- Geographic Data Warehousing
  - Spatial Data Warehousing
  - Differences between SDW & DW
- Types of Spatial Data Mining
  A) - (SDM) Spatial Classification
  B) - Spatial association
- Discovering Co-location Patterns
- Discovering Co-location Patterns from Spatial
  C)- Spatial Classification and prediction
  D)- Spatial Clustering
  E)- Spatial outlier analysis
  F)- Geovisualization
- Resource Geographic Knowledge
- Important researches in GDM
  - Georeferenced multimedia
  - Parallel geographic data mining
  - Developing Structure of GDM
- References and books
- Conclusion
Mapping/searching is the main important spatial analysis of the geographic data mining.

For example: in Google Map Researches
- Moving from displaying an accurate and comprehensive images that supercomputers have managed, built, and allowed to access the images as 3D to ...
The result:
- to create images by storing and analysing each pixel on the model to generate textured 3D mesh for each building or item on the map.
Geographic Research

The result:
- A picture can be used to generate 3D "overfly" pictures, by analysing or mining each pixel and color of the original picture.
Outline

- Introduction
- Why Geographic Knowledge Discovery?
- Where we need it?
- Geographic Knowledge Discovery GKD
  - Geographic Information Science GIS
  - Geographic Research
  - Geographic applications
- Geographic Data Warehousing
  - Spatial Data Warehousing
  - Differences between SDW & DW
- Types of Spatial Data Mining
  A) - (SDM) Spatial Classification
  B) - Spatial association
  - Discovering Co-location Patterns from Spatial
  C)- Spatial Classification and prediction
  D)- Spatial Clustering
  E)- Spatial outlier analysis
  F)- Geovisualization
  - Resource Geographic Knowledge
  Important researches in GDM
    - Georeferenced multimedia
    - Parallel geographic data mining
    - Developing Structure of GDM
- References and books
- Conclusion
GDM - applications

- **MI**: Map interpretation (for Interpretation of Terrain Features on a Topographic Map)
- **RSI**: Remote sensing interpretation (for interpreting and measuring of objects on aerial photographs)
- **EMS**: Environmental mapping (soil type, water borders, oil level)
- **ESTPC**: Extracting spatiotemporal patterns for cyclones and crimes
- **SPIS**: Spatial interaction (movement/flow of people, capital, goods)
- Introduction
- Why Geographic Knowledge Discovery?
- Where we need it?
- Geographic Knowledge Discovery GKD
  - Geographic Information Science GIS
  - Geographic Research
  - Geographic applications
- Geographic Data Warehousing
  - Spatial Data Warehousing
  - Differences between SDW & DW
- Types of Spatial Data Mining
  A) - (SDM) Spatial Classification
  B) - Spatial association
- Discovering Co-location Patterns from Spatial
C) - Spatial Classification and prediction
D) - Spatial Clustering
E) - Spatial outlier analysis
F) - Geovisualization
Resource Geographic Knowledge
- Important researches in GDM
  - Georeferenced multimedia
  - Parallel geographic data mining
  - Developing Structure of GDM
- References and books
- Conclusion
**Geographic Data Warehousing (GDW)**

**GDW** is a database used for reporting and analysis.

**Geographic data warehouse** uses staging, integration, and access layers to house its key functions.

A **spatial data cube** is the GDW analog to the data cube tool for computing and storing all possible aggregations of some measure in OLAP, Online Analytical Processing.

The spatial data cube must include standard attribute summaries as well as pointers to spatial objects at varying levels of aggregation.

**Aggregating spatial objects** is nontrivial and often requires background domain knowledge in the form of a geographic concept hierarchy. Strategies for selectively pre-computing measures in the spatial data cube include **none, pre-computing rough approximations, and selective pre-computation** (Han, Stefanovic and Koperski 2000).
Outline

- Introduction
- Why Geographic Knowledge Discovery?
- Where we need it?
- Geographic Knowledge Discovery GKD
  - Geographic Information Science GIS
  - Geographic Research
  - Geographic applications
- Geographic Data Warehousing
  - Spatial Data Warehousing
  - Differences between SDW & DW
- Types of Spatial Data Mining
  A) - (SDM) Spatial Classification
  B) - Spatial association
  C) - Discovering Co-location Patterns from Spatial
  D) - Spatial Classification and prediction
  E) - Spatial outlier analysis
  F) - Geovisualization
- Developing Structure of GDM
- Resource Geographic Knowledge
- Important researches in GDM
  - Georeferenced multimedia
  - Parallel geographic data mining
  - References and books
- Conclusion
Spatial data warehouses (SDWs) are data warehouses that also include both spatial and aspatial data. SDWs often include geo-referenced data, (such as medical imagery) that can be archived using SDW techniques.

Examples of geographic SDWs: the U.S. Census database (Sequoia 2000) and archives from transportation operations centers (Shekhar and Chawla 2003).
Why SDW

1- The nature of geographic space => various distances and spatial relationships
   - Spatial objects are embedded in multidimensional levels
     - implicit distances    • directional relationships    • topological relationships
   - Many physical and human geographic processes exhibit spatial properties
     (e.g. migration, disease propagation, travel times in congested urban areas)

2- The complexity of spatial objects and relationships
   - Spatial dependency
     « Everything is related to everything else, but near things are more related than distant things » ((Tobler1979))
   - Spatial location
   - Size and Shape of geographic units have non trivial influences on geographic processes
     and their measurement
Outline

- Introduction
- Why Geographic Knowledge Discovery?
- Where we need it?
- Geographic Knowledge Discovery GKD
  - Geographic Information Science GIS
  - Geographic Research
  - Geographic applications
- Geographic Data Warehousing
  - Spatial Data Warehousing
- Differences between SDW & DW
- Types of Spatial Data Mining
  A) - (SDM) Spatial Classification
  B) - Spatial association
- Discovering Co-location Patterns from Spatial
C) - Spatial Classification and prediction
D) - Spatial Clustering
E) - Spatial outlier analysis
F) - Geovisualization
- Developing Structure of GDM
- Important researches in GDM
  - Georeferenced multimedia
  - Parallel geographic data mining
  - References and books
- Conclusion
The main functional difference between SDW and standard data warehouses is the capabilities for visualization and spatial aggregation. That means: how to manipulate pictures

The differences:
* Conventional OLAP methods such as the data cube generate summary crosstabs in tables.
** Spatial data requires capabilities for data summaries in cartographic form.
* Conventional OLAP tools have clear standards for aggregation and cross-tabulation, namely, the one-dimensional attributes associated with each data object.
** Spatial aggregation is more complex, and standards for aggregation operators on geometric data types have not yet emerged (Shekhar and Chawla 2003).
Outline

- Introduction
- Why Geographic Knowledge Discovery?
- Where we need it?
- Geographic Knowledge Discovery GKD
  - Geographic Information Science GIS
  - Geographic Research
  - Geographic applications
- Geographic Data Warehousing
  - Spatial Data Warehousing
  - Differences between SDW & DW
- Types of Spatial Data Mining
  A) - (SDM) Spatial Classification
  B) - Spatial association
  C) - Discovering Co-location Patterns from Spatial
  D) - Spatial Classification and prediction
  E) - Spatial Clustering
  F) - Spatial outlier analysis
  G) - Geovisualization
  Resource Geographic Knowledge
  Important researches in GDM
  - Georeferenced multimedia
  - Parallel geographic data mining
  - Developing Structure of GDM
- References and books
- Conclusion
Spatial data mining (SDM)

One of the most common methods in GDM is SDM.

SDM is the application of data mining methods to spatial data.

The end objective of spatial data mining is to find patterns in data with respect to geography.

Pattern types such as **classes**, **associations**, **rules**, **clusters**, **outliers** and **trends** all have spatial expressions since these patterns can be conditioned by the morphology as well as spatial relationships among these objects.
A)- Spatial classification

Spatial Classification is to learn the concept associated with each class on the basis of the interaction of two or more spatially-referenced objects or space-dependent attributes, according to a spacing or set of arrangements. Koperski et al (1998) use spatial buffers to classify objects based on attribute similarity and distance-based proximity. (Each Near From Each) Ester et al. (1997-2001) generalize this approach through a spatial classification learning algorithm that considers spatial relationships defined as path relationships among objects in a defined neighborhood of a target object. These paths are highly general and can be defined using any spatial relationship.
Spatial association is a relationship between two or more archaeological items (objects or structural elements) because of their physical proximity and/or location.

Spatial association rules contain spatial predicates in their precedent or antecedent.

Koperski and Han (1995) pioneered this concept, providing detailed descriptions of their formal properties as well as a top-down tree search technique that exploits background knowledge in the form of a geographic concept hierarchy.

For example, a specific type of association rule is a co-location pattern.

Huang, Shekhar and Xiong (2002) develop a multi-resolution filtering algorithm for discovering co-location patterns in spatial data.
Discovering Co-location Patterns from Spatial
C)- Spatial classification and prediction

Malebra et al (2001) use inductive learning algorithms to extract information from general purpose topographic maps such as the type produced by national surveying and cartographic organizations. A search heuristic builds logical predicates based on the spatial objects, background knowledge, defined higher-level concepts and a performance criterion.

Qi and Zhu (2003) apply a decision tree induction algorithm to extracting knowledge about complex soil-landscape processes. Their system combines background knowledge in the form of a soil survey map with other environmental data to extract the expert’s judgments underlying the subjective map.

Gopal, Liu and Woodcock (2001) use a type of artificial neural known as adaptive resonance theory networks to extract knowledge from a remotely sensed imagery. They also illustrate the use of visualization to support interpretation and insights into neural network performance.
D)- Spatial clustering

Spatial clustering algorithms exploit spatial relationships among data objects in determining inherent groupings of the input data. Since finding the optimal set of k clusters is intractable (where k is some integer much smaller than the cardinality of the database)

Many of these can be adapted to or are specially tailored for spatial data (Han, Tung and Kamber 2001).
E)- Spatial outlier analysis

Methods which integrate the impact of spatial properties to the outlierness measurement.

Shekhar, Lu and Zhang (2003) define a spatial outlier as a spatially-referenced object whose non-spatial attributes appear inconsistent with other objects within some spatial neighborhood by size and shape.

Ng (2001) uses distance-based measures to detect unusual paths in two-dimensional space traced by individuals through a monitored environment.

These measures allow the identification of unusual trajectories based on size and shape.
Geographic visualization (GVis), or the integration of scientific visualization with traditional cartography, is highly complementary to the GKD process and can be exploited at all stages, including data pre-processing, data mining and knowledge construction (Gahegan et al 2001).

The problem is how to preserve the richness of this information space when restricted to the low dimensional information spaces that can be easily related to geographic space by the user (these being two or three spatial dimensions, and possibly time through animation).

For example, dense and complex symbols and colors within low dimensionality spaces can create visual interaction effects that are poorly understood and can confound knowledge discovery and communication (Gahegan 1999).
Outline

- Introduction
- Why Geographic Knowledge Discovery?
- Where we need it?
- Geographic Knowledge Discovery GKD
  - Geographic Information Science GIS
  - Geographic Research
  - Geographic applications
- Geographic Data Warehousing
  - Spatial Data Warehousing
  - Differences between SDW & DW
- Types of Spatial Data Mining
  A) - (SDM) Spatial Classification
  B) - Spatial association
  - Discovering Co-location Patterns from Spatial
  C) - Spatial Classification and prediction
  D) - Spatial Clustering
  E) - Spatial outlier analysis
  F) - Geovisualization
- Resource Geographic Knowledge
- Important researches in GDM
  - Georeferenced multimedia
  - Parallel geographic data mining
  - Developing Structure of GDM
- References and books
- Conclusion
Resources geographic knowledge

There is a rich source of existing geographic models, theories, laws and knowledge that can be codified to serve as background knowledge for the GKD process.

- **Data market centers**

There are market centers that have much information that can be applied geographic data at scales from local to global.

- **Physical geography systems**

That contains sophisticated geographic concept hierarchies based on geomorphology, river networks, biotic regimes and so on.
Outline

- Introduction
- Why Geographic Knowledge Discovery?
- Where we need it?
- Geographic Knowledge Discovery GKD
  - Geographic Information Science GIS
  - Geographic Research
  - Geographic applications
- Geographic Data Warehousing
  - Spatial Data Warehousing
  - Differences between SDW & DW
- Types of Spatial Data Mining
  A) - (SDM) Spatial Classification
  B) - Spatial association
- Discovering Co-location Patterns from Spatial
C)- Spatial Classification and prediction
D)- Spatial Clustering
E)- Spatial outlier analysis
F)- Geovisualization
Resource Geographic Knowledge

Important researches in GDM
- Georeferenced multimedia
- Parallel geographic data mining
- Developing Structure of GDM
- References and books
- Conclusion
Georeferenced multimedia

**Geo-multimedia database system** stores and manages large collections of georeferenced multimedia objects such as audio, image, and video, including metadata about where and when the media was collected, or the locations and times described by the media with most basic features and structures (Han and Kamber 2001).
Parallel geographic data mining

Parallel algorithms and distributed infrastructures for geographic data mining: Parallel processing platforms such as grid computing environments can be exploited in GKD.

- Spatial data mining techniques can sometimes be decomposed into parallel tasks.
- Even if task parallelism is impossible, geo-referenced data can often be divided into spatial subsets for parallel processing (Healy et al. 1998; Wang and Armstrong 2003).
Georeference should be unambiguous, unique and shared (agreed)

- **Metrics** (measures of distance from fixed places)
  - Latitude, Longitude (distance from the Equator or from the Greenwich Meridian)
  - XY map coordinates
- **Ordinal**
  - street addresses in most parts of the world order houses along streets
- **Nominal** (placenames)
  - administrative units, municipalities, regions, cities,...
Structure of Geographic Information

**Control points:** points which provide the connection to the national geodetic coordinate system

**Land cover:** buildings, roads, hydrology

**Single objects:** walls, masts, bridges, etc.

**Heights:** digital terrain model

**Local names:** place names, locality names

**Ownership:** land parcels

**Pipelines:** oil and gas

**Territorial boundaries:** municipal, district

**Landslips:** ground movement

**Building addresses:** geographical locations (road or street name, house number, postal code, locality name)
Structure of Geographic Information

Geographical objects models

- **Vector**
  - Point $(x, y)$
  - Line $(x, y)$
  - Area/polygon $(x, y)$

- **Raster**

  - Satellite images,
  - Air photos,
  - Temperatures, ...
  - Land cover

Specific operations based on topology info

Specific operations: neighbourhood, …
Geographic Information Systems (GIS)

A Data warehouse + specific geographic operations

A set of geographic layers

Points
Lines
Polygons
Raster

Operations

Intra-layer and inter-layers (overlay) operations
(Map/Geo-Algebra)
6- J. P. Wilson and A. S. Fotheringham (eds.) Handbook of Geographic Information Science, in press. Geographic Data Mining and Knowledge Discovery

8- Smyth and R. Ulthurusamy (eds.) Advances in Knowledge Discovery and Data Mining, Cambridge, MA: MIT Press, 1-34.


Books

- Geographic Data Mining and Knowledge Discovery
  By Harvey Miller and Jiawei Han

- Knowledge Discovery in Spatial Data
  By Yee Leung
Questions?!  

Thank you for listening