Spatial Data Mining in the Era of Big Data

Dr. Jin Soung Yoo
Associate Professor
Department of Computer Science
Indiana University-Purdue University Fort Wayne
Outline

■ Introduction to Data Mining
■ Works in Data Mining
What is Data Mining

- A computer-assisted process of discovering interesting, previously unknown, implicit, potentially useful, and non-trivial patterns or knowledge from large databases
  - Non-trivial search
    - Large (e.g., exponential) search space of plausible hypothesis
  - Interesting
    - Useful in certain application domain
  - Unexpected
    - Patterns is not common knowledge
    - May provide a new understanding of world
With rapid advances in data collection and storage technology, the explosive growth of data from many data sources:

- Purchases at grocery stores, customer services from call centers
- Web logs from e-commerce Web sites
- Bank/Credit card transactions
- Mobile phone contents
- Social networks
- World Wide Web: online news, digital images, YouTube

Source: various web sites
Why Data Mining (Conti.)

- Competitive pressure
  - Today business environment requires critical data analysis
    - Market analysis: targeted marketing, cross-selling, market segments, etc.
    - Risk management: forecasting, customer retention
    - Fraud detection and detection of unusual behavior
  - Business questions
    - “Who are the most profitable customers?”
    - “What products can be cross-sold?”
    - “How change if a new local store is added?”
Example: Target’s Finding

Target has figured out whether you have a baby on the way long before you need to start buying diapers.

“Women on the baby registry were buying larger quantities of unscented lotion around the beginning of their second trimester.”

“When someone suddenly starts buying lots of scent-free soap and extra-big bags of cotton balls, in addition to hand sanitizers and washcloths, it signals they could be getting close to their delivery date.”

Source: http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/
Why Data Mining - Scientific Viewpoint

Many science areas are collecting data for their new important discovery
- Earth observations from satellites
- Climate measurement
- Sloan Digital Sky Survey (SDDS)
- Cancer/epidemic data (SEER)
- Microarrays generating gene expression data
- Scientific simulations
- Animal behavior observation

Source: various web sites
# Scale of Data

<table>
<thead>
<tr>
<th>Organization</th>
<th>Scale of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walmart</td>
<td>~ 1 million customer transactions / hr</td>
</tr>
<tr>
<td>Facebook</td>
<td>~ 50 billion photos</td>
</tr>
<tr>
<td>Yahoo</td>
<td>~48 GB Web log data/hr</td>
</tr>
<tr>
<td>Falcon Credit Card Fraud Detection System (FICO)</td>
<td>2.1 billion active accounts world-wide</td>
</tr>
<tr>
<td>Business data worldwide, across all companies</td>
<td>Doubles every 1.2 years</td>
</tr>
<tr>
<td>NASA satellites</td>
<td>~ 1.2 TB/day</td>
</tr>
<tr>
<td>Sloan Digital Sky Survey (SDSS)</td>
<td>~140 TB (200GB /night)</td>
</tr>
<tr>
<td>NCBI GenBank</td>
<td>~ 22 million genetic sequences</td>
</tr>
</tbody>
</table>

“The great strength of computers is that they can reliably manipulate vast amounts of data very quickly. Their great weakness is that they don’t have a clue as to what any of that data actually means”  (Cass, IEEE Spectrum, Jan 2004)

- There is often information “hidden” in the data that is not readily evident
- Much of the data is never analyzed at all.

source: http://shawnwhatley.com/
Data mining is the art of finding treasures in the sea of data when you don't know what you're looking for or what you might find.
Confluence of Multiple Disciplines

Data Mining

- Artificial Intelligence
- Database Technology
- Machine Learning
- Statistics
- Visualization
- Pattern Recognition
- Other Disciplines
- Algorithm
Data Mining for Many Other Disciplines
Outline

Introduction to Data Mining

Works in Data Mining

- Spatial Data Mining
- Spatial Association Mining in Cloud Computing Environment
- Temporal Data Mining
- Spatiotemporal Data Mining
- Biological Data Mining
- Educational Data Mining
Spatial Data – Everywhere

Evolution of Location aware devices, Mobile computing, Wireless network

- Military, Homeland security
- Business (Location-Based Services)
- Criminology
- Earth Science
- Environmental Science
- Transportation
- Public Health

*Pictures from various sources*
Spatial Data Mining

- The process of discovering interesting, useful, non-trivial (as “automatized” as possible) patterns from large spatial or spatiotemporal data.

- Spatial Pattern Families vs. Techniques
  - Prediction
  - Hot spots
  - Interaction

Classification  Prediction  Clustering  Outlier detection  Association, Colocation

* Pictures from various sources
Examples of Spatial Patterns

- **Spatial relationships** (location, region, frontier, neighborhood, obstruction, field, basin, communication, diffusion, propagation) are importantly considered for the pattern discovery.

- **Historic Examples**
  - 1855 Asiatic Cholera in London; A water pump identified as the source.
  - Fluoride and health gums near Colorado river (with originally insufficient amount of fluoride).

Examples of Spatial Patterns

- Modern Examples
  - Cancer clusters to locate hazardous environments
  - Nile virus spreading from north east USA to south and west
  - Crime hotspots for planning police patrol routes
  - Colocation of a business with another franchise (such as colocation of a Pizza Hut restaurant with a Blockbuster video store)
  - Best locations for opening new hospitals based on the population of patients who live in each neighborhood.
  - Spatial region-based personalization
  - Unusual warming of Pacific ocean (El Niño) effects weather in USA
Outline

- Introduction to Data Mining
- Works in Data Mining
  - Spatial Data Mining
    - Spatial Association Mining
  - Spatial Association Mining in Cloud Computing Environment
- Summary
Spatial Association, Co-location, Correlation

- **Spatial association mining** discovers interesting spatial relationships and correlations among spatial objects.
  - Spatial correlation (or, *neighborhood influence*) refers to the phenomenon of the location of a specific object in an area affecting some nonspatial attribute of the object.
  - For example, the value (nonspatial attribute) of a house at a given address (geocoded to give a spatial attribute) is largely determined by the value of other houses in the neighborhood.
Spatial Co-location

- A **co-location** represents the presence of two or more spatial objects at the same location or at significantly close distances from each other.

- **Co-location mining** finds all subsets of spatial events (features) which are frequently observed in nearby areas.

- In the case of including non-spatial information:
  - For example, sales at franchises of a specific pizza restaurant chain were higher at restaurants colocated with video stores than at restaurants not colocated with video stores.

Find patterns from the above sample dataset?

**Answer:** \{Parking, Restaurant, Movie, Restaurant\}
Co-location Examples

- Which spatial events are related to each other?
- Which spatial phenomena depend on other phenomenon?

<table>
<thead>
<tr>
<th>Domain</th>
<th>Example Features</th>
<th>Example Co-location Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epidemiology</td>
<td>Disease types, environmental events</td>
<td>{West Nile disease, stagnant water sources, dead birds, mosquitoes}</td>
</tr>
<tr>
<td>Location-based services</td>
<td>Service type requests</td>
<td>{tow truck, police, ambulance}</td>
</tr>
<tr>
<td>Business</td>
<td>Local store types</td>
<td>{Burger King, MacDonald’s}</td>
</tr>
<tr>
<td>Transportation</td>
<td>Delivery service tracks</td>
<td>{US Postal Service, UPS, newspaper delivery}</td>
</tr>
<tr>
<td>Military</td>
<td>Critical points, events</td>
<td>{weapons Caches, IED factories}</td>
</tr>
<tr>
<td>Economics</td>
<td>Industry types</td>
<td>{suppliers, producers, consultants}</td>
</tr>
<tr>
<td>Ecology</td>
<td>Species</td>
<td>{Nile crocodile, Egyptian plover}</td>
</tr>
<tr>
<td>Earth Science</td>
<td>Climate and disturbance events</td>
<td>{wild fire, hot, dry, lightning}</td>
</tr>
<tr>
<td>Weather</td>
<td>Fronts, precipitation</td>
<td>{cold front, warm front, snow fall}</td>
</tr>
</tbody>
</table>
Preliminary - Key Terms

- **A co-location** $X$: A subset of spatial event types

- **A neighborhood**: A clique in a graph of neighbor relation

- **An instance $I$ of co-location $X = \{e_1 \ldots e_k\}$**: $I = \{o_1 \ldots o_k\}$
  - $o_j$ is object of $e_j (\forall j \in 1, \ldots, k)$
  - $I$ is a neighborhood.

* $A1 :$ event type A, object id 1
* — : neighbor relationship

Distance(A1, B1) < neighbor distance threshold
Preliminary - Interest Measures

- **Participation Index** $PI$ of co-location $X = \{e_1 \cdots e_k\}$

\[
PI(X) = \min_{e_i \in X} \{PR(e_i, X)\}
\]

* Strength of prevalence of co-location

- **Participation Ratio** $PR(e_i, X)$

\[
\frac{\text{# of objects of } e_i \text{ in instances of } X}{\text{# of objects of } e_i}
\]

* Strength of each event type in a co-location

If $PI > \text{prev\_threshold}$, $\{A, B, C\}$ is a frequent co-location.
Related Work: General Associations

<table>
<thead>
<tr>
<th>Trans</th>
<th>Items Bought</th>
<th>VS.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{socks, milk, beef, egg, …}</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>{ice-cream, muffin, …}</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>{pillow, toothbrush, …}</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>{juice, egg, chicken, battery, …}</td>
<td></td>
</tr>
</tbody>
</table>

* transaction: a set of items

**E.g., Diaper → Beer (0.5, 1)**

**E.g., Police → Tow, Ambulance (0.5, 0.8)**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Association Pattern</th>
<th>Co-location Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underlying Space</td>
<td><strong>Discrete Sets</strong></td>
<td><strong>Continuous Space</strong></td>
</tr>
<tr>
<td>Item Types</td>
<td>Product Types</td>
<td>Spatial Events(Features)</td>
</tr>
<tr>
<td>Item Collections</td>
<td><strong>Transactions T</strong></td>
<td><strong>Neighborhoods of L</strong></td>
</tr>
<tr>
<td>Prevalence (A, B)</td>
<td>Support: ( P(A \cap B \in T_i) )</td>
<td>Spatial Prevalence Measure</td>
</tr>
</tbody>
</table>
Related Work: Statistical Approach

- Ripley’s Cross K-Function [Cressie]
  - $K_{ij}(d) = \lambda_j^{-1}E$ [number of type $j$ feature within distance $d$ of a randomly chosen type $i$ feature]

- Limitations
  - Not proper for analysis of features of size $\geq 3$ e.g., triple features (K,T,R)
  - Not efficient in computation

Spatially correlated $K_{SR}(d)$
Spatial complete randomness $K_{MW}(d)$
8 Spatial Analysis and Mining

This chapter describes the Oracle Spatial support for spatial analysis and mining in Oracle Data Mining (ODM) applications.

8.4 Colocation Mining

Colocation is the presence of two or more spatial objects at the same location or at significantly close distances from each other. Colocation patterns can indicate interesting associations among spatial data objects with respect to their nonspatial attributes. For example, a data mining application could discover that sales at franchises of a specific pizza restaurant chain were higher at restaurants collocated with video stores than at restaurants not collocated with video stores.

Two types of colocation mining are supported:

- Colocation of items in a data mining table. Given a data layer, this approach identifies the colocation of multiple features. For example, predator and prey species could be collocated in animal habitats, and high-sales pizza restaurants could be collocated with high-sales video stores. You can use a reference-feature approach (using one feature as a reference and the other features as thematic attributes, and materializing all neighbors for the reference feature) or a buffer-based approach (materializing all items that are within all windows of a specified size).

- Colocation with thematic layers. Given several data layers, this approach identifies colocation across the layers. For example, given a lakes layer and a vegetation layer, lakes could be collocated with areas of high vegetation. You materialize the data, add categorical and numerical spatial relationships to the data mining table, and apply the ODM Association-Rule mechanisms.

The following functions and procedures, documented in Chapter 21, perform operations related to colocation mining:

- SDO_SAM.COLOCATED_REFERENCE FEATURES
- SDO_SAM.BIN_GEOMETRY
Challenges

- **No explicit transaction concept in spatial data**
  - Non-trivial to reuse association mining algorithms

- **Continuous neighbor relationship**
  - Especially, clique relations

- **Very large search space**
  - Given $n$ features, there are around $2^n$ possible candidate feature sets

- **Other workload**
  - Density, neighbor distance, prevalence threshold, etc.

- **Inherently too demanding of both processing time and memory requirements**
Why Not Use Modern Computational Framework

- Standard architecture emerging:
  - Cluster of commodity Linux nodes
  - Gigabit Ethernet interconnect

- Popular modern computational framework:
  - Cloud computing (distributed computing over a network)
  - Hadoop (MapReduce), a software framework for large-scale processing of data on clusters of commodity hardware.

- How to organize the mining computations on this architecture?
Problem Formulation

- **Given**
  - A spatial event dataset, <id, event type, location>
  - A spatial neighbor relationship (e.g., distance threshold)
  - A \( \text{prev\_threshold} \)

- **Find**
  Co-location patterns with participation index > \( \text{prev\_threshold} \)

- **Objectives**
  Develop a parallel/distributed co-location mining algorithm for spatial association analysis in cloud computing environment.
Outline

- Introduction to Data Mining
- My Works in Data Mining
  - Spatial Data Mining
  - Spatial Association Mining in Cloud Computing Environment
    - Background: Hadoop and MapReduce
    - Proposed Approach
    - Experimental Evaluation
  - Summary
Background: Hadoop

- Execution framework for running applications on large clusters of commodity hardware

  Includes
  - Storage: HDFS
  - Processing: MapReduce
    - Support the Map/Reduce programming model

- Characteristics
  - Economy: use cluster of commodity computers
  - Easy to use
    - Users: no need to deal with the complexity of distributed computing
  - Reliable: can handle node failures automatically
Background: MapReduce

- The heart of Hadoop
- A programming model for processing large data sets with a parallel, distributed algorithm on a cluster.
- The term MapReduce actually refers to two separate and distinct tasks that Hadoop programs perform.
  - The first is the map job, which takes a set of input data and converts it into another set of data.
  - The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples.

Figure source: [http://blog.sqlauthority.com/2013/10/09/big-data-buzz-words-what-is-mapreduce-day-7-of-21/](http://blog.sqlauthority.com/2013/10/09/big-data-buzz-words-what-is-mapreduce-day-7-of-21/)
Example: Word Count on MapReduce

- A **map function** process a **key/value pair** to generate a set of intermediate **key/value pairs**

  \[
  \text{map(key=null, val=record):}
  \]

  For each word \( w \) in contents, emit \((w, "1")\)

- The shuffling step merges all intermediate values associated with the same intermediate key and feed the key/values pairs to a **reduce function**. The reducer generates a set of result **key/value pairs**.

  \[
  \text{reduce(key=word w, values=[1, 1, \ldots, 1]):}
  \]

  Sum all “1”s in values list

  Emit result \((\text{word, sum})\)

---

**Figure source:** http://www.alex-hanna.com/tworkshops/lesson-5-hadoop-and-mapreduce/
MapReduce Model is Widely Applicable

- Example uses:
  - distributed grep
  - term-vector / host
  - document clustering
  - distributed sort
  - web access log stats
  - machine learning
  - web link-graph reversal
  - inverted index construction
  - statistical machine translation

Outline

- Introduction to Data Mining
- My Works in Data Mining
  - Spatial Data Mining
  - Spatial Association Mining in Cloud Computing Environment
    - Background: Hadoop and MapReduce
  - Proposed Approach
    - Experimental Evaluation
- Summary
Co-location Mining on MapReduce

**INPUT**
A spatial dataset

**Job 1: Neighboring object search**
- Map: Assign a grid no to data object
- Reduce: Search all neighboring pairs

**Job 2: Spatial neighborhood materialization**
- Map: Check neighbor constraints
- Reduce: Generate the neighborhood record

**Job 3: Event object count**
- Map: event=type of o_i
- Reduce: Count event objects

**Job 4: Co-located event set search**
- Map: Drop unprevalent event objects in N
- Reduce: Compute prevalence measures

**OUTPUT**
<co-located event set, prevalence>

**Preprocess**
- Neighborhood records
- Neighbor pairs

**Co-location pattern mining**
- k=1
- k=2
- k=k+1
Job1: Neighbor object search

**INPUT**
A spatial dataset

A spatial dataset:

<table>
<thead>
<tr>
<th>ID</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-85.1113430</td>
<td>41.1016631</td>
</tr>
<tr>
<td>2</td>
<td>-85.1085207</td>
<td>41.1026886</td>
</tr>
<tr>
<td>3</td>
<td>-85.1036761</td>
<td>41.1018515</td>
</tr>
<tr>
<td>4</td>
<td>-85.1084791</td>
<td>41.1017347</td>
</tr>
</tbody>
</table>

**Map**
Assign a grid no to data object

**Reduce**
Search all neighboring pairs

**Shuffle and Sort**

Overlapping space partition

Plane-sweep algorithm
O(nlogn)
Job 2: Neighbor object search

Job 2: Spatial neighborhood materialization

Map
Check the neighbor constraint
o_i's type < o_j's type

Reduce
Generate the neighborhood record

CONCEPTUAL FIGURE for Job1 and Job2

Disjoint neighborhood partition

Pivot object | Sub-star
---|---
A1 | B1 - C1
A2 | B4 - C2
A3 | B3 - C1
A4 | A4 - C1
B3 | C1 - B3 - C3
B4 | B4 - C2

Conditional neighborhood transactions

Job 1 Output

A1
B3
A3
C3
B5
A4
B1
B2
B3
C1
C2
Job 3: Event Object Count (Optional)

**Map**
\[ \text{event} = \text{type of } o_i \]

**Reduce**
\[ \langle \text{event type, count} \rangle \]

Job 2
Output

Neighborhood records
\[ \langle o_i, N(o_i) \rangle \]

(A1, (A1, B1, C1))
(A2, (A2, B4, C2))
(A3, (A3, B3, C1))
(A4, (A4, C1))

(B1, (B1))
(B2, (B2))
(B3, (B3, C1, C3))
(B4, (B4, C2))
(B5, (B5))

(C1, (C1))
(C2, (C2))
(C3, (C3))
Job 4: Prevalent Co-located Event Set Search

Map
- Drop unprevalent event objects in N
- Collect size k candidate instance
- Filter true instance

Reduce
- Compute prevalence measures
- Find prevalent co-located event sets

Output
- \(\langle\text{co-located event set}, \text{prevalence}\rangle\)
  - \(\langle\{A, B\}, \frac{3}{5}\rangle\)
  - \(\langle\{A, C\}, \frac{2}{3}\rangle\)
  - ...
Outline

- Introduction to Data Mining
- My Works in Data Mining
  - Spatial Data Mining
  - Spatial Association Mining in Cloud Computing Environment
    - Background: Hadoop and MapReduce
    - Proposed Approach
  - Experimental Evaluation
- Summary
Experimental Environment

- For real resizable clusters, Amazon Web Services (AWS) Elastic MapReduce (EMR) platform

  - Clusters with 1 ~ 20 nodes
  - Node type: m1.small (1 CPU, 1.7GB memory, 160GB storage), m1.large
  - Software: Hadoop 1.0.3, Hbase
  - Programming language: Java, MapReduce API, Hbase API

* Source: Amazon AWS
Results with Synthetic Datasets

EXP1-2 (30K points)
EXP1-3 (50K points)
EXP1-4 (70K points)

Execution time (sec)

Job3 (Frequent set search)
Job2 (Neigh. trans. generation)
Job1 (Neighbor search)

EXP1-4: f100p70K

* For other results, please refer the paper in http://users.ipfw.edu/yooj/publication.html
Results with Real-world Datasets

- Theft incidents (5218 points) and Point of interests (765 points, 16 types) in Fort Wayne area, and 1km for neighbor distance

- Point of Interests in Washington D.C. (17000 points, 87 types)

Graphs showing:
- Bar chart with execution time (sec) vs number of cluster nodes for different experiments (EXP5-1-2 for Frequent set mining and EXP5-1-1 for preprocess).
- Line graph showing execution time vs number of nodes for different distances: distance=0.1 mile, distance=0.15 mile, distance=0.2 mile.
Finding from Washington DC POI Data

- Point of Interests in Washington D.C. (17000 points, 87 types), 0.2 mile for neighbor distance

- {Pitch, Tennis} (0.7)
- {Museum, Public pieces of art} (0.55)
- {Park, Parking} (0.54)
- {Hostel, Recreation ground} (0.5)
- {Bicycle rental, Public pieces of art} (0.49)
- {Building, Parking} (0.48)
- {Attraction, Waste disposal} (0.46)
- {Bus Stop, Fast Food Restaurant} (0.44)
- {10-pin Blowing, Court house} (0.42)
- {Basketball, Tennis} (0.42)
- ...

- {Bank, Pub, Restaurant} (0.4)
- {Bank, Café, Pub} (0.4)
- {Bank, Café, Fast Food Restaurant} (0.43)
- {Bank, Pub, Public pieces of art} (0.41)
- {Bank, Café, Hotel} (0.41)
- {Building, Café, Restaurant} (0.44)
- {Building, Café, Fast Food Restaurant} (0.42)
- {Building, Restaurant, Fast Food Restaurant} (0.42)
- ...

- ...
Visualization of data points of selected 27 types

Visualization of data points of some co-located types
Outline

- Introduction to Data Mining
- My Works in Data Mining
  - Spatial Data Mining
  - Spatial Association Mining in Cloud Computing Environment
    - Background: Hadoop and MapReduce
    - Proposed Approach
    - Experimental Evaluation
- Summary
Summary of Our Contributions

- Develop computational efficient methods to discover spatial association patterns
  - Parallel/distributed approaches in cloud computing environment *(IEEE BigData’13, IEEE BigData’14, Int’l Conf. in Adv. In Big Data Analytics’14)*
  - Incremental update approach *(PATTERN’14)*
  - Join-less approach *(ICDM’05, TKDE’06)*
  - Partial join approach *(ACM-GIS’04)*

- Propose variant co-location patterns
  - Reduce sets of co-locations *(DMKD’13 accepted)*
  - Different framework of co-location mining *(DMKD’12)*
  - Top-k closed co-location patterns *(ICSDM’11)*
  - Maximal co-location patterns *(DaWak’11)*
  - N-most prevalent co-location patterns *(DaWak’09)*
  - Co-location mining for extended objects such as line and polygon *(SDM’04)*
Acknowledgement

- The recent work is supported by Air Force Research Laboratory (AFRL), Griffiss Business and Technology Park, and SUNY Research Foundation.
Thank you
Questions

Email to yooj@ipfw.edu
Homepage: http://users.ipfw.edu/yooj