Association Analysis for Large Spatial / Spatio-temporal Data

Ph.D student, Jin Soung Yoo

Spatial Database/Data Mining Group
Department of Computer Science & Engineering
University of Minnesota

Advisor : Dr. Shashi Shekhar
Spatial Association Pattern: Co-locations

(a) Epidemiology (death from cholera, water pump, 1854)  
(b) Ecology (symbiotic species)  

(c) Business (Burger King, MacDonald's)  
(d) Location-based Service (Mobile services)
Spatial Co-location - Definition

- A subset of spatial events whose instances are frequently located in a neighborhood.

- **An input spatial dataset**
  
  Records of location-based Services

<table>
<thead>
<tr>
<th>ID</th>
<th>Position</th>
<th>Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>xxx</td>
<td>(14975,27020)</td>
<td>Weather</td>
</tr>
<tr>
<td>xxx</td>
<td>(16723,24301)</td>
<td>Timetable</td>
</tr>
<tr>
<td>xxx</td>
<td>(14421,26441)</td>
<td>Ticket</td>
</tr>
<tr>
<td>xxx</td>
<td>(14433,27771)</td>
<td>Restaurant</td>
</tr>
<tr>
<td></td>
<td>(15533, 22221)</td>
<td>Sales</td>
</tr>
<tr>
<td>xxx</td>
<td>(14373,26752)</td>
<td>Timetable</td>
</tr>
</tbody>
</table>

- **Spatial events**
  (service request types)

  - Ticket: K
  - Timetable: T
  - Sales: S
  - Restaurant: R

  Co-located event sets: \{K, T\}, \{K, R\}
Interest Measures

- **Co-location rule** $C : C_1 \rightarrow C_2 (p, cp)$, where $C = C_1 \cup C_2$, $C_1 \cap C_2 = \emptyset$

  - **Prevalence measure** : Participation index ($p$) [Shekhar’01]

    $p = Pi(C) = \min \{Pr(C, e_i)\}$

    where $Pr(C, e_i)$ is the participation ratio of $e_i$ in co-location $C$.

    $Pr(C, e_i) = \frac{\# \text{ of distinct objects of } e_i \text{ in instances of } C}{\text{total } \# \text{ of objects of } e_i}$

    e.g., Co-location $C = \{A, B\}$

    $Pr(\{A, B\}, A) = 3/4$, $Pr(\{A, B\}, B) = 3/5$,

    $Pi(\{A, B\}) = \min(Pr(\{A, B\}, A), Pr(\{A, B\}, B)) = \min(3/4, 3/5) = 3/5$  

    if $> a$ threshold,

    co-location $\{A, B\}$ is prevalent.

- **Conditional Probability** ($cp$)

  $cp = Pr(C_1|C_2) = \frac{\# \text{ of distinct instances of } C_1 \text{ in instances of } C_1 \cup C_2}{\# \text{ of instances of } C_1}$

  e.g., Co-location rule $C : A \rightarrow B$,

  $Pr(C_1|C_2) = Pr(A|B) = 3/4$
Problem Definition

- **Given**
  - A spatial dataset: A set of spatial event types and a set of their instances <instance id, event type, location>.
  - A spatial neighbor relationship $R$
  - A minimum prevalence threshold ($min\_prev$) and a minimum conditional probability threshold ($min\_condi$)

- **Find**
  - A set of co-location rules whose prevalence value > $min\_prev$ and conditional probability > $min\_condi$

- **Objective**
  - Find a correct and complete set of co-location rules.
  - Reduce the computation cost.

- **Constraints**
  - $R$ is a distance based neighbor relationship and has symmetric property.
  - The spatial dataset is a point dataset.
Related Work

- **Spatial Statistics : Ripley’s Cross K-Function [Cressie,93]**
  - Spatial correlation analysis for pairs of spatial features
  - Not well defined for size $\geq 3$ features

- **Association Rule Mining [Agarwal, VLDB’94]**
  - Transaction is core concept.
  - No explicit transaction concept in continuous space

- **Spatial Association Rule Mining [Han, SSD’95]**
  - A reference feature (e.g., cancer)-based association analysis
  - Not proper for modeling for co-locations having clique relationship

- **Co-location Pattern Mining [Huang, SSTD’01]**
  - Join-based operating for finding instances
  - Computationally expensive
Challenges

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

- Continuous neighbor relationship, large numbers of features and instances.
  - Computationally expensive for finding all co-location instances having clique relationships.
  - Naïve approach: All maximal clique finding problem (NP-Complete)

- Need schemes for pruning candidate co-locations and filtering co-location instances efficiently.
Our Contributions

- Propose two novel models which materialize spatial neighbor relationships without any info loss.
  - Clique neighborhood partitioning
  - Star neighborhood partitioning

- Develop efficient co-location mining algorithms
  - A partial join method
  - A join-less method

- Validation
  - Prove the algorithms are correct and complete.
  - Provide analytical comparison of the computation costs.
  - Experimentally evaluate the algorithms.
    - Join-less < Partial-join < Join-based
Main Concepts

- **Key idea**: Materialization of spatial neighbor relationships for efficient co-location pattern mining

- **Main challenges**
  - Cut neighbor relationships
  - Storage: Duplicate relationships
  - How to find Clique relationships
  - Materialization cost

- **Main concepts**
  - Materialize disjoint clique neighborhoods.
    + Trace co-location instances related to cut relationships.
  - Materialize disjoint star neighborhoods (without any loss)
    + Check the cliqueness.

How can find clique relationships?
Clique Neighborhood Materialization

- Clique neighborhood partitioning (node partition)

Disjoint clique neighborhoods

<table>
<thead>
<tr>
<th>Clique No</th>
<th>Neighbor objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B.2, B.5</td>
</tr>
<tr>
<td>2</td>
<td>A.1, B.1</td>
</tr>
<tr>
<td>3</td>
<td>A.3, A.4, C.1</td>
</tr>
<tr>
<td>4</td>
<td>B.3, C.3</td>
</tr>
<tr>
<td>5</td>
<td>A.2, B.4, C.2</td>
</tr>
</tbody>
</table>

Cut neighbor relationships

- \{A.1, C.1\}
- \{B.3, C.3\}
- \{B.3, C.1\}

- **Cost**: depends on clique partitioning method. → Disjoint grid partitioning.
- **Model clique relationships directly**: Gather co-location instances simply.
- **Have ‘false dismissals’**: Need to trace cut co-location instances.

- \* \(d\) : neighbor distance
- \* nodes : spatial objects
- \* edges : neighbor relationships \(\leq d\)
Partial Join Co-location Mining Algorithm

Clique neighborhoods

<table>
<thead>
<tr>
<th>Neighbor objects</th>
<th>intra instances of {A, B, C}</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.2, B.5</td>
<td>{A.2, B.4, C.2}</td>
</tr>
<tr>
<td>A.1, B.1</td>
<td></td>
</tr>
<tr>
<td>A.3, A.4, C.1</td>
<td></td>
</tr>
<tr>
<td>B.3, C.3</td>
<td></td>
</tr>
<tr>
<td>A.2, B.4, C.2</td>
<td></td>
</tr>
</tbody>
</table>

Size 2 inter instances

\[\{A.1, C.1\}, \{B.3, C.3\}, \{B.3, C.1\}\]

\[
\text{Participation index: } \frac{2}{4} \quad \frac{2}{5} \quad \frac{2}{3}
\]

If \(>\) threshold, \(\{A, B, C\}\) is a prevalent co-location.
Partial Join vs. Join-based

Partial Join approach

Clique neighborhood set

<table>
<thead>
<tr>
<th>Neighbor objects</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B.2, B.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A.1, B.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A.3, A.4, C.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B.3, C.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A.2, B.4, C.2</td>
<td></td>
</tr>
</tbody>
</table>

Size 2 inter instances

A.3  B.3  join

\{A.2, B.4, C.2\} \rightarrow \{A.2, B.4, C.2\}, \{A.3, B.3, C.1\}

Join-based approach [Huang'01]

Size 2 instances

A.1  B.1  join

\{A.2, B.4, C.2\} \rightarrow \{A.2, B.4, C.2\}, \{A.3, B.3, C.1\}

Size 3

Co-location \{A, B, C\}

\{A.2, B.4, C.2\}, \{A.3, B.3, C.1\}
Star Neighborhood Materialization

- Star neighborhood partitioning (edge partition)

Cost: cheap.
- Does not model clique relationships directly → ‘False positives’
  → Need check of cliqueness

Disjoint star neighborhoods

<table>
<thead>
<tr>
<th>Center object</th>
<th>Neighbor objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A.1, B.1, C.1</td>
</tr>
<tr>
<td></td>
<td>A.2, B.4, C.2</td>
</tr>
<tr>
<td>A.2</td>
<td>A.3, B.3, C.1</td>
</tr>
<tr>
<td>A.3</td>
<td>A.4, C.1</td>
</tr>
<tr>
<td>A.4</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>B.1</td>
</tr>
<tr>
<td>B.1</td>
<td>B.2</td>
</tr>
<tr>
<td>B.2</td>
<td>B.3, C.1, C.3</td>
</tr>
<tr>
<td>B.3</td>
<td>B.4, C.2</td>
</tr>
<tr>
<td>B.4</td>
<td></td>
</tr>
<tr>
<td>B.5</td>
<td>B.5</td>
</tr>
</tbody>
</table>
Join-less Co-location Mining Algorithm

Star neighborhoods

<table>
<thead>
<tr>
<th>Center object</th>
<th>Neighbor objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A.1, B.1, C.1, B.2, A.2, B.4, C.2, A.3, B.3, C.1, A.4, C.1</td>
</tr>
<tr>
<td>B</td>
<td>B.1, B.2, B.3, C.1, C.3, B.4, C.2, B.5</td>
</tr>
<tr>
<td>C</td>
<td>C.1, C.2, C.3, C.4</td>
</tr>
</tbody>
</table>

Co-location \{A, B, C\}

**star instance** of \{A,B,C\}

- \{A.1, B.1, C.1\}
- \{A.2, B.4, C.2\}
- \{A.3, B.3, C.1\}

Check if clique

- \{A.1, B.1, C.1\}
- \{A.2, B.4, C.2\}
- \{A.3, B.3, C.1\}

**instance** of \{A,B,C\}

- \{A.2, B.4, C.2\}
- \{A.3, B.3, C.1\}

Participation ratio:

\[
\begin{array}{ccc}
2/4 & 2/5 & 2/3 \\
\end{array}
\]

If > threshold,
\{A, B, C\} is a prevalent co-location.
Join-less vs. Join-based

Join-less approach

<table>
<thead>
<tr>
<th>Center object</th>
<th>Neighbor objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A.1, B.1, C.1</td>
</tr>
<tr>
<td></td>
<td>A.2, B.4, C.2</td>
</tr>
<tr>
<td></td>
<td>A.3, B.3, C.1</td>
</tr>
<tr>
<td></td>
<td>A.4, C.1</td>
</tr>
<tr>
<td>B</td>
<td>B.1</td>
</tr>
<tr>
<td></td>
<td>B.2</td>
</tr>
<tr>
<td></td>
<td>B.3, C.1, C.3</td>
</tr>
<tr>
<td></td>
<td>B.4, C.2</td>
</tr>
<tr>
<td></td>
<td>B.5</td>
</tr>
</tbody>
</table>

Star neighborhoods

Join-based approach

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.1</td>
<td>B.1</td>
<td>C.1</td>
</tr>
<tr>
<td>A.2</td>
<td>B.4</td>
<td>C.2</td>
</tr>
<tr>
<td>A.3</td>
<td>B.3</td>
<td>C.1</td>
</tr>
</tbody>
</table>

Size 2

Instances of \{A, B, C\}

\{A.2, B.4, C.2\}

\{A.3, B.3, C.1\}

Co-location {A, B, C}

Size 3

star instance of \{A, B, C\}

\{A.1, B.1, C.1\}

Instances of \{A, B, C\}

\{A.2, B.4, C.2\}

\{A.3, B.3, C.1\}
Completeness and Correctness

- **Definitions**
  - **Completeness**:
    Find all co-location rules whose prevalence $> prev\_threshold$ and conditional probability $> condi\_threshold$.
  - **Correctness**:
    All co-location rules found have prevalence $> prev\_threshold$ and conditional probability $> condi\_threshold$.

- **Theorem 1.** The partial join co-location mining algorithm is complete and correct.
  - Proof sketch of completeness: The join method to trace inter instances with objects having at least one cut relationship are complete.

- **Theorem 2.** The join-less co-location mining algorithm is complete and correct.
  - Proof sketch of completeness: The set of star instances are a upper set of clique instances and the clique check method is correct.
Scalability of the join-less algorithm

- Increase overall neighbor density with number of spatial point objects, neighbor distance and prevalence threshold
- Join-less < Partial-join < Join-based
Retail Analysis with Spatio-temporal Data

- Wal-Mart has studied customer buying patterns in hurricane-prone areas.

- Wal-Mart sent hundreds of thousands of cases of bottled water, Strawberry Pop-Tarts, generators, etc. to distribution centers in southern Florida.

'THE ONLY LIFELINE WAS THE WAL-MART'
The world’s biggest company flexed its massive distribution muscle to deliver vital supplies to victims of Katrina. Inside an operation that could teach FEMA a thing or two.

By DEVIN LEONARD
October 3, 2005

FORTUNE

There was little guesswork involved. Wal-Mart has studied customer buying patterns in hurricane-prone areas. Some of the company’s findings are obvious. When a storm is on the way, customers stock up on bottled water, flashlights, generators, and tarps. Afterward, they buy chain saws and mops. But there have been surprises too. Customers also load up on Strawberry Pop-Tarts. Why is that? “They are preserved until you open them, the whole family can eat them, and they taste good,” says Dan Phillips, Wal-Mart’s vice president, information systems division.
Time-profiled Association Analysis

- **Time-profiled Association Mining**
  - Discovers subsets of items (events) whose prevalence variations are similar to a given specific query sequence.

- **Examples**
  - Climate event sets / El Nino index value, Item sets / Hurricane strength

- **Challenges**
  - Composite interest measure: A sequence of prevalence values
  - Search spaces: similarity, temporal, space

1. From temporal transaction dataset (Time-profiled associations)

2. From spatio-temporal dataset (Co-evolving spatial events)