Discovering Spatial Co-location Patterns: A Summary of Results

Yan Huang
Spatial Database Lab
Department of Computer Science
University of Minnesota
huangyan@cs.umn.edu

Full paper by Shashi Shekhar and Yan Huang at:
http://www.cs.umn.edu/research/shashi-group

Co-location Rules (Shashi Shekhar and Yan Huang), July 2001
Application Domains

* Ecology
  * Lansing woods tree data [Diggle 83]
  * Predator-prey species, symbiosis

* Immunogold Labeling

* Epidemiology
  * food-types, obesity, heart disease

* Examples from climatology [Potter01]

<table>
<thead>
<tr>
<th>Pattern #</th>
<th>Variable A</th>
<th>variable B</th>
<th>Examples of interesting patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Cropland Area</td>
<td>Vegetation</td>
<td>Higher cropland area alters NPP</td>
</tr>
<tr>
<td>P2</td>
<td>Smoke Aerosol Index</td>
<td>Precipitation</td>
<td>Smoke aerosols alter the likelihood of rainfall in a nearby region</td>
</tr>
<tr>
<td>P3</td>
<td>Sea Surface Temperature</td>
<td>Land Surface Climate and NPP</td>
<td>Surface ocean heating affects regional terrestrial climate and NPP</td>
</tr>
</tbody>
</table>
Motivating Example

* Given:
  * A collection of different types of spatial events

* Illustration

* Find: Co-located subsets of event types
Event Centric Model

* Non-transaction approach

An example: $A$ and $B$ happen in a neighborhood $\rightarrow C$ happens in their neighborhood with 80% conditional probability

* Application domain: Ecology
## Association Rules - An Analogy

* Association rule e.g. (Diaper in T → Beer in T)

<table>
<thead>
<tr>
<th>Trans.</th>
<th>Items Bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{socks, (\text{milk}), (\text{beef}), (\text{egg}), ...} }</td>
</tr>
<tr>
<td>2</td>
<td>{pillow, (\text{milk}), toothbrush, ice-cream, muffin, ...} }</td>
</tr>
<tr>
<td>3</td>
<td>{ (\text{milk}), (\text{beef}), pacifier, formula, blanket, ...} }</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>n</td>
<td>{battery, juice, beef, egg, chicken, ...} }</td>
</tr>
</tbody>
</table>

* Support: pr(Diaper and Beer in T)
* Confidence: pr(Beer in T|Diaper in T)

* Algorithm Apriori [Agrawal, Srikant, VLDB94]
  * Support based pruning using monotonicity

* Note: **Transaction is a core concept!**
Related Work: Spatial Association Rules

* Force-fit notion of transaction
* Reference feature centric model [Koperski, Han, SSD95]
  * All relevant co-locations reference to one feature

* Item types = boolean spatial features
* Transactions = Instances of reference feature
* An example: B and C are close to A \(\Rightarrow\) D is close to A with 60% conditional probability

* Application domain:
  - Focus on a specific boolean spatial feature, e.g. cancer
  - Need a reference feature
* Q: will it discover co-location rules?
New Challenges

* Association Rules Vs. Co-location Rules

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Association Rule</th>
<th>Co-location Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underlying Space</td>
<td>Discrete Sets</td>
<td>Continuous Space</td>
</tr>
<tr>
<td>Item Types</td>
<td>Product types</td>
<td>Spatial Features(Boolean)</td>
</tr>
<tr>
<td>Item Collections</td>
<td>Transactions ${T_i}$</td>
<td>Neighborhoods</td>
</tr>
<tr>
<td>Prevalence $(A \rightarrow B)$</td>
<td>Support: $p(A \cup B \in T_i)$</td>
<td>Participation Index</td>
</tr>
<tr>
<td>Conditional Probability $(A \rightarrow B)$</td>
<td>$p(B \in T_i</td>
<td>A \in T_i)$</td>
</tr>
</tbody>
</table>

* Items? transactions?
  * Spatial transactions may not be nature!
  * Support is not defined
  * Support based pruning (Apriori) not defined
Overview

* Introduction
  \rightarrow Problem Formulation
* Co-location Miner Algorithm
* Conclusions
Problem Formulation

* Given:
  * $K$ Boolean spatial feature types
  * Instances $<id, \text{feature type } t, \text{location } l>$
  * A neighbor relation $R$ over locations
  * Thresholds: prevalence and conditional probability

* Find:
  * Co-location rules with high prevalence and high conditional probability

* Objectives:
  * Completeness, Correctness, Efficiency

* Constraints:
  * Symmetric $R$
  * Monotonic prevalence measure
  * Event centric model
  * Sparse data set
Example of Key Concepts

* Co-locations of size 2
  * (A,B), (A,C), (B,C)

* Table instances of co-locations of size 2
  * (A,B): (1,1), (2,4), (3,4)
  * (A,C): (1,2), (3,1)
  * (B,C): (2,1), (4,1), (5,3)

* Participation Indexes of co-locations of size 2
  * (A,B): min(3/4, 2/5) = 2/5
  * (A,C): min(2/4, 2/3) = 2/4
  * (B,C): min(3/5, 2/3) = 3/5

* Example Dataset:

Legend:
  T.i: instance i with feature type T
  -: neighbor relationships
Key Concepts

* A Co-location $C$:
  * A subset of boolean spatial features

* A co-location rule $C_1 \rightarrow C_2(p, cp)$:
  * $C_1$ and $C_2$ are co-locations
  * $p =$ prevalence measure
  * $cp = \Pr[C_2 \in N(L) \mid C_1 \in L]$

* A Neighborhood:
  * A clique in a graph of neighbor relation $R$

* A row instance $I$ of a co-location $C = \{f_1, \ldots, f_k\}$:
  * $I = \{i_1, \ldots, i_k\}$
  * $i_j$: instance of $f_j (\forall j \in 1, \ldots, k)$
  * $I$ is a neighborhood

* Table instance (co-location $C = \{f_1, \ldots, f_k\}$):
  * Collection of all its row instances
  * Spatial join interpretation
Key Concepts cont ...

* Participation ratio
  * \( pr(C, f_i) = \frac{|\pi_{f_i} instance(C)|}{|instances(f_i)|} \)
  * \( C = \{f_1, f_2, \ldots, f_k\} \)
  * Monotonically decreasing

* The participation index
  * \( pi(C) = \min_{i=1}^k pr(C, f_i) \)

* Statistical interpretation: ongoing work
Overview

* Introduction
* Problem Formulation
  ⇒ Co-location Miner Algorithm
* Conclusions
Co-location Miner Algorithm

* Co-location Miner
  * Initialization
  * Generate size 2 co-location rules
  * for \( k \) in \((2, 3, \ldots, K - 1)\) do
    - 1. Generate size \( k \) candidate co-locations (\( apriori\_gen \))
    - 2. Generate table instances/prune based on neighborhood
    - 3. Prune based on prevalence of co-locations
    - 4. Generate co-location rules of size \( k \)
  * end

* Note: Step 2 not needed in mining association rules
  * because item collections (i.e. transactions) are given

* Execution Trace

<table>
<thead>
<tr>
<th>Size</th>
<th>Cand. Co-loc</th>
<th>Tables</th>
<th>Table Instances</th>
<th>Par. Ind.</th>
<th>Prev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(A)</td>
<td>T1</td>
<td>(1),(2),(3),(4)</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>(B)</td>
<td>T2</td>
<td>(1),(2),(3),(4),(5)</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>(C)</td>
<td>T3</td>
<td>(1),(2),(3)</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>(A,B)</td>
<td>T4=,T1,\times,T2</td>
<td>(1,1),(2,4),(3,4)</td>
<td>min(3/4,2/5)=2/5</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>(A,C)</td>
<td>T5=,T1,\times,T3</td>
<td>(1,2),(3,1)</td>
<td>min(2/4,2/3)=2/4</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>(B,C)</td>
<td>T6=,T2,\times,T3</td>
<td>(2,1),(4,1),(5,3)</td>
<td>min(3/5,2/3)=3/5</td>
<td>Y</td>
</tr>
<tr>
<td>3</td>
<td>(A,B,C)</td>
<td>T7=,T4,\times,T5</td>
<td>(3,4,1)</td>
<td>min(1/4,1/5,1/3)=1/5</td>
<td>?</td>
</tr>
</tbody>
</table>
Details of Co-location Miner

* Apriori-gen

  * Join step:
    insert into $C_{k+1}$
    select $p.f_1, \ldots, p.f_k, q.f_k, p.table_id, q.table_id$
    from $C_k p$, $C_k q$
    where $p.f_1 = q.f_1$ and $\ldots$ and $p.f_{k-1} = q.f_{k-1}$
    and $p.f_k < q.f_k$

  * Prune step:
    forall co-locations $c \in C'_{k+1}$ do
    forall size $k$ subset $s$ of $c$ do
      if ($s \notin C_k$) then
        delete $c$ from $C_{k+1}$;
Details of Co-location Miner cont...

* Generate table instance
  * forall co-location $c \in C_{k+1}$
    insert into $T_c$
    select $p.i_1, p.i_2, \ldots, p.i_k, q.i_k$
    from $c.table_id_1.p, c.table_id_2.q$
    where $p.i_1=q.i_1$ and $\ldots$ and $p.i_{k-1}=q.i_{k-1}$
    and $(p.i_k, q.i_k) \in R$;
  
end;

* Participation Indexes Calculation
  * Bitmap index based
  * One scan of table instances in current iteration

* Co-location rule generation:
  * See paper ...
Completeness and Correctness

* Definition:
  * Completeness:
    Find all rules with prevalence and conditional probability > thresholds
  * Correctness:
    Any rules found have prevalence and conditional probability > thresholds

* Theorem
  * Co-location Miner is complete and correct

* Proof:
  * Monotonic participation index
  * Any prevalent co-location’s subset is prevalent
  * Table join will not miss any instance
Conclusions

* Our Contributions
  * A NATURAL spatial association model, i.e. co-location rules
  * Eliminates need to transactionize spatial data
  * Co-location Miner Algorithm
  * Proof of correctness and completeness

* Future Work
  * Statistical interpretation
  * Performance evaluation
  * Other spatial data types: polygons, lines, etc.
  * Spatio-temporal datasets
  * Shameless plug for “Spatial Database: A Tour” book :)
  * Questions?