A Framework for Discovering Co-location Patterns in Data Sets with Extended Spatial Objects

Hui Xiong

Department of Computer Science & Engineering
University of Minnesota - Twin Cities
Overview

- Introduction
  - General Problems
  - Related Works
  - Research Motivations
- A Buffer-based Model
- A Filter-and-Refine Co-location Pattern Mining Framework
- Experimental Evaluation
- Conclusions and Future Work
Introduction & Background
### Examples of Co-location Patterns

<table>
<thead>
<tr>
<th>Domains</th>
<th>Example Features</th>
<th>Example Co-location Patterns</th>
</tr>
</thead>
<tbody>
<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>Economics</td>
<td>industry types</td>
<td>(suppliers, producers, consultants)</td>
</tr>
<tr>
<td>Epidemiology</td>
<td>disease types and environmental events</td>
<td>(West Nile disease, stagnant water sources, dead birds, mosquitoes)</td>
</tr>
<tr>
<td>Location-based service</td>
<td>service type requests</td>
<td>(tow, police, ambulance)</td>
</tr>
<tr>
<td>Weather</td>
<td>fronts, precipitation</td>
<td>(cold front, warm front, snow fall)</td>
</tr>
<tr>
<td>Transportation</td>
<td>delivery service tracks</td>
<td>(US Postal Service, UPS, newspaper delivery)</td>
</tr>
</tbody>
</table>
Related Works

- Spatial Statistics
  - Use measures of spatial correlation to characterize the relationship between spatial features
    * the cross-K function with Monte Carlo simulation
    * mean nearest neighbor distance
    * spatial regression model
  - Computationally expensive

- Data Mining Approaches
  - A clustering based approach by Estivill-Castro et al.
    * Features can be completely spatially random or declustered.
    * Sensitive to the choices of clustering algorithms.
  - Association-rule based approaches.
Related Works - Cont.

- Association-rule based approaches.
  - Transaction-based approaches.
    * A reference-feature centric model by Koperski et al.
      - Generalizing this paradigm to the case where no reference feature is specified is non-trivial.
      - May yield duplicate counts for many candidate associations.
  - Distance-based approaches.
    * k-neighboring class sets by Morimoto.
      - the number of instances for each pattern is used as the prevalence measure
    * an event centric model by Shekhar et al.
  - All these approaches are for point spatial features.
Motivation

- Identifying co-location patterns in data sets with extended spatial objects (e.g. polygons and line strings).
  - Highway often have frontage road nearby in large metropolitan.
  - nomandale Road $\Rightarrow$ highway 100
Problem Formulation

- Given:
  - A set $T$ of $K$ spatial feature types $T = \{f_1, f_2 \ldots, f_k\}$ and spatial data types can be point as well as other extended spatial objects, such as line strings and polygons.
  - A set of $N$ instances $P = \{p_1 \ldots p_N\}$, each $p_i \in P$ is a vector $<\text{instance-id}, \text{spatial feature type}, \text{location}>$ where spatial feature type $\in T$ and location $\in$ spatial framework $S$.
  - A buffer size, a minimum prevalence threshold, a minimum conditional probability threshold.

- Find: Co-location Patterns and Co-location Rules.

- Objective: Computational Efficiency.

- Constraints: Correctness and Completeness.
A Buffer-based Model

Definition 1 Buffer is a zone of specified distance around spatial objects. The boundary of the buffer is the isoline of equal distance to the edge of the objects.

- Motivation
  - Objects in space frequently have sort of impact on the objects and areas around them
    - freeways create “noise pollution” that can be heard blocks away.
    - factories emit fumes that can affect people for miles around.
A Buffer-based Model

Definition 2 $N(p)$, the size-$d$ Euclidean neighborhood of a point location $p$, is a circle of side $d$ with $p$ as its center.

Definition 3 $N(o)$, the size-$d$ neighborhood of an extended spatial object (e.g. polygon, line-string), is defined by the buffer operation.
A Buffer-based Model - Cont.

Definition 4 The coverage ratio $Pr(f_1 f_2 \ldots f_k)$ for a co-location $C = \{f_1, \ldots, f_k\}$ is $\frac{N(f_1 f_2 \ldots f_k)}{\text{The total area of the plane}}$, where $N(f_1 f_2 \ldots f_k)$ is the Euclidean neighborhood of the co-location $C$.

Definition 5 The conditional probability $Pr(C_2|C_1)$ of a co-location rule $C_1 \rightarrow C_2$ is the probability of finding the neighborhood of $C_2$ in the neighborhood of $C_1$. It can be computed as $\frac{N(C_1 \cup C_2)}{N(C_1)}$ using the neighborhoods of co-locations $C_1$ and $C_1 \cup C_2$.

Lemma 1 The coverage ratio for co-location patterns is monotonically non-increasing with the size of the co-location pattern increasing.
A Coarse-Level Co-location Pattern Mining Framework

**Definition 6** $BN(o)$, the bounding neighborhood of a spatial object is defined as $MBBR(Buffer(MOBR(Spatial Object O), d))$, where MOBR is the minimum object bounding box, Buffer is the buffer operation with a buffer size as $d$, and MBBR is the minimum buffer bounding box.

**Definition 7** The Euclidean bounding neighborhood $BN(f_j)$ of a spatial feature $f_j$ is the union of $BN(i_l)$ for every instance $i_l$ of the spatial feature $f_j$. 
Definition 8 The Euclidean bounding neighborhood $BN(f_1 f_2 \ldots f_k)$ for a coarse-level co-location pattern $CC = \{f_1, \ldots, f_k\}$ is the intersection of $BN(f_i)$ for every spatial feature $f_i$ in $CC$.

Definition 9 The coarse-level coverage ratio $CPr(f_1 f_2 \ldots f_k)$ for a coarse-level co-location pattern $CC = \{f_1, \ldots, f_k\}$ is $\frac{BN(f_1 f_2 \ldots f_k)}{The \ total \ area \ of \ the \ plane}$, where $BN(f_1 f_2 \ldots f_k)$ is the Euclidean bounding neighborhood of the coarse-level co-location pattern $CC$. 
A Coarse-Level Co-location Pattern Mining Framework - Cont.

Lemma 2 The coarse-level coverage ratio for coarse-level co-location patterns is monotonically non-increasing with the size of the coarse-level co-location pattern increasing.

Lemma 3 For any spatial feature set $F = \{f_1, f_2, \ldots, f_k\}$, the coarse-level coverage ratio $CPr(F)$ is larger than or equal to the coverage ratio $Pr(F)$. 
A Coarse-Level Co-location Pattern Mining Framework - Cont.

- $CP^r(A) = \frac{BN(A)}{\text{The total area of the plane}} = \frac{35}{200} = 0.175$

- $CP^r(AB) = \frac{BN(AB)}{\text{The total area of the plane}} = \frac{12}{200} = 0.06$. 
Lemma 3 For any $n$ spatial events $A_1, \ldots, A_n$,

$$\bigcup_{i=1}^{n} BN(A_i) = \sum_{i=1}^{n} BN(A_i) - \sum_{i<j} BN(A_iA_j) + \sum_{i<j<k} BN(A_iA_jA_k) - \sum_{i<j<k<l} BN(A_iA_jA_kA_l) + \ldots + (-1)^{n+1} BN(A_1A_2\ldots A_n).$$

(c) University of Minnesota - Twin Cities
April 8, 2004
Geometric Challenges and Solutions - Cont.

Theorem 2 Given any \( n \) spatial events \( A_1, A_2, \ldots, A_n \) and the corresponding bounding neighborhoods \( ((x_{1lb}, y_{1lb}), (x_{1rt}, y_{1rt})) \), \( ((x_{2lb}, y_{2lb}), (x_{2rt}, y_{2rt})) \), \ldots, \( ((x_{nlb}, y_{nlb}), (x_{nrt}, y_{nrt})) \), where the bounding neighborhood of the event \( A_i \), \( 1 \leq i \leq n \), is represented by the left bottom point \( (x_{ilb}, y_{ilb}) \) and the right top point \( (x_{irt}, y_{irt}) \), if the bounding neighborhoods of these \( n \) spatial events have the common intersection area, then this intersection area can be computed by Equation 2.

\[
BN(A_1A_2\ldots A_n) = (X_2 - X_1) \times (Y_2 - Y_1)
\]

where

\[
X_2 = \min\{x_{1rt}, x_{2rt}, \ldots, x_{nrt}\},
\]

\[
X_1 = \max\{x_{1lb}, x_{2lb}, \ldots, x_{nlb}\},
\]

\[
Y_2 = \min\{y_{1rt}, y_{2rt}, \ldots, y_{nrt}\},
\]

\[
Y_1 = \max\{y_{1lb}, y_{2lb}, \ldots, y_{nlb}\}.
\]
• DCS: A Direct Combinatorial Search Algorithm.
• EXCOM: An Extended Co-location Mining Algorithm.
Experimental Setup

- **Experimental Data Set**: MN/DOT base map.
- **Experimental Design**

  ![Diagram]

  - **Candidates**: DCS, EXCOM
  - **Parameters**: Coverage Ratio, Buffer Size
  - **Co-location ratio analysis for line-string co-location patterns**
  - **Summary**
The geometric filter can speed up the prevalence-based pruning approach by a fact of 30 - 40.
Line-String Co-location Patterns for Test Route Selection

- Evaluate the positional accuracy of digital roadmap databases.

- Co-located roads are the most challenging test sites for evaluating the ability of global positioning systems (GPS) systems to identify correct roads from a digital roadmap.
Contributions

- Generalize the concept of co-location patterns to extended spatial objects, e.g. polygons and line-strings.
- Propose a novel buffer-based model for mining co-location pattern. This model has three advantages over the event centric model and is transaction-free.
- Propose a geometric filter-and-refine co-location mining framework.
- Experiment evaluation with a real data set shows that the geometric filter-and-refine approach can speed up the prevalence-based pruning approach by a fact of 30 to 40.
- The application of applying line-string co-location patterns for selecting test routes has been provided to show the usefulness of co-location patterns.
Conclusions and Future Work

- Extending the notion of co-location pattern
  - de-colocation pattern
  - co-incidence pattern
- Applications of Co-location Pattern