Spatial Interestingness Measures for Co-location Pattern Mining

Christian Sengstock
Michael Gertz
Tran Van Canh

Database Systems Research Group
Heidelberg University, Germany

December 10, 2012
**Co-location pattern mining:** Finding subsets of spatial features located together

- Discovery of interesting spatial correlation patterns and rules
- Model multivariate spatial correlation structure
■ **Interestingly Measures**

- Number of mined patterns and rules is typically large
- Natural measure to filter: frequency/support

■ **Spatial Interestingness Measures**

- Patterns have inherent spatial characteristics (randomly distributed, clustered, regional, landmark)
- Spatial measures allow to filter patterns and rules by their spatial characteristics

■ **Applications:**

- Mine descriptive patterns for global/local/clumped phenomena
- Find predictive rules for arbitrary distributed phenomena (e.g. city population)
Related work

- **Regional** (zonal) co-location patterns [Celik, Kang, Shekar; ICDM '07], [Mohan, Shine, et. al.; ACMGGIS '11]

Our work

(a) General description of **spatial distribution of co-location patterns**

(b) Derivation and discussion of basic **spatial interestingness measures**

(c) **Computational considerations** to measure spatial characteristics

(d) **Preliminary experiments** comparing standard and spatial measures
Outline

1. Definitions
2. Spatial Measures
3. Experiments
4. Conclusions
Outline

1. Definitions
2. Spatial Measures
3. Experiments
4. Conclusions
Co-location Patterns

- $O : F \times W$: Spatially distributed features $F$ in window $W$
- $h$: Distance threshold of clique-based neighborhood relation
- $C \subseteq F$: Co-location pattern
- $rowset(C)$: Co-location instances of $C$
- $PI(C)$: Participation index (frequency-based measure)
- $p(C_2|C_1)$: Confidence of co-location rule $C_1 \rightarrow C_2$
Spatial Distribution

- $p(u')$: Spatial density at locations $u' \in W$ of point set $U = \{u_1, \ldots, u_n\} \subset W$

- Non-parametric estimation: Estimation based on distribution function characteristics (smooth, peaky)

- 2D-histogram estimator: Normalized number of points in cells $u' \in G_b$ defined by grid with bandwidth parameter $b$
Instance $i \in \text{rowset}(C)$ has a location $\text{centroid}(i)$

$\text{centroid}(i)$ is only a fraction of $h$ away of instance feature points

$\text{centroid} \{A, B\}$ and $\text{centroid} \{A, B, C\}$ are only a fraction of $h$ apart

$h \ll \sqrt{W}$

$\text{centroid}(i)$ is a good approximation for instance location

$C_1 \subset C_2, i_1 \in \text{rowset}(C_1), i_2 \in \text{rowset}(C_2) \Rightarrow \text{centroid}(i_1) \approx \text{centroid}(i_2)$
**Spatial Pattern Distribution**

- \( U(C) = \{ \text{centroid}(i) | i \in \text{rowset}(C) \} \): Co-location pattern instance points
- \( p_C(u') \): Density of pattern instance points at location \( u' \)
- \( C_1 \subseteq C_2 \Rightarrow U(C_2) \subseteq U(C_1) \): Instance points of \( C_2 \) are a sample of \( C_1 \)
Spatial Intr Measure: Pattern Entropy

- Entropy of $p_C$ over $m$ grid cells $u' \in G_b$ describes the spatial distribution of the pattern

$$\text{entropy}(C) := - \sum_{u' \in G_b : p_C(u') > 0} p(u') \log p(u')$$

Entropy $= 0$ $\Rightarrow$ Pattern only occurs at one location
Small entropy $\Rightarrow$ Pattern has a peaky distribution
Large entropy $\Rightarrow$ Pattern has a smooth distribution
Entropy $= \log(m)$ $\Rightarrow$ Pattern occurs uniformly

<table>
<thead>
<tr>
<th>$C$</th>
<th>$\text{rowset}(C')$</th>
<th>$PI(C')$</th>
<th>$\text{entropy}(C')$</th>
</tr>
</thead>
<tbody>
<tr>
<td>${A}$</td>
<td>8</td>
<td>8/8</td>
<td>1.04</td>
</tr>
<tr>
<td>${A, B}$</td>
<td>4</td>
<td>4/8</td>
<td>1.04</td>
</tr>
<tr>
<td>${A, C}$</td>
<td>4</td>
<td>4/8</td>
<td>0.00</td>
</tr>
<tr>
<td>${B, C}$</td>
<td>2</td>
<td>2/4</td>
<td>0.00</td>
</tr>
<tr>
<td>${A, B, C}$</td>
<td>2</td>
<td>2/8</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Spatial Intr Measure: Pattern KL-Divergence

- KL-divergence measures the similarity of two distributions
- $p_{C_1 \cup C_2}(u')$: Spatial distribution of rule $C_1 \rightarrow C_2$
- $p_{C_1}(u')$: LHS (baseline) distribution of rule
- Spatial similarity of $C_1 \cup C_2$ distribution to $C_1$ represents a spatial characteristic of the rule:

$$KL(C_1 \cup C_2 || C_1) := \sum_{u' \in G_b : p_{C_1 \cup C_2}(u') > 0} p_{C_1 \cup C_2}(u') \log \frac{p_{C_1 \cup C_2}(u')}{p_{C_1}(u')}$$

| $C_1 \rightarrow C_2$ | $p(C_2|C_1)$ | $KL(C_1 \cup C_2 || C_1)$ |
|-----------------------|--------------|--------------------------|
| \{A\} $\rightarrow$ \{B\} | 4/8 | 0.00 |
| \{A\} $\rightarrow$ \{C\} | 4/8 | 0.69 |
| \{A, B\} $\rightarrow$ \{C\} | 2/4 | 0.69 |
| \{C\} $\rightarrow$ \{B\} | 2/4 | 0.00 |
Observations:
- Entropy and frequency measures (e.g. participation index) are related
- KL and confidence are related

Intuition:
- Choose grid with $b$ that small that each instance falls into a unique cell:
  - Higher frequency will lead to higher entropy (pos. correlation)
  - Higher conditional frequency (confidence) will lead to lower KL-div (neg. correlation)

To mine meaningful spatial characteristics: $h \ll b$
Outline

1. Definitions
2. Spatial Measures
3. Experiments
4. Conclusions
Data and Setup

- Small OSM POI data set:
  \[ |F| = 201, |O| = 5840, W = \text{Los Angeles}, h = 0.11 \text{ km} \]
- Evaluation on two grids:
  - High-res grid: \( 101 \times 59 \) (\( b \approx 0.48 \text{ km} \))
  - Low-res grid: \( 11 \times 7 \) (\( b \approx 4.37 \text{ km} \))
- Comparison of entropy and KL-div to support and confidence
## Entropy

<table>
<thead>
<tr>
<th>by support</th>
<th>by entropy (high res)</th>
<th>by entropy (low res)</th>
</tr>
</thead>
<tbody>
<tr>
<td>school place-of-worship</td>
<td>school place-of-worship</td>
<td>school place-of-worship</td>
</tr>
<tr>
<td>fast-food restaurant</td>
<td>convenience fuel</td>
<td>fuel fast-food</td>
</tr>
<tr>
<td>cafe restaurant</td>
<td>fuel fast-food</td>
<td>library place-of-worship</td>
</tr>
<tr>
<td>cafe fast-food</td>
<td>fast-food restaurant</td>
<td>convenience fuel</td>
</tr>
<tr>
<td>convenience fuel</td>
<td>cafe restaurant</td>
<td>fire-station place-of-worship</td>
</tr>
<tr>
<td>bank restaurant</td>
<td>library place-of-worship</td>
<td>park school</td>
</tr>
<tr>
<td>fuel fast-food</td>
<td>supermarket fast-food</td>
<td>convenience fast-food</td>
</tr>
<tr>
<td>cafe fast-food restaurant</td>
<td>level-crossing place-of-worship</td>
<td>fast-food restaurant</td>
</tr>
<tr>
<td>supermarket fast-food</td>
<td>convenience fast-food</td>
<td>park place-of-worship</td>
</tr>
<tr>
<td>bank fast-food</td>
<td>cafe fast-food</td>
<td>library park</td>
</tr>
</tbody>
</table>

- Mined patterns by decreasing (1) support, (2) entropy on high-res grid, (3) entropy on low-res grid
- Patterns with more global distribution climb up
**KL-Divergence**

<table>
<thead>
<tr>
<th>by conf</th>
<th>by kldiv (high res)</th>
<th>by kldiv (low res)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cafe → restaurant</td>
<td>cafe → restaurant</td>
<td>cafe → restaurant</td>
</tr>
<tr>
<td>cafe → fast-food</td>
<td>cafe → fast-food</td>
<td>cafe → fast-food</td>
</tr>
<tr>
<td>cafe → restaurant fast-food</td>
<td>cafe → restaurant fast-food</td>
<td>cafe → bank</td>
</tr>
<tr>
<td>cafe → bank</td>
<td>cafe → bank</td>
<td>cafe → restaurant fast-food</td>
</tr>
<tr>
<td>cafe → bank restaurant</td>
<td>cafe → bank restaurant</td>
<td>cafe → pharmacy</td>
</tr>
<tr>
<td>cafe → bank fast-food</td>
<td>cafe → bank fast-food</td>
<td>cafe → museum</td>
</tr>
<tr>
<td>cafe → bank restaurant fast-food</td>
<td>cafe → bank restaurant fast-food</td>
<td>cafe → bank restaurant</td>
</tr>
<tr>
<td>cafe → convenience</td>
<td>cafe → school</td>
<td>cafe → restaurant pharmacy</td>
</tr>
<tr>
<td>cafe → station</td>
<td>cafe → fuel</td>
<td>cafe → place-of-worship</td>
</tr>
<tr>
<td>cafe → pub</td>
<td>cafe → pharmacy</td>
<td>cafe → bank pharmacy</td>
</tr>
</tbody>
</table>

- Mined rules with LHS='cafe' by (1) decreasing confidence, (2) increasing KL-div on high-res grid, (3) increasing KL-div on low-res grid
- Rules where RHS is spatially more similar to LHS climb up
Conclusions

- **Summary**
  - Introduced measures based on spatial distribution of pattern instance centroids
  - Entropy and KL-divergence are related to support and confidence (choosing $h$ and $b$ matters)

- **Ongoing Work**
  - Derivation of more spatial measures
  - Large-scale extrinsic evaluation
Thank you.