Spatial Interestingness Measures for Co-location Pattern Mining

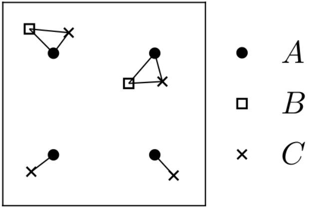
Christian Sengstock
Michael Gertz
Tran Van Canh

Database Systems Research Group
Heidelberg University, Germany

December 10, 2012

UNIVERSITÄT HEIDELBERG

DBS HEIDELBERG



- **Co-location pattern mining**: Finding subsets of spatial features located together
  - Discovery of interesting spatial correlation patterns and rules
  - Model multivariate spatial correlation structure

- **Interestingness 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/clustered 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**

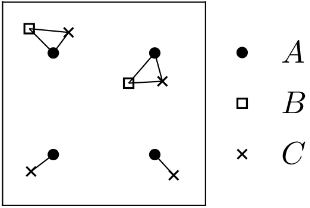
(c) **Computational considerations** to measure spatial characteristics

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

Outline

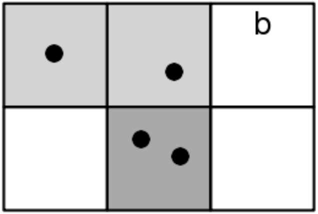
Outline

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 $C_1 \rightarrow C_2$

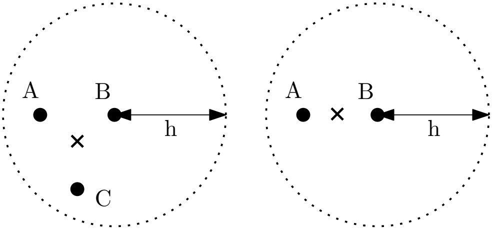
Spatial Distribution



- $p(u')$: 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

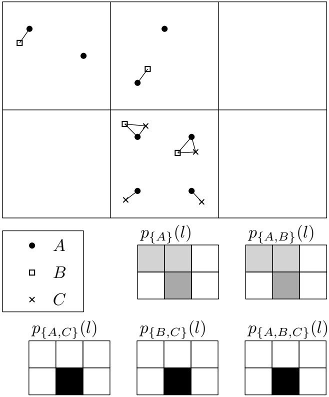
Outline

Co-location Instance Point



- Instance $i \in rowset(C)$ has a location $centroid(i)$
- $centroid(i)$ is only a fraction of h away of instance feature points
- $centroid(\{A, B\})$ and $centroid(\{A, B, C\})$ are only a fraction of h apart
- $h \ll \sqrt{W}$
  - $centroid(i)$ is a good approximation for instance location
  - $C_1 \subset C_2, i_1 \in rowset(C_1), i_2 \in rowset(C_2) \Rightarrow centroid(i_1) \stackrel{\sim}{=} centroid(i_2)$

Spatial Pattern Distribution



- $U(C) = \{centroid(i) | i \in rowset(C)\}$: 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

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

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

C	$\lvert rowset(C) \rvert$	$PI(C)$	$entropy(C)$
$\{A\}$	8	8/8	1.04
$\{A, B\}$	4	4/8	1.04
$\{A, C\}$	4	4/8	0.00
$\{B, C\}$	2	2/4	0.00
$\{A, B, C\}$	2	2/8	0.00

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:

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

$C_1 \rightarrow C_2$	$p(C_2 \vert C_1)$	$KL(C_1 \cup C_2 \vert\vert C_1)$
$\{A\} \rightarrow \{B\}$	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

Discussion

- **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:
    $\Rightarrow$ Higher frequency will lead to higher entropy (pos. correlation)
    $\Rightarrow$ Higher conditional frequency (confidence) will lead to lower KL-div (neg. correlation)
- To mine meaningful spatial characteristics: $h \ll b$

Outline

Data and Setup

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

Entropy

by support	by entropy (high res)	by entropy (low res)
school place-of-worship	school place-of-worship	school place-of-worship
fast-food restaurant	convenience fuel	fuel fast-food
cafe restaurant	fuel fast-food	**library place-of-worship**
cafe fast-food	fast-food restaurant	convenience fuel
convenience fuel	cafe restaurant	**fire-station place-of-worship**
bank restaurant	**library place-of-worship**	**park school**
fuel fast-food	supermarket fast-food	**convenience fast-food**
cafe fast-food restaurant	**level-crossing place-of-worship**	fast-food restaurant
supermarket fast-food	**convenience fast-food**	**park place-of-worship**
bank fast-food	cafe fast-food	**library park**

- 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

by conf	by kldiv (high res)	by kldiv (low res)
cafe → restaurant	cafe → restaurant	cafe → restaurant
cafe → fast-food	cafe → fast-food	cafe → fast-food
cafe → restaurant fast-food	cafe → restaurant fast-food	cafe → bank
cafe → bank	cafe → bank	cafe → restaurant fast-food
cafe → bank restaurant	cafe → bank restaurant	**cafe → pharmacy**
cafe → bank fast-food	cafe → bank fast-food	**cafe → museum**
cafe → bank restaurant fast-food	cafe → bank restaurant fast-food	cafe → bank restaurant
cafe → convenience	**cafe → school**	**cafe → restaurant pharmacy**
cafe → station	**cafe → fuel**	**cafe → place-of-worship**
cafe → pub	**cafe → pharmacy**	**cafe → bank pharmacy**

- 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

Outline

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.