Spatial Interestingness Measures for Co-location Pattern Mining

Christian Sengstock
Insitute of Computer Science
Heidelberg University, Germany
sengstock@informatik.uni-heidelberg.de

Michael Gertz
Institute of Computer Science
Heidelberg University, Germany
gertz@informatik.uni-heidelberg.de

Tran Van Canh
Institute of Computer Science
Heidelberg University, Germany
canh.tran.van@informatik.uni-heidelberg.de

***Abstract*—Co-location pattern mining aims at finding subsets of spatial features frequently located together in spatial proximity. The underlying motivation is to model the spatial correlation structure between the features. This allows to discover interesting co-location rules (feature interactions) for spatial analysis and prediction tasks. As in association rule mining, a major problem is the huge amount of possible patterns and rules. Hence, measures are needed to identify interesting patterns and rules. Existing approaches so far focused on finding frequent patterns, patterns including rare features, and patterns occurring in small (local) regions.**

In this paper, we present a new general class of interestingness measures that are based on the spatial distribution of co-location patterns. These measures allow to judge the interestingness of a pattern based on properties of the underlying spatial feature distribution. The results are different from standard measures like participation index or confidence. To demonstrate the usefulness of these measures, we apply our approach to the discovery of rules on a subset of the OpenStreetMap point-of-interest data.

***Keywords*-Co-location pattern mining, interestingness measures, density estimation**

I. Introduction

Co-location pattern mining tries to find subsets of spatial features frequently located together in spatial proximity in some geographic space. Example applications of co-location pattern mining include, among others, services and queries on mobile phones frequently requested and located together [1], interactions between symbiotic species in ecology [2], and public safety and health [3].

The frequency of a co-location might, however, not be sufficient to identify interesting patterns. As proposed in [3], [4], [5] patterns might only occur in small areas. Then the pattern has a low support, but it is still interesting for a particular region. Furthermore, a pattern might have a low support because a feature does not occur sufficiently often in the dataset, i.e., it is rare. A rule with a rare feature on the left-hand side (LHS) will, however, still be of high confidence and hence might be interesting [6], [7].

In this paper, we study a general class of interestingness measures based on the spatial distribution of co-locations. A co-location pattern naturally has an inherent spatial distribution based on the distribution of the features it describes. For example, a number of features might co-occur over the whole space. Such a co-location then has a meaning in the whole space and might be seen as exhibiting a global phenomenon. On the other hand, some features might occur in the whole space, but they only co-occur in a certain region. Such a co-location can be seen as exhibiting a certain regional phenomenon.

In our approach, we focus on a general formulation of spatial interestingness measures, which allows to define measures to mine co-location patterns and rules based on their spatial characteristics. We do not discuss algorithmic extension and justify our focus based on the following assumption:

- We assume that mining co-locations with small support-like thresholds (like the prevalence index) is possible for very large datasets using large cluster computing environments (e.g., MapReduce [11]) for transactionalization (e.g., like in feature-centric approaches or as shown in [12]) and parallel itemset mining methods like PFP-Growth [13].
- Statistics to compute spatial interestingness measures can easily be tracked in the mining process by extending the pattern data structures, as shown in Section IV. By that, the resulting patterns have all the necessary information to compute the spatial measures when the mining process has been completed. Moreover, we can always compute these statistics based on the mined patterns in a subsequent step in reasonable time, also using, e.g., cluster environments.

In this work, we therefore do not consider the problem of identifying a large number of patterns efficiently, but to filter out interesting patterns from the results. Therefore, our focus is to formulate, track and compute spatial characteristics of patterns for subsequent filtering and exploration tasks.

A. Related Work

Existing work on co-location pattern mining focuses on efficient algorithms to mine patterns with low support-like measure thresholds, e.g., [1], [2], [8]. The proposed techniques allow to efficiently identify patterns by exploiting the anti-monotonic property of support-like measures, such as the participation index. A superset of a pattern will only be considered if the pattern itself is frequent. A natural ordering of the patterns is then done based on decreasing support measure to identify the most interesting co-locations.

Extensions to the above techniques include the identification of co-locations having rare events, which still lead to rules with high confidence [6], [7]. The extensions allow to efficiently mine such rules, but they even increase the number of results. Moreover, in these approaches there is no formal notion of why an event is rare, might it be rare because it just occurs a few times globally, or because it just occurs in a small region.

Recent work studied the efficient identification of regional co-locations [3], [4], [5]. A regional co-location is unlikely to be found by support-like measures because it occurs less often than features that are distributed in the whole space. Beside efficiency considerations, local patterns are interesting because they exhibit local phenomena. The measures of spatial interestingness we introduce in this paper cover the description of local phenomena, among other spatial characteristics.

In [9] a clustering-based visualization of co-locations has been proposed. The motivation is to support the exploration and interpretation of mined patterns. We also consider the spatial distribution of patterns as an important characteristic. However, different from visualization-based analysis, our aim is to identify interesting patterns by measures describing the spatial characteristics. Furthermore, our proposed representation of a co-location pattern as a non-parametric distribution can easily be used for visualization purposes.

B. Contributions

In this paper, we present a new class of interestingness measures describing the spatial characteristics of mined co-location patterns, which we call *spatial interestingness measures* of co-locations. Our contributions are: (a) We formalize the spatial distribution of co-location patterns, (b) based on that we develop a pattern entropy measure and a Kullback-Leibler divergence rule measure to describe their spatial characteristics, (c) we give examples of how spatial measures can be computed based on data structures capturing statistics of the spatial pattern distribution, and (d) we show initial results of mined co-location patterns and rules using an OpenStreetMap (OSM) dataset based on our measures.

II. Background

In this section, we introduce the basic concepts underlying our idea of spatial interestingness of co-locations, which are formally presented in the next section.

A. Co-location Pattern Mining

In the following, we use the event-based approach introduced in [2] to define a co-location pattern mining framework. We note, however, that such measures are not restricted to the event-based approach but can easily be adapted to other methods, such as listed in [8].

As input we are given a set of p discrete spatial features $F = \{f_1, \ldots, f_p\}$. The features are distributed in geographic space $W \subseteq \mathbb{R}^2$ described by the relation $O \subseteq F \times W$. We describe the relation as a set of *feature instances* $o_i = (id, u, f)$ with id being a unique instance id, $u \in W$, and $f \in F$.

Spatial proximity is described by a spatial relation R between feature instances. In this paper, we assume a simple distance-based relation. Two feature instances are related if they co-occur within a given radius h. A neighbor set L is a set of instances such that all pairwise instance locations are spatially related in R, hence forming a clique. A *co-location* then is defined as a subset of features $C \subseteq F$ such that every feature in C appears in at least one neighbor instance of L, and there exists no proper subset in L doing so. In [2] the neighbor set instances of a co-location C are called the *row instances* of C, in the following denoted $rowset(C)$.

To efficiently mine co-locations being interesting regarding their frequency of occurrence the participation index is introduced in [2]. The participation index indicates whether spatial features in a co-location likely show up together. It is defined as the minimum participation ratio $pr(C, f)$ among all features $f \in C$. The participation ratio for a feature f in a co-location C is defined as follows.

$$pr(C, f) := \frac{\text{\# of } f \text{ instances in any row instance of } C}{\text{\# of } f \text{ instances}} \tag{1}$$

The participation index then is defined as the minimum participation ratio among all features in a co-location C,

$$PI(C) := \min_{f \in C} pr(C, f) \tag{2}$$

A *co-location rule* $C_1 \rightarrow C_2$ describes how likely one will find feature instances of C_2 in spatial proximity to feature instances of C_1. The rule has a frequency described by the participation index $PI(C_1 \cup C_2)$ and a conditional probability $p(C_2|C_1)$. This probability is also called the *confidence* of a rule and is simply computed as follows:

$$P(C_2|C_1) = \frac{\text{\# of } C_1 \text{ row instances in any } C_1 \cup C_2 \text{ row instance}}{\text{\# of } C_1 \text{ row instances}} \tag{3}$$

Given a co-location miner, the set of possible co-location patterns is typically very large. Measures to reduce the size of the patterns may include different tasks, such as looking only at patterns with certain features (projection), using higher prevalence/support thresholds, or using higher confidence thresholds for co-location rules. In this paper, we introduce an additional class of measures, allowing to identify patterns based on their spatial characteristics.

B. Non-parametric Density Estimation

We make use of non-parametric density estimation to describe the spatial distribution of co-location patterns. First, we generally assume a set of spatial points U. The density distribution of that point set describes how likely one will find a point $u \in U$ at a certain location l or inside a given area A. The density of the set U at a location l is denoted $p_U(l)$, and the density inside an area A is denoted $p_U(A)$. In parametric approaches the density is estimated by fitting a parametric model (like a Gaussian mixture model). Non-parametric approaches solely make assumptions on the characteristics of the distribution. An important characteristic is the smoothness of a distribution, which is usually described by a bandwidth parameter b. Using a small bandwidth the distribution will be very peaky and shows much small-scale variation, while using a large bandwidth the distribution will be very smooth and describes the large scale variation. In situations where no prior knowledge about the domain of the distribution is available the bandwidth can be chosen using cross-validation [10]. In the geographic domain the bandwidth reflects the scale-level of interest at which one wants to describe a spatial phenomena (e.g, country level, city level, street level). Then the bandwidth b is a function of the scale-level (e.g., 50 km for country level and 5 km for city level).

Among the most prominent estimators are Kernel Density Estimation (KDE) and 2D histograms (also called window counting). KDE estimates the density at a given location or inside an area by summing up the covered point influences described by a Kernel function [10]. While KDE gives a smooth, non-discrete estimation for an arbitrary location l, window counting approximates the distribution by a grid $G_b = \{l_1, \ldots, l_m\}$ over geographic space. Through this the number of points in U falling inside a grid cell $l_1, \ldots, l_m$ are counted. The density at a location $p_U(l)$ then is described by the normalized number of points inside cell $l_i \in G$ in which the location l falls. The width/height of a grid cell b represents the bandwidth parameter of the estimator. In the following, we use the window counting method by employing a sparse grid over geographic space. Smoothness of the distribution might be obtained by kernel convolution on the grid, resulting in KDE-like distributions.

III. Spatial Interestingness

In this section, we describe the spatial distribution of co-location patterns in a general way and develop and discuss spatial interestingness measures.

A. Pattern Distribution

We assume that a co-location instance $i \in rowset(C)$ can be represented by a point location in geographic space, called the *co-location instance point.* This assumption is reasonable, following the definition of an instance i as a clique with a distance-based neighborhood relation. All feature instance locations in a clique occur pairwise within distance h. The centroid of the clique's feature points of co-location instance i is denoted $centroid(i)$. The centroid roughly represents the co-location instance location as the center of a disc with radius h. Other location representations of an instance can easily be employed, e.g., the convex hull of the points, the cell id in discretized space covering the points, or just the set of points itself.

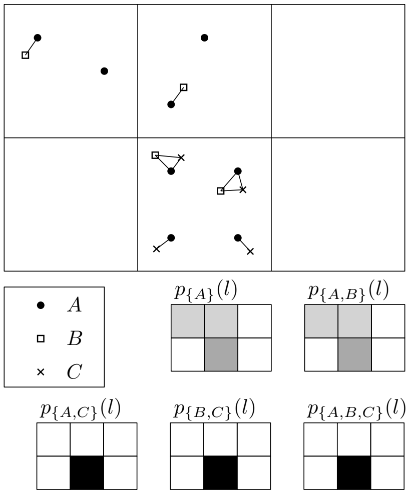

Figure 1. Top: Plot of the spatial feature instances A, B, C and their spatial relation (connecting lines). Bottom: 5 density distribution plots of the row instances of patterns $\{A\}, \{A, B\}, \{A, C\}, \{B, C\}, \{A, B, C\}$.

C	$\|rowset(C)\|$	$PI(C)$	$entropy(C)$
$\{A\}$	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

Figure 2. Table of patterns and their measures.

Given a co-location instance i described by a point location $centroid(i)$, a co-location pattern C defines a set of its instance points:

$$U(C) := \{centroid(i) | i \in rowset(C)\} \tag{4}$$

The point set $U(C)$ allows to derive spatial properties of the pattern. For example, the points might be randomly distributed in geographic space, they tend to cluster around several locations, or they just show up at a single location. Besides the frequency of the pattern in the whole space, we are now able to describe the pattern's spatial distribution and derive measures to describe its spatial characteristic. Additionally, given a density estimation of $U(C)$, we can easily visualize the co-location patterns for visual exploration such as discussed in [9].

Given the co-location instances of $\{A, B\}$ and $\{A, B, C\}$, as shown in the example in Figure 1. Clearly, the row instances $i \in rowset(\{A, B, C\})$ will always be supersets of those row instances $j \in rowset(\{A, B\})$ for which they cover the same feature instances. The centroids $centroid(i)$ and $centroid(j)$ for $j \subseteq i$ will, of course, not be the same, because j is made of two features instances while i is made of three feature instances. However, again referring to the assumption of a co-location instance being a clique, the centroids can just be apart by a fraction of h. In the following we neglect this difference. This can be justified if h is small in relation to the area of interest. Because

$C_1 \to C_2$	$p(C_2 \vert C_1)$	$KL(C_1 \cup C_2 \vert\vert C_1)$
$\{A\} \to \{B\}$	4/8	0.00
$\{A\} \to \{C\}$	4/8	0.69
$\{A, B\} \to \{C\}$	2/4	0.69
$\{C\} \to \{B\}$	2/4	0.00

Figure 3. Table of rules and their measures.

then the points $U_1 = U(\{A, B, C\})$ will always represent co-location instances that are also co-location instances of $U_2 = U(\{A, B\})$, the points U_1 are a subset of U_2. Hence, U_1 can be seen as being generated by a sampling from U_2.

B. Spatial Measures

Using the notion of a point set description $U(C)$ of a co-location pattern C, we can define a variety of measures to describe the spatial characteristics of such a set.

1) Pattern Entropy: The frequency of a pattern describes how often a co-location occurs in the whole space. However, it does not describe how the pattern is distributed. The entropy is a measure in information theory describing how much information is needed on average to encode the observations of a distribution. To identify an observation in a distribution that is almost random, one needs a lot of information (because the possible number of values is high), while identifying an observation in a distribution that describes just a single event needs no information at all (because we already know the value).

Put into the geographic domain with discrete space, the entropy tells us if the point distribution is more uniform (large entropy, globally and smoothly distributed), has a lot of peaks (medium entropy, clusters), or just a single peak (zero entropy, local). To calculate the entropy for a co-location pattern C we define the density of a co-location pattern over the cells of a grid $G_b = \{l_1, \ldots, l_m\}$ as $p_C(l)$. The density can be derived by window counting as explained in Section II-B. The entropy of a pattern C is then defined over all locations $p_C(l_i) > 0$ as follows:

$$entropy(C) := - \sum_{l_i:\ p_C(l_i)>0} p_C(l_i) \log p_C(l_i) \qquad (5)$$

Table 2 shows the entropies of the patterns $\{A, B\}$ and $\{A, C\}$, both having a support of 4 and a participation index of 0.5. However, the entropies differ because $\{A, B\}$ is distributed over 3 cells, while $\{A, C\}$ occurs just in a single cell. The entropy is hence a spatial interestingness measure of a co-location pattern. We can use the entropy to distinguish between global patterns and patterns that have a more peaky distribution (clustered or local).

2) KL-Divergence: The conditional probability $p(C_2|C_1)$ of a rule $C_1 \to C_2$ states how likely one will find features of distribution C_2 within the neighborhood of distribution C_1. Figure 3 shows the rules $\{A\} \to \{B\}$ and $\{A\} \to \{C\}$. Both rules have a confidence of 0.5. However, B co-occurs with A equally likely over the whole space, while C only co-occurs with A in a single cell. In the following, we describe how the spatial characteristic of a rule can be described.

The Kullback-Leibler (KL) divergence $K(P||Q)$ is a measure to describe how much additional information is needed on average to encode observations of a distribution P using a baseline distribution Q. The measure is zero if the distributions P and Q are equal (no additional information is needed). The more different the distribution P is from Q, the higher is the KL-divergence. We can use the KL-divergence to describe the spatial characteristic of a rule similar to the confidence. The distribution $p_{C_1 \cup C_2}(l)$ describes the density of the co-location induced by the rule $C_1 \to C_2$. The LHS distribution $p_{C_1}(l)$ is the baseline distribution. The KL-divergence can then be understood as the similarity of the induced co-location distribution, conditional to the LHS distribution. If the divergence is zero, the rule is valid equally likely at all locations where C_1 occurs. If the divergence is a high number, the rule is only valid at some locations where C_1 occurs. Note that $p_{C_1}(l)$ has a density greater zero at all locations where $p_{C_1 \cup C_2}(l)$ has a density greater zero; this is a necessary precondition to compute the KL-divergence.

$$KL(C_1 \cup C_2||C_1) := - \sum_{\substack{l_i: \\ p_{C_1 \cup C_2}(l_i)>0}} p_{C_1 \cup C_2}(l_i) \log \frac{p_{C_1 \cup C_2}(l_i)}{p_{C_1}(l_i)} \qquad (6)$$

C. Discussion on Measures

The relationship between a frequency measure of co-location patterns and the entropy is important. Assume we choose a bandwidth of our grid so small that each row instance falls into a distinct cell. Then, the number of cells having a value is equal to the number of row instances. The entropy will be higher the more cells have a value. Hence, the entropy then is proportional to the frequency measure.

Now if we increase the bandwidth, co-location instances in close spatial proximity fall into the same cell. The entropy will then be lower because the instances fall into a smaller number of cells. Assume the case where all instances fall into a single cell. We might still have a high frequency (because inside the cell many co-location instances occur), but an entropy of zero, stating that the co-location is clustered in a single cell.

The same observation is true for the KL-divergence. If each row instance falls into a distinct cell, the confidence of a rule will be anti-proportional to the KL-divergence. Consequently, to identify meaningful spatial characteristics, the measures of patterns and rules should be determined with a resolution much higher than the neighborhood distance threshold.

IV. Implementation Considerations

In general, spatial interestingness measures can be tracked during the mining process or they can easily be computed on the basis of the results. To make this clear we need to recall that each co-location instance represents a spatial point, and that a co-location pattern represents a set of points. A spatial interestingness measure of a pattern is based on statistics of the point set. Hence, we need to extend the patterns by data structures that are able to update the pattern statistics on the basis of the instance points. Then, we can compute the spatial interestingness measure based on these statistics. Hence, the main extension of a co-location miner is to use proper data structures to capture the point set statistics of the patterns. In the following we give a short description of examples for three different kinds of measures.

A. Area Statistics

An area measure of a co-location $area(C)$ describes the area that is covered by the co-location instances. Several approaches, probabilistic or geometric ones, to derive such a measure are possible. Geometric methods include the computation of a bounding box for each pattern C, captured in a simple two point data structure (southwest and northeast point). Each new instance extends the southwest and/or northeast point with a constant runtime complexity $O(\log(1))$. Similarly, the convex hull can easily be tracked by a point set data structure. Each instance is then checked to see if it is already included in the convex hull or if it should extend the set. This operation has the runtime complexity $O(n \log(n))$. For both approaches, we can efficiently compute the area covered by the pattern.

A probabilistic approach to derive the area is to compute the sufficient statistics of a Gaussian for the instance points of a pattern. These are the number of points n, the linear sum vector $ls \in \mathbb{R}^2$, and the squared sum matrix $ss \in \mathbb{R}^{2\times 2}$. From the tuple (n, ls, ss) it is possible to efficiently compute the covariance Σ of the Gaussian described by the points. Then, a confidence interval (e.g., 95%) can be used to compute the area in which points of this set fall, given a confidence threshold.

B. Clustering Statistics

A different measure is the amount of clustering a pattern shows. In the simple case of checking if the points cluster at a single location the area statistics are tracked. To check if the point set shows several clusters a promising statistic is the cluster feature tree [14], computed online for each pattern. A cluster feature tree describes clusters by sufficient statistics of a Gaussian, iteratively computed for each new incoming point. Setting up the tree for a number of n points has runtime complexity of $O(n \log(n))$. The clustering then can be described, given a branch-width and a distance split parameter, by the number of emerging clusters.

C. Non-parametric Distribution Statistics

To derive a non-parametric distribution of co-location patterns, e.g., to compute the entropy or the KL-divergence between pairs of patterns, a sparse grid with a given bandwidth can be employed (e.g., a hashmap data structure). Thereby the grid cell counts are updated for each incoming instance. Cells without any counts do not consume memory. The distribution is then easily extracted by normalizing the counts of each cell by the total number of counts.

V. Experiments

In this section, we demonstrate initial results of mining co-location patterns and rules using spatial interestingness measures.

A. Dataset and Computation

As data we use an OSM points-of-interest dataset covering the Los Angeles city area. The dataset consists of $|O| = 5840$ points of interest. The number of features is $|F| = 201$. To compute the pattern entropy and the rule KL-divergence we use two grids: (a) A low resolution grid with a bandwidth of 4.37 km resulting in a 11×7 grid, and (b) a high resolution grid with a bandwidth of 0.48 km resulting in a 101×59 grid. To compute the co-location patterns we use an FP-Growth implementation on a transactionalized dataset using support-based pruning. The neighborhood transactions are naively generated for each point of interest using a distance threshold of 0.11 km. This approach overestimates the frequency of co-location patterns with frequent features compared to an event-based approach. However, the results still show frequent co-locations. As analytically shown in this work, the measures will also identify spatial characteristics when an event-based approach is used. We plan to compare the measures with such an approach in our ongoing work.

B. Entropy Measure

Mined co-location patterns in the OSM dataset are shown in Table 4. The table shows the co-location patterns ordered by decreasing support, decreasing entropy using the high resolution grid, and decreasing entropy using the low resolution grid. As explained in Section III-C we expect the entropy of the patterns to be more similar to the support measure using a high resolution grid for density estimation. The results show that some patterns change their order with respect to the support ordering because they are more or less uniformly distributed. For example, the pattern {*fuel, fast-food*} gets a higher position because it is more uniformly distributed than some patterns above in the support ordered list. More interestingly we see new patterns being in the top-10 list. Using the high resolution grid there are three distinct patterns. For instance, {*library, place-of-worship*} is a pattern occurring rather uniformly but is not in the top-10 support list. Using the low resolution grid the results become even less similar to the support ordered list. E.g., the pattern {*library, park*} is considered. The results show that different patterns are identified if we use a distribution-based measure. In this example, we consider patterns as interesting that are most uniformly distributed, exhibiting a pattern valid in the whole area of interest.

C. KL-Divergence Measure

The rules mined from the OSM dataset are listed in Table 5. The table only shows rules having *cafe* on the LHS, ordered by decreasing confidence, increasing KL-divergence

		by support			by entropy (high res)			by entropy (low res)
sup	entr		sup	entr		sup	entr	
0.051	0.023	school place-of-worship	0.051	0.023	school place-of-worship	0.051	0.464	school place-of-worship
0.040	0.006	fast-food restaurant	0.021	0.007	convenience fuel	0.017	0.234	fuel fast-food
0.034	0.005	cafe restaurant	0.017	0.006	fuel fast-food	0.010	0.233	**library place-of-worship**
0.024	0.003	cafe fast-food	0.040	0.006	fast-food restaurant	0.021	0.229	convenience fuel
0.021	0.007	convenience fuel	0.034	0.005	cafe restaurant	0.005	0.192	**fire-station place-of-worship**
0.019	0.002	bank restaurant	0.010	0.004	**library place-of-worship**	0.007	0.175	**park school**
0.017	0.006	fuel fast-food	0.014	0.004	supermarket fast-food	0.011	0.173	**convenience fast-food**
0.017	0.002	cafe fast-food restaurant	0.007	0.004	**level-crossing place-of-worship**	0.040	0.170	fast-food restaurant
0.014	0.004	supermarket fast-food	0.011	0.004	**convenience fast-food**	0.006	0.167	**park place-of-worship**
0.013	0.002	bank fast-food	0.024	0.003	cafe fast-food	0.005	0.161	**library park**

Figure 4. Co-location pattern results of OSM dataset. Table shows patterns ordered by decreasing support, entropy (high resolution), and entropy (low resolution). Patterns in boldface in the entropy ordered lists are not in the support ordered list. Ordering by decreasing entropy means the patterns are ordered by increasing peaky distribution.

			by conf				by kldiv (high res)				by kldiv (low res)
sup	conf	kldiv		sup	conf	kldiv		sup	conf	kldiv	
0.034	0.645	0.304	cafe → restaurant	0.034	0.645	0.304	cafe → restaurant	0.034	0.645	0.178	cafe → restaurant
0.024	0.451	0.605	cafe → fast-food	0.024	0.451	0.605	cafe → fast-food	0.024	0.451	0.339	cafe → fast-food
0.017	0.319	0.869	cafe → restaurant fast-food	0.017	0.319	0.869	cafe → restaurant fast-food	0.012	0.232	0.560	cafe → bank
0.012	0.232	1.071	cafe → bank	0.012	0.232	1.071	cafe → bank	0.017	0.319	0.568	cafe → restaurant fast-food
0.010	0.200	1.239	cafe → bank restaurant	0.010	0.200	1.239	cafe → bank restaurant	0.004	0.083	0.624	**cafe → pharmacy**
0.008	0.151	1.387	cafe → bank fast-food	0.008	0.151	1.387	cafe → bank fast-food	0.004	0.087	0.666	**cafe → museum**
0.007	0.141	1.442	cafe → bank restaurant fast-food	0.007	0.141	1.442	cafe → bank restaurant fast-food	0.010	0.200	0.754	cafe → bank restaurant
0.005	0.109	1.899	cafe → convenience	0.003	0.061	1.742	**cafe → school**	0.003	0.071	0.790	**cafe → restaurant pharmacy**
0.005	0.103	1.933	cafe → station	0.003	0.064	1.842	**cafe → fuel**	0.003	0.064	0.880	**cafe → place-of-worship**
0.005	0.096	2.021	cafe → pub	0.004	0.083	1.884	**cafe → pharmacy**	0.002	0.041	0.900	**cafe → bank pharmacy**

Figure 5. Co-location rule results. Table shows rules ordered by decreasing confidence, increasing KL-divergence (high resolution), and increasing KL-divergence (low resolution). Rules in boldface in the KL-divergence ordered lists are not in the confidence ordered list. Ordering by increasing KL-divergence means the rules are ordered by increasing spatial dissimilarity.

using the high resolution grid, and increasing KL-divergence using the low resolution grid. As in the pattern example, the KL-divergence list shows a reordering and the emergence of new rules based on the similarity of the rule distribution to its LHS. As can be seen for the low resolution grid, e.g., the rules with *pharamacy, museum, place-of-worship* on the RHS are similar to the *cafe* distribution. Hence they are interesting co-location rules at all location where *cafe*s occur, even if they have a rather low frequency.

VI. Conclusions and Ongoing Work

In this paper, we introduced a class of interestingness measures that are based on the spatial distribution of co-location patterns. The measures are based on density estimations of the instance locations of co-location patterns. We showed that patterns can easily be extended by data structures capturing instance point statistics, which can in turn be used to compute spatial interestingness measures. The defined entropy and KL-divergence measures are promising candidates to identify patterns based on global and local characteristics. We plan to define more spatial measures and evaluate their applicability for spatial analysis using large-scale real world datasets.

References

[1] Y. Morimoto, "Mining frequent neighboring class sets in spatial databases," *Proc. KDD*, 353–358, 2001.

[2] Y. Huang, S. Shekhar, and H. Xiong, "Discovering colocation patterns from spatial data sets: A general approach," *IEEE TKDE*, vol. 16, no. 12, 2004.

[3] P. Mohan, J. A. Shine, J. P. Rogers, and N. Wayant, "A neighborhood graph based approach to regional co-location pattern discovery: A summary of results," *Proc. ACM SIGSPATIAL*, 122–132, 2011.

[4] M. Celik, J. M. Kang, and S. Shekhar, "Zonal co-location pattern discovery with dynamic parameters," *Proc. ICDM*, 433–438, 2007.

[5] C. F. Eick, R. Parmar, W. Ding, T. F. Stepinski, and J.-P. Nicot, "Finding regional co-location patterns for sets of continuous variables in spatial datasets," *Proc. ACM SIGSPATIAL*, pp. 30, 2008.

[6] Y. Huang, H. Xiong, S. Shekhar, and J. Pei, "Mining confident co-location rules without a support threshold," *Proc. ACM SAC*, 497–501, 2003.

[7] Y. Huang, J. Pei, and H. Xiong, "Mining co-location patterns with rare events from spatial data sets," *Geoinformatica*, vol. 10, no. 3, 2005.

[8] J. S. Yoo and S. Shekhar, "A joinless approach for mining spatial colocation patterns," *IEEE TKDE*, vol. 18, no. 10, 2006.

[9] E. Desmier and D. Gay, "A clustering-based visualization of colocation patterns," *Proc. IDEAS*, 70–78, 2011.

[10] L. Wasserman, *All of Statistics*, Springer, 2004.

[11] J. Dean and S. Ghemawat, "MapReduce: Simplifed data processing on large clusters," *Proc. OSDI*, 137–150, 2004.

[12] Y. Huang, L. Zhang, and P. Yu, "Can we apply projection based frequent pattern mining paradigm to Spatial Co-location Mining?" *Proc. PAKDD*, 719–725, LNCS 3518, 2005.

[13] Y. Wang and E. Y. Chang, "PFP : Parallel FP-Growth for query recommendation," *Proc. ACM RecSys*, 107–114, 2008.

[14] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," *Proc. ACM SIGMOD*, 103–114, 1996.