Geo-Social Co-location Mining

Michael Weiler* Klaus Arthur Schmid* Nikos Mamoulis+ Matthias Renz*
*Institute for Informatics, Ludwig-Maximilians-Universität München
+Department of Computer Science, University of Hong Kong
{weiler,schmid,renz}@dbs.ifi.lmu.de
nikos@cs.hku.hk

ABSTRACT

Modern technology to capture geo-spatial information produce a huge flood of geo-spatial and geo-spatio-temporal data as a new user mentality of utilizing this technology to voluntarily share information. This location information, enriched with social information, is a new source to discovery new and useful knowledge. This work introduces geo-social co-location mining, the problem of finding social groups that are frequently found at the same location. This problem has applications in social sciences, allowing to research interactions between social groups and permitting social-link prediction. It can be divided into two sub-problems. The first sub-problem of finding spatial co-location instances, requires to properly address the inherent uncertainty in geo-social network data, which is a consequence of generally very space check-in data, and thus very space trajectory information. For this purpose, we propose a probabilistic model to estimate the probability of a user to be located at a given location at a given time, creating the notion of probabilistic co-locations. The second sub-problem of mining the resulting probabilistic co-location instances requires efficient for large databases having a high degree of uncertainty. Our approach solves this problem by extending solutions for probabilistic frequent itemset mining. Our experimental evaluation performed on real (but anonymized) geo-social network data shows the high efficiency of our approach, and its ability to find new social interactions.

1. INTRODUCTION

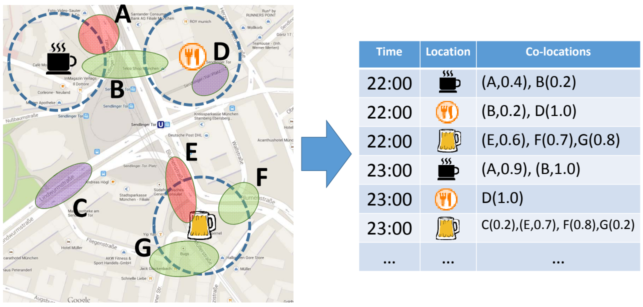
Spatial features describe the presence or absence of geographic object types at different locations. Examples of spatial features include plant species, animal species, road types, cancers, crime, and business types, or features of individuals, such as personal preferences, or simply their id. A spatial co-location pattern represents a subset of spatial features whose instances are frequently located in a spatial neighborhood. For example, “botanists may have found that there are orchids in 80% of the area where the middle-wetness green-broad-leaf forest grows” (example taken from [26]). Spatial co-location patterns may yield important insights for many applications. For example, a mobile service provider may be interested in services frequently requested by geographical neighbors, and thus gain sales promotion data. Other application domains include Earth science, public health, biology, transportation and geo-social networks. Traditional solutions for the problem of frequent co-location mining [26] considers classical spatial data, where each data record has a (certain) spatial location.

In this project, which we wish to discuss with a broad audience at GeoRich'15, we want to take the problem of spatial co-location mining into a new context, by considering spatio-temporal data, i.e., trajectory data of individuals. Thus, the problem now is to find groups of users which frequently co-locate in geo-space over time, creating the notion of geo-social co-location mining. There is already an abundance of public data sets that can be mined, including data sets from geo-social networks [7] and from social networks using geo-tags such as Twitter. Frequent co-location mining on such data may yield interesting patterns, such as “Members of LMU and HKU are frequently to be found at the same location, while members of some other university are often found in solitude or among themselves”. In such an application, each instance of a co-location corresponds to a (l, t, S) triple, where S denotes the set of individuals that have been at the same location l at the same time t. The problem of geo-social co-location mining introduces two major new challenges which have not been sufficiently covered in existing work on traditional co-location mining. Firstly, the temporal dimension leads to very large sets of co-location instances, since every location and time pair leads to a possibly non-empty co-location instance, secondly existing solutions do not consider the uncertainty which is inherent in spatial data: Spatial data may be imprecise (e.g., due to measurement errors), data can be obsolete (e.g., when the most recent position update is already minutes old), data may originate from unreliable sources (such as crowd-sourcing), or it may be blurred to prevent privacy threats and to protect user anonymity [8]. For example, the oval regions in Figure 1 may correspond to individual persons, while the color of each person may represent the individual's affiliations. Here, the location of each person is a conservative approximation based on the users GPS history. It is important to note that we are considering historic data. Thus, for a given point of time t, both past and future GPS positions of a user may be available.[1] Given these approximations, it becomes possible to estimate which point of interest each user is currently visiting, yielding probability distribution as shown in the table in Figure 1 for depicted point of time (22:00) and for a point of time one hour later.

Given such data, we can immediately envision a number of useful applications:

- Find groups of people often co-locating. In the setting de-

[1] A probabilistic model to estimate the position of a mobile user given past and future observations can be found in [21].



Time	Location	Co-locations
22:00	☕	(A,0.4), B(0.2)
22:00	🍴	(B,0.2), D(1.0)
22:00	🍺	(E,0.6), F(0.7),G(0.8)
23:00	☕	(A,0.9), (B,1.0)
23:00	🍴	D(1.0)
23:00	🍺	C(0.2),(E,0.7), F(0.8),G(0.2)
...	...	...

Figure 1: Spatial co-location mining in uncertain spatio-temporal data.

scribed above, two individuals being located at the same location may not actually be there together, but if the two individuals co-locate very often, it becomes highly unlikely that their co-locations are independent random events.

- Find groups of people visiting the same types of points of interest, even not at the same time. This allows to cluster user's by their points of interest, thus allowing to predict new locations that a user might find interesting, based on other users in the same cluster.
- Classify points of interest by the people that visit these. For example, if a point of interest that is unlabeled is being visited by a significant fraction of user's which (individually or even together) visit Italian restaurants, then it might be possible to predict the label of this point of interest.
- Visiting a new city, such as Melbourne, new people, that are similar to people you hang out with at home, and new point of interest, that are similar locations you visit at home, can be predicted to you and to the people that you now hang out with in Melbourne.

2. PROBLEM DEFINITION

In traditional co-location mining, the location of an object is known for certain. Under this assumption, a lot of work has been published in the last decade [30, 13, 28, 22, 12, 30]. A survey on the field of co-location mining on certain spatial data be found in [19, 20]. However, in many real applications such as plant disease diagnosis, environmental surveillance and geo-social networks, the location of objects is uncertain. In the following, the problem of probabilistic spatial collocation mining on uncertain spatial data is defined.

To formally define the problem of spatial co-location mining, we first have to define the concept of a spatial co-location. In this work, a neighbor relation will be used that is particularly important in social network applications. This relation uses a set of interesting spatial locations, such as bars, restaurants and football stadiums. Two individuals are co-located if they are sufficiently close to the same location, formally.

DEFINITION 1. *Let $\mathcal{L}$ be a set of spatial locations, and let $\mathcal{D}$ be a database of spatial objects. The neighbor relation $\mathcal{R}$ is defined as follows*

$$(o_i \in \mathcal{D}, o_j \in \mathcal{D}) \in \mathcal{R} \Leftrightarrow \exists l \in \mathcal{L} : dist(o_i, l) \leq \epsilon \wedge dist(o_j, l) \leq \epsilon$$

An example of the problem of frequent co-location mining in uncertain spatial data using the neighborhood relation of Definition 1 is given in the following.

EXAMPLE 1. *Consider uncertain positions of individuals in a geo-social network application. The task is to find groups of people that commonly spend time at the same locations, in order to predict missing social links in the underlying social network, or in order to direct special deals to such groups. Figure 1 exemplarily shows the position of a individuals $A, ..., G$, together with three locations: a café, a restaurant and a bar. For simplicity, each of these locations is represented by an oval region, but in practice, these uncertainty region can have arbitrary shapes [21]. It is not possible to tell for certain, whether user A is located inside the café, or just barely outside of it. In contrast, user D is certainly inside the restaurant, while user C is certainly outside all three places. The probability $P(U \text{ in } l)$ that a user U is located inside a location l can be computed using techniques for range queries on uncertain data [5].*[2] *Exemplary probabilities $P(U \text{ in } l)$ for all users U and all locations l are shown in the table of Figure 1. At the time stamp 22:00, the users E, F and G are co-located at the bar with a high probability. However, at time 23:00, user G is likely no longer located with users E and F. In contrast, it is likely that the group consisting of users B and D remained together, in the restaurant.*

Clearly, the number of co-locations may be extremely large, since in an application like this, there may be one non-empty co-location for each combination of time stamp and location. The task of probabilistic co-location mining is to find groups of users (objects), having a significantly high probability of having spent time at the same location for a sufficiently large number of times. Formally, the problem of probabilistic frequent spatial Co-location mining is defined as follows.

DEFINITION 2. *Given a set $\mathcal{F} = \{f_1, ..., f_k\}$ of k spatial features, given a database $\mathcal{D} = \{o_1, ..., o_N\}$ of N uncertain spatial objects each having a set $f(o_i \in \mathcal{D}) \subseteq F$ of spatial features, and given a positive integer $minSup$ and a probability threshold τ, a probabilistic frequent spatial co-location mining algorithm returns all sets $S \subseteq F$ of features such that the probability there exists at least $minSup$ spatial co-locations instances I such that $S \subseteq \bigcup_{o \in I} f(o)$ is at least τ.*

To find probabilistic frequent spatial co-locations, consider the following example.

EXAMPLE 2. *Returning to Example 1 assume that $minSup = 2$ and $\tau = 0.5$ and consider the spatial features $F = \{$red, green, purple$\}$, depicted by the corresponding colors in Figure 1, which may e.g., correspond to the affiliations of mobile users. In this example, we have two possible co-locations instances of features red and green, in the café and in the bar at times 22:00 and 23:00. Assuming independence between uncertain objects*[3]*, the probability of a co-location of green and red at time 22:00 can be computed by the product of marginal probabilities $P(A \wedge B) = P(A) \cdot$*

[2] For the case proximity to a location is not modelled by a circle, an adaption of the techniques in Section 4 can be made easily, by replacing distance calculation by intersection tests between points and polygons.

[3] We argue that in many applications, this assumption holds true. Note that the position of mobile objects can be strongly correlated, as for example friends are more likely to travel together. However, the assumption that measurement errors are mutually independent does often hold. For example, GPS errors between different devices can be assumed to be independent, and uncertainty regions the are added deliberately for privacy preservation should be independent as well. Nevertheless, this assumption of independent random variables of spatial locations can be a base for discussions at GeoRich'15 workshop.

$P(B) = 0.4 \cdot 0.2 = 0.08$. *At the bar at time 22:00, the probability of a co-location between red and green can be computed by* $P(E \wedge (F \vee G)) = P(E \wedge \neg(\neg F \wedge \neg G)) = P(0.6 \cdot (1 - 0.3 \cdot 0.2)) = 0.564$. *At time 23:00, can obtain the co-location probabilities red and green at the café and the bar of* 0.9 *and* 0.588, *respectively. Given these probabilities, we can compute the probability that at least* $minSup = 2$ *co-location instances exist by applying the generating functions technique of [16, 17], yielding a probability of* 0.778 *which is greater than* $\tau = 0.5$. *Thus the set of spatial features red and green will be returned as a probabilistic frequent co-location.*

In the following section, we propose solutions to compute the probabilities of probabilistic frequent co-locations efficiently.

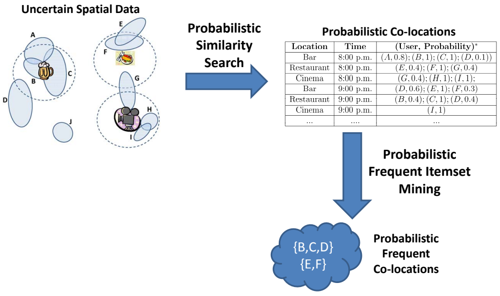
3. RELATED WORK

Traditional co-location mining on (certain) spatial data has been studied in the past [27, 30, 12]. These works define a spatial neighborhood relation on pairs of objects not exceed a given distance threshold. Due to the assumption of certain objects, the works can solve the problem of frequent co-location mining by applying traditional frequent pattern mining solutions such as Apriori-algorithm [2] combine the discovery of spatial neighborhoods with the mining process.

The problem of probabilistic co-location mining in uncertain spatial data is related to the problem of frequent itemset mining in uncertain transaction databases. Existing solutions for this problem transform uncertain items into certain ones by thresholding the probabilities. For example, by treating all uncertain items with a probability value higher than 0.5 as being present, and all others as being absent in a transaction. Such an approach loses useful information and leads to inaccuracies. Existing approaches in the literature are based on expected support ([9, 10, 1]). Itemsets are considered frequent if the expected support exceeds *minSup*. Effectively, this approach returns an estimate of whether an object is frequent or not with no indication of how good this estimate is. Since uncertain transaction databases yield uncertainty w.r.t. the support of an itemset, the probability distribution of the support and, thus, information about the confidence of the support of an itemset is very important. This information, while present in the database, is lost using the expected support approach.

There is a large body of research on Frequent Itemset Mining (FIM) but very little work addresses FIM in uncertain databases [9, 10, 15]. The approach proposed by Chui et. al [10] computes the expected support of itemsets by summing all itemset probabilities in their U-Apriori algorithm. Later, in [9], they additionally proposed a probabilistic filter in order to prune candidates early. In [15], the UF-growth algorithm is proposed. Like U-Apriori, UF-growth computes frequent itemsets by means of the expected support, but it uses the FP-tree [11] approach in order to avoid expensive candidate generation. In contrast to our probabilistic approach, itemsets are considered frequent if the expected support exceeds *minSup*. The main drawback of this estimator is that information about the uncertainty of the expected support is lost; [9, 10, 15] ignore the number of possible worlds in which an itemsets is frequent. [29] proposes exact and sampling-based algorithms to find likely frequent items in streaming probabilistic data. However, they do not consider itemsets with more than one item. To the best of our knowledge, our approach in [3] was the first that is able to find frequent itemsets in an uncertain transaction database in a probabilistic way.

However, this publication has stimulated research on the field of probabilistic mining of frequent itemsets in uncertain transaction data, creating a large number of follow up publications. A detailed survey can be found in [23]. In [24, 25], an approach is presented to approximate the support PDF of an itemset using a Poisson distribution. Approach yields a very small error if the database is sufficiently large. This approximation furthermore allows to compute the support PDF of an item much faster than the exact approach presented in [3] and in this chapter. The idea of [24, 25] is used to study a variety of approximation techniques in Chapter 4, including Expected support, Normal approximation and Poisson approximation. An approach to accelerate the computation of our approach in [3] was presented by [14], using massive parallelization exploiting GPGPU (General-Purpose computation on GPU). Furthermore, the related problem of mining frequent subgraphs over uncertain graphs [18, 32, 31] has gained alot of research interest in the last years. Finally, an approach for probabilistic frequent itemset mining on uncertain data avoiding multiple database scans incurred by the candidate generation step of [3] has been proposed by us in [4].


Uncertain Spatial Data
A
B
C
D
E
F
G
H
I
J
Probabilistic Similarity Search
Probabilistic Co-locations
Probabilistic Frequent Itemset Mining
{B,C,D}
{E,F}
Probabilistic Frequent Co-locations

Location	Time	(User, Probability)*
Bar	8:00 p.m.	$(A, 0.8); (B, 1); (C, 1); (D, 0.1))$
Restaurant	8:00 p.m.	$(E, 0.4); (F, 1); (G, 0.4)$
Cinema	8:00 p.m.	$(G, 0.4); (H, 1); (I, 1);$
Bar	9:00 p.m.	$(D, 0.6); (E, 1); (F, 0.3)$
Restaurant	9:00 p.m.	$(B, 0.4); (C, 1); (D, 0.4)$
Cinema	9:00 p.m.	$(I, 1)$
...	....	...

Figure 2: Workflow of probabilistic spatial co-location mining.

Only recently, the research community has tackled the challenge of spatial co-location mining in uncertain data. Recent work ([26]) considers existential uncertainty in spatial data. In this model, each object has a probability to be present in the database. The solution of [26] has a run-time polynomial in the number of possible worlds, thus exponential in the number of uncertain objects. The reason for this high complexity is the neighborhood relation $R(.,.)$ used in [26] is arbitrary, i.e., this approach can be applied to any neighborhood relation. This fact makes efficient co-location mining hard: For three uncertain objects A, B and C, the predicates $R(A, B)$ and $R(B, C)$ are stochastically dependent, despite the assumption of independence between objects.

4. PROBABILISTIC FREQUENT CO-LOCATION MINING

In a nutshell, the problem of probabilistic co-location mining requires two subtasks to be solved, as illustrated in Figure 2:

- First, for each location l and each time interval t, probabilistic instances have to be computed and derived. This requires to compute the probabilities of all objects, to be close to location l at time t. This task requires to utilize probabilistic similarity search methods on uncertain spatial data to derive the probability that a given object is a member of a co-location instance. For the neighbor relation given in Definition 1, this step requires to perform probabilistic range queries, using the locations $\mathcal{L}$ as query points. As a result of the first step, an uncertain spatial database is transformed into a probabilistic co-location database such as depicted in Figure 4.

- Second, all probabilistic co-location instances need to be mined in order to detect subsets of spatial features having a statistically significantly high probability to be co-located fre-

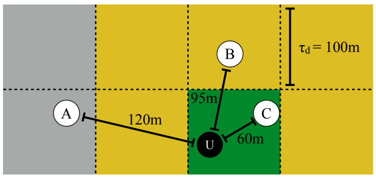


Figure 3: Grid based spatial index for efficient nearest neighbour queries

quently in the database. For this subtask, we can assume that a database $\mathcal{D}$ of probabilistic co-locations such as featured in Figure 4 is given as a result of solving the first subtask. Given such a database, the task of finding probabilistic frequent co-locations in such a database is equivalent to the problem of probabilistic frequent itemset mining [3] in uncertain transaction data. Both problems, of probabilistic mining of spatial co-locations in uncertain spatial data, as well the the problem of probabilistic frequent itemset mining in uncertain transaction data, are formally defined in the following.

4.1 Occurrence Probability Estimation

At each time interval t we estimate the probability of a user u being at a certain location l. Therefore we first determine the geographical distance $d_{u,l}$ by using the Haversine formula defined as follows:

$$2r \arcsin \sqrt{\sin^2\left(\frac{\phi_l - \phi_u}{2}\right) + \cos(\phi_u)\cos(\phi_l)\sin^2\left(\frac{\lambda_l - \lambda_u}{2}\right)}$$

where ϕ_u, ϕ_l represent the latitude of the coordinates of user u and location l and λ_u, λ_l the longitude respectively. The probability $P_{t,u,l}$ that a user u occurs at location l at time interval t is then given by

$$P_{u,l} = \frac{\rho\left(d_{u,l}\right)}{\sum\limits_{l_t \in L_{u,t}} \rho\left(d_{u,l_t}\right)}, \rho\left(d_{u,l}\right) = \begin{cases} 0 & d_{u,l} \geq \tau_d \\ \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} & d_{u,l} < \tau_d \end{cases}$$

where ρ is the density (PDF) regarding a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with $\mu = 0$ and $\sigma^2 = \frac{\tau_d}{3}$. Thereby τ_d denotes a distance threshold parameter (e.g. 100 meters) to cut the long tail of the probability distribution to 0 for each user with a distance $d > \tau_d$. We utilize this threshold in our implementation for a simple yet effective spatial index based on grid cells of the size τ_d. For a latitude ϕ and longitude λ of a user or location we determine the corresponding (x, y) grid cell as such $\left(\left\lfloor \frac{\phi}{tau_d} \right\rfloor, \left\lfloor \frac{\lambda}{tau_d} \right\rfloor\right)$. An example is shown in Figure 3: For the user U only locations within the neighbour cells ± 1 around the user's cell are candidates. Thereby locations that share the same cell with the user (e.g. location C) are certain hits because their distance must smaller than τ_d. Locations like A which are at least 1 complete cell away must not be considered as their distance to the user must be greater than τ_d.

4.2 Transformation to Probabilistic Frequent Itemset Mining

The definition of a uncertain co-location database can be mapped to the definition of an uncertain transaction database defined in [3].

Definition 3 (Uncertain Transaction Database). *Let I be a set of items. An* uncertain transaction *database $\mathcal{T}$ is a set of probabilistic transactions. Each transaction $T = (i \in I, P(i)) \in \mathcal{T}$ contains a set of items, each associated with a probability. For each pair $(i \in I, P(i))$, the probability $P(i)$ describes the likelihood that i is present in the probabilistic transaction T.*

This equivalence between Definition 2 and Definition 3 allows to interpret the problem of probabilistic frequent co-location mining in uncertain spatial data, as the problem of probabilistic frequent itemset mining in uncertain transaction data. This can be done by interpreting a spatial feature as an item or a probabilistic co-location instance as a transaction. Thus, solutions for the problem of probabilistic frequent item-set mining can now be applied. In fact, a large body of efficient algorithms (e.g. [24, 25, 14, 6]) have been proposed for the problem definition of [3]. Yet, a main common problem of these works is the lack of a real world application for the problem of probabilistic frequent itemset mining. We argue, that probabilistic frequent itemset mining and probabilistic spatial co-location mining can bridge this gap, thus providing spatial applications. In the following subsections, we will briefly outline a mapping of existing solutions to the problem of probabilistic co-location mining. Firstly, as a baseline a naive solution is presented, omitting the uncertainty information. Then, the exact solutions of [3] is reviewed and mapped to co-location mining. Finally, the same is done for the approximate solutions of WanCheLee10. For these solutions, an initial experimental run-time evaluation is presented in Section 5, by using a real-world data set consisting of geo-tagged tweets.

4.2.1 Naive Probabilistic Co-Location Mining

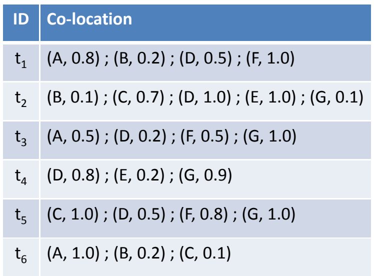
One *naive* approach is to transform an uncertain database into a non-uncertain database by setting the item probabilities to 0 or 1 and then applying a traditional frequent itemset detection method. For example, probabilities less then 0.5 could be mapped to 0 and probabilities above 0.5 could be mapped to 1. However, such a transformation obviously involves loss of information and accuracy. Furthermore, we would have no idea how confident we could be in the results. In particular, itemsets that are often associated with probabilities close to 0.5 yield a very large error in the result. Another approach is to use the probabilities associated with the itemsets in order to compute the expected support of an itemset.

To avoid incurring a biased result, previous work was based on the expected support [9, 10, 15], i.e., the expected number of spatial co-locations of a group of spatial features.

Definition 4. *Given a set $\mathcal{F}$ of spatial features and a database of co-location instances I, the* expected support *$E(X)$ of a set of spatial features $X \subseteq F$ is defined as $E(X) = \sum_{i \in I} P(X \subseteq i)$.*

The expected support of set of spatial features X can be efficiently computed by a single scan over all co-location instances. An itemset is considered frequent if its expected support is above *minSup*. However, the later step has the major drawback that the uncertainty information is forfeited when using the expected support approach. Thus, information is lost about the likelihood that X is frequent.

Example 3. *As an example, consider the database depicted in Figure 4, containing a set of uncertain co-location instances. Treating each co-location instance as a transaction, the expected support of the itemset $\{D\}$ is $E(\{D\}) = 3.0$. The fact that $\{D\}$ occurs for certain in one transaction, namely in t_2, and that there is at least one possible world where X occurs in five transactions are totally ignored when using the expected support in order to evaluate the frequency of an itemset. Indeed, suppose $minSup = 3$; do we call $\{D\}$ frequent? And if so, how certain can we even be that $\{D\}$ is frequent? By comparison, consider itemset $\{G\}$.*



ID	Co-location
t_1	(A, 0.8) ; (B, 0.2) ; (D, 0.5) ; (F, 1.0)
t_2	(B, 0.1) ; (C, 0.7) ; (D, 1.0) ; (E, 1.0) ; (G, 0.1)
t_3	(A, 0.5) ; (D, 0.2) ; (F, 0.5) ; (G, 1.0)
t_4	(D, 0.8) ; (E, 0.2) ; (G, 0.9)
t_5	(C, 1.0) ; (D, 0.5) ; (F, 0.8) ; (G, 1.0)
t_6	(A, 1.0) ; (B, 0.2) ; (C, 0.1)

Figure 4: Example of an uncertain co-location database.

This also has an expected support of 3*, but its presence or absence in transactions is more certain. It turns out that the probability that* $\{D\}$ *is frequent is* 0.7 *and the probability that* G *is frequent is* 0.91*. While both have the same expected support, we can be quite confident that* $\{G\}$ *is frequent, in contrast to* $\{D\}$*. An expected support based technique does not differentiate between the two.*

Concepts to evaluate the co-location instances in a probabilistic way are presented in the following.

4.2.2 Exact Probabilistic Support

A co-location is a *frequent co-location* if it occurs in at least *minSup* co-location instances, where *minSup* is a user specified parameter. The number of instances of a co-location is denoted as the support $supp(S)$ of S. In uncertain co-location databases however, the support of a co-location is uncertain; it is defined by a discrete probability distribution function (PDF).

DEFINITION 5 (PROBABILISTIC SUPPORT). *Let* $\mathcal{D}$ *be an uncertain co-location database and let* $X \subseteq \mathcal{F}$ *be a set of spatial features. The support of* X *is a probability density function*

$$supp(X) : IN_0 \to [0,1]$$

$$n \mapsto P(supp(X) = n).$$

that maps each non-negative integer n *to the probability that the support of features* X *equals* n*.*

Therefore, each set of spatial features has a *frequentness probability* – the probability that it is frequent.

The number of possible worlds $|W|$ that need to be considered for the computation of $P_i(X)$ is extremely large. In fact, we have $O(2^{|T|\cdot|I|})$ possible worlds, where $|I|$ denotes the total number of items. In the following, we show how to compute $P_i(X)$ without materializing all possible worlds [3].

LEMMA 1. For an uncertain transaction database T with mutually independent transactions and any $0 \leq i \leq |T|$, the support probability $P_i(X)$ can be computed as follows:

$$P_i(X) = \sum_{S \subseteq T, |S|=i} (\prod_{t \in S} P(X \subseteq t) \cdot \prod_{t \in T-S} (1 - P(X \subseteq t))) \quad (1)$$

Note that the transaction subset $S \subseteq T$ contains exactly i transactions.

PROOF. The transaction subset $S \subseteq T$ contains i transactions. The probability of a world w_j where all transactions in S contain X and the remaining $|T - S|$ transactions do not contain X is $P(w_j) = \prod_{t \in S} P(X \subseteq t) \cdot \prod_{t \in T-S}(1 - P(X \subseteq t))$. The sum of the probabilities according to all possible worlds fulfilling the above conditions corresponds to the equation given in Definition 5. □

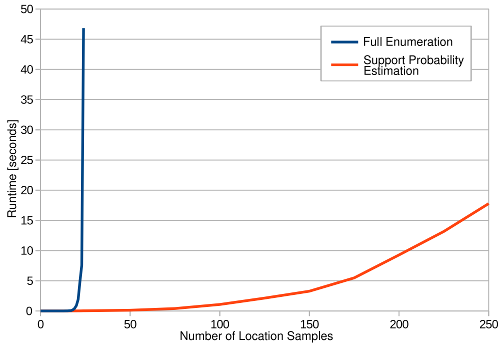


Figure 5: Calculation runtime comparison between enumeration and an estimation of support probability.

4.2.3 Support Probability Estimation

An approximation of the probabilistic support has been proposed in [24]. Here, the idea is to approximate the probabilistic support (cf. Definition 5) by a Poisson distribution. For each set of spatial features X, the single parameter λ of the Poisson distribution $Po(\lambda)$ used to approximate the support distribution of X corresponds to the expected support of X, which can be computed analogously to solutions using expected support (cf. Subsection 4.2.1). Then, the probability that the support of X exceeds $minSup$ can be computed by evaluating the cumulative distribution function of $Po(\lambda)$:

$$P(Po(\lambda) \geq minSup) = 1 - P(Po(\lambda) < minSup) =$$

$$e^{-\lambda} \sum_{i=0}^{minSup-1} \frac{\lambda^i}{i!}.$$

5. EXPERIMENTS

We chose to use data from Twitter[4] to prove the concept of our approach. For this, we collected over 8 Million tweets issued between September 2014 and March 2015 (approx. 1954 per hour) that were geo-tagged within the county of Los Angeles, USA using their public streaming API. Note that the openly available API only reveals one tenth of total tweets, and out of those only the ones carrying a geotag were useful. Therefore we decided to use Los Angeles since the tweet density is fairly high there. We discretized time into slots of one hour – smaller timeslots would have resulted in fewer co-locations, whereas larger values would yield less interesting results, e.g., users a and b patronized the same restaurant within the same day.

We cross-referenced this data with points of interest out of OpenStreetMap[5], out of which around 16 thousand were within the investigated region and of a fitting type (we excluded points like traffic lights or garbage bins). We paired each of these points of interest with all observations (tweets) within their τ_d-meter neighborhood and selected those pairs of PoIs p and timeslots t that contained at least two distinct observations. Each of these 184.452 (p, t)-pairs also specifies a list of the observed users with their respective sojourn probability at p.

For evaluation we implemented an algorithm based on Apriori[2] to estimate support probabilities. Figure 5 shows a performance evaluation against a simple enumeration of user combinations, which becomes practically unusable after surpassing only a

[4] http://www.twitter.com/

[5] http://www.openstreetmap.org/

few observations. For an input of between 10 and 250 distinct observations, we recorded the runtime to calculate support. As the graph shows, a full enumeration exhibits a quick super-exponential growth after about 25 observations, while the estimation approach terminates in interactive time.

Please note that our existing implementation is far from efficient due to the use of inefficient libraries and a slow runtime, since its purpose is merely to show feasibility of our concept.

6. CONCLUSIONS

In this paper we developed an efficient solution for finding probabilistic co-location patterns in uncertain locations, e.g., inaccurate observations derived from social media data. Our solution is mainly based on techniques used for probabilistic frequent itemset mining. In our experiments we show that the proposed methods enable co-location mining in data sets significantly larger than possible using straightforward methods.

7. REFERENCES

[1] C. Aggarwal, Y. Li, J. Wang, and J. Wang. Frequent pattern mining with uncertain data. In *Proc. KDD*, pages 29–38, 2009.

[2] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993.

[3] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Züfle. Probabilistic frequent itemset mining in uncertain databases. In *Proc. KDD*, pages 119–128, 2009.

[4] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Züfle. Probabilistic frequent pattern growth for itemset mining in uncertain databases. In *Scientific and Statistical Database Management*, pages 38–55. 2012.

[5] T. Bernecker, H.-P. Kriegel, M. Renz, and A. Züfle. Hot item detection in uncertain data. In *Proc. PAKDD*, 2009.

[6] T. Calders, C. Garboni, and B. Goethals. Approximation of frequentness probability of itemsets in uncertain data. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 749–754. IEEE, 2010.

[7] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[8] C.-Y. Chow, M. F. Mokbel, and W. G. Aref. Casper*: Query processing for location services without compromising privacy. *ACM TODS*, 34(4):24, 2009.

[9] C. K. Chui and B. Kao. A decremental approach for mining frequent itemsets from uncertain data. In *Proc. PAKDD*, pages 64–75, 2008.

[10] C.-K. Chui, B. Kao, and E. Hung. Mining frequent itemsets from uncertain data. In *In Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-2007), Nanjing, China, May 22-25*, pages 47–58, 2007.

[11] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. *Proc. SIGMOD*, 29:1–12, 2000.

[12] Y. Huang, J. Pei, and H. Xiong. Mining co-location patterns with rare events from spatial data sets. In *Geoinformatica*, volume 10, pages 239–260, 2006.

[13] Y. Huang, S. Shekhar, and H. Xiong. Discovering co-location patterns from spatial data sets: A general approach. In *IEEE TKDE*, volume 16, pages 1472–1485, 2004.

[14] Y. Kozawa, T. Amagasa, and H. Kitagawa. Gpu acceleration of probabilistic frequent itemset mining from uncertain databases. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 892–901, 2012.

[15] C. K.-S. Leung, C. L. Carmichael, and B. Hao. Efficient mining of frequent patterns from uncertain data. In *ICDMW '07: Proceedings of the Seventh IEEE International Conference on Data Mining Workshops*, pages 489–494, 2007.

[16] J. Li, B. Saha, and A. Deshpande. A unified approach to ranking in probabilistic databases. *Proc. VLDB*, 2(1):502–513, 2009.

[17] J. Li, B. Saha, and A. Deshpande. A unified approach to ranking in probabilistic databases. *VLDB Journal*, 20(2):249–275, 2011.

[18] J. Li, Z. Zou, and H. Gao. Mining frequent subgraphs over uncertain graph databases under probabilistic semantics. *The VLDB Journal*, 21:753–777, 2012.

[19] N. Mamoulis. Co-location pattern. In *Encyclopedia of GIS*, page 98. 2008.

[20] N. Mamoulis. Co-location patterns, algorithms. In *Encyclopedia of GIS*, pages 103–107. 2008.

[21] J. Niedermayer, A. Züfle, T. Emrich, M. Renz, N. Mamoulis, L. Chen, and H. Kriegel. Probabilistic nearest neighbor queries on uncertain moving object trajectories. *PVLDB*, 7(3):205–216, 2013.

[22] S. Shekhar and Y. Huang. Co-location rules mining: A summary of results. In *Proc. SSTD*, pages 236–256, 2001.

[23] Y. Tong, L. Chen, Y. Cheng, and P. S. Yu. Mining frequent itemsets over uncertain databases. 5(11):1650–1661, 2012.

[24] L. Wang, R. Cheng, S. D. Lee, and D. Cheung. Accelerating probabilistic frequent itemset mining: a model-based approach. In *Proc. CIKM*, pages 429–438, 2010.

[25] L. Wang, D.-L. Cheung, R. Cheng, S.-D. Lee, and X. Yang. Efficient mining of frequent item sets on large uncertain databases. *Knowledge and Data Engineering, IEEE Transactions on*, 24(12):2170–2183, 2012.

[26] L. Wang, P. Wu, and H. Chen. Finding probabilistic prevalent colocations in spatially uncertain data sets. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):790–804, 2013.

[27] H. Xiong, S. Shekhar, Y. Huang, V. Kumar, X. Ma, and J. S. Yoo. A framework for discovering co-location patterns in data sets with extended spatial objects. In *SDM*, pages 78–89, 2004.

[28] L. Z. Yan Huang and P. Zhang. Finding sequential patterns from a massive number of spatio-temporal events. In *Proc. SDM*, 2006.

[29] Q. Zhang, F. Li, and K. Yi. Finding frequent items in probabilistic data. In *Proc. SIGMOD*, pages 819–832, 2008.

[30] X. Zhang, N. Mamoulis, D. W. Cheung, and Y. Shou. Fast mining of spatial collocations. In *Proc. KDD*, pages 384–393, 2004.

[31] Z. Zou, H. Gao, and J. Li. Discovering frequent subgraphs over uncertain graph databases under probabilistic semantics. In *Proc. KDD*, pages 633–642, 2010.

[32] Z. Zou, J. Li, H. Gao, and S. Zhang. Mining frequent subgraph patterns from uncertain graph data. *Knowledge and Data Engineering, IEEE Transactions on*, 22(9):1203–1218, 2010.