Interactive Probabilistic Post-mining of User-preferred Spatial Co-location Patterns

Lizhen Wang[#], Xuguang Bao[#], Longbing Cao[*]

[#]*School of Information Science and Engineering, Yunnan University, Kunming, P. R. China*

[*]*Advanced Analytics Institute, University of Technology Sydney, Sydney, Australia*

***Abstract*—Spatial co-location pattern mining is an important task in spatial data mining. However, traditional mining frameworks often produce too many prevalent patterns of which only a small proportion may be truly interesting to end users. To satisfy user preferences, this work proposes an interactive probabilistic post-mining method to discover user-preferred co-location patterns from the early-round of mined results by iteratively involving user's feedback and probabilistically refining preferred patterns. We first introduce a framework of interactively post-mining preferred co-location patterns, which enables a user to effectively discover the co-location patterns tailored to his/her specific preference. A probabilistic model is further introduced to measure the user feedback-based subjective preferences on resultant co-location patterns. This measure is used to not only select sample co-location patterns in the iterative user feedback process but also rank the results. The experimental results on real and synthetic data sets demonstrate the effectiveness of our approach.**

I. INTRODUCTION

The extraction of spatial co-location patterns is a rising and promising field in spatial data mining. A spatial co-location pattern is composed of a set of spatial features frequently observed together within geographical neighborhoods [1], [2]. Spatial co-location pattern mining yields important insights for various applications such as Earth science [3], public transportation [4], and air pollution [5].

Typically, spatial co-location pattern mining methods use the frequencies of a set of spatial features participating in a co-location pattern to measure a pattern's prevalence (known as *participation index*, *PI* for short) and require a user-specified minimum prevalence threshold *min_prev* to filter prevalent co-location patterns [1], [2], [6]. However, User's preferences are often subjective, a pattern preferred by one user may not be favoured by another, thus cannot be measured by existing objective-oriented *PI* measures. Therefore, it is necessary and advantageous to involve user's preferences [7], [8], [9], [10], [11].

This work proposes a framework to discover user-preferred co-location patterns by iteratively involving user's interactive feedback and probabilistically quantifying user-preferences on co-location patterns. As shown in Fig. 1, our system takes a set *PC* of mined prevalent co-locations as input. First, the top-*k* (e.g., *k*=5, in prevalence value order) co-locations in *PC* are presented to the user as sample co-locations, and the system then asks the user for his/her preferences. The user chooses a set of preferred co-locations and so the first set $PC_{feedback}$ of selected co-locations is collected. Based on $PC_{feedback}$, the prevalent co-locations in *PC* are estimated for their subjective preference by a *probabilistic model*, and ranked by their estimated subjective preferences. Then, as the sample co-locations, the top-*k* co-locations are fed to the user again. After several rounds of the interactive process, the system refines the output that is closest to the user's preference on co-location patterns.

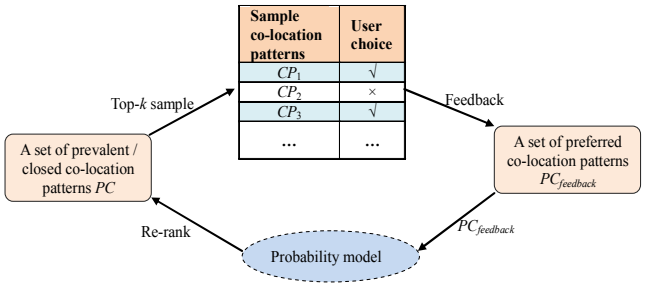

Fig. 1. A framework for interactively post-mining preferred co-locations

The rest of the paper is organized as follows. First, we review concepts related to traditional co-location mining, and formally defines our problem. Next, the probabilistic model method is proposed, and then presents our evaluation strategy and results.

II. PROBLEM STATEMENT

A. Co-location Patterns

In a spatial database, let *F* be a set of *n features* $F=\{f_1, f_2, \ldots, f_n\}$, *D* be a set of *instances* of *F*, where each *instance* is a tuple <feature type, instance ID, location>, and *R* be a *neighbour relationship* over locations of instances, where *R* is symmetric and reflexive.

A *co-location pattern c* is a subset of the feature set *F*. The number of features in *c* is called the *size* of *c*. A *row instance I* of a co-location *c* is a set of instances in *D*, which includes the instances of all features in *c* and forms a clique under the neighbour relationship *R*. The set of all row instances of *c* is called *table instance* of *c,* denoted as *TI*(*c*).

The *participation ratio* of feature f_i in a co-location *c*, denoted as $PR(f_i, c)$, is the fraction of the instances of f_i that participates in table instance *TI*(*c*) of *c*. The *participation index* of a co-location *c*, denoted as *PI*(*c*), is the minimum participation ratio $PR(f_i, c)$ among all features f_i in *c*. A co-location pattern *c* is a *prevalent co-location pattern*, if its

participation index PI is no less than a given prevalence threshold *min_prev*, that is, $PI(c) \geq min\ prev$.

The PI and PR measures satisfy the *anti-monotonicity property* (*downward closure property*) [6]. The introduction of *closed co-location patterns* creates a lossless condensed representation. A prevalent co-location c is *closed* if there is no co-location c' such that $c \subset c'$ and $PI(c) = PI(c')$ [12].

B. Subjective Preference Measure

We assume that a set PC of prevalent or closed co-location patterns has been mined which forms the input to our system. In PC, we suppose that there is a set PC_I of ideal co-location patterns of interest to the user, and then PC_I is the preferred co-location set and $PC_{II} = PC - PC_I$ is the non-preferred co-location set.

We know that the value $PI(c)$ ($c \in PC$) is an objective interest measure in the prevalent co-location pattern mining. In reality, it is not possible that the objective interest can be substituted for subjective preference. PC might contain a large number of mined prevalent co-location patterns, which may not be actionable and useful for users, since they may just be general knowledge, their prevalence may have been enhanced by the instances' autocorrelation, or they are just not preferred by the user.

In order to learn the prior knowledge of the user, we interactively ask for user's feedback concerning preferred co-locations. User's feedback is combined into a set of preferred co-location patterns which is denoted $PC_{feedback}$. The set $PC_{feedback}$ is updated every time the system obtains user's feedback. Therefore, in the interactive process, we use a similarity measure $SIM(c, PC_{feedback})$ between a co-location pattern c in PC and the selected co-location patterns $PC_{feedback}$ to evaluate the degree of subjective preference of any co-location pattern c which has not yet been judged.

C. Problem Statement

The problem of *post-mining preferred spatial co-location patterns through interactive feedback* can be stated as follows. Given a set of prevalent or closed co-location patterns, can the system return the ideal co-location patterns of user's preference, according to user's feedback about preferred co-locations, and at the same time minimize the user's efforts in providing feedback?

Considering the uncertainties of the user's ideal preference, in this paper we use a *classic probabilistic model* to model the prior knowledge of the user. The basic idea of this method is: given a set PC of prevalent or closed co-location patterns, there exists a set PC_I of ideal preference co-location patterns in PC for a user. However, the system does not know the characteristics of the set PC_I at the beginning of the interactive process. It needs to make a guess. According to this guess, the system will identify a result set PC_I as an initial hit. Then the user or system judges the initial result PC_I. Based on the feedback, the system can optimize and improve the initial result PC_I incrementally in an interactive process so that, after repeated interactions, the resultant PC_I should be close to the user's ideal preference result set.

The essence of the above probabilistic model is to estimate the probability of the similarity $SIM(c, PC_{feedback})$ between the selected co-locations $PC_{feedback}$ per user's feedback and a co-location pattern c in the set of prevalent or closed co-locations PC whose preference level has not yet been judged.

III. Probabilistic Model

Assume, for a user, there is a preferred co-location set PC_I and also a non-preferred co-location set PC_{II} in the prevalent or closed co-location set PC. After obtaining a set $PC_{feedback}$ of user's feedback, the similarity $SIM(c, PC_{feedback})$ between a co-location pattern c in PC and the set $PC_{feedback}$ per user's feedback is defined as the ratio of the probability of c being of preference to the user compared to the probability of c not being of preference to the user. i.e.,

$$SIM(c, PC_{feedback}) = \frac{P(PC_I \mid c)}{P(PC_{II} \mid c)} \tag{1}$$

where $P(PC_I \mid c)$ represents the probability of c being of preference to the user, and $P(PC_{II} \mid c)$ represents the probability of c not being of preference to the user.

Since the values of $P(PC_I \mid c)$ and $P(PC_{II} \mid c)$ cannot be computed directly, they need to be estimated with known values. Assume there is an initial guess about the user's ideal preference set PC_I, so Eqn. (1) can be converted per the Bayes' rule:

$$SIM(c, PC_{feedback}) = \frac{P(c \mid PC_I) \times P(PC_I)}{P(c \mid PC_{II}) \times P(PC_{II})} \tag{2}$$

where $P(c \mid PC_I)$ represents the probability that c belongs to PC_I; $P(c \mid PC_{II})$ represents the probability that c belongs to PC_{II}, and $P(PC_I)$ and $P(PC_{II})$ represent the prior probabilities that any co-location in PC belongs to PC_I or PC_{II} respectively.

For a given prevalent or closed co-location set PC, the two values $P(PC_I)$ and $P(PC_{II})$ are related only to the user but not to c. Additionally, we are just concerned about the relative values in computing SIM. So Eqn. (2) can be simplified to:

$$SIM(c, PC_{feedback}) = \frac{P(c \mid PC_I)}{P(c \mid PC_{II})} \tag{3}$$

The probability of randomly selecting c from PC_I or PC_{II} (i.e., the probability that c belongs to PC_I or PC_{II}) can be calculated by the distribution of each 2-size co-location c_i in PC_I and PC_{II}:

$$P(c \mid PC_I) = \prod_{i=1}^{m} p(c_i \mid PC_I)^{w_i(c)} P(\overline{c}_i \mid PC_I)^{(1-w_i(c))} \tag{4}$$

$$P(c \mid PC_{II}) = \prod_{i=1}^{m} p(c_i \mid PC_{II})^{w_i(c)} P(\overline{c}_i \mid PC_{II})^{(1-w_i(c))} \tag{5}$$

where $m = \frac{n(n-1)}{2}$, where n is the number of features in F. $w_i(c) \in \{0, 1\}$: $w_i(c) = 1$ when the i-th 2-size co-location c_i of F is in $PC_{feedback}$ and c at the same time; otherwise $w_i(c) = 0$, and $\overline{c}_i$ represents "not containing 2-size co-location c_i."

Eqns. (4) and (5) can be interpreted as follows: when the 2-size co-location c_i is in $PC_{feedback}$ and c at the same time, i.e., $w_i(c)=1$, the probability that c_i appears randomly in a co-location pattern of PC_I is regarded as a contribution to the process of determining whether c and PC_I are related. In the contrary situation, when the 2-size co-location c_i is not the

same as in $PC_{feedback}$ and c, the probability that c_i does not appear randomly in a co-location pattern of PC_I is also regarded as a contribution.

Based on Eqns. (4) and (5), Eqn. (3) can be converted to:

$$SIM(c, PC_{feedback}) = \frac{\prod_{i=1}^{m} p(c_i \mid PC_I)^{w_i(c)} P(\bar{c}_i \mid PC_I)^{(1-w_i(c))}}{\prod_{i=1}^{m} p(c_i \mid PC_{II})^{w_i(c)} P(\bar{c}_i \mid PC_{II})^{(1-w_i(c))}} \quad (6)$$

Considering the meanings of $P(c_i \mid PC_I)$ and $P(\bar{c}_i \mid PC_I)$, we have $P(c_i \mid PC_I) + P(\bar{c}_i \mid PC_I) = 1$. Accordingly, $P(c_i \mid PC_{II}) + P(\bar{c}_i \mid PC_{II}) = 1$ holds. We take these relations into Eqn. (6) and by taking logarithms, it is converted to:

$$SIM(c, PC_{feedback}) = \sum_{i=1}^{m} w_i(c) \log_{10} \frac{p(c_i \mid PC_I)(1 - P(c_i \mid PC_{II}))}{p(c_i \mid PC_{II})(1 - P(c_i \mid PC_I))} + \sum_{i=1}^{m} \log_{10} \frac{1 - P(c_i \mid PC_I)}{1 - p(c_i \mid PC_{II})} \quad (7)$$

As in Eqn. (7), the expression $\sum_{i=1}^{m} \log_{10} \frac{1 - P(c_i \mid PC_I)}{1 - p(c_i \mid PC_{II})}$ is not related to c, so Eqn. (7) can be further simplified as:

$$SIM(c, PC_{feedback}) = \sum_{i=1}^{m} w_i(c) \log_{10} \frac{p(c_i \mid PC_I)(1 - P(c_i \mid PC_{II}))}{p(c_i \mid PC_{II})(1 - P(c_i \mid PC_I))} \quad (8)$$

That is to say, we can compute the similarity of co-locations c with $PC_{feedback}$ by Eqn. (8), and rank them with the values $SIM(c, PC_{feedback})$. However, as mentioned before, the user's preference set PC_I is not known initially. We need a method to calculate the probabilistic values $p(c_i \mid PC_I)$ and $p(c_i \mid PC_{II})$.

A simple method for calculating the probabilistic values $p(c_i \mid PC_I)$ and $p(c_i \mid PC_{II})$ is that:

$$\begin{cases} p(c_i \mid PC_I) = 0.5 \\ p(c_i \mid PC_{II}) = n_i / N \end{cases} \quad (9)$$

where n_i and N represent the number of co-locations containing 2-size c_i and the number of total co-locations in PC respectively. We can then calculate the $SIM(c, PC_{feedback})$ for each c in PC by Eqn. (8).

However, Eqn. (9) is too arbitrary. After obtaining user's feedback information, based on the feedback principle we propose two improved methods (Eqn. (10) and Eqn. (11)) for calculating $p(c_i \mid PC_I)$ and $p(c_i \mid PC_{II})$ which improve the computation of $SIM(c, PC_{feedback})$ and help minimize the user's efforts in providing feedback.

$$\begin{cases} p(c_i \mid PC_I) = (r_i + 0.5)/(r+1) \\ p(c_i \mid PC_{II}) = (n_i - r_i + 0.5)/(N - r + 1) \end{cases} \quad (10)$$

Or, by adding less arbitrary adjusting factors, we have Eqn. (11).

$$\begin{cases} p(c_i \mid PC_I) = (r_i + n_i / N)/(r+1) \\ p(c_i \mid PC_{II}) = (n_i - r_i + n_i / N)/(N - r + 1) \end{cases} \quad (11)$$

where r is an adjusting factor (it can be pre-specified by the user) used to get a collection V of the top-r co-location patterns under the values $SIM(c, PC_{feedback})$, and r_i is the number of co-location patterns containing 2-size co-location c_i in V.

Note that Eqn. (9) needs to be used at the beginning of the interactive process when using Eqn. (10) or Eqn. (11).

IV. Experimental Results

We conduct comprehensive experiments to evaluate the proposed approach from multiple perspectives on both real and synthetic data sets. Due to space limitations, we present only a subset of our full results here.

A. Experimental Setting

We set up an experimental environment, called Simulator, to simulate user's feedback. Since our goal is to discover preferred co-location patterns interactively and rank the results, our accuracy measure favours high-rank co-location patterns in the results. Let top-l(*learned_set*) be the top-l results reported by the ranking *learned* from the interactive feedback and *target_set* be the results in the *target* co-locations constructed by our Simulator, the accuracy measure is defined as follows.

$$Accurary = \frac{top - l(learned\_set) \cap target\_set}{l} \quad (12)$$

where l is given m/5, 2m/5, 3m/5, 4m/5 or m (m=|*target_set*|) in the experiments. It is obvious that the accuracy values in Eqn. (12) are the percentages of the top-l ranked co-locations in the *target_set*.

B. Accuracy Evaluation on Real Data Sets

Using the Simulator discussed above, our first task is to examine the accuracy of the results learned from the interactive feedback. We use three real data sets with different distributions in the experiments. Real-1 is from the rare plant data of the Three Parallel Rivers of Yunnan Protected Areas whose instances form a zonal distribution, which has a small quantity of instances. Real-2 is a spatial distribution data set of urban elements whose instances' distribution is both even and dense, and which has a large quantity of features as well as instances. Real-3 is a vegetation distribution data set of the Three Parallel Rivers of Yunnan Protected Areas, which has the fewest features but the most instances, and instance distribution presents various clusters.

We summarize the main lessons from the experiments.

First, over the three real data sets, we observed that F-10 and F-11 have better accuracy than F-9 (F-9, F-10 or F-11 means using Eqn. (9), (10) or (11) respectively to calculate the probabilistic values $p(c_i \mid PC_I)$ and $p(c_i \mid PC_{II})$ in calculating by Eqn. (8)) because F-10 and F-11 add some adjusting factors for computing $p(c_i \mid PC_I)$ and $p(c_i \mid PC_{II})$. The accuracy with F-11 is a little better than F-10 also because of the more reasonable probabilistic values. The accuracy estimated in closed co-locations is better than that in prevalent co-locations because closed co-locations are a form of compression of prevalent co-locations which can help effectively discover the interesting co-locations. We also find that: (1) as *iter* (number of iterations of feedback) increases, the accuracy increases, and this is because each iteration supplies new samples to the

user, and the new feedback from the user updates the SIM values of co-locations in *PC*, bringing them closer and closer to the user's real preference; (2) a larger k (number of sample co-locations for feeding to user) causes a higher accuracy because more samples can be fed to the user; (3) a smaller l (number to get top-l(*learned_set*) for accuracy measure in experiments) can reach higher accuracy because the co-locations in the front of *target_set* have been already chosen by the user.

Second, the main observations on Real-2 are similar to Real-1, although the accuracy estimated for Real-2 is higher than that for Real-1 with the same parameter values, but the accuracy gap between prevalent co-locations and closed co-locations is not as obvious as Real-1, because the compression of closed co-locations on Real-2 is much lower than that of Real-1, which makes a smaller gap between them.

Third, the accuracy on Real-3 can reach 100% within a few rounds. The reason for the high accuracy in Real-3 is that there are only 15 features, and the smaller number of features makes it easier to find the combinations preferred by a user.

C. Accuracy Evaluation on Synthetic Data Sets

Synthetic data sets are generated to test the accuracy and efficiency of our algorithm when data size changes. Figs. 2 and 3 show the accuracy and efficiency w.r.t. different number of features. We observe the following results.

First, the accuracy in a dense data set is higher than that in a sparse data set. The reason is that dense data sets can generate longer co-location patterns which have more chance of containing the preferred combination of features (rules), which means that preferred co-locations can be selected more easily in each round, further improving the accuracy of our algorithm.

Second, as the number of features increases, F-10 and F-11 show much better accuracy than F-9 in Fig. 2, and the gap of accuracy between F-10/F-11 and F-9 also increases. This is because the adjusted factors added in F-10 and F-11 play a bigger role as the data set gets bigger and bigger. Note that in this experiment there are about 1000000 spatial instances with 100 features.

Third, Fig. 3 shows the average running time of F-9, F-10 and F-11 per round and the number of closed co-locations (PC_count). It can be seen that F-9 has a much higher efficiency than either F-10 or F-11. When the number of closed co-locations reaches almost 700000, F-9 only costs no more than 20 seconds, and this is because F-9 only needs to calculate n_i, and n_i can be updated based on the last round value. While F-10 and F-11 need to calculate not only n_i but also r_i, and r_i cannot be updated as n_i because in each round the top-r co-locations based on the SIM values may change greatly, thus increasing the running time. But even with 100 features, the running time per round is only around 70 seconds.

Acknowledgements This work was supported in part by grants (No. 61472346, No.61662086) from the National Natural Science Foundation of China, in part by a grant (No. 2016FA026, No. 2015FB114) from the Science Foundation of Yunnan Province, and in part by the Project of Innovation Research Team of Yunnan Province.

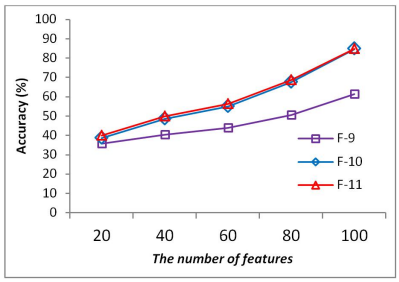

Fig. 2. Accuracy evaluation on synthetic data sets with different number of features

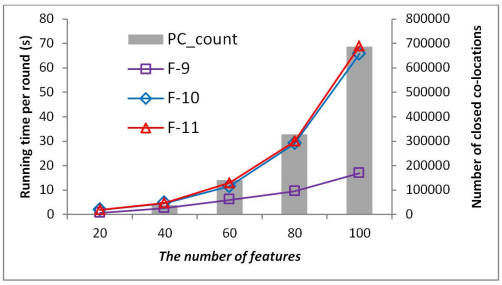

Fig. 3. Running time per round and the number of closed co-locations with different number of features

REFERENCES

[1] S. Shekhar, and Y. Huang, "Co-location rules mining: a summary of results," in *Proc. SSTD 2001*, LNCS, vol. 2121, pp. 236-256.

[2] X. Zhang, N. Mamoulis, D. Cheung, and et al, "Fast mining of spatial co-locations," in *Proc. ACM SIGKDD 2004*, pp. 384-393.

[3] F. Verhein, G and Al-naymat, "Fast mining of complex spatial co-location patterns using GLIMIT," in: *Proc. ICDMW 2007*, pp. 679-684.

[4] W. Yu, "Spatial co-location pattern mining for location-based services in road networks," *Expert Systems with Applications*, vol. 46, pp. 324–335, 2016.

[5] M. Akbari, F. Samadzadegan, and R. Weibel, "A generic regional spatio-temporal co-occurrence pattern mining model: a case study for air pollution," *J. Geograph. Syst.*, vol. 17, pp. 249–274, 2015.

[6] Y. Huang, S. Shekhar, H. Xiong, "Discovering colocation patterns from spatial data sets: a general approach," *IEEE Trans. Knowl. Data Eng (TKDE)*. vol. 16, no. 12, pp. 1472-1485, Dec. 2004.

[7] L. Wang, X. Bao, L. Zhou, "Redundancy reduction for prevalent co-location patterns," *IEEE Trans. Knowl. Data Eng (TKDE)*, vol. 30, no. 1, pp. 142-155, Jan. 2018.

[8] L. Cao, Y. Zhao, H. Zhang, D. Luo, C. Zhang. "Flexible frameworks for actionable knowledge discovery,'' *IEEE Trans. Knowl. Data Eng (TKDE),* vol. 22, no. 9, pp, 1299-1312, Sep. 2010.

[9] D. Xin, X. Shen, Q. Mei, and J. Han, "Discovering interesting patterns through user's interactive feedback," in *Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, pp. 773-778, 2009.

[10] X. Bao, L. Wang, and H. Chen, "Ontology-based interactive post-mining of interesting co-location patterns," in *Proc. of the 18th Asia-Pacific Web Conference (APWeb'16)*, LNCS vol. 9932, pp. 406-409, 2016.

[11] X. Bao, and L. Wang, "Discovering interesting co-location patterns interactively using ontologies," in *DASFAA Workshops 2017*, LNCS vol. 10179, pp. 75-89, 2017.

[12] J. S. Yoo and M. Bow, "Mining top-k closed co-location patterns," in *Proc. IEEE Int'l Conf. Spatial Data Mining and Geographical Knowledge Services (ICSDM'11)*, pp. 100-105, 2011.