# Can We Apply Projection Based Frequent Pattern Mining Paradigm to Spatial Co-location Mining?

Yan Huang, Liqin Zhang, and Ping Yu

Department of Computer Science and Engineering,
University of North Texas,
P.O. Box 311366, Denton, Texas 76203
{huangyan, lzhang, py0003}@unt.edu

**Abstract.** A co-location pattern is a set of spatial features whose objects are frequently located in spatial proximity. Spatial co-location patterns resemble frequent patterns in many aspects. Since its introduction, the paradigm of mining frequent patterns has undergone a shift from a generate-and-test based frequent pattern mining to a projection based frequent pattern mining. However for spatial datasets, the lack of a transaction concept, which is critical in frequent pattern definition and its mining algorithms, makes the similar shift of paradigm in spatial co-location mining very difficult. We investigate a projection based co-location mining paradigm. In particular, we propose a projection based co-location mining framework and an algorithm called **FP-CM**, for **FP-growth Based Co-location Miner**. This algorithm only requires a small constant number of database scans. It out-performs the generate-and-test algorithm by an order of magnitude as shown by our preliminary experiment results.

## 1 Introduction

We focus on a recent spatial data mining problem: finding spatial features that tend to be located in spatial proximity. This problem is also referred to as *spatial co-lcoation patterns mining* [7, 4, 2, 10, 9]. Let $\mathcal{F} = \{f_1, f_2, \ldots, f_l\}$ be a set of spatial features. consider a number of $l$ spatial datasets $\{SD_1, SD_2, \ldots, SD_l\}$, such that $SD_i, i \in [1, l]$ contains all and only the objects that have the spatial feature $f_i$, Let $\mathcal{R}$ be a given spatial neighbor relation (e.g. distance less than 1.5 miles). A set of spatial features $X \subseteq \mathcal{F}$ is a co-location if its value $im(X)$ of an interesting measure, is above a threshold $min\_im$. The problem of finding the complete set of co-location patterns is called the co-location mining problem. Mining *spatial co-location patterns* is an important spatial data mining task with broad applications.

Spatial co-location patterns resemble frequent patterns [5], a more general problem of mining association rules [1] in many aspects. Since its introduction, the problem of mining frequent patterns from large databases, has been subject

T.B. Ho, D. Cheung, and H. Liu (Eds.): PAKDD 2005, LNAI 3518, pp. 719–725, 2005.

of numerous studies. The paradigm of frequent pattern mining algorithms has undergone a fundamental shift from generate-and-test approaches [1] to projection based approaches [5]. Projection based approaches have major advantages over generate-and-test approaches and avoids multiple database scans by compressing transactional data into compact structures. However, the lack of prematerialized transactions becomes a major obstacle in adopting projection based algorithms in spatial co-location pattern mining. A natural question to ask is: can we push the same paradigm shift for mining spatial co-location patterns?

Many algorithms for co-location mining proposed in literature [7, 4, 10, 9, 3] employ an generate-and-test co-location mining paradigm, which utilizes the anti-monotone property of interestingness measures. In a clustering-based map overlay approach[4, 3], every spatial feature is treated as a map layer and point-data in each layer are clustered into regions. In a reference feature based approach [7], transactions are created according to different algorithms, then a level wise algorithm is applied . Under this model, a frequent pattern based algorithm can be applied straightforwardly due to the fact that the interestingness measure is defined based on the generated transactions. In distance based approaches [9, 10], the number of instances for each spatial feature set is used to define the interestingness measure. In an event centric model [10], a participation index was defined as the interestingness measure. The participation index of a pattern is defined as the minimal participation ratio of the objects of each feature in the pattern.

The contribution of this work is to study how to use a projection based paradigm for event based spatial co-location pattern mining (CM). We proposed a projection based framework for CM, which can incorporate any fast frequent pattern mining algorithm. In particular, we developed an FP-growth based algorithm for spatial co-location mining(FP-CM) based on the proposed framework. We provide preliminary experiment results to show that the FP-CM is an order of magnitude faster than the generate-and-test algorithm *Co-location Miner*.

**Paper Outline:** Section 2 recalls important concepts of co-location and frequent pattern mining. Section 3 proposes our projection based FP-CM framework and a FP-growth based co-location mining algorithm. We present the preliminary experimental results in section 4 and summarize our work and present future work in section 5.

## 2 Background

We review basic concepts of co-location patterns, a traditional generate-and-test co-location mining algorithm, and a projection based frequent pattern mining algorithm in this section.

In an event centric model [10], a participation index was defined as the interestingness measure. For a set of spatial features $X \subseteq \mathcal{F}$, a set of objects $\{o_1, o_2, \ldots, o_k\}$ is an *instance* of $X$ iff $(\forall i, i \in [1, k], o_i \in SD_i)$ and $(\forall i \forall j, 0 < i < j \leq k, (o_i, o_j) \in \mathcal{R})$. The *participation ratio* $pr(f, X)$ of a feature $f$ in a pattern $X$ is defined as: