Information Sciences 179 (2009) 3370–3382

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

An order-clique-based approach for mining maximal co-locations

Lizhen Wang [a,*], Lihua Zhou [a], Joan Lu [b], Jim Yip [b]

[a] Department of Computer Science and Engineering, School of Information Science and Engineering, Yunnan University, Kunming 650091, China
[b] Department of Informatics, School of Computing and Engineering, University of Huddersfield, Huddersfield HD1 3DH, UK

ARTICLE INFO

Article history:
Received 26 May 2008
Received in revised form 21 May 2009
Accepted 29 May 2009

Keywords:
Spatial data mining
Co-location patterns mining
Maximal ordered co-locations
Table instances
Order-clique-based approach

ABSTRACT

Most algorithms for mining spatial co-locations adopt an Apriori-like approach to generate size-k prevalence co-locations after size-$(k-1)$ prevalence co-locations. However, generating and storing the co-locations and table instances is costly. A novel *order-clique-based approach for mining maximal co-locations* is proposed in this paper. The efficiency of the approach is achieved by two techniques: (1) the spatial neighbor relationships and the size-2 prevalence co-locations are compressed into extended prefix-tree structures, which allows the order-clique-based approach to mine candidate maximal co-locations and co-location instances; and (2) the co-location instances do not need to be stored after computing some characteristics of the corresponding co-location, which significantly reduces the execution time and space required for mining maximal co-locations. The performance study shows that the new method is efficient for mining both long and short co-location patterns, and is faster than some other methods (in particular the join-based method and the join-less method).

1. Introduction

Co-location patterns mining is a new branch studied in the spatial data mining recently. A spatial co-location pattern represents a subset of spatial events (or features) whose instances are frequently located in a spatial neighborhood. Spatial co-location patterns may yield important insights in many applications. For example, a mobile service provider may be interested in mobile service patterns frequently requested by geographical neighboring users. The locations at which people congregate can be used for providing attractive location-sensitive advertisements. Botanists may be interested in symbiotic plant species in an area. A certain virus disease may frequently happen in the area where mosquitoes overrun, etc. Other application domains include Earth science, public health, biology, transportation, etc. [7,9,20].

Co-location pattern discovery is challenging due to the following reasons: first, it is difficult to find co-location patterns with traditional association rule mining algorithms since there is no concept of traditional "transaction" in most of the spatial datasets [10,18]. Second, the instances of a spatial event are distributed in a certain area and share complex spatial relationships with other spatial instances [2–4,11] so that a large fraction of the computation time of mining co-location patterns is devoted to generating table instances of co-location patterns [7]. Third, the Apriori-like algorithms to generate size-k prevalence co-locations after size-$(k-1)$ prevalence co-locations may suffer from the following two non-trivial costs:

(1) It is costly to handle a huge number of candidate co-locations. For example, if there are 10^3 spatial events, the Apriori-like algorithms will need to generate more than 10^5 size-2 candidates and test their occurrence prevalence. Moreover, to discover a prevalence co-location of size-100, such as $\{f_1,\ldots,f_{100}\}$, it must generate $2^{100}-2 \approx 10^{30}$ candidates in total. This is an inherent cost of generating the set of all prevalence co-locations, no matter what implementation technique is applied.

* Corresponding author. Tel.: +86 871 4602567.
E-mail address: lzhwang@ynu.edu.cn (L. Wang).

0020-0255/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2009.05.023

(2) It is wasteful to store excessive table instances of co-locations, especially so when the number of table instances is very large.

Can the number of co-locations generated in co-location mining be substantially reduced whilst preserving the complete information regarding the set of prevalence co-locations? Can a method be developed to utilize some novel data structures and algorithms to avoid storing a huge number of table instances? These questions motivated this study.

1.1. Related works

In previous work on mining co-location patterns, Morimoto [12] defined distance-based patterns as k-neighboring class sets. In his work, the number of instances for each pattern is used as the prevalence measure, which does not possess an anti-monotone property by nature. Nevertheless, Morimoto used a non-overlapping instance constraint to get the anti-monotone property for this measure. In contrast, Shekhar and Huang [13] developed an event centric model which abolishes the non-overlapping instance constraint, and a new prevalence measure called the *participation index (Pi)* is defined which has the desirable anti-monotone property. At the same time, Huang et al. [7] proposed a general mining approach: join-based approach mining co-locations (called *join-based approach*), which established the basis of co-location mining. This approach works well on sparse spatial datasets. However, when dealing with dense datasets, it is inefficient as the computation time of the join increases with the growth in co-locations and table instances. Yoo and Shekhar proposed two improved algorithms—called the *partial-join approach* and the *join-less approach*, respectively—to overcome the efficiency disadvantage of the join-based approach in [21,22].

The partial-join approach is to build a set of *disjoint clique* in spatial instances to identify the intraX instances of co-location (belonging to a common clique) and interX instances of co-location (all instances have at least one cut neighbor relation), and merge them to calculate the value of the *Pi*. This approach reduces the number of expensive join operations in identifying table instances dramatically. However, the performance depends on the distribution of the spatial dataset, more precisely the number of cut neighbor relations.

The join-less approach puts the spatial neighbor relationships into a compressed *star neighborhood*. All the possible table instances for every co-location pattern are generated by scanning the star neighborhood, and by a three-time filtering operation. The join-less co-location mining algorithm is efficient since it uses an instance-lookup scheme instead of an expensive spatial or instance join operation for identifying co-location table instances. However, the star neighborhood structure is not a suitable structure for generating table instances, as the table instances generated from this structure have to be filtered. Therefore, the computation time of generating co-location table instances will increase with the growing length of co-location patterns.

Besides the representative co-location mining algorithms above, there are some other interesting works in mining co-location patterns. Huang et al. addressed the problem of mining co-location patterns with rare spatial events [6]. In their paper, a new measure called the maximal participation ratio (maxPR) was introduced and a weak monotonicity property of the maxPR measure was identified. Verhein and Al-Naymat studied mining complex spatial co-location patterns from spatial datasets [15]. They introduced the idea of a maximal clique and applied the GLIMIT itemset mining algorithm to their task [14]. GLIMIT is a very fast and efficient itemset mining algorithm that has been shown to outperform Apriori and FP-Growth. Celik et al. studied the zonal co-location patterns discovery problem [1]. Wang et al. presented a new join-less method for co-location patterns mining [17]. In their paper, a new data structure called CPI-tree (Co-location Pattern Instance Tree) was introduced, which materializes spatial neighbor relationships. Related works about mining maximum frequent itemsets and maximal cliques were discussed in [5,8].

1.2. Our contributions

The following contributions are made by this paper:

(1) This is the first work to study maximal co-location mining. Some novel, compact data structures, P_2-*tree*, CP_m-*tree*, *Neib-tree* and *Ins-tree*, are introduced, which extend prefix-tree structures to store crucial, quantitative information about the size-2 prevalence co-locations, about the candidate maximal ordered co-locations, about the spatial neighbor relationships, and about the table instances. To ensure that the tree structures are compact and informative, the tree nodes are arranged in an ascending order (the spatial events sort in alphabetical order, and then the different instances of the same spatial event sort in numerical order).
(2) We propose a novel *order-clique-based* approach to mine candidate maximal co-locations and identify table instances. The order-clique-based approach is efficient since (a) the tree structures support a pile-instance-lookup scheme and an effective strategy of pruning the branches whose child-node number is less than a relevant value to identify table instances, and (b) it does not need to store excessive table instances for identifying next level table instances.
(3) A performance study has been conducted to compare the performance of the *order-clique-based* method with two representative co-location mining methods, the *join-based* and the *join-less*. The study shows that *order-clique-based* method is faster than *full-join* and *join-less*, especially when the spatial dataset is dense (containing many table instances) and/or the prevalence co-locations are long.