Expert Systems With Applications 46 (2016) 324–335

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Spatial co-location pattern mining for location-based services in road networks

Wenhao Yu*

School of Marine Science and Technology, Tianjin University, 92 Weijin Road, Tianjin 300072, PR China

A R T I C L E I N F O

Keywords:
Spatial data mining
Spatial co-location patterns
Network space
Network analysis
Location-based services

A B S T R A C T

With the evolution of geographic information capture and the emergency of volunteered geographic information, it is getting more important to extract spatial knowledge automatically from large spatial datasets. Spatial co-location patterns represent the subsets of spatial features whose objects are often located in close geographic proximity. Such pattern is one of the most important concepts for geographic context awareness of location-based services (LBS). In the literature, most existing methods of co-location mining are used for events taking place in a homogeneous and isotropic space with distance expressed as Euclidean, while the physical movement in LBS is usually constrained by a road network. As a result, the interestingness value of co-location patterns involving network-constrained events cannot be accurately computed. In this paper, we propose a different method for co-location mining with network configurations of the geographical space considered. First, we define the network model with linear referencing and refine the neighborhood of traditional methods using network distances rather than Euclidean ones. Then, considering that the co-location mining in networks suffers from expensive spatial-join operation, we propose an efficient way to find all neighboring object pairs for generating clique instances. By comparison with the previous approaches based on Euclidean distance, this approach can be applied to accurately calculate the probability of occurrence of a spatial co-location on a network. Our experimental results from real and synthetic data sets show that the proposed approach is efficient and effective in identifying co-location patterns which actually rely on a network.

1. Introduction

With the ubiquity of wireless Internet access, GPS-enabled mobile terminals and the advance in spatial database management systems, a new generation of mobile services, known as location-based services (LBS), has been developed. These services are capable of delivering geographic information and geoprocessing power to mobile users according to their current location (Beatty, 2002). Understanding the geographic context is the essential question in the LBS area. Geographic context, the information of the surroundings or circumstances, has been an important topic in many applications. By introducing the spatial data analysis and geo-sensor data collection into the geographic context-awareness, the service provider could establish reliable and high-quality services to help users in their trip planning, activity re-scheduling, and the decision-making process.

A spatial co-location pattern is a set of spatial features that are frequently located together in spatial proximity (Shekhar & Huang, 2001). As an important concept for spatial analysis and geographic context awareness, the spatial co-location pattern mining has been popularly applied in discovering the spatial dependency of objects. Spatial dependency is a tendency of observed objects located close to another in the geographic space to show a higher degree of similarity or dissimilarity (Miller & Han, 2009). All computed dependency patterns concerning objects in space depend on the definition of spatial neighborhood. In the literature, closeness can be defined by using different types of distance metric, like Euclidean distance, path distance, etc. To provide users realistic geographic information, providers should choose the distance measure which best fits the geographic context in study regions.

So far, many different studies on spatial co-location mining have been carried out. However, all of the previous studies are based on the Euclidean (or planar) space assumption. The researchers (Huang, Shekhar, & Xiong, 2004; Shekhar & Huang, 2001; Yoo & Shekhar, 2004) assume that the patterns of interest are occurred in an infinitely homogeneous and isotropic space, and spatial proximity between two objects is measured by the straight-line Euclidean distance. Miller (1994) pointed out that this can be ill-suited as many human activities are constrained only to the network portion of the planar space. Network distance is indeed a more meaningful and reliable distance measure for analyses related to social and economic

* Tel.: +86 15972171198; fax: +86 27 68778893.
E-mail address: ywh_whu@126.com, wdyuwenhao@gmail.com

http://dx.doi.org/10.1016/j.eswa.2015.10.010
0957-4174/

process (Okabe, Boots, Sugihara, & Chiu, 2009; Okabe, Okunuki, & Shiode, 2006; Yamada & Thill, 2007). For description simplicity, we term the spatial co-location pattern mining based on network distance as *NM-Coloc*, while the spatial co-location pattern mining based on Euclidean distance as *PM-Coloc*.

Road transport networks are emerging nowadays at an urban or extra-urban level. There are many cases that need to capture co-location patterns with network distance. For example, in mobile service applications, clients usually request services with regard to the co-location patterns of urban facilities. These patterns need to be measured using the distance along road network instead of Euclidean distance, because the driven vehicles are restricted to move on predefined roads under certain transformation condition. In fact, the computed patterns of facilities that take the limit of the movements of people into account can help service providers provide more attractive location-sensitive advertisements, recommendations, etc. As another example, these co-location patterns would also benefit location choosing for a business and advertisement. By extracting the network-constrained patterns of facilities of different types, the decision makers from a company can plan a better location for a new store considering the profitability of similar retail stores depending on the surrounding objects.

Motivated by the great potential demand and lack in research of network-constrained co-locations, we develop an efficient spatial data mining technique to help domain experts discover useful knowledge from the given network data set. As opposed to the Euclidean distance measure used in Huang et al. (2004) and Shekhar and Huang (2001), we try to find dependencies among features with the network distance. Besides, this study also takes the particular characteristics of network space into consideration and we got two major conclusions: (1) in many LBS applications, the distance of shortest path is more suitable to measure the spatial connection between two locations in the urban environment; (2) spatial neighbors searching in networks is generally a time-consuming task, and an efficient method needs to be performed for building the neighborhood relationship graph.

We conduct a series of comparisons and analyses using our experimental system of knowledge extraction. The results show that our method is effective, efficient and scalable for mining large spatial data sets without the repeated computation of the shortest path. Notably, for creating high order rules, the classical Apriori-like algorithm is also used in our research, but it is not a focus in our topic.

The remainder of the paper is organized as follows. Section 2 highlights related works. Section 3 describes the data model and a new method of *NM-Coloc*. Experimental results are analyzed in Section 4. Finally, Section 5 concludes this paper.

2. Related work

The co-location patterns mining is an important offshoot of spatial association rules mining. The notion of spatial association rule was first defined in Koperski and Han (1995), where a model named as reference feature centric model is proposed with attention paid to a particular type of spatial objects. Each set of spatial instances that have neighbor relationships with an instance of the reference feature is considered as a transaction, and traditional rules mining methods (e.g. *Apriori* algorithm) are applied to these transactions for discovering the association rules related to the reference feature. To focus on the rules which are relatively strong, i.e. which occur frequently, Agrawal and Srikant (1994) also introduced the measures of *support* and *confidence*. However, there is a limitation to these approaches, i.e. interesting rules do not always have a certain reference feature. Then, for discovering all of the interesting patterns, sometimes we should turn to the approach of co-location patterns mining.

Spatial co-location mining has many potential applications in various fields, including business, ecology, engineering, public health, environmental studies, transportation, and earth science. For example, some may focus on the co-occurrence relation of West Nile disease and stagnant water sources in epidemiology, and others may emphasize the dependence pattern of traffic jam and car accident in transportation (Shekhar & Huang, 2001). Huang et al. (2004) proposed an object-centric model for co-location mining. Compared to the reference feature centric model, this model defines clique instances as co-location instance and does not require specified reference features. The authors proposed a statistically meaningful interest measure (i.e. *participation index*) and a join-based algorithm which is similar to *apriori_gen* (Agrawal & Srikant, 1994). The measure is naturally anti-monotone, and the algorithm takes advantage of this desirable property for generating co-location instances. However, this approach requires a large amount of instance join operations, which are expensive in spatial data sets. To reduce the number of instance join operations, different algorithms were proposed in Yoo and Shekhar (2004, 2006) . Among them, Yoo and Shekhar (2004) proposed an algorithm based on plane partitioning. It partitions a plane to cells of neighborhood size such that, the maximum distance between any two feature instances in a partition is smaller than or equal to the neighborhood distance threshold. The spatial instances that are located in the same partition automatically form a clique, and instance join operations are only needed for neighborhood relationships not modeled by the partition cells. Although this approach can achieve a better performance, there are still a large amount of instance join operations required. To totally avoid the instance join operations, later they proposed a join-less algorithm to discover spatial co-location patterns. In Zhang, Mamoulis, Cheung, and Shou (2004), using an extension of the algorithm of spatial join, the authors unify the process of finding spatial neighborhoods with the process of mining. In fact, other extensions of spatial co-location mining algorithm also are possible (Flouvat, Soc, Desmier, & Selmaoui-Folcher, 2015; Huang, Pei, & Xiong, 2006; Phillips & Lee, 2012; Yoo & Bow, 2012).

The basic element of spatial data mining is the definition of neighborhood. In the literature, except for the studies that used a fixed distance and space partition discussed above, there are many other conceptualizations of spatial relationship used in spatial association pattern mining. They include buffer zones (Appice, Ceci, Lanza, Lisi, & Malerba, 2003; Mennis & Liu, 2005), topological relationships (Santos & Amaral, 2005), k nearest neighbors (Wan & Zhou, 2009), Delaunay diagrams (Bembenik & Rybiński, 2009), broad boundary (Clementini, Felice, & Koperski, 2000), etc. However, to the best of our knowledge, no one has yet focused on the co-location mining with network distances (e.g. shortest path distance).

The application of network spatial analysis is not a new topic in the field of GIS (Geographic Information Science). Many recent studies have been developed to illustrate the greater efficiency of network distance versus Euclidean distance measures in the analysis of network-constrained objects or phenomena (Yu, Ai, & Shao, 2015). For instance, to detect the clusters of vehicle crash, Yamada and Thill (2007) proposed to use the K function-based method with network distance rather than traditional method with Euclidean distance. To computing the proximity between urban facility points, Okabe, Satoh, Furuta, Suzuki, and Okano (2008) and Ai, Yu, and He (2015) proposed Voronoi diagram generating methods that effectively combine constraint from network distance measure. On the other hand, Xie and Yan (2008) and Shiode (2011) found that measuring distance by connecting the straight line between locations could possibly overestimate the clustering tendency of network phenomenon (traffic accident or street crime). However, in the field of spatial data mining, there are no studies that evaluate the effect of using network distance instead of Euclidean distance in computing spatial co-location patterns. Furthermore, most algorithms of network distance calculating are designed to work for a small amount of objects. They would be time consuming in network-constrained co-location mining, which usually involves many spatial-join operations.