Application of Space Co-location Mining in Urban Construction

Guoshuai Lv1*
1Yunnan University, Kunming, Yunnan, 650500, China
*Corresponding author’s e-mail: 1584238774@qq.com

Abstract: with the continuous development of the world economy, people's living quality has been improved significantly, but meanwhile a series of urban construction problems have occurred. Due to the increasing population, the number of residential areas, schools, hospitals and other facilities is increasing, and unreasonable distribution will inevitably take place. The main purpose of urban planning is to coordinate the specific arrangement of urban spatial layout and various concrete construction by rationally utilizing the urban land according to the determined nature, scale and development direction of the city. From this point of view, by planning the city to make the population of the city rationally distributed, and fully analyzing the factors of people's gathering according to the existing layout of the city, these elements can be planned and constructed in a centralized manner so as to solve the problems of transportation, environment, etc. in the city. This paper analyzes the reasonable and advanced urban layout and extracts useful knowledge from the spatial data, thus helping the relevant agencies of the government to plan and develop the new urban area.

1. Background information
Data mining has attracted the concern of scholars from various fields in recent years, which is a process of using some mining algorithms to find out some knowledge meaningful to people from a database where mass data are stored. For the last few years, people have accumulated vast amounting of data containing spatial features. How to extract meaningful knowledge from the abundant data has aroused people's great interest, and then spatial data mining came into being at this point.

2. Overview of spatial co-location mining
Spatial co-location mining is the main method used in this paper. Compared to the traditional data, the spatial distribution of spatial data has the feature of co-location, i.e. if the closer the two objects tend to be to each other, the more similar their properties are.

2.1. Related concepts of spatial co-location mode
• Spatial object and spatial object instance. The spatial object refers to the different things in a space. For instance, if Chengdu is taken as an example, Sichuan University can be regarded as a space object in Chengdu. The instance of a spatial object is a feature of a spatial object. For instance, School of Mathematics of Sichuan University can be treated as the spatial instance of the spatial object-Sichuan University.
• Spatial neighbor relation. Spatial neighbor relation is a concept that expresses the positional relationship between the spatial objects. Spatial neighbor relation can be spatial topological relation, spatial distance relation, spatial mixed relation, etc.
Cluster. For a collection of instances of a spatial object, if any two of them accord with spatial neighbor relation, the instance set of the spatial object is called a cluster.

Row instance and relation instance. If a group contains all the objects in a co-location pattern, and there is no subset of the group being able to contain all the objects in this pattern, then this group is called a row instance of the co-location pattern. A collection of all row instances in a co-location is called a relation instance of this pattern.

Participation rate and engagement. Like traditional association rule mining, spatial co-location pattern mining also introduces the concept of participation rate and engagement. The participation rate represents the ratio of the number of instances of a spatial object that are not repeated in a spatial co-location mode to the total number of instances of the spatial object, which can be expressed as follows:

\[ PR(c, f_i) = \frac{\pi(f_i(table \_ins \tan ce(c)))}{\pi(table \_ins \tan ce(f_i))} \]  \hspace{1cm} (1)

\( PR(c, f_i) \) denotes the participation rate of \( f_i \) in mode \( c \), and \( \pi \) denotes a projection operation of the relationship. The engagements is the minimum value of the participation rate of all spatial objects in the spatial co-location mode. and expressed in formula as:

\[ PI(c) = \min_{i=1}^{k} \{ PR(c, f_i) \} \]  \hspace{1cm} (2)

\( PI(c) \) represents the engagement in co-location mode \( c \). If \( PI(c) \) is greater than or equal to the minimum engagement given by the user, the co-location mode \( c \) is called to be frequent.

Co-location rules and conditional probability. The representation of the Co-location rule is basically the same as the association rule. As for the conditional probability, the credibility of \( c_2 \) is derived from the co-location mode \( c_1 \), which can be expressed as:

\[ cp(c_1 \Rightarrow c_2) = \frac{\pi(c_1(table \_ins \tan ce(c_1 \cup c_2)))}{\pi(table \_ins \tan ce(c_1 \cup c_2))} \]  \hspace{1cm} (3)

Where \( \pi \) represents the projection operation of the relationship. So far, all the related concepts about spatial co-location pattern mining have been roughly introduced. In the following, the kernel theory of this paper and the full-join algorithm of spatial co-location pattern mining will be introduced.

2.2. Spatial co-location pattern mining algorithm based on full join

The full-join algorithm for spatial co-location pattern mining was proposed by Huang Y, Shekhar S, Xiong H [3] in 2004. As a branch of association rule mining [4], the algorithm is similar to the Apriori algorithm.

According to the description of Huang Y, Shekhar S, Xiong H [3], the core of the full-join algorithm is to generate the candidate pattern and instance. The generating process of candidate pattern is: in the two K-1 order modes, if the first K-2 order is the same, then the two K-1 order modes are joined to generate the candidate K-order mode.

The process of generating the K-order mode instance is: if the first K-2 items of the relation instance corresponding to the two K-1 order modes are the same, and the last two relation instances satisfy the spatial neighbour relation, the two mode relation instances are joined to generate the relation instance of the K-order mode.

2.3. Description of the fully-join algorithm

Variable:
K: co-location
\( C_k \): k-order candidate co-location mode.
\( T_k \): set of table instances of k-order co-location.
\( P_k \): k-order co-location frequent set.
R_k: k-order co-location rule set.
T_{C_k}: A set of rough table instances of k-order co-location in C_k.

Step:
Co-location size k=1; C_1=ET, P_1=ET;
T_1=generate_table_instance(C_1,E);

IF(fmul=TRUE) then
T_{C_1}=generate_table_instance(C_1,k);

Initialize data structure C_k,T_k,P_k,R_k,T_{C_k} to be empty for 1<k≤K;

While(not empty P_k and k<K ) do{
C_{k+1}=generate_candidate_colocation(C_k,k);
IF(fmul=TRUE) then
C_{k+1}=multi_resolution_pruning(θ, C_{k+1}, T_{C_k}, multi_rel(R));
T_{k+1}=generate_table_instance(θ, C_{k+1}, T_k, R);
P_{k+1}=select_prevalent_colocation(θ, C_{k+1}, T_{k+1});
R_{k+1}=generate_colocation_rule(θ, P_{k+1}, T_{k+1});
K=k+1;
}

In order to improve the efficiency of the algorithm, the candidate patterns generated can be filtered by the means of pruning. At this time, the minimum engagement threshold set by the user plays a key role. There are generally two pruning strategies, namely pruning and discriminating pruning based on engagement. The relevant pruning strategy is not going to be detailed here. For the details, please refer to Huang Y, Shekhar S, Xiong H [3].

2.4. Method and process of spatial co-location pattern mining
This paper mainly discusses the application of spatial co-location pattern mining in urban planning and studies how the various elements in the city are related to the distribution of spatial location, and how to take advantage of these links to provide suggestions for the expansion of the new urban area, so as to make the new urban area and the old city maintain the consistent development, which can facilitate people's lives.

Therefore, this article takes Chengdu as an example to study the spatial positional relationship of schools, hospitals, shopping malls, banks, etc. in Chengdu. The form of the data will be elaborated here.

Since the co-location relationship of each element is going to be sought, in order to use the full-join algorithm, the data form of all elements is as follows:

| Identification | Latitude  | Longitude |
|----------------|-----------|-----------|
| SC0001         | 30.67     | 104.06    |

The above data show that there are three columns in total, and the identifier refers to the nature of the data. For example, SC0001 in the example stands for primary school. Columns 2 and 3 represent the coordinates of the location of the elementary school.

3. Urban Element Analysis Based on Spatial Co-location Pattern Mining
This section will be divided into the following steps:
- Data extraction.
- Data preprocessing.
- Spatial co-location pattern mining.

3.1. Data extraction
The data in this part are obtained from the vector map of Chengdu through Mapinfo software [5]. In order to make the research data more meaningful, the elements of entertainment, shopping malls, community, school, hospital, and bank are selected as the data extraction objects. It can be found that these spatial
elements are related to people's lives, and it is easier to make people gather. Therefore, it is quite significant to study the relationship between these data.

Mapinfo software is used to open the vector map of Chengdu, and the school distribution map of Chengdu can be gained by selecting school. Using the coordinate extraction tool can help you get the coordinates of each school. By clicking "Export to txt format", the txt format file of the school coordinates can be obtained.

3.2. Data preprocessing

Through the above operation, the txt format coordinate file of Chengdu School can be obtained. The file is going to be preprocessed to make the required data more accurate. The data are stored in the Ecel, and deleting the duplicate items can help delete the useless repeating data.

3.3. Data analysis

Based on the data collected previously, by running the composed program, selecting the appropriate distance threshold and support threshold, and setting the different thresholds to observe the mining results, the following tables are used to represent the rules generated under the different thresholds. The min_conf represents the minimum confidence coefficient set by the user, and user_dist represents the distance threshold given by the user:

- Set min_prev=50%, min_conf=40%, user_dist=0.3.

Table 2. Mining results

| Number | Rule       | Sup  | Conf |
|--------|------------|------|------|
| 1      | SP→SC^HO  | 59%  | 44%  |
| 2      | SP^SC→HO  | 71%  | 40%  |
| 3      | SP^HO→SC  | 69%  | 68%  |

Among them, SP is a supermarket, SC is a school, and HO is a hospital. It can be seen that when the threshold set by the user is small, the rules that are mined are few and the confidence coefficient is low, it is a need for us to raise the threshold.

- Set min_prev=50%, min_conf=40%, user_dist=0.8.

Table 3. Mining results

| Number | Rule       | Sup  | Conf |
|--------|------------|------|------|
| 1      | BA→HO^SP  | 75%  | 69%  |
| 2      | SP^SC→CO  | 71%  | 65%  |
| 3      | BA→CO^PA  | 78%  | 70%  |
| 4      | CO^SP→HO^BA^SC | 73% | 58% |
| 5      | SP→SC^CO^PA | 70% | 55% |
| 6      | HO^BA^PA→CO^SP | 65% | 52% |
| 7      | HO^PA→SP^CO^BA | 72% | 65% |
| 8      | SC^CO^BA^PA→SP | 71% | 63% |

Among them, SP is a supermarket, SC is a school, HO is a hospital, CO is a community, BA is a bank, and PA is a leisure and entertainment area.

4. Conclusions

Through the above mining results, it can be known that in Chengdu, banks, supermarkets, hospitals, and leisure places are more likely to appear, and there is a greater possibility of supermarkets and communities appearing around the school. According to the above analysis of mining results, the characteristics of urban crowd gathering in Chengdu can be gained, which will provide a great reference
significance for the reorganization of the planning of urban old towns of other cities and the construction of new urban areas.

Acknowledgments
Finally, I would like to thank everyone who helped me throughout the writing process. Especially my teacher, Lizhen Wang, who has offered me a lot of help and guidance. When setting the title of the thesis, she first confirmed the direction of my topic which has given me the specific direction in writing. When formulating the outline of the thesis, I did not have a clear mind. Her help made me come up with the clear thinking all of a sudden. After completing the first draft, she carefully reviewed my writing and pointed out many problems existed. I am so grateful to her for her careful guidance which helps me successfully complete this paper.

References
[1] J, Han, J, Pei, Y, Yin. (2000) Mining frequent patterns without candidate generation. In: SIGMOD. Dallas, TX. pp. 1-12.
[2] Ester, M., Hans-Peter, K., Sander, J. (2000) Discovery in Spatial Database. Institute for Computer Science, 4: 193–216.
[3] Huang, Y., Shekhar, S., Xiong, H. (2004) Discovering Colocation Pattern from Spatial Data Sets: A general Approach. IEEE Transactions on Knowledge and Data Engineering, 16:1472-1485.
[4] Knorr, E.M., Ng, R.T. (1988) In: Algorithms for Mining Distanced-Based Outliers in Large Datasets. New York. pp. 392-403.
[5] S, Spielman, J, Thill. (2008) Social area analysis, data mining, and GIS. Computers, Environment and Urban Systems, 32: 110-122.