2016 International Conference on Electronic Information Technology and Intellectualization (ICEITI 2016)
ISBN: 978-1-60595-364-9

A Kind of Improved VGI Spatial Association Rule Mining Algorithm Based on Multi-level Semantic Constraints

Lingli Zhao, Shuai Liu, Junsheng Li and Hongwei Guo

ABSTRACT

VGI information is a kind of multi-source data absolutely, which contains spatial data and tagging data. VGI data have multi-level semantic. Spatial data mining is a demanding field since huge amounts of spatial data have been collected in various applications, ranging from remote sensing to geographical information systems, VGI data, computer cartography, environmental assessment and planning. The paper proposes a kind of VGI spatial association rule mining, which can extract frequent set quickly from VGI data. The experiment shows that the proposed algorithm is valid and efficient.

INTRODUCTION

Web 2.0 technologies enable users of social media to make contributions or to communicate with each other, especially Volunteered Geographic Information (VGI) data [1,2,3] . Among the various types of information contributed and shared by users on social media, the geographic one is called Volunteered Geographic Information (VGI) [4].The most common providers of VGI are Flickr, Open Stree tMap, Twitter, Facebook, YouTube, Wikimapia, Foursquare, etc. The amount of valuable data in VGI and different structures grew and technical progress made it possible to link these different systems and different structures, the wish to exchange and share these data arose and became more and more important.

Geographic data are real-world entities, also called spatial features, which have a location on the Earth's surface. Spatial features (e.g. Brazil, Belgium) belong to a feature type (e.g. country) and have both non-spatial attributes (e.g. name, population) and spatial attributes (geographic coordinates x, y). Besides the spatial attributes, there are implicit spatial relationships[5], which are intrinsic to geographic data, but usually not explicitly stored in geographic databases (GDB). Because of spatial relationships, real-world entities can affect the behavior of other features in the neighborhood, becoming the main characteristic to be considered in spatial data mining[6,7]. Spatial relationships are the main aspect in which knowledge discovery in geographic databases differs from knowledge discovery in transactional databases.

Spatial association rule mining is important component of spatial data mining. The idea of spatial association rule was presented by Koperski and Han firstly. Many association rule-mining algorithms have been proposed in the last few years, based on IPL(Inductive Logic Programming) method[8], Voronoi Diagram[9], Spatial Analysis[10], Immune Algorithms[11], overlay analysis and area calculation[12], multi-level relation[13], co-location patterns[14], cluster analysis[15]. Existing algorithms have only considered the data, while the multi-level thematic has not been considered. The paper proposes a kind of algorithm for extracting spatial association rules based on the multi-level semantic constraints, which can extract frequent set quickly from VGI data, and the experiments confirmed that the proposed algorithm is valid and efficient.

MINING SPATIAL ASSOCIATION RULES

Spatial association rules are implications of the form$X \to Y(c, s)$, where X and Y are sets of predicates, and at least one element in X or Y is a spatial predicate, and $X \cap Y = \emptyset$ [2]. Let $D\{d_1, d_2, \dots, d_k, \dots, d_n\}$be a set of items, and T be a set of rows $\{w_1, w_2, \dots, w_k, \dots, w_m\}$, where each $w \in T$ is a set of items such that $T \in D$. The support of the rule s, measures the percentage of transactions containing both the antecedent and consequent of the rule. The confidence of the rule c indicates that c of transactions that satisfy the antecedent of the rule will also satisfy the consequent of the rule. The support s of an item set X is the percentage of rows in which the item set X occurs as a subset. The support of the rule $X \to Y$ is given as $\text{support}(X \to Y)$. The confidence c is the probability factor that an item that contains the item set X also contains the consequent. The support of the rule $X \to Y$ is given as $\text{confidence}(X \to Y)$.

RELATIONS PREDICATES AND SEMANTIC HIERARCHY

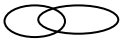
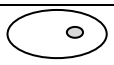
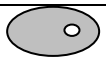
Spatial relation specifies how some object is located in space in relation to some reference object. The most well-known model of spatial relation is RCC-8[16],

which uses a 9-Intersection framework to identify the spatial relation between two regions.

Spatial Topological Relations Predicates

The region connection calculus (RCC) is intended to serve for qualitative spatial representation and reasoning. RCC abstractly describes regions (in Euclidean space, or in a topological space) by their possible relations to each other. RCC-8 consists of 8 basic relations, also is extended by 9 relations. The spatial configuration can be formalized in RCC-9 as shown in Table I. Spatial topological relations predicates are Boolean functions that return true if a test passes and false, otherwise, to determine if a specific relationship exists between a pair of geometries. 9 type of relations and predicates is descripted in Table I.

TABLE I. 9 TYPE OF RELATIONS AND PREDICATES.

Predicate	Representation	Description	RCC-9
close_to		If the second geometry is far from the first geometry.	Close To
far_from		If the second geometry is close to the first geometry.	Far From
adjacent_to		If none of the points common to both geometries i ntersect the interiors of both geometries.	EC
overlap_to		If none of the points common to both geometries i ntersect the interiors of both geometries.	PO
within		If the first geometry is completely within the second geometry.	PPI
contains		If the second geometry is completely contained by the first geometry.	NPPI
cover		If the second geometry is covered by the first geometry.	NTPP
cover_by		If the first geometry is covered by the second geometry.	TPP
equal_to		If two geometries of the same type have identical X,Y coordinate values.	EQ

Semantic Hierarchy of VGI

VGI information is a kind of multi-source data absolutely, which contains spatial data and tagging data. VGI data have multi-level semantic. The Semantic Hierarchy of VGI is a model of knowledge of large-scale space consisting of multiple interacting representations.

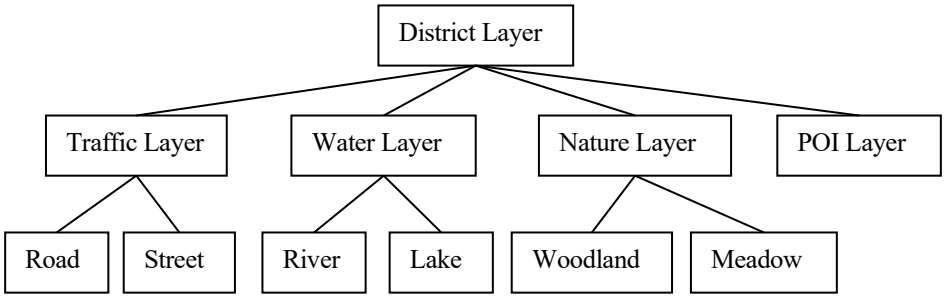


Figure 2. Semantic Hierarchy of VGI.

SPATIAL ASSOCIATION RULES MINING

Frequent set generation procedure is important component in spatial association rules mining. In this paper, the Apriori algorithm is improved based on the semantic constraints of VGI data. Table II shows the candidate generation procedure. Table III shows the frequent set generation procedure, which take care of the thematic constraints. The flow chart of frequent set generation algorithm is shown in Figure 3.

TABLE II. PSEUDO-CODE OF THE CAND IDATE-GENERATING FUNTION.

Function: apriori_gen
Description: Generate candidate set
Input: $D\{d_1, d_2, \ldots, d_k, \ldots, d_n\}$, L_{k-1}
Output: candidate sets C_k
Method:
For(i=1; i<k; i++)
{
If($l_i \cap l_k \neq \emptyset$)
Break;
}
If(i==k)
$C_k = L_{k-1}$ Join l_k;
Return C_k.

TABLE III. PSEUDO-CODE OF FREQUEST SET-GENERATION.

Function: apriori_freq
Description: Generate frequent set
Input: $D\{d_1, d_2, \ldots, d_k, \ldots, d_n\}$, minimum support, minimum confidence, thematic constraints w
Output: frequent sets L_k
Method:
L_1=predicate sets;
For(k=2; $L_{k-1} != \emptyset$; k++)
{ C_k= apriori_gen(L_{k-1});
C_w=subset(C_k,w);
If (candidates $c \in C_w$) c.count++;
else Remove c from C_k
L_k={ $c \in C_w$ |c.count≥minsup};}
Return L_k.

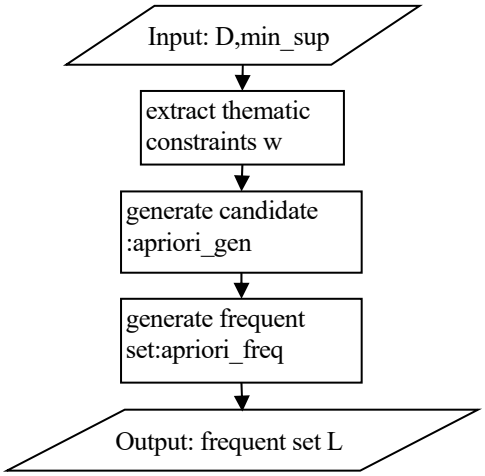


Figure 3. The flow chart of frequent set generation algorithm.

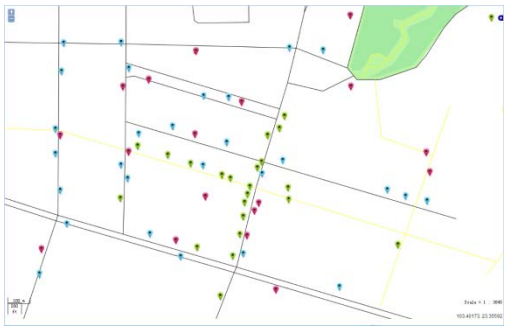


Figure 4. Experimental data of Open Street Map.

EXPERIMENTS

In order to verify the above algorithm, the experimental data of this paper uses the open source VGI data from Open Street Map, as shown in Figure 4. There are four attributes in every POI, namely tagging type, tagging information, user, tagging timestamp, which can be converted into predicates. In order to automatically check the effectiveness of POI information, it is necessary to extract the spatial association rules associated with it. For example, spatial association rules can be extracted as shown in Table IV.

Table IV. LARGE K-PREDICATE SETS AT THE SECOND LEVEL(500 POI).

K	Predicate set	Count
1	<adjacent to, school>	119
1	<adjacent to, downtown>	172
1	<adjacent to, community>	132
1	<close_to, road>	416
1	<far_from, highway>	457
2	<adjacent to, school><close_to, road>	114
2	<adjacent to, downtown><close_to, road>	152
2	<adjacent to, community><close_to, road>	129
3	<adjacent to, school><close_to, road><far_from, highway>	114
3	<adjacent to, community><close_to, road><far_from, highway>	152

CONCLUSION

Spatial data mining is used in areas such as remote sensing, traffic analysis, climate research, biomedical applications including medical imaging and disease diagnosis. The algorithm presented in this paper discusses efficient mining procedures for VGI.

ACKNOWLEDGMENT

This research is supported by National Natural Science Foundation (No. 41301442, 41201418), Honghe University Academic Leaderhead Reserve Talent Foundation (No.2014HB0201,2015GG0203), Honghe University Foundation (No.XJ15B06, XJ15B07).

REFERENCES

1. Sui D., Elwood S., Goodchild M.F. Crowdsourcing geographic knowledge: Volunteered geographic information (VGI) in theory and practice. Int. J. Geogr. Inf. Sci. 2014, 28, 847–849.
2. Clark A. Where 2.0 Australia's environment? Crowdsourcing, volunteered geographic information, and citizens acting as sensors for environmental sustainability. ISPRS Int. J. Geo-Inf. 2014, 3, 1058–1076.
3. Fast V., Rinner C. A systems perspective on Volunteered Geographic Information. ISPRS Int. J. Geo-Inf. 2014, 3, 1278–1292.
4. Goodchild M.F. Citizens as sensors: the world of volunteered geography [J]. Geo Journal, 2007, 69(4): 211-221.
5. Guting R.H. An introduction to spatial database systems [J]. The International Journal on Very Large Data Bases, 1994, 3(4): 357-399.
6. Koperski K., Han J. Discovery of spatial association rules in geographic information databases [C]. Advances in spatial databases, Springer Berlin Heidelberg, 1995: 47-66.
7. Ester M., Frommelt A., Kriegel H.P., et al. Spatial data mining: database primitives, algorithms and efficient DBMS support [J]. Data Mining and Knowledge Discovery, 2000, 4(2-3): 193-216
8. Hong Li, Zhihua Cai. ILP Method Applied in Spatial Association Rule Mining [J]. Computer Engineering and Applications, 2003, 16: 188-191, 197.
9. Guangqiang Li, Min Deng, Jianjun Zhu. Spatial Association Rules Mining Methods Based on Voronoi Diagram [J]. Geomatics and Information Science of Wuhan University, 2008, 12:1242-1245.
10. Chen Jiangping, Fu Zhongliang, BianFuling, et al. Mining Spatial Association Rules with Spatial Analysis [J]. Computer Engineering, 2003, 29(11): 29-31
11. Yu Zhu, Hong Zhang, Lingdong Zong. A New Spatial Association Rules Mining Method Based on Immune Algorithms [J]. Geomatics and Information Science of Wuhan University, 2009, 34(12): 1485-1489.
12. Dong Lin, Shu Hong, Niu Xiao. Spatial association rule mining based on overlay analysis and area calculation [J]. Geomatics and Information Science of Wuhan University, 2013, 38(1): 95-99.
13. Chen Jiangping, Li Pingxiang. An Algorithm about Spatial Association Rule Mining Based on Thematic [J]. Journal of Remote Sensing, 2006, 10(3): 289-293.

14. Shekhar S., Huang Y. Discovering spatial co-location patterns: A summary of results [M]. Advances in Spatial and Temporal Databases. Springer Berlin Heidelberg, 2001: 236-256.
15. Estivill-Castro V., Lee I. Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data [C]. The 6th International Conference on Geo computation. 2001: 24-26.
16. Randell, D.A., Cui, Z. and Cohn, A.G.: A spatial logic based on regions and connection, Proc. 3rd Int. Conf. on Knowledge Representation and Reasoning, Morgan Kaufmann, San Mateo, 1992: 165–176.