THE PRELIMINARY RESEARCH ON SPATIAL DATA-MINING THEORY AND METHOD BASED ON FORMAL CONCEPT ANALYSIS

Kun QIN[a] , Zequn GUAN[a], Deren LI[b], XinZhou WANG[c]

[a] The School of Remote Sensing Information Enginerring, Wuhan University,Wuhan,China,
qqqkkk@263.net ,zequng@public.wh.hb.cn
[b] National Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing,, Wuhan University ,Wuhan ,China ,
dli@wtusm.edu.cn
[c] The Ministry of Science and Technology,Wuhan University,Wuhan,China,
xzwang1@public.wh.hb.cn

Commission II, WG II/3

KEY WORDS: spatial data-mining, formal concept analysis, concept lattice, concept clustering, remote sensing image classification

ABSTRACT:

This paper emphatically discussed the theory of Formal Concept Analysis, and discussed the basic theory of spatial data-mining. According to the thought of Formal Concept Analysis theory, this paper put forward some new thoughts of spatial data-mining. Formal Concept Analysis (FCA), is a kind of very efficient tool for data analysis. It is a set-theoretic model that mathematically formulates the human understanding of concepts, and investigates the possible concepts in a given domain. Each node of concept lattice is a formal concept which includes two sections: extent, intent. Through Hasse graph, concept lattice can lively, simply and clearly embody the generalization relationship and characterization relationship among these concepts. The procedure to form concept from datasets materially is a kind of procedure of concept clustering. Formal Concept Analysis not only can find out the hierarchical clustering, but also can find out a good description about concept. Concept lattice can be used in many machine learning tasks, its main shortcoming is high complicity. To control the increase of node is very necessary. Spatial Data-mining, means extracting interesting spatial modes and characters, the universal connection between spatial data and non-spatial data from spatial databases, and other universal data characters which implied in the spatial databases. Spatial databases implies large knowledge about the surface features and their relationships among each other, These knowledge have very important values to remote sensing image understanding, GIS spatial analysis, spatial decision support etc.. It is a very good method to apply Formal Concept Analysis in spatial data-mining tasks.

1. INTRODUCTION

With the development of spatial data acquisition technologies (RS, GIS, GPS etc.) and computer information technologies, the quantity, size and complexity of spatial database are all increasing rapidly. People are submerged by spatial data. For example, remote sensing technologies have developed into a global stereo earth-oriented observation network which is multi-layer, multi-angle, all-azimuth and all-day; and have formed a configuration which realizing the combination of high, middle, low orbits and the integration of high, middle, low resolutions. The abundant data greatly satisfies the potential needs to research earth resources and environment, widens the information resources which can be utilized. However, because of the relatively draggling of spatial data processing technologies, large quantity of data are set aside. People increasingly need to utilize the spatial data in higher level. In order to adapt to this kind of need, Spatial Data Mining technology was put forward. Spatial Data Mining, which also called Knowledge Discovery from Spatial Databases, means extracting interesting spatial modes and characters, the universal connection between spatial data and non-spatial data from spatial databases, and other universal data characters which implied in the spatial databases. Spatial databases implies large knowledge about the surface features and their relationships among each other, such as the surface features' locations, shapes, characters, states and spatial distribution, spatial association, spatial clustering and spatial evolvement etc.. These knowledge have very important values to remote sensing image understanding, GIS spatial analysis, spatial decision support etc.. The research on spatial data-mining and knowledge discovery from databases has been paid more and more attention.

The traditional data-mining methods is process-centered, knowledge discovery in databases is considered an interactive and iterative process between a human and a database that may strongly involve background knowledge of the analysing domain expert. Much attention and effort has been focused on the development of data-mining techniques but only a minor effort has been devoted to the development of tools that support the analyst in the overall discovery task. So it is need to emphasize the process-orientation of KDD tasks and argue in favour of a more human-centered approach for successful development of knowledge-discovery support tools.Human-centered KDD refers to the constitutive character of human interpretation for the discovery of knowledge, and stresses the complex, interactive process of KDD as being lead by human thought.

Formal Concept Analysis (FCA), which also called concept lattice (or Galois lattice), is a kind of very effective tool for data analysis, and developed very rapidly in recent years. Formal Concept Analysis is a set-theoretic model that mathematically formulates the human understanding of concepts, and investigates the algebraic structure, Galois lattice, of possible concepts in a given domain. Formal Concept Analysis provides an effective tool to support data analysis. Each node of concept lattice is a formal concept, which includes two sections: extent, the examples covered by the concept; intent, the description of concept, representing the common characters of those examples covered by concept. Through Hasse graph, concept lattice can lively, simply and clearly embody the generalization relationship and characterization relationship among these concepts. Therefore, Concept lattice is considered as the efficient tool to analyze data. The procedure to form concept from datasets (called formal background in concept lattice) materially is a kind of procedure of concept clustering. Formal Concept Analysis is a top-down, increment, hill-climbing classification method. The main difference between Formal Concept Analysis and traditional non-supervise clustering lies in that the former not only can find out the hierarchical clustering, but also can find out a good description about concept. Concept lattice can be used in many machine learning tasks. Now, concept lattice has been applied to many fields, such as information retrieving, digital library, software engineering and knowledge discovery etc.. The main shortcoming of Formal Concept Analysis is high complicity. In the worst conditions, the node number in concept lattice increase in exponent mode. To control the increase of node is very necessary. For some algorithms of building lattice like Bordat, the support threshold should be introduced. In the procedure of building lattice, in order to decrease branches, those nodes which support degree less than support threshold should not be developed. For some other algorithms of building lattice, the conditions are more complex, especially for increment algorithm, the pruning in the procedure of building lattice is needed. In order to maintain the characters of lattice, pruning only can be started to do from the bottom of the lattice, and the parents nodes number of these bottom nodes is 2 in the most. After the pruning of bottom nodes, the new layer's nodes which can be pruned displays. In this case, statistical methods can be used as the pruning principle.

Based on the discussion of spatial data mining theory and method and the analysis research of Formal Concept Analysis theory and method, this paper put forward to the methods to apply Formal Concept Analysis theory and method to spatial data mining. The method can realize several spatial data mining tasks, such as spatial association rule analysis, remote sensing classification etc..

2. FORMAL CONCEPT ANALYSIS

2.1 Theory Background

The theory of Formal Concept Analysis was firstly introduced by Rudolf Wille, a German mathematician in 1982.Until now, it has developed for 20 years. Formal Concept Analysis is a field of applied mathematics based on mathematization of concept and conceptual hierarchy. It thereby activates mathematical thinking for conceptual data analysis and knowledge processing. The underlying notions of "concept" evolved early in the philosophical theory of concepts and still has efforts today. In mathematics, it played a special role during the emergence of mathematical logic in the 19th century. Subsequently, however, it had virtually no impact on mathematical thinking. It was not until 1979 that the topic was revisited and treated more thoroughly. It is perfectly possible to use Formal Concept Analysis when examining human conceptual thinking. The adjective "formal" in the name of Formal Concept Analysis means we are dealing with a mathematical field of work. Through the connections with the structured concept related to these work, comprehensible, significative knowledge can be discoveried.

Formal Concept Analysis is based on the order theory, especially complete lattice theory, in mathematics. In order to thoroughly comprehend the theory and method of Formal Concept Analysis, Some related theories and concepts should be explained firstly.

Lattice theory: Lattice theory is a branch of algebra, it is a kind of important tool to research algebra, geometry, topology, measure, functional, combinatorics, digitial computer and fuzzy mathematics. Lattice theory is related to the following concepts.

Ordered sets: Assume P is a set, if the binary relation $\leqslant$ of P satisfies the following three conditions, then it is called **order relation**(or shortly an **order**), the set of （P，$\leqslant$） which is composed of a series of order relation is called **ordered sets.**

(1)reflexivity: $a \leqslant a$，（$\forall a \in P$）, means element a has a binary relation with itself.
(2)antisymmetry : $a \leqslant b$，$b \leqslant a \Rightarrow a = b$（$\forall a, b \in P$）.
(3)transitivity: $a \leqslant b$，$b \leqslant c \Rightarrow a \leqslant c$（$\forall a, b, c \in P$）.

Lattice: In a ordered set of （P，$\leqslant$）, if random two different elements x, y have supremum $x \vee y$ and infimum $x \wedge y$, then the ordered set of （P，$\leqslant$）is called a **lattice**.

Cover: Assume a, b are random two different elements in a ordered sets of （P，$\leqslant$）, if $a<b$, but no $x \in P$ satisfies $a<x<b$ （$x \neq a$，$x \neq b$）, then call a is a lower neighbour of b , and b is a above neighbour of a, or call "b cover a", and record: $a \prec b$.

Hasse graph(line grah): In a limited ordered sets（P，$\leqslant$）, order relation $\leqslant$ was completely decided by the element pairs which have cover relation(for example, $a \prec b$). If the every element in set P is denoted by a circle in a plane. If and only if b cover a, then draw b above a ,and connect a, b through a line, then the graph is called a **line graph** of ordered set of （P，$\leqslant$）, also called **Hasse graph**. Through Hasse graph, the relationship among the elements in ordered sets can be denoted. The following is a simple example which can explain the construction process of Hasse graph.

Assume M={1,2,3,4,…,12}，$\mid$ denotes the exactly dividing relation,（M, $\mid$）is a limited ordered set, its Hasse graph like figure 1.

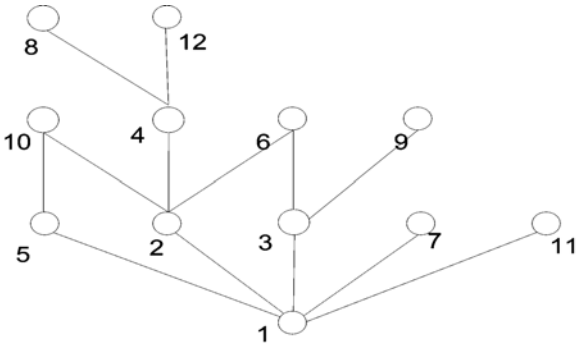

Figure 1. Hasse graph

2.2 Formal Concept Analysis

Concepts are necessary for expressing human knowledge. Therefore, the process of discovering knowledge in databases benefits from a comprehensive formalization of concepts which can be activated to communicatively represent knowledge coded in databases. Formal Concept Analysis offers such a formalization by mathematizing concepts that are understood as units of thought constituted by their extension and intension. To allow a mathematical description of extensions and intensions, Formal Concept Analysis always starts with a formal context defined as a triple (G, M, I), where G is a set of (formal) objects, M is a set of (formal) attributes, and I is a binary relation between G and M(i.e. $I \subseteq G \times M$); in general, gIm($\Leftrightarrow$ (g，m)$\in$ I) is read: "the object g has the attribute m".

A formal concept of a formal context (G, M, I) is defined as a pair (A, B) with $A \subseteq G$ and $B \subseteq M$, such that (A, B) is maximal with the property $A \times B \subseteq I$; the sets A and B are called the extent and the intent of the formal concept (A, B). The subconcept-superconcept relation is formalized by （A_1,B_1）$\leqslant$ (A_2,B_2)： $\Leftrightarrow$ $A_1 \subseteq A_2$ （$\Leftrightarrow$ $B_1 \supseteq B_2$）. The set of all concepts of a context (G, M, I) together with the order relation $\leqslant$ is always a complete lattice, called the concept lattice of (G, M, I) and denoted by $\underline{B}$ (G，M，I)

In a line graph of a concept lattice, the name of an object g is always attached to the circle representing the smallest concept with g in its extent (denoted by γg); dually, the name of an attribute m is always attached to the circle representing the largest concept with m in its intent (denoted by μm). This labelling allows us to read the context relation from the graph because gIm $\Leftrightarrow$ γg $\leqslant$ μm, in words: the object g has the attribute m if and only if there is an ascending path from the circle representing γg to the circle representing μm. The extent and intent of each concept (A, B) can also be recognized because A={g$\in$G$\mid$γg$\leq$(A,B)} and B={ m$\in$M$\mid$(A,B) $\leq$μm}.

Graphically represented concept lattices have proven to be extremely useful in discovering and understanding conceptual relationships in given data. Therefore a theory of "conceptual data systems" has been developed to activate concept lattices as query structures for databases. For example, based on Formal Concept Analysis theory, the system TOSCANA has been developed. Conceptual data systems activated by the management system TOSCANA can be considered as knowledge discovery support environments that promote human-centered discovery processes and representations of their findings.

3. SPATIAL DATA-MINING THOUGHTS BASED ON FORMAL CONCEPT ANALYSIS

The procedure to form concept from datasets (called formal background in concept lattice) materially is a kind of procedure of concept clustering. Formal Concept Analysis is a top-down, increment, hill-climbing classification method. The main difference between Formal Concept Analysis and traditional non-supervise clustering lies in that the former not only can find out the hierarchical clustering, but also can find out a good description about concept.

To apply Formal Concept Analysis theory and method to spatial data mining can realize several spatial data mining tasks, such as spatial association rule analysis, remote sensing classification etc.. Formal Concept Analysis can be applied to remote sensing image classification. The procedure includes the following steps: 1) carry through pre-processing of remote sensing images, extract some image characters, such as regional segmentation information, shape information, texture information, statistical information etc., and extract some auxiliary information through GIS analysis; 2) building up the Hasse graph of Formal Concept Analysis based on the above information; 3) through the concept clustering procedure of Formal Concept Analysis, carry out the remote sensing image classification. In the processing procedure, some key problems connected with the Formal Concept Analysis in spatial data mining must be dealt with, such as how to reduce the complicity of building concept lattice, how to extract the characteristic information of image, how to utilize the auxiliary information of GIS analysis etc.

Formal Concept Analysis theory can also be applied to spatial data-mining from GIS data. In general, GIS data consist of spatial(location) data and attribute(descriptive) data. The relationships of spatial-attribute is similar to the relationships of g(object) and m(attribute) in a triple of (G, M, I). So we can use the Formal Concept Analysis theory and methods to build a Hasse graph of GIS, through the concept hierarchy of Hasse graph, some spatial knowledge can be discoveried. For example , the spatial relationship, the relationships between geometry locations and their attributes etc..

4. CONCLUSIONS

This paper mainly discussed the theory and method of Formal Concept Analysis, and put forward some thoughts to apply Formal Concept Analysis to spatial data-mining.

Formal Concept Analysis is a kind of very efficient tool for data analysis, it is top-down, increment, hill-climbing classification method. This theory can be applied to spatial data-mining from GIS data, and remote sensing image classification.The main shortcoming of Formal Concept Analysis is high complicity. To control the increase of node is very necessary.

Further research on spatial data-mining based on Concept Concept Analysis aims at the following several points:

1. Research more efficient method of building lattice for spatial data.
2. Research how to decrease the complexity.
3. As Formal Concept Analysis and Description logics are closely related and have similar purposes, research how to

combin logic-based components with FCA to improve the efficiency of spatial data-mining.
4. Research how to incorporate statistical and computational components with FCA to improve the methods of spatial data-mining.

ACKNOWLEDGEMENTS

This paper is supported by the National Natural Science Fund of China (No.40023004) and by the award of the first times of excellent doctoral dissertation, China(No.199936) and by the award of National High Technology R&D Program(863)(No.2001AA135081).

REFERENCES

Bohm, C., Braunmuller, B. and Kriegel, H., 2000. The pruning power: theory and heuristics for mining databases with multiple k-nearest-neighbor queries. *Proc. Int.Conf. on Data Warehousing and Knowledge Discovery(DaWak 2000)*,Greenwich,U.K..

Brachman, R.J. et.al., 1993. Integrated support for data archaeology. *Int.J. of Intelligent and Cooperative Information System*s, 2(2), pp.159-185.

Ester, M., Kriegel, H.-P and Xu, X.,1995. Knowledge discovery in large spatial databases: focusing techniques for efficient class identification. In: *Proc.4th Int. Symp. On Large Spatial Databases(SSD'95)*, pp.67-82, Portland, Maine, August 1995.

Ester M., et al.,1996a. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proc of 2nd International Conference of Knowledge Discovering in Databases and Data Mining(KDD-96)*. Portland: AAA Press.

Ester, M., Kriegel, H., Sander, J. and Xu, X., 1996b. A density-based algorithms for discovery clusters in large spatial databases with noise. In: *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining(KDD-96)*.

Ester Martin, Frommelt Alexander, Kriegel Hans-Peter, Sander Jorg, 1998a. Algorithms for characterization and trend dection in spatial databases. In: *The Proceedings of 4th International Conference on Knowledge Discovery and Data Mining(KDD-98)*

Ester, M., Kriegel, H., Sander, J. and Xu, X., 1998b. Clustering for mining in large spatial databases. In: *Special Issue on Data Mining, KI-Journal*, ScienTec Publishing, 1998, Vol.1.

Ganter,·B. and Wille, R., 1999. *Formal concept analysis: Mathematical Foundations*. Berlin: Springer.

GUO Jiansheng, ZHAO Yi, SHI Peng-fei, 2001. An efficient dyanmic conceptual clustering algorithm for data mining. *Journal of Software*, 2001, No.4.

Ho, T.B., 1997. Incremental conceptual clustering in the framework of Galois lattice. In: Lu H, Motoda H, in H, eds. *KDD: Techniques and Applications*.World Scientific,1997, pp.49-64.

HU Changliu, 1990. *The basis of lattice theory*. The Press of Henna University.

HU Keyun, LU Yuchang, SHI Chunyi, 2000. Advanced in concept lattice and its application. *Journal of Tsinghua University(Sci & Tech)*, 2000, No.9.

Koperski, K. and Han, J., 1995. Discovery of spatial association rules in geographic information databases. In: *Advances in Spatial Databases, Proc. Of 4th Symp. SSD'95,* Springer-Vering, Berlin,1995, pp.47-66.

LI Deren and MA Fei, 1998. A general mathematical model for GIS spatial analysis via raster data. In: *Geoinformatics'98 Conference Proceedings*, Beijing.

LI Deren, WANG Shuliang, SHI Wenzhong, WANG Xinzhou, 2001. On spatial data mining and knowledge discovery(SDMKD). *Geomatics and Information Science of Wuhan University*, 2001, No.6.

LIN Hong Fei, ZHAN XueGuang, YAO TianShun, 2000. Text structure analysis based on concept. *Journal of Computer Research & Development*, 2000, No.3.

LIU Mingjie, WANG Xiufeng, LI Baolin, 2001. A knowledge discovery method based on multi-level concept generalization. *Journal of computer science*, Vol.28, No.3.

Marchisio, G.B., Krzysztof Koperski and Michael Sanella, Querying Remote Sensing and GIS Repositories with Spatial Association Rules.

Marchisio, G.B. and Alan Q. Li., 1999. Intelligent system technologies for remote sensing repositories. In: *Information Processing in Remote Sensing*, World Scientific Publishing,1999.

Soh, Leen-Kiat, 1999. Segmentation of satellite imagery of natural scenes using data mining. *IEEE Transactions on Geoscience and Remote Sensing*, 1999, No. 2.

Thomas Brinkholf, Hans-Peter Kriegel, 1994. The impact of global clustering on spatial database systems. In: *Proceedings of the 20th VLDB Conference*, Santiago,Chile

XIE Zhipeng, LIU ZongTian, 2000. Concept lattice and association rule discovery. *Journal of Computer Research & Development*, 2000, No.12.

XU XiaoWei, Ester Martin, Kriegel Hans-Peter, Sander Jorg, 1998. A distributed-based clustering algorithm for mining in large spatial databases. In: *The Proceedings of 14th International Conference on Data Engineering(ICDE'98)*.

ZHANG Huajie, QIANG Fangzuo, YUAN Guobin, 1997. The representation and clustering of numeric attributes in concept formation. *Journal of Software*, 1997, No.6.

ZHU Shaowen etc., 2001. A method of mining multiple-level quantitative association rules. *Journal of computer science*, 2001, Vol.28,No.2.

http://www.informatik,uni-muenchen.de

http://www.cis.unisa.edu.au

http://astro.u-strasbg.fr/~fmurtagh/mda-sw/online-sw.html