Image Mining for Generating Ontology Databases of Geographical Entities

Zhenfeng Shao 1, Jun Liu 2 and Xianqiang Zhu 1

1 State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan 430079, China.

2 School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China.

Abstract: This paper extracts the basic geographic information from remote sensing images at first, and then studies the resolution granularity of the remote sensing images which can be applied to distinguish the features of corresponding objects by adopting global-covered remote sensing images with multi-frequency spectra and multi-resolution. Thus necessary feature information for the geographical ontology database, such as texture characteristic information can be mined and through our data mining strategy from remote sensing images based on the formal concept analysis theory, data mining methods for texture features are achieved. The emphases of this paper are the mining method for texture characteristic for generating ontology database of the geographical entity. By mining the texture characteristics, we can find the partial structure that frequently appears in the remote sensing image data, and find the restriction relationship between the central pixel and its neighborhood pixels in partial regions of images. This process is constituted by the following four steps: sampling areas partition normalized processing, characteristic data mining, building Hasse graph and generating rules. Through the computation about remote sensing image data mining, we put the uncertainty problem about characteristics form data mining up to a height of information theory and study it, and find the consolidate mathematics expression between information quantity and uncertainty about the characteristics in order to resolve the quantitative evaluation problem between information quantity and uncertainty of remote sensing image. This paper introduces the concepts-driven data mining framework to uncertainty process, so as to guide the idiographic algorithm and process during the image mining procedure. According to the characteristic of remote sensing images, combining with all kinds of GIS data, we can describe the essential characteristics that build ontology database of the geographical entity.

Keywords: image mining, ontology database, semantic-based spatial information sharing and interoperability, uncertainty

1. Introduction

The emergence of Grid [1] technology provides the possibility to eliminate the obstacles among different geographic information systems. But it is still a problem that how to achieve spatial information sharing and spiritual services after connecting different spatial information systems seamlessly. Considering that people in different areas commonly may have different cognition of the real world or may describe the same geographical phenomenon or object from totally different angles, the generation of semantic heterogeneity is inevitable, resulting in so-called semantic gap or islands of information. Actually, this phenomenon even occurs in the stage of spatial data collection clearly. Take an application of land use classification as an instance. The agricultural sector and the forestry sector may divide the same block into different types in accordance with their industry standards in land use classification respectively, which results in the semantic heterogeneity for the same spatial object. How to eliminate the semantic heterogeneity has been an urgent problem for spatial information sharing.

Just as academician Li Deren once pointed out that semantic conflict is the most important one of the four pressing problems in the sharing of spatial information systems [3], it is difficult to support the function of deep mining of GIS Data only by traditional mapping products due to their limitations on the amount of
information. For this reason, this paper describes an effective application framework to eliminate semantic heterogeneity. In this paper a GIS sharing ontology database which has been proved to be adaptive for the description of the intrinsic characteristics of spatial objects maximally is established and through it the elements for this database can be mined from remote sensing images.

Remote sensing images contain total elements of the geographic information that can fully support the GIS data mining. From images, a real environmental, social, demographic, economic, cultural and other information can be acquired so as to help users complete in-depth development in the management, decision-making, tracking, and other advanced functions, and bring the function of GIS decision and supporting into play as effective as it possible, which had to be finished difficultly through interpreting the map including many arcane semantics in the past.

For example, the public security department can only collect about 20% police geographic information from traditional 4D products (DEM, DOM, DLG, DRG). The result of the absence of social attributes information is that it can not meet the needs of users of the public and the majority of trades directly. When we establish various professional GIS databases by remote sensing images classification, since semantic classification is not created for image content, the result is the generation of the semantic gap in different industries. However, data mining based on remote sensing images can make up the deficiency of 4D products and further can construct GIS ontology database directly for the purpose of mining mine total elements of sharing information for such industries as police, municipal services, transportation, telecommunications, finance, health care, navigation, LBS, and so on. By combing the attributes of remote sensing images including visual, measurable and minable with current 4D products, a lot of useful knowledge which are necessary for all industries can be acquired and the construction of GIS ontology database based on data mining from remote sensing images is possible.

The aim of this paper is to study how to extract necessary characteristics as more as possible for general ontology database from remote sensing images in order to achieve semantic spatial information sharing. Spatial semantic information sharing is our research focus and a proven effective solution about spatial intelligent mining based on remote sensing images, which is conducive to the establishment of GIS general ontology database and the sharing of spatial information and services, is proposed in this paper.

In the field of data mining, scholars and researchers from different countries have made unremitting efforts and exploration and they already have a series of fruitful achievements.

Dr. Fayyad proposed the idea of mining non-standard and multimedia image data (Fayyad, 1993) in 1993. He used decision tree and statistical optimization analysis method to classify the astronomical objects of the images. His idea was helpful to identify fuzzy stars in the astronomical images and had been applied to the study of detection of volcanic Venus.

Researchers from Simon Fraser University of Canada used multi-dimensional analysis technology to create image data cube, realized finding of many types of knowledge including knowledge of classification, knowledge of association rules and so on[5].

In American California University at Santa Barbara, a research team led by professor Manjunath carried out study on the association rules among the image objects based on special incident cube. The basic idea of their research was to divide an image with certain grid size, to analyze the content of images (color, texture), to label the content of the image, to establish the incident cube space according to a large number of image objects and association between them, and to analyze and mine the association rule of these image objects. Professor Li Qi, Du Mingyi, Li YiFan, Xu Mingjie at el. researched on the mining method from spatial dataware (including images) [6] [7] [8] [9].

Scholars from Georgia Institute of Technology in American studied the association rules mining of the images objects by content-based remote sensing image retrieval technology. Scholars in North Dakota university researched remote sensing image mining based on association rules, and applied it to the precision agriculture field. In the process of association rule mining, they took RGB value of each pixel of the major training data and agricultural output as the item sets. Ma Chaofei, a Dr. of Institute of Remote Sensing Application of the Chinese Academy of Sciences, studied mining and application based on association rules from remote sensing data. He mainly focused on the attribute data associated with remote sensing images.
such as soil erosion model parameters, the source of sandstorms ground temperature, and soil and water information [17]. A professor of this institute, Ma Jianwen, carried out researches on understanding and data mining from optical satellite remote sensing images in the view of spatial angel theory[18].

Zhou Chenghu, a professor of Geographic Institute of Chinese Academy of Sciences, carried out researches based on information entropy of the spatial data (including remote sensing data) mining model[19]. Dr. Zhang Jianting carried out researches based on information entropy of the data (including remote sensing data) mining and knowledge discovery model[20].

A research group of Wuhan University led by Academician Li Deren studied knowledge discovery based on inductive learning methods, applied them to remote sensing images classification, and carried out researches on data mining theory and methods based on the formula concept analysis[4] [10].In the application, NASA extracted semantic information of knowledge from the images automatically and applied to the crater terrain concretely and other aspects.

2. Framework for image mining

Fig.1 illustrates the framework of our research. Through data mining from GIS data and remote sensing images combining with ontology theory, we can establish the semantic sharing ontology database.

Data mining from Remote sensing images is a process of concept formation. The most abundant amount of information of remote sensing images is the basis of information mining. The purpose of remote sensing image data mining is to analyze and mine potential knowledge from massive data and other relevant data, to extract all level concepts and relationships between these concepts. In this paper, we extract characteristic information of geographic entities for further ontology database construction from the view of data mining. The main work in this paper is as follow:

2.1. Generation of the concept of a geographical entity

In psychology, the concept refers to thinking pattern which reflects common characteristics and nature attributes of the objects. It is the basic unit of higher cognitive activities. The things which are included in a concept share the common attributes and characteristics. Each concept of the geographical entity includes the connotation and extension. Connotation refers to the concept reflected in the essence of things, and extension refers to the scope of the concept. Different concepts have different hierarchies, or different levels. In this paper, there are two ways to form concepts:

The first one is artificial selection. Through the support of experts in the field, we can choose concepts of this filed, and give the corresponding semantic information, as well as the levels and relationships between concepts, so as to achieve the purpose of forming concept.

The second one is semi-automatic. As for the existing GIS data, such as land-use maps, plans, and other classification maps, different categories itself are the bottom concepts within the field, and formed by
classification of remote sensing images automatically. Concepts after classification have corresponding semantic information. Supporting by expertise, we can establish the levels and relationships between these concepts, so as to achieve the purpose of forming concept.

Concepts are organized by the form of Ontology, and ontology includes concepts as well as the relationships between them. Concepts formed by these two channels are incomplete, or, only with the level and contained relations between the concepts. It is necessary to specify the nature, attributes, connotation and extension, and to do so we can use data mining tools.

2.2. Geographical entity feature information mining

The process of extracting the concept and mining geographic entities feature information from remote sensing images is a process of continuous learning, thereby creating the knowledge. First, the training samples’ selection principles and process need to be solved. Training samples should follow the principles of accuracy, representativeness and statistical. Accuracy is to ensure the consistency of the selected plots and the actual objects. Representation on the one hand refers to that the choice of a particular geographical region is a representation of an entity, on the other hand that consideration should be taken to the complexity of the geographical entity itself. Statistics refers to the choice of training plots must have sufficient number of pixels. In this paper, choosing sample regions in a particular type of specific resolution remote sensing images through artificial outlined or automatically matched every concept, calculating the texture and shape features of the sample, producing the association rules reflected on the concept which generated in the image, and providing the basis of target recognition and matching, semantic sharing and interoperability.

Secondly it is necessary to solve the problem that how to define the conceptual level. Actually it is the problem that how to determine the granularity of remote sensing image data mining, in the image data mining it refers to the unit size of objects. In remote sensing image data mining the smallest granularity is the pixel granularity, and the unit of the object is the pixel. The coarser granularity is the pixel group granularity, the unit of the target is pixel group which is a collection of pixels. Different pixel Groups can be constructed by various different methods. Therefore according to the size and type of the mining target, we choose appropriate resolution remote sensing image for research.

3. Concept assignment and methodology

By mining texture association rules we can find the local structure frequently appeared in the remote sensing data, and also can find the mutual restraint relations between the local area around the center pixel and its neighbors in portion area of the images. Association rules is one of the most important contents in data mining, at present, most of the theory and technology about date mining focus on the study of association rules. The concepts of association rules were proposed by Agrawal, Imielinski and Swami. They are simple but very practical rules in data. The model of association rules is a description model. Mining association rules is to discover the relationship in the data collection, and to reveal the natural laws. With the collection and storage of massive data, mining the association rules in database becomes more and more important.

Now we consider some transaction involving many items, for example, object A appears in transaction one, object B appears in transaction two, and both A and B appear in transaction three, then, are there some rules about the appearance of object A and object B? In the knowledge discovery in database, association rule is the knowledge model which describes the rules of objects appearing in transactions at the same time. More exactly, association rules describe how the appearance of object A affects the appearance of object B using qualitative number.

Some definition key concepts are given as follows in this paper:

Association rule: set I={i1,i2,…,in} as sets of items. Set data D as sets of database transaction, of which every transaction T is the sets of items, that is \( T \subseteq I \). Every transaction has a identifier, named TID. Set A as a item, which is contained in transaction T, and \( A \subseteq T \). Association rule is the expression like \( A \Rightarrow B \), in which \( A \subseteq I, B \subseteq I \), and the intersection of A and B is null.

K-item sets: a set formed by k items.

Frequent item set: If the appearance frequency of an item set in transaction sets D is bigger or equal to the product of min_sup and total of transaction D, it is name frequent item set.
Root pixel: The center pixel of $n \times n$ neighborhood image is named root pixel.

Item: In the neighborhood of a specific root pixel, every pixel is an item, which is expressed by a ternary group $(X,Y,I)$, in which $X$ is the offset in horizon, $Y$ is the offset in vertical, and $I$ is the gray value of pixel with offset $X$ and $Y$. If it is the root pixel, its $X=0$ and $Y=0$. So, in a $n \times n$ neighborhood image, there are $n^2G$ items, where $G$ is the possible number of gray level. In order to avoid too much items, it is very necessary to compress the gray level for 8 bit images or 24 bit images. In general, they are compressed into 3 or 4 or 8 bit images, so $n$ is 3 or 4 or 5, thus, the number $n^2G$ will be much smaller. For the convenience of calculating, we suppose $n=3$ and $G=3$, then the number of item is $3 \times 3 \times 3 = 27$.

Item set: item set is the set of items, and the number of items is cardinal number of this item set. For texture image mining, a item set corresponded with a root pixel is a collection of items in its 8-neighborhood images. For example, considering a root pixel, we will get a 9-item set like: $\{(1,0,0), (0,0,0), (1,1,1)\}$.

Transaction: A specific transaction is relevant to a item set of root pixel. Every root pixel corresponds to only one transaction. For example, a transaction formed by a 9-item set is: $\{(1,0,0), (0,0,0), (1,1,1)\}$.

4. Implementation and experimental results

4.1. Image data preparation

Based on the size of the object, we select remote sensing images with appropriate resolution and corresponding GIS data, such as all kinds of classification maps, with the remote sensing images in the same area, cut out the images that corresponding to the object which is expressed with different concepts in different classification maps. With these images, we can prepare for revealing the nature characteristics of the object by association rules mining from the texture images, and also prepare for the building of ontology graphic. Taking the calculation into account, that the sample images should not be too big, we use the sample image with 30*30 pixels.

4.2. Association rules mining from texture images

By association rules mining, we can find the frequent local structure in remote sensing data, as well as the restraint relationship between the center pixel and its neighbor pixels. The concrete flow is: cutting sample image, normalizing, storing into database, generating Hasse graphic, generating rules.

- Normalized of image

  After calculating the maximum $n_{\text{Max}}$ and minimum $n_{\text{Min}}$ of the gray value of all pixels, we divide the values into three levels: $[n_{\text{Min}},n_{\text{Min}} + (n_{\text{Max}} - n_{\text{Min}})/3],[n_{\text{Min}} + (n_{\text{Max}} - n_{\text{Min}})/3,n_{\text{Min}} + (n_{\text{Max}} - n_{\text{Min}})/3*2],[n_{\text{Min}} + (n_{\text{Max}} - n_{\text{Min}})/3*2,n_{\text{Max}}]$, Corresponding to the value of 0,1,2 after treatment.

- Storage of texture characteristics

  Form a database record with every pixel on the normalized images and its corresponding 8 neighborhoods pixels that are also be normalized, and then import this record into the texture feature database, that is, texture features data storage.

  For a specific root pixel, if the coordinate offset is 0 and 1 in direction X and direction Y, that is to say, neighborhood size is $3 \times 3$, the corresponding number of the maximum data item is 9, they are: $(0,0,0),(1,0,0),(0,1,0),(1,1,0),(0,0,1),(1,0,1),(0,1,1),(1,1,1)$, $I$ is the gray value. For a image of 30*30, the number of root pixel is $(30-3+1)^2 = 784$. Because every root pixel has a corresponding transaction, so the image of 30*30 can be conversed to 784 transactions. Then mine the association rules based on the concept lattice.

- Generation of concept lattices and Hasse graphic

  According to the texture characteristic, using the concept lattice generation arithmetic, we generate the concept lattices that are corresponding to the texture characteristic and draw Hasse graphic.

  The methods of constructing concept lattices are divided into two classes: batch process algorithm and incremental algorithm. Due to their different lattice structure, batch processing algorithm can be divided into three categories, namely, the top-down algorithm, the bottom-up algorithm and the enumeration algorithm.
In the top-down algorithm, first construct the top of the lattice nodes, and then gradually go down, such as Bordat algorithm (Godin, 1995), OSHAM algorithm (Ho, 1995), and so on. Bottom-up algorithm is opposite, construct the bottom nodes, then expand upward, such as Chein algorithm (Godin, 1995). In enumeration algorithm, all the nodes are enumerated with a certain order, and then we can generate Hasse graphic which shows the relationship between the nodes. Ganter algorithm (Godin, 1995) and Nourine algorithm (Nourine and Raynaud, 1999) are of this type. An incremental concept formation is named that reasoning the level of concepts by an incremental approach. It is so called incremental algorithm that only scanning the database once and inserting into the specific lattice \( L \) a new transaction \( T \) so as to generate a new lattice \( L' \). The Godin algorithm is the classic incremental algorithm. It compares the nodes in new lattice \( L' \) to those in former \( L \), and divides the nodes into three types: steady nodes, updated nodes and incremental nodes. After a new node generated, the edges of new lattice must be updated. The generators of new nodes are also child nodes of new nodes, and the edges of former father nodes of generators must be connected again. There may be other child nodes for a new node, if yes, they are also new nodes, if no, the new node can be generated by its two child nodes.

- **Generation of association rules:**
  Setting support = 0.3 and confidence = 0.8, we can get 12 association rules as follows:

<table>
<thead>
<tr>
<th>Rule number</th>
<th>rule</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>{(1,0,0)} \Rightarrow {(0,0,0)}</td>
<td>0.409832</td>
<td>0.801980</td>
</tr>
<tr>
<td>(2)</td>
<td>{(1,1,0)} \Rightarrow {(1,0,0)}</td>
<td>0.320984</td>
<td>0.801619</td>
</tr>
<tr>
<td>(3)</td>
<td>{(0,1,2)} \Rightarrow {(0,0,2)}</td>
<td>0.296737</td>
<td>0.923469</td>
</tr>
<tr>
<td>(4)</td>
<td>{(0,-1,2)} \Rightarrow {(0,0,2)}</td>
<td>0.438436</td>
<td>0.932292</td>
</tr>
<tr>
<td>(5)</td>
<td>{(1,-1,2)} \Rightarrow {(1,0,2)}</td>
<td>0.280903</td>
<td>0.811448</td>
</tr>
<tr>
<td>(6)</td>
<td>{(-1,-1,2)} \Rightarrow {(-1,0,2)}</td>
<td>0.300826</td>
<td>0.927711</td>
</tr>
<tr>
<td>(7)</td>
<td>{(1,-1,0),(1,1,0)} \Rightarrow {(1,0,0)}</td>
<td>0.259364</td>
<td>0.938650</td>
</tr>
<tr>
<td>(8)</td>
<td>{(0,-1,0),(0,1,0)} \Rightarrow {(0,0,0)}</td>
<td>0.370809</td>
<td>0.855491</td>
</tr>
<tr>
<td>(9)</td>
<td>{(-1,1,0),(1,0,0)} \Rightarrow {(-1,0,0)}</td>
<td>0.237458</td>
<td>0.812925</td>
</tr>
<tr>
<td>(10)</td>
<td>{(1,-1,2),(1,1,2)} \Rightarrow {(1,0,2)}</td>
<td>0.384655</td>
<td>0.932961</td>
</tr>
<tr>
<td>(11)</td>
<td>{(1,-1,2),(-1,0,2)} \Rightarrow {(0,0,2)}</td>
<td>0.367746</td>
<td>0.803859</td>
</tr>
</tbody>
</table>

![Fig. 2: association rules](image)

- **Explaining of texture association rules:**
  After mining these texture association rules, we need to explain the specific meaning of these rules according to the knowledge in some domain. For example, the explain of rule 12 is: in a 3*3 neighborhood image, after normalized, if the value of lower-left corner pixel is 2, and the value of upper-left corner pixel is 2, then we can reason that the value of its very left pixel is 2, and the support for this rule is 0.348655, the confidence is 0.932961.

![Fig. 3: explanation of rules](image)

5. **Accuracy analysis**

In this paper, the uncertainty problem is mainly arrived from the following reasons:

- uncertainty of image data

  In the stage of remote sensing imaging, there exists such phenomenon as same objects with different spectrum and different objects with same spectrum. It results in uncertainty of spectrum and features of image data. In the stage of data acquisition, it is difficult to control all error. Take uniformity of atmospheric conditions as an example. Even if it is processed spectral and geometric correction, residual error still exist;
In the stage of image matching through geometric correction, it is unable to eliminate those error caused by ground control point or image control point, which will affect the next stage of texture characteristics extraction; With regard to the research work of this paper, uncertainty mainly arrived from the stage of he generation of classified maps due to the factor of technology level and human mistake since classified maps are used. Concrete uncertainty problems include uncleanness, disagreement of class definition, non-unified quantitative standards, or classification error. Precision controlling makes it can satisfy some basic needs, is carried through, and they can basically meet the use of sectors, the gap between them and the actual types is obvious, but the distance between classified result and actual object can not be neglected.

- uncertainty of extraction rules

In this paper, association rules data mining in the formula concept analysis is used to extract texture nature features of objects. There are five indicators used to control the uncertainty in extracting:

1. Confidence

If in the transactions D which support the item set A, there are c% of D simultaneously also support the item set B, c% is called the confidence of the association rule \( A \Rightarrow B \). Briefly speaking, confidence means that in the transaction T which contains item set A, how much probability item set B appears in T at the same time. It can be expressed by probability as \( P(B/A) \), that is, confidence(\( A \Rightarrow B \)) = P(B/A).

2. Support

If there are s% of transactions support item set A and B simultaneously, s% is called the support of association rule \( A \Rightarrow B \). The support describes how much probability the C, which is intersection of item sets A and B, appears in all transactions. It can be expressed by probability \( P(A \cup B) \), namely: support(\( A \Rightarrow B \)) = \( P(A \cup B) \).

3. Expected confidence

If there is e% of transactions support item set B, e% is called expected confidence of association rule \( A \Rightarrow B \). It shows the probability that item B contained in all transactions without any other conditions, and can be expressed by probability: expected confidence(\( B \)) = P(B).

4. Lift

It is the ratio of confidence and confidence expectations. It describes how much influence the emergence of item set A to B. As the probability B appears in all transaction is expected confidence, and the probability B appears in transactions in which A appears is confidence, the ratio of confidence to expected confidence reflects that how much changes take place in the probability of B’s appearance, after adding the condition “item set A appears”. It can be expressed as: Lift(\( B/A \)) = confidence(\( A \Rightarrow B \)) / Expected Confidence(\( B \)).

5. Interesting rules:

The purpose of introduction of interesting degree is to prune some uninteresting rules and avoid generating false association rules. Normally, the interesting degree of a rule is the ratio of real density to expected density based on statistical independence assumptions, and it is between -1 to 1. For the interest degree of a rule, the bigger it is than 0, the more we are interesting in it; Otherwise, The less it is than 0, the more we may be interesting in its opposite rule, that means, the actual value of its opposite rule is greater. It is the association rule whose interesting degree is around 0 that have no significance.

6. Conclusion

In this paper, formal concept analysis theory is introduces into remote sensing images mining. Concept-driven image data mining combined with the GIS data and remote sensing images is studied in detail. Furthermore mined texture association rules are adopted to build sharing ontology database subsequently. At the same time the uncertainty problem in the whole image data mining process is analyzed. Experimental results show that what we mined have the capability of describing the objects of the images in nature not only from the view of theory but also from that of practice. Similar research can be extended to other fields besides texture association rules. In the near future, sharing ontology databases adaptive for all industries based on these association rules for geographic entities will be built combined with the theory of ontology and an acceptable solution to a more theoretical and practical way of spatial resources sharing and
interoperation will be explored.

Main contribution of our research work is that we propose a concept-driven strategy to mine rules from images. Future work includes mining association rules to more geographic entities so as to describe the characteristics of geographic entities better and completely.

7. Acknowledgements

This paper is supported by the National Key Basic Research and Development Program of China (No.2004CB318206) and the Basic Research of Survey and mapping bureau Project (No.1469990711111).

8. References