RESEARCH ON SPATIAL DATA MINING BASED ON UNCERTAINTY IN GOVERNMENT GIS

Bin Li[1,2], Jiping Liu[1], Lihong Shi[1]

[1]Chinese Academy of Surveying and Mapping
16 beitaiping Road, Beijing 10039, China
libin@casm.ac.cn

[2]School of Resource and Environment
Wuhan Univ
119 Luoyu Road,Wuhan 430079, China

Abstract

Uncertainty is the intrinsic property of spatial data and one of important factors affecting the course of spatial data mining. There are diversiform forms for the essentiality and aspect of uncertainty in the spatial objects of geographic information system. Essentiality of uncertainty may consist of the components of randomicity, fuzzy, chaos, etc. And the latter, i.e. aspect of uncertainty, may include error uncertainty, location uncertainty, attribute uncertainty, topology uncertainty, inconsonance uncertainty, immaturity uncertainty and so on.

Spatial data mining, which is based on uncertainty, is the course of discovering knowledge in the spatial data including the attribute of uncertainty. Spatial data is the necessary object operated by spatial data mining and its uncertainty can exist in the whole process of data input, data processing, data output, etc. Uncertainty can affect directly or indirectly the quality of spatial data mining. Specially, uncertainty of spatial data can affect directly or indirectly the veracity and reliability of ultimate decision-makings and may lead to produce false results and even reverse conclusions. Spatial data mining based on uncertainty has taken synchronously into account the two associated factors of uncertainty of data and spatial data mining. Therefore, due to considering the uncertainty, it can enhance the reliability of discovering knowledge and improve the veracity and reliability of spatial as well as raise the efficiency and practicability for the decision-makings made by operation departments.

In this paper, aiming at massive spatial data belonging to the Government Geographic Information System, some work has been done under the conditions of summarizing existing relevant theory and technology of spatial data mining and uncertainty and carry

out necessary practice, which can be described as follows.

i) Considering the uncertainty of spatial data, peculiarity and quality of spatial data to be mined have been investigated with emphasis to analyze their corresponding the contents and characters. There are diversiform modes for the exterior representation of uncertainty on spatial data mining, such as geometry uncertainty, topology uncertainty, theme uncertainty, etc. Imprecise spatial data often concerns with relevant data types and is affected by the following factors of data processing method, classified arithmetic, position precision, and so on. Two methods, i.e. disposal of data collection and data cognizance, are used to deal with the above-mentioned problems.

ii) Affection of uncertainty to the whole process of spatial data mining and correlative technology on how to deal with uncertainty, such as spatial data cleaning, generalization, integration, etc, have also been studied. Owing to the existence of different kinds of errors, uncertainty of spatial data often spreads every phase of data acquirement, storage, update, transmission, query and analysis, etc. In fact, uncertainty cannot be eliminated radically and only lessens by all means to debase the relevant affection to the best of our abilities.

iii) Some technology and methods concerned with spatial data mining, such as classified analysis, clustering analysis, etc, have also been investigated to summarize and analyze existing deficiency. On the above-mentioned basis, an improved and synthetic arithmetic will be presented in the end. And an experimental technical and application platform will also be constructed to provide the laboratorial environment. Technical improvement may be hoped to achieve under the existing conditions to meet the needs of spatial aided decision-making in the Government Geographic Information System.

1、 Introduction

China has made and implemented the policy by realizing the information technology to stimulate industrialization, and put forward higher requirements to information technology, especially in the government sectors. It is greatly important to enrich and improve the content and quality of existing spatial data, and ensure the authority and reliability of source information in order to effectively improve the informationization construction, make the realization of information sharing and information-intensive construction.

"Mass Spatial data and poor knowledge" has become a gap for the development of geo-spatial information science including Government Geographic Information Systems. Large numbers data is stored in Government GIS. 80 percent of the data is concerned with spatial location. In fact, there are little applications of these data. A large number of data is idle, causing a huge waste of data, which requires necessary data mining. Uncertainty of data and SDM are used to consider at the same time. Therefore, taking into account the uncertainty of data and spatial data mining, it has become an efficient method to enhance the credibility of knowledge to enhance the accuracy and reliability and improve the aided decision-making efficiency for government sectors as well as provide referenced information with authority and science in the fields of economy, society, culture and education, etc.

2、 Uncertainty of spatial data and its dissemination

With the continuous improvement of the degree of automation of spatial data access, spatial data increase exponentially. However, there are keen-edged contradiction between mass spatial data and knowledge acquisition due to the shortcomings of remote sensing and geographic information system software functions in dealing with spatial data, which finally results in "massive spatial data and poor knowledge" [1].

As far as the observation is concerned, although with the improvement of observation techniques and means, the accuracy of observation has made continuous improvement, there are still differences between the observation value and real value, which is a kind of expression form. It can only be weakened, but not completely eliminated. Besides, the expression and description of objective reality is not accurate and not enough can also lead to a series of uncertainty [2]

2.1 Characters and analysis of spatial data uncertainty

Uncertainty is an inherent property of spatial data, and is one of the important factors of affecting spatial data mining(SDM). Spatial data is the operation object for SDM, the uncertainty of which can be introduced in the process of spatial data input. It can affect the quality of SDM either directly or indirectly. If uncertainty couldn’t get enough attention and reasonable disposal, which could lead to false final results, or even opposite conclusions.

Geographic information systems among spatial objects, there are various forms of nature and appearance of uncertainty. Internal uncertainties may include randomness, ambiguity, chaos, and so on. There are diversiform modes of the external manifestations

of uncertainty nature in SDM, such as geometric uncertainty, topology uncertainty, and subject uncertainty, and so on. Among them, geometric uncertainty deals with the uncertainty when making geometric description to spatial entities. Attribute uncertainty represents the qualitative and quantitative uncertainty of spatial entities. Topology uncertainty reflects the uncertainty of spatial relationships among a spatial entity and circumjacent entities. And subject uncertainty describes the uncertainty of entity types. In this paper, attribute uncertainty is the focus in studying.

Inaccurate spatial data is often concerned with corresponding data type and is affected the factors of data processing methods, classification algorithms, location accuracy, time-varying of attributes, and so on. Therefore, two types of processing methods of data acquisition and data cognition are used to deal with the problems. The contents of attribute uncertainty of spatial objects in Government GIS can be described in figure 1.

2.2 The dissemination of uncertainty

Due to the existence of a variety of errors, uncertainty of Government GIS spatial data not only stems from the instability existence of natural phenomenon of its own and the incompleteness of human understanding, but also spreads in the stages of spatial data acquisition, storage, update, transmission, inquiry, analysis, expression, etc. Moreover, through the entire process of data mining, uncertainty of a step can be passed to the next step, leading to continuous accumulation and spread of uncertainty and making a serious impact on the effects of mining. Both spatial data and non-spatial data themselves have uncertainty, and can continue to spread and accumulate, which may lead to information and knowledge extracted with a certain degree of error, bias or being even meaningless, etc. All the knowledge mined cannot be believed to be useful and certain.

3、 SDM based on uncertainty

Spatial data mining (SDM), or "find knowledge from spatial database ", means to extract the contents, which users are interested in, such as spatial patterns and features, general relations between space and non-spatial data, and general characters of data implicated in the database from spatial database[1]. It is the extension of knowledge discovery technology in the application of spatial database. SDM usually finishes its successive steps as follows. Firstly, measure the physical entity and then make the conversion, input, processing of data. Finally, express the above-mentioned contents with knowledge model. At present, research on SDM mainly is concentrated in the field of the principles and methods. While, there is a few reports in another important aspect, that is, uncertainty of SDM. SDM based on the uncertainty means to discovery

knowledge from the spatial data with uncertainty under the given requirements and reference. Compared with traditional data analysis, SDM is more emphasis on the rules of spatial data analysis under the implied or unknown conditions, which can gain the information with more summary and proficiency[2]. SDM technologies have the characters of complexity, scale, integration, space, visualization and diversity, etc.

3.1 Characteristics of SDM

i)Source data used in SDM often has different degrees of uncertainty. However, the data is regarded to be certain in the traditional method, which lacks of reasonable considerations of authenticity to the original spatial data.

ii)A lot of uncertainty will be brought about in the process of SDM, especially in the course of data discretization of being continuous, which may lead to the knowledge exhumed to exist errors or be even meaningless. However, effective means don't be used to deal with the uncertainty in the traditional methods.

iii)Spatial data often tends to spatial autocorrelation with high degree. While, some assumptions of sampling independence is often given in the traditional methods, which lacks of the measurement and consideration to spatial autocorrelation.

3.2 Steps of SDM

The process of SDM can be divided into four phases: data selection, data pre-processing, data conversion, data mining, knowledge representation and evaluation. In figure 1, the flow of SDM can be described.

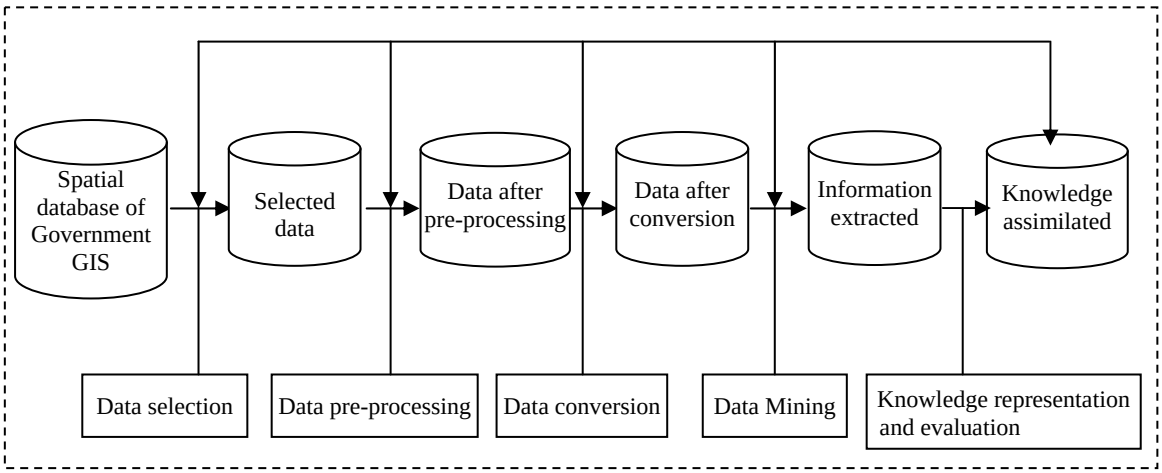

Figure 2.Flow of SDM

3.3 Uncertainty of SDM

Similarly, there is also considerable uncertainty accumulation and dissemination in the process of SDM, which is even more complex. Uncertainty of spatial data select mainly reflects the uncertainty in the process of subjective selection to target data in accordance with the requirements of the task of SDM, which includes what data should be selected and how many data should be sufficient[3]. In this phase, the uncertainty is mainly affected by the application of data mining technology to solve the problem needed to participate in the definition and the knowledge structure. Spatial data pre-processing mainly includes data cleaning, data transformation and data reduction. Data cleaning attempts to fill in vacant value, identify isolated points, and eliminate noises and correct the uncertainty of data. Data transform is to switch data into the form suitable for mining, which includes data smoothness, data assembly, and data aggregation. In this phase, on the one hand, uncertainty can be dealt with gradually. On the other hand, the process may bring about new uncertainty. Uncertainty of data mining itself mainly represents the limitations of data mining arithmetic, which may bring about the results caused by mining and the real situation not to be exactly the same. This is one of the important reasons of uncertainty caused by data mining. Each type of data mining algorithms has its advantages and disadvantages and the uncertainty of use scope and data. Uncertainty of knowledge denotation mainly reflects the uncertainty of implied knowledge itself including random, fuzzy, and so on. The same knowledge can be expressed in a number of ways. Some knowledge is better to be expressed in c certain method. And some knowledge may be appropriate to be denoted in another method. The knowledge gained by SDM is mainly the contents of being induced and abstracted, or the combination of qualitative and quantitative knowledge[3]. The best method of expressing the knowledge is to use natural language. At least, language values should be contained in the knowledge representation, that is, to use language values to express the qualitative concepts.

4、 Application of SDM based on uncertainty in Government GIS

This section takes the data stored in Government GIS as the experiment object, by reducing uncertainty and making use of spatial data mining technology, can mine corresponding potential information and provide the aided decision-making services for the department sectors of different levels. Here, this section takes the attribute data as an example.

4.1 Control of attribute uncertainty of data in Government GIS

Attribute uncertainty of data in Government GIS is the uncertainty change of the value of being measured or analyzed around the true value randomly in time or space during the course of collection, description and analysis of objective entities. It is the space extension of attribute error[4]. Attribute uncertainty can be affected by the factors of scale, resolution, sampling, etc. It has a variety of sources, and mainly roots in the introduction of uncertainty of attribute definition, data sources, data modeling and analysis of the. The main methods of dealing with attribute uncertainties can be described as follows:

i) Make the attribute definition as much as possible

Through the participation of experts, relevant theories and methods on Geo-ontology can be used to make reference, to maintain a comprehensive definition of attribute and reduce the emergence of ambiguity as far as possible.

ii) Select the accurate data source as far as possible

Data source is the cornerstone of constructing application applications, and its accuracy can directly affect the system operating efficiency, and thereby affect the quality of decision support system. The precise data sources should be selected to the full extent**.**

iii) Improve the accuracy of data modeling

The model of Government GIS is the similar expression of the objects of uncertainty. Take raster data as an example here, corresponding to domain model, which is used to describe the uncertainty of continuous region. Domain model can be used to describe the spatial entity by enduing with the properties of each unit, rather than realize by making corresponding abstract or description of the objects of the topological relations among them[5]. Domain model can represent the single-valued function defined in the continuous space by making use of spatial data. It is more suitable for the description of heterogeneous data as well as the uncertainty of regions with gradual change. Such as land classification, population distribution and other geographic phenomena.

iv) Uncertainty control of data analysis

Merge or change the classification.
Change the quantitative indexes.

Transform from one index to another or from one numerical value unit to another.

4.2 Application example on SDM based on uncertainty in Government GIS

In this section, a factual example on land utilization and land cover, which happened in a certain region of Guizhou Province, Southwest China, can be introduced to describe how to create a mining model of SDM. In fact, it is a kind of clustering algorithms. Clustering algorithm, which is also called aggregation algorithm, is an indirect data mining algorithms and does not use independent variables to get designated output. Different from classification model, clustering algorithm does not know beforehand that there are several categories to be divided and what these categories are. And it does not know how to define these categories according to some data items. Although it cannot predict unknown data value, while classification algorithm can, it provides a way to find similar records. These records can be considered the components of given clusters according to self-determined algorithm itself[7].In the processing course of constructing clustering algorithm, diverse data can be divided into different categories in order to make the difference between categories as big as possible and inner differences in the category as small as possible. Here, data of land utilization and land cover is mainly raster data, which is one of spatial data. Take raster data as the example. Details of SDM can be described in figure 2.

Figure 2. Spatial data mining based on clustering algorithm

As shown in figure 2, the original data set is divided into three categories. Thereinto, *C6, C8, C12* represent respectively land cover type of lush shrubbery, flourish grassland and cropland. The three land cover types hold the vast majority of the total number of all types. The amount is about 96.15%. Simultaneously, other land cover types has little proportion and even can be almost negligible. In-depth analysis to *Cluster 1* node can be

made. In this category, most of land cover types, of which proportion is about 95.51%, are *C6* and *C8*. That is, land cover type of this region is basically shrubbery and grassland. It is more suitable for the development of animal husbandry in a view of economic factor.

5、 Conclusions

In this paper, the spatial data in Government GIS is taken as the experimental objects. Both uncertainty of spatial data of its own and data mining process are combined to make some analysis to investigate the characters and realized process of SDM based on uncertainty. And this paper takes attribute data as an example with emphasis, and has described the spread process of uncertainty, control methods of uncertainty as well as the spatial data mining process under the conditions. At present, SDM based on uncertainty is a worldwide difficult problem with many challenges and involves lots of knowledge of multi-disciplinary and multi-field. More and more researchers have become an increasing emphasis on its application in the government sectors. With the development of further technology, there will be more in-depth applications in many fields.

6、 Acknowledgements

The work presented in this paper has been funded by China Fundamental Scientific Research Foundation (grant no.77738). And all the spatial data used in this paper comes from the database in Government GIS.

Reference

[1] Shuoben Bi, Huantong Geng. Research on the development and technology of civil spatial data mining. Geomatics World, 2008(1): 21-27

[2] Klir, G. J. Diversity and utility of uncertainty theories. Intelligent System, 2004. Proceeding of 2nd International IEEE Conference. 2004

[3] Xiaorong Fu.Research on uncertainty of spatial data mining. Master paper of Wuhan Technology Univ. 2007

[4] Binbin He, Tao Fang. Uncertain spatial data mining algorithms. Journal of China University of Mining & Technolog, Vol.36 No, 2007: 121-125

[5] Wenzhong Shi. Principle of modeling uncertainties in spatial data and analysis, Science Press, 2005

[6] Goodchild M F.1989.Modeling error in objects and fields[A].In: Goodchild M

F,Gopal S. Accuracy of spatial database.London:Taylor&Francis: 107-113
[7] Bin Li, etc. Study of construction and application of data warehouse for Government GIS", Bulletin of Surveying and Mapping, 2002(2): 4-6

Biography

Bin Li is an associate professor in Chinese Academy of Surveying and Mapping, Beijing, P.R. China. He was born in Sep. 1973 and graduated from Dept. of Map Cartography, Wuhan University in 1996. He received the master degree in GIS in 2002. His research interest is in the fields of digital cartography, geo-spatial database, geo-ontology, etc.