A STUDY ON GEO-CBR AND ITS APPLICATION IN SPATIAL DATA MINING FRAMEWORK

Yun yan DU[1] Fenzhen SU[1] Ce LI[2] Wei WEN[1]
1.The State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Science and Natural Resources Research, CAS, Beijing 100101
2.ESRI China(Beijing) Limited

Abstract: Currently, Geo-data-mining and knowledge discovering, new kernel of GIS spatial analysis study, which help breaking theoretic limitation of Geo-expert system and revealing an innovative research roadmap for new era Geo-information sciences, represent latest researching trend of GIS. Various research communities have tried to applying or revising mathematic tools as probability theory, spatial statistic, fuzzy set and rule based induction method to studies concerning specific geo-scientific problems. According to latest decade development in this study area, data mining method has absorbed, borrowed and revised latest mathematic tools and theories rising in AI study area; and focused both on theoretic research and its application in mining rules lying in spatial dataset. Development of Geo-data-mining couples tightly with AI and application mathematics with widely crossing and deeply fusion.

CBR (Case Based Reasoning), a traditional AI theory focusing on similarity reasoning, has been gaining growing concerns from study communities during 1990s. CBR, a new AI method that expands knowledge capturing channels, encapsulating problems by case, solving new problem by referencing historical similar ones, storing and re-using successful cases, has advantages as simplicity, flexibility, scalability, high efficiency, knowledge learning and accumulation, which enable CBR to analyst and reason complex geo-problems.

This paper mainly discusses Geo-CBR from a spatial data mining view and deems it as a kind of problem oriented spatial data mining method. Firstly, a detailed Geo-CBR definition and its encapsulating method are given as well as discrimination between spatial data mining and problem oriented Geo-CBR. Then, considering physical geography zonal and regional variation effect, inter-dependent and mutually condition relationships between geo-cases are examined in depth. And a quantitative data-mining method to explore intrinsic spatial relationships from geo-cases is presented based on tough set theory. In addition, due to variation of spatial feature types and their spatial relationships in geo-case representative model, 3 categories of spatial similarity calculating model are derived. Finally, a pilot study for LU is provided with purposes of land use problems quantitative analysis and deduction and demonstration of Geo-CBR's characteristics and advantages in solving and analysis spatial related problems.

In this paper, Geo-CASE, a geographical event occurring at a specific geographic location in a specific time duration (or at a specific time), is defined as a group of case's attributes containing spatial information and a "problem, geographic environment and result" description pair composed by a group of raster/vector data describing spatial distribution of environmental variables and a spatial/non-spatial case solving solution. The concept model using "problem, geographic environment

and result" description pair to depict Geo-CASE not only introduces "geographic environment" to represent intrinsic spatial rule/relation in Geo-CASE, but also expand "result" into spatial domain with a consequent ability of spatially reasoning.

Geographic zonality and regional variation effect in geographic space determine inter-dependent and mutually condition relationships between Geo-CASE. Rough set theory is introduced in this paper to filter out determined spatial relation rules from various kinds of spatial relationships described discretely and qualitatively using knowledge reduction algorithm and then draw out dominate determining spatial relation rules implied in Geo-CASE so as to participate in spatial reasoning for Geo-CASE "result".

Determining factor for its result varies according to different geosciences problems, which suggests different spatial similarity models for Geo-CASE. This paper proposes 3 types of similarity models: 1) coded spatial relation enabled Geo-CASE similarity model; 2) case spatial geometry information enabled similarity model; and 3) hybrid similarity model combining previous 2 models.

In section 3 this paper, a pilot study on land use issues in Pearl River Estuary to verify theories and models proposed previously is given, which aims 2 types of reasoning problems: quantitatively reasoning for land use type in 2003 and quantitatively reasoning for land use change from 1995 to 2000.

For the first problem, we build a history case database storing information of 4996 typical land use patches, using perimeter, area, shape index, fractal dimension, topological neighborhood relation, distance relation and etc to depict spatial characteristics and spatial relations of Geo-CASE. According to the experiment of predicting 72 cases, the prediction accuracy is 77%. As for the later problem, we build a history case database for 397 typical land use change patches, using area, perimeter of those LU patches, land use types of neighbor patches in 1995 and 2000, nearest distances from each patch to resident area, river and high way respectively to depict spatial characteristics of Geo-CASE. According to the experiment of quantitatively predicting 30 LU change cases, the prediction accuracy is 80%.

Our experiments results show simplicity and flexibility of Geo-CBR with a substantial advantage in quantitative analysis and simulation for complex geosciences problem when data resource is abundant. In addition, updating and growing of case database enables Geo-CBR with adaptive capability for solving rapidly changing natural resource and environment problems.