Rules Extraction and Representation for Geographic Information Systems (Short Abstract)

Lemonia Ragia[1], Vladimir Berenzon[2]
[1]Information System Department, University of Geneva, 24 rue General-Dufour,
1211 Geneva 4, Switzerland
lemonia.ragia@cui.unige.ch
[2]Informatik V, RWTH Aachen, Ahornstr. 55,
52056 Aachen, Germany
berenzon@i5.informatik.rwth-aachen.de

With the rapid use of the World Wide Web a huge amount of spatial data are available on line obtained from different sources and in different time frames. It is, therefore, important for the users to understand the nature of available data in order to use them appropriately for a specific task. In this paper we investigate knowledge representation and sharing in Geographic Information Systems (GIS).

Knowledge extraction and representation is a known problem in Artificial Intelligence (AI). Machine learning techniques have been used in different kinds of data but there is very little work regarding spatial data. There is some work proposed in which tools from AI are integrated in a GIS (Albert, 1988) and other approach represents machine learning algorithm for spatial data classification from a GIS (Stearns, 1995). Most data, however, are not integrated in one GIS but are present in several.

In this paper we propose a method based on machine learning to extract rules from spatial databases and to import them in a readable form in a GIS. We use the Web Ontology Language (OWL) as a special language for semantic information and the Resource Description Framework Language (RDF) for rules representation and for making them available via the World Wide Web. As an example, we applied our prototype in an application on groundwater vulnerability. The data were obtained in three different scales and there were both in vector and raster format. The data were distributive and managed by different systems.

The prerequisites for our approach are: the rules should be amenable for representation in a GIS system, they must be imported in every system, they have to be independent of a specific system and they have to be easily understandable and integrated.

The system involves the following components:
a) The machine learning component where the rules are extracted
b) The description logical component where the rules are defined at the logical level
c) The ontology web language component
d) The resource description framework component

Machine learning includes four main topics: classification, association rules, clustering sequence and similarity. Given a data set with spatial objects the classification provides a partitioning of the data set into classes according to the attributes of the objects. In this work we apply a classification method using the Interval Classifier (IC) algorithm of Agrawal (Agrawal, 1992). The spatial data are stored in databases using the Relational model, there are tables with attributes and their values. The algorithm is tree-based and can handle numerical and categorical attributes. To deal with categorical attributes a branch is introduced for every value of the categorical attribute. For the numerical attributes the range of values is divided to k disjoint intervals. The definition of the numerical intervals is based on a histogram. The representation of the result is a decision tree in which the

internal nodes contain a criterion for an attribute, the branches depict the attribute values and the leafs designate the classification.

We use for rules representation one of the most common languages of knowledge representation, the description logics which represent the knowledge of an application domain. It can be determined by a set of concepts and roles. A concept indicates a class or a set of objects and the roles are the relationships between these objects. There are semantics to characterize them.

The rules are represented in a deductive form $A \rightarrow B$ using a logic language. If the rules assumption are $A=\{A_1, A_2, \ldots A_m\}$ and the rules conclusion are $B=\{B_1, B_2, ..B_n\}$ then the logical operators "and" "or" connect them. The characterization of an attribute is based on the form $A_i = (a_i$ op v) where a_i is an attribute of spatial data and v belongs to the domain a_i. For the numerical values interval the op can be one of $\{<, >, =, \leq, \geq\}$ and for categorical attributes the op can only be $\{=\}$. The terms for the conclusion of a rule differ form the terms of the rule premise.

The Web Ontology Language describes the semantics of classes and properties and uses the ontology concept. This concept is especially used by professional people to share information about a domain. In geographical terms a domain could be, e.g. archeology, geology, urban planning, etc. In the OWL modeling we have the classes with the properties, the property restriction, value restrictions, cardinality restrictions and the relationships between the properties.

We also use a framework for representing information in the Internet the Resource Description Framework (RDF). This language is a collection of triples and includes the following:

- Resources which are the objects. Every object has a Uniform Resource Identifier.
- Properties which represent a binary relationship between two resources or between resource and resource value.
- Literals which are the values of the characteristics of the resources.

A statement is a resource, a property and a literal together which depict facts for the resources. The RDF model is a set of those statements.

We have outlined an approach for the extraction and representation of rules inherent in a spatial data set. We then indicate how these rules can be imported in a GIS and finally made available through the Web. We claim that with this approach we can incorporate more knowledge in a GIS. In this way a user can understand better the nature of the spatial data and is in a better position to use them appropriately in a wide variety of applications.

BIBLIOGRAPHY

Albert T. M. 1988. Knowledge-based geographic information systems (KBGIS): New analytic and data management tools. Mathematical Geology, Vol. 20, No. 8, pp. 1021-1035.

Agrawal R., Ghosh S., Imielinski T., Iyer B. R., and Swami A., 1992. An Interval Classifier for Database Mining Applications. In Proccedings of the 18th International Conference on Very Large Data Bases. L. Yuan (ed.) Very Large Data Bases. Morgan Kaufmann Publishers, San Francisco, CA, pp. 560-573.

Stearns, S. and St. Clair, D. C. 1995. Rule-based machine learning of spatial data concepts. In Proceedings of the 1995 ACM Symposium on Applied Computing. K. M. George, J. Carroll, and D. Oppenheim, (eds.) ACM Press, New York, pp. 242-247.