

SEMANTIC SIMILARITY EVALUATION MODEL IN CATEGORICAL DATABASE GENERALIZATION

Liu Yaolin¹, Martin Molenaar², Menno-Jan Kraak²

¹School of Resource and Environmental Science
Wuhan University, Wuhan, 430070, P.R. China

²Department of Geo-Informatics
International Institute for Aerospace Survey and Earth Science
7500 AA, Enschede, The Netherlands
Email {yaolin, Molenaar, [Kraak](mailto:Kraak@itc.nl)}@itc.nl}

Commission IV, WG IV/3

KEY WORDS: Hierarchical structure, Similarity Model, Categorical Database, Classification System,

ABSTRACT

Database generalization process will be used to derive a new database with less detail for some application purposes from a single detailed database. In a database generalization process, semantic similarity measures among objects and among object types play a key role in object aggregation process. In this paper a generalization procedure will be developed based on object aggregation. The decision to aggregation objects will be based on the geometric properties of the objects, on their spatial relationships and on their thematic similarity. Normally, this similarity among objects acts as decisive rule that controls the generalization operations. This paper focuses on semantic similarity evaluation model for categorical database generalization. It presents a semantic evaluation model after reviewing current similarity models. The models are based on classification hierarchies, set theory, and attribute structure of classes. It is a computation model and can not only be used to compute the similarity between objects at same level, or at different levels, but also the similarity between object types at the same level and at different level.

1. INTRODUCTION

The similarity measures among objects and among object types play a key role in object aggregation in database generalization. The aggregating two objects not only depends on the geometric properties of the two objects, but also the thematic properties of the two objects. Using Set theory, Tversky (1977) defines a similarity measure in terms of matching process. This measure produces a similarity value that is not only the result of the common, but also the result of the different characteristics between objects, which is in agreement to an information theory definition of similarity (Lin, D 1998, Rodriguez and Egenhofer 1999, Bishr, 1997, Chakroun et al, 2000, Rodríguez and M. Egenhofer, 1999, Rodríguez, M. Egenhofer and Rugg, 1999). Although many models of similarity are defined in the literature, the similarity measure method based on set theory, hierarchy structure and attribute structure of class is still lack. This paper mainly discusses the similarity measure method in the database generalization used

set theory, hierarchy structure and attribute structure of class. The paper is organized as following. First, the brief review is given; secondly, the basic concepts for similarity measure are given and followed by the computing similarity model is proposed; finally an example is used to test the model.

2. BASIC CONCEPTS

Some basic concepts must be clear before discussing similarity computing model.

2.1 Class, Object Type and Object

In this paper, the class and object type have the same meaning. A class or object type determines a set of attributes to form its attribute structure. Each class c_j or object type c_j has its own attribute structure $List(c_j)$ as following:

$$List(c_j) = \{A_1, \dots, A_i, \dots, A_n\}$$

A_i denotes one of the attributes of class c_j . Each attribute will have a name, a domain that will be specified by defining

the range of the attribute values and scale type of the domain which indicates whether these values are from a nominal, an ordinal, an interval or ratio scale. An attribute A_i can be specified by a three tuple (Molenaar 1998):

$$A_i = \{\text{NAME}(A_i), \text{SCALETYPE}(A_i), \text{DOMAIN}(A_i)\}$$

For each object which belong to class c_j , the attribute structure defined by a class specifies the description structure of the object. A value is assigned to every attribute in the attribute structure of object. For any object, its direct class is unique and lowest in a classification hierarchy since different classes have different attribute structures. These values must fall within the range of the attribute domain, which must be defined prior to the actual assignment of attribute values.

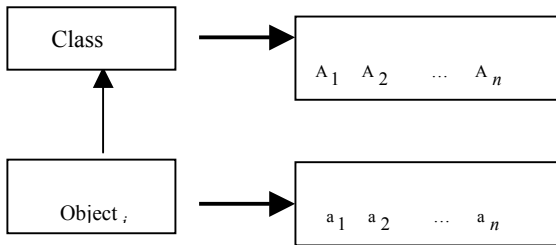


Figure 1 Diagram representing the relations between objects, classes and attributes (after Molenaar 1998)

Each object has an attribute structure list containing one value a_j for every attribute of its class. The thematic description of an object can now be specified by its class (which specifies the attributes of the object) together with the list of attribute values.

2.2 Intension and Extension of A Class

The intension of a class c_j can be expressed as a condition for the value of a combination of attributes. The condition specifies a subset of the set containing all the possible combinations of the values of the attributes, denoted as $\text{Int}(c_j)$, while the set of all the objects that belong to c_j with the same attribute structure is commonly identified as the extension of the class c_j , denoted as $\text{Ext}(c_j)$.

2.3 Formalizing Classification Hierarchy and Aggregation Hierarchy

The class hierarchy has been studied for many years in databases and knowledge bases, especially in relation to data abstraction and generalization (Smith and Smith 1977, Yee, Leung, Kwong, S.L. and He J.Z 1999, Molenaar 1996).

However there are still lack of formalization of classification hierarchy and aggregation hierarchy based on Set theory and properties of class. In the following part, the formalizations of classification hierarchy and aggregation hierarchy are discussed.

The object types in a geo-spatial model are normally determined by classification hierarchy and aggregation hierarchy. Changing classification hierarchy and aggregation hierarchy will results in changing geo-spatial model associated with database, and in turn changing the contents of the database. Changing the attribute structure and extension of classes at different level in classification hierarchy and aggregation hierarchy associated with a database will induce a new classification hierarchy and aggregation hierarchy and define a new data model of database and rebuild the corresponding contents of database.

Let S be the set of objects $\{o_1, o_2, o_3, \dots, o_i\}$ in space, U be the set of classes $\{c_1, c_2, c_3, \dots, c_j\}$ for the database space. Before formalizing classification and aggregation hierarchy, the relations between classes must be identified.

Let $c_i, c_j \in U$ be two arbitrary spatial classes, there should be no objects that belong to the extensions of the two different classes of U . Based on the definition of classification hierarchy last chapter and relations among classes, we can formalize classification hierarchy with intension and extension of the classes.

- $c_i \Psi c_j$, only if $\text{Ext}(c_i) \cap \text{Ext}(c_j) \neq \emptyset$; (Relations of consistent among classes symbolized by Ψ), (c_i is consistent with c_j);
- $c_i \equiv c_j$, only if $\text{Ext}(c_i) = \text{Ext}(c_j)$, denoted as $c_i = c_j$; (Relations of equivalence among classes symbolized by \equiv), ($\text{Ext}(c_i)$ of c_i is equal to $\text{Ext}(c_j)$ of c_j and c_i is identical to c_j);
- $c_i \subseteq c_j$, only if $\text{Ext}(c_i) \subseteq \text{Ext}(c_j)$, denoted as $c_i \leq c_j$; (Relations of inclusion among classes symbolized by \subseteq), ($\text{Ext}(c_i)$ of c_i include $\text{Ext}(c_j)$ of c_j and c_i belongs to c_j) or

- $c_i \subset c_j$, if $\text{Ext}(c_i) \subset \text{Ext}(c_j)$, denoted as $c_i < c_j$. We also call c_i a sub-class of c_j and c_j a super-class of c_i . (Relations of complete inclusion among classes symbolized by \subset), ($\text{Ext}(c_i)$ of c_i completely include $\text{Ext}(c_j)$ of c_j and c_i completely belongs to c_j).
- $c_i \neq c_j$, only if $\text{Ext}(c_i) \cap \text{Ext}(c_j) = \emptyset$. (Relations of disjunction among classes symbolized by \neq), (there is no common object between c_i and c_j , and c_i is different from c_j).

A class c_i is included by a class c_j if and only if the extension of c_i is subsumed by the extension of c_j . We call c_i 'IS-A' c_j .

Obviously, \leq_c is a partial binary relation on U , called the 'Belong to' relation, and (U, \leq_c) is a partially ordered set that we call a classification hierarchy. The process of formalization of classification hierarchy depicts how the object types (classes) and super object types can be formed into a hierarchical structure. For creation of a new super object type in the classification hierarchy there will be:

- If any $A, B \in H$ and no $D \in H$ satisfying $\text{Ext}(D) = \text{Ext}(A) \cup \text{Ext}(B)$, then generate such a class D , and let $\text{Ext}(D) = \text{Ext}(A) \cup \text{Ext}(B)$, $\text{LIST}(D) = \text{LIST}(A) \cap \text{LIST}(B)$ and $D \in H$, noted as IS-A links;

The upward connections from objects to classes and classes to super classes are is-a links, which express that an object is an instantiation of a class and that a class is a special case of more general super-class. At each level, the classes inherit the attribute structure of their super-classes at the next higher level and propagate it normally with an extension to the next lower level. At lowest level in the hierarchy are elementary objects (Molenaar 1998).

2.4 Hierarchic Semantic Similarity Matrix

For a hierarchical structure as shown in Figure 2. a semantic similarity matrix as shown in Table 1 can be defined based on

the properties of the hierarchical structure. A, B and C in Figure 2 represent the different branches in the hierarchical structure. For later use, they are called sub tree. T in the same Figure is called the top of the structure and c_i are object or object type.

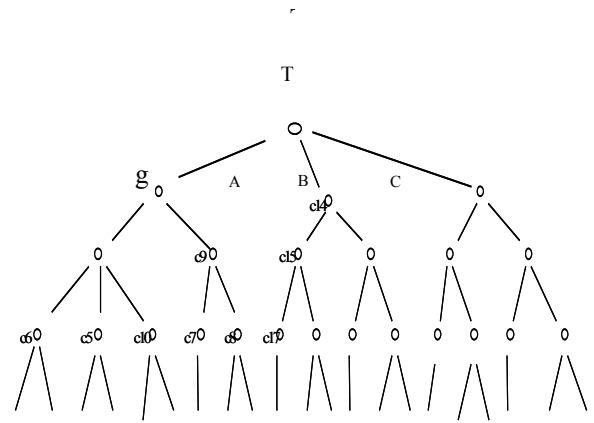


Figure 2 Example of a hierarchical structure

The semantic similarity matrix represents the similarity between object types.

Table 1 Example of semantic similarity matrix

SIMILARIT Y	Sub-type1	...	type1	type2	...	Sup-type1	..
Sub-type1	s_{11}	...	s_{14}	s_{15}	...	s_{17}	..
Sub-type2		...	s_{24}	s_{25}	...	s_{27}	..
...	
type1			s_{44}	s_{45}	...	s_{47}	..
type2				s_{55}	...	s_{57}	..
...				
Sup-type1						s_{77}	..
...							..

Where:

sub-type1 etc denote different elementary object types;

type1,type2 denote different object types;

sup-type1 etc denote (super) composite object type;

s_{ij} denotes similarity value among object types.

The larger the value of an element in the matrix is, the more the similarity between two object types that the element links is. The matrix is symmetric and reflexive one, and has the property that s_{ij} is equal to s_{ji} ($s_{ij} = s_{ji}$) and s_{ii} is equal to

s_{jj} ($s_{ii} = s_{jj} = 1$) in the matrix. s_{ji} is a value between 0 and 1.

This matrix shows the similarity among different levels of object types. It will provide potential possibility to choose objects of different types to be merged or aggregated. The similarity matrix will be used as a look-up table for guiding or governing the aggregation process of spatial objects in semantics to a certain application.

The value of element s_{ij} in the matrix can be given by expert knowledge or by calculation (to be discussed in Section 6.3.2) based on aggregation hierarchy and classification hierarchy.

3. COMPUTING MODEL OF SIMILARITY

A computational model that assesses similarity among objects and object types based on set theory, classification hierarchy and attribute structure of class is proposed. There are three distances. One is the distance (number of the link edges) from immediate super object type that subsumes c_i and c_j to the top of a hierarchy such as object type g to top of the tree T in Figure 2 which represents common part of attribute structure between two object types c_i and c_j . Another distance (number of the link edges) from immediate super object type that subsumes c_i and c_j to c_i such as object type g to c_1 in Figure 2 which represents the different part of attribute structure between object type c_i and c_j ($|c_i - c_j|$). And the third is the distance (number of the link edges) from immediate super object type that subsumes c_i and c_j to c_j such as object type g to c_5 in Figure 2 which represents the different part of attribute structures between object type c_i and c_j ($|c_j - c_i|$). The proposed model is shown in Equation 2. It suits for two cases. One is for two given objects or object types belonging to the same sub tree such as sub tree A in Figure 2 and the other is for two given object types belonging to two different sub tree such as A and B as shown in Figure 2. For the first case, the model uses two types of distances to define the common and difference properties between the given object types. One is the distance between given objects or object types

and immediate super object types that subsumes them which reflects difference properties between two given object types and the other is the distance between immediate super object types that subsumes two given object types and the top of hierarchical structure which reflects the common properties of two given object types. For the second case, the distance between immediate super object types that subsumes two given object types and the top of hierarchical structure will be zero since the two given object types belong to different sub tree such as c_1 and c_{15} in Figure 2. So this distance will be replaced by the correlation value between two sub trees in the Equation 2.

$$s_{ij}(c_i, c_j) = \begin{cases} \frac{l}{l + \alpha(c_i, c_j) * d_{ci} + (1 - \alpha(c_i, c_j)) * d_{cj}} \\ \frac{\beta}{\beta + \alpha(c_i, c_j) * d_{ci} + (1 - \alpha(c_i, c_j)) * d_{cj}} \end{cases} \quad (1)$$

Where:

- l : the shortest distance (number of the link edge) from immediate super object type that subsumes c_i and c_j to the top of a hierarchy;
- d_{ci} : the shortest distance (number of the link edges) from immediate super object type that subsumes c_i and c_j to c_i ;
- d_{cj} : the shortest distance (number of the link edges) from immediate super object type that subsumes c_i and c_j to c_j ;
- α : a function of the distance (number of the link edge) between immediate super object type that subsumes c_i and c_j to the class c_i and c_j .

β : correlation degree among different sub-trees, such as similarity among agriculture land use, forest land use and building up land use, and its value can be given by experts based on application requirement.

A natural approach to comparing the degree of generalization between object types is to determine the distances from these object types to the immediate super object types that subsumes

them in a classification hierarchy as shown in Figure 1, that is, their least upper bound in a partially ordered set. In a sense, the difference in the distances from these object types to the immediate super object types that subsumes them in a classification hierarchy reflects the difference in attribute structure between two object types. The $\alpha(c_i, c_j)$ can be expressed as a function of the distance d_{ci} and d_{cj} . In order to get final values of α , the function (Equation (2)) is defined as following:

$$\alpha(c_i, c_j) = \begin{cases} \frac{d_{c_i}}{d_{c_i} + d_{c_j}} \\ 1 - \frac{d_{c_i}}{d_{c_i} + d_{c_j}} \end{cases}$$

where:

d_{ci} : the shortest distance (number of the link edges) from immediate super object type that subsumes c_i and c_j to c_i ;

d_{cj} : the shortest distance (number of the link edges) from immediate super object type that subsumes c_i and c_j to c_j .

This similarity function yields values between 0 and 1. The

5. CONCLUSIONS

The computing model that has been proposed has taken set theory, classification hierarchy and attribute structure into account. It can be used to measure the semantic similarity among object types in classification hierarchy. The example shows that the computing result of the model are reasonable and efficient. How to decide the correlation parameter β will still need to research in the future work.

The authors would thanks the financial support from ITC, Ministry of Education, P.R. China and State Bureau of Surveying and Mapping.

6. REFERENCES

Bishr, Y., 1997, Semantic aspects of interoperable GIS. ITC, Publication No. 56, Enschede: International Institute for Aerospace Survey and Earth Sciences (ITC).

extreme value 1 represents the case that the two entity classes are completely the same, whereas the value 0 occurs when the two entity classes are completely different.

4. EXAMPLE

An example for computing element of similarity matrix from classification hierarchy is as following: Taking Figure 3 as example to calculate the similarity among object types. The class code can be seen in appendix. The correlation value among construction land, agriculture land and unused land that is given by the experts is 0.5.

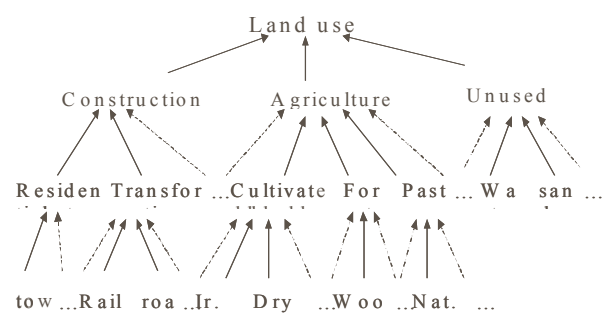


Figure 3 Land use classification hierarchy before generalization

The computing result of similarity among object types based on the model (see in Table 2.)

Chakroun, H., Benie, G.B., Oneill, N.T., and Desilets, J. 2000, Spatial analysis weighting algorithm using Voronoi diagrams. International Journal of Information Science, Vol.14, No.4 pp.319-336.

Lin, D. 1998, An Information theoretic definition of similarity (eds.) International conference on machine learning, ICML'98.

Molenaar, M., 1998, An Introduction to the Theory of Spatial Object Modelling for GIS, Taylor & Francis Ltd. London.

Molenaar, M., 1996, The Role of Topologic and Hierarchical Spatial Object Models in Database Generalization, In Molenaar, M. (Eds.) Methods for the Generalization of Geo-databases, Delft: Netherlands Geodetic Commission, New Series, Nr 43, pp.13-36.

Rodríguez and M. Egenhofer, 1999, Putting similarity assessment into context: matching functions with the user's intended operations. Modeling and Using Context, CONTEXT'99, Trento, Italy. In: P. Bouquet, L. Serafini, Patrick Brézillon, Massimo Benerecetti, and Francesca Castellani (eds.), Lecture Notes in Computer Science, Vol. 1688, Springer-Verlag, pp. 310-323.

Rodríguez, M. Egenhofer, and R. Rugg , 1999, Assessing semantic similarity among geospatial feature class definition Interoperating Geographic Information Systems, Second International Conference, INTEROP'99, Zurich, Switzerland. In: . Vckovski, K. Brassel, and H.-J. Schek (eds.), *Lecture Notes in Computer Science*, Vol. 1580, Springer-Verlag, pp. 189-202,.

Smith, J.M. and Smith, D.C.P, 1977, Database abstractions: aggregation and generalization, *ACM Transactions on Database System*, 2, pp.105-133

Tversky,A. 1977, Features of similarity. *Psychological review*, 84(40:PP.327-352).

Appendix

Land use classification (code)

1. Agriculture Land (100)

11□ Cultivated land

111	Irrigated paddy fields
112	Rain fed paddy fields
113	Irrigated land
114	Dry land
115	Vegetable plots

12□ Garden Land

121	Orchards
122	Mulberry fields
123	Tea fields

	124	Rubber plantation
	125	Other
13□ Forest		
	131	Wood land
	132	Shrubbery land
	134	Young forestation land
	135	Slashes
15□ Water area		
	151	Rivers
	153	Reservoir
	154	Pond
	156	Beaches and flats
	158	Hydraulic building

2. Construction Land (200)

21 Residential quarters and industrial and mining land

211	Area of cities and towns
212	Residential quarters in rural areas
213	Isolated industrial and mining land

3. Unused lands (300)

310	Waste lands
380	Others

