Semantic Heterogeneity in Distributed Geographic Databases

M.F.Worboys
Department of Computer Science
Keele University
Keele, Staffs ST5 5BG UK

S.M.Deen
Department of Computer Science
Keele University
Keele, Staffs ST5 5BG

Abstract

This paper considers the special problems of semantic heterogeneity in a distributed system of databases containing spatially referenced information. Two forms of semantic heterogeneity are identified. *Generic* semantic heterogeneity arises when nodes are using different generic conceptual models of the spatial information. *Contextual* semantic heterogeneity is caused by the particular local environmental conditions at nodes. It is contextual heterogeneity which is especially the consideration with geographic databases and to which the paper devotes most attention. Two possible solutions are proposed, one founded on transforming processors between models and a second using a canonical model which is a generalization of existing generic spatial models.

1 Introduction

Geographic databases are characterized by very large quantities of often complexly structured data about a locality. This local information space might contain multi-layered information about a restricted region of the earth's surface (for example, environmental data on the Highlands of Scotland) or a stratum of information for a larger area (for example, land use data for the UK or census information for the US). In both cases it would be greatly advantageous for local information contained in a set of nodes to be linked into a distributed spatial information system. However, when this integration is attempted, problems are encountered regarding the physical and semantic heterogeneity of the nodal systems.

Within distributed systems, semantic heterogeneity exists when 'there is a disagreement about the meaning, interpretation, or intended use of the same or related data' (see [10]). This paper will discuss specific problems of semantic heterogeneity in a distributed spatial system. Two types of semantic heterogeneity are identified:

- generic semantic heterogeneity;
- contextual semantic heterogeneity.

Unlike databases built on the relational model, spatial databases have more complex and varied logical models of spatial data. Generic semantic heterogeneity occurs when different nodes are using different generic models of the spatial information. For example one may use a layer-based approach and a second may use an object-based approach. An example is given in [5] where the city of San Jose has a collection of data sets (e.g. road network, locations of city features, hospitals, fire stations, police stations, etc.) stored in vector format (i.e. using the object-based approach). The local university has a database of seismological information used for the prediction of earthquakes. This information is stored using a quadtree data structure and is based upon the layer model. The problem arises when the city wishes to integrate the two systems in order to answer such questions as, 'Are any hospitals near fault lines?', or 'How many fire stations are within 10 mile by major road of this fault line?'.

Contextual semantic integrity occurs when the semantics of schemas depend upon the local conditions at particular nodes. An example will illustrate some of the considerations here. Imagine two databases holding information on the road networks for the regions in which they are sited. Each database will contain relatively static information about the topology of the networks, distances between centres, road types, tolls, etc. But each will also hold time-dependent information about local traffic conditions, road repairs, etc. A global query might be 'Plan a route between a centre in one region to a centre in the second region'. Semantic heterogeneity of the generic type would occur if the underlying models of the two databases are distinct. However, further semantic discrepancies will be present if, for example, road types do not correspond or units of measurement are different. It is also possible that there will be discrepancies between the two systems' views of the networks at the boundary (or region of intersection) of the regions. There are problems integrating geographic databases modelled in different contexts.

The remainder of this paper is structured to examine these two aspects of heterogeneity in turn. The next section considers generic heterogeneity. After a formal distinction is made between layer-based and object-based approaches, we consider problems of interchangeability between the two approaches, and provide the basis of possible frameworks in which such conversions may take place. It is generic semantic heterogeneity which is specifically the problem for ge-

ographic databases and to which emphasis is given in this paper. Section Three discusses contextual heterogeneity with reference to some examples. In the conclusion, the paper considers whether spatial data places any special demands on the semantic heterogeneity problem for distributed databases.

2 Generic semantic heterogeneity

2.1 Layer-based spatial models

Spatial data models fall into two main categories: layer-based and object-based. In the former case, a layer-based model consists of a finite collection of layers, $\{L_i \mid 1 \leq i \leq n\}$. Assume as underlying spatial framework a set F of spatial references. For example, F might be the points of a regular grid. For $1 \leq i \leq n$, each *layer* L_i is a function from set F to an attribute set A_i. Two examples of attribute sets are a set of topographical altitudes and a set of vegetation types. A layer would then associate with each point of a grid a topographical altitude or a type of vegetation.

Given a framework F, a *neighbourhood function* n : $F \rightarrow \mathcal{P}(F)$ may be defined, which associates with each location x a set of locations within a specified distance and/or bearing of x.

Given a layer L_i, a *zone* is a subset of F, the values of whose attributes satisfy a predefined condition. For example, with the altitude layer, the zone *below sea-level* is the subset of F consisting of those locations which have negative altitude. A *zoning* of F is a partition of F into disjoint zones whose union is F.

The algebra of a layer-based model is specified by giving the layers and the operations on the layers. Such operations take as arguments existing layers and produce a new layer. They can be of essentially three types (see [11]):

local operations An attribute is created at location x which depends on the attributes of x associated with the layers which are arguments of the operation. For example, given layers of populations and lung cancer mortalities, create a new layer of ratios of lung cancer mortalities per unit population.

focal operations Here, the attribute created at x depends not only on the appropriate attributes of x but also on the attributes in the neighbourhood of x. For example, from a topographical altitude layer create a layer of gradients.

zonal operations The resultant attribute created at location x is dependent upon the appropriate attributes within the zone containing x. For example, given a layer of temperatures and a zoning into regions, create a layer of average temperatures for each region.

2.2 Object-based spatial models

For object-based spatial models, we assume the usual object data model with the addition that the objects have attributes which inherit properties from generic spatial objects, such as point, line and polygon. Thus, a lake will have a polygonal spatial structure for its boundary, as well as having instance variables for its name, depth, capacity, etc. The generic spatial objects have been modelled as objects with their own complex structure and relationships. Let the most general spatial object type be denoted *spatial*. Then *spatial* may lead to specialized classes depending on dimension, closure, connectness, etc. For further details, see [13].

A widely used model of spatial objects embedded in a plane is the node-arc-polygon (or the 2-D winged-edge) model, where an arc has a start and end node and has polygons to its left and right (see, for example, [7]).

Much discussion has taken place in the literature (see, for example, [9, 12]) regarding a primitive set of spatial operations. Recent work [1] suggests that such operations fall into the following four classes. These classes are ordered in the sense that each Euclidean space is a metric space; each metric space is a topological space; and so on.

set-based operations Operations which treat the spatial objects purely as sets. The intersection of two regions would be an example of a set-based operation.

topological operations Operations which depend upon the topological structure of the space. For example, the boolean operation which determines whether two regions are adjacent, or the boolean operation which determines whether a region is connected.

metric operations Operations which involve measurement of length. For example, the distance between two points.

Euclidean operations Operations which involve measurements not only of length but also of angle. For example, the bearing of one point from another, or the area of a region with a polygonal boundary.

2.3 Integrating the two approaches

The general distinction between layer-based and object-based models is that a layer is a function from a set of spatial references to an attribute set, thus providing information on global variation of the attribute over the layer, while the object-based approach models the information structure as populated with constituent objects which have as attributes references to spatial objects. Thus the one model is, in a sense, the inverse of the other. The issue of generic semantic heterogeneity arises since spatial information systems are built using different underlying models. Some types of

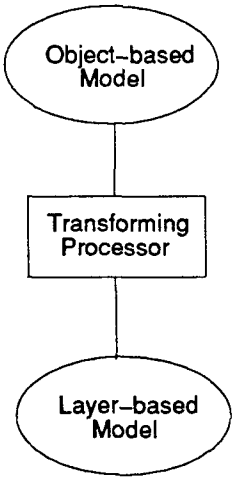

Figure 1: Transforming processor approach

spatial data are more appropriately handled in a layer-based system (for example, spot-heights throughout a region), while others (road networks, for example) are more appropriately handled in an object-based system. Some spatial operations are more efficiently performed in a layer-based system. For example, given information on variation of topographical altitude and rainfall for a region, it is possible to model this as two layers and perform an overlay operation to make comparisons. This is natural in a layer-based system but cumbersome in an object-based system. On the other hand, to compute an optimum route through a road network, it is clear that an object-based model is preferable. There are applications which may require a hybrid approach. Some classes of diffusion problems fall into this category, where there is both a discrete and continuous component (as in the spread of an epidemic). Thus, any optimization strategy would require interchangability between the two models.

2.3.1 Transforming processor approach

Above, we have described the structures of the layer-based and object-based approaches. It is possible to move between these two models by using a transforming processor (see [10] and Figure 1). In this case, there must be a mechanism for the transformation of the data representations between the two models. Use of zoning within layers and aggregation of attributes into objects provides a strategy for moving from layers to objects. A layer of variation of soil type may be transformed into a set of areal objects, one of whose attributes will be soil type. However, difficulties remain; for example, identification of lineal features and error considerations. Errors constitute a most important issue in geographic databases since errors proliferate much more quickly (and unpredictably) than in traditional numerical applications. With regard to lineal features, algorithms for layer to lineal object transformation exist (see, for example, [8]). Thus layer-based seismic information, which is functional from a set of locations to seismic levels, may be inverted into fault lines, as may topographical altitude variations into contour lines.

Movement from objects to layers is accomplished by forming single layers from common attributes of objects. Thus all objects of type 'road' would have their spatial references amalgamated into the general frame for the layer-based model. Again, there are difficulties, not the least of which are error considerations and the efficiency of such transformations.

There is also the problem of translating constraints from one model to the other. An example of a layer-based constraint is the imposition of an upper and lower bound upon the gradient of a layer function. Using the notation of Section 2.1,

$$\forall p. \quad k_1 \leq \nabla L(p) \leq k_2$$

where L is a layer (assumed to be a field for this example) and k_1, k_2 are scalar values and p is an element of the underlying spatial framework. An example of an object-based constraint is a pair of bounds on one or more dimensions of the spatial references of objects of a given type. Let $\mathcal{O}$ be an object class and s be the spatial reference attributed to $\mathcal{O}$. Then, the example constraint is given by,

$$\forall O \in \mathcal{O}. \quad k_1 \leq O.s \leq k_2.$$

In neither case is it natural to express the constraint using the other model.

The above examples lead one to the conclusion that both models have their advantages with geographic databases. To force all geographic information into one of the generic models is to lose the modelling and representational power of the other. Thus most single node systems at present, while concentrating on one or other of the models, provide some of the functionality of the other. An example of a prototype system where both models are given equal status is Relational Image-Based Geographical Information System (RIGIS) [14].

2.3.2 Canonical model approach

Another approach to the integration of the two models is to create a model which is a generalization of both layer-based and object-based models. Such a *cross-representational algebra* or *canonical model* (see [10] and Figure 2) has been proposed as a long-term goal in [4] where it is suggested that point-set topology might provide a sufficiently general model.

Such a structure would provide a general integration of heterogeneous components and is a current topic for research. This paper briefly considers the role that simplicial complexes (objects in the realm of algebraic topology) might have in a general model. Assume for simplicity that the underlying spatial framework is the Euclidean plane. Thus, the constructions performed in this subsection are all planar, but the ideas can easily be generalized to higher dimensional structures.

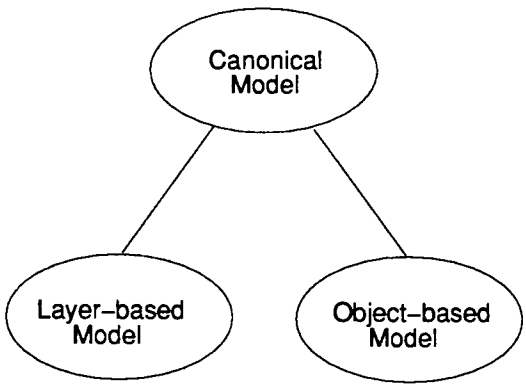

Figure 2: Canonical model approach

0-simplex A *0-simplex* is a set consisting of a single point in the Euclidean plane.

1-simplex A *1-simplex* is a set consisting of all the points on a straight line between two distinct points in the Euclidean plane, including the end points.

2-simplex A *2-simplex* is a set consisting of all the points on the boundary and in the interior of a triangle with vertices three non-collinear points.

Thus an n-simplex is the convex hull of $n+1$ linearly independent points. The number n is the *dimension* of the simplex. The *boundary* of a simplex S, ∂S, is the union of all the lower dimensional simplexes contained within it.

Simplexes are the building blocks of simplicial complexes. A *simplicial complex* C is a finite set of simplexes satisfying the properties:

- A face of a simplex in C is also in C.
- The intersection of two simplexes in C is either empty or is also in C.

Intuitively, a simplicial complex is a union of simplexes which do not impinge on each others' interiors.

Simplicial complexes define a set of connected regions, lines and points in the plane. They have the advantage of simplicity and yet contain a complete specification of the spatial object which they represent. They enable efficient computation of spatial operations upon them (for example, it is a straightforward matter to calculate the boundary of a complex). Non-spatial attributes can be associated with simplicial complexes or with their component simplexes.

Collections of simplicial complexes, together with their associated attributes (at both simplex and complex levels), may be used as the foundation for both layer-based and object-based models. In the layer-based approach, the spatial framework, F, is itself a simplicial complex. Each layer is an association of attributes with individual simplexes over the entire complex. In the object-based approach, objects have spatial references which are sub-complexes of F. A great advantage here is that lineal objects are themeselves complexes, and therefore sub-complexes of F. Therefore, there is no need to approximate lines by areas, as there was in the conversion processes in the preceding section. Simplicial complexes were first considered as a data model for geographic databases in [3].

A disadvantage of the canonical model approach is the potentially high data volumes ensuing. How serious this disadvantage is will be determined by ongoing research.

3 Contextual semantic heterogeneity

By contextual semantic heterogeneity is meant local differences in nodal conceptual schemas which are dependent upon the environment in which the node is placed and from which it takes its data. Distributed geographic databases inherit all the problems that any distributed database system might have in this respect. Thus, problems of different semantics for names, scale, units, etc. at different nodes must be resolved as for general distributed database systems. Such matters are considered by Deen *et al* in [2]).

To illustrate the special flavour of contextual semantic heterogeneity in geographic systems, we consider areal spatial units. Information is often referenced with respect to an areal unit (for example, population density and morbidity ratio). However, the areal unit depends upon the context. In the UK there are local government administrative units, postcode units, health authority units, and others [6]. The global conceptual schema should be a coherent picture in which all these units are integrated. Sometimes, the units nest into each other, in which case it is easy to aggregate data from smaller units to larger. For example, counties are partitioned into districts, which are themselves partitioned into wards. Clearly, there is no problem in calculating the populations of counties if we are given data on ward populations. However, in most cases this simple nesting does not occur and approximate conversions have to be found. To illustrate with a simple example, one of the authors' work on a medical application has involved the integration of a hospital database where spatial references are to postcode sectors with a database of census information where spatial references are to administrative areas. Thus to calculate ratios per unit population it is necessary to disaggregate data referenced to one set of units and aggregate to the second. The transformation module is only approximate.

Such examples point to a general problem within all databases, but particularly with geographic systems. The general consideration is that the information is within the context of the local space. With regard to spatial units, conversions based upon overlay, point-in-polygon, or aggregation/disaggregation are required, along with estimations of accuracy. It has not been found feasible to construct a canonical model in this case (although the UK government has at times encouraged data collectors to use postcode sectors as

a standard areal unit). Thus, we have built conversion factors and estimates of accuracy into an interface schema, which is associated with the local schema to give a *component schema* (see [10]).

4 Conclusions

The aim of this short paper has to been to examine the problems of semantic heterogeneity in the context of geographic databases. We have found that two major sources of such heterogeneity arise; one caused by the dichotomy of underlying model for geographic systems and the other arising from local contexts. Both sources of semantic heterogeneity cause real problems for the geographic database community. The solution to the former is still a major topic for research, and the paper has indicated a promising direction using a canonical model based upon simplicial complexes, while the latter may be solved in specific cases by adding interface units to local schemas, thus forming component schemas.

The paper has concentrated upon the problems of generic semantic heterogeneity because it is more suggestive of the types of issues that researchers in the area of geographic databases face. Future work will formally define the canonical model and its transformation to layer-based and object-based models.

References

[1] B. Dawson, K.T. Mason and M.F. Worboys. An object-based paradigm for an intelligent geographical database system. In *Proc. Workshop: Spatial Databases, User Needs and Solutions*, DAKE Centre, Keele University, 1991.

[2] S.M. Deen, R.R. Amin and M.C. Taylor Data integration in distributed databases. *IEEE Trans. Software Engineering*, 13(7):860–864, July 1987.

[3] M. Egenhofer, A. Frank and J. Jackson. A topological data model for spatial databases. In *Proc. Symp. Design and Implementation of Large Spatial Databases*, pages 271–286, Springer-Verlag, 1989.

[4] O. Guenther and A. Buchmann. Research issues in spatial databases. *SIGMOD Record*, 19(4):61–68, Dec 1990.

[5] L. Haas and W. Cody. Exploiting extensible dbms in integrated geographic information systems. Technical Report RJ 8173 (75132), IBM, June 1991.

[6] H.M. Hearnshaw, D.J. Maguire, and M.F. Worboys. An introduction to area-based spatial units. Technical Report 1, Midlands Regional Research Laboratory, Univ. Leicester, Leicester LE1 7RH UK, 1989.

[7] S.D. Morehouse. ARC/INFO: A geo-relational model for spatial information. In *Proc. Auto Carto 7*, 1985.

[8] T. Pavlidis. *Algorithms for graphics and image processing.* Springer-Verlag, Berlin, 1982.

[9] D.J. Peuquet. Towards the definition and use of complex spatial relationships. In *Proc. 3rd Int. Symp. on Spatial Data Handling*, 1988.

[10] A.P. Sheth and J.A. Larson. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, 22(3):183–236, Sept 1990.

[11] C. Dana Tomlin. *Geographic Information Systems and Cartographic Modeling.* Prentice Hall, 1990.

[12] M.F. Worboys, H.M. Hearnshaw, and D.J. Maguire. Object-oriented data and query modelling for geographical information systems. In *Proc. 4th Int. Symp. on Spatial Data Handling*, pages 679–689, Zurich, Switzerland, 1990.

[13] M.F. Worboys, H.M. Hearnshaw, and D.J. Maguire. Object-oriented data modelling for spatial databases. *Int. J. Geographical Information Systems*, 4(4):369–383, 1990.

[14] Q. Zhou and B.J. Garner. On the integration of GIS and remotely sensed data: Towards an integrated system to handle the large volumes of spatial data. In *Proc. 2nd Symp., SSD91*, pages 63–72, Zurich, Switzerland, 1991.