Finding Analogous Structures in Cartographic Data

Diarmuid O'Donoghue, Adam Winstanley

Department of Computer Science,
National University of Ireland, Maynooth
Co. Kildare, Ireland
{diarmuid.odonoghue; adam.winstanley}@may.ie

Abstract. In this paper we describe the application of analogical structure matching techniques to the domain of Geographic Information Systems (GIS). Automatic categorisation of map data into roads, buildings *etc.* is currently based on isolated objects. We describe how identifying analogous clusters of objects can categorise ambiguous polygons by introducing context into the categorisation process. We describe a number of GIS classification tasks that can be performed by analogical structure matching, resulting in a useful classification tool that operates in a cognitively plausible manner.

1. Introduction

Manually recorded topographic data consists primarily of boundary definitions, where lines combine to form polygons enclosing parcels of land. But these "line drawings" are of limited usefulness to both cartographers and the general public. Classifying individual polygons as buildings, roads, made-land, or unmade-land *etc.* vastly increases the usefulness to these data, but is expensive to perform manually on terra-bytes of cartographic data.

Automatic classification of geometric topographical data into object types (and/or feature codes) can be partially accomplished through isolated-shape recognition [Keyes & Winstanley, 2000], by focusing on parameters of individual object such as total area, boundary length, and shape. Performance can be improved by extending the classification mechanism with contextual information. This improves the accuracy of automatic classification, because we can frequently resolve ambiguous data by examining its context to provide evidence for category membership (thus informally, for example, we can say that a square on a map is more likely to depict a house if it is near a road).

We describe a method of matching clusters of topographical objects against a known prototype cluster by identifying *analogous structures*, and in this way we automatically infer the identify of unclassified polygons. An analogy is a comparison between a well known *source* and a problem *target.* The source acts as a predictor for the target, because the source supports inference about that target. The ability of analogical comparisons to support inference is the prime reason for our use of the analogy process to perform topographical classification. Our categorisation technique involves reasoning with collections objects, and thus the reliability of our category assignment process improves by incorporating neighbourhood data in the classification.

A variety of cognitive studies have been carried out to ascertain the exact nature of the analogy process. A widely studied example (Duncker, 1945) concerns the problem of treating a patient suffering from an inoperable tumour. One set of subjects are given the source domain of a country ruled by an evil

dictator and whose fortress can only be reached by sending troops down multiple roads to simultaneously converge on the fortress thereby overwhelming it. Subjects that do not receive the "fortress" information have a 10% solution rate, while 80% of subjects that receive the fortress domain give the required convergence solution (Gick and Holyoak, 1980). Thus, analogies have a profound effect on peoples ability to generate inferences and solve problems. Inferring the category of unclassified topographic objects is just one of many domains that may be solved using analogical comparison.

Classifying unknown topographical objects through structure matching requires two complimentary tasks. The central, or core, activity concerns generating the largest possible mapping between the problem data the some pre-stored prototype. Since Gentner [1983] identified that analogies are built on an implicit parallelism between two information structures, computational modelling of the analogy process has been the focus of much work. This work has largely focused on creating more efficient algorithms for identifying the largest structure mapping - including [Falkenhainer, Forbus and Gentner, 1989; Keane, and Brayshaw 1988; Veale, O'Donoghue and Keane, 1999; Salvucci and Anderson, 2001].

The second activity revolves around determining the boundaries between this problem data and "irrelevant" background information. This requires identifying clusters of information that can be (temporarily) isolated from the remainder of the map data, in order to allow the core mapping process to proceed. Category prototypes play a significant role in boundary identification, with the efficiency of the matching process being reliant on domain selection. This retrieval activity is a necessary precursor to the matching process and has received comparatively little attention - see [Forbus, Gentner and Law, 1994; Plate, 1998; Crean and O'Donoghue, 2001].

2. Geometric Analogies

In this paper we focus on comparisons between sets of geometric objects, broadly similar to those comparisons found in IQ tests (Evans, 1967; Bohan and O'Donoghue, 2000). These have the structure A is-to B as C is-to an unknown D. (If a square within a circle(A) changes to a striped square within a circle(B), what do we do with a triangle within a square(C)?). Solving geometric analogies is founded upon identifying a matching on the information structures between A and C. Its is only by comparing the relationships between objects (and *not* comparing similar objects themselves) that we see the square (from A) and the triangle (from C) play the same role in Figure 1. It is the information structure and the relationships between object dictates that the square and the triangle are matching objects in this problem. Thus, the solution will involve a small striped triangle. Were we to compare similar objects, we would align the two squares and we would end up with a large striped square - which is clearly incorrect! Algorithms to perform the structure matching task are more complex, and are based on analyses of and comparisons between information structures - rather than on the content of this information itself.

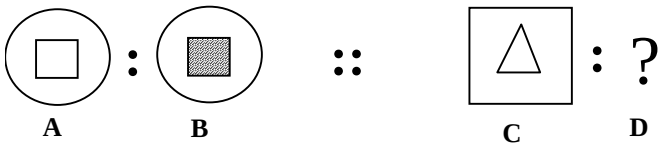

Figure 1 : *A Simple Geometric Analogy problem*

Computational modelling of the geometric analogy process can be achieved to two steps. Firstly, (find and) represent the known information about both problem domains using predicate calculus assertions. Secondly, find the largest isomorphism between the two sets of data (Gentner, 1983; Veale, O'Donoghue and Keane, 1999). For the problem in Figure 1, domain A might be represented as "`contains (circle, square)`" while C might be "`contains (diamond, triangle)`". Aligning these predicate structures will identify the *mapping* indicated by Figure 2. Solving the above geometric analogy is built upon identifying this inter-domain mapping (between A and C).

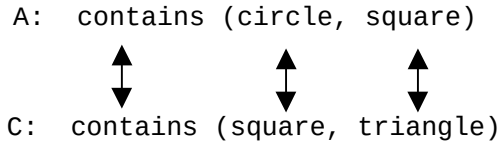

Figure 2 : *A predicate mapping identified on structural similarity*

Next we identify the attribute transformation (Bohan and O'Donoghue, 2000) that occurs in the source domain, changing between parts A and B above and denoted (A -> B). For this we need only note the attribute alterations that must be applied. In the previous example we note that the square changes from plain to striped. (Later, this shall correspond to identifying the required classification for an unclassified topographic object). Applying the attribute transformation (A->B; read, A changes-to B) to the equivalent objects in C. (So, the square maps onto the triangle, so if the square becomes striped then so must the triangle.) The predicate information does not change in this problem, as it doesn't for topographical classification problems. While the general process of analogy is far more complex, we shall not analyse it in further detail here.

2.1 Domain Representation

Before the analogy process can be applied to the problem described, we must first represent the problem information in a suitable manner. This involves quite an amount of analysis of the underlying topographic data, and shall be discussed later. We identify two relationships between topographic objects.

i) `adjacent (a, b)` indicates that two polygons share a common boundary.

ii) `point-adjacent (a, b)` indicates that the adjacent two polygons meet only at a single point.

Generally these two sets of information are mutually exclusive, although this may not be the case, for example, when two objects touch at multiple locations. It is these two simple relations that form the basis of out structure matching process, that drives the category assignment procedure. A typical topographic domain that consists of say 5 topographic objects may contain, approximately, five of the first predicate and two examples of the second predicate. The analogy process takes this problem information and applies the analogy process to this information.

2.2 Analogical Comparisons in Topographic data

Applying analogy to topographic data is even simpler than the geometric analogies previously described. First lets consider the problem of categorizing a single unclassified polygon contained within a cluster of otherwise categorized polygons. Let us also assume that the correct source domain has been selected for use with the given target problem. Finding the correct classification for the unclassified polygon requires two simple steps.

i) Identify the largest possible structure matching between the problem and the solution template.

ii) Find the object that maps with the unclassified polygon, and apply the class of that polygon to the unclassified polygon.

As with the analogy process itself, the new classification is derived by a process of pattern completion applied to the inter-domain mapping. So, if the unclassified object matches with an "unmade land" polygon, then the unclassified polygon also assumes the classification of "unmade land". Pattern completion itself is a relatively straight-forward algorithm, but its simplicity belies the fact that it must be only applied after the pre-requisite process have been carried out.

Of course the usefulness of this categorization technique is completely reliant on the applicability of the identified source domain. This has two implications, first great care must be taken in constructing the store of candidate source domains to ensure that the inferences mandated by each comparison are valid. Secondly, we need a reliable technique to retrieve the most appropriate domain from the stock of candidate source domains. Retrieval is currently initiated by identification of a target problem containing a single unclassified polygon. We use an attribute based retrieval scheme to select the most appropriate source domain. This retrieves the most similar source domain that contains not just polygons of the required types, but also in the required configuration. However, in this paper we focus on the use of polygon attributes to effect retrieval. This causes retrieval of a similar source domain with the same contents, which differing by the classification of a single polygon. For a description on structure based retrieval see [Crean and O'Donoghue, 2001].

3. System Architecture

Having described the central theory of category assignment by structure matching, we now describe a software realisation of this system. This technique requires that the underlying geographical information is already at least partly classified. This is because structure without some content information (*ie* classification) is too vague to identify categorisation for any but the simplest of cases. For that reason this structure matching approach is ideally after a partial classification has already been achieved. Figure 3 illustrates the basic system architecture.

The first process involves selecting suitable collections of polygon information to be passed on to the analogy process itself, and this involves much more than ensuring only one unclassified polygon is included in each of these clusters. The basic cluster of polygon information used in this project is referred to a *locality*, consisting of a root polygon, all adjacent polygons and all adjacency information between these polygons. Localities are the basic unit of process for all subsequent structure matching activities. A number of factors motivated this choice of structural primitive. Firstly, structure matching is an NP-complete problem, and thus using small information domains is considerably more efficient than large domains. Even though large domains may support more robust classification, the resultant computational expense would be too severe for practical application on a typical desktop computer. Secondly, localities include sufficient information for a variety of ambiguous classification tasks - including error detection as well as classification. Third, explicit storage of one locality can expedite computation for adjacent localities. Fourth, localities offer the possibility of easily including more detailed information on individual polygons with structure information at some later date.

Separating locality identification from the structure matching process itself has some addition benefits. Under the described architecture, we explicitly store each locality and additionally we index each locality according to the category of each polygon contained within that locality. Though there is a large degree of overlap between adjacent localities, explicit storage helps expedite the subsequent matching process and the locality identification process itself. Before proceeding to structure matching itself, we see how certain types of localities support categorisation without applying computationally expensive structure matching algorithms.

More importantly, a very significant classification advantage emanates from explicit storage and indexing of locality information. This advantage concerns classification polygons with *flexible structure matching*. This classification process extends the power of the matching process by allowing a number of problems with differing structure to all match the same source domain. For example, a number of structures can be identified as erroneous without recourse to detailed structure matching. Regardless of the structure of certain localities, we may identify the error based purely on the contents of the locality itself. Such errors can be identified by direct examination of the locality information, and this process is referred to Adjacency Matrix Classification - see Figure 3.

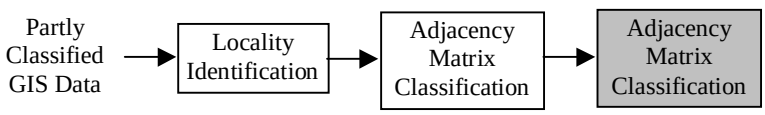

Figure 3 : *System Architecture*

Clearly, one cannot have a “waterway” polygon completely enclosed within a polygon classified as a building - regardless of the structure of these or any other polygons with the locality. A house located entirely within another house can be flagged as a probable error, and this surprisingly accounts for several errors in the supplied “Purbeck” data alone. This technique offers the unique capability of altering the original map data - this is vital for extending existing boundaries to ensure feature codes don’t *run* into adjacent objects, causing mis-classification.

We see two different uses of this structure matching approach to topographical classification, first for completing classification of partly classified polygons, and secondly for correcting certain categories of classification error. Examples of both uses are described later in this paper.

A major problem with topographic data is known as “category bleeding”, where the categorisation of one object bleeds into attached objects because not all objects perfectly enclose an area. This results in two polygons being connected by a narrow neck of land - often invisible to the cartographer. We see the identification of known localities as providing a potential solution. By recognising a known structure, particularly a problematic one, we may be able to detect the error and even offer the ability to extend one boundary to stop this category bleeding.

4. Detecting Adjacency Errors

In this section we examine in detail two applications of the flexible structure matching technique. These examples illustrate the technique use in identifying (pre-existing) classification errors - this activity being a necessary precursor to reliability estimation and to error correction.

Consider the illegal-prototype representing "no building may directly and completely enclose another building" - though intuitively obvious, this simple rule cannot be represented by isolated object identification. The flexible structure matching technique identifies (amongst others) the following misclassification. In figure 4 (below) the lighter colored objects identify building, while the darker objects represent any other categorization.

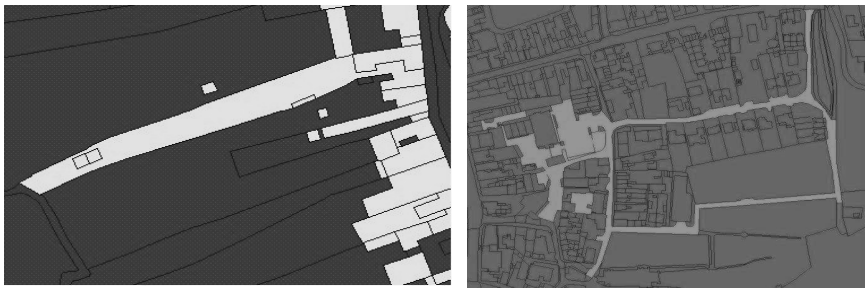

Figure 4: *Examples of the “Building within building” error*

Another illegal neighborhood that is identifiable from the adjacency matrix is the "no road my be completely detached from all other roads". This effectively enforces the rule that all roads must be attached to other roads - hence a roads’ usefulness. This template identifies the error, which can be easily corrected by isolated shape-recognition techniques.

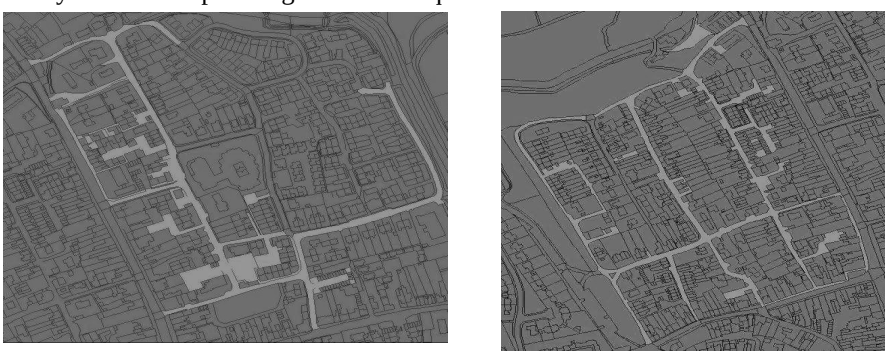

Figure 5 : *Examples of the "Isolated road" error*

5. Structure Matching

Detailed structure matching is more expensive to perform, but is more generally applicable to polygon classification. Structure matching allows us to infer and assign a classification to unclassified polygons. The matching process identifies a 1-to-1 correspondence between some problem structure (containing an unidentified polygon) and a similar template. Given the identical structures and sufficient similarity in each polygons categorisation, we may infer the identity of the unclassified polygon.

The following figure illustrates the template for a "semi-detached house". Again we define (a number of) template structures for each class of object we wish to classify. The dots in this diagram indicate that neighbouring polygons are adjoined at one point rather than being joined by a line - thus diagonally adjoining polygons are included in a locality.

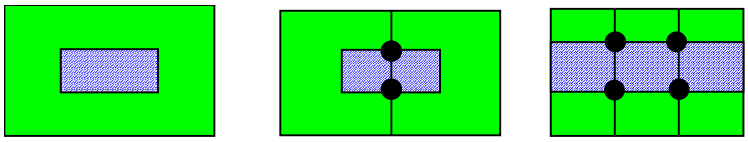

Figure 6 : *Simple Geographic objects*

We represent all the adjoins relationships in the chosen locality. The incremental matching algorithm [Keane, 1994] identifies the structural isomorphism between the problem and template. If an isomorphism exists, wherein every problem object is matched to one (and only one) template object - and if the same juxtapositions exists in each domain, then we have a structural match between domains. This is the first part of our requirement.

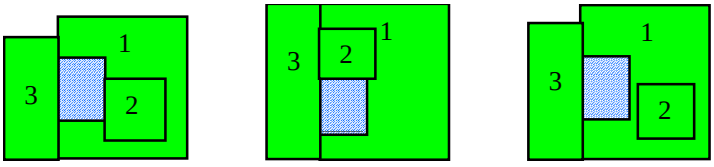

Figure 7 : *Labelled Geographic localities*

The second requirements is that matching objects are of the same categories; this requirement obviously does not apply to the problem polygon which must be unclassified. Given the structural match, and the subsequent category matching, we infer the identify of the unclassified polygon.

The Structure matching algorithm for matching GIS structures is based on the Incremental mapping model [Keane et al, 1994]. However, matching is this domain presents some unique problems. Firstly, each domain represents adjacency information between locality objects. Structure matching domains described entirely with commutative predicates introduces problems of structural ambiguity - because `adjacent(a,b)` may also be written `adjacent(b, a)` and the matching algorithm must be aware of this problem. Secondly, we need to integrate attribute information into the matching process - this requires a fundamental extension to the theory of Gentner [1983]. Finally, different source domains must be applied to solve different problem structures. Thus our mapping process must include a retrieval phase to select the most appropriate template. However, we shall not describe our algorithm further. The significant factor here is that it identifies the largest possible 1-to-1 correspondence between the problem and the template data.

Results are not yet available for detailed structure matching, but initial results are very promising. The extra precision provided by detailed structure matching brings extra refinement to the classification

process, and thus many ambiguous object are categorised by virtue of the classes of the surrounding objects.

6. Conclusion

Current techniques for automatically classifying topographical data are based on analysis of isolated objects - and thus cannot process ambiguous polygons. We described the technique of structure matching and how it may sustain topographical object classification. We also showed how contextual information can use used to identify typographical classification errors, even without detailed information on individual objects. This is achieved by matching object neighbourhoods against known templates - representing problematic clusters of topographic information.

Acknowledgements

We would like to thank the Ordnance Survey of Great Britain (Southampton) for access to their "Purbeck" topographic database and their for support for this project.

References

Bohan, A. O'Donoghue, D. 2000, "LUDI: A Model for Geometric Analogies using Attribute Matching", AICS-2000 11th Artificial Intelligence and Cognitive Science Conference, Aug. 23-25, NUI Galway, Ireland.

Crean B. P. O'Donoghue, D. "Features of Structure for Analogy Retrieval", Applied Informatics 2001, Innsbruck, Austria.

Duncker, K. "On problem solving", Psychological Monographs, 58 (whole, no. 270), 1945.

Evans, T. G. 1968, "A program for the solution of a class of geometric-analogy intelligence test questions", In "Semantic Information Processing", (Ed.) M. Minsky, MIT Press.

Falkenhainer, B. Forbus, Gentner, D. 1989 "The Structure Mapping Engine: Algorithm and Examples", Artificial Intelligence, 41, 1-63.

Gentner, D. 1983, "Structure-Mapping: A Theoretical Framework for Analogy", Cognitive Science, 7, 155-170.

Gick, M. Holyoak, K. 1980, "Analogical Problem Solving", Cognitive Psychology, 12, 306-355.

Keane, M. T. Brayshaw, M. 1988, "Indirect analogical Mapping: A Computational Model of Analogy", in Third European Working Session on Machine Learning. Ed. D. Sleeman, London Pitman,.

Forbus, K. Gentner, D. Law, K. "Simulating Similarity-Based Retrieval: A Comparison of ARCS and MAC/FAC", Proc. 14th Cognitive Science Society, 543-548, 1994.

Plate T. “Structured operations with Distributed Vector Representations”, in “Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational and Neural Sciences”, New Bulgarian University, Sofia, Bulgaria, July, 1998.

Salvucci D. D. Anderson J. R. 2001, “Integrating Analogical Mapping and general problem solving: the path-mapping approach”, Cognitive Science, 25, 67-110.

Veale, T. O'Donoghue, D. and Keane, M. T. 1999, “Computability as a limiting cognitive constraint : Complexity concerns in metaphor comprehension about which cognitive linguists should be aware”, in Cultural, Psychological and Typological Issues in Cognitive Linguistics, Ed. M. Hiraga, C. Sinha and Wilcox., S. John Benjamins Publ. Amsterdam/Philadelphia. pp 129-155 - ISBN: 90 272 36569.

Winstanley, A.C. and Keyes L. 2000, “Applying Computer Vision Techniques to Topographic Objects”, XIXth International Archives of Photogrammetry and Remote Sensing, 33 (B3), 480-487.