This paper was selected by a process of
anonymous peer reviewing for presentation at

COMMONSENSE 2007

8th International Symposium on Logical Formalizations of Commonsense Reasoning

Part of the AAAI Spring Symposium Series, March 26-28 2007,
Stanford University, California

Further information, including follow-up notes for some of the
selected papers, can be found at:

www.ucl.ac.uk/commonsense07

Grounding a Geographic Ontology on Geographic Data

David Mallenby

School of Computing
University of Leeds
davidm@comp.leeds.ac.uk

Abstract

Vagueness is prevalent within the geographical domain, yet it is handled poorly in existing ontology approaches. A proposed way to rectify this is to ground the ontology upon the data. By grounding the ontology, we make an explicit link between the ontology and the data, and thus allow reasoning to be made within the context of the particular data. In order to ground the ontology upon the data, we must first decide how to represent the data and how to handle the vagueness with reasoning. This paper illustrates the stages required to prepare geographical data for an ontology to be grounded upon, including considering how to reason about the vagueness, how to represent the data in a more efficient manner and how to reason about relations within the data to extract attributes that would be used within an ontology.

Introduction

There is a huge amount of geographical data available today, in a variety of formats from classical cartographic maps to satellite imagery. This data can be analysed, combined and reasoned with in Geographic Information Systems (GIS). In order to reason about geographical features we need a method of representing the data and the meanings attached in a logical manner. The use of ontologies has become a popular method of representing such data [9, 33, 11].

The use of ontologies in GIS has been proposed in [9, 27] amongst others. Existing methodologies do not adequately handle vagueness, which is inherent to the geographical domain. Features are often dependant on the context in which they are made, with local knowledge affecting definitions. Geographical objects are often not a clearly demarcated entity but part of another object [9, 27]. The individuation of entities is therefore more important to geographical domains than to others.

One approach proposed to improve the handling of vagueness is to ground the ontology upon the data [17]. By grounding the ontology, we make an explicit link between the ontology and the data, thus allowing reasoning to be made within the context of the particular data. Grounding the ontology upon the data requires the data to be represented in a manner that will allow the link between data and ontology. We require an approach that allows the ontology to segment the data accordingly, based on user specifications.

In this paper we will examine the stages that are required in order to convert geographical data into a suitable form upon which terms in the ontology can be grounded. The data to be looked at is of The Hull Estuary, with the aim being to obtain a method of reasoning about the hydrological features which are implicit in the data. It is important to note that the particular formats and segmentation processes applied here may not necessarily apply to other features within the geographical domain. Rather, the aim is to show the process of preparing such data for an ontology.

Motivation

One of the key considerations for geographical ontologies is the handling of vagueness [31]. Vagueness is inherent to the geographical domain, with many features defined without precise definitions and boundaries. Such definitions are dependant on the context in which they are made.

Vagueness is handled inadequately in present GIS; some approaches such as [9, 27] choose to ignore the size quantifier and categorise a river simply as a waterbody, whilst others have sets of quantifiers [31]. Both approaches base the size quantifier on a predefined perspective that may not be agreed upon or may be based on a particular context that isn't applicable in all situations.

Vagueness is not a defect of our language but rather a useful and integral part. Rather than attempting to remove vagueness, it is better to develop an approach that allows the user to decide what makes up a vague feature. By improving the handling of vagueness, we improve the functionality of GIS, allowing vague features to be reasoned about in an effective manner.

Vagueness in Geography

As discussed by Bennett [2], vagueness is ubiquitous in geographical concepts. Both the boundaries and definitions of geographical concepts are usually vague, as well as resistant to attempts to give more precise definitions. For example, the definition of a river as given by the Oxford English Dictionary [1] is:

> A large natural flow of water travelling along a channel to the sea, a lake, or another river.

This is clearly vague, with the most obvious example being the use of 'large', although there are other parts of the definition that are vague also.

The sorites paradox can be easily adapted to illustrate vagueness in geography, as shown in [32, 33]. So, whilst there are some things that are definitely rivers and some that are definitely not, there does not exist an explicit boundary between the two sets, thus classical reasoning can not state if something is or isn't a river.

Geographical definitions are dependant on the context in which they are made. For example, in the UK rivers are defined usually as permanent flows, but in Australia they may not contain water all year round, thus there is a temporal requirement to the definition [29].

The principal approaches for handling vagueness at present are fuzzy logic and supervaluation theory. Both approaches offer a method of reasoning over vague features. It is usually the case that the two are presented as opposing theories. However, this in part assumes that vagueness can only take one form, which as discussed in Dubois [7] is not true. Rather, there are instances where it is more appropriate to use fuzzy logic and instances where supervaluation theory is better.

Fuzzy logic is the popular approach to handling vagueness, and has been used in a variety of applications since its conception by Lotfi Zadeh [37, 35, 36]. The underlying concept is to allow a method of processing data by allowing partial set membership rather than strict set membership or non-membership. Fuzzy logic is especially adept at handling situations where we do not want to generate an explicit boundary between two sets, but rather represent a gradual transition between the two.

Initially proposed by Fine [8], supervaluation theory proposes that there exist many interpretations of the language. Statements could therefore be true in some interpretations and false in others. In supervaluation semantics, 'precisifications' are used to determine the boundary points at which statements are considered true or false in a given interpretation. Supervaluation theory is suited to situations where we wish to generate a boundary between sets that we know exists but are not able to permanently mark as such.

In our proposed system, we wish to segment, individuate and label hydrological features. We therefore require a method of reasoning that marks explicit boundaries depending on user preferences.

If we were wishing to mark features with transitional boundaries, then fuzzy logic would be suitable, as we would have fuzzy boundaries between features. However, an attempt to return crisp boundaries would not be suited to fuzzy logic due to logical rules used in reasoning.

Supervaluation theory on the other hand, is suited to return a crisp boundary for given preferences. With fuzzy logic we take the stance that there is not a boundary between features so we show a gradual range, whereas with supervaluation theory we assume that there is a boundary, we just don't know for certain (or agree upon) where it is. The user preferences therefore become the precisifications. Supervaluation theory is therefore preferable for this problem.

Ontology Grounding

The ontology level is usually seen as separate to the data level; we reason within the ontology, and return the data that matches our queries. Thus the ontology is devoid of the data context, despite any impact this may have. This has a clear impact upon handling vagueness, where attributes are based heavily upon the context in which they are made.

A proposed improvement to this is to ground the ontology upon the data [17]. By grounding the ontology, we make an explicit link between the ontology and the data, thus allowing reasoning to be made within the context of the particular data.

The symbol grounding problem as proposed by Harnad [13] suggests that computers do not actually understand knowledge they are provided, as meanings are merely symbols we attach to objects. There have been no adequate solutions to this problem as yet and it remains an open problem [28]. Ontology grounding does not solve the problem. Rather, it allows the user to decide the meaning of concepts to some extent.

Grounding the ontology upon the data allows reasoning with the data in particular context. Thus in a particular context a river could be a channel that contains water for a particular period of time as opposed to a permanent flow.

To ground the ontology upon the data, we need to work at both the data level and the ontology level. At the ontology level, we need to consider what attributes we require in order to identify or reason about a feature, whilst at the data level we need to consider how we will obtain such attributes. For example, linearity is an important concept when analysing geographical domains, as the way a feature's shape changes is often used to classify that feature.

So by identifying linear stretches within data, we have an attribute that can be passed to a grounded ontology to facilitate reasoning about that feature. Because linearity is dependant on the data and the context it is used, we must ground the ontology upon the data to collect such an attribute.

Data representation

In order to ground the ontology upon the data, we need to represent the data in an appropriate manner. We need to consider what attributes we require and how these may be collected from the data provided. This is crucial to geographical objects, as often a feature is part of a larger feature, as opposed to being a unique object. Individuation is therefore more important in the geographical domain than in other domains.

The case study looked at here is for inland water networks. Previous work on an ontology for water networks was done in [3]. Here, formal concept analysis was used to determine the attributes required to reason

about water networks. The key attributes included flow, size and linearity, with flow and linearity are closely linked. So, we require a method of extracting linear stretches that could be passed to an ontology. We start with our initial polygon that represents the water network, and need to analyse the geometry to determine linear stretches. Linearity is a vague concept, so we will use techniques based upon supervaluation to determine when exactly a particular part of a polygon is considered linear. Thus the user sets the precisification for linearity.

The initial polygon of the water network is insufficient to reason about aspects such as flow or linearity effectively, so we require a better representation of the polygon. The medial axis of a polygon as first proposed by Blum [4] is defined as the locus of the centre of all the maximal inscribed circles of the polygon. Here, a maximal inscribed circle is a circle that cannot be completely contained within any other inscribed circle in the polygon [10].

The benefits of using the medial axis in relation to river networks is discussed in [22], and was suggested in [3] as a way of determining the linearity of stretches of river. The medial axis (or skeleton) has also been used in similar problems to determining river junctions, such as road networks [16].

There are numerous methods for calculating the medial axis, such as extraction from the Voronoi diagrams [5, 14, 19], fast marching methods [30], the divergence of flux [6], and use of the Euclidean distance transform [10, 15, 23, 26].

A Voronoi diagram based approach offers a relatively simple and efficient method of obtaining the medial axis, as the medial axis is a sub graph of the Voronoi diagram for a simple polygon, and so we need only delete the unnecessary Voronoi edges. The VRONI approach and program developed by Held [14] produces Voronoi diagrams and associated derivations such as the medial axis.

The Voronoi/medial axis approach could also be suitable for other areas of the geographical domain. For example, the density of buildings within a village could be analysed using a voronoi diagram, whereby the size of the cells represents the density of buildings.

Figure 1 shows the result of calculating the medial axis of our input file of the Hull Estuary. Because we are only interested in inland water features, the medial axis of the sea was removed, leaving only the medial axis corresponding to the inland water network and a small extension beyond the river mouth.

Attribute collection

At an abstract level, the medial axis provides us with a useful and meaningful representation of the original shapes. For example, in Figure 1 the centre line of the river

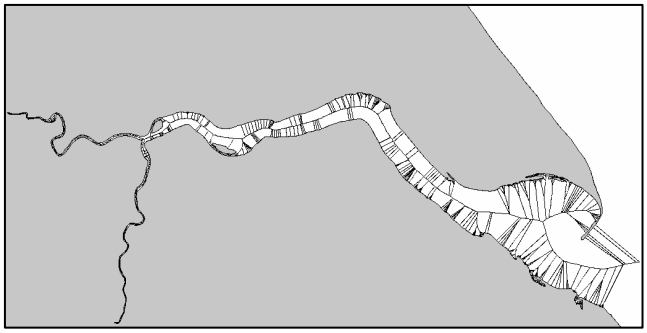

Figure 1: Medial Axis of The Hull Estuary

is easy to locate, and the number of lines in certain sections gives us an idea of the variation in shape in that area.

However, in order to extract any meaningful attributes to pass to an ontology, we must consider the relations between the data and determine the attributes to be extracted. The aim is to collect all the attributes required by an ontology grounded upon the data to reason about the features.

The medial axis is easily translated into a graph. The output from VRONI is a series of arcs, where the radius at one end of the arc is the smallest of the maximal discs on the arc, and the largest radius at the other end. The radii in between are therefore a transition between the two. By recording a point each side of the arc that these min-max radius touch the original polygon sides, we can construct a polygon from an arc or series of connected arcs. We can therefore translate the VRONI arcs into a graph, with the ends of the arcs being the nodes.

We also want to consider series' of arcs, by joining arcs considered to be part of the same channel. One method of determining what arcs to join together is to use approaches used to determine flow through the river network [12, 24, 25]. There are limitations to such approaches, as they assume that lakes and islands do not occur within the network, although Mark [20, 21] suggests that except in rare circumstances lakes do in fact have only one downstream flow. By applying the algorithm to our graph structure, we have an efficient method of determining what arcs to join together into 'superarcs'. We now have an effective method of representing the river network, and a basis from which to collect attributes.

Marking linear stretches

We could calculate whether a stretch is linear in a variety of ways. For our case study, we require the method to be scale invariant, as the size of the channels may vary dramatically.

To determine if a point is linear, we first find all the medial axis points that are on the same superarc within the maximal inscribed circle at that point. We then examine the radius at each of the points, determining the variance between minimum and maximum. If the variation of these widths is below some threshold, then the point is linear.

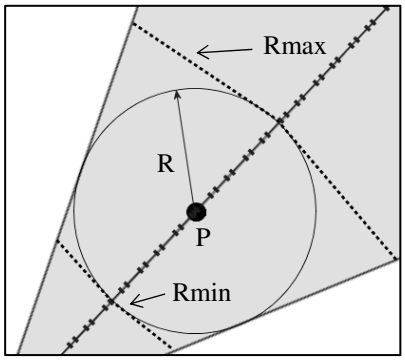

Figure 2: Example of linearity testing

Figure 2 demonstrates this process. Suppose we have the polygon as coloured grey. The medial axis at this section is a simple bisector of the two sides, represented by the dashed line that point P resides on. The maximal inscribed disc is the circle with radius R as shown. To determine if P is linear, we take all points on the medial axis that are within R distance of the point (all points of the medial axis contained by the maximal inscribed disc of P). We then find the radius of the maximal inscribed disc at each of the points, searching for the maximum and minimum values. In our example, these values will clearly be at the points a distance R from P. These are the dotted lines in the figure, marked Rmin and Rmax. If the variation between R-Rmin and R-Rmax is below a certain threshold, then we say the point is linear, as the width of the channel is only varying by a small amount.

So we now have a method of measuring the linearity in relation to the width. The approach is scale invariant, since larger rivers will require more points and smaller rivers will require fewer. This stage can therefore output sets of connected arcs within a superarc that are linear.

Marking gaps to be filled in

Depending on the precisification used, the previous stage may not find all the required stretches. Gaps may occur at sharp bends in a channel or sudden bulges. We could eliminate gaps by changing the degree of linearity required by the program, but in doing so we may end up classifying other sections as linear that we did not want to do so.

It is therefore intuitive to have 'gap' as an attribute that can be collected, whereby if a gap exists between two linear stretches and this gap is small enough (and thus been marked as 'gap'), we can join the stretches together into a major stretch. As with linearity, we require the measurement to be scale invariant.

We first search superarcs for any gaps between linear stretches. Given our graph structure, these are easily found, as each superarc represents a cycle-free path between the start and end nodes of that superarc. We can therefore simply traverse this path searching for gaps between linear stretches.

We calculate the length of the gap by adding the lengths of the arcs within the gap, and calculate the mid-point of the gap, obtaining the radius of the maximal inscribed disc at this point. If this value multiplied by a given threshold is greater than the length of the gap, then the gap is deemed sufficiently small and is marked with the 'gap' attribute. This approach is scale invariant, since larger gaps will require larger radius values at the mid-point in order to be marked as 'gaps'.

Result of marking major stretch

Arcs within the model are now labelled depending on the linearity and gap precisifications, and allow segmented polygons to be generated. We can now classify three simple features; linear stretches, gaps between stretches, and finally major stretches. Here, major stretches are defined as the union of linear stretches and gaps between that are sufficiently small.

The reason these are separate is because principal reasoning of features is to occur at the ontological level. This stage is to collect the attributes that are to be reasoned about at a higher level. The definitions of these attributes is now grounded upon the data, as the attributes 'linear' and 'gap' are unclear unless they are defined within the context of the data. This further ensures that the ontology will be grounded upon the data.

Figure 3 shows the results of these stages, having taken The Hull Estuary as input, with major stretches marked grey and the original medial axis shown in black. The system was developed in Prolog.

Despite only using two attributes, the system is able to mark major stretches stretching along the channels, including around islands. However, there are additional interesting results. First, the polygon generated as major stretches does not always go fully to the edge of the polygon, with occasional inlets missed out.

An example of this is shown in Figure 4, where a small inlet is not part of the major stretch. In some cases, we may want such an inlet to be part of the stretch, but there exist other cases where we would want that to be a separate feature; for example the inlet may in fact stretch out a long

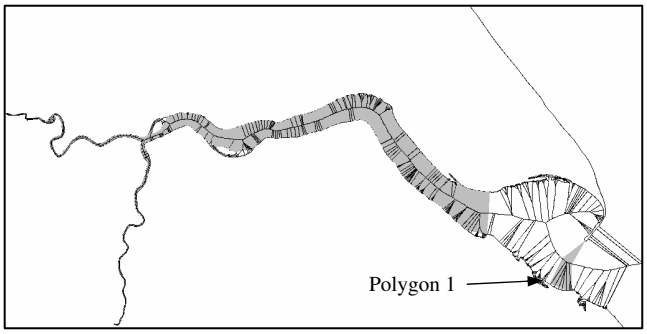

Figure 3: The result of marking major stretches, marked grey, with the original medial axis shown again in black. Polygon 1 represents a surprising result

way or be another channel. Therefore the most appropriate way to deal with this is in a similar fashion to gaps as previously discussed, and design an attribute for such inlets.

The other interesting result is the polygon occurring at the river mouth at Spurn Head, labelled polygon 1. At a first glance this does not seem to be a linear polygon. However, if we imagine travelling in a boat and attempting to remain roughly equidistant from both sides, we would find ourselves travelling in an arc that kept us equidistant from Spurn Head and the south bank of the river, as Spurn Head would be the closest point to the north of us.

To rectify this, we require a reconsideration of our definition of linearity. This particular result suggests that in order for a shape to be linear, we require both the variation in the width to remain small and also the variation in the curvature of the sides.

Future work

The present attributes used would only allow two different features to be considered; linear and non-linear stretches. However, there are many other attributes to be considered (including size, islands or temporal attributes), which in turn will allow us to reason about other features.

A more important stage is to feed the results in an ontology, so reasoning over the features can occur. This grounded ontology will be able to handle the vague entities contained within depending on the user's preferences. The ontology could be built in existing ontology languages such as OWL, as OWL can be inputted into Prolog for the reasoning stage [34, 18].

Conclusion

In this paper we have shown how geographical data can be represented and attributes collected to allow the grounding of an ontology. We have compared fuzzy logic and supervaluation theory, showing why they are suited to different tasks and why supervaluation theory is best suited to our particular problem.

We have also shown how the representation of the data is an important consideration, and that we must find the most effective method of representing the data.

Finally, we used the new representation to collect simple attributes that could then be passed to an ontology to reason about the features. In doing so, we have shown that adding these stages to the design process will allow a manner of reasoning about vague geographical features.

Acknowledgements

I am grateful for comments from Brandon Bennett and Allan Third who greatly helped with the presentation of the paper and my work. I am also grateful to our industrial partner Ordnance Survey, for their contribution towards PhD funding. Finally, I appreciated the comments of the reviewers that helped me prepare the final text.

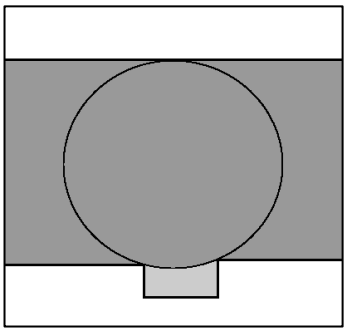

Figure 4 Simple example of how the stretch segmentation doesn't necessarily fill out the whole polygon. Dark grey represents a linear stretch, light grey a part marked as non-linear. The circle represents a maximal inscribed disc to illustrate why this occurs

References

[1] *Concise Oxford English Dictionary*, 2004.

[2] B. Bennett, *What is a forest? On the vagueness of certain geographic concepts*, Topoi-An International Review Of Philosophy, 20 (2001), pp. 189-201.

[3] B. Bennett, G. Sakellariou and P. Santos, *Supervaluation Semantics for an Inland Water Feature Ontology*, (2004).

[4] H. Blum, *Biological Shape And Visual Science.1*, Journal Of Theoretical Biology, 38 (1973), pp. 205-287.

[5] F. Chin, J. Snoeyink and C. A. Wang, *Finding the medial axis of a simple polygon in linear time*, Discrete & Computational Geometry, 21 (1999), pp. 405-420.

[6] P. Dimitrov, C. Phillips and K. Siddiqi, *Robust and efficient skeletal graphs*, *IEEE Conference On Computer Vision And Pattern Recognition, Proceedings, Vol I*, 2000, pp. 417-423.

[7] D. Dubois, F. Esteva, L. Godo and H. Prade, *An information-based discussion of vagueness*, *10th Ieee International Conference On Fuzzy Systems, Vols 1-3 - Meeting The Grand Challenge: Machines That Serve People*, 2001, pp. 781-784.

[8] K. Fine, *Vagueness, Truth and Logic*, Synthese, 30 (1975), pp. 265--300.

[9] F. Fonseca, M. J. Egenhofer, C. Agouris and C. Cmara, *Using Ontologies for Integrated Geographic Information Systems*, Transactions in Geographic Information Systems, 6 (2002).

[10] Y. R. Ge and J. M. Fitzpatrick, *Extraction of maximal inscribed disks from discrete Euclidean distance maps*, *1996 Ieee Computer Society*

Conference On Computer Vision And Pattern Recognition, Proceedings, 1996, pp. 556-561.

[11] N. Guarino, *Understanding, building and using ontologies*, International Journal Of Human-Computer Studies, 46 (1997), pp. 293-310.

[12] R. M. Haralick, S. Wang, L. G. Shapiro and J. B. Campbell, *Extraction Of Drainage Networks By Using The Consistent Labeling Technique*, Remote Sensing Of Environment, 18 (1985), pp. 163-175.

[13] S. Harnad, *The Symbol Grounding Problem*, Physica D, 42 (1990), pp. 335-346.

[14] M. Held, *VRONI: An engineering approach to the reliable and efficient computation of Voronoi diagrams of points and line segments*, Computational Geometry-Theory And Applications, 18 (2001), pp. 95-123.

[15] W. H. Hesselink, M. Visser and J. Roerdink, *Euclidean skeletons of 3D data sets in linear time by the integer medial axis transform*, *Mathematical Morphology: 40 Years On*, 2005, pp. 259-268.

[16] W. Itonaga, I. Matsuda, N. Yoneyama and S. Ito, *Automatic extraction of road networks from map images*, Electronics And Communications In Japan Part Ii-Electronics, 86 (2003), pp. 62-72.

[17] A. Jakulin and D. Mladenić, *Ontology Grounding*, *SiKDD 2005*, Ljubljana, Slovenia, 2005.

[18] L. Laera, V. Tamma, T. Bench-Capon and G. Semeraro, *SweetProlog: A system to integrate ontologies and rules*, *Rules And Rule Markup Languages For The Semantic Web, Proceedings*, Springer-Verlag Berlin, Berlin, 2004, pp. 188-193.

[19] D. T. Lee, *Medial Axis Transformation Of A Planar Shape*, Ieee Transactions On Pattern Analysis And Machine Intelligence, 4 (1982), pp. 363-369.

[20] D. M. Mark, *On The Composition Of Drainage Networks Containing Lakes - Statistical Distribution Of Lake In-Degrees*, Geographical Analysis, 15 (1983), pp. 97-106.

[21] D. M. Mark and M. F. Goodchild, *Topologic Model For Drainage Networks With Lakes*, Water Resources Research, 18 (1982), pp. 275-280.

[22] M. McAllister and J. Snoeyink, *Medial Axis Generalization of River Networks*, CaGIS, 27 (2000), pp. 129 -138.

[23] A. Meijster, J. Roerdink and W. H. Hesselink, *A general algorithm for computing distance transforms in linear time*, *Mathematical Morphology And Its Applications To Image And Signal Processing*, 2000, pp. 331-340.

[24] J. A. C. Paiva and M. J. Egenhofer, *Robust Inference of the Flow Direction in River Networks*, Algorithmica.

[25] J. A. C. Paiva, M. J. Egenhofer and A. U. Frank, *Spatial Reasoning about Flow Directions: Towards an Ontology for River Networks*, XVII International Congress for Photogrammetry and Remote Sensing, 24 (1992), pp. 318-224.

[26] E. Remy and E. Thiel, *Exact medial axis with euclidean distance*, Image And Vision Computing, 23 (2005), pp. 167-175.

[27] B. Smith and D. M. Mark, *Ontology and geographic kinds*, Proceedings, International Symposium on Spatial Data Handling (1998).

[28] M. Taddeo and L. Floridi, *Solving the symbol grounding problem: a critical review of fifteen years of research*, Journal Of Experimental & Theoretical Artificial Intelligence, 17 (2005), pp. 419-445.

[29] M. P. Taylor and R. Stokes, *When is a River not a River? Consideration of the legal definition of a river for geomorphologists practising in New South Wales, Australia*, Australian Geographer, 36 (2005), pp. 183-200.

[30] A. Telea, *An Augmented Fast Marching Method for Computing Skeletons and Centerlines*, in D. Ebert, P. Brunet and I. Navazo, eds., *EG/IEEE TCVG Symposium on Visualization*, ACM Press, 2002.

[31] E. Tomai and M. Kavouras, *From "onto-geoNoesis" to "onto-genesis": The design of geographic ontologies*, Geoinformatica, 8 (2004), pp. 285-302.

[32] A. C. Varzi, *Vagueness in Geography*, Philosophy & Geography, 4 (2001), pp. 49-65.

[33] A. C. Varzi, *Vagueness, Logic, and Ontology*, The Dialogue. Yearbooks for Philosophical Hermeneutics 1 (2001).

[34] J. Wielemaker, *An optimised Semantic Web query language implementation in prolog*, *Logic Programming, Proceedings*, Springer-Verlag Berlin, Berlin, 2005, pp. 128-142.

[35] L. A. Zadeh, *Fuzzy Algorithms*, Information And Control, 12 (1968), pp. 94-&.

[36] L. A. Zadeh, *Fuzzy Sets*, Information And Control, 8 (1965), pp. 338-&.

[37] L. A. Zadeh, *Information And Fuzzy Sets*, Proceedings Of The American Society For Information Science, 13 (1976), pp. 83-83.