Geographical Information Retrieval with Ontologies of Place

Christopher B. Jones[1], Harith Alani[2] and Douglas Tudhope[3]

[1]Department of Computer Science, Cardiff University
Queens Buildings, PO Box 916, Newport Road, Cardiff CF24 3XF, United Kingdom
email: c.b.jones@cs.cf.ac.uk

[2] Department of Electronics and Computer Science, University of Southampton
[3]School of Computing, University of Glamorgan

Abstract. Geographical context is required of many information retrieval tasks in which the target of the search may be documents, images or records which are referenced to geographical space only by means of place names. Often there may be an imprecise match between the query name and the names associated with candidate sources of information. There is a need therefore for geographical information retrieval facilities that can rank the relevance of candidate information with respect to geographical closeness as well as semantic closeness with respect to the topic of interest. Here we present an ontology of place that combines limited coordinate data with qualitative spatial relationships between places. This parsimonious model of place is intended to support information retrieval tasks that may be global in scope. The ontology has been implemented with a semantic modelling system linking non-spatial conceptual hierarchies with the place ontology. An hierarchical distance measure is combined with Euclidean distance between place centroids to create a hybrid spatial distance measure. This can be combined with thematic distance, based on classification semantics, to create an integrated semantic closeness measure that can be used for a relevance ranking of retrieved objects.

Keywords: relevance ranking, similarity measures, thesauri, gazetteers

1 Introduction

For users of the world-wide web, information retrieval has become an everyday activity. The search engines associated with web browsers are used to find information relating to most domains of human activity. A large proportion of that information may be regarded as embedded in geographical space and, as a consequence, many users will wish to specify geographical place names as part of their query. A characteristic of much of the research into data access methods for geographical information systems is that it has been targeted towards handling coordinate-based geometric representations of space. Yet most people use place names to refer to geographical locations, and will usually be entirely ignorant of the corresponding coordinates.

The previous emphasis upon coordinates is fully justified by the need to retrieve, analyse and display graphically the wealth of map-based and primary-surveyed data that are referenced to geographical and map grid coordinates. However, as more information becomes available both to environmental and social scientists and to non-specialists, the need arises to provide more intelligent information retrieval facilities that can recognise natural language concepts of space and time (Agosti et al. 1993). It is also the case that much geographically-referenced information is identified by place names and other non-spatial classification terms that are not directly accessible by coordinate-based indexing methods. Examples of such data arise in modern and historical textual documents, including records of cultural and natural-

environmental events and descriptions of material kept in museums, research institutes and other archival repositories.

Simple word matching of the sort that is used in search engines is not adequate for many purposes of geographical retrieval. There is a need for systems that perform imprecise matching of place-name terminology. Thus if the user specifies a place name in a query, then the retrieval system should find references to the same or similar places that may be referred to by different names, or may be at different levels of the administrative or topographical hierarchy, or may be nearby due to connectivity or to some other measure of proximity (Walker et al. 1992; Jones et al. 1996; Larson 1995; Moss et al 1998). Having found candidate matches for the query it should also be possible to rank them according to relevance to the user. This requires the use of similarity or closeness measures expressed in terms of place name concepts. Assuming that the user has expressed an interest in some non-spatial concept then available information should also be matched for relevance to that concept. A final ranking of search results might then be expected to combine spatial and "thematic" concepts.

The earliest attempts to enable users of geographical information to refer to place names when making queries were based on the use of simple gazetteers in which each place name is associated with a map-grid or geographical coordinate. The coordinate is then used typically as the basis of a conventional coordinate-based query to a GIS. The information processing is therefore brought immediately back into the world of coordinate space. The potential of the simple gazetteer-based retrieval facilities may be increased by encoding relationships between names. The Thesaurus of Geographic Names (TGN) was a notable development in this respect (Harpring 1997). In the TGN hierarchical relationships between names relating to administrative areas and to some physical features are recorded. In addition the TGN stores alternative versions of place names as well as a single geographical coordinate. A metadata approach to gazetteer standardisation has been proposed in Hill et al. (1999) in the context of the Alexandria Digital Library. This supports the possibility of encoding semantic or spatial relationships between places. It also leaves flexibility regarding the storage of a geometric "footprint". In practice a footprint may be no more than a single representative point, or centroid, that is located within the areal extent of the place, but it might also be a polygon or a set of points.

The need for information retrieval facilities that recognise domain-specific terminology has led to various efforts to construct and exploit ontologies that model the associated concepts (e.g Guarino et al. 1999). In the field of information science, research into thesauri has led to the development of a range of semantic net and thesaurus-mediated information retrieval techniques, for which a variety of semantic closeness metrics have been designed (e.g. Rada et al. 1989; Lee et al. 1993; Richardson et al. 1994). Although the importance of ontologies for representing aspects of geographical information has been clearly highlighted in several studies, notably with regard to the representations of boundaries of geographical phenomena (e.g. Smith 1995 and Smith and Mark 1998) relatively little progress has been made on the practical representation of concepts of place specifically for purposes of information retrieval.

This paper is concerned with creating a model of place that may be exploited for purposes of information retrieval. In this respect place names play an essential role. The motivation here is to provide a method for matching a specified place name with place names that refer to equivalent or nearby locations. We are not concerned in the present study with the problem of finding places that are conceptually similar but possibly entirely separate in location. Within the substantial body of literature on the subject of place (e.g. Relph 1977; Yuan 1977; Gould and White 1986; Johnson 1991; Curry 1996; Jordan et al 1998) a common theme is that it reflects human experience of space and the meanings that we impose upon space. Thus Couclelis (1992) locates the term place in the context of "experiential space". There is little doubt that individual places may be imbued with personal associations for users of the place name, but these experiential aspects cannot be assumed to be relevant in the use of the place name as a locator. Many place names refer to regions representing official administrative categorisations of space and these regions are typically embedded within regional hierarchies that have considerable potential to assist in expanding queries that employ place names to include geographically related places. The use of the term place in this context may be regarded therefore as perhaps nearer to that of Johnson (1991) who adopts for place some of the concepts of Paasi concerned with institutionalisation of regions. With regard to the various types of

ontology (Guarino in press), our use is that of a quite narrowly focused domain ontology. It is intended however that the approach should be extensible to support richer models of place that may be used in a wider variety of applications.

An objective here is to rank the results of imprecise geographical queries using place names, that might be global in scope across a wide range of scales. This raises questions of the appropriate types of relationships and semantic attributes to maintain for such applications and leads to the idea of parsimonious spatial models that record the least amount of information necessary to process the queries. Equally important for us is the development of similarity measures that are based on the model of place and that can be used for ranking search results.

In the next section of the paper we discuss some of the options for place ontologies, for use in information retrieval. In Section 3 we introduce a place schema that has been implemented in the context of a cultural heritage information system, using a semantic data modelling system. Sections 4 and 5 propose some measures of semantic distance for purposes of ranking search results based on queries that include place names and concepts. Section 6 summarises some experimental results of ranking search results on the basis of the proposed closeness measures. The paper concludes in Section 7 with a discussion of the results and future work.

2. Ontologies of Place

2.1 Requirements

As indicated above, we are concerned with a conceptualisation of place that supports the measurement of locational similarity between named places. The objective is to implement procedures that match a given named place to named places that are equivalent or similar in geographical location. It is assumed that a place may refer to any geographical phenomenon, provided that it has been given a name or literal description. Examples of *referents* of place include therefore physical features of the Earth's surface such as forests, lakes, rivers and mountains, in the natural realm, and cities, counties, roads, and buildings in the human-made environment. This scope of types of place encompassses those with *fiat* and *bona fide* boundaries in the terminology of Smith (1995), that may have either crisp or fuzzy boundaries, as determined by physical, political, social or other cognitive parameters.

As regards the interpretation of similarity, this will be context dependent. Thus it should be possible to expand a search for similar places to any distance from the given named place. This expansion should be possible with respect to contained and containing places and with respect to overlapping, connected and separate places.

In the search for an appropriate ontology of place for information retrieval, we must also take account of the requirement for great, possibly global, geographical extent across a wide range of scales. An important question in this context is whether there is a requirement for inclusion of references to mapped coordinate-based representations of named places. If coordinate data for a global representation of geographic space were ever to be available in a single database, that database would of course be extremely large. Given the need to deal with places for which there may be no exact coordinate-based representation then, even if such a database existed, the procedures that operated upon it would still only apply to a part of the domain of interest.

2.2 Parsimonious Spatial Models

It is proposed here that for many practical purposes, detailed geometric data are not necessary, as well as not being desirable for the reasons just stated. A more pragmatic approach may be to maintain a parsimonious spatial model of geographical place, in which minimal coordinate data, such as centroids, are maintained in combination with qualitative spatial relationships of topology and proximity. This model of

place is related therefore very much more closely to that of geographical thesauri and "rich" gazetteer models than it is to a conventional GIS.

The approach adopted is based on the assumption that a great deal of important spatial information for purposes of information retrieval can be pre-computed from coordinate-based data, or computed on-the-fly from the sparse coordinate data that are stored. Firm topological relationships of connectivity and overlap between regions with digitised boundaries can easily be extracted from digital map datasets. Similarly attributes that might be of significance when evaluating locational similarity measures, such as area and the length of common boundaries, can again be derived from the digital map data.

2.3 Approximating Spatial Footprints

Examples of Euclidean distance measures that might not so easily be pre-computed and stored are those of the distance of a centroid to a boundary, and a boundary to another boundary, since for any given centroid or boundary there may be many other boundaries to which distance might need to be computed. A solution to this problem, that avoids storing boundary data explicitly, is to compute approximate boundaries on-the-fly from stored centroids, prior to calculating the distances. One method of doing this is to construct Voronoi diagrams of the set of centroid-referenced locations known to be inside a region of interest, in combination with nearby centroids known to be outside. The region is then approximated by the set of Voronoi polygons associated with the internal sites. Knowledge of containment and exclusion can be derived from stored topological relations of region containment and of region connectivity. Experiments reported in Alani et al. (2001) demonstrate that these methods can result in reasonably good quality estimates of regional area and boundary location, with area and boundary length approximation typically within 5% of the corresponding values measured on the original digital map data.

An attraction of Voronoi-based methods of region approximation is that they can be applied to the estimation of imprecise regions for which there may be no existing digitised boundary but for which there may be knowledge of included and external places.

3 Modelling Place in OASIS

Here we explain how place has been modelled in an experimental terminology system called OASIS (Ontologically-Augmented Spatial Information System). It has been built using the Semantic Index System (SIS) which is an object-oriented hypermedia management system developed by ICS-FORTH (Doerr and Fundulaki 1998). SIS provides multiple classification levels starting with the Token level, above which are a Simple Class and successive Meta Class levels. Both classes and their associated attributes are treated in SIS as objects that can have names, attributes and relationships to other levels. Access to the SIS database can be programmed with the Programmatic Query Interface (PQI) functions, in combination with C++ programs.

OASIS has been used to maintain cultural information about archaeological finds and historic buildings that have been classified using terms from the Art and Architecture Thesaurus (AAT) and referenced geographically using place names associated with the data and linked into the Thesaurus of Geographical Names. A *Place* class has been implemented in OASIS as a type of *Geographical Concept.* The thematic categories of place are specified by current and historical place types. A place may have multiple place types allowing instances of place to be characterised by physical, cultural or administrative classes. In our implementation, instances of place are taken mostly from the Thesaurus of Geographic Names, augmented with Bartholomews digital data for the UK, and the associated place types are from the AAT.

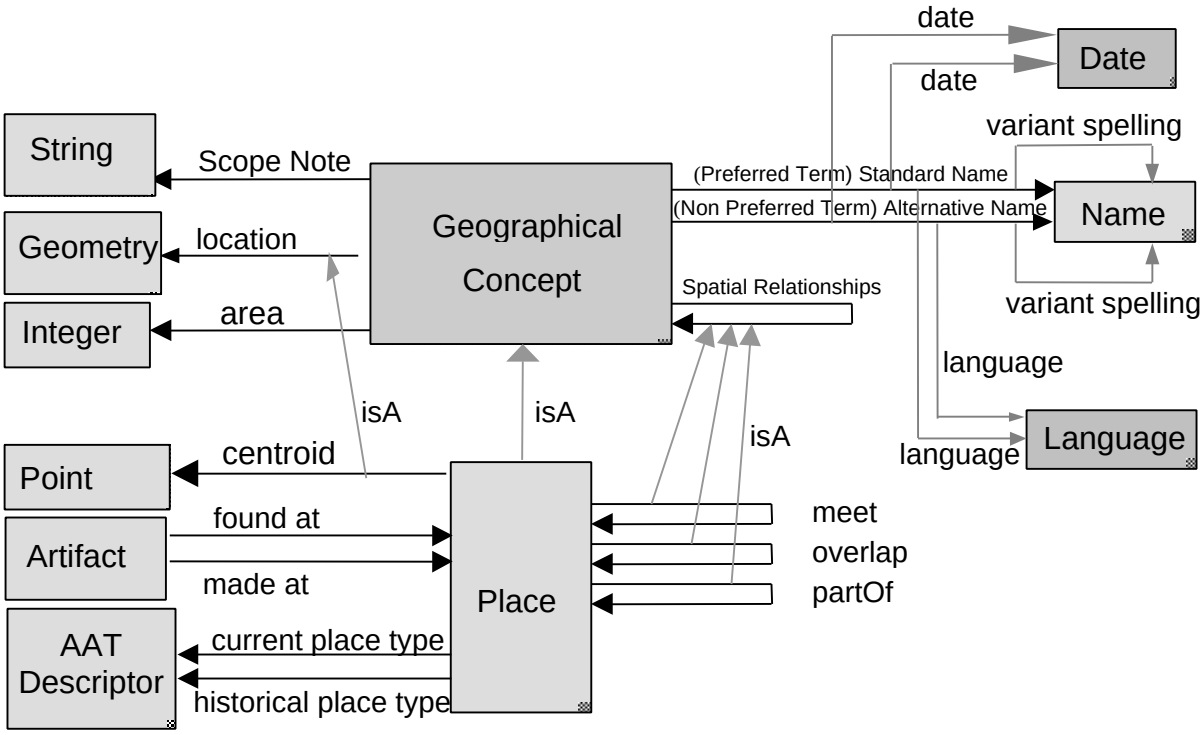

Fig. 1. Place as a type of geographical concept

Figure 1 illustrates the schema for *Place* and shows how it inherits various attributes and relationship types from *Geographical Concept*. A *Geographical Concept* has a *Standard Name* (or Preferred Term) and *Alternative Names* (Non-Preferred Terms). These names are associated with alternative spellings, a date of origin and a language. A scope note provides a verbal explanation of the concept. *Geographical Concept*s may be associated with a location defined by a geometry object, an area measurement value and spatial relationships. In the *Place* class, location is specialised to a centroid, defined by latitude and longitude coordinates, and spatial relationships are specialised into the *meet*, *overlap* and *partOf* relationships. The *Artefact* class has attributes of *date-found*, *type* and *description* and relationships of *made-of* to the *Material* Class (not shown here). As illustrated in the figure, the *Artefact* class is associated with the *Place* class via *found_at* and *made_at* relationships.

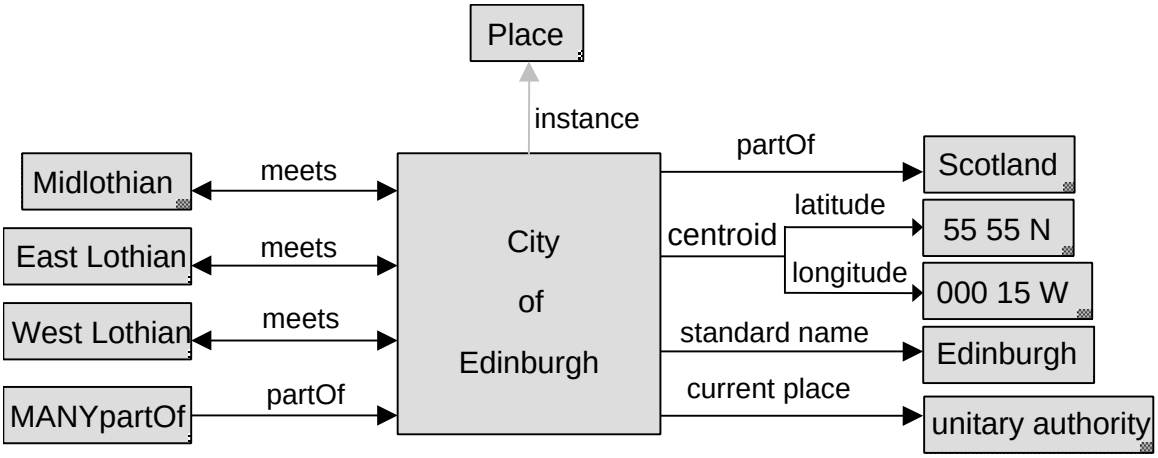

Fig. 2. An example of an instance of place

Figure 2 illustrates the classification of City of Edinburgh as an instance of the *Place* class. Note that its has three *meets* topological relationships with the administrative adjoining regions of Midlothian, East

Lothian and West Lothian and that it is the parent for a set of *partOf* relationships with multiple places that are referenced in the figure by the *MANYpartOf* object. It should be noted here that the *partOf* relationship itself has attributes, notably of *dates*, enabling temporally-specific query expansion.

When there is more than one place with the same name, only one instance of the repeated name is maintained, but a unique name for each of the associated places is created. These unique name objects are then linked via *Standard Name* or *Alternative Name* relationships to the non-unique name. An example is provided in Figure 3 in which the name Hull is the standard name for the Hull in Canada (Canada_Hull) while it is the alternative name for the Hull in the UK (UK_Hull), the official (standard) name of which is Kingston upon Hull. This approach facilitates place name disambiguation, since access to the non-unique place name leads directly to all associated places. The user can then select the place of interest.

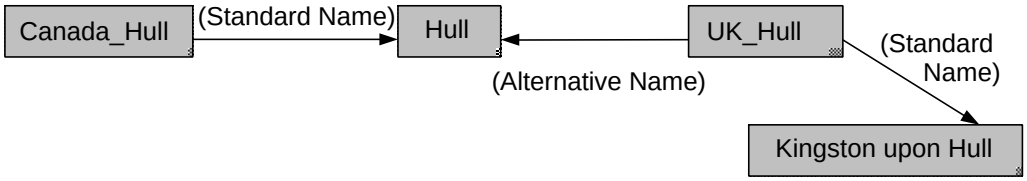

Fig. 3. Example of the use of a single place name to that refers to multiple places

OASIS has been populated with cultural heritage data from the Royal Commission on the Ancient and Historical Monuments of Scotland (RCHMS). An example of the schema for a particular artefact, axe number DE121, is illustrated in Figure 4 which illustrates several relationships including the *found_at* relationship to a place.

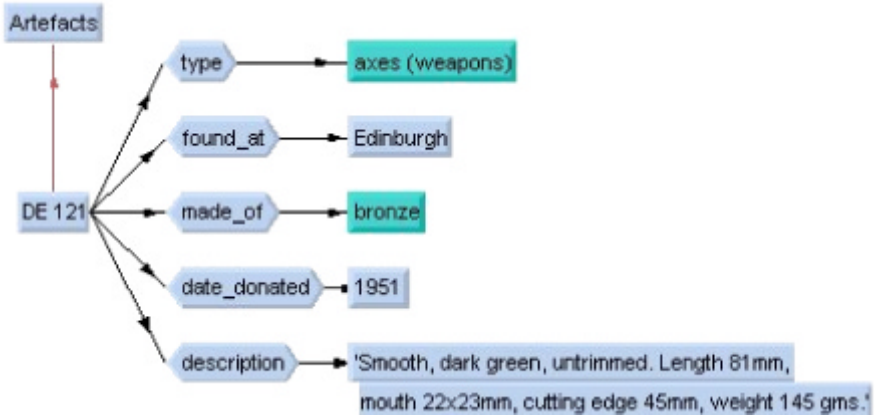

Fig. 4. Example of a particular artefact as illustrated in the OASIS user interface.

4 Locational Similarity Metrics

4.1 Relevant Criteria

In developing or choosing closeness metrics for ranking places in information retrieval we focus upon those aspects of geographical space that are concerned primarily with location and proximity. We are not concerned with matching places that may be similar in shape or structure unless they refer to the same or similar locations on the earth's surface. We treat place names as descriptors for regions of space. Thus if location is regarded as a substrate for concrete objects (Guarino 1997), then place names refer to specific locations. The referents of place serve to locate specified phenomena that are of interest to users of the place names. For example an archaeological query may request "axes in Edinburgh", in which the user may

be interested in occurrences of axes in places that are inside Edinburgh, or refer to the same location as Edinburgh or its constituent parts, as well perhaps as places that are somehow near to Edinburgh.

Potential criteria for assessing locational similarity between a specified place and a candidate place, when searching for information, include the following:

- distance in map or geographical coordinate space between query and candidate;
- travel time between query and candidate;
- number of intervening places;
- spatial inclusion of the candidate within the query place;
- containment of the query place by the candidate;
- containment of candidate within, or overlap of candidate with, regions that contain or overlap the query place;
- boundary connectivity between query and candidate.

Because the motivation here is information retrieval we do not make any assumptions about the user's familiarity with the places which they specify. In this context therefore cognitive measures of closeness as perceived by people living in or familiar with the places may not be relevant to determining locational similarity. The criteria are therefore based on more concrete attributes of places. This is not to rule out the potential of cognitive measures but we regard them as a special case.

In the present study we have focused initially on the use of geographical hierarchies in combination with Euclidean distances. Search expansion methods based on these aspects of space will automatically find connected places and will tend to give them higher priority than disconnected places. Euclidean distance whether in map-grid space or as measured on the Earth's surface leads to a ranking based on physical proximity, and introduces the possibility of constraining the expansion of a search for similar places according to specified distance thresholds. While Euclidean distance is undoubtedly valuable for measuring locational similarity, it fails to take account of characteristics of *place* as determined by physical, social and political factors. Regional hierarchies have great potential in distinguishing between the relevance of nearby places since they record facets of interpretation of the place of interest with regard to the topographic and human environment and allow for the possibility of making distinctions between the importance of different aspects of geographical space according to the interests of the user. In this regard the possibility of inclusion within multiple hierarchies is essential (as opposed to the single hierarchies imposed for example by the TGN).

Travel time is not considered here at this stage since it is so dependent upon the means of transport, which would need to be assumed, or specified by the user. Euclidean distance however acts as a surrogate for travel time that may or may not be appropriate depending on the local topography and the interests of the user.

The number of intervening places between query and candidate place provides the possibility of a qualitative measure of distance. It is a familiar concept in analysing geographical accessibility and has been proposed as a measure of spatial similarity for purposes of information retrieval (Jones et al 1996). It has not been included in the present study, partly due to negative results of some preliminary user tests of its significance. However it is still regarded as potentially useful.

One method of determining similarity between two objects is based on their common and distinctive (non-common) features (Tversky 1977). An example of the application of a feature-based spatial entity class similarity measure is found in Rodriguez et al. (1999), who integrated measures representing parts, functions and attributes respectively. If a conceptual hierarchy is available then terms within the hierarchy can be compared by measuring the distance between them along the branches of the corresponding graph. Following Rada et al. (1989) this distance is then equal to the number of connecting links in the shortest path in the graph. Calculation of distance within hierarchies can be refined by applying weights to links in the corresponding graph (Kim and Kim 1990). The weight or importance of a node in the graph may also be related to the inverse of its depth in the hierarchy.

In the context of hierarchies and poly-hierarchies, an alternative approach to similarity or distance measurement is to consider all the non-common parents (at whatever level) of the respective nodes, each of which may have a weight inversely proportional to the depth in the hierarchy. This introduces a measure of the degree to which the nodes differ in their inherited categories, that is sensitive to the hierarchical levels. Semantic distance between a pair of terms increases in proportion to the number of distinctive (non-common) parents. The use of non-common super-classes has been advocated by for example Spanoudakis and Constantopoulos (1994) and Sintichakis and Constantopoulos (1997) who have demonstrated the approach in the context of similarity metrics based on a combination of generalisation relations, classification and attributes.

4.2 Hierarchical Distance Measure

Here we adapt the methods based on non-common super-classes to geographical poly-hierarchies using generic *part-of* relations that may be interpreted spatially as inside or overlap. The parent regions provide units of space, or substrates, that assist in determining commonality of location, if places lie in the same parent region, and differences of places that belong to separate regions. It is assumed that a place is characterised by the sum of the geographical regions, or other parent places, to which it belongs either directly or by inheritance within a hierarchy. A town for example may be inside or overlap a county that itself is part of the formal hierarchical administrative subdivision of a nation, which is itself part of a global geopolitical hierarchy. The same town might belong to a physiographical hierarchy based on predominant features of the landscape, giving rise to descriptors such as "Great Plains", "Mackenzie Delta", "Southern Uplands" within which might be subdivisions consisting of particular named river valleys, mountains or marshes. There might be single level subdivisions, based for example on national parks, military installations or ethnic groups that further characterise the place either completely or partially.

We define the Hierarchical Distance Measure (HD) between query place q and candidate place c as follows:

$$HD(q,c) = \sum_{x \in \{a.PartOf - b.PartOf\}} \frac{a}{L_x} + \sum_{y \in \{b.PartOf - a.PartOf\}} \frac{b}{L_y} + \sum_{z \in \{a,b\}} \frac{g}{L_z} \quad (1)$$

The L_x, L_y and L_z values represent the hierarchical levels of the individual places within their respective hierarchies. The set of places x are those distinctive super-parts of the query term that belong to it but not to the candidate, while places y are the distinctive super-parts of the candidate that are not shared with the query. The places z are the query and candidate terms themselves. The sets of terms *q.PartOf* and *c.PartOf* refer to the transitive closure of the super-parts of q and c respectively in the part-of hierarchy. The weights α, β and γ provide control over the application of the measure. In particular the weights α and β provide the option of asymmetry, making candidates that are sub-parts of the query more (or less if required) similar to the query than are candidates that are super-parts. Tversky (1977) reported experiments indicating that asymmetry is observed in people's perception of the similarity of terms where one is more important in some sense than the other, for example if one term is the category or super-class of the other (e.g. bird *vs* sparrow).

The purpose of the weight γ is to provide control over the use of the query and candidate term levels in the distance measure. It is envisaged that if both the query and candidate are members of the same hierarchy then γ should be a non-zero value. Thus if q and c are both sub-regions of the same parent region within a particular hierarchy, this would result in a non-zero distance between them. However, if q and c are not both members of a regional hierarchy then in general γ would be zero. The consequence of this is that if both of the latter places belonged to the same parent region(s) in a hierarchy there would be no difference between them *with respect to the regional hierarchy*.

In general when applying the hierarchical distance measure, distance between query and candidate increases according to the number of non-common parents, i.e. the distinguishing regions. The level values increase with increasing depth in the hierarchies with the result that there are smaller differences between

pairs of places deeper down the hierarchy than there would be higher up. This is intended to reflect the idea that branches higher up a hierarchy introduce more significant differences than lower down. It should be noted that the formula measures distance explicitly with regard to distinguishing super-parts, while closeness is regarded as implicit within the branching structure of the hierarchies. Examples of the application of the measure are provided in section 6.

4.3 Euclidean Distance Measure (ED)

Because we are concerned here with applications that may be global in extent, we base measurement of Euclidean distance on latitude and longitude values for centroids. The Euclidean Distance measure calculates the great circle distance. As pointed out above, use of only the centroids of places that have areal extent produces ED values that reflect the separation of the approximated centres of the respective places. As indicated in Section 2.3 it is possible to create region approximations that may be used to measure distances between approximated boundaries, or between point sites and approximated boundaries.

4.4 A combined spatial closeness measure

The two locational distance measures can be combined in a weighted combination referred to as the Total Spatial Distance (TSD) as follows:

$$\mathrm{TSD}(q,c) = w_e\mathrm{ED}(q,c) + w_h\mathrm{HD}(q,c) \tag{2}$$

where w_e and w_h are weights of the ED and HD respectively. These weights lie in the range 0 to 1. In order to calculate a weighted combination of the individual distance measures as above, it is necessary to normalise both of the measures to a range between 0 and 1 prior to use.

5 Thematic Distance

In order to measure the semantic similarity between non-spatial concepts we introduce a thematic distance measure that is applied here within the mono-hierarchical classifications of the AAT. The purpose of the measure is to determine similarity between the query-specified phenomenon of interest and a candidate information object, with a view to combining this measure with those of spatial similarity. It may be noted however that a non-spatial similarity measure could also be used to compare places with regard to their place type categories, which as previously noted are also taken from the AAT in our current system.

The primary semantic relationships in the AAT are those of broader term (BT) which relates a term to its parent term within a hierarchy, narrower term (NT) which does the converse, and related term (RT). The BT and NT relations are the fundamental links between terms in individual hierarchies. In the context of the AAT, the BT relationship is one of semantic generalisation and hence is equivalent to the *is-a* relationship. The RT relationship records associations between terms that may be in different hierarchies. A detailed discussion of the application of the RT terms in OASIS can be found in Tudhope et al (2001).

Here we apply a weighted shortest path procedure based on Tudhope and Taylor (1997) that is a modification of the method of Rada et al (1989). It is based on the principle of measuring the weighted distance between a pair of classification terms by the shortest number of links that separate them in the semantic net of classification terms. The weighting is affected by an inverse hierarchical depth factor that is analogous to the hierarchical level in the hierarchical spatial distance measure. Thus the thematic distance between two terms a and b is given by:

$$\mathrm{TD}(\mathrm{a},\mathrm{b}) = \left(\frac{\mathrm{C}_{\mathrm{a},\mathrm{x}_1}}{\mathrm{L}_{\mathrm{x}_1}} + \frac{\mathrm{C}_{\mathrm{x}_1,\mathrm{x}_2}}{\mathrm{L}_{\mathrm{x}_2}} + \frac{\mathrm{C}_{\mathrm{x}_2,\mathrm{x}_\mathrm{n}}}{\mathrm{L}_{\mathrm{x}_\mathrm{n}}} + \cdots + \frac{\mathrm{C}_{\mathrm{x}_\mathrm{n},\mathrm{b}}}{\mathrm{L}_{\mathrm{b}}} \right) \tag{3}$$

where each ratio on the right hand side refers to a link in the shortest path between a and b. The values L_i represent the levels of the terms i, while the values $C_{j,k}$ represent weights attached to those links between the respective terms j and k. Each different type of relationship may be given a different weight. In our experiments BT and NT relationships are given equivalent weights, while the RT relationship is given a larger weight, resulting in greater computed distance.

6 Ranking results

6.1 Examples

We now provide examples of the application of the spatial similarity measures. An example hierarchy is illustrated in Figure 5 in which several hills are associated via *partOf* and *overlap* relationships with four administrative region places (Scottish Borders, West Lothian, Midlothian and City of Edinburgh). Note that the four administrative regions share a common parent in the region Scotland. Scotland is at hierarchical level 4 as it is part of the United Kingdom, that is part of Europe, that is part of the World. Thus the other two levels are numbered 5 and 6 respectively.

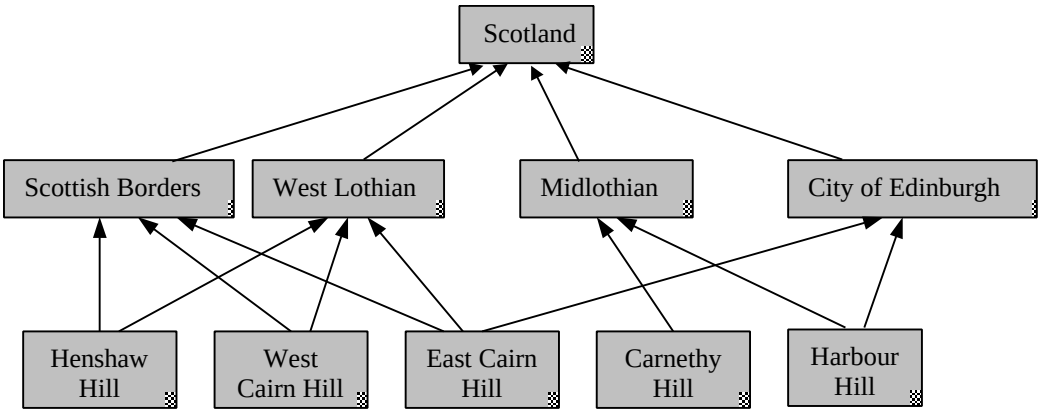

Fig. 5. An example of part of a place name poly-hierarchy

Examples of the application of the hierarchical distance measure are as follows:

1. HD (Henshaw Hill, West Cairn Hill) = 0
reflecting the fact that the two places both overlap the same two regions of Scottish Borders and West Lothian and no other regions.

2. HD (Henshaw Hill, East Cairn Hill) = $\frac{1}{level\ \ of\ \ City\ \ of\ \ Edinburgh}$

$= 1/5 = 0.2$

reflecting the fact that East Cairn Hill overlaps the City of Edinburgh, but Henshaw Hill does not.

3. HD (Henshaw Hill, Carnethy Hill) =

$$\left(\frac{1}{level\ of\ \ Scottish\ Borders} + \frac{1}{level\ of\ West\ \ Lothian}\right) + \frac{1}{level\ of\ \ Midlothian}$$

$= (1/5 + 1/5) + 1/5 = 0.6$

In this case the Scottish Borders and West Lothian are both distinctive super-parts of Henshaw Hill, while Midlothian is a distinctive super-part of Carnethy Hill.

4. HD (Henshaw Hill, Harbour Hill) = $(1/5 + 1/5) + (1/5 + 1/5) = 0.8$

This example shows that Harbour Hill is at a larger distance from Henshaw Hill due to the former's two distinctive super-parts of Midlothian and City of Edinburgh in addition to the two super-parts of Henshaw Hill.

To illustrate the application of asymmetry, the values of α and β may be set to 1 and 0.5 respectively. When the query term of Scotland is compared with the candidate term Henshaw Hill we obtain the following result:

5. $$\text{HD (Scotland, Henshaw Hill)} = 0 + 0.5\left(\frac{1}{level\ of\ Scottish\ Borders} + \frac{1}{level\ of\ Scotland}\right)$$
$$= 0.5\ (1/5 + 1/6) = 0.18$$

Thus Scotland has no distinctive super-part, while Scottish Borders and Scotland are distinctive super-parts of Henshaw Hill. Note that Scotland is not regarded as a super-part of itself.

6. $$\text{HD (Henshaw Hill, Scotland)} = 1\left(\frac{1}{Level\ of\ Scottish\ Borders} + \frac{1}{level\ of\ Scotland}\right) + 0$$
$$= (1/5 + 1/6) = 0.37$$

which indicates that Scotland is found to be more distant from Henshaw Hill than *vice versa*. For the purposes of finding things that are in Scotland or in Henshaw Hill this appears to be appropriate. It is important to note however for other types of query referring to these places this might not be an appropriate setting of the weights.

Finally we introduce a score that combines normalised values of the thematic distance measure and the spatial distance measures:

$$\text{Score} = 100 - (w_t\text{TD}_n + (w_s\,(w_e\text{ED}_n + w_h\text{HD}_n)\,) \qquad (4)$$

where each of the measurements refer to the distance between the query and the relevant candidate terms whether places or non-spatial objects. The subscript n refers to normalisation and w_t, w_s are weights for thematic and spatial measures respectively. These weights can all be set via the OASIS user interface. Figure 6 illustrates the results of ranking retrieved data for a query that requested "axes in Edinburgh". Here weights for w_t and w_s have been set to 0.4 and 0.6 respectively and the weights for w_e and w_h have been set to 0.6 and 0.4 respectively. Note that in the figure, place names are paired with their immediate parent. Thus "Edinburgh'Currie" refers to the town of Currie which is inside the City of Edinburgh, while "Edinburgh'Edinburgh" refers to the candidate place of Edinburgh itself. Due to the sparse distribution of certain types of artefact in our database, several simulated data objects have been inserted in to the database for demonstration purposes. These include "tomahawks".

6.2 Query expansion and weight specification

The search for matches with the query terms involves traversal of the place name poly-hierarchies and of the thematic, conceptual and poly-hierarchies belonging to the phenomenon of interest. In order to constrain expansion, threshold values for the TD, ED and HD can be set in the user interface. A further control on search is the number of links traversed from the query term. The user interface also includes provision to set weights for the thematic BT, NT and RT relationships as well as the weights for the combination of the ED and HD and for the combination of the spatial and thematic similarities.

As each node in the respective hierarchies is encountered, the distance measures are calculated and compared with the relevant threshold values. Search will continue to expand in the respective hierarchies until all thresholds have been met or there are no more paths to follow.

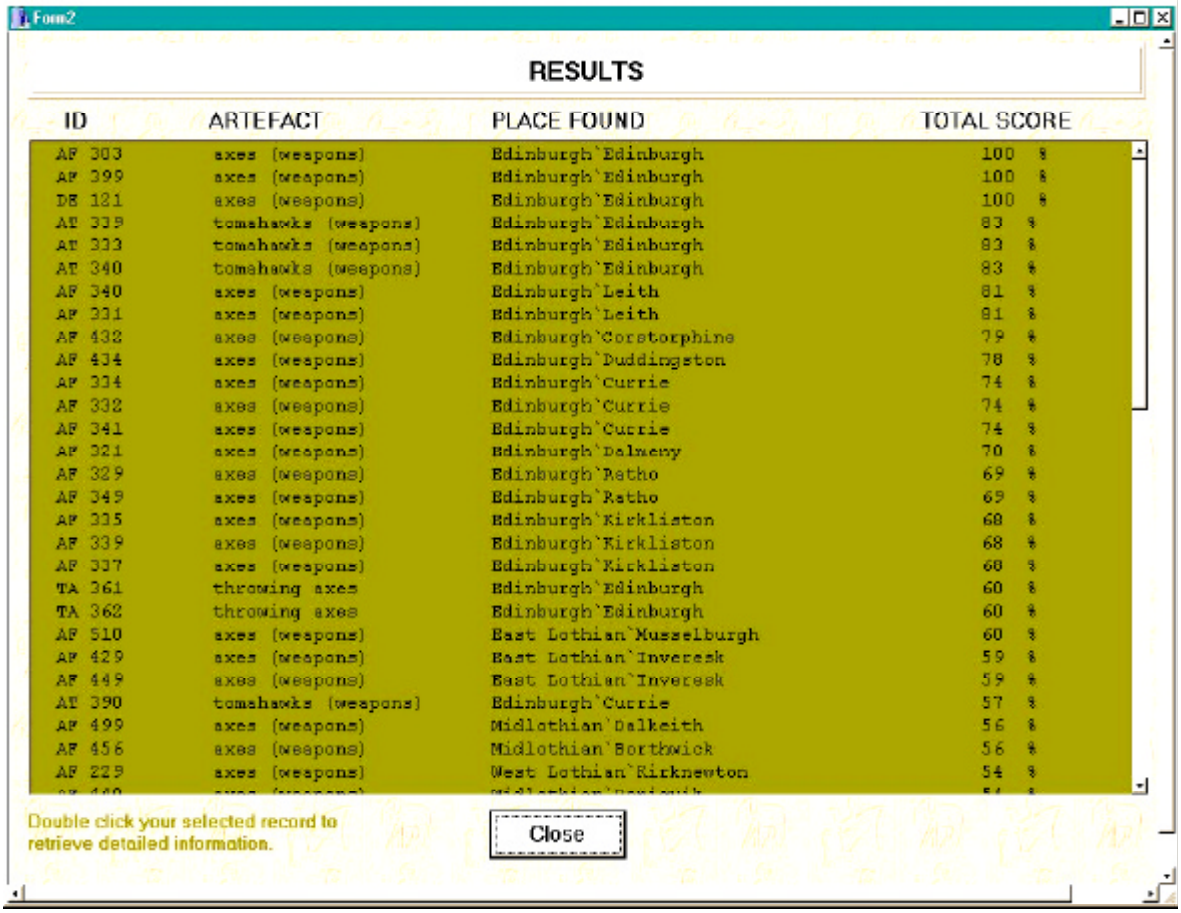

Fig. 6. Ranked results for a query on "axes in Edinburgh" using both spatial and thematic distance measures.

7 Conclusions and Discussion

In this paper we described an ontology of place that may be used to derive semantic distance measures for use in geographically-referenced information retrieval. The proposed ontology is characterised by a mix of qualitative and quantitative spatial data including topological relations and sparse coordinate data representing the spatial footprints of places. Places are classified according to their geographical categories and are linked to instances of non-geographical phenomena classified by conceptual hierarchies. Places are associated with other places via topological relationships. Terminology is classified into standard (preferred) and non-standard (or non-preferred) terms, and terms and relationships are qualified with dates. In using a parsimonious model of space that records very limited coordinate data, the approach is closely related to information retrieval based on gazetteers and geographical thesauri. The primary contribution lies in diversifying the types of information maintained and hence developing integrated semantic closeness measures that can be used for automatic ranking of query results.

In an implementation of the ontology, using data from existing thematic and geographical thesauri, from a cultural history source, and some simulated data, we have demonstrated how a combination of a hierarchical distance measure and an Euclidean distance measure can be used to rank retrieved archaeological site information in terms of geographical relevance. An overall ranking is obtained via a weighted combination of the spatial distance measures and a thematic measure. The hierarchical distance

measure is notable for distinguishing between places according to the extent to which they belong to or overlap with different geographical regional hierarchies.

While the techniques presented here appear to have considerable potential, there is no doubt that the distance measures could be refined and extended in various ways. The hierarchical distance measure presented here combines "proper part of" (inside) relations with overlap relations. It would be possible to weight overlap relations according to the degree of overlap as proposed by Beard and Sharma (1997). To do so would require measurement of that overlap with coordinate-based data. With the spatial model proposed here, that would not be possible directly from the centroid data. However, an approximation of overlap could be obtained using Voronoi methods that approximated regions by the union of the Voronoi cells of their contained places. An alternative would be to pre-compute overlap of all represented regions using more detailed map data.

The presented scheme gives equal weight to super-parts that are at the same hierarchical level. It may be that some geographical hierarchies are of more significance than others for a particular application, leading to the possibility of weighting members of individual hierarchies differently, or omitting them, according to context.

The hierarchical distance measure distinguishes between places that are in different parts of a poly-hierarchy, but it does not take account of whether adjacent places are connected or not. If two candidates are at equal distances from the query place, but one was regionally connected to the query place while the other was not, then the former might be regarded as closer. Clearly the meet relation could be employed to introduce a weighted connectivity term into the total spatial distance measure. The weighting of this term could be a function of the length of the common boundary. The Euclidean distance measure employed here is based solely on measurement between centroids, but it may be more appropriate to take account of distances between boundaries of regional places, or between regional boundaries and the centroid of a point-referenced place. Voronoi techniques again may help in these measurements.

This paper has focused on the use of place as a locator, and for purposes of similarity measurement the only thematic (non-spatial) information taken into account when comparing places is the categories of the regions to which places are referenced via part-of and overlap relationships. It is possible that the place class may be relevant when comparing places for purposes of location and it would certainly be expected to be relevant when searching for places that were similar to the query place. It would be quite simple to extend the existing measures to include a thematic similarity measure based on the place types. This could employ the existing TD measure presented here. Alternatively, assuming that places may be associated with multiple place types, it may appropriate to consider a method based on non-common superclasses, analogous to the hierarchical distance measure, or on the feature-based methods exemplified by Tversky's ratio model.

A further issue is whether weights in distance measures can be determined automatically using machine learning.

Acknowledgements

We would like to thank the J. Paul Getty Trust and Patricia Harpring in particular for provision of their TGN and AAT vocabularies; Diana Murray and the Royal Commission on the Ancient and Historical Monuments of Scotland for provision of their dataset; and Martin Doerr and Christos Georgis from the FORTH Institute of Computer Science for assistance with the SIS.

References

AAT (2000) Art & Architecture Thesaurus
http://www.getty.edu/research/tools/vocabulary/aat/

Agosti, M., F. Crivellari, G. Deambrosis and G. Gradenigo (1993). “An architecture and design approach for a geographic information retrieval system to support retrieval by content and browsing.” *Computers, Environment and Urban Systems* 17: 321-335.

Alani, H., C. B. Jones and D. S. Tudhope (2001). “Voronoi-based region approximation for geographical information retrieval with gazetteers.” *International Journal of Geographical Information Science*: accepted for publication.

Beard, K. and V. Sharma (1997). “Multidimensional ranking for data in digital spatial libraries.” *International Journal of Digital Libraries* 1: 153-160.

Couclelis, H. (1992) Location, place, region and space. *Geography's Inner Worlds*, R.F. Abler, M.G. Marcus and J.M. Olson (eds), 215-233. Rutgers University Press, New Jersey

Curry M.R. (1996) *The Work in the World - Geographical Practice and the Written Word.* University of Minnesota Press, Minneapolis.

Doerr, M. and Fundulaki, I. (1998) "SIS - TMS: A Thesaurus Management System for Distributed Digital Collections". In *Second European Conference on Research and Advanced Technology for Digital Libraries, ECDL'98* (Eds, Nikolaou, C. and Stephanidis, C.) Heraklion, Crete, Greece, 215-234.

Gould P. and R. White (1986) *Mental Maps.* Allen and Unwin, London.

Guarino, N. (1997). Some organizing principles for a top-level ontology. Padova, National Research Council, LADSEB-CNR Int. Rep. 02/97.

Guarino, N. (in press). Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration. In M. T. Pazienza (ed.) *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*. Springer Verlag: 139-170.

Guarino, N., C. Masolo and G. Vetere (1999). “ OntoSeek: Content-Based Access to the Web.” *IEEE Intelligent Systems* 14(3): 70-80.

Harpring, P. (1997). “Proper words in proper places: The Thesaurus of Geographic Names.” *MDA Information* 2(3): 5-12.

Hill, L. L., J. Frew and Q. Zheng (1999). “Geographic Names. The implementation of a gazetteer in a georeferenced digital library.” *Digital Library* 5(1): www.dlib.org/dlib/january99/hill/01hill.html.

Johnson, R.J. (1991) *A Question of Place: Exploring the Practice of Human Geography.* Blackwell.

Jones, C. B., C. Taylor, D. Tudhope and P.Beynon-Davies (1996). "Conceptual, spatial and temporal referencing of multimedia objects". *Advances in GIS Research II*. M. J. Kraak and M. Molenaar (eds). London, Taylor and Francis**:** 33-46.

Jordan T, M. Raubal, B. Gartrell and M.J. Egenhofer (1998) "An affordance-based model of place in GIS". *Proceedings 8th International Symposium on Spatial Data Handling,* T.K. Poiker and N. Chrisman (eds), International Geographical Union, 98-109.

Kim, Y. W. and Kim, J. H. (1990) "A Model of Knowledge Based Information Retrieval with Hierarchical Concept Graph". *Journal of Documentation,* 46(2)**,** 113-136.

Larson, R., R. (1995) Geographic Information Retrieval and Spatial Browsing. In *GIS and Libraries Patrons, Maps and Spatial Information,* Linda Smith and Myke Gluck (eds), Urbana-Champaign : University of Illinois, 1996. (p. 81-124).

Lee, J. H., M. H. Kim and Y. J. Lee (1993). “Information retrieval based on conceptual distance in IS-A hierarchies.” *Journal of Documentation* 49(2): 113-136.

Moss A., E. Jung and J. Petch (1998) "The construction of WWW-based gazetteers using thesaurus techniques". *Proceedings 8th International Symposium on Spatial Data, International* Geographical Union, 65-75.

Rada, R., H. Mili, E. Bicknell and M. Blettner (1989). “Development and application of a metric on semantic nets.” *IEEE Transactions on Systems, Man and Cybernetics* 19(1): 17-30.

Relph E. (1977) Place and Placelessness. Pion Limited.

Richardson, R., A. F. Smeaton and J. Murphy (1994). "Using WordNet for conceptual distance measurement". *Information Retrieval: New Systems and Current Research***:** 100-123.

Rodriguez, M. A., M. J. Egenhofer and R. D. Rugg (1999). "Assessing semantic similarities among geospatial feature class definitions". *Interop'99*. A. Vckovski, K. Brassel and H.-J. Schek (eds). Berlin, Springer. Lecture Notes in Computer Science 1580**:** 189-202.

Sintichakis, M. and P. Constantopoulos (1997). A method for monolingual thesauri merging. *20th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, 129-138.

Smith, B. (1995). "On drawing lines on a map". *Spatial Information Theory A Theoretical Basis for GIS*. A. U. Frank and W. Kuhn. Berlin, Springer. Lecture Notes in Computer Science 988**:** 475-484.

Smith, B. and D. M. Mark (1998). "Ontology and geographic kinds". *8th International Symposium on Spatial Data Handling SDH'98*, Vancouver, International Geographical Union.

Spanoudakis, G. and P. Constantopoulos (1994). *Measuring Similarity Between Software Artifacts*. 6th International Conference on Software Engineering & Knowledge Engineering (SEKE '94), Jurmala, Latvia, 387-394.

TGN (2000) Getty Thesaurus of Geographic Names. . http://www.getty.edu/research/tools/vocabulary/tgn/

Tuan Ti-Fu (1977) *Space and Place: the Perspective of Experience*. Edward Arnold.

Tudhope, D. and C. Taylor (1997). “Navigation via Similarity: Automatic Linking Based on Semantic Closeness.” *Information Processing and Management* 33(2): 233-242.

Tudhope, D., H. Alani, C. Jones (2001) "Augmenting thesaurus relationships: possibilities for retrieval". Journal of Digital Information Vol. 1(8): http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Tudhope/

Tversky, A. (1977). “Features of similarity.” *Psychological Review* 84(4): 327-352.

Walker, D., I. Newman, D. Medyckyj-Scott and C. Ruggles (1992). “A system for identifying datasets for GIS users.” *International Journal of Geographical Information Systems* 6(6): 511-527.