Geographical Terminology Servers - Closing the Semantic Divide

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

[1]Department of Computer Science
Cardiff University
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
The user interfaces of most geographical information systems are project specific, imposing upon the user a particular organisation's view of the information content. As geographical information becomes more widely available across multiple domains, the requirement arises to provide information retrieval facilities that recognise geographical terminology from multiple perspectives at multiple levels of abstraction. This paper elaborates upon these issues and describes techniques for making imprecise matches between user terminology for geographical places and concepts and the relevant stored terms. These techniques exploit an experimental ontology that derives semantic relations and classification terms, from existing thesauri of place names and cultural information concepts, and includes a parsimonious spatial data model combining qualitative relations with sparse quantitative geometry.

1. Introduction

Most aspects of human experience may be regarded as having a geographical dimension. Thus everything that we do usually takes place somewhere in the vicinity of the Earth's surface and communication between people frequently requires that we refer to particular places. A consequence of this is that many types of information either have, or else need to be given, some explicit geographical context. The last couple of decades have seen great advances in the development of technology referred to as geographical information systems (GIS). In practice these systems are typically concerned with handling digital maps in which location is recorded primarily by geographical (latitude and longitude) or map grid coordinates. For the most part, individual GIS are domain specific, often project-based, serving the needs of organisations that have traditionally relied upon map-based recording of information.

Undoubtedly GIS have made significant contributions to improving the information retrieval and analysis capabilities of these organisations. When viewed with regard to the need for public access to geographically referenced information, the contribution has been less significant. For those with access to the internet, one of the commonest methods of seeking information is to employ a

search engine within a web browser. When a geographical term, such as a place name, is typed in, it is usually treated the same as any other term or phrase, with the result that documents are retrieved if there is an exact match with the whole or some part of the query phrase. The consequence is that we will often fail to find information that we are interested in, because it has been given geographical terminology different from that of the query, even though it refers to a similar location. This may happen due to the hierarchical nature of geographic space, so that a particular place may have sub-parts or super-parts referred to by different names. In fact there are multiple overlapping hierarchies of geographical place, varying according to political, topographic and cultural perspectives. Equally there may be places that are close to the specified place and hence potentially of interest, as well as place names that may differ due to historical change or to language.

Ideally, when we use a place name to refer to some location, we should be able to retrieve information about places that are equivalent or nearby and rank the results according to their relevance to the query. When we specify a term or phrase referring to the thing of interest we should be able to find things that have equivalent or similar descriptors. At present web search engines are weak in handling all types of terminology where there is a need for imprecise matching between query and target. These shortcomings are widely recognised and have led for example to the development of mark up languages that tag data with terms that clarify meaning, exemplified by XML and its applications, as well as improved levels of intelligence in the search engine itself (see for example Guarino 1999). In this paper we are concerned with improving the level of intelligence of information retrieval tools with regard to geographical terminology.

In conventional GIS the most common way of accessing information by location is to point to somewhere on a map or to specify coordinates explicitly. Frequently spatial objects may also have their name as an attribute that can be used for search, but the associated query procedure is normally based on precise match. Some GIS include a simple gazetteer that allows the user to specify a place name that is then used automatically to specify a map coordinate for purposes of coordinate-based search. The importance of place names as a way of allowing users to search for computer-based information was recognised at the Getty Institute in the mid 1990s and led to the development of the Thesaurus of Geographic Names (TGN) (Harpring 1997). The TGN is hierarchically structured and hence allows for the possibility of expanding a place name query term by finding its contained places and the parent places. It also maintains different versions of the same name, along with their associated dates, and the geographical coordinates of a representative point location, i.e. a *centroid*. Places in the TGN are either geopolitical or topographic and are associated with place types using terms taken from the Art and Architecture Thesaurus (AAT). In parallel with the TGN, various other gazetteers and place name lists have been produced on regional and international levels. In association with the Alexandria Digital Library a gazetteer metadata standard has been developed in which all places are

associated with a coordinate-based spatial footprint (such as a minimum bounding rectangle), while other relationships such as of administrative hierarchy may be recorded but are not mandatory (Hill 1999).

The introduction of gazetteers and the TGN has gone some way to addressing the requirement for access to information by place name, but there has been very little research to investigate how they may be exploited automatically and indeed what sort of geographical information they should record to maximise their utility in information retrieval. The problem of providing intelligent support for geographically referenced query on the web, as well as within specialised GIS, is a challenging one (Walker et al 1992; Larson 1995; Jones et al 1996; Beard and Sharma 1997; Moss et al 1998). In principle, for web search it is desirable to be able to recognise place names at all levels of generalisation for anywhere on Earth, find equivalent co-located places and nearby places and rank the results. Many questions arise with regard to the information to be stored: what types of spatial relationship between stored places; how much coordinate-based data; how can imprecise regions be represented; what types of information characterise places from cultural and environmental perspectives? Decisions about what should be stored affect the capability of relevance ranking procedures.

In this paper we address some of these issues and we describe an experimental cultural information system that integrates geographical and thematic thesauri in the context of a semantic modelling system. In section 2 we elaborate on the subject of defining place for purposes of information retrieval, before summarising in section 3 the potential contribution of previous work on thesauri for encoding semantic relations between terminology. The metadata schema of OASIS are presented in section 4 and the subject of semantic closeness measures is considered in section 5 in which we describe techniques for ranking spatial and non-spatial information employed in the experimental system. Concluding remarks and some issues for further research are presented in section 6.

2. Encoding Places for Information Retrieval

In general when place names are specified in a query to retrieve information, the place may serve to specify location, either by itself, or as part of a spatial expression, or it might be used for comparative purposes to find similar types of place. Here we focus on the former role of place as locator. Thus we assume that a query searches for "something" geographically located "somewhere", where the "somewhere" may include a place name. When modelling place for information retrieval, we need to identify characteristics that will assist in expanding the set of query terms to include co-located places (including places with different names but similar spatial extent) and neighbouring or nearby places, and in ranking the results with respect to the query terms. It may be noted that the final ranking should take account both of the phenomena of interest and the geographic location. We consider modelling non-locational concepts in a subsequent section.

Much has been written about the nature of geographic place (e.g. Relph 1977; Yuan 1977; Gould and White 1986; Johnson 1991; Curry 1996; Jordan et al 1998). It is one of several "basic concepts" in geography (Couclelis, 1992) alongside location, region and space, of which space may be regarded as the most fundamental. Thus space may be considered the substrate within which locations, regions and places and defined. We are concerned particularly with place here because it is associated with names for specific parts of space and hence allows us to refer to location in natural language, as opposed to the typically more formalised expressions of location in terms of coordinates and spatial objects. An issue often raised is that of the human characterisation of place, the fact that places materialise in response to events and experiences and hence are essentially a human construction. Examples of distinguishing properties of a place are the name, the categories that reflect its physical or social features, familiar landmarks that become symbolic, activities, personal experiences and opportunities.

In building an ontology of place to support public information retrieval it may be regarded as a priority to store information that is generic in the sense of not being specific to a few individuals. In this regard experiential aspects of place that are personal in nature may be of low priority. From a pragmatic standpoint, assuming that we may need globally extensive coverage, it may be important to select characteristics that are relatively easily obtainable. This may result in a biased view of place that reflects to some extent that of Johnson (1991), who adopts for place some of the concepts of Paasi concerned with the institutionalization of regions. The reason for this bias is that existing sources of lists of place names such as those of national mapping agencies typically confine their non-geometric attributes to those of administrative authority (and hence typically a containing region) and perhaps the size of the population. The distinction between regions and places may be of significance for purposes of information retrieval in that regions relate to a partition, often hierarchical, of some parts of space. In doing so they provide a representation of that can be exploited for query expansion. In so far as regions label parts of the space with commonly used names, we regard them as a type of place. The only significant distinction for our purposes is the fact that regions form systematic partitions and hierarchies while other places could be isolated, while still referenced in some way to regions.

2.1 Identity

As we are concerned here with specific instances of places, it is essential to maintain the name or names that are typically used to refer to the place. Names may be formal administrative terms, that typically will correspond to a precise boundary such as that of a city or parish, and informal terms that reflect common means of referring to places that may be fairly precise in extent, such as a building, or be imprecise such as a mountain range. Place names change over time and in doing so may come to differ somewhat in exactly what territory or phenomenon they refer to. Certainly knowledge of the temporal extent of a place

name may be important when searching for information that itself may be temporally specific. Names may also differ simply due to differences in language. A single name is sometimes used to refer to different places which means that a unique identifier must be found or created. The need for explicit unique identification may vary according to the nature of associated data that are stored. If places are always linked to parent regions, or if a geographical coordinate is stored, then the presence of these attributes may serve to obtain uniqueness.

2.2 Spatial Data

Query expansion with respect to location can be supported using coordinate-based methods of conventional GIS in which a search is expanded with increasing Euclidean distance from the query object. Standard GIS methods assume the presence of point, line and polygon spatial objects defined by coordinates. If a model of place were to be maintained for the entire globe as might be required for general web browser querying, then the amount of coordinate data required to represent both the smallest and the largest places would be massive, and on first impression impracticable. There is a motivation therefore for a parsimonious spatial model that encompasses much of geographic space but in a way that minimises the amount of stored information. An alternative to dependency upon coordinates is to encode qualitative relations, such as those of containment/inclusion, overlap and contiguity, all of which can be derived from vector map data, which need not subsequently be stored as part of the model. This would then facilitate query expansion to contained and containing places as well as to neighbouring places. Assuming the presence of multiple (overlapping) regional hierarchies, then if a place was registered in the ontology, its containment and overlap relations to all hierarchies could be determined and recorded. An advantage of encoding and exploiting qualitative spatial relations is that it enables some historical places to be recorded for which documentary evidence provides regional containment information in the absence of a cartographic representation. In section 5 we discuss the issue of measuring similarity of place using hierarchical relationships.

It should be noted that contiguity relationships can only easily be derived from maps based on polygonal partitions of space (as in administrative regions) and hence will lead directly to query expansion only within the map regions. For purposes of determining proximal and directional relations between isolated places, it may still be very useful to employ some coordinate data. At the least this could be a single representative point or centroid, as in a simple gazetteer. Storage of centroids facilitates ranking of places with respect to Euclidean distance and the determination of nearest neighbouring places using Voronoi diagram or Delaunay triangulation methods (Aurenhammer 1991). Distances calculated between places based on centroids will of course be approximations as they take no account of the location of the boundary. It is however possible to estimate the locations of boundaries of places given data on contained and neighbouring places for which centroids are available (Alani et al, in press).

2.3 Accessibility

When considering "nearness" of place a factor that might be considered is accessibility and therefore the types of information required to measure it. Accessibility is a function of available methods of transport and the properties of the transport routes. In general it appears to be most relevant if the subject is considering visiting the place or its neighbours. Clearly this may be relevant to some types of query and not to others. Support for accurate measurement of accessibility would require a network data structure with relevant impedance or cost factors attached to all links. If accessibility is a weak requirement then it may be that the approximate Euclidean distances derivable from limited coordinate data may serve as a surrogate measure.

2.4 Non-Spatial Concepts of Place

If we assume that whenever a person specifies a place name they deposit some conceptual baggage that they associate with the name, then in modelling place it is reasonable to suppose that we should record attributes that may reflect the baggage. For example, if someone uses a named mountain range as a locator, then in expanding the search it is possible that they might regard hilly places bordering the mountains as more relevant than equally close neighbouring cities. If a named city were used as locator, it might be that neighbouring settlements that were in the same country as the named city were more relevant than those in another country that were equally near in Euclidean space.

In these examples, potentially relevant non-spatial attributes are topographic land-cover categories, and administrative (geopolitical) regions. If these types of attribute form the basis of regional hierarchies to which a place was referenced (by being inside or overlapping with an individual region) or of which it was a member, then the use of these hierarchies for query expansion would automatically result in the inheritance of their properties by the places that are related to them. In the absence of an association with a regional hierarchy then it would appear important to attach classification terms to places in addition to the regional hierarchy relationships of containment, overlap and adjacency, and to be able to exploit them in search procedures.

Several authors have referred to the characteristics of place that offer opportunities and indeed constraints on the activities that may be performed there (Jordan et al 1998). It is possible to envisage storing classification terms for place that reflect these opportunities or *affordances* directly. Alternatively and more economically it may be that certain types (such as port, mountain, river) may imply opportunities and actions and that for purposes of information retrieval the use of the classification terms to measure similarity with respect to class may serve as a rough surrogate for affordance.

3. Modelling Conceptual Terminology with Thesauri

There is long history of the use of thesauri in modelling terminology to assist in indexing and retrieval of information within particular domains. They are relevant to geographical information retrieval in that it may be possible to associate a specialised model of place, based on ideas referred to above, with existing thesauri representing non-spatial concepts associated with geographical locations. In this section we review briefly the principle types of relationships encoded within thesauri, in order to provide some background to our exploitation of thesaural relationships for purposes of non-spatial concept matching which is employed in association with place matching.

A major part of a thesaurus is usually one or more domain-specific classifications, derived either from a single source or resulting from a process of merging multiple classifications within the same domain. Individual classification hierarchies may be grouped together within facets. One of the earliest suggested sets of facets is that of Ranganathan who proposed the five-fold division into Personality, Matter, Energy, Space and Time (PMEST) in the context of the Colon Classification system. The Art and Architecture Thesaurus (AAT) includes facets for physical attributes, styles and periods, agents, activities, materials and objects. The objects facet for example includes a Settlements and Landscapes Hierarchy which includes a variety of terms that express different types of place.

Classification structures are encoded in thesauri by means of generalisation and specialisation relations, referred to as broader term (BT) and narrower term (NT). In order to denote their application to hierarchical encoding of generic relationships they may be described more specifically as BTG (Broader Term Generic) and NTG (Narrower Term Generic). Hierarchical relationships are also typically encoded in thesauri to describe part-whole relationships. Again broader and narrower relations are distinguished, in this case as BTP (Broader Term Partitive) and NTP (Narrower Term Partitive). Aitchison and Gilchrist (1987) distinguish four categories of part-whole relationships. These are a) systems and organs of the body (e.g. ear BTP internal ear); b) geographical locations (USA BTP California); disciplines or fields of study (e.g. archaeology BTP marine archaeology); and hierarchical social structures (e.g. methodist church organisation BTP methodist district).

A third type of BT relationship is that of the instance relationship between an object and its class. The fourth type is that of polyhierarchical relationships, whereby some terms may be related to more than one parent class.

In acknowledging that users may refer to a very similar concept by means of different words, one term for a concept is usually designated the preferred term for purposes of encoding the term relationships. This leads to the converse equivalence relations of preferred-term and non-preferred-term that are referred to

as the UF (or USE_FOR) and USE relationships. For example, if of the two terms 'lake' and 'mere', the former was the preferred, then the following relationships could be encoded in a thesaurus: *lake UF mere*; *mere USE lake.* This type of relationship is applicable to place name terminology as well as other concepts. Thus it allows multiple names for the same place to be referred to a single standard name. In the event of generating a unique name for a place that shares its name with other places, the unique name can be associated with the standard name via a USE relationship.

Clearly many terms are related to each other by relations other than the main hierarchical ones, while not being synonyms. This has given rise to the use of the associative relationships referred to as the related-term (RT) relationship. An important function of RT relationships is to link terms that may occur in separate facets but may be logically associated. For example the association *masons* RT *bricklaying* links terms in the Agents and Activities facets respectively of the AAT.

4. Place and Concepts in OASIS

Some of the characteristics of place and concept described here have been implemented in OASIS (Ontologically Augmented Spatial Information System) to provide a basis for experimenting with geographical retrieval techniques. OASIS has been built using the Semantic Index System (Doerr et al 1998) which is an object-oriented hypermedia system that supports a number of semantic modelling constructs, a graphical user interface and API functions for data access. Schema creation can be performed with the TELOS language (Mylopoulos et al 1990) or via data entry forms. Information can be classified at several levels including Token, Simple Class, and four levels of Meta Class. Both classes and objects are treated as objects and can have names, attributes and relationships to other objects.

The application area for which OASIS has been developed is that of cultural heritage and there is support therefore for the maintenance of archaeological artifacts that are linked to place by relationships of *found_at* and *made_at*. A Place type has been defined in OASIS as a subtype of Geographical Concept and its properties are illustrated in Figure 1. Places can be classified with one or more current and historical place types that, in our implementation, are mostly derived from the Art and Architecture Thesaurus (AAT). Of particular importance are place types that belong to regional hierarchies that can be expanded for purposes of information retrieval. The name of a place is associated with it via Standard Name and Alternative Name relations that include the attributes of *variant spelling*, *date* and *language.* Because of the possibility of duplication of names, places are given unique names that are referenced to their conventional name via the Standard Name relationship.

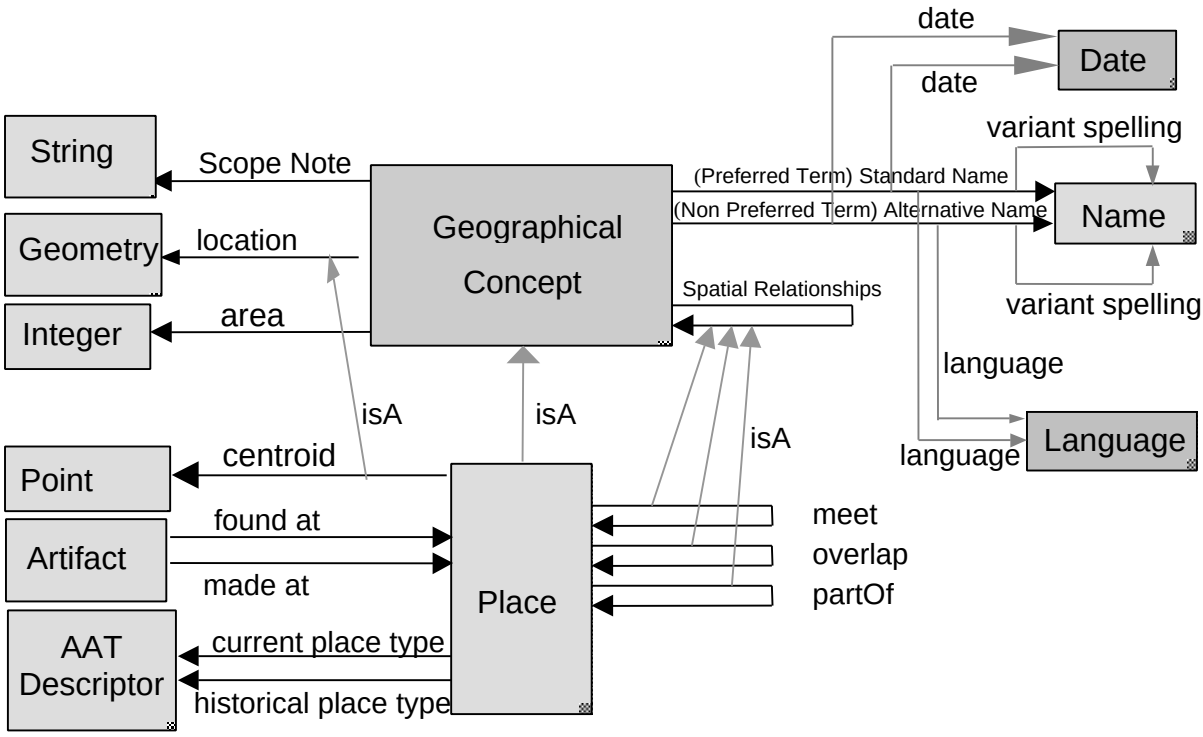

Figure 1. The schema for place in OASIS

The location of a place is represented quantitatively by a centroid (single point) defined by latitude and longitude values. It is treated as a specialisation of the location. The spatial representation of place includes the topological relations of *meets, overlaps* and *part-of* that are treated as specialisations of *spatial relationship.* These relationships are intended to allow an isolated place to be linked, by *part-of* or *overlap* to a regional hierarchy, as well as serving to encode the structure of the hierarchies in terms of *part-of* and *meets* relations.

5. Semantic Distance Measures

5.1 Vector Space Methods

The are various methods for measuring the similarity between terms when matching a query expression with a target object. With vector space methods indexed documents are allocated coordinates in a multidimensional (term) space determined by the occurrence of terms in the document (Salton 1989). A distance is then measured in vector space between a candidate document and the query expression, that is also located in the vector space. The approach is commonly applied to documents that may contain many terms, where the frequency of occurrence of individual terms may be taken into account in determining the location in vector space. The approach suffers from the disadvantage that query and target will only ever be regarded as similar if they share terms (at least) that are identical (either in full or when stemmed). It takes no account of the semantic relationships between terms that may be similar or related in meaning. It is based

on the assumption that a document may have multiple occurrences of a term and as such it is not directly applicable to the comparison of concepts that may be defined by a set of attributes and relationships.

5.2 Feature-Based Methods

In the feature matching methods introduced by Tversky (1977) an object is associated with a set of features which are compared with regard to their commonality and their difference. Similarity between two objects is measured as some function of the intersection of the features they have in common, the features unique to one of the objects and the features unique to the other object. By attaching different weights respectively to the features that belong to one object but not the other, a matching function can express the asymmetry commonly observed in relationships between objects where one is either more important or more prototypical than the other. Thus for example a house may be regarded as more similar to a building than a building is to a house, as a house is a subclass of building. Alternatively a settlement may be more similar to the state to which it belongs than is the parent state to the settlement.

5.3 Thesaural Methods

Tversky's feature matching methods been shown to have potential for measuring similarity of spatial entities (Rodriguez et al 1999), but they do have some limitations. One of these is that the results of matching operations will be skewed if one object has a different number of features than the other. The approach is intended for comparison of objects for which there are sets of descriptive features and consequently it cannot be applied directly to comparison of objects such as classification terms unless they are accompanied by a set of features. The approach also breaks down if there are differences in the terminology used to describe similar or equivalent features. Differences in terminology are widespread and may arise for example due to multiple organisations developing their own classification systems to refer to the same real-world domain. It is because of such differences in terminology that thesauri are widely used for indexing purposes. They provide a means for standardisation of terminology, for automatically identifying matches between equivalent terms via USE/EF relations, and for identifying terms that are similar, but not equivalent, in meaning by traversing the hierarchical and associative relations.

The use of thesauri or similar semantic nets has led to the development of various semantic distance measures based on the traversal of the semantic relationships. A simple approach is to base the semantic distance between two terms within a thesaurus on the shortest path between them (Rada, Mili, Bickell, Blettner 1989; Lee et al 1993). In a classification hierarchy this is the smallest number of is-a links between the two terms. A variation on the method is to attach different weights to links according to their type or their depth in the hierarchy (Kim and Kim 1990; Richardson et al 1994; Tudhope and Taylor 1997). In OASIS this approach, with weighted links, has been applied to the determination of similarity between AAT terminology that we have used to define non-spatial concepts. We

use the following formula to determine the thematic distance TD between two terms *a* and *b*:

$$TD(a,b) = \left(\frac{C_{a,x_1}}{L_{x_1}} + \frac{C_{x_1,x_2}}{L_{x_2}} + \frac{C_{x_2,x_n}}{L_{x_n}} + \cdots + \frac{C_{x_n,b}}{L_b} \right)$$

It is based on a summation of the weighted links in the shortest path from a to b. $C_{j,k}$ is the weight of the relationship between intermediate terms j and k in the shortest path between a and b, and is related to the thesaural type of the relationship. L_i is, by default, the hierarchical level of the term i, hence resulting in smaller distances between terms lower down a hierarchy. An example of the application of the method is given in section 5.7.

5.4 Non-Common Super-Classes

An alternative approach to measuring similarity of classification terms, in the context of a thesaurus or some other semantic net that includes hierarchical relationships, is one based on the non-common super-classes of pairs of terms (Spanoudakis and Constantopoulos 1994). The non-common super-classes of two objects *a* and *b* consist of parent classification terms that belong to *a* but not to *b* and those that belong to *b* but not to *a*. These terms may be regarded as analogous to the distinctive features of Tversky's methods. While the feature-based methods include an explicit measure of the common features, this is implicit in the non-common super-classes method, since the semantic net encodes relations of class generalisation or of part-whole directly, so that by definition if a pair of terms has no non-common super-classes then they must be closely related within the semantic space of the ontology. A further difference from the feature-based methods is the use of level-specific values whereby differences between terms decrease with increasing depth in the hierarchy, just as in the shortest distance methods referred to above.

5.5 An Hierarchical Spatial Distance Measure

In our treatment of place we regard the non-common super-classes method as applicable to the measurement of similarity between places with regard to the regional hierarchies to which they belong, via *part-of* or *overlap* relations. It is considered appropriate as it leads to measures of similarity that reflect differences in inherited properties of place as determined by the multiple hierarchies to which a particular place may belong. A limitation of the non-common super-classes method compared to the feature-based method is that it cannot express asymmetry of similarity. In order to address this shortcoming we propose adapting the method by including separate weights α, β for the distinctive super-classes of the two terms respectively. We also introduce a further weighting term γ to provide flexibility with regard to inclusion of the query and candidate terms in the measurement formula. The hierarchical distance HD of query place *a* from candidate place *b* is

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

where L_x and L_y are the hierarchical levels of the distinctive super-parts of *a* and *b* respectively, while L_z are the hierarchical levels of *a* and *b*. *a.PartOf* and *b.PartOf* refer to all super-parts of *a* and *b* respectively, i.e. at all higher hierarchical levels. If *a* and *b* are to be included in the measurement then γ takes on a non-zero value, otherwise it is zero. If either *a* is a sub-part of *b* or *b* is a sub-part of *a* (separated by one or more hierarchical levels) then α is set larger than β. Otherwise α and β are equal. Thus in a measurement of the distance between *a* and *b,* if *a* is the super-part it will have no non-common super-parts and hence the distance will be biased by the smaller weight. Conversely if *a* is a sub-part of the candidate term, its non-common parents (that include *b*) will be biased by the larger weight, resulting in a greater distance value.

It is envisaged that γ should be set non-zero when both *a* and *b* are members of the regional hierarchies, as opposed to one of then simply being referenced to a member of a hierarchy. Thus two sub-regions with a common parent will be separated by a finite distance, reflecting the fact that they are not the same region. However if two non-regional places belong to the same parent region then, purely with regard to the regional hierarchy, there is no difference between them. Clearly they will have a difference in Euclidean space and they may have a difference with regard to their individual place classes.

5.6 Coordinate-based distance

As indicated in the description of the OASIS schema for place, we attach a centroid to each place consisting of two coordinates. In order to support global applications, we encode the coordinates as latitude and longitude. Earth surface distances are then calculated along great circles. We refer to this measure as Euclidean Distance. In using centroids, the resulting distances may be regarded as somewhat error prone, particularly in the case of places with considerable area extent. As is explained in Alani et al, centroids can be used to approximate the boundaries of regions provided there is knowledge of both contained and neighbouring external places that are associated with centroids. The approximated boundaries may then be used to determine distance between boundaries or between points and boundaries.

5.7 Examples

In this section we illustrate the use of thematic distance and the hierarchical distance measure and show how they can be combined with Euclidean distance to produce an integrated ranking measure.

5.7.1 Thematic distance

As explained in the previous section, we have applied the thematic distance (TD) measure to comparison of concepts that belong to the AAT. The costs of traversal

of the different types of relationships have been set to BT 3, NT 3 and RT 4. When traversing RT relationships the level is taken as that of the originating term rather than the destination term of a relationship.

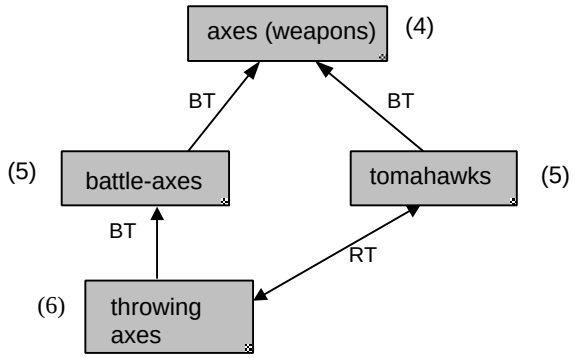

Figure 2

Referring the example in Figure 2, there are two possible paths between axes(weapons), at level 4, and throwing axes, at level 6. Thus:

$$\mathrm{TD}(\text{axes (weapons)}, \text{throwing axes}) = \frac{C_{NT}}{\text{level of battle axes}} + \frac{C_{NT}}{\text{level of throwing axes}} = \frac{3}{5} + \frac{3}{6} = 1.1$$

Note that NT relationships are simply the converse of BT relationships.

The second path produces the distance:

$$\mathrm{TD}(\text{axes (weapons)}, \text{throwing axes}) = \frac{C_{NT}}{\text{level of tomahawks}} + \frac{C_{RT}}{\text{level of throwing axes}} = \frac{3}{5} + \frac{4}{5} = 1.4$$

Since it is the first path that has the lowest cost the value of TD in this case is 1.1.

5.7.2 Hierarchical distance measure

We illustrate the use of the hierarchical distance measure with regard to an example scenario in Figure 3, in which several places of type hill are associated with members of an administrative regional hierarchy. In order to illustrate the application of a polyhierarchy, the association between hills and administrative regions represents both part-of and overlap relationships. The regional hierarchy is built entirely from part-of relationships.

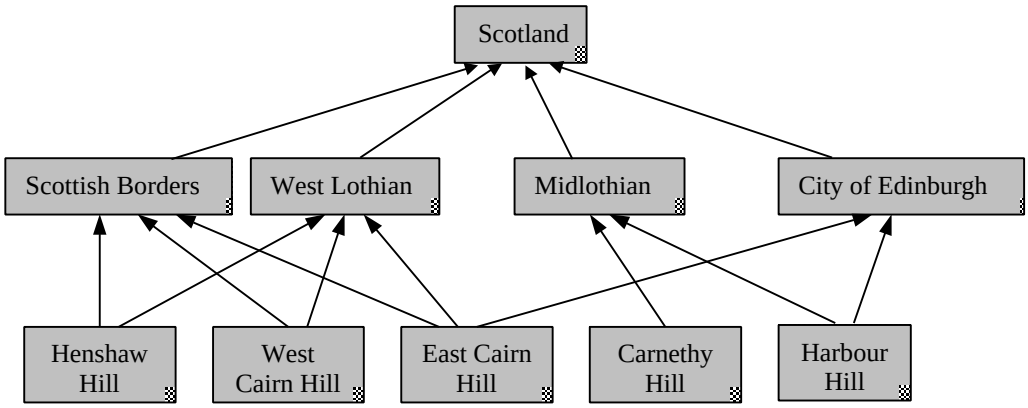

Figure 3. Example of hill places referenced by part-of and overlap relationships to an administrative hierarchy.

In this scenario Scotland is placed at hierarchical level 4 (Scotland *part of* United Kingdom *part of* Europe *part of* World), its sub-regions are at level 5, and the hills are therefore at level 6. In the following examples of distances between hills the weights α and β have been set equal to 1, while γ has been set to zero, giving the following results:

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) =

$$\left(\frac{1}{Level\ of\ S\cot tish\ Borders} + \frac{1}{level\ of\ WestLothian} \right) + \frac{1}{level\ of\ Midlothian}$$

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

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

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

5. HD (Henshaw Hill, Scotland) =

$$1\left(\frac{1}{level\ of\ Scottish\ Borders}+\frac{1}{level\ of\ West\ Lothian}+\frac{1}{level\ of\ Scotland}\right)+0$$

$$= (1/5 + 1/5 + 1/6) = 0.57$$

whereas with the comparison of Henshaw Hill with the query term of Scotland we obtain:

6. HD (Scotland, Henshaw Hill) =

$$0 + 0.5\left(\frac{1}{level\ of\ Scottish\ Borders}+\frac{1}{level\ of\ West\ Lothian}+\frac{1}{level\ of\ Scotland}\right)$$

$$= 0.5\ (1/5 + 1/5 + 1/6) = 0.28$$

which indicates that Henshaw Hill is nearer to its containing place than *vice versa*.

5.7.3 Combining measures

Given the existence of several distance measures relating both to non-spatial concepts and to qualitative and quantitative ("Euclidean") space, some means is required to combine such measures to provide a single overall ranking. In experiments with OASIS, the Euclidean and hierarchical distance measures were combined by normalising the individual measures before applying weights to the two components to produce a total spatial distance (TSD) measure defined as

$$TSD = (w_e ED_n + w_h HD_n)$$

where ED_n and HD_n are the normalised measures and w_e and w_h are the respective weights that sum to one. The TSD may then be combined with a normalised thematic distance as follows:

$$Score = 100 - (w_t TD_n + w_s\, TSD_n)$$

to produce a value between 0 and 100 where w_t and w_s are the weights for theme and space that also sum to one. Figure 4 illustrates an example of applying the score to rank the results of a query for "axes in Edinburgh", where axes has been specified as belonging to the weapons hierarchy. In this case the weights were set as follows:

$$Score = 100- (0.4 * TDn + 0.6 * (0.6*EDn + 0.4*HDn))*100$$

In the experiment, asymmetry was not taken into account in the hierarchical distance measure. Due to a paucity of real data in some geographic regions, some imaginary data items were added, referring for example to tomahawks. Taking the example of an occurrence of tomahawks (weapons) in the region of Currie (a part of Edinburgh in the administrative regional hierarchy), the normalised thematic distance between tomahawks and axes (weapons) was 0.428, while the normalised

Euclidean distance and hierarchical distances values of Currie from Edinburgh were 0.321 and 0.615 respectively. This results in a calculation of the score of:

Score = 100-(0.4*0.428 + 0.6 * (0.6 * 0.321 + 0.4 * 0.615))*100 = 57%

In the example it is apparent that places of Edinburgh and the contained places of Edinburgh are ranked before neighbouring places outside Edinburgh such as East Lothian. The ranking has resulted in some cases in the regional hierarchy modifying ranking that would be produced with Euclidean distance alone. Note also for example that because throwing axes are semantically more distant from axes (weapons) than are tomahawks, occurrences of the former that were found in Edinburgh are relegated to a lower score than tomahawks and exact matches of axes (weapons) that are referenced directly to Edinburgh or to the constituent parts of Edinburgh.

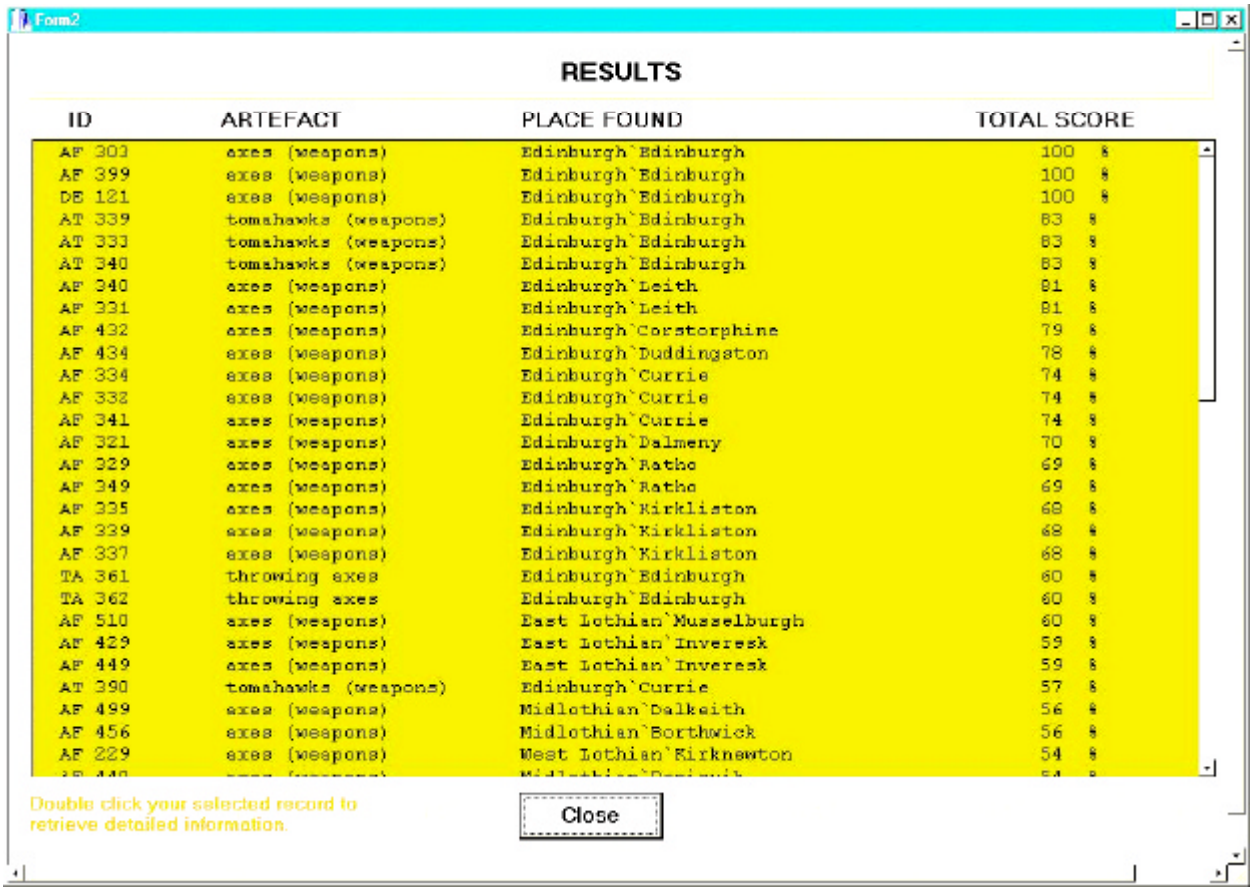

Figure 4. Example of ranking the results of a query for "axes(weapons) in Edinburgh"

6. Concluding Remarks

This paper has addressed the problem of developing facilities for geographical information retrieval in which the user may employ place names and concept terms that may not match precisely with the terms used to describe information of interest. A model of place has been proposed in combination with semantic closeness measures that can be used to rank the relevance of retrieved information with regard to the user's query terms. The model of place adopts a parsimonious approach to storage of spatial data with a view to providing potentially global coverage of geographic place names. The place names are associated with alternative versions of their name and one or more place type categories. Instances of place are linked to other places via qualitative spatial relations and are linked to geographical coordinate space with a single centroid. A hierarchical spatial distance measure is introduced that determines the distance between two places in terms of the number of non-common, parent places to which they belong, as determined by relationships of containment and overlap. The measure is combined with Euclidean distance to create an integrated spatial distance measure. A semantic distance measure based on weighted shortest paths within a thesaurus of classification terms is combined with the spatial measures to obtain an overall ranking of the results of queries that specify a thematic query term in combination with a place name.

The techniques presented are intended to make some progress towards handling natural language terms in geographical queries. Some preliminary user experiments, not reported here, have given support to the validity of the methods presented. This paper has focused on the use of place name as a locator and the inherent thematic or non-spatial aspects of place have only played a significant role with regard to measurement of semantic distance on the basis of non-common parent places. There is clearly scope to employ an explicit thematic distance measure that uses the place type terms to provide more sensitive distinctions between place, with regard to cultural, socio-economic and historic perspectives, than that provided by the parent places. This would be of particular importance if the purpose of a query were to find places similar to a specified query place, as opposed to finding some phenomena that are located at a specified place. There is also scope for experimenting with a wider variety of spatial closeness measures. It would be possible, for example, to employ qualitative measures of distance based on contiguity of neighbouring places (Jones et al 1996) and it would also be possible to weight distance measures according to degrees of overlap between places following the approach of Beard and Sharma (1997).

The methods presented here have been motivated by problems of information retrieval, but they have a wider application. In particular the semantic closeness measures may have potential for assisting in solving problems of geographical data integration in which data from different sources may employ different classification terminology and different place names to refer to similar locations. In these contexts the similarity measures could be used to help identify the equivalence of multiple representations of the same real-world phenomena.

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

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.

Aitchison, J. and Gilchrist, A. (1987) *Thesaurus Construction: a practical manual,* Aslib. London.

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

Aurenhammer, F. (1991) Voronoi Diagrams - A Survey of a Fundamental Geometric Data Structure. *ACM Computing Surveys,* **23**(3)**,** 345-405.

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), Rutgers University Press, New Jersey, 215-233.

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., 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.

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 Handling*, International Geographical Union, 65-75.

Mylopoulos, J., Borgida, A., Jarke, M. and Koubarakis, M. (1990) Telos: A Language for Representing Knowledge About Information Systems. *ACM Transactions on Information Systems,* **8**(4)**,** 325-362.

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.

Salton, G. (1989) *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer,* Addison-Wesley.

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.

Spanoudakis, G. and P. Constantopoulos (1994). "Measuring Similarity Between Software Artifacts". *6th International Conference on Software*

Engineering & Knowledge Engineering (SEKE'94), Jurmala, Latvia. pp. 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.