Associative and Spatial Relationships in Thesaurus-based Retrieval

Harith Alani, Christopher Jones, Douglas Tudhope
School of Computing, University of Glamorgan, Wales, CF37 1DL, UK.

Abstract
The OASIS (Ontologically Augmented Spatial Information System) project explores terminology systems for thematic and spatial access in digital library applications. A prototype implementation uses data from the Royal Commission on the Ancient and Historical Monuments of Scotland, together with the Getty AAT and TGN thesauri. This paper describes its integrated spatial and thematic schema and discusses novel approaches to the application of thesauri in spatial and thematic semantic distance measures. Semantic distance measures can underpin interactive and automatic query expansion techniques by ranking lists of candidate terms. We first illustrate how hierarchical spatial relationships can be used to provide more flexible retrieval for queries incorporating place names in applications employing online gazetteers and geographical thesauri. We then employ a set of experimental scenarios to investigate key issues affecting use of the associative (RT) thesaurus relationships in semantic distance measures. Previous work has noted the potential of RTs in thesaurus search aids but the problem of increased noise in result sets has been emphasised. Specialising RTs allows the possibility of dynamically linking RT type to query context. Results presented in this paper demonstrate the potential for filtering on the context of the RT link and on subtypes of RT relationships.

1. Introduction

Recent years have seen convergence of work in digital libraries, museums and archives with a view to resource discovery and opening up access to digital collections. Various projects are following standards-based approaches building upon terminology and knowledge organisation systems. Concurrently, within the web community, there has been growing interest in vocabulary-based techniques, with the realisation of the challenges posed by web searching and retrieval applications. This has manifested itself in metadata initiatives, such as Dublin Core and the proposed W3C Resource Description Framework. In order to support retrieval, provision is made in such metadata element sets for thematic keywords from vocabulary tools such as thesauri (ISO 2788, ISO 5964). Metadata schema (ontologies) incorporating thesauri or related semantic models underpin diverse ongoing projects in remote access, quality-based services, cross domain searching, semantic interoperability, building RDF models and digital libraries generally (Amann and Fundulaki 1999; Chen et al 1997; Doerr and Fundulaki 1998; Michard and Pham-Dac 1998)

Thesauri define semantic relationships between index terms (Aitchison and Gilchrist 1987). The three main relationships are Equivalence (equivalent terms), Hierarchical (broader/narrower terms: BT/NTs), Associative (Related Terms: RTs) and their specialisations. A large number of thesauri exist, covering a variety of subject domains, for example MEdical Subject Headings and the Art and Architecture Thesaurus (AAT 2000). Various studies have supported the use of thesauri in online retrieval and potential for combining free text and controlled vocabulary approaches (Fidel 1991). However there are various research challenges before fully utilising thesaurus structure in retrieval. In particular, the ‘vocabulary problem’ – differences in choice of index term at different times by indexers and searchers (Chen et al 1997) poses problems for work in cross domain searching and retrieval generally. For example, indexer and searcher may be operating at different levels of specificity, and at different times an indexer(s) may make different choices from a set of possible term options. While conventional narrower term expansion may help in some situations, a more systematic approach to thesaurus term expansion has the potential to improve recall in such situations. In the work described here, we have employed the Getty AAT and TGN (Thesaurus of Geographic Names - Harpring 1997) vocabularies. The AAT (Soergel 1995), is a large, evolving thesaurus (nearly 120,000 terms) widely used in the cultural heritage community, organised into 7 facets (and 33 hierarchies as subdivisions) according to semantic role. Harpring (1999) gives an overview of the Getty’s vocabularies with examples of their use in web retrieval interfaces and collection management systems. Examples are given of their use as a source of variant names of a concept. It is suggested that the AAT’s RT relationships may be helpful to a user exploring topics around an information need and the issue of how to perform query expansion without generating too large a result set is also raised.

The work described here is part of a larger project, OASIS (Ontologically Augmented Spatial Information System), exploring terminology systems for thematic and spatial access in digital library applications. One of our aims concerns the retrieval potential of geographical metadata schema, consisting of rich place name data but with locational data limited to a parsimonious approximation of spatial extent, or footprint. Such geographical representations may be appropriate for online gazetteers, geographical thesauri or geographic name servers, where conventional GIS datasets are unavailable, unnecessary or pose undesirable bandwidth limitations (Jones 1997). Notable projects include the Alexandria Digital Library (Frew et al 1998). Another aim explores the potential of reasoning over semantic relationships to assist retrieval from terminology systems. Measures of semantic distance make possible imprecise matching between query and information item, or between two information items, rather than relying on an exact match of terms (Tudhope and Cunliffe 1999). Previous work investigated hybrid query/navigation tools based on semantic closeness measures over the purely hierarchical Social History and Industrial Classification (Cunliffe et al 1997). This paper describes an integrated spatial and thematic schema and discusses two novel approaches to the application of thesauri, from both spatial and thematic points of view.

In section 2 we discuss our schema, illustrating how the spatial relationships in the thesaurus can be used to provide more flexible retrieval for queries incorporating place names. The second topic (sections 3 and 4) concerns the use of associative thesaurus relationships in retrieval. Existing collection management systems include access to thesauri for cataloguing with fairly rudimentary use of thesauri in retrieval (mostly limited to interactive query expansion/refinement and Narrower Term expansion). In particular, there is scope for increased use of associative (RT) relationships in thesaurus-based retrieval tools. RTs are non-hierarchical and are sometimes seen as weaker relationships. There is a danger that incorporating RTs into retrieval tools with automatic query expansion may lead to excessive 'noise' being introduced into result sets. It has been argued that semantic distance measures over RT relationships are less reliable than over hierarchical relationships, unless the user's query can be closely linked to the RT relationship. We discuss results from scenarios with semantic distance measures in order to map key issues affecting use of RTs. Conclusions are outlined in section 5.

2. OASIS Overview and Spatial Access Example

Thematic data was taken mainly from the Royal Commission on the Ancient and Historical Monuments of Scotland (RCAHMS) database, which contains information on Scottish archaeological sites and historical buildings (Murray, 1997). The OASIS ontology was linked to the AAT which provided thematic descriptors such as 'town', 'arrow', 'bronze', 'axe', 'castle', etc. The spatial data in the OASIS system includes information on hierarchical and adjacency relations between named places, in addition to place types, and (centroid) co-ordinates. This information was taken from the TGN, augmented with data derived from the Bartholomew's (Harper Collins 2000) digital map data for Scotland.

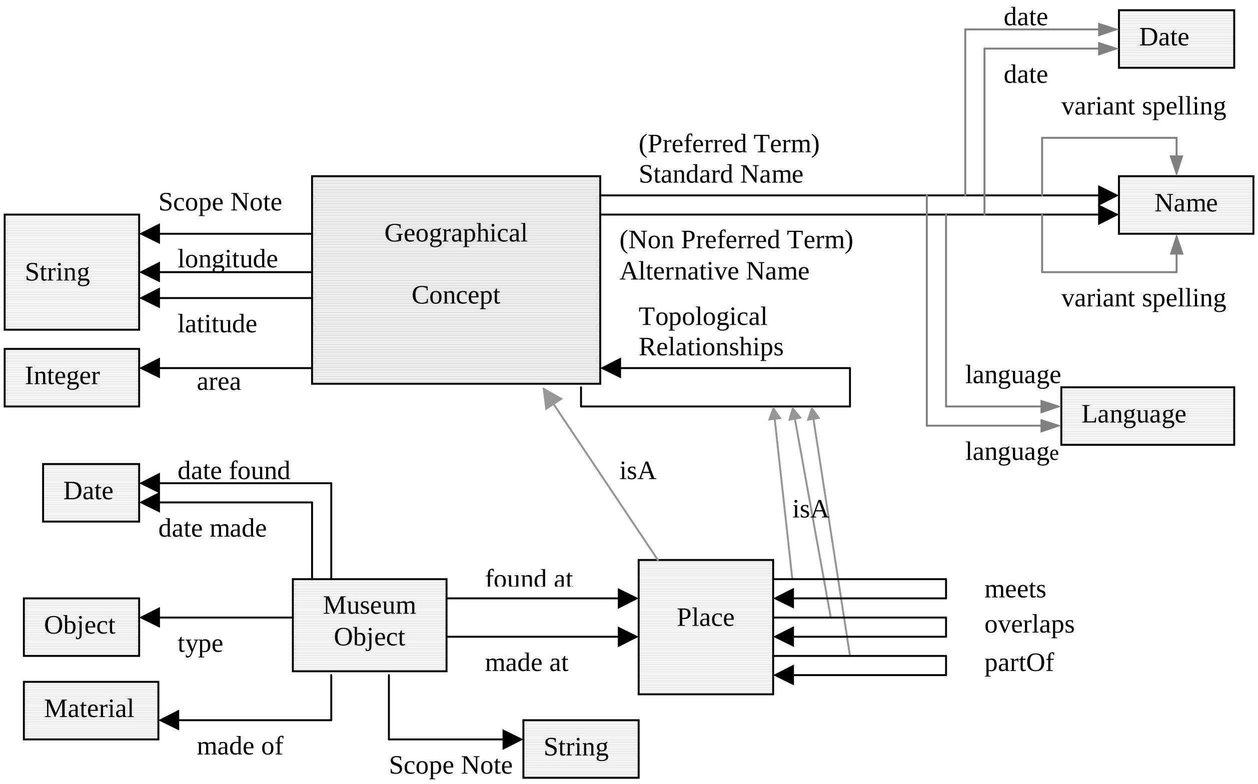

Figure 1. The Classification schema of *Place* and *Museum Object* in the OASIS system.

The term 'ontology' has widely differing uses in different domains (Guarino 1995). Our usage follows that of Amann and Fundulaki (1999), in that we see an ontology as a conceptualisation of a domain, in effect providing a connecting semantics between thesaurus hierarchies with specifications of roles for combining thesaurus elements. The OASIS schema (Figure 1) encompasses different versions of place names (e.g. current and historical names, different spellings, etc.), place types (e.g. Town, Building, Port, River, Hill), latitude and longitude co-ordinates, and topological relationships (e.g. meets, part of). The schema is implemented using the object-oriented Semantic Index System (SIS - Constantopolous and Doerr 1993) also used to store the data, and which provided the AAT implementation. The SIS has a meta modelling capability and an application interface for querying the schema. Figure 1 shows the meta level classification of the classes *Place* and *Museum Object*. As we discuss later in relation with RTs, relationships can be instantiated or subclassed from other relationships. Thus, *meets*, *overlaps*, and *partOf* are subclasses of *Topological Relationships*. The relationships *Standard Name* and *Alternative Name* are instances of the relationships *Preferred Term* and *Non Preferred Term* respectively (shown in brackets). The relationships *Standard Name* and *Alternative Name* are associated with the relationships; *variant spelling*, *date*, and *language*. For example, the *variant spelling* relationship links the place name (standard or alternative) to its spelling variations. *Place* inherits relationships such as *longitude*, *latitude*, *area*, etc, from its superclass *Geographical Concept*. The information stored in the OASIS database can be accessed using a set of functions through which it is possible to find all the information related to a given place, or find all the places with specific relationships, or to find objects at a place made of a certain material. For example to find all the places that are part of the City of Edinburgh, the system would return a set of all the places that are linked with a *partOf* relationship pointing to the City of Edinburgh.

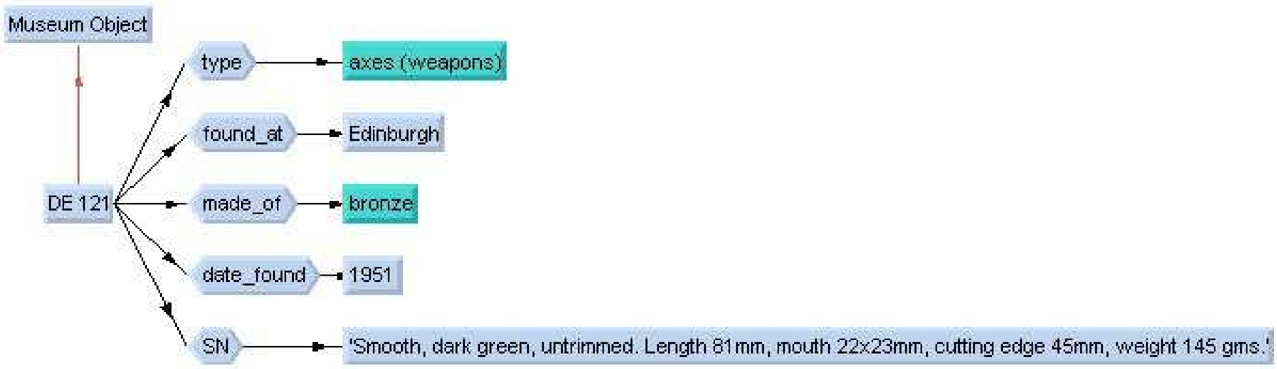

Figure 2. Classification of the axe artefact NMRS Acc. No. DE 121.

Figure 2 shows the OASIS classification of an axe artefact from the RCAHMS dataset. OASIS implements a set of thematic and spatial measures that enables query expansion to find similar terms. Consider the query *Do you have any information on axes found in the vicinity of Edinburgh?.* An exact match to the query would only return axes indexed by the term *Edinburgh*, such as the artefact represented in Figure 2. To search for axes found in the vicinity, spatial distance measures can be applied to expand the geographical term *Edinburgh* to spatially similar places, where axes have been found. Conventional GIS measures could be applied in situations where a full GIS polygon dataset is available. However, there are contexts where a GIS is either not appropriate (due to lack of co-ordinate data or bandwidth limitations) or where qualitative spatial relationships are important, eg remote access to online gazetteers and application contexts where administrative boundaries are important (Jones 1997).

In our database, a query on axe finds would return several places, including *Carlops, Corstorphine, Harlow Muir, Hermiston, Leith, Penicuik, Tynehead, West Linton*. These places can be ranked by spatial similarity using the *Part-of* spatial containment relationship, which in OASIS is based on the spatial hierarchies in the TGN. Given the term *Edinburgh*, the OASIS spatial hierarchy distance measure ranks *Corstorphine, Leith, Tynehead* equally and ahead of the other places listed, since (like *Edinburgh)* they are districts within the region *City of Edinburgh*. Similarly, since *Carlops* etc are places in Scotland, they would be returned ahead of any axe finds in England. In fact, the TGN provides centroid co-ordinate data for places/regions and our larger project explores the integration of different spatial distance measures and boundary approximation methods, based on geographical thesaurus relationships and limited locational footprint data (Alani et al 2000).

3. Semantic Distance Measures

A thesaurus can be used as a search aid to a user constructing a query by providing a set of controlled terms that can be browsed via some form of hypertext representation (eg Bosman et al 1998; Pollitt 1997). This can assist the user to understand the context of a concept and how it is used in a particular thesaurus and feedback on number of postings for terms (or combinations of terms) in a particular collection can also be provided. The inclusion of semantic relationships in the index space, moreover, provides the opportunity for knowledge-based approaches where the system takes a more active role in building a query by automatic reasoning over the relationships. Candidate terms can automatically be suggested for a user to consider in refining a query or various forms of query expansion are possible, making possible imprecise matching between query and media item, or between two media items (ie ‘More like this one’), rather than relying on an exact match of terms (Tudhope and Cunliffe 1999). The various Okapi projects conducted a number of experiments with thesauri as part of an underlying probabilistic retrieval model (Beaulieu 1997), investigating the extent to which the thesaurus should play an interactive or automatic role in query expansion.

The basis for such automatic term expansion is some kind of semantic distance measure. Semantic distance between two terms is often based on the minimum number of semantic relationships that must be traversed in order to connect the terms (Rada et al 1989). Each traversal has an associated cost factor. In poly-hierarchical systems, variations have been based on common or uncommon superclasses (Richardson et al 1994; Spanoudakis and Constantopoulos 1994, 1996), or have employed spreading activation (Chen and Dhar 1991, Cohen and Kjeldsen 1987; Croft et al 1989; Paice 1991). Rada et al (1989) assigned an identical cost to each traversal, whereas other work has assigned different weights depending on the relationship involved (McMath et al 1989, Kim and Kim 1990, Lee et al 1993). Sometimes depth within the hierarchical index space has been a factor, with distance between two connected terms considered greater towards the top of a hierarchy than towards the bottom, based on arguments concerning relative specificity, density or importance (Richardson et al 1994; Spanoudakis and Constantopoulos 1994). Other issues include similarity coefficients between sets of index terms (Smeaton and Quigley 1996, Tudhope and Taylor 1997). However our focus in this paper is upon the factors particularly relevant to the use of RTs in retrieval.

RTs represent a class of non-hierarchical relationships, which have been less clearly understood in thesaurus construction and applicability to retrieval than the hierarchical relationships. At one extreme, an RT is sometimes taken to represent nothing more than an extremely vague ‘See-also’ connection between two concepts. This can lead to an introduction of excessive noise in result sets when RT relationships are expanded. Rada et al (1989) argue from plausible demonstration scenarios that semantic distance measures over RT relationships can be less reliable than over hierarchical relationships, unless the user's query can be closely linked to the RT relationship - a medical expert system example is given in Rada et al (1991). As we discuss later, structured definitions of RTs (eg Aitchison and Gilchrist 1987) offer potential for systematic approaches to their use. There is some evidence that RTs can be useful in retrieval situations. The basic assumption of a cognitive basis for a semantic distance effect over thesaurus terms has been investigated by Brooks (1997), in a series of experiments exploring the relevance relationships between bibliographic records and topical subject descriptors. These studies employed the ERIC database and thesaurus and consisted of purely linear hierarchies, as opposed to tree hierarchical structures (as with the AAT) or indeed poly-hierarchies. However the results are suggestive of the existence of some semantic distance effect, with an inverse correlation between semantic distance and relevance assessment, dependant on position in the subject hierarchy, direction of term traversal and other factors. In particular, a definite effect was observed for RTs (typically less than for hierarchical traversal). An empirical study by Kristensen (1993) compared single-step automatic query expansion of synonym, narrower-term, related term, combined union expansion and no expansion of thesaurus relationships. Thesaurus expansion was found to improve recall significantly at some (lesser) cost in precision. Taken separately, single step RT expansion results did not differ significantly from NT or synonym expansion (specific results showing a 12% increase in Recall over NTs, but with 2.8% decrease in Precision). In another empirical study (Jones et al 1995), a log was kept of users’ choices of relationships interactively expanded via thesaurus navigation while entering a query. In this study of users refining a query, a majority of terms retrieved from the thesaurus came from RTs (although it should be noted that the INSPEC thesaurus employed at that time contained many more RTs than hierarchical relationships).

4. RT Scenarios and Discussion

This section maps key issues affecting use of RTs in term expansion algorithms for retrieval. Results are given from a series of scenarios applying different versions of a semantic distance algorithm to terms in the AAT (AAT 2000). The distance measure employed a branch and bound algorithm, with weights for relationships given below and a depth factor which reduced costs according to hierarchical depth. It was implemented in C++ using the SIS function library to query the underlying schema given in Figure 1.

Our aim was to investigate different factors relevant to RT expansion, rather than relative weighting of relationships. In general the purpose of weighting relationships is to achieve a ranking in 'semantically close' terms to allow a user to either choose a candidate term to expand a query or to select an information item from a result set deriving from an automatic query expansion. If a useful ranking is produced then the weighting may be said to have performed its purpose. When assigning weights to relationships it should be noted that there may be a dependency on type of application and particular thesaurus involved. The choice of threshold to truncate expansion is an associated factor, which may in practice be made contingent on some user indication of the amount of flexibility desired in results. The weights chosen for this experiment were selected to reflect some broad consensus of previous work. Commercial collection management or retrieval systems employing a thesaurus tend to be restricted to narrower term expansion (if any), thus favouring NTs. McMath et al (1989) assigned costs of 10, 15 to NT and BT respectively. Chen and Dhar (1991) employed weights of 9, 5, and 1 for NT, RT, and BT relationships respectively. Their weights were set according to the use frequency of relationships during empirical search experiments. Cohen et al's (1987) spreading activation algorithm traversed NT before BT. Our weights (BT 3, NT 3, RT 4), taken together with a depth factor inversely proportional to the hierarchical depth of the destination term, assign lowest costs to NTs and favour RTs over BTs at higher depths in the hierarchy (following an AAT editorial observation that RTs appear to work better at fairly broad levels). The threshold used to terminate expansion was 2.5.

We developed a series of experimental scenarios based around term generalisation involving RT traversal. Building on the example in Section 2, we focus on the AAT's *Objects* Facet: *Weapons & Ammunition* and *Tools & Equipment* hierarchies. The introductory scenario supposes a narrowly defined information need for items concerning axes used as weapons (mapping to AAT term *Axes (weapons)*). In this initial scenario, expansion is limited and restricted to NT relationships only: *tomahawks (weapons)*, *battle-axes*, *throwing axes*, and *franciscas*.

The second scenario supposes an information need for items more broadly connected with axes used as weapons – thus allowing for some flexibility in expansion. We first consider expansion only over hierarchical relationships and then discuss expansion with RTs. Table 1 shows results from BT/NT expansion only, with path and semantic distance shown for each term.

Term	Dist.	Path	Term	Dist.	Path	Term	Dist.	Path
axes (weapons)	0	()	halberds	2.35	(BT NT NT)	poniards	2.35	(BT NT NT)
tomahawks	0.6	(NT)	pollaxes	2.35	(BT NT NT)	stilettos (daggers)	2.35	(BT NT NT)
battle-axes	0.6	(NT)	gisarmes	2.35	(BT NT NT)	trench knives	2.35	(BT NT NT)
edged weapons	1	(BT)	bills (staff weapons)	2.35	(BT NT NT)	arm daggers	2.35	(BT NT NT)
throwing axes	1.1	(NT NT)	corsescas	2.35	(BT NT NT)	fighting bracelets	2.35	(BT NT NT)
franciscas	1.53	(NT NT NT)	glaives	2.35	(BT NT NT)	finger hooks	2.35	(BT NT NT)
staff weapons	1.75	(BT NT)	integral bayonets	2.35	(BT NT NT)	finger knives	2.35	(BT NT NT)
sword sticks	1.75	(BT NT)	knife bayonets	2.35	(BT NT NT)	brass knuckles	2.35	(BT NT NT)
harpoons	1.75	(BT NT)	plug bayonets	2.35	(BT NT NT)	switchblade knives	2.35	(BT NT NT)
bayonets	1.75	(BT NT)	socket bayonets	2.35	(BT NT NT)	dirks	2.35	(BT NT NT)
daggers (weapons)	1.75	(BT NT)	sword bayonets	2.35	(BT NT NT)	bolos (weapons)	2.35	(BT NT NT)
fist weapons	1.75	(BT NT)	left-hand daggers	2.35	(BT NT NT)	bowie knives	2.35	(BT NT NT)
knives (weapons)	1.75	(BT NT)	cinquedeas	2.35	(BT NT NT)	Landsknecht daggers	2.35	(BT NT NT)
swords	1.75	(BT NT)	ballock daggers	2.35	(BT NT NT)	<swords by form>	2.35	(BT NT NT)
partisans	2.35	(BT NT NT)	baselards	2.35	(BT NT NT)	<swords by function>	2.35	(BT NTNT)
spears (weapons)	2.35	(BT NT NT)	eared daggers	2.35	(BT NT NT)	weapons	2.5	(BT BT)
leading staffs	2.35	(BT NT NT)						

Table 1. BT/NT expansion only.

When term expansion is extended to RTs in a distance measure including a depth factor, it becomes important to base RT depth on the starting (not destination) term. Otherwise, two terms one link away could appear at different distances if they came from different hierarchical levels and this distortion is propagated to subsequent BT/NT expansions. Table 2 shows the effect of introducing RT expansion. Note that in this scenario staff weapons related to axes are brought closer *(halberds*, *pollaxes*, *gisarmes)* and new terms, (such as *axes (tools)*, *chip axes*, *ceremonial axes)* are introduced. The latter set of terms could be relevant to broader information needs or to situations when a thesaurus entry term was mismatched (in this case the information need might relate more to tool use). In some situations however, the RTs could be seen as 'noise'.

Term	Dist.	Path	Term	Dist.	Path	Term	Dist.	Path
axes (weapons)	0	()	adze-hatchets	1.9	(NT RT NT)	sword bayonets	2.35	(BT NT NT)
tomahawks (weapons)	0.6	(NT)	hewing hatchets	1.9	(NT RT NT)	left-hand daggers	2.35	(BT NT NT)
battle-axes	0.6	(NT)	lathing hatchets	1.9	(NT RT NT)	cinquedeas	2.35	(BT NT NT)
edged weapons	1	(BT)	shingling hatchets	1.9	(NT RT NT)	ballock daggers	2.35	(BT NT NT)
axes (tools)	1	(RT)	<cutting tools>	2	(RT BT)	baselards	2.35	(BT NT NT)
halberds	1	(RT)	fasces	2	(RT RT)	eared daggers	2.35	(BT NT NT)
pollaxes	1	(RT)	Pulaskis	2	(RT RT)	Landsknecht daggers	2.35	(BT NT NT)
gisarmes	1	(RT)	<ceremonial weapons>	2	(RT BT)	poniards	2.35	(BT NT NT)
ceremonial axes	1	(RT)	<wood-cutting and -finishing tools>	2.15	(NT RT BT)	stilettos (daggers)	2.35	(BT NT NT)
throwing axes	1.1	(NT NT)	arrows	2.33	(BT RT)	trench knives	2.35	(BT NT NT)
hatchets	1.4	(NT RT)	machetes	2.33	(BT RT)	arm daggers	2.35	(BT NT NT)
franciscas	1.53	(NT NT NT)	darts	2.33	(BT RT)	dirks	2.35	(BT NT NT)
chip axes	1.6	(RT NT)	partisans	2.35	(BT NT NT)	fighting bracelets	2.35	(BT NT NT)
berdyshes	1.6	(RT NT)	spears (weapons)	2.35	(BT NT NT)	finger hooks	2.35	(BT NT NT)
staff weapons	1.75	(BT NT)	leading staffs	2.35	(BT NT NT)	finger knives	2.35	(BT NT NT)
sword sticks	1.75	(BT NT)	bills (staff weapons)	2.35	(BT NT NT)	brass knuckles	2.35	(BT NT NT)
harpoons	1.75	(BT NT)	corsescas	2.35	(BT NT NT)	switchblade knives	2.35	(BT NT NT)
bayonets	1.75	(BT NT)	glaives	2.35	(BT NT NT)	bolos (weapons)	2.35	(BT NT NT)
daggers (weapons)	1.75	(BT NT)	integral bayonets	2.35	(BT NT NT)	bowie knives	2.35	(BT NT NT)
fist weapons	1.75	(BT NT)	knife bayonets	2.35	(BT NT NT)	<swords by form>	2.35	(BT NT NT)
knives (weapons)	1.75	(BT NT)	plug bayonets	2.35	(BT NT NT)	<swords by function>	2.35	(BT NT NT)
swords	1.75	(BT NT)	socket bayonets	2.35	(BT NT NT)	weapons	2.5	(BT BT)
<projectiles with nonexplosive propellant>	1.77	(NT NT RT)						

Table 2. RT expansion included.

One method of reducing noise introduced by RT expansion is by filtering on the original term's (sub)hierarchy - RTs to terms within different sub-hierarchies are not traversed (or could be penalised). Table 3 shows a set of terms (and their hierarchies) which are removed from the above example (distances are from Table 2). Note that instances of axes serving both as tools and as weapons *(hatchets, machetes)* are now excluded, since due to the mono-hierarchical nature of the AAT they are located within the *Tools&Equipment* hierarchy.

Term	Dist.	Sub-hierarchy	Term	Dist.	Sub-hierarchy
axes (tools)	1	Tools & Equipment	<cutting tools>	2	Tools & Equipment
hatchets	1.4	Tools & Equipment	fasces	2	Information Forms
chip axes	1.6	Tools & Equipment	Pulaskis	2	Tools & Equipment
adze-hatchets	1.9	Tools & Equipment	<wood-cutting and - finishing tools>	2.15	Tools & Equipment
hewing hatchets	1.9	Tools & Equipment	machetes	2.33	Tools & Equipment
lathing hatchets	1.9	Tools & Equipment			
shingling hatchets	1.9	Tools & Equipment			

Table 3. Terms excluded when inter-hierarchical traversals are not allowed.

The next scenario explores an alternative approach to filtering based upon selective specialisation of the RT relationship according to retrieval context. This is in keeping with the recommendation of Rada et al (1991) that automatic expansion of non-hierarchical relationships be restricted to situations where the type of relationship can be linked with the particular query, and also with Jones' (1993) discussion of using sub-classifications to help distinguish relationships according to strength. The aim is to take advantage of more structured approaches to thesaurus construction where different types of RTs are employed. For example, common subdivisions of RTs include partitive and causal relationships (Aitchison and Gilchrist 1997). In some circumstances it may be appropriate to consider all types of associative relationships as a generic RT for retrieval purposes (as in the above scenarios). However, under other contexts it may be desirable to treat RT sub-types differently, permitting some RT traversals but forbidding or penalising (via weighting) others. Thus heuristics may selectively guide RT expansion, depending on query model and session context. The AAT is particularly suited to investigation of this topic, since its editors followed a systematic, rule-based approach to the design of RT links (Molholt 1996). The AAT RT editorial manual specifies a set of rules to apply to the relevant hierarchical context and scope notes in order to identify valid RT relationships between terms when building the vocabulary or enhancing it. This includes a set of specialisations of the RT relationships (AAT 1995), following their notation: *1A and 1B)* Alternate hierarchical (BT/NT) relationships (since AAT is not polyhierarchical); *2A and 2B)* Part/Whole relationships; *3)* Several Inter/intra Facet relationships (eg Agents-Activities and Agents-Materials); *4)* Distinguished From relationship (the scope note evidences a need to distinguish the sense of two terms); *5)* frequently Conjuncted terms (eg Cups AND Saucers). We have extended the original SIS AAT schema to specialise the associative relationship. See Figure 3, where (for example) *AAT_RT_4* represents the Distinguished From relationship (the 19 *AAT_RT_3* subtypes are not displayed separately in interests of space). RTs in our model can optionally be treated as specialised sub-relationships, or as generic RTs via *associative_relation_Type.*

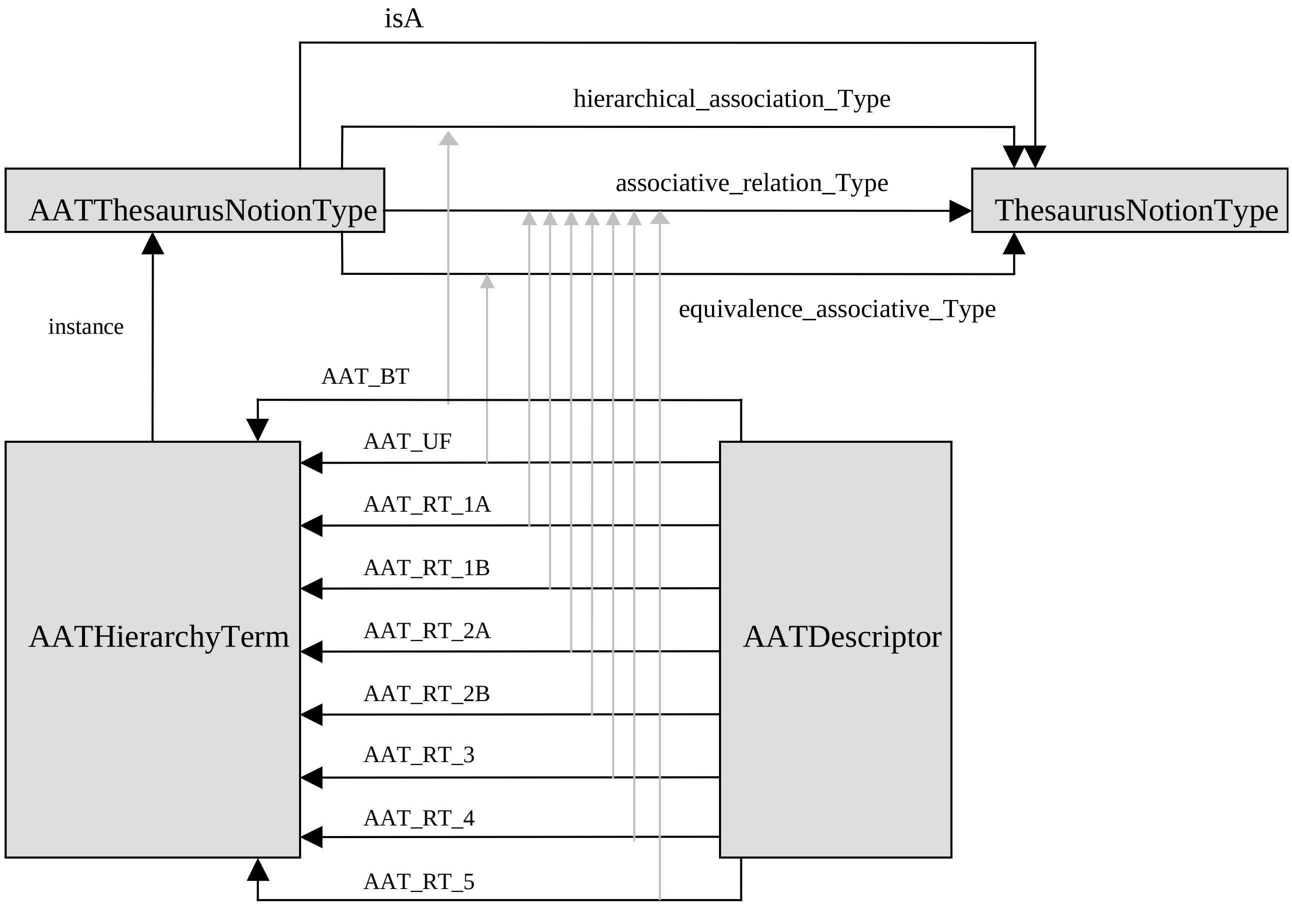

Figure 3. Specialisation of the associative relationship.

The editorial rules for creating specific associative relationships are not retained in electronic implementations of the AAT to date. Thus, for this experiment we manually specialised all RT relationships 3 links away from *axes (weapons)* into their corresponding sub-types by following sample extracts of AAT Editorial Related Term Sheets and applying the editorial rules. In the scenario, the distance algorithm was set to filter on the subtype of RT, only permitting traversal over the Alternate BT and Alternate NT relationships. Table 4 summarises the differences (terms included and excluded) with the hierarchy filtering approach (Table 3) – all terms of course were present in the unfiltered Table 2. This might correspond to a reasonably strict information request but results retrieved now include terms, such as *machetes*, *hatchets* from the *Tools & Equipment* hierarchy, excluded when narrowly filtering on the hierarchy. For example, an alternate NT relationship exists between *tomahawks* and *hatchets*. Since they are classed as both tools and weapons, *hatchets* might well be regarded as relevant to the scenario.

Terms Included			**Terms Excluded**		
Term	**Dist.**	**Path**	**Term**	**Dist.**	**Path**
hatchets	1.4	(NT RT)	axes (tools)	1	(RT)
adze-hatchets	1.9	(NT RT NT)	chip axes	1.6	(RT NT)
hewing hatchets	1.9	(NT RT NT)	<cutting tools>	2	(RT BT)
lathing hatchets	1.9	(NT RT NT)	fasces	2	(RT RT)
shingling hatchets	1.9	(NT RT NT)			
<wood cutting and -finishing tools>	2.15	(NT RT BT)			
Pulaskis	2.2	(NT RT RT)			
machetes	2.33	(BT RT)			

Table 4. Filtering by RT specialisation

This specialisation allows for retrieval purposes a treatment of the AAT as a poly-hierarchical system. Some reviewers have been critical of its mono-hierarchical design (Soergel 1995). By filtering on the subtype of RT relationship it can be taken as treated for retrieval as a mono or poly hierarchy accordingly. It may well be preferable to weight such alternate hierarchical RT relationships identically to BT/NTs, but this is an issue for future investigation.

The AAT Scope Note for *axes (weapons)* reads:
"Cutting weapons consisting basically of a relatively heavy, flat blade fixed to a handle, wielded by either striking or throwing. For axes used for other purposes, typically having narrower blades, use axes (tools)."
Thus the associative relationship between *axes (weapons)* and *axes (tools)* is of subtype *Distinguished From* and is not traversed in this scenario when filtering only on alternate hierarchical RT subtypes. We can see in Table 4 that the term *axes (tools)* and tool-related terms derived solely from this link *(chip axes, cutting tools, etc)* are excluded. Under some contexts, such terms might be considered of relevance but in a stricter weapons-related scenario they might well be seen as less relevant and can now be suppressed. The point is that this control can be passed to the retrieval system.

Other scenarios illustrate the potential for filtering on other types of RT relationship. For example, an information need relating to *archery and its equipment*, would justify traversal of AAT RT inter-facet subtype *Activity - Equipment Needed or Produced*. This would in turn yield the terms *arrows* and *bows (weapons)*, which could be expanded to terms such as *bolts (arrows)*, *crossbows*, *composite bows*, *longbows*, and *self bows*. The same approach can be applied to scenarios relating to parts or components of an object, using the RT *Whole/Part*, and *Part to Whole* subtypes. Here, a query on *arrows* would yield the terms *nocks* and *<arrow components>* which could be expanded to terms such as *arrowheads*, and *feathers (arrow components).*

The effect of combining RT and BT/NT expansion, or chains of hierarchical and non-hierarchical relationships warrants some future investigation. Should all possible chains of relationships be considered equally transitive for retrieval purposes?. For example, in our scenarios RT-BT traversal chains led to some tenuous links (<cutting tools>, <ceremonial weapons>). One approach to reducing noise might be to consider penalising certain combinations or vary RT weighting depending on order of relationship traversal, although it is difficult to argue from individual cases. Support for this can be found in the AAT RT editorial manual which stresses a guiding inheritance principle when identifying RT relationships: RT links from an initial terms must apply to all NTs of the target term. RT-BT chains could be seen as less valid and RT-NT chains as more valid from consideration of the inheritance principle – however the topic needs further investigation.

5. Conclusions

Semantic distance measures operating over thesaurus relationships can underpin interactive and automatic query expansion techniques. It may be impractical to expect non-specialist users to manually browse very large thesauri (for example, there are 1792 terms in the AAT's *Tools&Equipment* hierarchy). Ranked lists of candidate terms can assist query expansion or automatic ranking of information items by a matching function. Results are presented in this paper from novel approaches to semantic distance measures for associative relationships and geographical thesauri. Online gazetteers and geographical thesauri may not contain co-ordinate data for all places and regions or, if they do, associate place names with a limited spatial footprint (centroid or minimum bounding rectangle). In such situations, the ability to rank places within a vicinity according to hierarchical (or other) relationships in a spatial terminology system can be useful. In contexts where administrative boundaries are highly relevant, distance measures could combine quantitative and qualitative spatial relationships. Related work has noted the potential of RTs in thesaurus search aids but the problem of increased noise in result sets has been emphasised. Experimental scenarios (Section 4) exploring different factors relating to incorporation of RTs in semantic distance measures demonstrate the potential for filtering on the context of the RT link in faceted thesauri and on subtypes of RT relationships. Specialising RTs allows the possibility of dynamically linking RT type to query context and, in cases like the AAT, treating alternate hierarchical RT relationships more flexibly for retrieval purposes. Thus RT subtypes can be selectively filtered in or out of distance measures, depending on cues derived from an expression of information need or from information elicited by a query editor. In practice, it is likely that a combination of filtering heuristics will be useful. The ability in retrieval to either specialise RTs or to treat them as generic retains the advantages of the standard minimal set of thesaurus relationships for interoperability purposes, while allowing an option of a richer set of RT sub-relationships.

There are implications for thesaurus developers and implementers. A systematic approach to RT application in thesaurus design, as in the AAT, has potential for retrieval systems. Information (eg of relationship subclasses) used in thesaurus design should be retained in data models and database design for later use in retrieval algorithms. In future work, we intend to build on the underlying semantic distance measures and explore how best to incorporate thesaurus semantic distance controls in the user interface - we will investigate the performance, utility and usability of resulting retrieval systems. The issue of RT specialisations expressing thesaurus inter-facet links and its retrieval implications is a promising area we intend to pursue, which converges with work on broader ontological conceptualisations attempting to more formally define the roles played by entities in the schema.

Acknowledgements
We would like to thank the Getty Information Institute for provision of their vocabularies and in particular Alison Chipman for information on Related Terms; 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 1995. The AAT Editorial Manual: Related terms. User Friendly, 2(3-4), 6-15. Getty Art History Information Program.

AAT 2000. http://shiva.pub.getty.edu/aat_browser/

Aitchison J., Gilchrist A. 1987. Thesaurus construction: a practical manual. ASLIB: London.

Alani H., Jones C., Tudhope D. 2000. Voronoi-based region approximation for geographical information retrieval with online gazetteers. Working Paper.

Amann B., Fundulaki I. 1999. Integrating ontologies and thesauri to build RDF schemas. Proc. 3rd European Conference on Digital Libraries (ECDL'99), (S. Abiteboul and A. Vercoustre eds.) Lecture Notes in Computer Science 1696, Springer-Verlag: Berlin, 234-253.

Beaulieu M. 1997. Experiments on interfaces to support query expansion. Journal of Documentation, 53(1), 8-19.

Bosman F., Bruza P., van der Weide T., Weusten L. 1998. Documentation, cataloguing, and query by navigation: a practical and sound approach. Proc. 2nd European Conference on Digital Libraries (ECDL'98), (C. Nikolaou and C. Stephanidis eds.) Lecture Notes in Computer Science 1513, Springer-Verlag: Berlin, 459-478.

Brooks T. 1997. The relevance aura of bibliographic records. Information Processing and Management, 33(1), 69-80.

Chen H., Dhar V. 1991. Cognitive process as a basis for intelligent retrieval systems design. Information Processing and Management, 27(5), 405-432.

Chen H., Ng T., Martinez J., Schatz B. 1997. A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the Worm Community System. Journal of the American Society for Information Science, 48(1), 17-31.

Cohen, P. R. and R. Kjeldsen (1987). Information Retrieval by Constrained Spreading Activation in Semantic Networks. Information Processing & Management 23(4): 255-268.

Constantopolous P., Doerr M. 1993. The Semantic Index System - A brief presentation. Institute of Computer Science Technical Report. FORTH-Hellas, GR-71110 Heraklion, Crete.

Croft W., Lucia T., Cringean J., Willett P. 1989. Retrieving documents by plausible inference: an experimental study. Information Processing and Management, 25(6), 599-614.

Cunliffe D., Taylor C., Tudhope D. 1997. Query-based navigation in semantically indexed hypermedia. Proc. 8th ACM Conference on Hypertext, 87-95.

Doerr M., Fundulaki I. 1998. SIS-TMS: A thesaurus management system for distributed digital collections. Proc. 2nd European Conference on Digital Libraries (ECDL'98), (C. Nikolaou and C. Stephanidis eds.) Lecture Notes in Computer Science 1513, Springer-Verlag: Berlin, 215-234.

Fidel R. 1991. Searchers' selection of search keys (I-III), Journal of American Soc. for Inf. Science, 42(7), 490-527.

Frew J., Freeston M., Freitas N., Hill L., Janee G., Lovette K., Nideffer R., Smith T., Zheng Q. 1998. The Alexandria Digital Library Architecture. Proc. 2nd European Conference on Digital Libraries (ECDL'98), (C. Nikolaou and C. Stephanidis eds.) Lecture Notes in Computer Science 1513, Springer-Verlag: Berlin, 61-73.

Guarino N. 1995. Ontologies and knowledge bases: towards a terminological clarification. In: Towards very large knowledge bases: knowledge building and knowledge sharing, 25-32. IOS Press.

Harper Collins, 2000, Bartholomew. http://www.bartholomewmaps.com

Harpring P. 1997. The limits of the world: Theoretical and practical issues in the construction of the Getty Thesaurus of Geographic Names. Proc. 4th International Conference on Hypermedia and Interactivity in Museums (ICHIM'97), 237-251, Archives and Museum Informatics.

Harpring P. 1999. How forcible are the right words: overview of applications and interfaces incorporating the Getty vocabularies. Proc. Museums and the Web 1999. Archives and Museum Informatics. http://www.archimuse.com/mw99/papers/harpring/harpring.html

Jones C. 1997. Geographic Interfaces to Museum Collections. Proc. 4th International Conference on Hypermedia and Interactivity in Museums (ICHIM'97), 226-236, Archives and Museum Informatics.

Jones, S. 1993. A Thesaurus Data Model for an Intelligent Retrieval System. Journal of Information Science 19: 167-178.

Jones S., Gatford M., Robertson S., Hancock-Beaulieu M., Secker J., Walker S. 1995. Interactive Thesaurus Navigation: Intelligence Rules OK?, Journal of the American Society for Information Science, 46(1), 52-59.

Kim Y., Kim J. 1990. A model of knowledge based information retrieval with hierarchical concept graph. Journal of Documentation, 46(2), 113-136.

Kristensen J. 1993. Expanding end-users' query statements for free text searching with a search-aid thesaurus. Information Processing and Management, 29(6), 733-744.

Lee J., Kim H., Lee Y. 1993. Information retrieval based on conceptual distance in ISA hierarchies. Journal of Documentation, 49(2), 113-136.

McMath C. F., Tamaru R. S., Rada R. 1989. A graphical thesaurus-based information retrieval system, 31, 121-147.

Michard A., Pham-Dac G. 1998. Description of Collections and Encyclopaedias on the Web using XML. Archives and Museum Informatics, 12(1), 39-79.

Molholt P. 1996. Standardization of inter-concept links and their usage. Proc. 4th International ISKO Conference, Advances in Knowledge Organisation (5), 65-71.

Murray D. 1997. GIS in RCAHMS. MDA Information 2(3): 35-38.

Paice C 1991. A thesaural model of information retrieval. Information Processing and Management, 27(5), 433-447.

Pollitt A. 1997. Interactive information retrieval based on facetted classification using views. Proc. 6th International Study Conference on Classification, London.

Rada R., Mili H., Bicknell E., Blettner M. (1989). Development and Application of a Metric on Semantic Nets. IEEE Transactions on Systems, Man and Cybernetics, 19(1), 17-30.

Rada R, Barlow J., Potharst J., Zanstra P., Bijstra D. 1991. Document ranking using an enriched thesaurus. Journal of Documentation, 47(3), 240-253.

Richardson R., Smeaton A., Murphy J. 1994. Using Wordnet for conceptual distance measurement, Proc. 16th Research Colloquium of BCS IR Specialist Group, 100-123.

Smeaton A., & Quigley I. 1996. Experiments on Using Semantic Distances Between Words in Image Caption Retrieval, Proc. 19th ACM SIGIR Conference, 174-180.

Soergel. D 1995. The Art and Architecture Thesaurus (AAT): a critical appraisal. Visual Resources, 10(4), 369-400.

Spanoudakis G., Constantopoulos P. 1994. Similarity for analogical software reuse: a computational model. Proc. 11th European Conference on AI (ECAI'94), 18-22. Wiley.

Spanoudakis G., Constantopoulos P. 1996. Elaborating analogies from conceptual models. International Journal of Intelligent Systems. 11, 917-974.

Tudhope D., Taylor C. 1997. Navigation via Similarity: automatic linking based on semantic closeness. Information Processing and Management, 33(2), 233-242.

Tudhope D., Cunliffe D. 1999. Semantic index hypermedia: linking information disciplines. ACM Computing Surveys, Symposium on Hypertext and Hypermedia. Forthcoming.