Representation and Retrieval in Faceted Systems

Abstract: This paper discusses two inter-related themes: the retrieval potential of faceted thesauri and XML representations of fundamental facets. Initial findings are discussed from the ongoing ‘FACET’ project, in collaboration with the National Museum of Science and Industry. The work discussed seeks to take advantage of the structure afforded by faceted systems for multi-term queries and flexible matching, focusing in this paper on the Art and Architecture Thesaurus. A multi-term matching function yields ranked results with partial matches via semantic term expansion, based on a measure of distance over the semantic index space formed by thesaurus relationships. Our intention is to drive the system from general representations and a common query structure and interface. To this end, we are developing an XML representation based on work by the Classification Research Group on fundamental facets or categories. The XML representation maps categories to particular thesauri and hierarchies. The system interface, which is configured by the mapping, incorporates a thesaurus browser with navigation history together with a term search facility and drag and drop query builder.

1. Introduction

Our research investigates the extent to which a search system can play an active role in the retrieval process by automatic traversal of the semantic relationships in a thesaurus. Rather than relying on an exact match of terms, semantic term expansion makes possible flexible matching on 'close' terms. This offers various possibilities for retrieval tools beyond interactive browsing of the hierarchies. Items can be ranked by degree of similarity in a result or destination set, the system might automatically suggest terms to be considered for inclusion in a query, or having found one item of interest, a 'More like this' option can yield a set of similar but not necessarily identically indexed items.

Facet analysis is a key technique in knowledge organisation; concepts are decomposed into elemental classes, or facets, which form homogenous mutually exclusive groups. Faceted systems include MeSH, BLISS, PRECIS and the one discussed in this paper, the J. Paul Getty Trust’s Art and Architecture Thesaurus (AAT). Recent times have seen interest in applying faceted approaches to the web and online retrieval generally (Broughton 2001; Chan et al., 2001; Pollitt, 1998). HIBROWSE (Pollitt, 1997) demonstrates the potential of faceted thesauri for information retrieval; hierarchies can be browsed and terms from different, "mutually constraining" facets interactively combined, with the number of postings shown dynamically as the query is constructed and refined.

The work discussed here seeks to take advantage of the structure afforded by faceted schemes for multi-term queries and flexible matching. Frequently queries contain multiple terms. Multi-term descriptors have potential for very specific item descriptions and high precision in retrieval. However, their full potential has yet to be exploited. The lack of flexible retrieval tools that can deal with ranked matches
of sets of terms is a disincentive to use of multi-term descriptors in indexing and search: "The major problem lies in developing a system whereby individual parts of subject headings containing multiple AAT terms are broken apart, individually exploded hierarchically, and then reintegrated to answer a query with relevance" (Petersen, 1994).

With the aim of increasing the universality of the work, our intention is to drive the system from general representations that can be mapped to particular thesauri and collections. To this end, we are developing an XML representation of important aspects of thesaurus structure and key system parameters. We report here on initial findings from this work, part of the ongoing ‘FACET’ project at Glamorgan (FACET, 2002), in collaboration with the National Museum of Science and Industry, which includes the National Railway Museum (NRM) whose collections database forms the dataset for the project. The system is implemented on a SQL-Server platform with a C++ retrieval engine and a Visual Basic interface incorporating a thesaurus browser with navigation history and ‘Favourites’ together with a term search facility and drag and drop query builder (Figure 1). While the thesaurus is stored as relational tables in the database, a parallel in-memory representation of the semantic network of relationships underlying the thesaurus permits real-time semantic term expansion, as outlined in Section 2. The main thesaurus in the project to date is the AAT (AAT, 2002, Soergel, 1995), although we are also incorporating smaller more specialised thesauri such as the NRM’s Railway Terminology Thesaurus. The AAT is a large thesaurus (over 120,000 terms). Descriptors are organised into 7 facets (and 33 hierarchies as subdivisions): Associated Concepts, Physical Attributes, Styles&Periods, Agents, Activities, Materials, Objects.

2. Semantic Term Expansion

Semantic term expansion is based on a measure of distance over the semantic index space formed by thesaurus relationships. Different cost factors are associated with traversal across the different relationship types, also taking into account depth in the hierarchy. Starting from a given term, the algorithm spreads over connecting relationships until a ‘semantic closeness’ threshold is reached (Tudhope & Taylor, 1997). Measures of distance between queries and item descriptors with multiple terms pose particular problems. The number of index terms may differ and terms may not match exactly. Here we outline work on a multi-term matching function that includes semantic term expansion (for details of the algorithm see Tudhope et al, 2002). Figure 2 shows the interface main window with list of queries and ranked results for the selected query. The top ranking result has 75% match since three out of four terms match and there is no penalty for extra item descriptor terms, whereas the 50% matches have just two matching terms. Item 1988-7335 has a higher match than 50% due to the additional partial matches on Carver chairs and Queen Anne Style. Similarly, item 1986-7794 fully matches on two terms but is ranked above the 50% items due to a partial match between upholstering and buttoning after semantic term expansion. For example, Edwardian, Regency and Victorian belong to a mini-hierarchy in the
Styles & Periods facet and Queen Anne Style is also a modern British style period. The bolded term (armchairs) in Figure 1 indicates a 'focus term' which must yield a match (after expansion). The choice of focus term can be dynamically changed with corresponding effect on results. Instead of one unified ranking, it would also be possible to produce rankings by facets separately. Queries and their result sets are stored persistently, using an XML representation.

Figure 1: FACET interface - Browser and Query Builder with query used in Section 2

Figure 2: Queries with results (selected query: leather Edwardian upholstered armchairs)
3. Fundamental Facets and Categories

The faceted approach to subject analysis began in 1933 with Ranganathan's Colon Classification (Personality, Matter, Energy, Space and Time). Ranganathan's analytico-synthetic approach was subsequently elaborated by the (British) Classification Research Group (CRG) who evolved an extended set of commonly occurring fundamental categories (Table I) from the experience of developing a series of special classifications in the 1960s and 1970s and more recently BC2, the revised general BLISS Classification (Broughton 2001; Mcllwaine and Broughton 2000; Vickery 1960). Broughton (2001) discusses the advantages of a faceted approach over enumerated classifications: avoids having to preassign all compound terms to the classification; avoids arbitrary assignment of complex subject headings to a single place in a hierarchical classification (or complicated poly­hierarchies); subject headings can be synthesised from underlying facet elements. Entity is the main focus of an application domain - it could be types of mammal in a zoological taxonomy, subjects in library applications and in museum collections it will tend to refer to objects. It should be emphasised that facet analysis is more a technique than any one prescribed structure and that the CRG categories are not meant as absolutes but to be adapted for different applications. However the categories appear to overlap with several faceted schemes (eg Aitchison, Gilchrist & Bawden, 2000, p. 70) and capture useful higher level concepts. The current ISO standard also mentions common categories and work on the semantic network for the UMLS metathesaurus project (UMLS, 2002) has produced a set of categories suitable for various (medical) applications.

<table>
<thead>
<tr>
<th>CRG Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity/Thing</td>
</tr>
<tr>
<td>Kind</td>
</tr>
<tr>
<td>Part</td>
</tr>
<tr>
<td>Property</td>
</tr>
<tr>
<td>Material</td>
</tr>
<tr>
<td>Process</td>
</tr>
<tr>
<td>Operation</td>
</tr>
<tr>
<td>Patient</td>
</tr>
<tr>
<td>Product</td>
</tr>
<tr>
<td>By-product</td>
</tr>
<tr>
<td>Agent</td>
</tr>
<tr>
<td>Space</td>
</tr>
<tr>
<td>Time</td>
</tr>
</tbody>
</table>

Table I

4. XML mapping of categories

Various projects have investigated XML thesaurus representations (eg Howarth 2000; Mitchell & Vizine-Goetz 2000; VOCML 2000). For our purposes, we wanted a level of representation more general than specific facets in particular thesauri or classifications. We wish to facilitate mapping from (eventually and as far as possible) a common query structure and interface to different thesauri and to different parts of the same thesaurus for different collections or different user profiles. One aim is to be able to shield a user from some of the complexity of a large thesaurus and dynamically map to particular hierarchies for a particular collection or information need. Another aim is to allow a user to express queries using a standard terminology that can be mapped to different thesauri.

An external XML file maps a category to a hierarchy or concept in a particular thesaurus. For figure 1, the XML mapping reproduced the AAT facets. Figure 3 shows the query discussed in Sections 2 and 3 but this time the browser interface is driven by a different XML mapping based on the CRG categories. Whereas before the AAT Components hierarchy was located under the Objects facet, the CRG categories give more prominence to Parts as a top level category and Parts has been mapped to Components. The point is that different mappings are possible for different applications, collections or user profiles. We employed a
subset of the CRG categories but any standard set of categories is possible. The XML representation offers the potential for a Query/Indexing tool to call up just part of a large thesaurus at some stage in the index/search process where browsing might be offered. Another possibility would be for the categories to drive ‘Wizards’ that encouraged users to construct standard faceted expressions of queries or index descriptors.

Figures 4 and 5 show an example of how the mapping of categories can affect retrieval functionality. Figure 4 shows the Find Term utility with no restriction on the scope of the search within the thesaurus (note the detection of compound terms, such as Queen Anne Style). Figure 5 shows the same search with the Parts category selected and resulting shorter list of terms from the mapping to the AAT Components hierarchy. This could be extended to restrict the scope of
automatic traversals in a retrieval system component.

5. Conclusions and Future Work

Thesauri have traditionally been intended to be read by human indexers and searchers, either on computer displays or in print form. The growth of digital collections has resulted in increased opportunities for automated processing of thesaurus representations. A matching function incorporating semantic term expansion can produce ranked results for queries with multiple thesaurus terms where individual terms are expanded and recombined. This may be useful for complex queries on large thesauri where manual browsing may be cumbersome. A semantic distance measure can also assist interactive query refinement. To reflect the move to online thesauri and assist automatic processing, there is a need for revised thesaurus standards which explicitly represent thesaurus structure, since we cannot reply on human interpretation of context. For example, it should be possible to distinguish automatically between Hierarchical Generic and Hierarchical Instance relationships and perhaps different kinds of associative relationship in electronic representations (Tudhope, Alani & Jones, 2001).

The initial experiments with higher level external XML representations of facet structure discussed in this paper are part of a longer term project to model the underlying semantic categories of concepts following a facet analysis approach. In future work, we intend to extend the representation to thesaurus relationships and query/descriptor structure. A cataloguing or retrieval system could have different XML representations for different facet structures. The aim is not to build into the system implementation a dependency on the particular facets of the AAT (for example). Rather, we attempt a more universal approach, by seeking to identify general categories that can be mapped to the facets and hierarchies of the AAT or other thesauri.

Acknowledgements

We would like to acknowledge the support of the UK Engineering and Physical Sciences Research Council (Grant GR/M66233/01) and the support of HEFCW for the Internet Technologies Research Lab. We would like to thank Helen Ashby, Ann Borda, Sarah Norville, Charlotte Stone and other staff from the National Museum of Science and Industry for their assistance and the J. Paul Getty Trust for provision of the AAT.

References
http://www.getty.edu/research/tools/vocabulary/aat/  
http://web.glam.ac.uk/schools/soc/research/hypermedia/facet_proj/index.php


