Structured Access to Scientific Information

Caterina Caracciolo and *Maarten de Rijke*

Abstract

We report on an ongoing project aimed at providing an exemplary architecture for an electronic dissemination environment for scientific handbooks. We focus on our way of facilitating navigation through and access to electronic handbooks by using a WordNet-like concept hierarchy consisting of synsets that are connected to each other and to external sources by semantic relations for navigational purposes.

1 Introduction

Electronic publishing offers many opportunities, for readers, authors, and publishers alike. While technical reports, conference proceedings, and journals are increasingly being made available in an electronic form, sometimes even exclusively, other kinds of scientific publications have, by and large, not been recast for electronic dissemination yet. In particular, for scientific knowledge of a unifying kind, as traditionally found in a handbook, several proposals for a suitable architecture are only now being tried out or explored.

There are many reasons that justify electronic versions of scientific handbooks such as the *Handbook of Logic and Language* (van Benthem and ter Meulen, 1997) or the *Handbook of Automated Reasoning* (Robinson and Voronkov, 2001). It makes distribution easier and quicker, and readers can be helped considerably when searching for information: even simple keyword searches are more useful than scanning tables of contents or indexes, especially for large handbooks, and tracking down a reference can be as simple as a mouse click. Also, electronic publications are less rigid than their paper counterparts: the publication can be adaptable to the reader, thus better satisfying her information needs. Electronic availability also facilitates integration with other media types: e.g., computer simulations and visualizations, movies, and tools.

Electronic books facilitate a more modular way of reading than traditional paper books. Indeed, many web sites consist of many relatively small modules connected by hyperlinks. (Harmsze, 2000) proposes a modular structure for articles in experimental sciences, but it is not clear whether this approach can be adapted to handbooks that contain more abstract content. A potential problem with the ‘small modules – many links’ structuring of information is disorientation of the reader (Conklin, 1987). A reader should know where she is in the hypertext network, and how to get to other locations: high quality navigation tools are essential.

In our approach to the development of an electronic dissemination environment for scientific handbooks, we intend to facilitate navigation through and access to electronic handbooks by using a WordNet-like concept hierarchy. It consists of synsets that are connected to each other and to external sources by semantic relations.

2 Strategy

Topic or concept hierarchies are often used for the purpose of navigating through large collections of documents. They are very useful for the organization, display and exploration of large amounts of information. Well-known examples include Yahoo!’s topic hierarchy for exploring the Web, and Google’s directories (based on the DMOZ open directories initiative).

One of the advantages of using concept hierarchies is that users do not have to know exactly what information they are looking for. Without having to phrase their information need in precise terms, they can browse from general to more specific categories, or from example to counterexample, and thus get a clearer idea of the information being sought. This is especially helpful for readers that are less familiar with the topic for which they consult the handbook.

It has been shown that users in a hypertext search task who had hierarchical browsing patterns through the hypertext performed better than users who had sequential browsing paths (McEneaney, 1999). Therefore, it is very important that architectures for electronic handbook allow, or even enforce, such hierarchical patterns, and a concept hierarchy is a good way of doing this.

Based on these considerations, it was decided to investigate the use of fine-grained concept hierarchies for navigation through and access to scientific handbooks within the *Logic & Language Links* project. To make matters concrete, the project aims to develop an electronic version of the *Handbook of Logic and Language* (van Benthem and ter Meulen, 1997). However, the envisaged results and our discussion below are applicable to many other domains.

3 Organization of the Hierarchy

The building blocks of the concept hierarchy are the concepts, and its cement consists of several semantic relationships. In line with WordNet (Fellbaum, 1998), we make a distinction between words or terms on the one hand, and concepts on the other hand: a concept is denoted by a *synset*, a set of synonymous words. Words are synonymous if they have (more or less) the same meaning in some setting. The semantic relationships come in two kinds: *internal* to concept hierarchy, and ones that link the concepts to *external* resources.

3.1 Internal Architecture

Concepts in the hierarchy are annotated with a gloss: for instance, *The study of language meaning* is a gloss for *semantics*. Moreover, they come with multiple, increasingly more technical descriptions, only one of which will be served to an individual reader, depending on her level of expertise.

The main semantic relationship that structures our concept hierarchy is *related subtopic*, or simply *subtopic*. While it covers the usual ‘is-a-kind-of’ relation (e.g., *epistemic_logic* is a related subtopic of *modal_logic* in this sense), the subtopic relation extends it in a number of ways. For instance, we allow the subtopic relation to cover meronymic cases as well: the relationship between *compactness* and *logic* (in the sense of ‘a system of reasoning’) is of this kind.

We don’t require that the concept hierarchy be a strict tree. Except the root, every node may have multiple parents. We do not allow cycles in the subtopic hierarchy, since cycles disorientate readers. Moreover, we don’t allow any concept to be unconnected to the rest of the hierarchy. At present, every concept is below one of four ‘beginners’ (*computer science*, *mathematics*, *linguistics*, *philosophy*).

As to additional (non-hierarchical) navigational relations, these include the following:

Similarity: concepts are similar if they share some properties or are somehow analogous to each other. For instance, *finite state machine* is similar in this sense to *regular language*.

Antonymy: learning the antonym of a concept not only teaches us more about the meaning of the antonym, but also about the concept itself (Muehleisen, 1997).

Sibling: informal experiments have convinced us that readers find it useful to know what the siblings of a given concept are; it helps prevent the ‘lost in space’ problem.

Other meanings: for every name in the synset of a given concept, we provide links to other concepts in whose synset the name string occurs.

3.2 External Connections

In addition to the internal links, our concept hierarchy has *external* links in the sense that they are between concepts and targets outside the hierarchy. We distinguish between *handbook* links (to information in the handbook but outside the concept hierarchy), and *web* links (to information sources on the web). Again, we make this distinction for the benefit of the reader; we have found that it is important for a reader to know whether a link target is outside the space controlled by us.

The target of a handbook link can be of different levels of granularity (a part, a chapter, a subsection, a definition, etc.). Ideally, concepts higher in the hierarchy refer to larger fragments in the handbook, while lower concepts refer to smaller parts. However, as the handbook chapters are written by different authors, resulting in a different structuring and writing style for every chapter, this is hard to achieve.

Internal and external links in the concept hierarchy take advantage of a set of metadata that is automatically generated. Internal links are established using the references given by the unique identifier associated to each node.

Handbook links come with metadata describing crucial information about the publication linked (e.g., author, editor, publisher), enriched with an indication of the link type (e.g., definition, theorem, discussed-in, example/counterexample). As to web links, they too come in a small number of types, including research group, home pages, tools (links to software related to the concept; for example, *information_retrieval* has a link to the `mg` system), and publications.

We mostly adhere to the Dublin Core, even though in our approach a large part of the metadata suggested by the Dublin Core plays the role of actual data. Even data about authors and editors of the concept hierarchy is meant to be available as user user. Data about creation/modification of concepts in the concept hierarchy has a somehow different role. At present it is metadata, but in a future scenario it will be user data as well.

4 Building the Hierarchy

The *Logic & Language* concept hierarchy is currently being built, by hand. Our efforts are based on *The Bluffer's Guide to Computational Semantics* (Fracas, 1996), and the glossary of the *Handbook of Logic and Language* (Groeneveld, 1997).

At present the focus of our work on constructing the hierarchy is on creating the hierarchical relationships, complete with glosses and internal relations; extensive descriptions and external links have mostly been left out at this stage. Domain experts at the authors' home institute are about to be involved in the process of building the hierarchy as large-scale community building effort.

4.1 The Current Prototype

The current version of our hierarchy is populated with close to 1000 concepts, provided by us. For every concept we maintain a single XML file. From these XML files a *Logic & Language* web site is generated at regular intervals, to incorporate changes to the underlying XML files.

We are currently setting up a web-based interface to enable domain experts to easily add further concepts or modify existing ones. Editors are being approached to take responsibility for subparts of the hierarchy (such as *computational_logic* or *dynamic_semantics*). We plan to launch a version of the *Logic & Language* concept hierarchy on a publicly accessible web server in early 2002.

4.2 Support Tools: Hierarchy Development

In our approach it is essential that the concept hierarchy be constructed by a team of human editors to guarantee high quality. Nevertheless, this activity is time consuming and error prone. For this purpose, in constructing the hierarchy and in linking it, we intend to provide authors with tools that make suggestions and check for coherence of the inserted data.

Completeness and *correctness* are two important criteria in our hierarchy development efforts: the concepts in the hierarchy should cover the information in the domain covered by the handbook, and they should only cover information in that domain. We have carried out two kinds of experiments to help us ensure these criteria. First, we have used ideas based on inverse document frequencies of terms in collections of arbitrary scientific papers versus collections of papers in logic and language area.

Second, we have explored methods for automatically generating concept hierarchies. Research on the latter comes in three flavors: pattern matching, based on partial parsing, and based on statistics and cooccurrence. Well-known work on using methods based on pattern-matching for extending WordNet were described by (Hearst, 1998), while (Manning, 1993) is an example of work based on partial parsing. (Sanderson and Croft, 1999) aim to generate a hierarchy with the same subtopic relation we employ in our concept hierarchy, a mixture of hypernymic and meronymic relationships. In response to a suitable query, they consider the set of 500 top-ranked documents. From these a term collection is built based on similarity to the query, which is then used to build the hierarchy.

We have carried out small-scale experiments with the Sanderson and Croft algorithm. To be applied to the construction of a concept hierarchy for the Logic & Language area, some adaptations had to be made. First, we needed a sufficiently large corpus of papers in the area. Second, we do not have a query whose terms can be used as potential concepts in the hierarchy; we simply use hierarchy terms that are already present.

Having automatically generated a hierarchy, we may be faced with the need to merge (parts of) it with the existing hand built one. Recent work on ontology development and enlargement, such as Chimaera (McGuinness et al., 2000) and Prompt (Noy and Musen, 2000), is particularly relevant to us for this purpose.

4.3 Support Tools: Linking the Hierarchy

Finally, let's turn to the task of generating links from the concept hierarchy, which is another natural task begging for automation. So far, we have experimented with automatically generating hypertext links from concepts in the hierarchy to (electronic versions of) the chapter in the handbook. We used the vector space model, exploring a variety of options. As the documents to be retrieved, we have taken pages of the original handbook; while arbitrary, this choice was forced upon us by the diversity of the writing styles of the contributing authors. Some preliminary experiments indicated that cosine similarity provides the best weighting scheme for this setting, with normalization for the queries, but not for the documents. For the queries we explored several possibilities (term, term plus description, term and description plus additional weightings on the term). We have found that best retrieval results were obtained by factoring the key terms (taken from the concept hierarchy) with a higher constant than the descriptions of the terms (Monz et al., 2000).

5 Conclusion

We have reported on ongoing work aimed at providing an exemplary architecture for an electronic dissemination environment for scientific handbooks. We focused on facilitating navigation through and access to electronic handbooks by means of a WordNet-like concept hierarchy consisting of synsets connected to each other and to external sources by various semantic relations.

Acknowledgments This research was supported by Elsevier Science Publishers. The second author was supported by the Spinoza project 'Logic in Action' and by grants from the Netherlands Organization for Scientific Research (NWO), under project numbers 612-13-001, 365-20-005, 612.069.006, 612.000.106, and 220-80-001. We thank Joost Kircz and the referees for helpful comments and suggestions.

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION
UNIVERSITY OF AMSTERDAM
NIEUWE ACHTERGRACHT 166
1018WV AMSTERDAM
THE NETHERLANDS
{caterina,mdr}@science.uva.nl

References

J. Conklin. 1987. Hypertext: An introduction and survey. *IEEE Computer*, pages 17–41.

C. Fellbaum, editor. 1998. *WordNet, an Electronic Lexical Database*. MIT Press.

L.R.E. Fracas. 1996. The bluffer's guide to computational semantics, January.

W. Groeneveld. 1997. Logic and language: A glossary. In (van Benthem and ter Meulen, 1997), pages 1179–1213.

F. Harmsze. 2000. *A Modular Structure for Scientific Articles in an Electronic Environment*. Ph.D. thesis, Universiteit van Amsterdam.

M.A. Hearst. 1998. Automated discovery of WordNet relations. In (Fellbaum, 1998), pages 131–151.

C.D. Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proc. 31st ACL*, pages 235–342.

J.E. McEneaney. 1999. Visualizing and assessing navigation in hypertext. In *Proc. 10th ACM Conference on Hypertext and Hypermedia*, pages 61–70.

D.L. McGuinness, R. Fikes, J. Rice, and S. Wilder. 2000. The Chimaera ontology environment. In *Proc. AAAI 2000*.

C. Monz, J. Ragetli, and M. de Rijke. 2000. Concept-based computer-aided link generation for electronic handbooks. In *Proc. DIRW'00*.

V.L. Muehleisen. 1997. *Antonymy and Semantic Range in English*. Ph.D. thesis, Northwestern University.

N. Fridman Noy and M.A. Musen. 2000. Prompt: Algorithm and tool for automated ontology merging and alignment. In *Proc. of AAAI 2000*.

J. Robinson and A. Voronkov, editors. 2001. *Handbook of Automated Reasoning*. Elsevier.

M. Sanderson and B. Croft. 1999. Deriving concept hierarchies from text. In *Proc. SIGIR'99*, pages 206–213.

J. van Benthem and A. ter Meulen, editors. 1997. *Handbook of Logic and Language*. Elsevier.