Organizing and accessing online handbooks

Caterina Caracciolo and Maarten de Rijke

We report on an ongoing project aimed at defining an electronic environment for the digitalization and dissemination of scientific handbooks. Such an environment is expected to combine agile organization of information and ease of retrieval; moreover it can be useful as a didactic tool as much as a research instrument.

In order to fulfill these requirements, we opted for a WordNet−like conceptual hierarchy of the subjects, for navigation through and access to scientific handbooks. It has been shown that, when committed with a hypertext search task, users perform better in a hierarchical browsing pattern (through the hypertext), than in a sequential one (McEneaney, 1999). Therefore it is very important that architecture for an electronic handbook makes use of a hierarchical pattern.

To make things concrete, we focus on an architecture of the fields of linguistic and logic, using the 'Handbook of Logic and Language' (van Benthem and ter Meulen, 1997) as primary landmark, then extending our work to cover other handbooks in the fields.

Our WordNet−like conceptual hierarchy consists of a collection of synsets, internally connected to each other, and externally connected to other sources by means of a variety of semantic relations introduced for navigational purposes. In line with WordNet (Felbaum 1998), we make a distinction between words and 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 the same meaning in a certain setting. In order to facilitate the reader, each concept−synset comes with a gloss; for instance, one of the concepts labeled with *semantics* has *the study of language meaning* as its informal gloss.

The backbone of the hierarchy is the relation of *related subtopic*, or simply *subtopic*: it covers the usual 'is−a−kind−of' relation but can also be extended to cover other cases, like the meronymic relation. Moreover, also domain specific relations can be included in the subtopic relation, for instance the 'is−a−theorem−of' relation is especially defined for the field of logic. A node can have more than one parent (supertopic), but cycles are not allowed, in order not to disorientate the reader.

Subtopic relations are manually defined by human authors. We are aware of automatic systems for classification and extraction of hierarchies from texts, but we believe that a manually populated hierarchy can give us a more meaningful tool: its structure can be easily generalized to other fields, while its contents are

accurate enough for scientific and didactic purposes.

Other non hierarchical relations are introduced in order to facilitate the navigation within the hierarchy and give the reader a better insight into the subject. In fact, informal experiments have convinced us that, when examining a concept, the reader benefits from having explicitly given siblings (subtopics of the same topic), similar concepts (sharing some properties), antonym concepts (having opposite meaning) and homonym concepts (concepts sharing one or more words in their synset). Some of these relations are defined by the authors, while others are automatically extracted; an interesting point in study is the definition of one or more automatically computable relations of similarity, to be added to the ones manually pointed out. We pay great attention to similarity relations because, according to our personal experience, an important part of the learning process consists of establishing connections with similar concepts, likewise learning about the antonym of a concept also teaches us about the concept itself.

One of the advantages of having a navigable hierarchy is that the reader can look up for a concept without phrasing his/her information's need. This is especially helpful for readers that are least familiar with the subject. For example s/he can browse the hierarchy by going from generic to more specific categories, or by examining a succession of similar nodes. The reader can also submit queries and rely on a highly interactive environment in order to identify the right concept, or highlight only some of the available relations. Moreover, since concepts are provided with multiple descriptions, of increasing technical complexity, the reader can choose the level of detail that better fits his/her need.

In addition to the internal relations in the hierarchy, we also provide external links: namely handbook−links and web−links. We make this distinction for the benefit of the reader, because we have found that it is important for the reader to know what kind of source the link leads to. Links to the web allow the hierarchy to be enriched with the most recent results and with different kind of sources (e.g. audio, video, database of different nature, theorem prover).

Handbook−links connect the hierarchy to the handbook: they are computed by an information retrieval system especially tailored to cope with a handbook−style text. Because a handbook tends to be divided in long chapters (sections, subsections) often overwhelming the reader, we expect that the information retrieval system points out passages with different bounds than in the original text. Here the issue is to define optimal levels of granularity and modularity of the extraction of passages.

After having defined the structure of the concept hierarchy, we are now populating it and creating web−based tools to enable domain experts to add further concepts or modify existing ones. For each node, we maintain a single XML file; from these XML files a *Logic and Language* web site is generated at regular intervals, to incorporate changes to the underlying XML files. During the second stage of the project, we plan to define the information retrieval system, and evaluate its effectiveness in generating handbook−links. We plan to elaborate tests, to be submitted to three groups of users: students, teachers and researchers. Since those groups have very specific needs, we expect to get important feedback from them regarding: (a) the capability to prevent the user from falling in the feeling of "lost in hyper space", especially dangerous for students; (b) the usefulness of a modular organization of information in order to define a teaching programme, (c) the possibility of offering a peers−to−peers exchange among researchers, and the flexibility of the electronic medium to accommodate various sources.

Bibliografia

J.E. McEneaney, 1999. Visualizing and assessing navigation in hypertext. In Proceedings 10th ACM Conference on Hypertext and Hypermedia, pg. 61−70.

C. Felbaum, (ed.) 1998. WordNet: an Electronic Lexical Database. MIT Press.

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