OntoSeek: Using Large Linguistic Ontologies for Accessing On-Line Yellow Pages and Product Catalogs

Nicola Guarino
National Research Council, LADSEB-CNR, Corso Stati Uniti 4, I-35127 Padova, Italy
Nicola.Guarino@ladseb.pd.cnr.it
Claudio Masolo
University of Padova, Dept. of Electronics and Computer Science, Italy
Claudio.Masolo@ladseb.pd.cnr.it
Guido Vetere
IBM, Rome Tivoli Labs, Roma, Italy
gvetere@tivoli.com

Introduction

To exploit effectively the mass of information available today on the Web, the key problem is that of *content matching*: the relevant information must be selected according to the user needs, independently of the vocabulary and the syntax used to express it. Content matching seems to be an intrinsic problem for textual documents or web pages: current information retrieval techniques either rely on an encoding process that describes a given item according to a certain perspective or classification scheme, or perform a full-text analysis based on the search for user-specified words. Neither case guarantees content matching, because an encoded description may reflect only part of the content, and the mere occurrence of a word (or even a sentence) does not necessarily reflect the document's content.

For general documents, there doesn't yet seem to be a much better option than some sort of lazy full-text analysis, leaving us to sift through endless result pages. There is however a relevant class of information repositories–on-line yellow pages and product catalogs–where content matching can be both feasible and crucial.

In this paper[1], we first analyze the peculiarities of these repositories with respect to generic Web documents, and then we discuss the role that current linguistic ontologies like WordNet (Miller, 1995) can play to support content matching. We then present the architecture of a system called OntoSeek, specifically targeted to on-line yellow pages and product catalogs. The system is the result of a two-year cooperation between CORINTO (national research consortium for object technology, a partnership of IBM Semea, Apple Italia, and Selfin Spa) and LADSEB-CNR, as part of a project on retrieval and reuse of object-oriented software components (Borgo et al., 1997). OntoSeek adopts a language of limited expressiveness for content representation, and exploits a large linguistic ontology based on WordNet (namely SENSUS, developed at ISI-USC) for content matching. In general, with respect to standard word-matching systems, expressing the content structure by means of a simple representation language increases the precision of the retrieval, while adopting a hierarchy of keywords increases both recall and precision. In OntoSeek, the use of a linguistic ontology results in two further advantages: a decoupling between the user vocabulary and the encoding terminology, and an additional increase of recall and precision due to synonymy handling and sense disambiguation. Our conclusion is that yellow pages and product catalogs constitute a strategic *niche*, where retrieval techniques based on simple representation capabilities and large linguistic ontologies appear to be particularly effective.

The peculiarities of yellow pages and product catalogs

The peculiarities of yellow pages and product catalogs with respect to generic Web documents are reported in Table 1. If a repository addresses a huge domain, the vocabulary size is necessarily high: this is true for yellow pages, Web documents and heterogeneous product catalogues. Product catalogs employ a technical and detailed vocabulary and have a high description complexity due to the presence of a large number of semantic relations and constraints–more so than yellow pages. Despite the high variability of each description's content, description structure is fixed for homogeneous product catalogs, because they use the semantic relations to form each description. So, description heterogeneity is low for homogeneous catalogs, but high for heterogeneous catalogs. Only yellow pages have high query specificity, because in other cases users do not need to know all of a desired product's characteristics in advance.

[1] The full version of this paper is going to appear on *IEEE Intelligent Systems* (in press)

Parameter	*Yellow Pages*	*Homogeneous Product Catalogs*	*Heterogeneous Product Catalogs*	*Web Documents*
Vocabulary Size	Moderate/High	Moderate	High/Very High	Very High
Description Complexity	Low	Moderate/High	Moderate/High	Very High
Description Heterogeneity	Low	Low	High	Very High
Query Specificity	High	Low/Moderate	Low/Moderate	Very Low

Table 1. Yellow pages and product catalogs compared to generic Web documents. The four categories of resources, from left to right, correspond to increasing levels of difficulty of content-based information retrieval. We distinguish between homogenous and heterogeneous product catalogs based on whether they describe single or multiple product lines.

OntoSeek's basic capabilities

In knowledge retrieval systems, a good ontology can significantly increase both recall and precision. However, especially for yellow pages and product catalogs, three main factors limit the practical adoption of ontologies:

- Data's intrinsic dynamics requires a continually updated ontology to keep track of new terms;
- both the query and encoding process crucially depend on understanding a rigid set of terms;
- the vocabulary size and descriptions' heterogeneity require a broad coverage ontology.

OntoSeek overcomes these limitations thanks to the use of a large linguistic ontology such as WordNet. It has been specifically designed for yellow pages and on-line catalogs, and combines an ontology-driven content-matching mechanism with a moderately expressive representation formalism. Its main characteristics can be summed up as follows:

- Option to use arbitrary natural language terms for accurate resource descriptions in the encoding phase.
- Complete terminological flexibility for the queries, thanks to ontology-driven semantic match between queries and resource descriptions
- Interactive assistance on query formulation, disambiguation, generalization, and specialization.
- High precision, thanks to structured representation of queries and descriptions and interactive sense disambiguation;
- High recall, thanks to synonymy handling and hierarchy exploitation
- State-of-the-art Internet architecture.
- Good scalability and portability.

The system is designed to handle both homogeneous and heterogeneous product catalogs. As we have seen from Table 1, the latter are more difficult to handle than yellow pages, mainly because of their higher description complexity and heterogeneity. For this reason a structured representation formalism based on a simplified version of conceptual graphs has been adopted for queries and resource descriptions. Compared to simple attribute-value lists, these conceptual graphs are extremely more flexible and significantly more expressive, although the expressiveness is still rather moderate with respect to, for instance, current versions of description logics. With conceptual graphs, the problem of content matching reduces to ontology-driven graph matching, where individual nodes and arcs match if the ontology tells that a subsumption relationship holds between them. However, in order to exploit the advantages of a linguistic ontology, we need to make sure that the graphs can be actually linked to it. This is achieved by constraining their labels to be *lexical items*, and introducing suitable semantic constraints. We call these graphs Lexical Conceptual Graphs (LCGs), as they are simplified variants of Sowa's conceptual graphs (Sowa, 1984). An interesting technical aspect of LCGs is the way of labeling binary relations. Usually, people are forced to invent for them ad-hoc labels like "part-of" or "has-function". In OntoSeek, these labels are replaced with ordinary nouns like "part" or "functions", which however have a special *relational interpretation* (Guarino, 1992). In this way we always guarantee a *lexical handle* to interpret the meaning of our descriptions according to the linguistic ontology.

In conclusion, we point out that linguistic ontologies offer immense potential for gathering information resources from the Web. Just taken as they are in their present status (i.e., with their poor ontological structure) they can provide substantial improvements to current search systems. Converting them into understandable, clean, and coherent ontologies suitable to drive future information-search systems is not a trivial task, but one well worth addressing.

Bibliography

Borgo, S., Guarino, N., Masolo, C., and Vetere, G. 1997. Using a Large Linguistic Ontology for Internet-Based Retrieval of Object-Oriented Components. In *Proceedings of 1997 Conference on Software Engineering and Knowledge Engineering*. Madrid, Knowledge Systems Institute, Snokie, IL, USA: 528-534.

Guarino, N. 1992. Concepts, Attributes and Arbitrary Relations: Some Linguistic and Ontological Criteria for Structuring Knowledge Bases. *Data & Knowledge Engineering*, **8**(2): 249-261.

Miller, G. A. 1995. WORDNET: A Lexical Database for English. *Communications of ACM*, **2**(11): 39-41.

Sowa, J. 1984. *Conceptual Structures: Information Processing in Mind and Machine.* Addison-Wesley, Ready, Massachussets.