The use of Description Logics in the Condorcet conceptual information retrieval system

Jeroen Nijhuis
Knowledge-Based Systems Group
Faculty of Computer Science
University of Twente
Enschede, The Netherlands
jnijhuis@cs.utwente.nl

Introduction

This position paper will describe the intended use of *Description Logics (DLs)* in a particular conceptual information retrieval system, the Condorcet system. First the Condorcet system is introduced, focusing on knowledge representational and reasoning aspects and criteria. Second, based on these requirements, we will argue that DLs are particularly promising candidates for knowledge representation and reasoning in a "real" and knowledge-intensive application like the Condorcet system. Last I will summarize my interest in DLs and the DL workshop.

The Condorcet project

The Condorcet project, carried out by six members of the Knowledge-Based Systems group at the University of Twente, aims at developing an automatic conceptual information retrieval system. The Condorcet project is funded by the Technology Foundation (STW). This Dutch foundation finances application-oriented research. The Condorcet project should deliver in 1999 a prototype of a conceptual information retrieval system designed to handle 30000 documents and tested on two document collections: abstracts of scientific articles about the treatment of epilepsy and abstracts about the mechanical properties of ceramic materials. At this moment a small demo exists for the ceramic materials domain. We are extending this demo to capture also the epilepsy domain. BACK is used to represent the needed knowledge.

A *Conceptual Information Retrieval System (CIRS)* assigns *index concepts* to documents out of some document collection, in such a way that they express what the document is about. These index concepts are not directly taken from the document's text, but are instead retrieved from a conceptual system. Rather than using flat index concepts we will use *structured* index concepts, structured in the sense that index concepts can be related to each other by what we will call a *coordinator*[1]. For example a document which is about the treatment of epilepsy by phenytoin should not only be indexed with the concepts *phenytoin* and *epilepsy*, but also by the structured index concept *treats(phenytoin,epilepsy)*. In order to be able to assign this kind of index concept automatically, the conceptual system should have a clear and formal semantics. All possible index concepts are specified in the conceptual system by defining concepts and the possible coordinators between them.

The conceptual system specifies the structured index concepts that are possible. These terms however, have to be 'discovered' in the abstracts of the documents. In Condorcet an extensive *syntactic and semantic analysis* is performed. In order to assign the appropriate structured index concepts in the semantic analysis domain knowledge has to be applied. At this moment we are investigating what this knowledge should be. We assume that the shallow domain knowledge introduced by defining the index concepts is not enough for a satisfactory semantic analysis of the abstracts. For example, if it is unclear which property of a certain material is described in an abstract, the measurement unit named in the abstract can be used to decide which index concept has to be assigned. The measurement unit however, shall not be used as an index concept and therefore it is not defined in the conceptual system. We will need deep knowledge of the domain too. To separate this two kinds of knowledge we introduce the *index concept knowledge base* and the *domain knowledge base*. The former contains the definitions of the index concepts, for example *treats* is restricted to coordinate only concepts of type *medicine* and *disease*. The domain knowledge base contains domain knowledge to be used in the semantic analysis.

Another needed knowledge base is a knowledge base containing *background knowledge* e.g. knowledge to translate non-standard measurement units to standard ones.

The matching process compares a query with index concepts assigned to the documents. The result of this comparison is a set of documents which are supposed to be relevant to the query. Apart from the semantic analysis, the *matching process* uses the index concept knowledge base too e.g. if an index concept indicates that a document is about phenytoin, then it should be able to infer that this document is also about the more general concept medicine.

In figure 1 an outline is given of the Condorcet system.

[1] A term borrowed from library science.

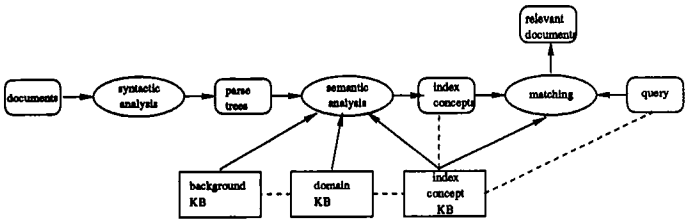

Figure 1: The Condorcet information retrieval process. The (sub)processes and KBs are represented by ovals and boxes. The horizontal arrows describe the flow through the indexing subprocesses. The other arrows illustrate the support of the KBs. The dotted lines indicate relations between various KBs.

Description Logics and the Condorcet system

To represent the index concepts and the definitions of the index concepts, a formalism is needed which is expressive enough and has appropriate inferences like subsumption. Description Logics seem to be very promising candidates: (1) they are sufficiently expressive concept languages, (2) they have a clear and sound theoretical foundation, (3) the computational properties of the inference algorithms are extensively studied, and (4) a number of DLs are implemented into software systems.

These are all requirements for the Condorcet application. (1) In the previous section I have explained that we need an expressive concept language. (2) In order to maintain large knowledge bases (remember the system should be able to handle 30000 documents) the semantics of the knowledge structures have to be clear [Speel, 1995]. (3) Furthermore, to guarantee reliable performance, the computational properties of the inference algorithms have to be known. This is especially evident in the matching process: in case the Condorcet system will be used on-line a user should get an answer to a query within a few seconds. Given the size of the knowledge bases this will put a severe constraint on the knowledge representation system. (4) The time-constraint of the deliverance of the prototype requires that we reuse as many existing resources as possible. A great advantage of DLs is the availability of many implemented knowledge representation system based on them.

Workshop interest

The main reason for my interest in the DL workshop is the possibility to discuss the use of DLs in conceptual information retrieval systems like Condorcet. To be more specific, the use of DLs in knowledge-intensive applications (e.g. [Borgida, 1995], [Doyle and Patil, 1991], [Speel, 1995] etc.), the use of DLs in semantic analysis (e.g. [Quantz *et al.*, 1995]), the use of DLs in information retrieval modeling (e.g. [Meghini *et al.*, 1993], [Buongarzoni *et al.*, 1995]) and the combination of those.

[Meghini *et al.*, 1993] use one DL, MIRTL, to model the queries, the index terms, the index terminological system, and the matching process. An important advantage of using one DL is that it is ensured the knowledge bases will participate in the retrieval process in a uniform fashion [Meghini *et al.*, 1993]. In the Condorcet project, however, more knowledge bases have to be used. The trade-off between expressiveness and tractability comes in two guises in the Condorcet project: (1) high-expressiveness in the domain KB - not too demanding on worst-case tractability, and (2) relative low-expressiveness in the index concept KB - high performance and reliability. This is a design discussion I would like to discuss: stick to one DL language to represent all knowledge since this knowledge is highly interrelated or use more languages to capture the ontological, expressiveness and performance differences between the knowledge bases.

References

[Borgida, 1995] A. Borgida. Description logics in data management. *IEEE transactions on knowledge and data engineering*, 7:671–682, 1995.

[Buongarzoni *et al.*, 1995] P. Buongarzoni, C. Meghini, R. Salis, F. Sebastiani, and U. Straccia. Logical and computational properties of the description logic MIRTL. In *International workshop on Description Logics*, pages 80–84, Roma, Italy, June 1995.

[Doyle and Patil, 1991] Jon Doyle and Ramesh S. Patil. Two theses of knowledge representation: language restrictions, taxonomic classification, and the utility of representation services. *Artificial Intelligence*, 48:261–297, 1991.

[Meghini *et al.*, 1993] C. Meghini, F. Sebastiani, U. Straccia, and C. Thanos. A model of information retrieval based on a terminological logic. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, pages 298–307, Pittsburgh, PA, 1993.

[Quantz *et al.*, 1995] J. Joachim Quantz, Guido Dunker, Manfred Gehrke, Uwe Kuessner, and Birte Schmitz. FLEX-based disambiguation in VERBMOBIL. In *International workshop on Description Logics*, pages 112–118, Roma, Italy, June 1995.

[Speel, 1995] Piet-Hein Speel. *Selecting knowledge representation systems.* PhD thesis, Universiteit Twente, 1995.