Distributed Search in the Semantic Web

Umberto Straccia
ISTI-CNR, Pisa, ITALY
Umberto.Straccia@isti.cnr.it

The *Semantic Web*[1], is widely regarded as the next step in the evolution of the Web. It aims at enhancing content on the Web with meta-data, enabling *agents* (machines or human users) to *process*, *share* and *interpret* Web content. *Ontologies* will play a key role in the Semantic Web. They provide a source of shared and precisely defined terms (using an *ontology language* [2]) that can be used in such meta-data. An ontology consists of a conceptual schema of a domain, e.g. a hierarchical description of important concepts, along with the description their properties.

As building an ontology is an expensive process, the added value of semantic annotation over an ontology should compensate, in terms of better retrieval effectiveness, the enormous labour to construct it. While the construction of an ontology may accepted to be manual, semantic annotation *should be automatic in the long run* [6].

However, providing the Internet with more capabilities of processing and understanding the semantics of information will not be sufficient to bring Semantic Web to its full potential. In particular, the way in which information is accessed on the Internet has to undergo significant changes. Indeed, today's searching on networks mostly relies on *centralized* systems, which have the limitations in terms of coverage and freshness of Web information. Much rather, an agent would like to search among those information sources that hold *relevant* information *directly* and *immediately* [3]. This task is called *distributed search*.

Our objective of our work program is to address the issue of distributed search in the context of the Semantic Web, where we assume that an agent may have access to a large number of heterogeneous and distributed information sources. In order to effectively cope with such masses of knowledge, the task of distributed search may be defined in terms of the following sub-tasks. Assume that the agent A has to satisfy an information need Q_A expressed in a query language $\mathcal{L}_A$, whose basic terms belong to an ontology O_A, defined using the ontology language $\mathcal{O}_A$. Assume that there are a large amount of Web sources $\mathscr{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$ accessible to A, where each Web source $\mathcal{S}_i$ provides access to its Web pages by having its own ontology O_i, ontology language $\mathcal{O}_i$ and query language $\mathcal{L}_i$. Then the agent has to perform the following steps: (i) the agent has to *select* a subset of *relevant* sources $\mathscr{S}' \subseteq \mathscr{S}$, as it is not reasonable to assume to access to and query all sources (*source selection*); (iii) for every

[1] www.semanticweb.org

[2] E.g., DAML+OIL [10], RuleML [2] and the OWL [5]. Their semantics is model-theoretic with close relationships to *Description Logics* and their combination with Logic Programming [7]. This has many advantages: they are well-established, well-understood, computational complexity of reasoning in it is known and implemented systems exists.

[3] This is likely a similar desiderata in so-called Peer-to-Peer networks [1].

selected source $\mathcal{S}_i \in \mathscr{S}'$ the agent has to *reformulate* its information need Q_A into the query language $\mathcal{L}_i$ provided by the source (*schema mapping*); (iii) the results from the selected sources have to be merged together (*data fusion*). As information sources continue to proliferate, these problems of source selection, schema mapping and data fusion become major obstacles to information access. This is an ineffective manual task for which accurate automated tools are desired. That is, an agent must know and automatically learn *where* to search, automatically learn how to query different sources, and how to combine information from diverse sources. Our vision is that any successful solution to distributed search in the Semantic Web, should envisage a fully *automatic* process in the *large scale*.

Our aim is to transfer the solutions proposed to the problem of distributed search in the context of Information Retrieval (IR), where *keywords based search is supported only*, to the Semantic Web. Investigations addressed the problem both globally [11], as well as locally in terms of its sub-tasks (source-selection [3]; schema mapping [8]; data-fusion [9]). In IR, both the automated source selection problem and the schema-mapping problem are highly correlated and are based on the so-called *query-based source sampling methodology* [4]. This method consists in computing automatically an approximation of the content of a source, relying on a sampling technique. In automated source selection, this approximation is used then to decide whether a source may contain relevant information or not with respect to the agents' information need [3], while in the schema mapping problem, this information is used in order to establish automatically *uncertain* mappings, between the agents' query language and the query language of the source [8].

References

[1] P. Bernstein, F. Giunchiglia, A. Kementsietsidis, J. Mylopoulos, L. Serafini, and I. Zaihrayeu. Data management for peer-to-peer computing: A vision. In *In WebDB 2002*, 2002.

[2] H. Boley, B. Grosof, M. Sintek, S. Tabet, and G. Wagner. RuleML design. http://www.dfki.uni-kl.de/ruleml/indesign.html.

[3] J. Callan. Distributed information retrieval. In W.B. Croft, editor, *Advances in Information Retrieval*, pages 127–150. Kluwer Academic Publishers, Hingham,MA, USA, 2000.

[4] J. Callan and M. Connell. Query-based sampling of text databases. *ACM Transactions on Information Systems*, 19(2):97–130, 2001.

[5] M. Dean, D. Connolly, F. van Harmelen, J. Hendler, I. Horrocks, D.L. McGuinness, P.F. Patel-Schneider, and L.A. Stein. OWL web ontology language 1.0 reference. http://www.w3c.org/TR/owl-ref.

[6] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J.A. Tomlin, and J.Y. Zien. SemTag: and Seeker: Bootstrapping the semantic web via automated semantic annotation. In WWW-03, 2003.

[7] B. Grosof and I. Horrocks. Description logics programs: Combining logic programs with description logics. In WWW-03, 2003.

[8] H. Nottelmann and N. Fuhr. Combining DAML+OIL, XSLT and probabilistic logics for uncertain schema mappings in MIND. In ECDL-03, 2003.

[9] M. E. Renda and U. Straccia. Web metasearch: Rank vs. score based rank aggregation methods. In ACM SAC-03, 2003.

[10] F. van Harmelen, P.F. Patel-Schneider, and I. Horrocks. A model-theoretic semantics for DAML+OIL. http://www.daml.org/2001/03/model-theoretic-semantics.html.

[11] C. Yu, W. Meng, King-Lup, W. Wu, and N. Rishe. Efficient and effective metasearch for a large number of text databases. In CIKM-99, 1999.