The Semantic Web Workshop at the 11th International WWW Conference (WWW-2002)

Martin Frank
–
Information Sciences Institute of the University of Southern California, Marina del Rey, California 90292 USA
frank@isi.edu

Natalya F. Noy
Stanford Medical Informatics
Stanford University
Stanford, CA, 94305, USA
noy@smi.stanford.edu

Steffen Staab
Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany
sst@aifb.uni-karlsruhe.de

The original goal of the World-Wide Web was to help *people* find information in heterogeneous hypertext systems on the network. Nowadays, however, a large fraction of the information on the Web, such as "web pages" created as a result of queries to database system, is hardly for human consumption. The amount of data represented in this way far outweighs static HTML/XML documents. However, software agents today use this information in a very limited way performing full text information retrieval or scraping the data from HTML. Alternatively, in the *Semantic Web* machine agents can roam and find information in a way that is understandable to them.

This idea of the Semantic Web has become very popular recently. The Spring and Summer of 2002 alone saw more than a dozen Semantic Web conferences and workshops. Therefore, there was no lack of venues for presenting new research results. However, there was the general feeling that we were not dealing with a community, but rather with a broad range of researchers attracted by the topic but coming from very different backgrounds. Naturally, they frequently lack the knowledge or understanding of the state-of-the-art research in other disciplines that are relevant to the Semantic Web. Therefore, our main question was 'What should the Semantic Web look like?" We organized the Semantic Web workshop at the 11th International WWW Conference in Hawaii in May 2002 primarily as a forum for discussing this question. Our goal was to gather researchers and practitioners in the area of the Semantic Web, as well as representatives from other related fields and to have a one-day discussion on what can we build on and what do we need to do in order to have the Semantic Web deliver on its promises. We also solicited research and application papers to have presentations of current ideas in the field.

The workshop generated a lot of interest: There were 55 registered participants, by far the most-attended workshop at the conference. Seventeen research papers and 11 position statements were submitted. The program committee selected ten papers for publication in the proceedings. The proceedings of the workshop are available at http://www.ceur-ws.org/Vol-55/

The workshop had two invited talks, by Rakesh Agrawal (IBM Almaden Research Center) and Mike Uschold (The Boeing Company). The birds-of-feather sessions discussing possible killer applications for the Semantic Web were the focal point of the workshop. Authors of accepted papers had the opportunity to present their research in a brief teaser session to the overall auditorium and in a very lively interactive poster session. The authors of the paper that received the highest rating from reviewers presented their paper in the main session.

Invited Talks

Rakesh Agrawal talked about "Building Blocks for the Semantic Web" based on his database and machine-learning background. The building blocks ranged from efficient storage and querying up to the integration of different taxonomies. He shared insights about efficient access to Semantic Web-like data, the data that has a large number of different relations, but sparse data entries. The ability to access such data efficiently was especially encouraging since it fundamentally questioned a widely held, but possibly erroneous, believe that the Semantic Web cannot scale up to the dimensions of the

World-Wide Web today. Only little adaptation of standard database technology to this new type of data patterns provides a lot of scalability. Agrawal's conclusion was that Semantic Web can fruitfully exploit these and many other techniques. To reach its full potential, the Semantic Web will need automated techniques with performance similar to that of humans (both of which are far from perfect). However, the Semantic Web will really take off when we take advantage of the core driver of the Web—participation on a large scale.

Mike Uschold based his talk on his long research experience with ontologies, focusing on the semantic continuum that exists between lean semantics on one extreme and heavy ontologies on the other.[1] He predicted that the Semantic Web will evolve by moving along the semantic continuum: we are already seeing applications exploiting light-weight ontologies; more and more we will see machine processing of formal semantics which will increase the range of tasks that agents can perform on the Web.

He also described what different people mean by "semantic integration" of two independently written agents, and classified the possible architectures for such semantic integration based on who generates the agent-to-agent mappings, on when the mappings are generated, on the topology of agent interactions (point-to-point or mediated), and on the degree of agreement between the agents. He sees the holy grail of semantic integration in architectures that allow two agents to generate needed mappings between them on the fly without a priori agreements and without them having built-in knowledge of any common ontology.

Discussion Sessions

We then discussed as a large group near-term (1-3 year) Semantic Web applications the participants were interested in. The questions were as follows. (1) What will drive information providers to semantically mark up their pages? (2) What is the value of the semantic mark-up for information consumers? (3) In a nutshell, what will make the Semantic Web take off as the original Web did? We eventually ended up with five breakout session groups discussing the following questions :

(a) How do we do semantic mark-up for the majority of today's Web pages which are not static but generated on demand?

(b) How do we handle the different semantics that the same document (such as a contract) has for different departments of the same organization (e.g. legal, financial, human resources, technical management)?

(c) How can Semantic Web technology support better search and matching based on user-provided keywords?

(d) How do we support search of class catalogs across different universities?

(e) How could RDF support better email services (either by embedding RDF information in email headers, or by externally annotating existing email stores)?

Overlapping results were produced in the discussion groups. One of the concerns was that 2 billion web pages are static, but roughly 500 billion web pages are transient dynamic! One should probably not reproduce all the content in RDF or similar formats, but rather annotate the database directly and leave the data where it is. For instance, one might rather annotate the entry of Lynda Hardman's book on SMIL directly in the shop entry than to gather it from there. What then comes into play are the processes. Rather than to annotate content, it may become necessary to provide semantic information about how to access data in a process—Web Services based on semantic descriptions.

Given such information sources, there is the question of whether the Semantic Web can do any better at facilitating the retrieval of information than the best search engines. Actually, experiences from participants show that there are classes of problems, like retrieval of product information, where "classical" information retrieval engines on the Web do very poorly. These problems could be a good starting point for improving search by using semantics, especially when the context is narrow enough to warrant searching with one or several ontologies.

[1] Cf. http://semanticweb2002.aifb.uni-karlsruhe.de/USCHOLD-Hawaii-InvitedTalk2002.pdf

Obviously, such search possibilities may not solve every information integration and federation problem. Nevertheless, for some domains like educational purposes, the usage of semantic metadata seems to lend itself not only to centralized search schemes (e.g. a central semantic search engine), but also to federated ones like in metadata-based peer-to-peer systems.

Accepted Papers

Several papers discussed applications dealing with the traditional problem of the world where there is an abundance of electronic information: *finding your way around this information*.

Middleton, Alani, and De Roure (University of Southampton, United Kingdom) discussed this problem in the context of recommender systems (their paper got the highest ratings from the reviewers). The authors investigated the synergy between ontologies and recommender systems. An ontology provides recommender systems with the initial information, thus helping the system overcome the cold-start problem: a significant effort that is required to annotate initial documents in a recommender system or to create a profile for a new user. At the same time an analysis by the recommender system helps instantiate an ontology of research interests.

Davies, Duke, and Stonkus (BTexact Technologies, United Kingdom) present a similar problem: a community of practice sharing information where the system chooses which information to present to each user based on their profiles. The users of OntoShare annotate documents using terms from a shared ontology, which is described in the paper.

Quan, Huynh, and Karger (Massachusetts Institute of Technology, US) address a different aspect of personal-information management in their Haystack system: the formal representation, organization, and querying of personal information, such as email messages, web pages, and so on. The authors describe the use of RDF to create a Personal Information Ontology and the user of RDF metadata to enable novice users to generate user interfaces.

We have already mentioned the cold-start problems in recommender systems. Two other papers described ways of overcoming what can be described as *a cold-start problem for the Semantic Web*: automatically generating semantic markup from the information that is already available. This problem is arguably the biggest problem facing the Semantic Web today.

Haustein and Pleumann (University of Dortmund, Germany) describe an InfoLayer system that creates RDF and HTML pages based on an ontology defined in UML. They argue that many UML ontologies are already available and leveraging this resource will help produce a critical mass of pages with semantic markup.

Frank, Szekely, Neches, Yan, and Lopez (Information Sciences Institute/USC, US) describe the WebScripter tool which monitors the users when they perform simple intuitive operations, such as creating entries in spreadsheet, to extract alignment information between different ontologies. The tool enables naïve users, who may not have any knowledge of ontologies or taxonomies, to generate the critical set of alignment data.

Two papers present tools for *extracting semantic information from sets of documents that were generated by tools that do not support semantic annotation*.

Melis, Büdenbender, Goguadze, Libbrecht, and Ullrich (Universität des Saarlandes, Germany) describe ActiveMath—an environment that supplements on-line authoring of educational math documents in LaTex by generating semantic representation. The representation includes semantic markup for mathematical information itself, as well as pedagogical information for mathematical knowledge, such as difficulty of exercises, prerequisites, and so on.

Punin and Krishnamoothy (RPI, US) use collections of linked HTML pages generated by word-processing software, such as LaTex, Word, or PowerPoint, to discover hierarchical structure in the collection.

One paper argued that *fully separating semantics from presentation is not always desirable*.

Van Ossenbruggen and Hardman (CWI, the Netherlands) argue that sometimes presentation and semantics are so intricately intertwined that it is useful to include presentation information in the semantic markup. The authors propose a

SmartStyle layer that adapts presentation of information to the presentation context.

Two papers in the proceedings address the issues of *querying the Semantic web information*.

Nejdl and Wolf (Learning Lab Lower Saxony (L3S) and University of Hannover, Germany) and Staab and Tane (L3S and University of Karlsruhe, Germany) present Edutella, an open-source project that provides RDF metadata infrastructure for P2P applications. Edutella provides annotation and query services, which are both built on a common query language exchange format.

Ahmedi and Lausen (Universität Freiburg, Germany) describe the use of Lightweight Directory Access Protocol (LDAP) to support queries over semistructured databases. The authors investigate a global query strategy based on ontological links among entities that are spread across different local or remote servers. LGAccess, their query system, brings together ontologies, XML, and LDAP.

Concluding Remarks

Should we have another such event? The Semantic Web was virtually omnipresent at WWW-2002. The following question arose: Is its scope too wide to be accommodated in a one-day workshop? That is, would it be better to break into sub-workshops about different aspects of the Semantic Web? Judge for yourself—in Budapest 2003.

Acknowledgments

We would like to thank the organizers of the WWW2002 conference for their help. The hard work of the program-committee members ensured the high quality of the proceedings. The workshop was sponsored in part by OntoWeb.