

Which Future Web?

An interview with Frank van Harmelen

The semantic Web will be a considerable part of the future Web. What is the difference between the semantic Web and artificial intelligence? And what about Web 2.0? Frank van Harmelen, computer scientist in the Netherlands and a specialist in the semantic Web, answers some questions.

The semantic Web initiative is often said to address the same issues that have already been approached for thirty years in Artificial Intelligence. What makes the semantic Web, with its focus on ontologies and reasoning so different?

There is indeed a widespread misconception that the semantic Web is ‘AI all over again’. Even though the two may have tools in common (eg ontologies, reasoning, logic), the goals of the two programmes are entirely different. In fact, the goals of the semantic Web are much more technical and modest. Rather than seeking to build a general-purpose all-encompassing global Internet-based intelligence, the goal is instead to achieve interoperability between data sets that are exposed to the Web, whether they consist of structured, unstructured or semi-structured data.

Tim Berners-Lee (W3C) devoted an entire presentation to the confusion between AI and semantic Web in July 2006 (see the link below). This presentation also does a very good job of busting some of the other myths surrounding the semantic Web, such as that the semantic Web is mainly concerned with hand-annotated text documents, or that the semantic Web requires a single universal ontology to be adopted by all.

Web 2.0 appears to be the new kid on the block - everybody's darling, loved both by academia and industry. The semantic Web, on the other hand, has fallen from grace, owing to numerous unmet promises. How do you regard the coexistence of these two Webs and what role will Web 2.0 assume in the semantic Web's story?

My feeling is that this question is based on a false premise, namely that “the semantic Web has fallen from grace, owing to numerous unmet promises”. The SemTech conference, an annual industry-oriented event organized in the past three years in San Jose, California, attracted 300 attendants in 2005, 500



Frank van Harmelen.

attendants in 2006, and 700+ attendants in 2007. Its European counterpart, the European Semantic Technologies Conference, attracted 200+ attendants to its first event in Vienna in May 2007, of whom 75% were from companies. This would suggest that research and interest in the semantic Web are alive and well.

In fact, semantic technology is in the process of an industrial breakthrough. Here is a quote from a recent (May 2007) Gartner report, the industry watcher not known for its love of short-lived hypes:

“Key finding: During the next ten years, Web-based technologies will improve the ability to embed semantic structures in documents, and create structured vocabularies and ontologies to define terms, concepts and relationships. This will offer extraordinary advances in the visibility and exploitation of information - especially in the ability of systems to interpret documents and infer meaning without human intervention.” And: “The grand vision of the semantic Web will occur in multiple evolutionary steps, and small-scale initiatives are often the best starting points.”

Turning to the substance of your question: There is widespread agreement in the research world that Web 2.0 and semantic Web (or Web 3.0) are complementary rather than in competition. For example, a science panel at the WWW07 conference in May 2006 in Edinburgh came to the following consensus: Web 2.0 has a low threshold (it's easy to start using it), but also has a low ceiling (folksonomies only get you so far), while Web 3.0 has a higher threshold (higher startup investments), but has a much higher ceiling (more is possible).

The aforementioned Gartner report has useful things to say here as well. It advises the combination of semantic Web with Web 2.0 techniques, and predicts a gradual growth path from the current Web via semantically lightweight but easy to use Web 2.0 techniques to higher-cost/higher-yield Web 3.0 techniques.

And what about automated means of learning ontologies, relationships between entities, and so forth - that is, resorting to natural language processing, text mining, and statistical means of knowledge extraction and inference. Do you regard these techniques as complementary to the manual composition of ontologies or rather inhibitory?

My attitude towards the acquisition of ontologies and the classification of data objects in these ontologies is: if it works, it's fine. Clearly relying solely on the manual construction of ontologies puts a high cost and a low ceiling on the volume of knowledge that can be coded and classified. Hence, I expect that the techniques you mention will play an ever-bigger role in the range of semantic technologies. I see no reason why such techniques are ‘bound to fail’? instead I am rather optimistic about their increasingly valuable contribution.

All great technological inventions and milestones are marked by the advent of a killer application. What could/will be

the semantic Web's killer app? Will there be one at all?

I find the perennial question for the 'killer app' always a bit naive. For example: we can surely agree that the widespread adoption of XML was an important technical innovation. But what was XML's 'killer app'? Was there a single one? No. Instead there are many places where XML facilitates progress 'under the hood'. Semantic Web technology is primarily infrastructure technology. And infrastructure technology is under the hood, or in other words, not directly visible to users. One simply notices a number of improvements. Web sites become more personalized, because under the hood

semantic Web technology is allowing your personal interest profile to be interoperable with the data sources of the Web site. Search engines provide a better clustering of results, because under the hood they have classified search results in a meaningful ontology. Desktop search tools become able to link the author names on documents with email addresses in your address book, because under the hood, these data formats have been made to interoperate by exposing their semantics. However, none of these applications will have 'semantic Web technology' written on their interface. Semantic Web technology is like Nikasil coating in the cylinders of a car: very few car drivers are aware of it, but they are

aware of reduced fuel consumption, higher top speeds and the extended lifetime of the engine. Semantic Web technology is the Nikasil of the next generation of human-friendly computer applications that are being developed right now.

Links:

<http://www.w3.org/2006/Talks/0718-aaai-tbl/Overview.html>

<http://www.cs.vu.nl/~frankh/>

Please contact:

Frank van Harmelen

Vrije Universiteit Amsterdam, The Netherlands

Tel:+31 20 598 7731/7483 (secr)

E-mail: Frank.van.Harmelen@cs.vu.nl

KAON2 – Scalable Reasoning over Ontologies with Large Data Sets

by Boris Motik

Scalability of ontology reasoning is a key factor in the practical adoption of ontology technologies. The KAON2 ontology reasoner has been designed to improve scalability in the case of reasoning over large data sets. It is based on a novel reasoning algorithm that builds upon extensive research in relational and deductive databases.

Ontologies – vocabularies of terms often shared by a community of users – are being applied in science and engineering disciplines as diverse as biology, geography, astronomy, agriculture and defence. Nowadays, ontologies are usually expressed in the W3C standard language called the Web Ontology Language (OWL). OWL ontologies consist of a schema part, called a TBox, which describes the concepts and relationships in the domain of discourse, and a data part, called an ABox, which describes the actual data in the application. An efficient reasoner is the cornerstone of most OWL-based applications. It implements the formal semantics of OWL and thus provides the application with query-answering capabilities.

While reasoning over OWL ontologies is a provably intractable computational problem, it has been observed that rarely do ontologies encountered in practice involve a combination of constructs that leads to intractability. By relying on sophisticated optimizations, reasoners were developed that can handle ontologies with large Tboxes, yet

these still do not provide adequate performance on ontologies containing large ABoxes. This has so far prevented the usage of OWL in many applications that depend on large data sets, such as metadata management and information integration.

Parallel to the development of reasoning techniques for OWL, significant effort has been invested into improving the scalability of relational and deductive databases. In particular, numerous optimizations of query-answering in (disjunctive) datalog (a widely used deductive database language) are known and have proven themselves effective in practice. It is therefore natural to try to improve the scalability of ABox reasoning in OWL by building on this large body of existing work.

This idea has been realized in a new reasoner called KAON2. The architecture of the reasoner is shown in Figure 1. The central component of the system is the reasoning engine, which implements a completely new reasoning algorithm. A query Q over an ontol-

ogy consisting of a TBox T and an ABox A is answered by first reformulating T as a set of clauses in first-order logic, and then transforming the result into a disjunctive datalog program DD(T). The latter step is the key part of KAON2: based on certain novel results in resolution-based theorem proving, it ensures that all answers of Q over T and A can be equally computed by evaluating Q over DD(T) and A. The main benefit of such a transformation is that to evaluate Q w.r.t DD(T) and A, the disjunctive datalog engine can use the optimization techniques known from deductive databases, such as (disjunctive) magic sets or join-order optimizations.

This approach to query-answering has shown itself to be practical and effective in cases where the TBox is rather simple but the ABox contains large amounts of data. On such ontologies, KAON2 has shown performance improvements over the state of the art of one or more orders of magnitude. KAON2 has thus become the platform of choice for numerous research proj-