Web Science Workshop

12th-13th September, 2005

Hosted by the British Computer Society, London

Report

Executive summary

On 12th and 13th September, 2005, the British Computer Society hosted a workshop on Web Science, at their headquarters in Southampton Street, London. Attendance at the workshop was by invitation only, with 21 participants. The workshop was chaired by Sir Tim Berners-Lee, Wendy Hall and James Hendler.

There were three types of sessions on the agenda: lynchpin talks from invited speakers; panels in which short presentations were followed by discussions; and open discussion sessions.

The purpose of the workshop was to explore what the critical research challenges are for the future of the World Wide Web, and to outline what major breakthroughs will be needed to let its exponential growth continue into the future. It was structured to consider fundamental questions in a range of Web technologies including:

- The networking needs for continued scaling of decentralized technology
- The scientific challenges of managing and searching for information in the growing Web space
- The means for managing the space of concepts and semantics needed for understanding and processing the emerging Web of data
- The challenges that arise when attempting to discover, invoke and compose a globally distributed and constantly evolving set of services
- The development of peer-to-peer and other architectures needed for continued Web scaling
- The interface needs for exploring and using the increasingly complex information space that is the future Web.
- Social, cultural and legal challenges raised by the evolving Web and means by which computer scientists and others can shape technology to address such issues.

Photographs of the event have been posted by Ora Lassila and James Hendler.

Content

Panel: Public policy and architecture

Discussion

Lynchpin talk by Yorick Wilks

Panel: Large scale decentralised (information) systems

Lynchpin talk by Tim Berners-Lee

Discussion

12th September

Introduction

- Tim Berners-Lee (chair)
- Wendy Hall
- James Hendler

The participants were welcomed to the workshop. Berners-Lee discussed his expectations for the scientific agenda over the next 10 years, and described the task in hand, as we move from tree-based information structures to web-based ones, as philosophical engineering.

Hall said that Web Science had grown out of work to develop the Semantic Web (SW), and the need for an intellectual overview of web-like relationships. The aim is to set up a new discipline, or at least to bring people together in a new interdisciplinary space.

Hendler announced that the workshop should result in a white paper, to be written by Hendler, Shadbolt & Weitzner.

Lynchpin talk by Hal Abelson

Hal Abelson's talk began from the observation that statistical methods and machine learning tended to produce better results than reasoning from ontologies, and set out 5 challenges to the SW community.

1. We need real applications, not demos that work in the lab.
2. How do we communicate the ideas without being mired in syntax? We don't want new languages, new working patterns, or new panaceas.
3. How do we make ontologies robustly outside the lab, with real-world constraints, on the web scale?
4. Knowledge is not all classificatory, and is not always best modelling by ontologies. How should alternative methods of representation and reasoning be incorporated?
5. Can the Web be made more useful by incorporating human-like reasoning?

In the discussion, various types of reasoning or types of knowledge were discussed. Yorick Wilks recommended information retrieval techniques. Henry Thompson pointed up the need for metadescription, and warned that that made inference harder. Carole Goble said that representations of workflow, rather than ontological knowledge, were needed in her area of e-science.

Berners-Lee argued that large, complex, internally-consistent ontologies were not necessarily required on the SW. Actually, we can generally employ small overlapping ontologies that can be sewn together, and which are consistent with each other in the requisite areas. Guus Schreiber noted that it was important for advocates of the SW not to overclaim.

Panel: Representation (or not) on the Web

- Nigel Shadbolt (chair) – position paper
- Carole Goble – position paper
- Craig Knoblock – position paper
- Guus Schreiber
- Henry Thompson – position paper – slides

Shadbolt argued that content is primary, and that we must avoid overanalysis. Lightweight ontologies work, while distributed representations (cf. Brooks) are compelling. Labels must work stably for such ideas to work.

Goble has been working in molecular biology, which produces too much information for web formats to cope with. What does a community want to share? There are lots of ontologies about, but sharing knowledge demands workflows which are key for knowledge dissemination.

Knoblock looked at the issue of bootstrapping the Web from syntax to semantics. We need simple tools and immediate payoff. We need personalised integration tools and need integration frameworks which allow end users to build applications.

Schreiber discussed representation and consensus. In practice, one needs to work with pre-existing representations that may be problematic. Major research issues include partial alignment techniques, the semantics of visual representations, and reasoning under inconsistency.

Thompson discussed the requirements of a *science* of the Web: hypotheses, experiment, falsifiability, consensus methods and standards for measuring quality. As regards the Web, the key question is how the SW differs from AI, and whether analytic techniques can outdo, or even match, statistical techniques.

In the discussion, Berners-Lee made the point that underlying much of the potential of the SW is the large amount of data in structured relational databases, which the SW will be able to exploit. Wilks, Thompson and Hendler underlined the difficulties in achieving consensus. Hendler and Goble pointed out that automatically generated ontologies and hand-crafted ones had different properties, and worked in different domains and tasks: one task for Web Science would be to create the use cases to link them. Robin Milner and Thompson highlighted the need to represent provenance.

Lynchpin talk by Robin Milner

Robin Milner's talk described a multi-layered structure for Web Science, by analogy to the natural sciences' three levels of physics, chemistry and biology. At the lowest level would be the elements of computational interaction, basic processes such as naming or synchronised actions. In the middle would be non-trivial structures such as agents, data structures or workflow patterns. At the upper level would be complete individual systems of software. Software science should be about description as well

as prescription. Software engineering, in contrast, is about processes of development, not about what software actually *is*, because languages had gradually become divorced from theories in computing. One aim of Web Science should be to keep languages and formalisms in step with theory. We are in danger of understanding complex biological systems better than wholly engineered computing systems.

In the discussion, Wilks suggested parallels with Milner's layered structure with Braithwaite's notion of semantic ascent and supervenience of layers. Issues surrounding reductionism were discussed, including the effects of the Web scale, and of having humans in the loop.

Panel: Public policy and architecture

- Daniel Weitzner (chair) – slides
- Kieron O'Hara
- Joan Feigenbaum – position paper
- Mark Ackerman
- Jonathan Zittrain – position paper

Weitzner noted the importance of debate in this area. Architecture constrains public policy in various ways. It allows copying of information, for instance, and allows anonymity. Similarly, policy has been poorly understood. Two important concepts are transparency and accountability. Can we build large scale systems that include transparency (the history of data uses) and accountability (the ability to check whether the policies that govern data manipulations and inferences were in fact adhered to)? Can you prove that you have used the information correctly?

O'Hara noted the difficulties in applying political and philosophical constructs to online spaces, which are analogous to but different from offline spaces. There is also an important distinction between normative and de facto accounts – how are these to be related?

Feigenbaum argued that personal information should be termed 'sensitive' not 'private'. The focus should shift from hiding information to saying what misuse is. How can you balance consultation and correction with automation and not bothering people all the time? On copyright, creative commons isn't a total solution, as it doesn't work for legacy content.

Ackerman discussed the issue of how things become resources. How can you make things that can be incorporated into people's lives? It is hard to understand what people do and what the intent of systems is in that context.

Zittrain discussed the notion of subversive and enabling technologies. For instance, RSS is an enabling technology, while the URL is a subversive idea, making every web address equal and one click away. Rules about privacy or copyright on the Web would have been wrong at the beginning, because it is hard to predict how people use information.

In the discussion, a number of points were made. Most privacy policy is concerned with government intrusion on the citizen's privacy, although there are other types of intrusion. An important issue, highlighted by Zittrain, Weitzner, Hendler and Ora Lassila, is that sites such as Google are bottlenecks, and the functioning of the Web depends to a large extent on their good behaviour. But often it is the badly-behaved

(hackers, illegal downloaders, spammers) who put in the effort to understand the technology.

Discussion

The summary discussion of the first day's work noted a number of issues.

- Transparency. Cookies or redirects make the Web less transparent for the user. Links are important invariants of the Web experience, but they are being undermined in various ways. Preserving such invariants is important for trust.
- Metadata are important. Does the SW knowledge representation architecture need a coherent story about provenance etc?
- How do we link the technical architecture with the human/social effects that follow?

These three issues are linked by suggestions such as e.g. time-stamping code encapsulated by a URI, to make it invariant. Changes could be noted in the metadata, in that case – how would this affect reasoning? TAG at W3C has discussed such matters in some detail.

- One other opportunity comes from technology to map knowledge, and spot trends and developments.

13th September

Lynchpin talk by Yorick Wilks

Yorick Wilks' talk focused on the relationship between NLP and the SW, in the wider context of the relationship between language and knowledge representation. If the SW is, or depends on, the traditional AI knowledge representation task, then there is no particular reason to expect progress in this new form of presentation as all the problems and challenges of logic reappear and it will be no more successful outside the narrow scientific domains where KR seems to work and the formal ontology movement has brought some benefits. Alternatively, if the SW is the WWW with its constituent documents annotated so as to yield their content or meaning structure, then NLP/IE will be central as the procedural bridge from texts to KR. This view was discussed in some detail and YW argued that this is in fact the only way of justifying the structures used as KR for any SW.

In the discussion, Berners-Lee argued that the SW rests not on NLP but logic. Logic and ontologies will suffice to extract much of the value from the data held in structured relational databases. Wilks responded that the unstructured (legacy) part of the Web needs NLP for annotation, even in other media, at least until visual recognition systems become more reliable.

Panel: Large scale decentralised (information) systems

- James Hendler (chair) – position paper
- David de Roure – position paper
- Ora Lassila – position paper – slides
- Dieter Fensel – position paper – slides
- Wendy Hall – position paper

Hendler focused on the information that is *not* on the Web. Automatic annotation is important, but basically tells the machine what it already knows. Getting humans into the annotation loop is important, involving providing applications into which people are willing to put some work. Semantic enabling will allow unifying approaches to creating, browsing and searching text, and integrating information that is not stored in a centralised way.

De Roure discussed the relationship between the SW and grid computing, as a method of getting to a world where large quantities of data were open to large amounts of computational power. There is an important need to get semantics into the middleware, and to allow scientists to formulate and answer more complex queries.

Lassila noted that sharing across application boundaries is very difficult. It is hard to anticipate all the application pairings. We need to focus on the information, not the application. Also, implicit information is inaccessible. Ontologies and reasoning services are needed for such things. Also SW concepts are useful for mobile and ubiquitous computing, for similar reasons: awkward usage situations and unanticipated encounters make application-centred approaches difficult.

Fensel argued that engineers relate the usage of resources to achievement of goals. We can abstract from hardware, and now we are abstracting from software, using service oriented architectures (SOA). SOAs will not scale without mechanising service discovery, etc, so machine-processable semantics are required for services. Later we may even talk of *problem*-oriented architectures, focusing on the customer's problem, not the provider's service.

Hall discussed the importance of linking, and interacting with big information spaces, multimedia environments, etc. We need to link things, add metadata to archives, put links in video, etc. The current web is mostly linkless, and we need to make it better. A query is in effect a non-existent link. We need the web to disappear behind the scenes.

In the discussion, it was agreed that open-ended architectures are essential, but complicate the task of designing the next generation of the Web. Semantic specifications of, e.g. services are still an open research question, which makes the task of working in such a distributed environment very complex.

Lynchpin talk by Tim Berners-Lee

Tim Berners-Lee's talk highlighted different phases of information structure, moving from characters, to linear structures, to trees and hierarchies, and lastly webs, which can grow without a central bottleneck. The goal of a web is serendipitous reuse, but a minus is that it comes with global obligations, such as maintaining Web content, which are important for allowing the serendipity to happen. The SW is a web of logic, very different from hypertext. We need the same standard for symbol space as the WWW. We need to be able to map URIs to anything. Looking up URIs is still the critical architecture.

Berners-Lee also offered some comments on the NLP/SW debate stemming from Wilks' lynchpin talk, arguing that the two are very separate.

NLP	**SW**
Words	Terms of logic

NLP	**SW**
Meaning is use	Meaning is defined in words, or code, or specific use
No ownership of words	URI ownership
“Hydrogen”	pt:Hydrogen
Defining words in ontology is never complete and a waste of time	Defining terms is never perfect but useful
NL is constantly changing	Ontologies are basically static
Can’t benefit from injecting logic	Can’t benefit from cloudy statistics
Machine finds stuff	Machine infers stuff

Following on from this, some further misconceptions of the SW were set out.

- *Is it RDF “syntax”?* No –XML etc can also be used.
- *Is it OWL?* No, there are alternative ontology representation languages.
- *Is it built on NLP extraction?* No.
- *Does it depend on manual annotation of legacy content?* No.
- *Actual SW data deployment will be mainly on existing RDBs?* Yes.
- *I will have to do everything?* No, it will be a collective effort. The effort of developing ontologies will scale. If we model the effort of building ontologies, relative to the scale of the organisation, number of users and number of cooperating builders, the effort per person of ontology building grows very small as the scale increases.
- *Is it slow?* No – many of the prototype systems are not optimised. Furthermore, simply dumping RDF into a triplestore isn’t going to solve information processing problems.
- *Everyone will have to mark up web pages to generate content?* No.
- *Management of data will have to change?* No.

Finally, Berners-Lee outlined some important large-scale phenomena and discoveries.

- Eigenvector analysis is wonderful: no-one assumed Google would be possible.
- Plume tracing. New trends can be studied to find originators of trends. This technology should be available beyond government.

Issues to do with decentralisation.

- How decentralised is the web, given for example the US domination of the DNS?
- We want to encourage P2P experience. Could we make this an HTTP fallback? When servers are unavailable, P2P could locate cached pages, removing the dependence on DNS, to make the web even more decentralised.

Issues to do with user interfaces.

- The ultimate SW browser would be every application rolled into one.

- Speech and NL interfaces to SW.
- Security/trust models, rules and policies. UI issues are the tricky ones here. It is easy to write a trust policy, but hard for a neophyte to do it.

In the discussion, the argument about the relationship between NLP and the SW continued, with Wilks arguing that on Berners-Lee's view there would be serious problems with the fuzziness of terms (in domains such as law, for example), and only certain areas of very hard science would be amenable to the SW approach, and Berners-Lee responding that some areas of law are relatively non-fuzzy, such as regulations. On Wilks' view, even ownership of URIs allowing the retrieval of relatively definitive information would not remove the fuzziness, and would also create the problem of trusting the institutions managing the URIs.

On the topic of trust, Hendler pointed out that trust is a problem throughout the WWW – authority cannot be created, and people make more or less informed judgements. Understanding the semantics of trust is useful but not necessary. Thompson worried that on the SW it is the computer that makes the judgement on the basis of metadata, not a person. Berners-Lee argued that it was another misconception about the SW that most applications would draw information from a wide range of unknown information sources (thereby exacerbating the problem of trust); in fact many applications will draw information from a small number of well-known and trusted sources. Joan Feigenbaum detected an issue for Web Science, which is that the architecture is so open that we lack computational models for investigating trust requirements.

Discussion

The final discussion section centred on two topics, architecture and the relation between the Web and society.

Architecture

The discussion on architecture, chaired by Shadbolt, followed an issue that had been prominent in the workshop, together with representation, decentralisation, service orientation, inference and large-scale structure.

Questions were raised by Fensel and Goble about whether the discussions previously had focused too much on browsers, whether demand for browsers (e.g. in e-science) was high or low, and whether an SW browser was vital. Berners-Lee argued that a generic browser looks attractive, but because of the richness and diversity of applications integrated by the SW, there will be a great variety of ways that people will want to interact with it.

Milner introduced the orthogonal topic of process: can we replace procedure names with URIs? What SW platforms need to be in place, and what can be developed by users? Berners-Lee contrasted the Web as a (nearly) static information space and Web services as a different protocol. Milner posited another type of Web Science looking at interaction calculi, such as Choreography or process algebras.

Various representational requirements were discussed, including probability, coreference resolution, persistence and process. Thompson raised the strategic question of whether there are things that OWL + RDF don't do that other KRLs can. Milner noted that P2P needs private links between parties, and so models of

decentralisation were needed. We don't have the notion of an abstract process, as we do of computation.

Further issues concerned negotiation, agents and trust; Hendler and Hall argued that the agent trust literature addressed different questions, and for the SW large-scale decentralised trust mechanisms were needed. Goble argued that the grid community was a better source of insight into trust than the agent world. Weitzner argued that there is much about trust that can't be modelled totally.

Web and society

The discussion on the Web and society was introduced by the chairman, Weitzner, reviewing earlier discussions on public policy questions of privacy and copyright, and arguing that we need to give people control over how they interact with information resources, while adding transparency. Feigenbaum argued that questions of accountability and transparency are linked: transparency enables accountability. It was generally agreed that it was possible to write rules to cover policies and policy violations, but that a problem was how to get people to use such rules.

Feigenbaum also noted that accountability was controversial, particularly at high levels of abstraction. Alternative goals, such as fairness, could be pursued. Weitzner distinguished between accountability and enforcement, and noted that the pursuit of particular goals needed experimental and empirical work to address the question: what happens to social interactions with particular architectures?

Berners-Lee and Milner discussed the use of process calculi to model small parts of systems, as a potential method for helping with these social questions.

Goble noted that any kind of social regulation infrastructure would have to be lightweight. Feigenbaum suggested that many aspects of the infrastructure, such as identification and authentication, aren't hard to use, although Hendler cautioned that controlling identities is hard in distributed open systems. Weitzner noted the useful property of the general SW architecture that allows generic high level rules covering social interactions to be written.

Note

Ben Shneiderman was unable to attend, but contributed a position paper.