The Semantic Web

Ben Humphreys (bh704@soton.ac.uk)

Abstract

The semantic web was proposed in 1994 by the inventor of the original world wide web, Sir Tim Berners-lee. He outlined a web of information that provided meaning and structure to information in a way that the existing web did not. This meaning, or *semantics* would allow for the automation of a greater number of tasks, such as a hospital booking that involves many agents, trust and communication as described in the article.

Since then the semantic web has become the focus of a large amount of research: knowledge representation, multi-agent systems, semantic web services and trust. Each of these subject areas has made a significant contribution to the current state of the semantic web. This paper summarises the research of a key paper in each of these areas of semantic web research, as well as the original Berners-lee article in order to provide an overview of the current state of the semantic web.

Contents

1 Introduction

The idea of the semantic web was first proposed by the creator of the world-wide web, Sir Tim Berners-lee, at the first World Wide Web developers conference in 1994. In this and a subsequent Scientific American article, Berners-lee described a vision of the future in which many simple tasks were automated through interconnected systems [1].

This document summarises the original Semantic Web article, along with four of the most important papers from the considerable scientific research surrounding the semantic web: knowledge representation from artificial intelligence, multi-agent research, web services from distributed systems, and trust.

2 Semantic Web as a Whole

The Semantic Web published in *Scientific American* was more of a manifesto than a scientific article, but it was key in bringing the subject of the Semantic Web to the wider world [1]. The key idea of the semantic web is that of adding structure and *meaning* to data to support automated services. The existing Hypertext-based [2] web lacks structure, making automation reliant on difficult natural language parsing. The article proposes the use of ontologies – networks of nodes and arcs representing entities and the relationships between them – as a way of giving structure and meaning to information.

To illustrate the potential of semantically-enabled systems, the article describes the booking of a hospital visit involving many services and agents being used to find the best solution. It then goes on to outline some of the concepts behind ontologies including the Uniform Resource Identifier (URI) before mentioning the role of agents and trust.

Other key points of Berners-lee's idea of the semantic web are scattered throughout. Namely, that the semantic web is not intended as a replacement for the 'traditional' current web, nor as a completely separate web of information. By working with existing information and semantically annotating services, processes can be automated more easily.

3 Knowledge Representation

3.1 Introduction

The representation and storage of knowledge is arguably the most important topic within the semantic web. As with the original web, the power of the semantic web will come through its vast wealth of information. It is also one of the fastest changing areas within the sphere of semantic web research, with a number of major onotology specification languages being released in the last decade (SHOE [3], OIL [4], DAML+OIL [5]).

From SHIQ and RDF to OWL: The Making of a Web Ontology Language [6] is a historical paper that outlines how knowledge representation in the semantic web has progressed, and how the currently popular Web Ontology Language (OWL) [7] came into existence. Its historical overview is useful for explaining the influences on OWL, influences that are likely to affect successors.

OWL was designed based on existing web ontologies to be 'backwards-compatible'. By using previous frameworks it is subject to some of their weaknesses which are highlighted in the paper. In particular, OWL uses the Resource Description Framework (RDF) for expressing facts, and

RDF Schema (RDFS) for defining classes, properties and their hierarchies. On top of the limited expression offered by RDFS, OWL adds the ability to define more complex classes: logical combinations such as intersections, unions and complements. Properties are also extended, allowing for property hierarchies, domains and ranges for values, and transitive, symmetric and functional properties.

The paper attributes the knowledge representation methods of *description logics* and *frames* as two other main influences in the creation of OWL, beyond that of its predecessors. The emphasis on sound, complete and solvable reasoning present in description logic [8] has caused OWL to grow beyond a simple ontology definition language to supporting complex proofs.

3.2 Problems

Horrocks et. al. go on to detail some of the shortcomings of OWL, partly created through the difficulty associated with trying to combine many languages, each with quirks and problems.

OWL uses RDF to express facts, and by extension uses the XML-like syntax associated with it. While this format has become a standard for expressing text-based data, it is extremely verbose. For example the short statement $Student = Person \geq \sqcap 1 enrolledIn$, meaning a *a Student is a Person who is enrolledIn at least 1 thing* [6], requires 11 lines of RDF (Figure 1). This verbosity is also carried through to the triple representation of RDF in which multiple statements are required to express a single fact.

```
<owl:Class rdf:ID="Student">
  <owl:intersectionOf rdf:parsetype="Collection">
    <owl:Class rdfs:about="Person" />
    <owl:Restriction>
      <owl:onProperty rdf:resource="enrolledIn" />
      <owl:minCardinality rdfs:datatype="&xsd;Integer">
        1
      </owl:minCardinality>
    </owl:Restriction>
  </owl:intersectionOf>
</owl:Class>
```

Figure 1: Verbosity of RDF: (taken from [6])

4 Agents

Agents were mentioned in Berners-lee's original Scientific American article, but only in passing as the services that utilise the semantic web's knowledge. Even within the context of the semantic web, there has been a great deal of research [9, 10]. While not directly related to the semantic web, *Intelligent Agents: Theory and Practice* by Wooldridge and Jennings [11] is a paper that brings together the research of a huge number of authors to compare and contrast the research in the field to date.

The paper is divided into three broad sections, first defining what constitutes an agent through *agent theory*, before moving onto the practical implementation details of *agent architectures* and *agent languages*.

4.1 Agent Theory

The definition of what constitutes an agent is something still being debated. Jennings summarises a number of properties that agents have, grouped into weak and strong agency. Weak agency attributes are autonomy, social ability, reactivity, proactiveness, whereas strong agents are usually thought to have more complex internal processes similar to humans.

The issues of *belief*, *desire* and *intention* are core to agent research. Agents believe certain things based on evidence and have a desire to complete tasks. Jennings et. al. summarise a number of different papers in the field that have investigated the concepts of belief, desire and intention.

4.2 Agent Architectures

This section of the paper is concerned with the architectures created through artificial intelligence research, which shares some common ground with agent development. Starting with classical AI research dealing with physical-symbol systems and moving on to more recent and possibly controversial approaches to AI architectures, Jennings provides a broad spectrum of important research.

4.3 Agent Languages

An agent language is defined in *Intelligent Agents* as "a system that allows one to program hardware or software ... in terms of some of the concepts defined by agent theorists" [11]. In practical terms agent languages can define the mental state of agents, their beliefs and goals.

The paper concludes by briefly discussing some of the details of existing applications that use agents, such as interface and information retrieval agents.

5 Web Services

In *Agent-based Semantic Web Services*[12], Gibbins et. al. mention some of the shortcomings of traditional non-semantic web services and outline the benefits of separating intentional and domain-specific information in messages.

5.1 Discovery & Description

The loosely-coupled distributed nature of services and the rapid creation of new services means that service discovery is very important. Service discovery requires two things: a directory service and service descriptions. The directory service allows services to advertise themselves, as well as allowing others to search for certain services. Service descriptions are vital for enabling other services and users to find the kinds of services they want.

In their paper, Gibbins et. al. explain that messages to and from web services can be thought of as similar to those sent between agents. Within agent communication languages (ACLs) such as FIPA [13] and KQML [14, 15], there are a finite number of purposes to a message, such as requests, directives, statments of fact. However in current service descriptions using WSDL and DAML, the domain-independent *speech act* and domain-specific ontology of the description are mixed, leading to less reusability and more brittle services [16].

The paper goes on to describe the problem of finding appropriate services within a service directory. At the time of writing there was no standard query language for RDF, DAML+OIL or OWL, and so Gibbins et. al. were forced to use a restrictive query style based on bNodes. While not mentioned in the paper, the recently approved SPARQL [17] RDF query language may prove useful for solving the problem of finding semantically-described services.

6 Trust

Trust is an issue within the semantic web that has only recently come to the forefront of research surrounding it. When potentially conflicting information is gathered from a large number of sources, the issue of trust arises. How much credence should be put in each source, and concerning which topics? *Trust Strategies for the Semantic Web* by K. O'Hara et al. [18] investigates and summarises a number of trust strategies for the semantic web with a view to providing a guide for developers of agents using the semantic web.

6.1 Trust Strategies

There is no single perfect trust strategy, but a number of variations that are better suited to certain tasks.

Pessimistic At first the agent does not trust any sources, but slowly finds trustworthy sources as it it has evidence to do so.

Optimistic The logical opposite to the pessimistic approach, agents initially trust all sources until an event occurs to remove that trust.

Centralised A centralised trust strategy relies on a trustworthy institution to find trustworthy sources. This approach simplifies the problem of trust for individual agents, but creates other problems of authority, scalability and security.

Investigative Agents communicate their current ideas about which other agents are trustworthy. From this 'gossip' they build up a distributed network of agreed trust levels, and agents consult other agents when receiving information from an unknown source.

Transitive When agents have a direct connection to their source, the trust they place in that source is determined by them alone. However when trust information is gathered indirectly through one or more third-parties, it can be modified in a number of ways.

6.2 Costs and Benefits

The strengths and weaknesses of strategies are most commonly measured in their costs and benefits in various terms. O'Hara defines four key costs as well as the abstract concept of risk. The details of these costs are summarised in Table 1.

- Operational Cost: The general cost of the strategy.

- Opporunity Cost: The rewards gained by interacting with other agents and services.
- Deficiency Cost: Costs associated with being betrayed.
- Service Charges: Other costs associated with using other services.

	Optimism	Pessimism	Centralisation	Investigation	Transitivity
Operational Cost	Low - does not require much policing	Rises with complexity of filtering tests	Low - the cost is embedded in the centralised service	High -complex tests increase cost	Low if efficient methods are used
Opportunity Cost	Low - unlikely to miss many interactions	High - does not check all possibilities	High - decrease if many agents are certified	High -only interact with investigated agents	The larger the network, the lower the cost
Risk	High - more risk when less cautious	Low - won't interact with uncertain agents	Low - only legitimate agents are certified	Risk is low by lowering uncertainty	Rises with length of chains of referrals
Deficiency Cost	High - no check for malicious agents	Low - malicious agents will be caught up front	High - betrayal of one certified agent may affect the whole service	Low - malicious agents banned if discovered	High - behaviour of one agent might spread to network
Service Payments	High - if max price/no. of interactions not set	Limited interactions results in limited charges	Low - pay for the centralised service only	High - may need to consult several services	May rise when accessing fee-based services

Table 1: Estimated costs for trust strategies (taken from [18])

7 Critical Analysis

The original Semantic Web article in Scientific American was a bold manifesto showing Tim Berners-lee's vision of the semantic web. It has been criticised for 'oversimplifying the problem' and making an extremely complex task requiring research in many areas seem easy to achieve [19].

Parallels can be drawn between the current state of the semantic web and that of the early days of the world-wide web. The primary users are scientific researchers and it arguably has yet to make a large impact on the world. For the semantic web to grow, there must be a perceived benefit to using it for businesses and individuals.

As future knowledge representation work is likely to extend existing technology, the historical nature of *The Making of a Web Ontology Language* makes for an excellent starting point to understand the current state and future developments of knowledge representation.

Extremely detailed and exhaustive, *Intelligent Agents: Theory and Practice* provides an encyclopaedic overview of agents in general. However the paper is difficult to understand at times, requiring readers to be experts in artificial intelligence to understand some concepts and terminology.

Agent-based Semantic Web Services highlights the shortcomings of existing service descriptions and proposes a more robust semantic model. It is useful for providing an insight into the benefits of the semantic web, but is at times verbose and features large bodies of RDF that have little use.

Trust, beyond that of simple authentication, is a topic previously ignored by semantic web research. *Trust Strategies for the Semantic Web* provides a simple introduction to strategies and costs, but still does not address the issues of subjectivity, representation and practical transitivity.

There is still a long time before the semantically-enabled booking services and multi-faceted interactive agents mentioned in the Scientific American article are realised.

Word count: 1993 words, excluding figures and appendices.

References

[1] T Berners-Lee, J Hendler, and O Lassila. The semantic web. *Scientific American*, Jan 2001.

[2] V Bush. As we may think. *Computer-Supported Cooperative Work: A Book of Readings*, Jan 1988.

[3] Shoe: A knowledge representation language for internet applications. *Technical Report*, 1999.

[4] Oil: An ontology infrastructure for the semantic web. *IEEE Intelligent Systems*, 2001.

[5] DAML+OIL reference description. *W3C Note*, 2001.

[6] I Horrocks, P Patel-Schneider, and F van Harmelen. From shiq and rdf to owl: the making of a web ontology language. *... : Science*, Jan 2003.

[7] D McGuinness and F van Harmelen.... Owl web ontology language overview. *W3C Recommendation*, Jan 2004.

[8] *The Description Logic Handbook*. Cambridge University Press, 2002.

[9] Michael Huhns. Agents as web services. Jul 2001.

[10] James Hendler. Agents and the semantic web. May 2001.

[11] M Wooldridge and N Jennings. Intelligent agents: Theory and practice. Sep 2005.

[12] N Gibbins, S Harris, and N Shadbolt. Agent-based semantic web services. *... : Science*, Jan 2004.

[13] Fipa communicative act library specification. *Foundation for Intelligent Physical Agents*, Oct 2002.

[14] T Finin, R Fritzson, D McKay, and R McEntire. KQML as an agent communication language. *Proceedings of the third international conference on*, Jan 1994.

[15] H Chalupsky, T Finin, R Fritzson, and D McKay. An overview of KQML: A knowledge query and manipulation language. *KQML Advisory Group. April*, Jan 1992.

[16] Web services description language (WSDL) version 2.0 part 1: Core language. *W3C Working Draft*, 2004.

[17] SPARQL query language for RDF. *W3C*, 2007.

[18] Trust strategies for the semantic web. Sep 2004.

[19] N Shadbolt, W Hall, and T Berners-Lee. The semantic web revisited. *IEEE Intelligent Systems*, Jan 2006.

A Appendix

A.1 Reflection

I began my research into the semantic web by finding and reading a number of summary papers on the topic. After familiarising myself with the concepts behind the semantic web, I began researching certain aspects in more depth. My primary sources of information were online repositories of scientific journals such as Springer, as well as the university's E-Prints and library facilities.

By checking the references in useful papers, the more prominent papers in each subject area became clear. As technologies for the semantic web are changing so rapidly, it was hard to choose a single paper in each section that had clearly moved the research forward *and* was still relevant. I tried to solve this problem by selecting papers that were either higher-level summary papers or ones that illustrated the core concepts behind the topic area, rather than a particular implementation or technique within it.

In order to write the report, I began by re-organising the notes I had taken on each selected paper into a more cohesive form, paying special attention to certain key points that the papers emphasised.

Through this research project, I have learnt a great deal about the techniques, prospects and problems of the semantic web. My skills at reading large bodies of information and summarising them concisely have also been improved.