Uta Priss
School of Library and Information Science, Indiana University Bloomington, USA

Alternatives to the "Semantic Web": multi-strategy knowledge representation

Abstract: This paper argues that the Semantic Web needs to incorporate both formal and associative structures (and possibly a multitude of other structures and strategies) to be successful. The arguments for this claim are based on an observation of successes and failures in the areas of artificial intelligence (AI) and natural language processing (NLP).

1. Introduction

The WWW provides numerous challenges for information and knowledge processing activities. Information may be available but not accessible or retrievable because of language barriers or insufficient search strategies. Data mining techniques may discover implicit information in explicit data but these techniques do not necessarily guarantee that the discovered information is relevant, significant and trustworthy. During the last several decades hundreds or thousands of computer and information scientists have developed probably thousands of natural language processing and artificial intelligence techniques that were aimed at solving problems related to intelligent information processing only to encounter more and more new obstacles along the way. The latest solution, the Semantic Web, appears as an open declaration of defeat: since natural language processing and AI techniques did not provide sufficient results, it is now proposed to put the burden on the shoulder of the authors of webpages who are expected to populate their pages with metadata and additional markup. Metadata is essentially a new form of controlled vocabulary; markup - at least in the form of XML, XSL, etc - is essentially a programming language. Existing studies of the use of controlled vocabularies and indexing practices in information science and studies of teaching programming languages to "everybody" (Python, 2002) have shown that both are difficult and full of unsolved problems. This can further dampen the expectations of the success of the Semantic Web.

In contrast to machines and despite numerous inter-cultural conflicts around the world, humans do communicate surprisingly successfully even across national, linguistic and cultural boundaries. The question then arises: why are humans successful at information processing tasks such as information integration, translation and communication, which computers find so difficult? One obvious answer is that human cognition is embodied and grounded in our shared experiences of living in the same world. AI researchers have theoretically explored the idea of symbol grounding in the early 1990's but so far, connectionist artificial agents with perceptual interfaces have not been integrated with a large-scale capability of symbolic representations.

A critical review of existing strategies and theories of knowledge representation from a theoretical perspective is essential. The central question is: which strategies or theories have shown promise for which tasks, and why have they shown promise or have failed. It can be argued that successful knowledge representation requires a multitude of structures and techniques which each function with respect to specific conditions and contexts. This is in analogy to human cognition and communication, which also appears to be based on a multitude of structures. Some of these structures are "grounded" and thus provide for universals and primitives. Other structures are rule-based and provide for patterns, inferences and linguistic structures. Last but not least, meta-structures guide the combination and application of more basic structures. It can be argued that there is still hope that carefully designed systems that employ equally carefully designed knowledge representation strategies may overcome many of the current challenges for information systems. But this will require sophisticated theoretical models from a variety of disciplines.

2. The Semantic Web

In summer 2002 there were two workshops related to the Semantic Web co-located at Stanford University. The first one, the Semantic Web Working Symposium (SWWS, 2001) was funded by NSF, DARPA, INRIA and corporate sponsors; attracted several hundred participants from academia and industry; and had a highly technical focus. Many of the participants in this workshop were heavily involved in the implementation and development of the Semantic Web, such as invited speaker Eric Miller who is the W3C Semantic Web Activity Leader. The atmosphere at this workshop was very optimistic. The second workshop, the PORT Semantic Web Workshop (PORT-ICCS, 2001) attracted a much smaller number of participants but among them were famous researchers from diverse disciplines, such as Doug Engelbart, Ted Nelson, Terry Winograd, Keith Devlin, Tim Lenoir and Geoffrey Nunberg. The goal of PORT-ICCS was to investigate the use of the Semantic Web as a collaborative infrastructure for the development of Peirce Online Resource Testbeds (PORT). An unexpected outcome of this workshop was a fairly pessimistic view of the possibilities for the Semantic Web.

The SWWS final report defines the Semantic Web as "a vision: the idea of having data on the Web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications" (SWWS Final Report, 2001). Eric Miller in his invited talk and introduction to SWWS stated that in the Semantic Web "everything has an online representation: people, places, things ...". According to the SWWS final report, the Semantic Web *can* be built: "in order to make this vision a reality for the Web, supporting standards, technologies and policies must be designed to enable machines to make more sense of the Web, with the result of making the Web more useful for humans." The SWWS participants thus see the Semantic Web as universal and technically feasible.

The more pessimistic views of the PORT-ICCS participants included the following: Terry Winograd remarked that only a small portion of the Web actually requires structure. Major parts of the Web have mostly recreational and entertaining use and probably do not require a Semantic Web. He further had doubts about the feasibility of mark-up; whether authors of web pages would be sufficiently motivated to generate high quality, consistent mark-up. Ted Nelson characterized the Semantic Web as "just another delusion of the techies" because he considers human thought to be highly interconnected, non-hierarchical and difficult to process by machines.

In general there was suspicion among the PORT-ICCS participants that the Semantic Web really is a declaration of failure of the outcomes of artificial intelligence (AI) and natural language processing (NLP) research. If AI and NLP facilitated the creation of autonomous artificial agents with adequate language processing skills, then these agents would be able to process any information on the Web without requiring further mark-up. But neither AI nor NLP have yet reached their initial goals. For example, Geoffrey Nunberg, a linguist from Xerox Parc, argued at PORT-ICCS that although NLP can solve some tasks, such as genre classification and topic segmentation, universal NLP for any types of texts and unlimited domains will never work. The Semantic Web attempts to counteract these shortcomings by incorporating immense amounts of manual labor because it requires page authors to provide meta-information and mark-up. The Semantic Web is also under time pressure because a developing e-commerce industry demands instant solutions for conducting efficient, secure and profitable business on the WWW.

The question is what can and what cannot realistically be expected from the Semantic Web? What can the Semantic Web learn from successes and failures in the AI and NLP research of the last decades? Is the general idea of the Semantic Web pointing in the right direction, or are there better alternatives?

3. Successes and Failures

It is of interest to analyze which aspects of AI and NLP have produced successes and which have not so far. The success stories can be grouped into the following categories:

Pattern recognition: AI software has been built that can recognize patterns in visual, auditive, other sensory or abstract information. This ranges from face recognition to detecting credit card fraud (Kahn, 2002) by analyzing abstract patterns of credit card use. Pattern recognition also plays a part in speech recognition and robots (such as the Japanese toy dog robot that came on the market last year) because robots use pattern recognition as part of their perceptive interfaces.

Complex dynamic systems: AI software has been successful in modeling dynamic systems and designing solutions for controlling variables in such systems. For example AI solutions have been applied to monitoring tire pressure in cars, scheduling airport traffic, adjusting the movement in hand-held cameras, and building two-legged robots that walk in a human like fashion. These solutions often employ distributed representations. In a sense, even the search engine Google falls into this category. Although Google does not explicitly employ AI techniques, its success builds on exploiting the distributed nature of the dynamic linkage structure of the WWW.

Simulation of complex behavior: AI has achieved simulations of complex behavior such as language and human behavior. But these cannot yet pass the Turing test, i.e.

simulate behavior that is indistinguishable from real human behavior. Examples are artificial agents that users of computer games engage with (Johnson, 2002) or "bots", such as the Alicebot, that can participate in on-line chat with human users.

Rule- and logic-based systems: these are AI systems, such as expert systems that encode facts in a formal representation (knowledge base or AI ontology) and compute inferences based on logical rules. Examples are chess computers and medical expert systems that support doctors in the diagnostic process (Kahn, 2002).

NLP techniques: NLP has been successful in achieving tasks that rely on linguistic pattern recognition and text summarization and classification. These are especially successful if applied to domains, which employ formalized language styles and vocabularies, such as medicine. Linguistic pattern extraction can be used to identify the language in which a document is written, for speech synthesis and to filter information. For example, Monster.com uses pattern extraction to filter job announcements and job applicants's information out of general WWW documents (Kahn, 2002). Data mining websites, such as KD nuggets (2002), list numerous software tools for clustering or classifying documents and similar tasks.

Failures: despite the achievements of NLP there are still no reliable tools available for translating between languages or for identifying the content of a document in more detail than a summary or classification. Less formalized linguistic domains, such as poetry and colloquial language, pose even more significant problems for any NLP tool. The Semantic Web attempts to remedy these failures by increasing the amount of information that can automatically be extracted from webpages.

4. Associative and Formal Structures

There are two types of conceptual representations: associative and formal ones (Priss (2001). Associative concepts are grounded (or embodied); they are dynamic and complex. Formal concepts are rule-based, abstract and aim for consistency. Human reasoning most likely involves both associative and formal structures (Sloman, 1996). But most of the AI and NLP techniques mentioned in the previous section are either exclusively formal (such as most NLP and rule- and logic-based AI) or exclusively associative (such as AI tools for pattern recognition, complex systems and simulations and statistical NLP tools). Maybe the failure of AI and NLP tools is due to the fact that they do not combine associative and formal structures? Humans shift seamlessly between both forms of representation. Clark (1997) argues that the combination of different representations can invoke feedback loops. Maybe these feedback loops are a defining factor in human cognition. Lakoff & Johnson (1999) argue that classical, formal philosophy can be misleading because it ignores the embodied nature of cognition. There can be no doubt that logical formalisms serve some purpose because, for example, most computer software is entirely logical-formal. Thus Lakoff & Johnson's argument for embodied structures is really an argument for the need to combine embodied, associative structures with formal ones.

It is interesting to observe that the Semantic Web in its current vision is entirely formal. Although it is intended to exploit the Semantic Web in data mining applications, many of which are associative, there is no discussion about employing associative structures directly in the creation process of the Semantic Web. For example, there are so far no suggestions to learn from Google's successes, which are based on the exploitation of the dynamic linkage structure of the web, and to incorporate dynamic structures into the generation of the Semantic Web itself. Instead the creation of the Semantic Web relies heavily on the development of formal standards which become longer and more complicated with each new version. But studies in the area of controlled vocabularies, and formal programming languages (Python, 2002) and the experiences with 15 years of manual construction of the formal ontology CYC (Lenat, 2000) have shown that it is difficult (or impossible) to construct and manage large-scale formal systems that are consistent and user-friendly.

5. Conclusion

Instead of creating yet another entirely formal structure as an attempt to cope with natural language complexity, it might be more promising if the Semantic Web creators would pay more attention to previous successes and failures among AI and NLP tools. Most existing tools are either exclusively formal or exclusively associative. These tools are successful for limited tasks which are also predominantly formal or associative, respectively. But natural language and human cognition are more complex and require a combination of formal and associative structures (and possibly a multitude of other structures). It can thus be predicted that the Semantic Web also needs to combine both formal and associative structures and in general utilize multi-strategy approaches to knowledge representation. The current emphasis on complicated formal standards may not be sufficient. The underlying collaboration among authors that are expected to understand and apply these formal standards to their webpages is associative in nature and needs to be explicitly incorporated, modeled and understood by the supporters of the Semantic Web.

References

Alicebot. (2002) [http://alicebot.org/]. Accessed 3/15/02.

Clark, Andy. (1997). Being There. Putting Brain, Body, and World Together. MIT Press.

Johnson, Steven. (2002) Wild Things. Wired, March 2002/10.03. [http://www.wired.com/wired/archive/10.03/aigames.html]. Accessed 3/15/02.

Kahn, Jennifer. (2002) Alive! – Wired, March 2002/10.03. [http://www.wired.com/wired/archive/10.03/everywhere.html]. Accessed 3/15/02.

KD nuggets. (2002). [http://www.kdnuggets.com/]. Accessed 3/15/02.

Lakoff, George; Johnson, Mark. (1999). Philosophy in the Flesh. Basic Books, New York.

Lenat, Douglas. (2000). The Dimensions of Context Space. [http://www.cyc.com/publications.html]. Accessed June 2000.

PORT-ICCS. (2001). PORT Semantic Web Workshop. [http://www.ksl.stanford.edu/iccs/SemanticWeb.htm]. Accessed 3/15/02.

Priss, Uta. (2001). Multilevel Approaches to Concepts and Formal Ontologies. In: Proceedings of the 12th ASIST SIG/CR Classification Research Workshop.

Python. (2002). Computer Programming for Everybody. [http://www.python.org/cp4e/]. Accessed 3/15/02.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. Psychological Bulletin, 119, p. 3-22.

SWWS. (2001). Semantic Web Working Symposium. [http://www.semanticweb.org/SWWS/]. Accessed 3/15/02.

SWWS. Final Report (2001).[http://www.semanticweb.org/SWWS/report/]. Accessed 3/15/02.