Language Technologies and the Evolution of the Semantic Web

Enrico Motta, Marta Sabou

Knowledge Media Institute, The Open University, Milton Keynes, UK
{e.motta, r.m.sabou@open.ac.uk}

Abstract

The availability of huge amounts of semantic markup on the Web promises to enable a quantum leap in the level of support available to Web users for locating, aggregating, sharing, interpreting and customizing information. While we cannot claim that a large scale Semantic Web already exists, a number of applications have been produced, which generate and exploit semantic markup, to provide advanced search and querying functionalities, and to allow the visualization and management of heterogeneous, distributed data. While these tools provide evidence of the feasibility and tremendous potential value of the enterprise, they all suffer from major limitations, to do primarily with the limited degree of scale and heterogeneity of the semantic data they use. Nevertheless, we argue that we are at a key point in the brief history of the Semantic Web and that the very latest demonstrators already give us a glimpse of what future applications will look like. In this paper, we describe the already visible effects of these changes by analyzing the evolution of Semantic Web tools from smart databases towards applications that harness collective intelligence. We also point out that language technology plays an important role in making this evolution sustainable and we highlight the need for improved support, especially in the area of large-scale linguistic resources.

1. The Evolution of the Semantic Web

The key intuition underlying the Semantic Web (Berners-Lee et al., 2001) is that the availability of huge amounts of formally described semantic markup, at a scale comparable to that of the current Web, will make it possible to achieve a dramatic improvement in terms of agent interoperability and user functionalities, which will be enabled by the technology. For instance, the application of semantics to Web Services (Hepp et al, 2005) will make it possible to achieve flexibility at scale, where services will be dynamically located, composed and executed, a process which currently is carried out manually and is therefore expensive. It is also easy to envisage all sorts of new ‘smart’ functionalities for Web users, which will become possible once semantic markup becomes truly ubiquitous. For instance, we will see new tools for business intelligence, new shopping services, and new forms of news generation, syndication, and personalization, just to list a few examples.

Unfortunately, while this is a compelling and exciting vision, we are still quite a long away from realizing it. In particular, one obvious problem is that all these applications require that a large-scale Semantic Web is there in the first place, which is not the case yet. At the same time it is obvious that the process of realizing the Semantic Web cannot proceed according to a waterfall model, whereby we first build the required infrastructure and produce large-scale semantic markup, and in a second phase we exploit such markup to produce exciting new applications. Clearly, the two processes, Semantic Web construction and application (or at least, demonstrator) development, have to go hand-in-hand. The result of this strategy so far has been that, by and large, the early demonstrators produced in the past few years lack many of the key elements which will characterize ‘real’ Semantic Web applications. Specifically, Semantic Web applications will operate in an ***open, large-scale, distributed and heterogeneous*** environment, while these early ‘proof-of-concept’ tools provide semantic techniques on top of rather small, homogeneous and centralised data stores. Consequently, they are more akin to traditional knowledge-based systems, than to ‘real’ Semantic Web applications.

Recently, however, we have reached a turning point in the history of the Semantic Web regarding its size and development. The Semantic Web is gaining momentum by registering a 300% growth in 2004 alone and thus outpacing the growth of the Web itself (Lee & Goodwin, 2004). There is now a reasonable amount of online semantic data, to such an extent that the need has arisen for a semantic search engine, Swoogle (Ding et al., 2005), which can crawl and index all these data. Hence, we are now slowly reaching a key point in the history of this very young discipline, where we can move away from the early, simplified applications and start developing the kind of applications, which will characterise the Semantic Web of the future. In this paper we will analyse the current state of the art of Semantic Web applications and in particular we will look at a number of existing demonstrators, with the aim of identifying and differentiating the elements typical of first-generation Semantic Web applications, from those which will characterise the ‘real’ Semantic Web. In the analysis we will emphasize the key role played by language technologies in the context of the Semantic Web and in the second part of the paper we will identify some key ‘missing bits’, especially in the area of large scale linguistic resources, which need to be more closely targeted to the new scenarios presented by the Semantic Web.

2. From Smart Databases to Harnessing Collective Intelligence

Magpie (Dzbor et al., 2003), was one of the first tools to envision new mechanisms for browsing and making sense of information on the Semantic Web. In the absence of available semantic markup, this tool automatically generates a semantic layer, by mapping items on the current web page to an ontology, by means of Named Entity Recognition technology. In this respect Magpie is a classic example of a first-generation tool. Because Magpie