Bringing Semantics to the Web

Craig A. Knoblock
University of Southern California
Information Sciences Institute and Computer Science Department
knoblock@isi.edu

Introduction

A key factor in the success of the Web today is its simplicity and ease of use. This simplicity allows just about anyone to write and publish an HTML page and immediately see the results of their efforts. The next obvious step in the evolution of the web is to bring in semantics to describe the contents. This would allow not just sharing of information, but also support improved searching, extraction, filtering, monitoring, and, most importantly, integration. This has clearly been a goal of the Semantic Web, and while the Semantic Web is a step in this direction and has achieved some success, it has not been embraced the larger Web community.

I believe that the two key reasons that the Semantic Web has not taken the Web by storm are that, unlike the syntactic web, there is a lot more to learn in order to use it and the end users do not immediately see the benefit of their efforts. This is a challenge that needs to be addressed. There are two parts to this challenge. First, how do we make it easy to bring the needed semantics to the information and sources on the Web. Second, how do we provide an immediate benefit to the users such that they see a direct benefit for their efforts.

As an example, consider the set of tasks that I undertook to book a hotel in London for the Web Science Workshop. First, I reviewed the recommended hotels and noticed that the cheapest workshop rate was 125 pounds. Given the weak dollar, this converts into over $225, which is more than double what I would typically spend on a hotel in the US. Next, I reviewed a few of the hotel aggregator web sites (e.g., Orbitz) and found that I could get a smaller room in the same hotel for less money, but not a lot less. Finally, I went to the BiddingForTravel.com web site, which lists winning bids for Priceline. (Priceline is a web site that provides steep discounts on hotel rooms and airline tickets, but you have to bid on these products.) After reading through dozens of posts on BiddingForTravel, I found that I could probably bid on a hotel on Priceline for less than $100. In the end I was able to book the Thistle Marble Arch, which is a 4* hotel about 3 miles from the workshop, on Priceline for $75 per night. However, despite knowing where to look, the research required to find this hotel took several hours of work.

Let's consider the time consuming parts of this task. First, I had to read through dozens of posts on BiddingForTravel to find the prices for hotels on similar dates in London. Second, since in Priceline you bid on an area and not on a hotel, I had to take the list of hotels organized by areas, and invoke GoogleMaps on each hotel to see where they were located relative to the workshop location. Third, I had to check the prices on the individual hotels for the required dates to be sure that I would not bid more than the hotels were charging. Finally, I had to combine all of this information together (largely in my head) to make a decision on what hotel to book or what to bid in Priceline. The first task is an information extraction problem, the second and third are tasks that could be automated given better models of the sources, and the last task could be greatly improved with better web integration tools. I will address each of these in turn.

Information Extraction

There are many unstructured and semistructured sources on the web that contain valuable information. The problem is how to turn this data into information that can be exploited and integrated in a larger context.

The problem of wrapping semistructured sources has been studied extensively and there are now commercial products for wrapping web sites. A more neglected problem is turning the unstructured data into usable information. In the case of the posts on BiddingForTravel, each post provides the hotel name, Priceline area, prices, and dates, but the information is provided by individual users and there is no fixed structure. One example approach to this problem is illustrated by a recent technique we developed for exploiting reference sets for information extraction [MK05]. The idea is to first link the posts on a site such as BiddingForTravel with the reference set and then use the reference data to improve the accuracy of the extraction. The result is the ability to accurately structure unstructured data so that it can then be processed and queried automatically.

Source Modeling

There are, of course, many web sources that are already structured, such as web services and databases, but do not provide any semantic information to facilitate the modeling. Even for many of the web services, there are very few that provide anything but the WSDL description of the source. So an important challenge is to develop techniques that automate as much as possible the modeling of structured sources and services. An example of this is some preliminary work we did on learning models of the inputs, outputs, and functions of a web service by exploiting background knowledge and related services [CK05]. The key idea in this work is to exploit background knowledge and experimentation to automatically formulate a model of a new source.

Web-based Integration Tools

The extraction and modeling address the problem of getting access to the relevant data in a usable form, but there is still the issue of how to integrate the data in novel ways. For the problem described above, there are a number of tasks that require integrating the data across sources. Having a semantic model of the sources greatly facilities these tasks, but the other part of the problem is allowing the user to interactively specify how the data should be integrated. Providing web-based integration tools will create the demand and need for more structure on the sources. An example of this is an approach that we are developing that allows a user to build integration applications [KST05]. Our approach is based on the idea of defining the integration tasks using constraints, which specify which sources to query and how to combine the sources.

Discussion

The high-level vision of the Semantic Web is on target, but the challenge is how to get there. I have argued that the key to success in bringing semantics to the Web hinges on addressing two issues. First, we need to develop tools and techniques that greatly simplify the task of structuring and modeling the data already available on the Web. The assumption that users today will be willing to provide semantic models of their sources and services is flawed. Second, we need to develop integration frameworks and related tools that allow end users to rapidly create their own personal integrated applications. This type of tool will provide immediate benefit to users and fuel the need to provide additional semantics for the existing sources.

References

[CK05] Mark James Carman and Craig A. Knoblock. Inducing source descriptions for automated web service composition. In *Proceedings of the AAAI 2005 Workshop on Exploring Planning and Scheduling for Web Services, Grid, and Autonomic Computing*, AAAI Press, 2005.

[KST05] Craig A. Knoblock, Pedro Szekely, and Rattapoom Tuchinda. A mixed-initiative system for building mixed-initiative systems. In *Proceedings of the AAAI Fall Symposium on Mixed-Initiative Problem-Solving Assistants*, 2005.

[MK05] Matthew Michelson and Craig A. Knoblock. Semantic annotation of unstructured and ungrammatical text. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-2005)*, 2005.