ABSTRACT

In order to combine data from various heterogeneous sources, software agents must first understand the semantics of the sources, expressed in the source model. Currently, source modeling is manual, but as large numbers of sources come online, it is impractical to expect users to continue modeling them by hand. We describe two machine learning techniques for automatically modeling information sources: one that uses source's metadata, contained in a Web Service Definition file, and one that uses the source's content, to classify the semantics of the data it uses. We go beyond previous works and verify predictions by invoking the source with sample data of the predicted type. We provide performance results of both methods and validate our approach on several live Web sources. In addition, we describe the application of semantic modeling within the CALO project.

Keywords: data integration; data modeling; data semantics; semantic data model

INTRODUCTION

Software agents will soon assist users in the office environment by carrying out complex everyday tasks, such as planning a trip to a meeting. To complete their task successfully, these agents will need to combine information from a variety of heterogeneous sources. Take, for example, the task of arranging a trip. A user has to follow a number of steps in a sequence, including:

- get the date and location of the meeting,
- find a convenient flight on those dates to that location,
- get the weather forecast,
- find all hotels in the city of the meeting with the required amenities (e.g., high speed Internet),
- keep hotels within 3 miles of the meeting site,
- keep hotels with rates within government-allowed per diem,
- book hotel from the list with the best reviews,
- rent a car, and so on.

Some of the information sources a user interacts with in the course of carrying out this task are shown in Figure 1. In order to make use of these sources—whether they are
Web services, HTML pages or databases—an intelligent agent needs to know the schema of the source. Currently, the user must manually model the source: specify the type of data it accepts and returns, and the functionality it provides. This is what is done by information integration systems (Thakkar, Ambite, & Knoblock, 2005) that, similarly to our envisioned intelligent office assistants, provide uniform access to heterogeneous online information sources in order to answer a specific user query or be composed with other sources as a new information service. However, as new sources come online or are discovered, it is infeasible to ask the user to explicitly model them. While various technologies, most notably the Semantic Web, have been proposed to enable programmatic access to new sources, they are slow to be adopted, and at best will offer only a partial solution, because information providers will not always agree on a common schema. Rather than rely on standards to which providers may or may not conform, we instead propose to automatically model a new source: that is, discover the semantics of its inputs and outputs, and the functionality it provides.

We divide source modeling tasks into two subproblems: recognizing the semantics of the data it uses (semantic labeling), and inducing the logical definition of the source (learning functionality) (Carman & Knoblock, 2007). This article addresses the first problem, which is similar to the schema matching problem encountered by the database community, where the schema of one database has to be linked with the schema of another before the first database can use the data provided by the second. Unlike schema matching, where the content of both databases is available to the matching algorithm, we must first figure out how to invoke the source whose input and output parameters we are trying to label. We will show in the article that we can use the source’s metadata, if available, to classify the input parameters used by the source and assign them to predefined semantic types. If the metadata is not easily available, but instead, a user has provided a few examples of the inputs by querying the source, we show that we can classify the data types used by the source using their content only. While others have used metadata to semantically classify the source’s data types (Hess & Kushmerick, 2003), we go further and verify a classifier’s
The page contains a call to action for purchasing the full version of the document, available at [www.igi-global.com/article/semantic-labeling-online-information-sources/2838?camid=4v1](http://www.igi-global.com/article/semantic-labeling-online-information-sources/2838?camid=4v1). It also mentions that the title is available in various InfoSci-Journals collections and an option to recommend the product to a librarian at [www.igi-global.com/e-resources/library-recommendation/?id=2](http://www.igi-global.com/e-resources/library-recommendation/?id=2).

**Related Content**

- **A Mobile Matchmaker for the Ubiquitous Semantic Web**
  Floriano Scioscia, Michele Ruta, Giuseppe Loseto, Filippo Gramegna, Saverio Ieva, Agnese Pinto and Eugenio Di Sciascio (2014). *International Journal on Semantic Web and Information Systems* (pp. 77-100).
  [www.igi-global.com/article/a-mobile-matchmaker-for-the-ubiquitous-semantic-web/129763?camid=4v1a](http://www.igi-global.com/article/a-mobile-matchmaker-for-the-ubiquitous-semantic-web/129763?camid=4v1a)

- **Efficient String Matching Algorithm for Searching Large DNA and Binary Texts**
  [www.igi-global.com/article/efficient-string-matching-algorithm-for-searching-large-dna-and-binary-texts/189771?camid=4v1a](http://www.igi-global.com/article/efficient-string-matching-algorithm-for-searching-large-dna-and-binary-texts/189771?camid=4v1a)

- **Semantic Web and Adaptivity: Towards a New Model**
  [www.igi-global.com/chapter/semantic-web-adaptivity/35746?camid=4v1a](http://www.igi-global.com/chapter/semantic-web-adaptivity/35746?camid=4v1a)