Chapter VII

XWRAPComposer: A Multi-Page Data Extraction Service

Ling Liu, Georgia Institute of Technology, USA
Jianjun Zhang, Georgia Institute of Technology, USA
Wei Han, IBM Research, Almaden Research Center, USA
Calton Pu, Georgia Institute of Technology, USA
James Caverlee, Georgia Institute of Technology, USA
Sungkeun Park, Georgia Institute of Technology, USA
Terence Critchlow, Lawrence Livermore National Laboratory, USA
David Buttler, Lawrence Livermore National Laboratory, USA
Matthew Coleman, Lawrence Livermore National Laboratory, USA

Abstract

We present a service-oriented architecture and a set of techniques for developing wrapper code generators, including the methodology of designing an effective wrapper program construction facility and a concrete implementation, called XWRAPComposer. Our wrapper generation framework has two unique design goals. First, we explicitly separate tasks of building wrappers that are specific to a Web service from the tasks that are repetitive for any service, thus the code can be generated as a wrapper library component and reused automatically by the wrapper generator

system. Second, we use inductive learning algorithms that derive information flow and data extraction patterns by reasoning about sample pages or sample specifications. More importantly, we design a declarative rule-based script language for multi-page information extraction, encouraging a clean separation of the information extraction semantics from the information flow control and execution logic of wrapper programs. We implement these design principles with the development of the XWRAPComposer toolkit, which can semi-automatically generate WSDL-enabled wrapper programs. We illustrate the problems and challenges of multi-page data extraction in the context of bioinformatics applications and evaluate the design and development of XWRAPComposer through our experiences of integrating various BLAST services.

Introduction

With the wide deployment of Web service technology, the Internet and the World Wide Web (Web) have become the most popular means for disseminating both business and scientific data from a variety of disciplines. For example, vast and growing amount of life sciences data reside in specialized Bioinformatics data sources, and many of them are accessible online with specialized query processing capabilities. Concretely, the Molecular Biology Database Collection currently holds over 500 data sources (DBCAT, 1999), not even including many tools that analyze the information contained therein. Bioinformatics data sources over the Internet have a wide range of query processing capabilities. Typically, many Web-based sources allow only limited types of selection queries. To compound the problem, data from one source often must be combined with data from other sources to provide scientists with the information they need.

Motivating Scenario

In the Bioinformatics and Bioengineering domain, many biologists currently use a variety of tools, such as DNA microarrays, to discover how DNA and the proteins they encode may allow an organism to respond to various stress conditions such as exposure to environmental mutagens (Quandt, Frech, Karas, Wingender, & Werner, 1995; Altschul et al., 1997; DBCAT, 1999). One way to accomplish this task is for genomics researchers to identify genes that react in the desired way, and then develop models to capture the common elements. This model will be used to identify previously unidentified genes that may also respond in similar fashion based on the common elements. Figure 1 illustrates a workflow that a genomics researcher has created to gather the data required for this analysis. This type of workflow