WrapIt: Automated Integration of Web Databases with Extensional Overlaps

Mattis Neiling*, Markus Schaal*, and Martin Schumann

Free University of Berlin, Department of Economics, Institute for Information Systems, Garystraße 21, D-14195 Berlin, Germany
mneiling@wiwiss.fu-berlin.de, schaal@cs.tu-berlin.de, info@schumannsoftware.de

Abstract. The world wide web does not longer consist of static web pages. Instead, more and more web pages are created dynamically from user request and database content. Conventional search engines do not consider these dynamic pages, as user input cannot be simulated, thus providing often insufficient results.

A new approach for online integration of web databases will be presented in this paper. Providing only one sample HTML result page for a source, result pages for new requests will be found by structural recognition. Once structural recognition is established for one source, other web databases of the same universe (e.g. movie databases) can be integrated on the fly by content-based recognition. Thus, the user receives results from various sources.

Global schemata will not be produced at all. Instead, the heterogeneity of the single sources will be preserved. The only requirement is given by the existence of an extensional overlap of the databases.

1 Introduction

The world wide web is increasing at an enormous rate. Already now it can be viewed as the largest source of information. According to Najork and Vien-\nne\-s\-e [MN01], the following estimation applied in October 2000: 2.5 billion pages were directly accessible and 550 billion pages were dynamically generated. Especially, more and more databases are made available via the web. In order to access information from these web databases, queries are entered into certain input forms (text boxes, check boxes, etc.). The query result is returned by dynamically generated result pages. Users need to access each source individually in order to formulate appropriate queries and to identify the results.

Automated searching across several web databases requires knowledge about the structure of each individual site. Nowadays, individual wrappers are developed for each source separately. This solution is not very satisfying, as new sources appear every day and legacy sources may change their layout and programming style. Instead of creating one wrapper per source, a spider is used for

* Part of this work was supported by the Berlin-Brandenburg Graduate School in Distributed Information Systems (DFG grant no. GRK 316)
querying sources. Following this wrapper-free approach, a method for recognition of result pages is required. Related work (e.g. Crescenci et. al. [CMM01]) has proposed the use of several (at least two) sample result pages per source in order to recognize other result pages. Our approach extends this idea to multiple sources, requiring only one sample page for one source.

Our idea is as follows: The structure of result pages of the first source can be derived by comparing the sample page with new query results (so-called structural recognition). Once the structure of the first source is known, domain-specific content can be extracted from this source. Then, the same query is sent to another source and result pages are recognized there by content-based matching (so-called content-based recognition).

Thus, wrapper-free search can be extended towards automated integration of previously unknown or changing web databases.

2 State of the Art

Semi-automated wrapper generation has been recognized as challenge in the scientific community. At least two approaches exist, namely Lixto (cf. Baumgartner et. al. [BFG01]) and W4F (World Wide Web Wrapper Factory, cf. Sahuguet et. al. [SA99]).

Lixto generates Wrappers by user interaction. A graphical user interface (Interactive Pattern builder) enables the user to select text parts within HTML pages. Each text part is an example for a kind of data which can be extracted. Lixto represents the HTML page internally as a tree where text parts correspond to partial trees. Additional knowledge has to be provided by the user in terms of so-called Lixto or Elog rules for the extraction of the actual data.

As it is the case for Lixto, W4F also uses a query language (HEL, HTML Extraction Language) for producing extraction rules. From these rules JAVA source code is produced. This source code is the actual wrapper and contains not only the extraction methods, but also methods for communication with the web server.

Besides semi-automated wrapper generation, there is also an approach for automated wrapper generation, namely Roadrunner (cf. Crescenci et. al. [CMM01]). Very similar to the approach described here, Roadrunner starts out with two sample result pages and analyzes the structure of such pages by parsing the HTML pages in parallel. Identical elements are ignored. Different text strings are used in order to localize content, while different HTML tags are used in order to analyze optional or repeating structures. The Roadrunner approach leads to a grammar for result pages.

While our approach builds upon the concepts realized in Roadrunner, we have gone one step further by extending the search to other sources without having previous knowledge except for their Unified Resource Locator (URL).