Chapter XVII

Machine Learning Techniques for Wrapper Maintenance

Kristina Lerman
University of Southern California, USA

Steven N. Minton
Fetch Technologies Inc., USA

Craig A. Knoblock
Fetch Technologies Inc. & University of Southern California, USA

ABSTRACT

The proliferation of online information has led to an increased use of wrappers for extracting data from Web sources and transforming it to a structured format. The resulting data can then be used to build new enterprise applications. While most of the previous research has focused on quick and efficient generation of wrappers, the development of tools for wrapper maintenance has received less attention. This is an important problem, because Web sources often change in ways that prevent the wrappers from operating correctly. In this chapter, we describe machine learning techniques for verifying that a wrapper is working correctly and repairing it if not. Our approach is to learn structural descriptions of data and use these descriptions to verify that the wrapper is correctly extracting data. The repair algorithm automatically recovers from Web source format changes by identifying data so that a new wrapper may be generated for this source.
INTRODUCTION

Companies have vast repositories of information, which they share among themselves as well as with outside users. Unfortunately, much of this information is presented in a form that can be easily read by humans, not computer applications. Although there are hopes that the extensible markup language (XML) will solve the information extraction problem, it is not yet in widespread use, and even in the best case, it will only address the problem within application domains where the interested parties can agree on the XML schema definitions. Enterprises are instead relying on Web wrappers to extract information from Web sources and convert it to a structured format that can be used by various applications.

Wrappers use extraction rules to identify the data field to be extracted. Semiautomatic creation of extraction rules, or wrapper induction, has been an active area of research in recent years (Knoblock et al., 2001a; Kushmerick, 1997). The most advanced of these wrapper generation systems use machine learning techniques to learn the extraction rules by example. For instance, the wrapper induction tool developed at USC (Knoblock et al., 2001a; Muslea, 1998) and commercialized by Fetch Technologies, allows the user to mark up data to be extracted on several example pages from an online source using a graphical user interface. The system then generates “landmark”-based extraction rules for these data that rely on the page layout. The USC wrapper tool is able to efficiently create extraction rules from a small number of examples; moreover, it can extract data from pages that contain lists, nested structures, and other complicated formatting layouts.

In comparison to wrapper induction, wrapper maintenance has received less attention. This is an important problem, because Web sites frequently change their layout, and even slight changes in the Web page layout can break a wrapper. In this chapter, we discuss our approach to the wrapper maintenance problem, which consists of two parts: wrapper verification and reinduction. A wrapper verification system monitors the validity of data returned by the wrapper. If the site changes, the wrapper may extract nothing at all or some data that are not correct. The verification system will detect data inconsistency and notify the operator or automatically launch a wrapper repair process. A wrapper reinduction system repairs the extraction rules so that the wrapper works on changed pages.

Figure 1 graphically illustrates the entire life cycle of a wrapper. As shown in the figure, the wrapper induction system takes a set of Web pages labeled with examples of the data to be extracted. The output of the wrapper induction system is a wrapper, consisting of a set of extraction rules that describe how to locate the desired information on a Web page. The wrapper verification system uses the functioning wrapper to collect extracted data. It then learns patterns describing the structure of data. These patterns are used to verify that the wrapper is correctly extracting data at a later date. If a change is detected, the system can automatically repair a wrapper by using this structural information to locate examples of data on the new pages and to rerun the wrapper induction system with these examples. At the core of these wrapper maintenance applications is a machine learning algorithm that learns structural information about common data fields. We will introduce the algorithm, DATAPROG, and describe its application to the wrapper maintenance tasks in detail. Though we focus on Web applications, the learning technique is not Web specific and can be used for data validation in general.

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.