ABSTRACT

Internet, often viewed as an enormous database for data searching, is growing at a speed of millions of web pages a day. With more and more applications developed, more data formats are introduced into the increasingly complicated Internet. As a result, it is becoming increasingly difficult for a user to inquire and extract the useful data from the abyss of web pages. The wrapper technology is proposed to solve this problem. A wrapper provides users the interface for inquiring and extracting data from a web page. However, the current wrapper technology has a major drawback, that is, a wrapper can be only used for a single web page. Different wrappers are needed for different web pages. That is, when the content of a web page is changed, a different wrapper needs to be developed and used. To solve the aforementioned problem, a generalized wrapper is proposed in this paper. Based on the production provided in the web page, a generalized wrapper can analyze the web page, extract the results of the user's query, and pack the extracted data in standard XML format. The major advantage of the generalized wrapper is that the same wrapper can be used even when the content of the web page is changed. That is, the users only need to change the production, and the same generalized wrapper can be used without any changes. Keywords: semi-structured data, wrapper, production, XML.


