The Generalized Wrapper

蔡秉承、邱紹豐

E-mail: 9225036@mail.dyu.edu.tw

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.

Keywords : semistructured data ; wrapper ; production ; XML

Table of Contents

REFERENCES

1. S. Abiteboul, “ Semi-structured data: from practice to theory, ” 16th Annual IEEE Symposium on Logic in Computer Science, pp. 379-386, 2001. 2. Dan Suciu, “ Semi-structured Data and XML, ” In Proceedings of International Conference on Foundations of Data Organization, 1998. 3. S. Abiteboul, D. Quass, J. McHugh, J. Widom, and J.L. Wiener, “ The Lorel Query Language for Semi-structured Data, ” International Journal on Digital Libraries, 1(1), pp. 68-88, 1997. 4. J.Hammer, H.Garcia-Molina, J.Cho, R.Aranha, A.respo, “ Extracting Semi-structured Information from the Web, ” In Proceedings of the Workshop on Management of Semi-structured Data held in conjunction with ACM SIGMOD'97, pp. 18-25, 1997. 5. L. Liu, W. Han, D. Buttler, C. Pu, and W. Tang, “ An XML-based Wrapper Generation Toolkit for Internet Information Sources, ” In Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data (SIGMOD'99) (short paper and software demo), pp 540 - 543, 1999. 6. Naveen Ashish and Craig A. Knoblock, “ Semi-automatic wrapper generation for Internet information sources, ” In Proceedings of the Second IFCIS International Conference on Cooperative Information Systems, pp. 160-169, 1997. 7. Xiaoying Gao and Leon Sterling, “ Autowrapper: automatic wrapper generation for multiple online services, ” In Proceedings of Asia Pacific Web Conference 1999 (APWeb99), World Wide Web-Technologies and Applications for the New Millennium, pp. 229-238, 1999. 8. Alberto Mendelzon, George Mihaila, Tova Milo, “ Querying the World Wide Web, ” In Proc. PDIS'96 (Full version in Int'l Journal on Digital Libraries

1,1997), pp. 54-67,1996. 9. Sergey Brin, “ Extracting Patterns and Relations from the World Wide Web, ” WebDB Workshop at EDBT'98, 6th International Conference on Extending Database Technology, pp. 172-183,1998. 10. Smith, T. F., and Waterman, M. S, “ Identification of Common Molecular Subsequences, ” Journal of Molecular Biology, vol. 147, pp. 195-197, 1981. 11. Boris Chidlovskii, Jon Ragetli, Maarten de Rijke, “ Automatic Wrapper Generation for Web Search Engines, ” Web-Age Information Management, First International Conference, WAIM, pp. 399-410, 2000. 12. Montebello, “ Wrapping WWW information sources, ” Database Engineering and Applications Symposium, 2000 International, pp. 431 —436, 2000. 13. Y. Papakonstantinou, S. Abiteboul, and H. Garcia-Molina, “ Object Fusion in Mediator Systems, ” In Proceedings of Twentieth International Conference on Very Large Databases, Bombay, India, pp. 413-424, 1996. 14. G. Wiederhold, “ Mediators in the architecture of future information systems, ” IEEE Computer, pp. 38-49,1992. 15. S. Abiteboul, “ Querying semi-structured data, ” In ICDT, pp. 1-18,1997. 16. K. Shoens, A. Luniewski, P. Schwarz, J. Stamos, and J. Thomas, “ The RUFUS System: Information Organization for Semi-Structured Data, ” In Proceedings of Nineteenth International Conference on Very Large Databases, Dublin, Ireland, pp. 97-107, 1993.