Web資料單一化包裝器之研究

蔡秉承、邱紹豐

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

摘 要

網際網路(Internet)可視為一個龐大的資料庫，儲存於其上的資料格式也不盡相同，且資料量每天以難以想像的速度在成長，更新速度也非常快。為了查詢以及擷取從網際網路獲得的大量資訊來源，目前一般都是每個網頁都有專屬的包裝器(wrapper)，而每個包裝器只適用於查詢以及擷取其所屬的網頁。當網頁資料改變或更新時，其包裝器便不再適用於擷取其所屬的網頁。 為了解決這樣的問題，在本研究中我們設計了一個Web資料單一化包裝器 (或generalized wrapper)，使用者輸入查詢指令給包裝器，Web資料單一化包裝器根據網頁所提供的規則(production)對網頁進行解析，進而從網頁擷取出查詢的資訊，最後將擷取出來的資訊封裝成可擴展標示語言(XML)格式。而當網頁資料改變時，使用者只需要更新規則(production)就可以，不需要再重新建立新的包裝器。 關鍵詞：半結構性資料、包裝器、規則、可擴展的標示語言。

關鍵詞 : 半結構性資料 ; 包裝器 ; 規則 ; 可擴展的標示語言

目錄

參考文獻

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.