以最小元素合併結構合併XML文件之研究

洪崇恩、邱紹豐

E-mail: 9314762@mail.dyu.edu.tw

摘 要

隨著電子商務的發展，XML文件已被廣泛地運用在組織間資訊的交換及存取上，因為文件的格式及結構可由使用者自訂，因此造成在整合XML文件上的困難。本研究是藉由比較XML文件的文件類型定義(Document Type Definition, or DTD)方式來達到整合XML文件的目的。因為每個元素(ELEMENT)在文件中必須遵循DTD，因此產生一個能夠包含所有原始文件的DTD，能讓原始資料依照新的DTD的定義來達成XML文件整合的效果。 在本研究中介紹以最小元素合併結構(minimum merged structure)來合併數個XML文件的DTD，這是為了讓XML文件的DTD能夠在合併後，保留原始文件結構所提出的方法。在產生最小元素合併結構時會遭遇到元素之間的元素宣告符號(如＊、？、+等)的不同和元素因順序出現無法比對的問題，因此我們在本研究中提出一個元素宣告符號合併的轉換機制和元素依序比對的插入合併機制，來幫助DTD的最小元素合併結構產生。 透過整合後DTD將可以獲得一個整合過的新XML文件，並依照新的元素定義表示出來。經整合後的新XML文件亦保留原始的XML文件所有格式，所以在查詢上、更新上仍可使用目前的XML查詢及更新的方式不需另外再重新設計。在本研究中我們提供了DTD的結構合併規則及其相關功能之演算法，如元素宣告符號合併轉換規則表、元素依序的插入比對等演算法。除了有效的解決結構上的合併問題外，我們所提出的方法中，是一個能夠快速何DTD結構的做法。

關鍵詞 : 最小的合併結構 ; 文件類型定義 ; 結構合併

目錄

參考文獻

[1] H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and Jennifer Widom, “ Integrating and Accessing Heterogeneous Information Sources in TSIMMIS, ” In Proceedings of the AAAI Symposium on Information Gathering, pp. 61-64, 1995.

[2] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom, “ The TSIMMIS Project: Integration of Heterogeneous Information Sources, ” In Proceedings of IPSJ Conference, pp. 7-18, 1994.

[3] P. M. D. Gray , L. Kerschberg , P. J. H. King, A. Poulovassilis, “ Functional Approach to Data Management - Modeling Analyzing and Integrating Heterogeneous Data, ” Springer, 2003.

[4] Y. Papakonstantinou, P. Velikhov, “ Enhancing Semistructured Data Mediators with Document Type Definitions, ” 15th International Conference on Data Engineering , 1999.

[5] Y. Papakonstantinou, H. Garcia-Molina, J. Ullman, “ Medmaker: A Mediation System Based on Declarative Specifications, ” In International Conference on Data Engineering, pp. 132 — 141, 1996.

[6] H. Garcia-Molina , Y. Papakonstantinou , D. Quass , A. Rajaraman , Y. Sagiv , J. Ullman , V. Vassalos , J. Widom, “ The TSIMMIS approach to mediation: Data models and Languages, ” In Journal of Intelligent Information Systems, 1997.

[7] Y. Papakonstantinou, P. and Velikhov, “ Enhancing Semi-structured Data Mediators with Document Type Definitions, ” In IEEE Data

Engineering Conf., 1999.
[8] Y. Papakonstantinou, S. Abiteboul, H. Garcia-Molina. "Object Fusion in Mediator Systems," In VLDB Conference, 1996.
[9] Roy Goldman, JenniferWidom, “ Approximate DataGides, ” In Proc. Of the Workshop onQuery Processing for semistructured Data and Non-standard Data Formats, pp. 436-445, 1999.
[10] Andy S. Chiou, Chilan Lin, “ The study of index on semi-structured Data, “ Communications of Institute of Information and Computing Machinery 6, pp. 99-112, 2003.
[11] Hong Su, Harumi Kuno, and Elke Rundensteiner, “ Automating the transformation of XML documents, ” Workshop on Web Information and Data Management (WIDM'01), 2001.
[12] E. Bertino, G. Guerrini, M. Mesiti, I. Rivara, and C. Tavella, “ Measuring the Structural Similarity among XML Documents and DTDs, ” Technical Report DISI-TR-02-02, Dipartimento di Informatica e Scienze dell'Informazione, 2001.
[13] Francois Bry, Dan Olteanu, Sebastian Schaffert, “ Grouping Constructs for Semistructured Data, ” In Proceedings of DEXA , 2001.
[14] Mong Li Lee, Liang Huai Yang, Wynne Hsu, Xia Yang, “ XClust: Clustering XML Schemas for Effective Integration, ” in 11th ACM International Conference on Information and Knowledge Management (CIKM), McLean, Virginia, November 2002.
[15] S. S. Chawathe and H. Garcia-Molina. “ Meaningful Change Detection in Structrued Data, ” In SIGMOD, page 26-37, 1997 [16] S. S. Chawathe and H. Garcia-Molina. “ Change detection in hierarchically structrued information, ” In SIGMOD, page 493-504, 1996 [17] K. Zhang and D. Shasha. “ Simple Fast Alogrithms for the Editing Distance Between Trees and Related Problems. ” SIAM J. Comput. 18(6)1245-1262, DEC. 1989.
[18] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein, “ introduction to Algorithms, ” kingsinfo, 2002.