半結構性資料索引之研究

林錡嵐、邱紹豐

E-mail: 9225031@mail.dyu.edu.tw

摘 要

近年來在網際網路興起之後，傳統關聯性的資料結構並不足以描述及整合其上的異質性的資料(heterogeneous data)。這些資料包含了網頁、e-mail、討論區的文章等等。因為他們並沒有固定的架構(schema)而且它的架構被允許可以是沒有規則或是不完全，這些資料稱之為半結構性資料(semi-structured Data)。在這種新的資料儲存模式中，快速的存取資料成為資料庫研究學者的一個新的挑戰。半結構性的資料通常以物件交換模型(Object Exchange Model, OEM)的形式來表示。OEM是一種圖形架構(graph)，資料的屬性以每一條路徑中所有edge上的label來表示，而資料則儲存於每一條路徑末端的節點上。因為每一條路徑都不一定相同，傳統在關聯性資料庫中為了固定架構資料形式所提供的索引機制，並不能滿足索引圖形所需要的條件。為了提供在這樣環境下的資料庫系統能擁有更快的查詢速度，在本研究中介紹路徑合併圖(Path Merge Graph, PMG)。PMG是以圖形為基礎在半結構性資料上有效的為路徑(path)建立索引，其基本原理是以相同的路徑來表達多種不同的涵義，藉由這種方式可以降低索引所需要的空間，並藉此提升查詢所需的時間。在本研究中我們提供了索引的資料結構及其相關功能之演算法，如新增、刪除、更新索引等，提供快速存取半結構性資料的機制。

關鍵詞 : 半結構性資料 ; 索引 ; 物件交換模型 ; 路徑合併圖 ; 資料存取

目錄

參考文獻

[1]P. Buneman,“ Semistructured Data, ”In Proc. of the 6th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 117-121. May 1997.

[2]S. Abiteboul,“ Object Database Support for Digital Libraries, ”European Conference on Digital Libraries, 1997.

[3]S. Abiteboul,“ Query Semistructured Data, ”ICDT, pp. 1-18. Jan. 1997.

[4]Y. Papakonstantinou, H. Garcia-Molina, and J. Widom,“ Object Exchange Across Heterogeneous Information Sources, ”In Proc. of the 11th International Conference on Data Engineering, pp. 251-260. Mar. 1995.

[5]S. Abiteboul, P. Buneman and D. Suciu,“ Data on the Web, ”Morgan Kaufmann Publishers, 2000.

[6]S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman and J. Widom,“ The TSIMMIS Project: Integration of Heterogeneous Information Sources, ”16th Meeting of the Information Proc. Society of Japan, pp.7-8. Oct. 1994.

[7]H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y. Sagiv, J.Ullman, and J. Widom,“ The TSIMMIS Approach to Mediation: Data Models and Languages, ”In Proc. of 2th International Workshop on Next Generation Information Technologies and Systems, pp. 185-193. Jun. 1995.

[8]G. Wiederhold, “ Mediators in the Architecture of Future Information Systems, ” IEEE Computer, Vol.25, pp. 38-49, 1992.
[9]J. McHugh, S. Abiteboul, R. Goldman, D. Quass, and J. Widom, “ Lore: A database management system for semistructured data, ” Technical report, Stanford University Database Group, 1997.
[10]D. Quass, A.Rajaraman, Y. Sagiv, J. Ullman, and J.Widom, “ Querying semistructured heterogeneous information, ” In Proc. of the Fourth International Conference on Deductive and Object-Oriented Databases, pp. 319-344. Dec. 1995.
[11]S. Abiteboul, D. Quass, J. McHugh, J. Widom, and J. Wiener, “ The Lorel query language for semistructured data, ” International Journal on Digital Libraries, Vol.1, pp. 68-88. Apr. 1997.
[12]J. McHugh, J. Widom, S. Abiteboul, Q. Luo, and A. Rajamaran, “ Indexing Semistructured Data, ” Technical report, Stanford University Database Group, Jan. 1998.
[13]V. Christophides, S. Cluet and G. Moerkotte, “ Evaluating Queries with Generalized Path Expression, ” In Proc. Of the ACM SIGMOD International Conference on Management of Data, pp. 413-422. June. 1996.
[14]J. McHugh and J. Widom, “ Query optimization for semistructured data, ” Technical report, Stanford University Database Group, 1997.
[15]J. McHugh and J. Widom, “ Query Optimization for XML, ” The VLDB Journal, pp. 315-326. Sep. 1999.
[16]R. Goldman and J. Widom, “ DataGuides: Enabling query formulation and optimization in semistructured databases, ” In Proc. of the 23th International Conference on VLDB, pp. 436-445. Aug. 1997.
[17]R. Goldman and J. Widom, “ Approximate DataGuides, ” In Proc. of the Workshop on Query Processing for Semistructured Data and Non-Standard Data Formats, pp. 436-445. Jan. 1999.
[18]R. Kaushik, P. Shenoy, P. Bohannon and E. Gudes, “ Exploiting Local Similarity for Indexing Paths in Graph-Structured Data, ” 18th ICDE, pp.129-140, 2002.
[19]B. Cooper, N. Sample, M. J. Franklin, G. R. Hjaltason, and M. Shadmon, “ A Fast Index for Semistructured Data, ” In Proc. of the 27th VLDB, pp. 341-350. Sep. 2001.
[20]F. Rizzolo, A. Mendelzon, “ Indexing XML Data with ToXin, ” In Proc. of the 4th International Workshop on the Web and Databases , pp. 49-54. May 2001.