The Study of Index on Semi-structured Data

林錡嵐、邱紹豐

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

ABSTRACT

As the Internet is becoming more important and treated as the data repository, traditional relational data model is insufficient to describe and integrate the heterogeneous data on the web, such as web page, e-mail, news group documents, and so on. This kind of data is called semi-structured data since they don't have fixed schema and incomplete schema is allowed among data. As the data volume increases dramatically, a new challenge of fast data retrieval is posed to the database researchers. Object Exchange Model, or OEM, normally models the semi-structured data. OEM is a graph data structure, in which the data attributes are represented by the edges of paths and the data are stored at the end nodes of the paths. Since the paths in a graph may differ from each other, the traditional indexing systems designed for fixed-schema data in relational databases are not suitable for this new type of data type. In order to accelerate semi-structured data retrieval, in this research we provide a new concept called Path Merge Graph, or PMG. PMG is based on the graph data structure to build indices on semi-structured data. To reduce the space required to store the indices, PMG utilizes paths to allow more than one paths embedded in a single path. By reducing the index size, yet still storing enough indexing information, the overhead of search index can be reduced as well. In this research, we provide the special data structure for storing the new indices and the functions, such as insertion, deletion, and updating, to maintain the index system.

Keywords : Semi-structured Data ; Index ; Object Exchange Model ; Path Merge Graph ; Data Access ; PMG

Table of Contents

REFERENCES

[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.