The Study of Index on Semi-structured Data

Lin Shih-Ann, Chiu Shao-Feng
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

第一章 前言..............................1
  1.1 研究動機..............................1
  1.2 研究目的..............................2
  1.3 論文結構..............................3

第二章 相關研究..............................4
  2.1 Object Exchange Model...............4
  2.2 TSIMMIS與LORE............................6
  2.3 OEM上架構與索引..........................10
  2.4 Extensible Markup Language...............16

第三章 路徑合併圖............................18
  3.1 路徑合併圖設計目的........................18
  3.2 路徑合併圖基本原理........................19
  3.3 路徑合併圖概述............................20
    3.3.1 Binary Search Tree.....................21
    3.3.2 Path Graph.............................22
  3.4 方法..............................23
    3.4.1 Search演算法...........................23
    3.4.1.1 一般查詢.............................24
    3.4.1.2 特殊查詢.............................25
    3.4.2 Insertion演算法........................26
    3.4.3 Deletion演算法.........................28
    3.4.4 達例....................................29

第四章 索引選擇與PMG文件化...................35
  4.1 系統架構..............................35
  4.2 PMG文件化..............................37
    4.2.1 BST文件................................38
    4.2.2 PG文件.................................40

第五章 效能評估與實驗........................46
  5.1 PMG與DG之間差異..........................46
  5.2 PMG與DG之間關係..........................49
  5.3 實驗......................................50

第六章 應用..............................54
  6.1 電子書包架構..............................54
  6.2 PMG的應用..............................56

第七章 結論..............................57

REFERENCES


