Chapter 7

Labelling-Scheme-Based Subgraph Query Processing on Graph Data

Hongzhi Wang
Harbin Institute of Technology, China

Jianzhong Li
Harbin Institute of Technology, China

Hong Gao
Harbin Institute of Technology, China

ABSTRACT

When data are modeled as graphs, many research issues arise. In particular, there are many new challenges in query processing on graph data. This chapter studies the problem of structural queries on graph data. A hash-based structural join algorithm, HGJoin, is first proposed to handle reachability queries on graph data. Then, it is extended to the algorithms to process structural queries in form of bipartite graphs. Finally, based on these algorithms, a strategy to process subgraph queries in form of general DAGs is proposed. It is notable that all the algorithms above can be slightly modified to process structural queries in form of general graphs.

INTRODUCTION

In many applications, data is naturally modeled as a labeled directed graph. For example, the XML document of the relationship of publications and authors adapts to graph structure since one paper may have more than one author and one author may have more than one paper. A fragment of such information is shown in Figure 1. Obviously, a graph can be represented in tree structure by duplicating the nodes with more than one incoming paths. But it will result in redundancy. If the information in Figure 1 is represented with a tree, the element “author” will be duplicated.

Among the queries on a graph, the subgraph queries are widely used. A subgraph query on a graph (subgraph query for short) is to retrieve the subgraphs matching the graph pattern in the
query. For instance, a subgraph query on the graph in Figure 1 is to retrieve the names of authors with publications both in proceedings and journals. Another subgraph query on the graph in Figure 1 is to retrieve the name of the journal with an author who published papers in the conference ICDE.

It is a big challenge to process subgraph queries efficiently. All the four kinds of traditional methods of processing structural queries on tree-structured data, structural join based methods (S. Al-Khalifa& H. V. Jagadish, et. al. (2002)), holistic Twigjoin based methods (Bruno, N. & Koudas, N. et.al. (2002)), structural index based methods (Milo, T. & Suciu, D. (1999)) and subsequence matching based methods (Wang, H. & Park, S. et. al. (2003)), cannot be used to process subgraph queries.

The structural join based methods and the holistic Twigjoin based methods both depend on the labelling scheme especially for tree-structured data. The encoding scheme of graphs is totally different from that of trees. As a result, they cannot be used to process subgraph queries.

Tree-structured index is enough for the structural query processing for tree-structured data. However, structural queries on graphs require graph-structured indices. Thus, the structural index-based methods for tree-structured data cannot be used to process the subgraph queries efficiently.

The subsequence matching based methods require that the tree-structured data and the queries on the documents must be converted into sequences before query processing. It is difficult to covert a graph into a sequence because the traversal of a graph can only cover a small share of edges while the traversal on a tree can cover all the edges. Therefore, it is not easy to process subgraph queries using the subsequence matching based methods.

To process subgraph queries effectively and efficiently, a natural way is to use interval-based labeling scheme (Agrawal, R. & Borgida, A. (1989)). The reasons for choosing the interval-based labeling scheme (Agrawal, R. & Borgida, A. (1989)) are as following.

- It contains only intervals and identifications (ids). All intervals and ids are numerical values so that there is an order on them, which makes query processing easier.
- Interval-based reachability and adjacent labeling scheme are compatible so that they can be used to process subgraph queries with both reachability and adjacent edges.
- The interval-based labeling scheme can be used to process subgraph queries in various forms, e.g. graphs with circles, non-planar graphs, on general graphs.

Our proposed method is designed step by step. Firstly, a hash-based join algorithm, HGJoin, is proposed for processing reachability queries on a graph to retrieve pairs of nodes in the graph with reachability relationship between nodes in each pair. Second, the HGJoin algorithm is extended to the IT-HGJoin and T-HGJoin algorithms to process the extensive reachability queries with multiple ancestors or multiple descendants. Then, Bi-HGJoin, the combination of IT-HGJoin and T-HGJoin, is designed to process queries in form...
Related Content

Comprehension of Hierarchical ER Diagrams Compared to Flat ER Diagrams
Revital Danoch, Peretz Shoval and Mira Balabaan (2005). Information Modeling Methods and Methodologies: Advanced Topics in Database Research (pp. 241-257). www.igi-global.com/chapter/comprehension-hierarchical-diagrams-compared-flat/23017?camid=4v1a

INDUSTRY AND PRACTICE: Initiating Change in Documentation Practices

View Maintenance for Materialized Transitive-Closure Relations

Objectification of Relationships
Terry Halpin (2006). Advanced Topics in Database Research, Volume 5 (pp. 106-123). www.igi-global.com/chapter/objectification-relationships/4389?camid=4v1a