A Design of Index and Query Schema for XML Documents Based on a Concept of Database Structure

Youngrok Song¹, Sangmin Lee²

¹Institute for Information and Electronics Research, Inha University, Incheon, Korea
²Department of Electronic Engineering, Inha University, Incheon, Korea
{gateway32, sanglee}@inha.ac.kr

Abstract. In this paper, the structural information between tree nodes of XML documents is represented without any structural changes of the tree by converting that number information added to the tree into bit streams. It is also shown that other structural information can be retrieved and added to index schema in the process. In the query schema, it is possible to recover ancestor nodes through a given node’s information, not only in the case of complete path query expressions but also in the case of relative path query expressions with using the index schema. Therefore, it has the advantage of making derivative query expressions with a given query. In this process, we confirm that the response time can be minimized by conducting a bit operation between the bit streams with index and the query schema in use, and accurate results can be reached by searching only with information of record sets by nodes in index files.

Keywords: Index schema, Query schema, XML, Database structure.

1 Introduction

In this paper, query processing in diverse forms is given flexibility by adding the structural information field that is needed for the structure of the existing index files when additional structural information is added to XML documents. As a result, the given query analysis time is reduced, and thus response time to output return is reduced by using query schema based on index schema in order to improve the query processing efficiency.

2 Index and Query Schema

This paper has gone through the following processes to achieve an index schema as in figure 1. 1) Build up XML documents into trees by using DOM (document object model) trees. 2) Give sequential numbers to the built-up trees by the node of each level. 3) Trees given numbers are rebuilt-up. 4) Data to be stored in index files is
acquired from the trees. 5) Make bit streams by using the number given to each node. 6) Store the data gained from [4), 5)] in the index files.

![Index schema based on a concept of database](image)

Fig. 1. Index schema based on a concept of database

When a query is initiated by a user, then the query should be converted into bit streams by referring to the table with stored data values gained from the expanded trees of the full query route that is true of the normal form expressions of XPath given to match the database. At this point, if the query expression is given in the relative route, not in the absolute route, then the full query path should be restored. Restoration is made by accessing the structural information table of the database with the basis of the names of primitive or attributes found in the query expressions, referring to the host node information field of the given node, and repeating the restoration through the absolute route until reaching the very top root base. Once perfect query expressions are restored, they are returned to users after drawing the appropriate query results from the gained database index information.

3 Experiments and Results

Experiment data was arranged and the performance was evaluated in order to determine the validity of index and query schema to process the XML documents presented so far. The accuracy test was conducted about the general query types by using 1,000 all different XML documents, and another test was carried out to compare the performances between the experiments of existing XRel and INRIA and the methods suggested in this paper. As shown in figure 2, the accuracy was more than 94%, though the number of primitives, attributes and texts in expressions increases. The error rate was less than 6%, the reason is that with the increased complexity of documents, the number of nodes by level in XML document trees increases.

To carry out the response time experiment of query processing, queries and Shakespeare's plays (Bosak Shakespeare Collection) used in INRIA and XRel's comparison experiment in XREL were adopted for experiment data and queries. The performance evaluation of the suggested query processing method was made by measuring the time from the point when the converted query SQL questions were delivered to the database up to the time where the query results were gained as output. The process was repeated 100 times for each query, and the average values from the
performed results were used. In query Q₆ in figure 3, the query operation time was so long that it was excluded from the experiment’s graph.

Fig.2. Accuracy of retrieval results for XML queries

Fig.3. Response time of retrieval result for XML queries

4 Conclusion

This paper proposed an effective index and query schema to search large amounts of XML document structural information. The query accuracy test with 1,000 various XML documents and the query-response time measured in the experiment over Shakespeare plays whose documents are deep and complicated were conducted, and this proved the efficiency of the suggested index and query methods.

Acknowledgments. This work was supported by Key Research Institute Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2011-0018394)