ABSTRACT

In the Internet, users can share their information in a visual, linked manner by creating documents using Hypertext Markup Language (HTML). It is useful for representing information for display on a browser or other display devices. It was not adequate for the extensible, data-oriented nature of the information exchange. But, the semi-structured nature of eXtensible Markup Language (XML) format allows users to store and process data.

To process the queries over XML database, several indexing methods are used. One such method of indexing is to label the nodes of XML data. The method of assigning the labels is called labeling scheme. The label values of the nodes serve the purpose of index of the data. In the XML document, searching for information involves checking for structural relationships. The labels show the structural information like ancestor-descendant relationship among the XML nodes.

The semi-structured nature of the XML document allows order-sensitive update operations in a flexible manner. If labeling is used as an indexing method for the document, the lexicographic order of the labels used for the siblings is to be maintained. It is useful for answering the queries using
the XPath expressions over the XML documents. The labels of the existing nodes of the document are affected due to some changes. Hence, the re-computation of labeling is needed for the affected nodes. This causes the system an inconvenient one.

To avoid the re-computation problem in labeling dynamic XML, this research work proposes a system which uses three approaches to provide persistent labels to the nodes of XML document. These approaches solve the problem of re-computing the labels as well as reduce the time to generate the persistent labels. Also, they reduce the size of the indexes. The experimental results show an improvement in the effectiveness of the proposed system.

The architecture of the proposed system consists of query processing, labeling and query mapping functions. The XPath expression for querying database is given to query processing method. With the help of labeling method, the required nodes can be found and returned as a response to the query.

The first approach called, New Labeling Scheme for XML (NLSX), is to avoid the re-computation problem in labeling method. The label values are the combinations of the characters and the digits. In NLSX, these combinations are made with the digits (0-9), lower case characters (a-z) and upper case characters (A-Z). The labels generated are permanently assigned to the nodes of XML document. The labels are made such that they maintain the
lexicographic order of siblings in the document. This system reduces the size of index as well as reduces the time taken for generating the label.

To get further improvement, the second approach, called New Labeling Scheme for XML using Unicode Characters (NLSXU) is proposed in the thesis. In this scheme, persistent labels are made with digits (0-9), both lower and upper case letters and the few of the characters in Unicode character set.

The final approach, called Bit Scheme Indexing for XML (BSIX), uses binary strings to make the values of persistent labels. It avoids the re-computation problem and reduces the size of index further. The query system reduces the time taken for answering the queries.