CHAPTER 1

INTRODUCTION

This chapter introduces the history and development of eXtensible Markup Language (XML), XML web services, applications of XML, motivation and objectives of this research and architecture of the proposed system.

1.1 XML

Today, in most of the business, exchanging of data over the internet is an important activity. Many users use different formats for storing their data which creates an issue of exchanging data over the internet. The problem with Electronic Data Interchange (EDI) is its rigid and inflexible data format. The format has to be agreed before any transaction can take place. It is an expensive process. To integrate a wide range of applications on different environments, several groups have developed markup languages. The most successful of these attempts was the creation of the Generalized Markup Language (GML) (Goldfarb 2002). From GML, Standard Generalized Markup Language (SGML) was developed. SGML allows documents to describe their own format. Also, users can specify tags which can handle large and complex documents. But, it is very complex to use the features of SGML (James 1997) in the applications due to limited number of predefined tags.

Next, Hypertext Markup Language (HTML) (Bosak 1997), a simple application of SGML, was developed for publishing hypertext on the web. It is useful for designing and presenting hyperlinked, formatted
information in web browser. But, it has no capability to represent metadata, provide validation, support extensibility by users or support even the basic needs of e-business. HTML has fixed set of tags. It does not allow users to use their own tags. The checking process for validating the structure and the correct content is not possible with HTML.

To overcome the limitations of HTML, the World Wide Web Consortium (W3C) has developed XML (Bray 2004). It allows users to invent and use their own tags to better describe the data. It has a tree structure which is powerful enough to express complex data and simple enough to understand. And, it is used for representing and exchanging information over the internet whereas HTML is used only for displaying. In traditional database systems, like relational or object-oriented systems, storing and querying is easier, because the structure of data is always predefined by a schema. Rarely, the changes are needed in the schema. The occurrence of elements is defined by using occurrence characters like *, +, etc. The schema of XML does not fully describe the structure of XML document. Optional values can also be used in XML. Users can flexibly do the manipulations on XML document. Thus, the structure of XML document can be more complex and much richer than that of data in traditional databases (Deitel 2001). The nature of XML allows an arbitrary number of occurrences and nested elements in an infinite number of levels.

1.2 XML AND WEB SERVICES

Ron Schmelzer et al (2011) have been discussed various schemes available for performing web services. In distributed computing on the internet, the fundamental building block is XML web services. Applications are constructed using multiple XML web services from various sources that work together regardless of where they reside or how they were implemented
XML web services provide a platform for application integration. The features are listed as below:

It exposes useful functionality to users in the web through a standard web protocol like Simple Object Access Protocol (SOAP).

It provides a way to describe the interfaces which allows a user to build a client application to talk to them. Web Services Description Language (WSDL) provides this description.

These are registered with Universal Discovery Description and Integration (UDDI), so that potential users can find them easily.

Thus, an XML web service can be defined as a software service exposed on the web through SOAP, described with a WSDL file and registered in UDDI.

1.3 APPLICATIONS OF XML

XML is used in many different areas. Few applications are given below:

Electronic Commerce

Communication between companies can be simplified by using XML. Using EDI in business requires additional technical equipment which is very expensive. Standardization of E-Commerce is possible with XML.

The eXtensible Business Reporting Language (XBRL) represents a new methodology for data information exchange. It is useful for preparing financial information, intermediaries in those preparation and distribution processes etc. It enhances the creation, exchange and comparison of business reporting information.
Multimedia

Scalable Vector Graphics (SVG) is a family of specifications of an XML-based file format. It is used for describing two-dimensional vector graphics, both static and dynamic. SVG images and their behaviors are useful in searching, indexing etc. SVG images can be created and edited with any text editor, but it is often more convenient to create these types of images with drawing programs. Virtual Reality Modeling Language (VRML) uses standard XML file format. It is useful for representing 3-Dimensional (3D) interactive vector graphics, designed particularly with the World Wide Web (WWW) in mind. Synchronized Multimedia Integration Language (SMIL), a markup language is used for describing multimedia presentations.

Science

Bio-informatics Sequence Markup Language (BSML) is proposed for communicating bioinformatics data. Mathematical Markup Language (MathML) makes mathematical formulae available for the web. Using Chemical Markup Language (CML), complex chemical molecules can be described easily. Also, many markup languages use XML format.

Wireless Markup Language (WML), is an XML derivative model for tagging content presentation on mobile handsets. Wireless Application Protocol (WAP) is a specification for the delivery and presentation of information and telephony services via wireless networks on wireless devices. These devices include mobile phones or other wireless devices. The WML script is a part of WML specification. It is widely used for creating scripts like client-side validation, customized interface design etc. Also, VoiceXML is a XML-based format that allows web sites for delivering the voice messages to users over telephones.
Health Level 7 (HL7) is XML based format, which covers the need for Electronic Patient Records (EPR), prescriptions, and medical insurance details etc. It is also used for medical imaging.

1.4 MOTIVATION FOR THE RESEARCH

One of the challenges in organizations is exchanging data between them over internet. They use different formats and semantics of meta-data to describe the data. Due to the platform independent nature of XML, it is easy to describe, store and exchange information. Information written in XML can be transmitted regardless of software and hardware used. The day to day activities includes manipulation and querying of data.

To process an XML document, a tree structure of the document is created in memory. The tree structure represents the hierarchical structure of the document. It includes one root element, child elements, attributes, processing instructions, comments etc. It depicts the ancestor-descendant relationships between the elements. An XML node represents an element of the XML document. Through proper Application Programming Interface (API), tree traversal can be made to process the data. To speed up the query processing, various indexing schemes are used. XML query processing methods involve various numbering and labeling schemes (Kaplan 2004) (Yu 2004) (Yun 2008). In labeling method, nodes are assigned with a string value which is used as an index. Labels provide the structural relationship among the elements in the document. In this thesis, the terms index and label refers to the same and are used interchangeably. The size of index grows too high, while using large size documents. Handling index of large size document becomes problem in query processing. A document may be static or dynamic in nature. For static XML, labels of the nodes are persistent and never changed. Semi-structured nature of XML allows user to modify the data in a flexible manner. Changes in the hierarchical structure of the document
results in changes of document index. (Korman 2007). Size of the index of a
document is based on the nature of that document. That is, it is based on the
fan out and depth of the hierarchical structure used in the document.

In dynamic XML, any change in the structure of XML document
affects the labels of the existing nodes (Jiaheng 2004). The changes in the
label values of the nodes can be done through re-computation process. This
process can also be known as re-labeling. In this thesis, both words
‘re-labeling’ and ‘re-computation’ refers to the same and are used
interchangeably. The process of re-computation of labels is a time consuming
and tedious one. Suppose, a document is frequently updated, then the re-
computation of labels occupies more time than that of query processing.
Providing persistent labels to the nodes become a problem in dynamic XML.

Also, time taken for generating persistent labels and responses for
the queries has a significant impact on the performance of the system.

1.5 ARCHITECTURE OF THE SYSTEM

The overall architecture of the proposed system is given in
Figure 1.1. The labeling method generates persistent labels for the nodes of
XML document. Varieties of methods are used for generating labels. Few
labeling methods use letters and digits to make labels. The combinations of
labels are made in such a way that it always maintains a specific order of the
nodes. The database is stored in a relational model along with its labels. The
label of a node is used as an index for the database. XPath expression is given
to the query processing method. The query mapping process maps the XPath
expression into an equivalent Structured Query Language (SQL) statement
which is useful for answering the query. With the help of the labels, the
required nodes can be found. And, the matches found can be returned as a
response to the user.
Similarly, modification of the nodes can be done with the help of labeling method. XML nodes can be modified through insert, delete and update operations. Inserting a node or a sub tree can change the hierarchical structure of the system. The changes in XML nodes affect the lexicographic order of labels assigned to the siblings. This can be avoided by using a suitable labeling method. The labeling method has to maintain the lexicographic order of the siblings in a document while changes are made. This is same for deletion too. Any such modifications affect the order of nodes in the document. However, updating the text contents of a node does not affect the hierarchical structure of the document.

The relationships like ancestor-descendant, parent-child among the nodes have been preserved such that the generated labels are useful for query processing. Also, the label of a node provides the information about the level of a node.

1.6  OBJECTIVES OF THE RESEARCH

To manage data efficiently in dynamic XML, many labeling schemes have been already addressed in literature. To overcome some of the issues (Amagasa 2003), (Fomichev 2006) (Li 2005) (Li 2006a) (O’Neil 2004) in the labeling methods, a different way of labeling is suggested in this proposal. The following major issues in the labeling schemes for dynamic
XML with order-sensitive updates have been addressed by the proposed scheme.

- Avoiding re-computation of labels
- Reducing the size of index
- Reducing time taken for generating labels
- Improving the response time for query processing

The thesis is orchestrated as follows:

Chapter 2 introduces syntax of XML, parsing and validating XML documents and query processing on XML. It provides the basic knowledge about XML.

Chapter 3 focuses the objective of labeling methods, re-computation problem, the different labeling methods used in the existing systems and their shortcomings.

Chapter 4 proposes a New Labeling Scheme for XML (NLSX) which avoids the re-computation of labels of the existing nodes. Here, the labels are created by combining digits (0-9), lower case characters (a-z) and upper case characters (A-Z). Such combinations assign persistent labels to the nodes of XML document and also, maintain the lexicographic order of the siblings in the document. This system reduces the size of the index as well as the time taken for generating the same. The response time for answering queries has been analyzed. Reusability of the deleted nodes has also been examined.

Chapter 5 suggests, a New Labeling Scheme for XML using Unicode characters (NLSXU), which improves the performance of the system
further. Here, the persistent labels are made with digits (0-9), both lower and upper case letters and a few of the characters in Unicode character set. The analysis has also been made on the size of the index, time for generating index and query processing.

Chapter 6 develops a Bit Scheme Indexing for XML (BSIX) which uses the binary strings to make the values of persistent labels. By doing so, the problem of re-computation has been avoided and also the size of index has been reduced further. Also, the time taken for querying has been reduced.

Chapter 7 provides the summary of the thesis. It also discusses further work possible on approaching other variations of the labeling problems.