Chapter 5
Index Structures for XML Databases

Samir Mohammad
Queen's University, Canada

Patrick Martin
Queen's University, Canada

ABSTRACT

Extensible Markup Language (XML), which provides a flexible way to define semistructured data, is a de facto standard for information exchange in the World Wide Web. The trend towards storing data in its XML format has meant a rapid growth in XML databases and the need to query them. Indexing plays a key role in improving the execution of a query. In this chapter the authors give a brief history of the creation and the development of the XML data model. They discuss the three main categories of indexes proposed in the literature to handle the XML semistructured data model and provide an evaluation of indexing schemes within these categories. Finally, they discuss limitations and open problems related to the major existing indexing schemes.

INTRODUCTION

XML is becoming the dominant method of exchanging data over the Internet. It was endorsed as a W3C recommendation in 1998 (Bary, Paoli, & Sperberg-McQueen, 1998). Its roots go back to SGML (Standard Generalized Markup Language) (Bary et al., 1998). XML poses a nested hierarchical nature. An example XML document is illustrated in Figure 1. It is based on DBLP (The DBLP Computer Science Bibliography, 2009), a popular computer science bibliography dataset. The data-tree shape in Figure 2 represents the data in the XML document of Figure 1.

The growth in the use of XML for data exchange has led to the introduction of native XML databases to store and manage the data directly in its XML representation. Since the repetition of XML data is irregular due to missing and/or repeated arbitrary elements, its storage structure can be scattered over many different locations on the disk, which decreases the performance of XML queries (Chung, Min, & Shim, 2002). Furthermore, the flexibility of specifications of the XML queries (e.g. use of

DOI: 10.4018/978-1-61520-727-5.ch005

Figure 1. DBLP like XML document

```
<Bib>
    <book>
        <author>Tim</author>
    </book>
    <paper> </paper>
    <paper>
        <author>Sarah</author>
    </paper>
    <paper @reviewer="Ahmad">
        <author>Wang</author>
    </paper>
</Bib>
```

wild cards) adds to the challenge of indexing methods (Wang, Park, Fan, & Yu, 2003; Zou, Liu, & Chu, 2004).

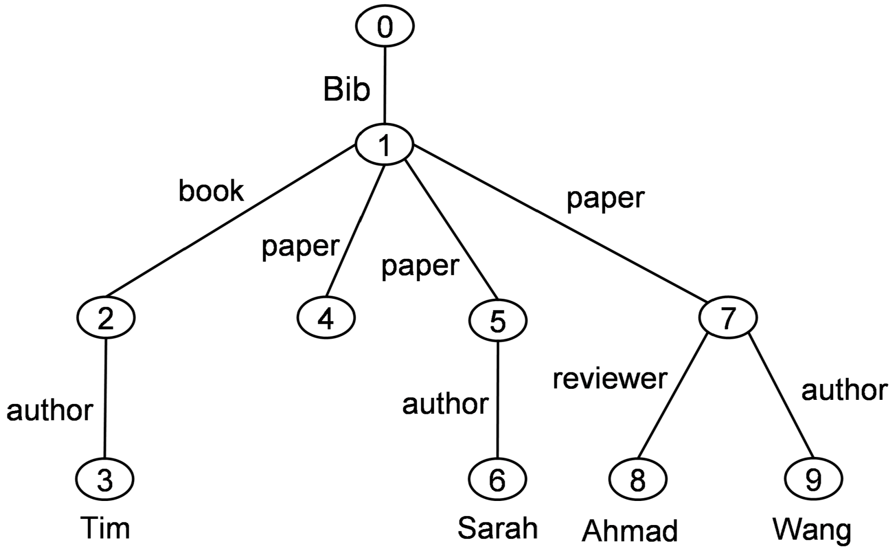
The best way to judge the strength of an indexing technique is to compare it with other techniques using common criteria that are applicable for all of them and can act as a benchmark. The main contributions in this chapter are:

- A set of common criteria to summarize the characteristics of the most popular indexing techniques used for XML databases.
- Classify novel classification of graph indexes that is based on the presence/degree of determinism and the bisimilarity direction(s) of indexing, which control the size of an index and its query answering power, respectively.

In the remainder of this chapter we discuss a number of approaches to XML indexing. We first give an overview of XML data models and the XPath query language. We next explain the three types of indexing techniques used for XML data, namely, Node index scheme, Graph index scheme, and Sequence index scheme; and compare approaches of each type. We divide the comparison criteria into four basic groups:

- **Retrieval power:** Which includes the precision and completeness of the result, and the type of queries supported.
- **Processing complexity:** Which involves the need to compute the relationship between elements (such as the parent/child and the ancestor/descendent relationships), the need for structural joins to answer a

Figure 2. Edge-labeled data-tree


0
Bib
1
book
paper
paper
paper
2
4
5
7
author
author
reviewer
author
3
6
8
9
Tim
Sarah
Ahmad
Wang