Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 115 (2017) 804–811

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Advances in Computing & Communications, ICACC-2017, 22-24 August 2017, Cochin, India

DGReLab+: Improving XML Path Query Processing by Avoiding Buffering Irrelevant Results

Samini Subramaniam*[a], Su-Cheng Haw[a], Lay-Ki Soon[a]

[a]*Multimedia University, Jalan Multimedia, Cyberjaya, 63000, Selangor, Malaysia*

Abstract

The impulsive headway in the use of XML attracted researchers to conduct researches on the optimization techniques of XML data management. One of the aspects that have been a challenge since then is the effective processing of user queries. Although many techniques have been proposed in the past, these techniques still suffer from large overhead due to buffering irrelevant results before producing the final output. Thus, in this paper, we propose a path query processing technique, DGReLab+, which evaluates queries without buffering irrelevant results. The evaluations revealed that DGReLab+ outperformed two other techniques which are TwigStack and QTwig.

Peer-review under responsibility of the scientific committee of the 7th International Conference on Advances in Computing & Communications.

Keywords: XML; Node Labeling; Annotation; Query Optimization; Path Query Processing

1. Introduction

eXtensible Markup Language (XML), is a semi-structured document that has been a crucial standard for data storage and exchange over the World Wide Web (WWW). Nested and tagged nodes are the building blocks of an XML document [1] (see Fig. 1.). The flexibility and self-descriptive structure of an XML document have made it as a great potential for data management. Accordingly, efficient query processing plays a vital role in data-centric

* Corresponding author. Tel.: +603-8312-5340; fax: +603-8312-5264.
E-mail address: samini.subra@mmu.edu.my

1877-0509
Peer-review under responsibility of the scientific committee of the 7th International Conference on Advances in Computing & Communications
10.1016/j.procs.2017.09.157

applications [2]. For this reason, the structural relationships like parent-child (P-C), ancestor-descendant (A-D), sibling, and level information among nodes needs to be preserved proficiently.

a

```
<item>
  <location></location>
  <quantity></quantity>
  <name></name>
  <payment></payment>
  <description>
    <parlist>
      <listitem>
        <text>
          <keyword></keyword>
        </text>
      </listitem>
    </parlist>
  </description>
  <shipping></shipping>
  <mailbox>
    <mail></mail>
  <mailbox>
</item>
```

b

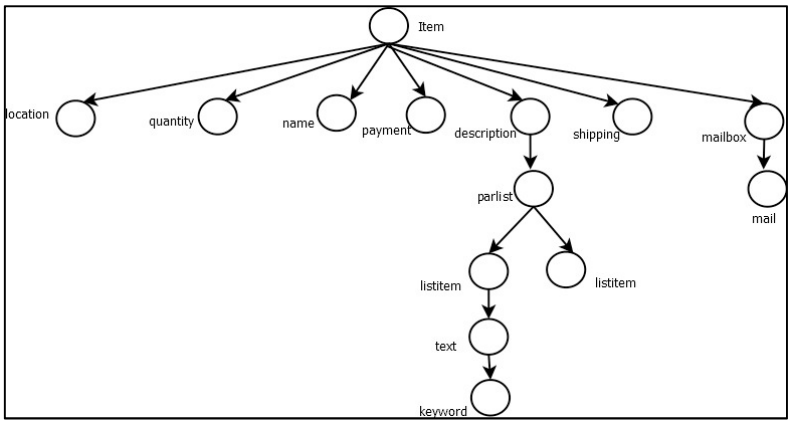

Fig. 1. (a) example of an XML document; (b) tree view of the XML document in (a).

Over the decades, there were many different types of node labeling techniques proposed by many researchers. Some of the significant properties of good labeling schemes are complexity, association, and correctness. The amount of computations required to generate distinct labels affects the performance of the labeling technique. The relationships between two nodes must be determined quickly and easily simply by examining their labels. Next, the association between nodes in the XML tree must be preserved through the labeling technique. It will be challenging for a query processing engine to identify relevant nodes of a query if the relationships are not well-preserved. Finally, the correctness of the results depend heavily on the association between the nodes. If the relationships between nodes are not properly preserved, then the query processing engine will not be able to produce flawless results.

Besides this, there were many studies done in the past to develop an efficient query processing technique using node labeling schemes. However, it is certainly crucial to develop a technique which can process queries as fast as possible, with less complexity and able to produce flawless results. Thus, in this paper, we propose a query processing technique named DGReLab$^+$, which adopts the concept of DataGuide [3] and uses a labeling technique, ReLab$^+$ [4].

The rest of the paper is organized as follows. Section 2 summarizes some of the existing works done in the past. Section 3 describes the proposed technique. Section 4 presents the experimental evaluations, while Section 5 concludes the paper.

2. Related Works

Extensive research works were conducted in the past to enhance the performance of XML query processing while producing seamless results. Some of those techniques will be discussed in this section.

One of the techniques that have caught researchers' attention was TwigStack[5]. This holistic query processing algorithm was proposed with the intention to avoid buffering large intermediate results. TwigStack was found to be optimal for queries with A-D edges but suboptimal for queries with P-C edges.

On the other hand, DataGuide summarizes all paths in the XML document that starts from the root. Each path in DataGuide is unique. This helps to match the exact path in the query, which helps to process queries in a swift manner. In spite of this, DataGuide was not designed to process queries with complex path expressions. This is