Chapter 4

Efficient Identification of Similar XML Fragments Based on Tree Edit Distance

Hongzhi Wang  
Harbin Institute of Technology, China

Jianzhong Li  
Harbin Institute of Technology, China

Fei Li  
Harbin Institute of Technology, China

ABSTRACT

Similarity detection between large XML fragment sets is broadly used in many applications such as data integration and XML de-duplication. Extensive methods are used to find similar XML fragments, such as the pq-gram state-of-the-art method which allows for relatively high join quality and efficiency. In this chapter, we propose pq-hash as an improvement to pq-grams. As the base of pq-hash, a randomized data structure, pq-array, is developed. With pq-array, large trees are represented as small fixed sized arrays. To efficiently perform similarity join on XML fragment sets, in this chapter we propose a cluster-based partition strategy as well as a sort-merge & hash join strategy to avoid nested loop join. Both our theoretical analysis and experimental results confirm that, while retaining high join quality, pq-hash gains much higher efficiency than pq-grams, and our strategies for approximate join are effective.

INTRODUCTION

Thanks to its ability to represent data from a wide variety of sources, XML has rapidly emerged as the new standard for data representation and exchange on the Internet. Given the flexibility of XML, data in autonomous sources which represent the same real-world object may not be exactly the same. Thus, similarity detection techniques are often applied to find XML fragments representing the same real-world object. In this chapter we refer to a real-world application in the Municipality of Bozen in order to illustrate the use of techniques for detecting similar XML fragments. In this con-
text, the GIS office in that municipality maintains maps of the city area, so that one would like to enrich such maps with information retrieved from various databases of the municipality as well as external institutions. Residential addresses turn out to play a pivotal role in this process since they have to be used to access and link relevant information. However, performing exact join on the street names would yield poor results since street names are different in different databases due to, e.g., spelling mistakes, different naming conventions, and renamed streets which are not always updated in all databases. Moreover, in the bilingual region of Bozen, each street has typically two names, and these are often used interchangeably. Since these data can be modeled as ordered, labeled trees, methods for the detection of similar XML fragments can be used to effectively match the data representing the same real-world data.

A widely used approach to the evaluation of similarity between XML documents is to compute their edit distance (Cobena, Abiteboul, & Marian, 2002; Guha, Jagadish, Koudas, Srivastava, & Yu, 2002; Lee, Choy, & Cho, 2004). Since XML documents are often modeled as ordered, labeled trees, the tree edit distance between any two such trees is defined as the minimum number of node insertions, deletions and relabelings to transform a tree into another (Tai, 1979). It is well-known that the edit distance behaves well but is computationally expensive. Many works such as (Zhang & Shasha, 1989; Klein, 1998; Chen, 2001; Demaine, Mozes, Rossman, & Weimann, 2007) have been proposed to improve the efficiency in the computation of tree edit distance. However, all of them have more than \(O(n^3)\) runtime, where \(n\) is the tree size, and hence they do not scale to large trees. Since it is hard to improve the efficiency fundamentally by optimizing the tree edit algorithms independently, transformation-based methods are often adopted in order to transform trees into other data structures whose similarities are easier to evaluate.

An example of computation of tree edit distance is shown in Figure 1. To transform the XML fragment \(T\) to the XML fragment \(T_3\), the following operations are made: insertion of node \(i\), relabeling of node \(c\) with \(x\), and deletion of node \(g\). Since this is the minimum cost transformation sequence, the tree edit distance between \(T\) and \(T_3\) is equal to 3.

The \(pq\text{-gram}\) method is known as an effective and efficient transformation-based tree similarity detection method (Augsten, Böhlen, & Gamper, 2005). In this method, each tree is split into a small subtree bag. The \(pq\)-distance between the \(pq\)-bags is used to describe the distance between their corresponding trees. Both theoretical analysis and experimental results confirm the detection quality based on \(pq\text{-grams}\). An “optimized join” (Augsten, Böhlen, Dyreson, & Gamper, 2008) was also presented to accelerate the detection process by taking advantage of the diversity of trees in a forest. However, this optimized join

---

**Figure 1.** Example of computation of tree edit distance computation: the left-most tree is transformed into the right-most tree.