



Tree Edit Distance and Maximum Agreement Subtree



Kilho Shin¹

GSAI, University of Hyogo, 7-1-28 Minatojima-Minami, Chuo, Kobe 6500047, Japan

ARTICLE INFO

Article history:

Received 12 May 2014

Received in revised form 29 August 2014

Accepted 1 September 2014

Available online 8 September 2014

Communicated by M. Yamashita

Keywords:

Graph algorithms

Pattern recognition

Tree edit distance

Maximum agreement subtree

ABSTRACT

This paper presents an interesting relation between the Maximum Agreement Subtree (MAST) problem and the Tree Edit Distance (TED) problem, both of which have been intensively studied in the literature. To be specific, we show that, for an arbitrary tree edit distance metric that is a derivative of the Taï tree edit distance metric, there exists a MAST-like pattern extraction problem, named Mostly Adjusted Sub-Forest (MASF) problem, such that computing a distance between trees x and y is equivalent to finding an optimal pattern shared between x and y . The MASF problem is different from the MAST problem in: (1) A pattern of the MASF problem may be a forest instead of a tree; (2) Instead of requiring exact match of labels, a pattern of a MASF problem is multi-labeled, and our flexible matching rule only requires that the label set of a vertex of a pattern includes the label of the corresponding vertex in a target tree; (3) To control ambiguity of matching caused by the flexible rule, the objective function includes a penalty function. As an application of this generic framework, we equate the Lu^* tree edit distance metric with a pattern extraction problem, named the Mostly Adjusted Agreement Sub-Tree (MAAST) problem. The MAAST problem aims to find optimal agreement subtrees under the flexible matching rule and is solved in $O(n^2\sqrt{d}\log d)$ -time, where n and d are the size and the minimum degree of the input trees. The MAST problem requires $O(n^{2.5})$ -time computation.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

There are many applications of data analysis where the target objects are represented as trees. For example, tasks that aim to learn some hidden structure from a text string are an important focus of natural language processing problems, and *parse trees* are typical representation of such target structures; *Mark-up languages* such as HTML and XML define tree generation syntax, and the resulting documents are naturally dealt with as trees; In biochemistry and structural biology, *secondary structures* of biopolymers such as proteins and nucleic acids have significant meanings, and, representing them as trees certainly yield practical advantages; In evolutionary biology, *evolutionary*

trees are used to represent relationships among biological species.

The Maximum Agreement Subtree (MAST) problem and the Tree Edit Distance (TED) problem have been intensively studied in the literature as fundamental tools to investigate trees. The MAST problem is a pattern extraction problem that aims to find the largest shared pattern, while the TED problem aims to measure similarity between trees using a metric function. We start with a quick review of the MAST and TED problems.

In this paper, by a tree, we mean a *rooted, unordered and labeled* tree: When Γ_x denotes the entire set of vertices of a tree x , the generation (top-to-bottom) order makes Γ_x a partially ordered set (poset); $v \leq w$ means either w is an ancestor of v or $v = w$; $v < w$ means $v \leq w$ but $v \neq w$; A label $\ell(v)$ is assigned to every vertex v . In some applications like evolutionary trees, only a part of vertices have labels. In such cases, we can assume that those vertices

E-mail address: yshin@ai.u-hyogo.ac.jp.

¹ Tel.: +81 78 303 1901.

Download English Version:

<https://daneshyari.com/en/article/10331109>

Download Persian Version:

<https://daneshyari.com/article/10331109>

[Daneshyari.com](https://daneshyari.com)