Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

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*

http://dx.doi.org/10.1016/j.ipl.2014.09.002 0020-0190/© 2014 Elsevier B.V. All rights reserved. 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); $v \le w$ means either w is an ancestor of v or v = w; v < w means $v \le w$ but $v \ne w$; 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.





Fig. 2. An example of edit scripts and mappings for Taï distance. To convert x into y, this edit script deletes five vertices and inserts ten vertices. The five dotted arrows indicate the substitution edit operations, and three of them actually change labels (numbers in circles). Therefore, the cost of this edit script turns out to be 5 + 10 + 3 = 18.

without labels commonly have the null label [4]. Furthermore, the notion of *nearest common ancestors* of vertices is defined as follows.

Definition 1 (Nearest common ancestors). For a tree x and $v, w \in \Gamma_x$, the nearest common ancestor $v \smile w$ of v and wis the minimum element of the non-empty totally ordered set $(\{u \mid v \le u, w \le u\}, \le)$.

Given a set of trees $\{x_1, \ldots, x_n\}$, an agreement subtree t is a tree such that there exist inclusion mappings $\phi_i : \Gamma_t \rightarrow$ Γ_{x_i} for i = 1, ..., n that satisfy $v < w \Leftrightarrow \phi_i(v) < \phi_i(w)$, $\phi_i(v \smile w) = \phi_i(v) \smile \phi_i(w)$ and $\ell(v) = \ell(\phi_i(v))$ for any v and w in Γ_t (Fig. 1). Then, the MAST problem is defined as follows.

Maximum Agreement Sub-Tree (MAST) Problem. Find $\tilde{z} \in \arg \max\{|\Gamma_z| \mid z \text{ is an agreement subtree}\}$ of $x_1, ..., x_n$.

Thus, the maximum agreement subtree found is the largest common pattern that relates and characterizes x_1, \ldots, x_n .

The Taï tree edit distance metric [6], on the other hand, is an important instance of the tree edit distance metric,

from which many derivatives have stemmed in the literature.

Given trees x and y, an edit script from x to y is a sequence of edit operations that converts x into y. Each edit operation is one of the following: (1) Substitution $\langle x' \rightarrow y' \rangle$ replaces $x' \in \Gamma_x$ with $y' \in \Gamma_y$; (2) Deletion $\langle x' \to - \rangle$ deletes $x' \in \Gamma_x$ from x and re-define the children of x' as children of the parent; (3) Insertion $\langle - \rightarrow y' \rangle$ inserts $y' \in \Gamma_y$ between a vertex and a (sub)set of children of the vertex. By sequentially applying the operations of an edit script to x_i we obtain y.

Furthermore, a cost is given to each edit operation. Although there are multiple different settings, we deploy the most common one: The costs of $\langle x' \to y' \rangle$ is $1 - \delta_{\ell(x'), \ell(y')}$, and the deletion and insertion cost 1. The cost $\gamma(\sigma)$ of an edit script σ is the sum of the costs of the edit operations of σ . Finally, the Taï Tree Edit Distance problem is formulated as follows.

Taï Tree Edit Distance (TTED) Problem. Find $\tilde{\sigma} \in \operatorname{arg\,min}\{\gamma(\sigma) \mid \sigma \text{ is an edit script from }$ x to y $\}$.

The Taï distance between x and y is defined as $d_{\mathsf{T}}(x, y) = \gamma(\tilde{\sigma}).$

Download English Version:

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

Download Persian Version:

https://daneshyari.com/article/10331109

Daneshyari.com