Traverse $D_h$:

T1: Set $level[h] = empty_list$;
T2: traverse $D_h$ in pre-order;
T3: on arrival at a vertex $d \in D_h$ such that $Tvertices[d] \neq \emptyset$ do
T3.1: $\max \xi = \max \xi + 1$;
T3.2: for all $v \in Tvertices[d]$ do
T3.2.1: $\xi[v] = \max \xi$;
T3.2.2: append $v$ to $level[h]$.

4 Hopcroft and Tarjan’s Canonical Order

The canonical order of Section 2.1 is defined from the leaves up. The same holds for the canonical indexes. As a result, two vertices $u$ and $v$ have the same canonical index if and only if $T_u$ and $T_v$ are isomorphic.

Hopcroft and Tarjan’s work [2] is aimed at tree isomorphism rather than subtree isomorphism. In their construction, the depth—the distance from the root—also distinguishes between vertices. Consequently, they assign the same index to vertices $u$ and $v$ if and only if they have the same depth and $T_u$ and $T_v$ are isomorphic (see Figure 1(b) for an example of their index).

This choice enables Hopcroft and Tarjan to overcome the difficulty (as mentioned in Section 2.3) of using bucket sort. They process the tree vertices by phases: in phase $i$ they process all vertices of depth $d_{\max} + 1 - i$. Since indexes are given in increasing order, the indexes of all the vertices of depth $d$ form a set of consecutive integers. Notice that all children of vertices at the depth $d$ are at the depth $d + 1$. Thus, the sets of children in distinct phases are disjoint. Therefore, the sum of their sizes is at most $n$, not $\Theta(n^2)$, as in the naive height oriented approach of Section 2.3.

Hopcroft and Tarjan’s algorithm yields a refinement of the subtree isomorphism classes, as generated by Zemlyachenko’s canonical indexes. Given the subtree isomorphism classes, it is easy to construct Hopcroft and Tarjan’s classes. Other than applying our algorithm, we do not know how to perform the reverse direction.

References


Initialization

Scanning $T$ from the leaves inwards,
determine for each $v \in V(T)$ the value of $\text{height}[v]$, and $\text{parent}[v]$;
For $h = 0, \ldots, \text{height}(T)$

determine $\text{level}[h]$;
Set $\text{children}[h] = \text{empty list}$;

The main loop
For $h = 0, \ldots, \text{height}[T]$ do

Phase $h$

Phase 0
For all $v \in \text{level}[0]$ /* the leaves of $T$ */
set $\xi[v] = 1$ and append $v$ to $\text{children}[\text{height}[\text{parent}[v]]]$;
set $\text{max}\xi = 1$

Phase $h$ ($h \geq 1$)
Assumptions: We have already constructed $\text{children}[h]$.

M1: Construct $D_h$ from $\text{children}[h]$. (See below.)
M2: Traverse $D_h$ to determine $\xi[v], v \in \text{level}[h]$,
and construct $\text{level}[h]$. (See below.)
M3: for $v \in \text{level}[h]$ do /* $\text{level}[h]$ should be processed by increasing order */
append $v$ to the list $\text{children}[\text{height}[\text{parent}[v]]]$.

Construct $D_h$ from $\text{children}[h]$.

C1: Initially $D_h$ consists of a single vertex—root.
C2: Let $\text{Tvertices[root]} = \text{level}[h]$;
C3: for $v \in \text{level}[h]$ do
C3.1 $D\text{vertex}[v] = \text{root}$;
C4: for $u \in \text{children}[h]$ do /* $\text{children}[h]$ should be processed by increasing order */
C4.1: let $v = \text{parent}[u]$;
C4.2: let $d = D\text{vertex}[v]$;
C4.3: if $d$ is a leaf or $\text{label[rightmost}[d]] \neq \xi[u]$ then
C4.3.1: add a new child $d'$ to $d$;
C4.3.2: $\text{label}[d'] = \xi[u]$;
C4.3.3: $\text{rightmost}[d] = d'$;
C4.4: else /* $\text{label[rightmost}[d]] = \xi[u]$ */
$d' = \text{rightmost}[d]$;
C4.5: move $v$ from $d$ to $d'$ /* update $D\text{vertex}[v], \text{Tvertices}[d]$ and $\text{Tvertices}[d']$ */.

Zemlyachenko

August 3, 1998
down the tree.

We now scan all the children of \( \text{level}[h] \) by nondecreasing index. When considering a vertex \( u \), we move its parent \( v \) from \( d \), the D-vertex to which it was associated, to \( d' \), the child of \( d \) whose label is equal to \( \xi'[u] \). If \( d \) has no such child, a new child is inserted. Since the children of \( \text{level}[h] \) are processed by nondecreasing index, \( d' \) has the largest label of all the current children of \( d \). (This observation eliminates the need to explicitly sort the children of vertex \( d \), thus allowing efficient insertions into \( D \).

When completed, the structure of \( D_h \) reflects the canonical order of the vertices of \( \text{level}[h] \): Let \( u, v \) be vertices of \( \text{level}[h] \) and let \( d_u \) (\( d_v \)) be the D-vertex associated with \( u \) (\( v \)). \( T_u \) is isomorphic to \( T_v \) if and only if \( d_u = d_v \). Also, \( u \) precedes \( v \) in the canonical order if and only if \( d_u \) precedes \( d_v \) in preorder. Therefore, we traverse \( D_h \) in preorder, assign the next consecutive integers as indexes to the vertices of \( \text{level}[h] \) and construct the sorted list \( \text{level}[h] \) of these vertices.

Finally, we have to prepare the lists \( \text{children}[h+1], \ldots, \text{children}[\text{height}(T)] \) of the children of vertices in levels \( h+1, \ldots, \text{height}(T) \). In iteration \( h \), we scan \( \text{level}[h] \): Let \( v \in \text{level}[h] \), \( f \) be the parent of \( v \) and \( j > h \) be the level of \( f \). We put \( v \) at the end of the list of the children of \( \text{level}[j] \). Note that while processing \( \text{level}[h] \), we will have partial lists of children of levels \( h+1, \ldots, \text{height}(T) \). The list of children of \( \text{level}[j] \) will be completed only after processing \( \text{level}[j-1] \). Since the vertices of \( \text{level}[h] \) have greater indexes than those of \( \text{level}[h'] \) for any \( h' < h \), and the vertices of \( \text{level}[h] \) are sorted by nondecreasing index, the lists of children are also sorted by nondecreasing index.

### 3.2 Detailed Description of the Algorithm

**Data structures:**

1. For each vertex \( v \in T \) we have a record with the following fields:
   - \( \xi \): the index of \( v \).
   - \( \text{parent} \): the parent of \( v \).
   - \( \text{height} \): the height of \( v \).
   - \( D\text{vertex} \): the vertex of \( D_h \) which currently contains \( v \).

2. For each vertex \( d \in D_h \) we have the following fields:
   - \( T\text{vertices} \): the set of vertices of \( T \) associated with \( d \).
   - \( \text{rightmost} \): the last child of \( d \).
   - \( \text{label} \): each vertex \( v \in T\text{vertices}[d] \) was moved from the parent of \( d \) to \( d \) when processing a vertex \( c_v \), which is the \( \text{depth}(d) \)-th child of \( v \) and has index \( \xi[c_v] \); all such children have the same index, the value of which is recorded in \( \text{label} \).

3. For \( h = 0, \ldots, \text{height}(T) \) we have:
   - \( \text{level}[h] \): The set of the vertices of height \( h \).
   - \( \text{level}[h] \): The vertices of \( \text{level}[h] \) sorted by nondecreasing index.
   - \( \text{children}[h] \): The children of vertices of \( \text{level}[h] \) sorted by nondecreasing index.
Figure 1: Bad tree for algorithm.

Figure 2: (a) The tree $T$, $level[2] = (u_1, u_2, u_3, u_4)$. (b) The tree $D_2$. 

Zemlyachenko

August 3, 1998
2.2 Canonical Index

We extend the $\leq_C$ order to subtree isomorphism classes, i.e., $S \leq_C S'$ if and only if for any $v \in S$ and $v' \in S', v \leq_C v'$.

In order to quickly check whether $u \leq_C v$ we assign to each vertex $w$ a canonical index $\xi[w]$. Let $S_1, \ldots, S_m$ be the subtree isomorphism equivalence classes ordered by the canonical order, i.e., $V(T) = \bigcup_{i=1}^m S_i$ and $S_1 \leq_C S_2 \leq_C \cdots \leq_C S_m$. The canonical index of a vertex $w$ is the index of the class to which it belongs, i.e., $\xi[w] = j$ if $w \in S_j$.

The indexes are computed from the leaves up. Since the leaves have minimum height and are subtree isomorphic to one another, the class $S_1$ consists of all leaves, i.e., $\xi[v] = 1$ if and only if $v$ is a leaf.

Assume that we have determined the indexes of all the vertices of height less than $h$ and assigned the numbers $1, \ldots, k$. We now consider all vertices of height $h$. For each such vertex $v$ we order its children, $v_1, \ldots, v_{d}$, by nondecreasing index, i.e., $\xi[v_1] \leq \cdots \leq \xi[v_d]$. We now lexicographically sort the vertices of height $h$ and assign an index to each of them according to this order. If $S_1^h \leq_C S_2^h \leq_C \cdots \leq_C S_m^h$ is the partition of the vertices of height $h$ into equivalence classes, then for $v \in S_i^h$ we set $\xi[v] = k + i$.

In particular, $\xi[u] = \xi[v]$ if and only if $(\xi[u_1], \ldots, \xi[u_d]) = (\xi[v_1], \ldots, \xi[v_d])$. Thus $\xi[u] = \xi[v]$ if and only if $T_u$ and $T_v$ are isomorphic.

2.3 On the Implementation

A straightforward implementation of the above procedure requires performing for each height $h$ a lexicographic sort of lists consisting of the children of vertices of height $h$. Let $n_h$ denote the number of vertices of height $h$. Then comparison-based methods require $O(n_h \log n_{h-1})$ time to sort all the lists of level $h$. Hence the total time will be $O(n \log n)$. Since the numbers to be sorted are in the range $1..n$, bucket-sort based methods can ensure only $O(n)$ for each sort. In Figure 1(a) both the range of indexes and the height are $\Theta(n)$; thus such an algorithm requires time $\Omega(n^2)$. In the next section we describe a linear time algorithm.

3 Canonical Indexes in Linear Time

3.1 Overview of the Algorithm

Let $\textit{level}[h]$ consist of all the vertices of height $h$. We construct the canonical indexes (indexes, for short) level by level, starting with $\textit{level}[0]$—the leaves. When processing $\textit{level}[h]$ we assume that we have already found the indexes of all the vertices at lower levels and have constructed a single list, $\textit{children}[h]$, consisting of all the children of vertices at $\textit{level}[h]$ sorted by their index (in nondecreasing order).

To assign the indexes to the vertices of $\textit{level}[h]$ we construct an auxiliary ordered tree—$D_h$, whose vertices are labeled by values of the indexes of the children of $\textit{level}[h]$,  (The data structures of the tree $D_h$ will appear in bold font.) The children of any vertex $d \in D_h$ have distinct labels and are ordered by increasing label. Each vertex $v$ of $\textit{level}[h]$ will be associated with a vertex $d \in D_h$ such that the path to $d$ from $\textit{root}$, the root of $D_h$, is labeled with the indexes of $v$’s children. More specifically, if we order the children of $v$ by nondecreasing index, then the label of the $i$-th vertex on the path is equal to the index of the $i$-th child of $v$. (See Figure 2.)

Initially, $D_h$ consists of a single vertex—$\textit{root}$, to which all vertices of $\textit{level}[h]$ are associated. In the course of the algorithm, new vertices are added to $D_h$, and the vertices of level $h$ proliferate.
1 Introduction

Free tree isomorphism is a well known problem which has many applications. A linear time algorithm to solve the problem has been developed by Hopcroft and Tarjan [2], who used it to devise a linear planarity test. Several other linear algorithms were suggested [1, pp.196-199], [3], [4]. The common approach is as follows.

First, free tree isomorphism is reduced to the rooted case. The root is set canonically at the central vertex or at the new vertex connected to the endpoints of the central edge (the old central edge is then deleted). Now, any two trees are isomorphic if and only if their (canonical) rooted versions are isomorphic.

Second, any rooted tree is assumed to be represented in the computer as an ordered tree, i.e., the children of each vertex are ordered. Evidently, ordered trees that differ only in the order of siblings represent isomorphic rooted trees. We call such ordered trees isomorphic.

Third, in any isomorphism equivalence class of rooted trees, a particular canonically ordered tree is defined. For any tree $T$, let us call such a distinguished tree in the isomorphism class of $T$ the canonical representation of $T$. Now, two rooted trees are isomorphic if and only if their canonical representations coincide. Therefore, to check whether two trees are isomorphic, it is sufficient to construct their canonical representations, and then compare them.

A natural generalization of the problem of checking whether two trees are isomorphic is that of partitioning a finite set of trees into isomorphism equivalence classes. Of particular interest is the set of rooted subtrees of a given rooted tree. The additional difficulty here is that the rooted subtrees are not disjoint and representing them explicitly may require $O(n^2)$ space for a tree of $n$ vertices.

In [5], Zemlyachenko describes a linear time algorithm for tree isomorphism, which partitions the subtrees of a given tree into isomorphism equivalence classes. Unfortunately, his description is terse and incomplete. In this note, we use modern data structures to explain and implement Zemlyachenko’s scheme. We give a full description of a free rendition of his method using some of his ideas and adding some new ones; in particular, the data structures are new.

2 Subtree Isomorphism

2.1 Canonical order

Given a rooted tree $T$ and a vertex $v$ of $T$ other than the root, removing the edge connecting $v$ to its parent, results in two subtrees. Let $T_v$, the subtree rooted at $v$, be the subtree that contains $v$. When $T_u$ and $T_v$ are isomorphic, we say that $u$ and $v$ are subtree isomorphic.

We start by defining a canonical order, $\leq_C$, on the vertices of a rooted tree $T$. The canonical order $\leq_C$ is defined from the leaves up. The leaves are the minimal elements of $\leq_C$.

Recall that the height of a vertex $v$ is the length of the path from $v$ to the furthest leaf of its subtree. To determine whether $u \leq_C v$ first compare their heights, if $\text{height}(u) < \text{height}(v)$ then $u \leq_C v$. If the heights are equal, we sort the children of $u$ and $v$ to create two lists $u_1 \leq_C u_2 \leq_C \ldots \leq_C u_{d(u)}$ and $v_1 \leq_C v_2 \leq_C \ldots \leq_C v_{d(v)}$. Now, $u \leq_C v$ if and only if the sorted list of $u$’s children is lexicographically less than the sorted list of $v$’s children.

Thus $u =_C v$ if and only if $\text{height}(u) = \text{height}(v)$, and $d(u) = d(v)$, and $u_i =_C v_i$ for $i = 1, \ldots, d(u)$. It is easy to see that $u =_C v$ if and only if $u$ and $v$ are subtree isomorphic.

Zemlyachenko August 3, 1998
On an Algorithm of Zemlyachenko for Subtree Isomorphism

Yefim Dinitz    Alon Itai    Michael Rodeh*
Computer Science Department
Technion, Haifa, Israel

June 25, 1998

Abstract

Zemlyachenko’s linear time algorithm for free tree isomorphism is unique in that it also partitions the set of rooted subtrees of a given rooted tree into isomorphism equivalence classes. Unfortunately, his algorithm is very hard to follow. In this note, we use modern data structures to explain and implement Zemlyachenko’s scheme. We give a full description of a free rendition of his method using some of his ideas and adding some new ones; in particular, the data structures are new.

*email: {dititz, itai, rodeh}@cs.technion.ac.il