Approximating Minimum Common Supertrees for Complete k-Ary Trees

Atsuko Yamaguchi
Satoru Miyano

March 25, 1993

Research Institute of Fundamental Information Science
Kyushu University 33
Fukuoka 812, Japan
E-mail: ako@rifis.sci.kyushu-u.ac.jp Phone: 092-641-1101 ex. 2329
Approximating Minimum Common Supertrees for Complete $k$-Ary Trees

Atsuko Yamaguchi  Satoru Miyano

Research Institute of Fundamental Information Science
Kyushu University 33, Fukuoka, 812, Japan
ako@rifis.sci.kyushu-u.ac.jp

Abstract

For a set $T$ of complete $k$-ary trees, we give a polynomial-time approximation algorithm for the problem of finding a $k$-ary common supertree with the minimum number of edges. This algorithm constructs a common supertree that has at most $(5/3)l$ edges, where $l$ is the number of edges in a minimum common supertree if $k \geq 2$.

1 Introduction

The shortest common superstring problem is to find the shortest string $u$ which contains all strings in a given set $S$ as substrings. This problem is known to be NP-complete [3, 4]. There have been some studies on approximation algorithms for the shortest common superstring problem [2, 5, 6, 7]. In particular, it is shown in [2] that a simple greedy algorithm produces a common superstring of length at most $3l_{opt}$, where $l_{opt}$ is the length of shortest common superstring.

In this paper we consider the problem of finding a minimum $k$-ary common supertree for a set of complete $k$-ary directed trees whose edges are labeled with some alphabet. For $k = 1$, the problem is exactly the same as the shortest common superstring problem. This paper deals with the case for $k \geq 2$. We shall show that the problem is NP-complete in Section 4. Thus an approximation algorithm for this problem interests us. The purpose of this paper is to present a polynomial time approximation algorithm for this problem. Let $l_A$ be the number of edges in the common supertree constructed by our approximation algorithm and $l$ the number of edges in the minimum common supertree. We prove that $l_A \leq (5/3)l$ for any $k \geq 2$.

2 Approximation algorithm

Let $\Sigma$ be a finite alphabet. In this paper, a tree over $\Sigma$ is a directed tree whose edges are labeled with symbols in $\Sigma$. When no confusion occurs, we simply it a tree. A $k$-ary tree is a tree such that each vertex has at most $k$ sons. A complete $k$-ary tree is a tree such that each vertex except leaves has exactly $k$ sons and the leaves are of the same depth. For a set $T$ of complete $k$-ary trees, a common supertree for $T$ is a $k$-ary tree such that each tree in $T$ is a subtree of it.
Step1. For each \( s, t \in T (s \neq t) \), compute \( \text{ov}(s, t) \) and \( h(s, t) \).

Step2. Initially, let \( E_A = \emptyset \). Examine \((s, t) \in T \times T\) in decreasing order of the overlap. If \((s, t)\) satisfies the following conditions, then add \((s, t)\) to \( E_A \) and mark a vertex \( v \in h(s, t) \) that satisfies (a).

(a) There is a vertex \( v \in h(s, t) \) such that none of \( v \) and its ancestors is marked in \( s \).

(b) For every \( s' \in T \), \((s', t)\) is not in \( E_A \).

(c) There is no cycle in \((T, E_A \cup \{(s, t)\})\).

Step3. From the spanning tree \( S = (T, E_A) \) constructed by Step2, compose a supertree of \( T \).

---

**Figure 1: GreedyOverlap**

**Definition 1.** The minimum common supertree problem for complete k-ary trees (MCSP(k)) is defined as follows:

**INSTANCE:** A finite set \( T \) of complete k-ary trees over some alphabet \( \Sigma \).

**PROBLEM:** Find a k-ary tree \( u \) with the minimum number of edges such that each tree in \( T \) is a subtree of \( u \).

MCSP(1) can be regarded as the shortest common superstring problem. In Section 4, we shall prove that the decision version of MCSP(k) is also NP-complete.

**Definition 2.** A set \( T \) of trees is called reduced if no tree in \( T \) is a subtree of another tree in \( T \).

Let \( T \) be a set of complete k-ary trees. Then \( T' = T - \{ t \mid t \text{ is a subtree of some } s \in T \text{ with } s \neq t \} \) is obviously reduced. Since the minimum common supertree for \( T \) is also that for \( T' \) and \( T' \) is computable from \( T \) in polynomial time, we deal with a reduced set of complete k-ary trees except where otherwise noted.

Let \( s \) and \( t \) be k-ary trees. \(|s|\) denotes the number of edges in \( s \). We introduce some notations.

1. \( s \twoheadrightarrow t \) is a minimum common supertree for \( s \) and \( t \) obtained by identifying the root of \( t \) with a vertex \( v \) of \( s \). If \( v \) is clear from the context, we may denote as \( s \rightarrow t \).

2. \( h(s, t) = \{ v \mid s \twoheadrightarrow t \} \).

3. \( \text{ov}(s, t) = |s| + |t| - |s \twoheadrightarrow t| \) (\( v \in h(s, t) \)). \( \text{ov}(s, t) \) is called the overlap of \( t \) on \( s \).

MCSP(k) can be regarded as the problem to find a common supertree \( u \) for \( T \) such that total overlap \( \sum_{t \in T} |t| - |u| \) of \( u \) is maximum. From this point of view, we shall give an approximation algorithm for MCSP(k) in Figure 1.

Step1 computes \( \text{ov}(s, t) \) and \( h(s, t) \) in time \( O(|s|^2|t|) \) using the algorithm by Akutsu [1]. By the conditions (b) and (c) of Step2, the directed graph \( S = (T, E_A) \) obtained by Step2
is a forest of the directed weighted graph $G = (T, E, ov)$ with $E = \{(s, t) \mid s, t \in T, s \neq t\}$. Suppose that $S$ is not connected. Let $S_1$ and $S_2$ be connected components in $S$ with $S_1 \neq S_2$, $r$ the root of the tree $S_1$, and $l$ one of the leaves of the tree $S_2$. Then the edge $(l, r)$ is not in $E_A$, but $(l, r)$ satisfies the conditions (a), (b), (c). Hence $S$ is connected. Thus $S$ is a directed spanning tree of $G$. For $(s, t) \in E_A$, $a(s, t)$ denotes the vertex marked when $(s, t)$ is added to $E_A$ at Step2.

Step3 constructs a common supertree $u$ for $T$ from $S = (T, E_A)$ as follows: Initially, let $V' = T$, $E' = E_A$ and $S' = (V', E')$. For a vertex $x$ of $S'$, we denote by $t_x$ the vertex $s_x$ locating at the position corresponding to $x$ in $S$. We assume the following conditions:

1. The vertex $x$ is a $k$-ary tree.
2. The $k$-ary tree $x$ contains the complete $k$-ary tree $t_x$ as a substree by sharing the root.

Initially, these conditions are obviously satisfied. Choose a leaf $w$ of $S' = (V', E')$. Let $p(w)$ denote the parent of $w$ in the tree $S' = (V', E')$. Let $p(w) \Rightarrow w$ be the tree obtained by overlapping the subtree $t_w$ of $w$ onto the subtree $t_{p(w)}$ of $p(w)$ by the way of $t_{p(w)} \leftarrow t_w$ with $v = a(t_{p(w)}, t_w)$. Since vertices $a(s, t)$ and $a(s, t')$ are not on the same path in the tree $s$ for any pair of edges $(s, t), (s, t') \in E_A$ with $t \neq t'$ and $T$ is reduced, we can see that the subtree of $p(w)$ rooted at $a(t_{p(w)}, t_w)$ is the region where $w$ overlaps on $p(w)$ that is exactly the same as the region where $t_w$ overlaps on $t_{p(w)}$. Hence $p(w) \Rightarrow w$ is a $k$-ary tree. Replace $p(w)$ by $p(w) \Rightarrow w$. Hence the edges incident to the vertex $p(w)$ are now incident to the vertex $p(w) \Rightarrow w$. Remove the leaf $w$ from $V'$ together with the edge coming into $w$. Then it is clear from the construction that $p(w) \Rightarrow w$ satisfies (1) and (2). There are no other changes for vertices in $S'$. Repeat this procedure until $E' = \emptyset$. When $E' = \emptyset$, we see $|V'| = 1$. The unique element of $V'$ represents a common supertree $u$ for $T$. It is obvious that Step2 and Step3 are computable in polynomial time.

### 3 Error bound

We discuss error of the approximation algorithm GreedyOverlap. For a directed weighted graph $G = (V, E, w)$, we denote $||G|| = ||E|| = \sum_{e \in E} w(e)$.

**Lemma 1.** Let $T$ be a reduced set of complete $k$-ary trees, $u$ the common supertree for $T$ composed by algorithm GreedyOverlap, and $(T, E_A)$ the spanning tree constructed in Step2. Then

$$|u| = \sum_{t \in T} |t| - \sum_{(s, t) \in E_A} ov(s, t).$$

**Proof.** Consider the tree $S' = (V', E')$ just before any iteration in Step3. Let $w$ be the leaf of $S'$ overlapped on its parent $p(w)$. Then by the argument in Step3, we see $|p(w) \Rightarrow w| = |p(w)| + |w| - ov(t_{p(w)}, t_w)$. By induction we have $|u| = \sum_{t \in T} |t| - \sum_{(s, t) \in E_A} ov(s, t)$. □

**Lemma 2.** Let $T$ be a reduced set of complete $k$-ary trees, $u_{opt}$ a minimum common supertree for $T$ with arity $k$, and $S = (T, E_A)$ the spanning tree composed by algorithm GreedyOverlap. Then

$$||S|| \geq \frac{1}{3} \left( \sum_{t \in T} |t| - |u_{opt}| \right).$$
Proof. For the directed weighted graph $G = (T, E, o_v)$, let $S_{\text{max}}$ be a maximum spanning tree of $G$. Since $\|S_{\text{max}}\| \geq \sum_{t \in T} |t| - |u_{\text{opt}}|$, this theorem can be shown by proving the following:

$$\|S\| \geq \frac{1}{3} \|S_{\text{max}}\|$$

Let $e_i$ be the edge added to $E_A$ at the $i$-th time and $E_i = \{e_1, \ldots, e_i\}$ for $i = 1, \ldots, m-1$, where $m$ is the number of trees in $T$. We define $J_i$ as follows:

$$J_i = \begin{cases} E_{\text{max}} & i = 0, \\ J_{i-1} - (\{e_i\} \cup D_i) & \text{otherwise}, \end{cases}$$

where $D_i$ is the set of edges in $J_{i-1}$ that shall be removed in the sequel by violating the conditions (a),(b),(c) of Step2 because of the existence of $e_i$ in $E_A$. It should be noted that there is at most one edge in $J_{i-1}$ which violates the condition (a) by the existence of $e_i$. The same holds for each of the conditions (b) and (c). Hence $D_i$ contains at most three edges. Since edges are examined in decreasing order of overlap, we have $o_v(e_i) \geq o_v(f)$ for all $f \in D_i$.

By induction on $i$ we shall show $3\|E_i\| + \|J_i\| \geq \|S_{\text{max}}\|$. For $i = 0$, since $E_0 = \emptyset$ and $\|J_0\| = \|E_{\text{max}}\| = \|S_{\text{max}}\|$, the claim is clear. Assume that the claim holds for $i-1$.

1. $e_i \in S_{\text{max}}$: Since $o_v(e_i) \geq 0$ and $\|D_i\| = \emptyset$,

$$3\|E_i\| + \|J_i\| = 3\|E_{i-1}\| + 3o_v(e_i) + \|J_{i-1}\| - o_v(e_i)$$

$$\geq 3\|E_{i-1}\| + \|J_{i-1}\|$$

$$\geq \|S_{\text{max}}\||. $$

2. $e_i \notin S_{\text{max}}$: Since $\|D_i\| \geq 3o_v(e_i)$,

$$3\|E_i\| + \|J_i\| \geq 3\|E_{i-1}\| + 3o_v(e_i) + \|J_{i-1}\| - \|D_i\|$$

$$\geq 3\|E_{i-1}\| + \|J_{i-1}\|$$

$$\geq \|S_{\text{max}}\||.$$

Since $J_{m-1} = \emptyset, E_{m-1} = E_A$, we have $3\|S\| \geq \|S_{\text{max}}\|$. □

Lemma 3. Let $T$ be a reduced set of complete $k$-ary trees with $k \geq 2$, and $u_{\text{opt}}$ be a minimum common supertree for $T$. Then

$$\sum_{t \in T} |t| \leq 2|u_{\text{opt}}|.$$

Proof. Let $t_1$ and $t_2$ be arbitrary complete $k$-ary trees in $T$. Let $j$ be the depth of $t_2$. Since $T$ is reduced and $t_1$ and $t_2$ are complete $k$-ary trees, at least $k^j$ edges of $t_2$ do not overlap on $t_1$.

Since $|t_2| = k^j + k^{j-1} + \cdots + k$ and $k^j-1 + \cdots + k < k^j$,

$$o_v(t_1, t_2) \leq k^j - 1 + \cdots + k < \frac{1}{2} (k^j + k^{j-1} + \cdots + k) = \frac{1}{2} |t_2|.$$

For the directed weighted graph $G = (T, E, o_v)$, let $S_{\text{max}}$ be a maximum spanning tree of $G$. Then

$$|u_{\text{opt}}| \geq \sum_{t \in T} |t| - \|S_{\text{max}}\| = \sum_{t \in T} |t| - \sum_{(t, t') \in E_{\text{max}}} o_v(t, t') \geq \frac{1}{2} \sum_{t \in T} |t|.$$

□
From Lemma 1, Lemma 2, Lemma 3, we have the following theorem:

**Theorem 1.** Let $T$ be a reduced set of complete $k$-ary trees with $k \geq 2$, $u_{opt}$ a minimum common supertree for $T$, and $u$ the common supertree for $T$ composed by algorithm GreedyOverlap. Then

$$|u| \leq \frac{5}{3}|u_{opt}|.$$

## 4 NP-completeness

We prove the NP-completeness of the following problem.

**Definition 3.** $MCS(k)$ is defined as follows:

**INSTANCE:** A finite set $T$ of complete $k$-ary trees over a finite alphabet $\Sigma$ and a positive integer $K$.

**QUESTION:** Is there a $k$-ary common supertree $u$ for $T$ such that the number of edges in $u$ is at most $K$?

**Theorem 2.** $MCS(k)$ is NP-complete for any $k \geq 1$.

**Proof.** We give a reduction from the directed Hamiltonian path problem [4], that is to decide if a given directed graph $G = (V, E)$ has a Hamiltonian path.

Let $G = (V, E)$ be a graph with $|V| = n$ and $|E| = m$. Let $\Sigma = V \cup \{v' \mid v \in V\} \cup \{\#, \}$. We construct a set $T$ of complete $k$-ary trees over $\Sigma$ from $G$ as follows: Let $d_v$ be the outdegree of $v \in V$ and $R_v = \{w_j \mid (v, w_j) \in E, 1 \leq j \leq d_v\}$. For $1 \leq j \leq d_v$, let $a^j_v$ denote a complete $k$-ary tree of depth 4 such that it has a path from the root to a leaf whose edges are labeled with $v', w_j, v', w_{j+1}$, where $w_{d_v+1} = w_1$, and the other edges are labeled with $\#$, (Figure 2). Let $c_v$ denote a complete $k$-ary tree of depth 3 such that it has a path from the root to a leaf whose edges are labeled with $v, \#, v'$ and the other edges are labeled with $\#$ (Figure 3).

Then we define as

$$T = \bigcup_{v \in V} (\{a^j_v \mid 1 \leq j \leq d_v\} \cup \{c_v \mid v \in V\}).$$

We show that the following statements (1) and (2) are equivalent.

1. There is a Hamiltonian path on $G$.

2. There is a common supertree $u$ for $T$ with at most $k + n(2k^2 + k^3 + m(k^3 + k^4))$ edges.

For $1 \leq j \leq d_v$, let $s^j_v = (\cdots (((c_v \rightarrow a^j_v) \rightarrow a^{j+1}_v) \cdots) \rightarrow a^*_v) \rightarrow a^j_v \cdots \rightarrow a^{j-1}_v$, where $a^*_v = a^0_v$. Since $ov(c_v, a^j_v) = k$ and $ov(a^j_v, a^{j+1}_v) = k + 2k^2$, we have $|s^j_v| = k + 2k^2 + k^3 + (k^3 + k^4)$. 

(1) $\Rightarrow$ (2): Let $(v_1, v_2), \ldots, (v_{n-1}, v_n)$ be a Hamiltonian path on $G$. Then $(v_i, v_{i+1})$ is in $E$ for each $1 \leq i \leq n - 1$. By the definition of $R_v$, we see $w_{j_i} = v_{i+1}$ for some $j_i$. Then the tree $(\cdots ((s^j_{v_1} \rightarrow s^j_{v_2}) \rightarrow s^j_{v_3}) \cdots) \rightarrow s^j_{v_{n-1}} \rightarrow s^j_{v_n}$ is a common supertree for $T$ for any $1 \leq p \leq d_v$. Since $ov(s^j_{v_i}, s^j_{v_{i+1}}) = k$, the number of edges in the tree is

$$\sum_{i=1}^{n-1} |s^j_{v_i}| + |s^j_{v_n}| - (n - 1)k = \sum_{i=1}^{n} (k + 2k^2 + k^3 + d_{v_i}(k^3 + k^4)) - (n - 1)k$$

$$= k + n(2k^2 + k^3) + m(k^3 + k^4).$$
(1) \iff (2): Let \( u \) be a common supertree for \( T \) with \(|u| \leq k+n(2k^2+k^3)+m(k^3+k^4) \). Then the total overlap of \( u \) is at least \((k+k^2)m-nk^2+(n-1)k \). Now we show that \( u \) is of the form \((\cdots((s_{v_1}^{a_1} \rightarrow s_{v_2}^{a_2}) \rightarrow s_{v_3}^{a_3})\cdots) \rightarrow s_{v_n}^{a_n} \) for some \( 1 \leq j \leq d_{v_k} \). Let \( t_\nu \) be the tree in \( T \) such that the overlap of \( a_v^j \) on \( t_\nu \) in \( u \) is maximum. Then

\[
\sum_{v \in V} \sum_{j=1}^{d_v} ov(t_\nu, a_v^j) \leq (k+k^2)(m-n) + nk = (k+k^2)m-nk^2.
\]

It is obvious that the equality holds if and only if, for each \( v \in V \), there is \( 1 \leq j \leq d_v \) such that \( a_v^j \) overlaps on \( c_v \) and \( a_v^p \) with \( 1 \leq p \neq j \leq d_v \) overlaps on \( a_v^{p-1} \). Let \( t_v \) be the tree in \( T \) such that the overlap of \( c_v \) on \( t_v \) in \( u \) is maximum. We can see that \( ov(c_v, u) = 0 \) for any \( \nu \neq v \), \( ov(a_v^j, c_v) = k \) for \((\nu, v) \in E \), and \( ov(a_v^p, c_v) = 0 \) for \((\nu, v) \notin E \). Let \( (t(u)) \in T \) be the subtree of \( u \) that shares the root with \( u \). Suppose that \( t(u) = a_v^j \) for some \( v \) and \( j \). We can see that \( \sum_{v \in V} \sum_{j=1}^{d_v} ov(t_\nu, a_v^j) \leq (k+k^2)(m-1) - nk^2 \). Moreover, \( \sum_{v \in V} ov(t_v, c_v) \leq nk \). Thus the total overlap is at most \((k+k^2)(m-1) - nk^2 + nk < (k+k^2)m-nk^2+(n-1)k \), a contradiction. Therefore \( t(u) = c_v \) for some \( v \). Then

\[
\sum_{v \in V} ov(t_v, c_v) \leq (n-1)k,
\]

and the equality holds if and only if for each \( v \in V \) with \( c_v \neq t(u) \), \( c_v \) overlaps on \( a_v^j \) for some \( \nu \) with \((\nu, v) \in E \) and \( j \).

Hence \( u \) is of the form \((\cdots((s_{v_1}^{a_1} \rightarrow s_{v_2}^{a_2}) \rightarrow s_{v_3}^{a_3})\cdots) \rightarrow s_{v_n}^{a_n} \) and \( ov(t_v, c_v) = k \) for any \( i \geq 2 \). Since edges \((v_1, v_2), \cdots, (v_{n-1}, v_n) \) are in \( E \), \( G \) has a Hamiltonian path. \( \square \)
5 Conclusion

Our approximation algorithm $\text{GreedyOverlap}$ for $\text{MCSP}(k)$ approximates the number $n$ of edges of the minimum common supertree by $(5/3)n$ for any $k \geq 2$. But we do not know whether $5/3$ is the best error bound for our algorithm. In this paper we have restricted our attention to complete $k$-ary trees. However, we do not know any approximation algorithm for the general problem, i.e., the problem of finding a minimum common supertree for a finite set $T$ of trees, where no restriction is put on the trees such as arity, etc.

References


About the Authors

Atsuko Yamaguchi (山口 敦子) was born in Nagasaki on February 11, 1969. She graduated from Department of Mathematics, Kyushu University. Presently, she is a graduate student at Department of Information Systems, Kyushu University. She is studying the computational complexity theory.

Satoru Miyano (宮野 悟) was born in Oita on December 5, 1954. He received the B.S. in 1977, the M.S. degree in 1979 and the Dr. Sci. in 1984 all in Mathematics from Kyushu University. Presently, he is a Professor of Research Institute of Fundamental Information Science, Kyushu University. His present interests include parallel algorithms, computational complexity, computational learning theory and Genome Informatics.

Research Institute of Fundamental Information Science, Kyushu University 33, Fukuoka 812, Japan.