Minimal Common Container of Tree Patterns

Junhun Wang
Griffith University, Australia
J.Wang@griffith.edu.au
Chaoyi Pang
CSIRO, Australia
Chaoyi.Pang@csiro.au

Jeffrey Xu Yu
Chinese University of Hong Kong, China
yu@se.cuhk.edu.hk
Chengfei Liu
Swinburne University of Technology, Australia
cliu@ict.swin.edu.au

ABSTRACT

Tree patterns represent important fragments of XPath. In this paper, we show that some classes of tree patterns exhibit such a property that, given a finite number of tree patterns \(P_1, \ldots, P_n\), there exists another pattern \(P\) (tree pattern or DAG-pattern) such that \(P_1, \ldots, P_n\) are all contained in \(P\), and for any tree pattern \(Q\) belonging to a given class \(C\), \(P_1, \ldots, P_n\) are contained in \(Q\) if and only if \(P\) is contained in \(Q\). We call such a property that, given a finite number of tree patterns \(P_1, \ldots, P_n\), there exists another pattern \(P\) such that \(P_1, \ldots, P_n\) are all contained in \(P\), and for any tree pattern \(Q\) belonging to a given class \(C\), \(P_1, \ldots, P_n\) are contained in \(Q\) if and only if \(P\) is contained in \(Q\) the Minimal Common Container (MCC) of \(P_1, \ldots, P_n\) with respect to \(C\). We provide a method for constructing such MCCs. However, we are not concerned with the complexity of the algorithms, as our focus is on the existence of MCCs.

Keywords
XPath, Tree Pattern, Containment

1. INTRODUCTION

Tree patterns represent important fragments of XPath. Over the last few years there have been extensive research on tree patterns. In particular, the containment of tree patterns has been investigated in several papers including [2], and some interesting structural properties of tree patterns have been observed, e.g., in [1, 3].

In this paper, we show that some tree patterns exhibit such a property that, given a finite number of compatible tree patterns \(P_1, \ldots, P_n\), there exists another pattern \(P\) (tree pattern or DAG-pattern) such that \(P_1, \ldots, P_n\) are all contained in \(P\), and for any tree pattern \(Q\) belonging to a given class \(C\), \(P_1, \ldots, P_n\) are contained in \(Q\) if and only if \(P\) is contained in \(Q\). We call such a property that, given a finite number of tree patterns \(P_1, \ldots, P_n\), there exists another pattern \(P\) such that \(P_1, \ldots, P_n\) are all contained in \(P\), and for any tree pattern \(Q\) belonging to a given class \(C\), \(P_1, \ldots, P_n\) are contained in \(Q\) if and only if \(P\) is contained in \(Q\) the Minimal Common Container (MCC) of \(P_1, \ldots, P_n\) with respect to \(C\). We provide a method for constructing such MCCs. However, we are not concerned with the complexity of the algorithms, as our focus is on the existence of MCCs. The existence of MCCs has a number of applications. For example, it can be used in testing containment of tree patterns under non-recursive, non-disjunctive DTDs as well as in query answering using caches.

2. PRELIMINARIES

Let \(\Sigma\) be an infinite set of tags. An XML tree (Xtree) is a tree with every node labeled with a tag in \(\Sigma\). A tree pattern (TP) is a tree with a unique distinguished node, and with every node labeled with a symbol in \(\Sigma \cup \{\ast\}\) (here \(\ast\) is the wildcard which represents any tag), and every edge labeled with either \(/\) or \(/\ast\). The path from the root to the distinguished node is called the distinguished path. Figure 1 shows two TPs \(P\) and \(Q\), where single and double lines represent \(/\)-edges and \(/\ast\)-edges respectively, and a circle indicates the distinguished node. Let \(P\) be a TP. We will use \(\delta(P)\) and \(\delta(P)\) to denote the distinguished node and the distinguished path of \(P\) respectively. Note: the TPs in our discussion correspond to the fragment \(P^{/\ast\ast}\) defined in [2]. Several subsets of \(P^{/\ast\ast}\) were studied in [2]. In particular, the subset \(P^{/\ast\ast}\) contains all TPs that do not have \(\ast\)-nodes. In this paper we are also interested in the subset of all TPs in \(P^{/\ast\ast}\) such that the root is not labeled \(\ast\), no \(\ast\)-node is incident on a \(/\ast\)-edge, and no leaf node is labeled \(\ast\). We denote this subset by \(P^{/\ast\ast}\).

In what follows, for any tree or rooted directed graph \(T\), we will use \(N(T)\) and \(\mathit{rt}(T)\) to denote the node set and the root of \(T\) respectively. We will also use \(\mathit{label}(v)\) to denote the label of node \(v\), and call a node labeled \(\tau\) a \(\tau\)-node. If \((v_1, v_2)\) is a \(/\)-edge (resp. \(/\ast\)-edge) in \(T\), we will say \(v_2\) is a \(/\)-child (resp. \(/\ast\)-child) of \(v_1\).

A matching of a TP, \(P\), in an Xtree, \(t\), is a mapping \(\delta\) from \(N(P)\) to \(N(t)\) satisfying the following conditions: (1) root-preserving, i.e., \(\delta(\mathit{rt}(P)) = \mathit{rt}(t)\), (2) label-preserving, i.e., \(\forall v \in N(P), \mathit{label}(v) = \ast \Rightarrow \delta(\mathit{label}(v)) = \mathit{label}(\delta(v))\), and (3) structure-preserving, i.e., for every edge \((x, y)\) in \(P\), if it is a \(/\)-edge, then \(\delta(y)\) is a child of \(\delta(x)\); if it is a \(/\ast\)-edge, then \(\delta(y)\) is a descendant of \(\delta(x)\), i.e., there is a path from \(\delta(x)\) to \(\delta(y)\). Each matching \(\delta\) produces a subtree of \(t\) rooted at \(\delta(\mathit{DN}(P))\), which is called an answer to the TP. We use \(P(t)\) to denote the set of all answers of \(P\) over \(t\). If \(T\) is a set of Xtrees, we use \(P(T)\) to denote \(\bigcup_{t \in T} P(t)\).
Let $P$ and $Q$ be TPs. $P$ is said to be contained in $Q$, denoted $P \subseteq Q$, if for every Xtree $t$, $P(t) \subseteq Q(t)$. The equivalence of two TPs is defined as two-way containment as usual.

3. MINIMAL COMMON CONTAINER OF TREE PATTERNS

Throughout this section (and without loss of generality), we assume that the root labels of tree patterns $P_1, \ldots, P_n$ are identical.

3.1 MCC for Boolean Patterns in $B^{(\text{//})}$

A boolean pattern $[2]$ is a tree pattern without distinguished node. Given an Xtree $t$ and a boolean pattern $P$, a matching of $P$ in $t$ is a mapping which is root-preserving, label-preserving, and structure-preserving. Given two boolean patterns $P_1$ and $P_2$, we say $P_1$ is contained in $P_2$, denoted $P_1 \subseteq P_2$, if whenever $P_1$ has a matching in $t$, for all $t$, $P_2$ will also have a matching in $t$. To explicitly distinguish boolean patterns from non-boolean patterns, we will use $B^{(\text{//})}$ to denote the set of boolean patterns in $P^{(\text{//})}$, that is, $B^{(\text{//})}$ is the set of all boolean patterns that do not have the wildcard $\ast$.

Definition 3.1: Let $P_1, \ldots, P_n$ be patterns in $B^{(\text{//})}$ with identical root labels. A minimal common container (MCC) of $P_1, \ldots, P_n$ wrt $B^{(\text{//})}$, denoted $\text{MCC}(P_1, \ldots, P_n)$, is a pattern $P \in B^{(\text{//})}$ such that

1. $P_1, \ldots, P_n \subseteq P$, and
2. for any pattern $Q \in B^{(\text{//})}$, if $P_1, \ldots, P_n \subseteq Q$, then $P \subseteq Q$.

Note that if $P$ is the MCC of $P_1, \ldots, P_n$, then for any $Q \in B^{(\text{//})}$, $P_1 \cup \cdots \cup P_n \subseteq Q$ if and only if $P \subseteq Q$. Furthermore, all MCCs of $P_1, \ldots, P_n$ are equivalent.

Next we show $\text{MCC}(P_1, \ldots, P_n)$ always exists. We will focus on the case $n = 2$ first, and extend it to the general case later.

Definition 3.2: (closest matching descendant pair) Two nodes $v_1 \in P_1$ and $v_2 \in P_2$ are said to be a matching pair if $\text{label}(v_1) = \text{label}(v_2)$. We use $[v_1, v_2]$ to denote the matching pair. Let $v_1, v_2$ and $[u_1, u_2]$ be matching pairs. If $v_1$ is a descendant of $u_1$, $v_2$ is a descendant of $u_2$, then we say $[v_1, v_2]$ is a matching descendant pair of $[u_1, u_2]$. If there is no matching descendant pair of $[x_1, x_2]$ such that $[v_1, v_2]$ is a matching descendant pair of $[x_1, x_2]$, then we say that $[v_1, v_2]$ is a closest matching descendant pair (CMDP) of $[u_1, u_2]$.

Note that every matching pair $[v_1, v_2]$, if $v_1 \neq rt(P_1)$ and $v_2 \neq rt(P_2)$, is a CMDP of at least one other matching pair.

Example 3.1: Consider the boolean patterns in Figure 2 (a) and (b). In the figure we use subscripts to distinguish nodes with the same label (e.g., $b_1, b_2$ and $b_3$ represent three nodes labeled with $b$). It can be seen that

$[a_1, a_2], [b_1, b_2], [b_1, b_3], [c_1, c_3], [c_2, c_3], [d_1, d_2], [e_1, e_2]$

are all matching pairs wrt $P_1$ and $P_2$, and $[b_1, b_2], [b_1, b_3], [c_1, c_3], [c_2, c_3], [d_1, d_2]$ are CMDPs of $[a_1, a_2]$. In addition, $[e_1, e_2]$ is a CMDP of $[b_1, b_3]$, but it is not a CMDP of $[a_1, a_2]$.

Given a matching pair $[v_1, v_2]$, we can construct a tree pattern $T(v_1, v_2)$ of height 1 as follows.

1. The root of $T(v_1, v_2)$ is labeled with $\text{label}(v_1)$.
2. For every CMDP $[u_1, u_2]$ of $[v_1, v_2]$, add a child $u$ to the root and label it with $\text{label}(u_1)$. If $u_1$ is a /-child of $v_1$, and $u_2$ is a /-child of $v_2$, then label the edge $(rt(T(v_1, v_2)), u) \times /$; otherwise label it with //.

For example, for the patterns $P_1$ and $P_2$ in Figure 2, $T(a_1, a_2)$ is the pattern consisting of the root and all of its children in $\text{MCC}(P_1, P_2)$ shown in Figure 2 (c).

Note that every node in $T(v_1, v_2)$ corresponds to a CMDP of $[v_1, v_2]$.

We can now construct a boolean pattern $P \in B^{(\text{//})}$ as follows.

Step 1: Initially, let $P = T(rt(P_1), rt(P_2))$.

Step 2: For every leaf node in $P$, if it corresponds to the matching pair $[u_1, u_2]$, then replace it with $T(u_1, u_2)$ (by merging it with the root of $T(u_1, u_2)$).

Step 3: Repeat Step 2 above until no more nodes can be added to $P$, i.e., every leaf node in $P$ corresponds to a matching pair which has no matching descendant pair.

Example 3.2: The pattern constructed for the patterns $P_1$ and $P_2$ in Figure 2 is as shown in Figure 2 (c). It is obtained by replacing the node corresponding to $[b_1, b_3]$ in $T(a_1, a_2)$ with $T(b_1, b_3)$.

Note that there is an 1:1 correspondence between the matching descendant pairs of $[rt(P_1), rt(P_2)]$ and descendants of $rt(P)$. We can prove the following theorem.

Theorem 3.1: The boolean pattern $P$ constructed above is a MCC of $P_1$ and $P_2$ wrt $B^{(\text{//})}$.

Now suppose there are boolean patterns $P_1, \ldots, P_n \in B^{(\text{//})}$. We can construct a MCC of $P_1, \ldots, P_n$ recursively using the formula $\text{MCC}(P_1, \ldots, P_n) = \text{MCC}(P_1, \text{MCC}(P_2, \ldots, P_n))$. Thus we have

Theorem 3.2: For any finite number of boolean patterns $P_1, \ldots, P_n \in B^{(\text{//})}$ that have identical root labels, there is a pattern $P \in B^{(\text{//})}$ which is a MCC of $P_1, \ldots, P_n$ wrt $B^{(\text{//})}$.

Note that, in the above theorem, there are no restrictions on the boolean patterns except that they do not involve $\ast$.

3.2 Non-boolean tree patterns in $P^{(\text{//})}$

We call tree patterns with the same label for the roots and the same label for the distinguished nodes compatible patterns. Suppose $P_1, \ldots, P_n \in P^{(\text{//})}$ are compatible tree patterns satisfying the following conditions:
Consider the patterns in Figure 5 (a) and Example 3.3: denote the full subtree rooted at \( v \) node on \( DP \) follows:

(A) the labels on the distinguished paths follow a fixed order, that is, there are no \( P_i, P_j \) and labels \( x, y \) such that in \( DP(P_i) \) an \( x \)-node is an ancestor of a \( y \)-node, and in \( DP(P_j) \) a \( y \)-node is an ancestor of an \( x \)-node;

(B) no label is repeated on the distinguished path of \( P_i \) (for \( i \in [1, n] \)).

With conditions (A) and (B) above, the common labels on \( DP(P_i) \) and \( DP(P_j) \) occur in the same order in \( DP(P_i) \) and \( DP(P_j) \), and for each common label there is a unique node in \( P_i \) with that label (for \( i, j \in [1, n] \)).

Before describing the method to construct the MCC, we need to introduce some notations. Given a pattern \( P \) and a node \( v \) on the distinguished path of \( P \), we use \( Sub_v(P) \) to denote the full subtree rooted at \( v \), and treat it as a boolean pattern (by disregarding the distinguished node). We also use \( S_v(P) \) to denote the boolean pattern corresponding to the subtree obtained from \( Sub_v(P) \) by removing \( Sub_v(P) \) if \( u \) is the child of \( v \) on the distinguished path. In other words, if \( v \) and \( u \) are nodes on \( DP(P) \), and \( v \) is a child of \( u \), then \( S_v(P) = Sub_u(P) - Sub_v(P) \). In the special case where \( v = DN(P) \), \( S_v(P) = Sub_v(P) \) (See Figure 3).

We can now construct a pattern \( P \) from \( P_1 \) and \( P_2 \) as follows:

1. Find the common labels on \( DP(P_1) \) and \( DP(P_2) \). Without loss of generality, we assume they are ordered as \( \tau_1, \ldots, \tau_n \) in \( DP(P_1) \) and \( DP(P_2) \). It follows that \( \tau_1 = label(rt(P_1)) = label(rt(P_2)) \) and \( \tau_n = label(DN(P_1)) = label(DN(P_2)) \).

2. Construct a temporary distinguished path \( x_1/x_2/\ldots/x_{n-1}/x_n \) such that \( label(x_i) = \tau_i \), with \( x_n \) being the distinguished node.

3. Suppose \( \tau_1, \ldots, \tau_n \) are the nodes labeled \( \tau_1, \ldots, \tau_n \) respectively in \( DP(P_1) \), and \( u_1, \ldots, u_n \) are the nodes labeled \( \tau_1, \ldots, \tau_n \) respectively in \( DP(P_2) \). For \( i \in [1, n-1] \), if \( u_{i+1} \) is a /-child of \( v_i \) and \( u_{i+1} \) is an /-child of \( u_i \), then we change the edge \( (x_i, x_{i+1}) \) to a /-edge.

4. For \( i \in [1, n] \), find MCC(\( Sub_v(P_1), Sub_u(P_2) \)), and add it under \( x_i \) (i.e., merge \( rt(MCC(Sub_v(P_1), Sub_u(P_2))) \) with \( x_i \)).

The final pattern \( P \) obtained is illustrated in Figure 4.

**Example 3.3:** Consider the patterns in Figure 5 (a) and (b). The pattern constructed using the above approach is shown in Figure 5 (c).

**Figure 3:** Illustration of \( S_v(P) \) and \( Sub_v(P) \)

![Figure 3](image-url)

**Figure 4:** Structure of the constructed pattern \( P \)

![Figure 4](image-url)

**Figure 5:** \( P_{3,4} \) is the constructed MCC for \( P_3 \) and \( P_4 \)

Note that we are not concerned with the complexity of the algorithm or the minimality of the constructed pattern, as our focus is the existence of a MCC.

We can show that \( P \) is a MCC of \( P_1 \) and \( P_2 \). (The definition of MCC for non-boolean patterns is similar to that for boolean patterns, thus omitted here).

**Theorem 3.3:** The above pattern \( P \) is a MCC of \( P_1 \) and \( P_2 \) wrt \( P(//;\bullet) \).

The above result can be easily extended to any finite number of compatible tree patterns in \( P(//;\bullet) \).

**Corollary 3.1:** For compatible patterns \( P_1, \ldots, P_n \in P(//;\bullet) \) such that \( P_i \subseteq P \) and \( (2) \) for any \( Q \in P(//;\bullet) \), \( P_1, \ldots, P_n \subseteq Q \) iff \( P \subseteq Q \).

One may wonder when the MCC of \( P_1, \ldots, P_n \) is equivalent to \( P_1 \cup \ldots \cup P_n \). In other words, when MCC(\( P_1, \ldots, P_n \)) = \( P_1 \cup \ldots \cup P_n \). From [3] we know that for tree patterns in \( P(//;\bullet) \), \( P \subseteq P_1 \cup \ldots \cup P_n \) iff there is \( i \in [1, n] \) such that \( P_i \subseteq P \). Therefore, we have

**Corollary 3.2:** For compatible patterns \( P_1, \ldots, P_n \in P(//;\bullet) \), if MCC(\( P_1, \ldots, P_n \)) exists, then MCC(\( P_1, \ldots, P_n \)) = \( P_1 \cup \ldots \cup P_n \) iff there is \( i \in [1, n] \) such that \( P_i \subseteq P_j \) for all \( j \neq i \).

### 3.3 Boolean Patterns in \( \bar{P}(//;\bullet) \)

We now consider boolean tree patterns in \( \bar{P}(//;\bullet) \), which is the set of boolean patterns corresponding to tree patterns in \( \bar{P}(//;\bullet) \). Recall that such tree patterns do not have *-nodes incident on //edges, or dangling *-nodes (i.e., *-nodes which do not have a non-*-descendant), and the root is not labeled *. As usual, we consider the case \( n = 2 \) first, and extend it to the general case later.

**Definition 3.3:** (label-matching pair and /-child pair)

Let \( P_1, P_2 \) be boolean patterns in \( \bar{P}(//;\bullet) \). Let \( v_1 \) and \( v_2 \) be nodes in \( P_1 \) and \( P_2 \) respectively. If \( label(v_1) = label(v_2) \neq s \), then we call \( [v_1, v_2] \) a label-matching pair. We call \( [v_1, v_2] \) a /-child pair, if \( v_1 \) is a descendant of its parent via a //edge. A pair refers to either a label-matching pair or a /-child pair.

Given a pair \( [u_1, u_2] \), if \( v_1 \) is a descendant of \( u_1 \), and...
Theorem 3.5: \( B \) is a pattern in \( \hat{\mathbb{P}} \).

Case, we can construct the MCC of \( B \), i.e., (1) the patterns in Figure 6 (a) and (b) is shown in Figure 6 (c).

Similar to boolean patterns in \( \hat{\mathbb{P}} \), if \( v_1, v_2 \) is a label-matching descendant pair of \( u_1, u_2 \).

Given a label-matching pair or \(/\)-child pair \( v_1, v_2 \), we can construct a boolean tree pattern \( T(v_1, v_2) \) of height 1 as follows.

1. The root of \( T(v_1, v_2) \) is labeled with \( \text{label}(v_1) \) if \( \text{label}(v_1) = \text{label}(v_2) \neq \ast \), otherwise the root is labeled \( \ast \).

2. For every \(/\)-child pair \( u_1, u_2 \) of \( v_1, v_2 \), add a \(/\)-child to the root. If \( u_1, u_2 \) is also a label-matching pair, then label the new node with \( \text{label}(u_1) \). Otherwise, label the new node with \( \ast \).

3. If \( v_1, v_2 \) is a label matching pair, then for every CLMDP \( u_1, u_2 \) of \( v_1, v_2 \), add a \(/\)-child labeled with \( \text{label}(u_1) \) to the root.

We can now construct a boolean tree pattern \( P \) as follows:

1. Initially, let \( P = T(\text{root}(P_1), \text{root}(P_2)) \).

2. For every leaf node in \( P \), if it corresponds to the pair \( u_1, u_2 \), then replace it with \( T(u_1, u_2) \) (by merging it with the root of \( T(u_1, u_2) \)).

3. Repeat the above step until no more nodes can be added to \( P \). i.e., the leaf nodes in \( P \) corresponds to a descendant pair which has no descendant parents.

4. Remove dangling \( \ast \)-nodes.

It is straightforward to see that the pattern \( P \) is in \( \mathcal{B}(\hat{\mathbb{P}}(\mathbb{P}_{\hat{\mathbb{P}}})^\ast) \).

Example 3.4: The pattern constructed for the boolean patterns in Figure 6 (a) and (b) is shown in Figure 6 (c).

Theorem 3.4: \( P \) is a MCC of \( P_1 \) and \( P_2 \) wrt \( \mathcal{B}(\hat{\mathbb{P}}(\mathbb{P}_{\hat{\mathbb{P}}})^\ast) \), i.e., (1) \( P_1, P_2 \subseteq P \), and (2) for every boolean pattern \( Q \in \mathcal{B}(\hat{\mathbb{P}}(\mathbb{P}_{\hat{\mathbb{P}}})^\ast) \), if \( P_1, P_2 \subseteq Q \), then \( P \subseteq Q \).

4. CONCLUSION

We showed the existence of MCCs for some tree patterns in \( \mathcal{P}(\mathbb{P}_{\hat{\mathbb{P}}}) \) and \( \mathcal{P}(\mathbb{P}_{\hat{\mathbb{P}}})^\ast \), and provided a way to construct such a MCC. As future work, we would like to investigate more efficient algorithms to compute the MCCs, as well as the boundary which divides tree patterns that have MCCs from those that do not.

5. REFERENCES

