Incremental Updates on Compressed XML

Citation for published version:

Digital Object Identifier (DOI):
10.1109/ICDE.2016.7498310

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
2016 IEEE 32nd International Conference on Data Engineering (ICDE)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Incremental Updates on Compressed XML

Stefan Böttcher*, Rita Hartel*, Thomas Jacobs* and Sebastian Maneth†
*University of Paderborn, Fürstenallee 11, 33102 Paderborn, Germany
Emails: {stb@, rst@, tjacobs@mail.}uni-paderborn.de
†University of Edinburgh, Informatics Forum, Crichton Street, Edinburgh, EH8 9AB, UK
Email: smaneth@inf.ed.ac.uk

Abstract—XML tree structures can be effectively compressed using straight-line grammars. It has been an open problem how to update straight-line grammars, while keeping them compressed. Therefore, the best previous known methods resort to periodic decompression followed by compression from scratch. The decompression step is expensive, potentially with exponential running time. We present a method that avoids this expensive step. Our method recompresses the updated grammar directly, without prior decompression; it thus greatly outperforms the decompress-compress approach, in terms of both space and time. Our experiments show that the obtained grammars are similar or even smaller than those of the decompress-compress method.

I. INTRODUCTION

Typical XML document trees compress well by means of dictionary-based methods. Buneman, Koch, and Grohe [1] show that XML document trees compress to about 10% of their edges, when represented as minimal DAGs (directed acyclic graphs). The idea of DAGs is to represent repeating subtrees only once. This idea was generalized to the sharing of repeated (connected) XML subgraphs by Busatto, Lohrey, and Maneth [2]. They show that straight-line tree (SLT) grammars compress typical XML trees to 3% of their edges. These ratios are further improved by the current state-of-the-art SLT compressor TreeRePair [3].

Dictionary-based compression methods are special, because they can provide in-memory representations which are queryable without decompression. TreeRePair provides the smallest queryable in-memory representation of XML document trees that we are aware of [4].

These features make SLT grammars ideal for in-memory XML processing. The downside however is: no previous method supports efficient updates, i.e., updates that can be executed on the grammar without decompression to the tree and that keep the updated grammar small. Updates lead to a significant blow-up in grammar size. In order to overcome this problem, the best known method is, after some updates, to decompress to a tree and then compress the resulting tree again. This can take exponential time and space. Therefore, SLT-compressed trees have only been applied in static scenarios. Many applications however rely on dynamic trees that change frequently; consider for instance DOM-trees used in web browsers.

This paper presents the first implementation of SLT-compressed trees with efficient updates. The idea is to generalize TreeRePair from running on trees to running on grammars. In this way, updates can be realized by first efficiently producing an updated grammar and then running GrammarRePair over this grammar. Our contributions are:

1) We devise the compression algorithm GrammarRePair which executes RePair compression directly on an SLT grammar, without prior decompression.
2) We show that on typical XML trees, GrammarRePair compresses as well as TreeRePair; on some trees it even compresses considerably better.
3) We investigate the update performance of SLT grammars under GrammarRePair compression:
   a) For sequences of thousands of updates (inserts and deletes), our grammars achieve virtually the same compression ratios as achieved by the update-decompress-compress (udc) method.
   b) For typical files, our space consumption overhead is less than 1% compared to the compression result of udc, while naive updates (without further compression) cause about 40% overhead. For a few files with extreme compression ratios, our overhead compared to the compression result of udc is a factor of two, while naive updates cause blow-ups of more than 100-fold.
   c) For XML trees with more than 100k edges, our update time is faster than that of udc. For trees with more than 200k edges, our update time is even faster than the mere compression time of udc.

Let us describe in a few more words the idea of our method. The idea of RePair compression is to repeatedly replace the most frequent digram by a new nonterminal. For strings [5] a digram simply consists of two adjacent symbols. Thus, on the string \( w = ababababa \) we can replace \( ab \) by a nonterminal symbol \( A \), obtaining \( AAAAa \), and then replace \( AA \) by \( B \), obtaining the final grammar \( G_w = \{ S \rightarrow BBa, B \rightarrow AA, A \rightarrow ab \} \). Note that the size of this grammar (= sum of lengths of right-hand sides) is 7, while the original string has length 9. In a tree, a digram is a triple \( (a, i, b) \) where \( a, b \) are labels of adjacent nodes and \( i \) is a child number. Note that a naive implementation that counts digrams after each round of replacements, does not run in linear time. A linear time implementation is non-trivial. It was solved by Larson and Moffat [5] through careful incremental updates of digram occurrence lists together with clever data structures such as a priority queue of length \( \sqrt{n} \) (where \( n \) is the length of the input string) holding frequent digrams. To see the challenge caused by an update, consider changing the \( w \) above into \( v = ba \). A grammar for this is \( H_v = \{ S \rightarrow bBBa, B \rightarrow AA, A \rightarrow ab \} \). However, this is not the grammar RePair would produce on \( v \), because initially \( ba \) is the most frequent digram. The desired smaller grammar would be \( G_v = \{ S \rightarrow BBa, B \rightarrow AA, A \rightarrow ba \} \). The question arises, how this grammar \( G_v \) can efficiently be obtained from the grammar \( H_v \)?
In our setting, we need to count digram occurrences not only in a string, but in a tree \( T \) generated by the SLT grammar \( G \), but without decompressing \( G \) to \( T \). We show how this is possible in one pass over an SLT grammar.

The main challenge is the efficient replacement of all occurrences of a given digram \( \alpha = (a, i, b) \) in a tree \( T \) generated by a grammar \( G \), in one pass through \( G \), and with minimal decompression of \( G \). One major technical issue is that an occurrence of \( \alpha \) need not reside in just one rule, but can span over several rules. This implies that we first have to apply these rules (thus locally decompressing) in order to make this occurrence explicit to be able to replace it. The delicate issue is to do as few rule applications as possible (by running them in a particular order), and to apply these rules in a “compact” way (by introducing new rules). At the end of the Preliminaries, we give a more technical outline of our technique.

Related Work

The idea of updates through path isolation is used by Fisher and Maneth [6]. For typical XML documents, they find that long update sequences of delete/inserts (on the same document) increase grammar sizes by around 40%. Their SLT grammars are produced by BPLEX [2].

Bätz, Böttcher, and Hartel [7] use path isolation to find that updates can be performed faster than using the decompress-update-compress method. Moreover, they show how to perform multiple updates in parallel during one grammar pass. This approach is further improved by Böttcher, Hartel, and Jacobs [8] by computing DAG-compressed representations of the update positions induced by an XPath query (when evaluated over the given SLT grammar).

For straight-line string grammars, digram replacement in a given grammar has been used by Jez [9] in his recompression framework. For this, he distinguishes “non-crossing” digrams that fully reside in rules (and thus are easy to replace) from “crossing” digrams which are more difficult to replace. He shows that all occurrences of a digram can be replaced in linear time with respect to the size of the grammar. Jez’s recompression approach has been generalized to trees by Jez and Lohrey [10] in order to present a compression algorithm by SLT grammars with the currently best proven approximation ratio (with respect to the minimal SLT grammar, which is NP-complete to compute, cf. [11]).

Succinct trees offer a compact tree representation, see, e.g., the work by Munro and Raman [12]. They have been used to build in-memory DOM representations for XML trees by Delpratt, Raman, and Rahman [13]; while this representation is space efficient and supports fast navigation, it does not support updates. Dynamic succinct versions are more complicated and efficient implementations are still missing, even though appropriate data structures have recently been implemented by Joannou and Raman [14], cf. the concluding remarks in [15].

II. Preliminaries

Our formal models of an XML tree and a tree grammar follow and slightly extend [3]. Like TreeRePair, we consider an input document as a labeled binary tree, where each non-leaf node has exactly two children, i.e., it has \( \text{rank} \) two: its first-child and its next-sibling (Figure 1). Different from TreeRePair, we introduce a leaf node with label \( \perp \) called the empty node to represent non-existing first-child or next-sibling nodes. Our compression works on ranked labeled ordered trees of which binary XML trees are just a special case.

A ranked alphabet \( \Sigma \) consists of a finite set of symbols each of which is associated a natural number, called the rank of the symbol. We fix the set \( Y = \{y_1, y_2, \ldots\} \) of formal parameters, and assume it to be disjoint with all other ranked alphabets. A finite tree (over \( \Sigma \)) with parameters is a tree \( t \), such that for every node \( u \) of \( t \), if the label of \( u \) is a symbol in \( \Sigma \) of rank \( k \), then \( u \) has exactly \( k \) children. Let params(\( t \)) denote the number of nodes of \( t \) that have their label in \( Y \). We demand that if params(\( t \)) = \( m \), then for each \( i \in \{1, \ldots, m\} \), there is exactly one node in \( t \) labeled by \( y_i \). Thus, our trees are linear in the parameters \( Y_m = \{y_1, \ldots, y_m\} \).

Let \( t, t' \) be trees and \( v \) be a node of \( t \). By \( t[v/t'] \), we denote the replacement of the subtree rooted at node \( v \) by the tree \( t' \). Now let \( \text{params}(t') = m \) and \( t_1, \ldots, t_m \) be trees. Then \( t[y_{1}/t_1, \ldots, y_{m}/t_m] \) denotes the tree obtained from \( t \) by replacing each occurrence of \( y_j \) by the tree \( t_j \), for \( 1 \leq j \leq m \). If \( t[v/t'][y_1/t_1, \ldots, y_n/t_n] \) = \( t \), then \( t' \) is a pattern of rank \( m \) occurring in tree \( t \) at node \( v \). For example, the pattern \( a(y_1,a(y_2,y_3)) \) occurs at four different nodes in the tree of Figure 1, namely, at the top \( a \)-node of each dashed box.

To compress a ranked labeled ordered tree by extracting patterns that occur multiple times, so-called straight-line linear context-free (SLCF) tree grammars can be used. Using the model of [3], a linear context-free tree grammar is a 4-tuple \( G = (\mathcal{F}, \mathcal{N}, P, S) \), where \( \mathcal{F} \) is the set of ranked terminal symbols (including the null pointer \( \perp \) with \( \text{rank}(\perp) = 0 \)), \( \mathcal{N} \) is the set of ranked nonterminal symbols, \( \mathcal{F} \cap \mathcal{N} = \emptyset \), \( P \) denotes the set of rules, and \( S \in \mathcal{N} \) is the start nonterminal symbol not occurring on the right-hand side of any rule. For each \( A \in \mathcal{N} \) of rank \( m \), there is exactly one rule \( (R \to t_B) \in P \), where \( t_B \) must not consist of a single node that is a parameter. Each such \( t_B \) is a tree over \( \mathcal{F} \cup \mathcal{N} \cup \mathcal{M} \), such that each \( y_j \in \mathcal{M} \) occurs exactly once in \( t_B \). The derivation relation of \( G \) is denoted by \( \Rightarrow_G \) (or by \( \Rightarrow \)). For trees \( t, t', t \Rightarrow_G t' \) if (1) \( t \) contains a subtree \( t_0 = A(t_1, \ldots, t_m) \) with \( A \in \mathcal{N} \) and (2) \( t' \) is obtained from \( t \) by replacing \( t_0 \) by the tree \( t_A[y_1/t_1, \ldots, y_m/t_m] \).

In order to identify a particular node \( v \) occurring in the right-hand side \( t_B \) of a grammar rule \( B \to t_B \), we denote \( v \) as \( (B,m) \), where \( n \) is the number of nodes (terminals, nonterminals, parameters) occurring in \( t_B \), \( 1 \leq m \leq n \), and \( (B,m) \) is the \( m \)-th node of \( t_B \) in preorder.

For example, consider the grammar \( G = (\mathcal{F}, \mathcal{N}, P, S) \) with \( \mathcal{F} = \{a, f, \perp\} \), \( \mathcal{N} = \{A, B, S\} \), and \( P = \{S \to \ldots \}

![Fig. 1. An XML tree and its binary tree](image-url)
$f(A(B(B), \bot), B \rightarrow A(\bot, \bot), A \rightarrow a(\bot, a(y_1, y_2)))$. Then, $B \Rightarrow A(\bot, \bot)$; we apply the $A$-rule to the tree $A(\bot, \bot)$ by replacing $y_1$ and $y_2$ in $t_A$ by $\bot$, i.e., $A(\bot, \bot) \Rightarrow a(\bot, a(\bot, \bot)) = \bot$. Then, $A(B(B) \Rightarrow \bot, A(\bot, a(B,B)) \Rightarrow a(\bot, a(t,t))$. Thus, $S \Rightarrow f(A(B,B), \bot)$ which derives to the binary tree $f(a(a(a(t,t)), \bot)$ which is shown in Figure 1.

Furthermore, $ref_c(Q) = \{(R,n) | (R \rightarrow t_R) \in P \land label(t_R,n) = Q\}$ denotes the set of the $Q$-labeled nonterminals within the rules of $G$. If $(R,n) \in ref_c(Q)$, we also say that $R$ calls $Q$, $calls_c(Q, R)$ for short. Then, the transitional closure, $calls^{\ast}_c(Q, R)$, contains all pairs of nodes, where $R$ directly or indirectly calls $Q$. We consider only grammars that are non-recursive, also called straight-line, i.e., $\forall Q : calls^{\ast}_c(Q, Q)$. We say that $Q$ occurs before $R$ in anti-straight-line order (anti-SL) in $G$, if $calls^{\ast}_c(Q, R)$.

To inline a rule, means for $(R,n) \in ref_c(Q)$ that occurs in the right-hand side of a rule $R \rightarrow t_R$, we modify tree $t_R$ by replacing $(R,n)$ by $t_Q$, where the $i$-th parameter node of $t_Q$ is replaced by the subtree rooted in the $i$-th child node of $(R,n)$. Formally, let $rank(Q) = n$ and let $t_i$ be the $i$-th subtree of $(R,n)$. That is, inlining $Q$ into $(R,n)$ is formally defined by the replacement $t_R[(R,n)/(t_{Q[y_1/t_1}, ..., t_{Q[y_n/t_n]})]$. If afterwards $|ref_c(Q)| = 0$, we delete rule $Q \rightarrow t_Q$ from grammar $G$. For example, if we inline rule $B$ for node $(S,3)$, we obtain $S \rightarrow f(A(A(\bot, \bot), B, \bot)$. Let $val_c(R)$ be the tree corresponding to $R$, i.e., the tree that we obtain when repetitively inlining nonterminals in $t_R$, until $t_R$ does not contain any nonterminals. As a special case, $val_c(G) = T$ is the tree generated by the grammar $G$. Let $b$ be a terminal node in any rule of a grammar $G$ with start symbol $S$, and let $T = val_c(S)$ be the tree generated by inlining all rules into $S$. We define a correspondence between $b$ and a set of nodes $v$ in $T$ by the following procedure. We mark $b$ in $G$, and whenever a rule containing a marked node $b$ is inlined, the mark is copied during the inlining step. When no more rules can be inlined, i.e., when we have transformed $G$ into $T$, all the marked nodes in $T$ are the nodes that correspond to $b$.

As defined in [3], a digram $\alpha = (a, i, b)$ is a triple, where $a,b \in \mathcal{F}$ and $i \in \{1, ..., rank(a)\}$. It denotes an edge from an $a$-labeled node to its $i$-th $b$-labeled child node. An occurrence of a digram $\alpha = (a, i, b)$ is a pair of nodes $v$ and $\alpha \in T$, such that $label(v) = a$ and $v$’s $i$-th child is $w$ with $label(w) = b$. Two occurrences of the same digram $\alpha$ are overlapping, if they share a common node in $T$. For two distinct occurrences of $\alpha$, overlapping can only happen if $a = b$.

For example, in the tree $T$ shown above in Figure 1, we have several occurrences of the digram $(a,2,a)$ marked by a dashed box. The occurrences marked with a blue box overlap with the occurrence marked by the red box.

Finally, let $\alpha = (a, i, b)$ be a digram with $rank(a) = m$ and $rank(b) = n$. The pattern $t_X$ representing $\alpha$ is defined to be $a(y_1, ..., y_i-1, b(y_i, ..., y_{i+n-1}), y_{i+n}, ..., y_{m+n-1})$.

$\alpha$ is called appropriate if

1) $rank(\alpha) = m + n - 1 \leq k_m$, where $k_m$ is a predefined constant limiting the maximum numbers of parameters of a rule $X \rightarrow t_X$ and
2) $\alpha$ has more than one occurrence within $T$.

Outline of the Proposed Solution. Recompression of a grammar $G$ is done stepwise by finding a most frequent digram $\alpha = (a, i, b)$ in $val_c(G)$, partially decompressing $G$, such that all digram occurrences of $\alpha$ are isolated, and replacing all digram occurrences of $\alpha$ with a new nonterminal $X$. In order to minimize the amount of decompression during partial decompression of $G$, we first minimize the amount of inlining steps, i.e., replace only those nonterminals $A$ by $t_A$, for which we can isolate a terminal $a$ or a terminal $b$ that belongs to a digram occurrence of $\alpha$. Second, instead of replacing each nonterminal $A$ determined in the first step by $t_A$, we transform $t_A$ into an equivalent tree $t_A'$ by recompressing parts of $t_A$ that are not needed to isolate $a$ or $b$, and then replace $A$ by the smaller tree $t_A'$.

III. Grammar Updates and Motivating Examples

XML tree structures can be updated using three atomic update operations: renames, deletes, and inserts. For this section, we mostly consider renames. Let $t$ be a binary tree representation of an XML tree, $u$ a node of $t$, and $\sigma$ be a label. Then rename($t, u, \sigma$) denotes the tree obtained from $t$ by relabeling the node $u$ by the label $\sigma$. To guarantee that the resulting tree is again a correct binary tree representation of an XML tree, we require that $\sigma \neq \bot$ and that the label of $u$ in $t$ is not $\bot$. As an example, let $t = f(d(\bot, b(\bot, a(\bot, b(\bot, \bot))))$ and $u = 2$, i.e., $u$ denotes the $d$-node in $t$. Then rename($t, u, a$) is the tree $f(a(\bot, b(\bot, a(\bot, b(\bot, \bot))))).

A. Path Isolation

In order to perform an update on a tree that is given as grammar, we first need to “isolate” the node at which the operation is to be performed. To see this, consider the following string grammar $G_5$:

$$\{A \rightarrow BB, B \rightarrow CC, C \rightarrow DD, D \rightarrow ab\}.$$ 

This grammar represents the string $(ab)^8$. We would like to rename the first $a$-symbol in this string by a $c$. Clearly, if we simply change the $a$-symbol of the $D$-production into $\alpha$, then the whole string changes to $(cb)^8$ which is not what we want. The issue is that $D$ appears many (eight) times in the string, and only the very first occurrence should be changed. Thus, we must “unfold” the grammar until the particular terminal symbol is available, and then perform the change. In the example, we can use this derivation:

$$A \Rightarrow BB \Rightarrow CCB \Rightarrow DDCB \Rightarrow abDCB.$$ 

Since the symbol $a$ occurs in its terminal form, at the correct position of the update, we may now change it, and obtain $cbDCB$. Thus, a grammar with rule $A \rightarrow cbDCB$ and the $B$, $C$, and $D$ rules as before, represents the renamed tree. We call this process of making a node terminally available a “path isolation”. As another example, consider this grammar $G_{exp}$:

$$A \rightarrow A_1 A_1, A_1 \rightarrow A_{i+1} A_{i+1}, \text{ for } 1 \leq i \leq 9, A_{10} \rightarrow a.$$ 

The grammar represents the string $a^{1024}$. Note that the size of the grammar is 21. Assume that we want to rename the letter at position 333 by the letter $c$. Since $A_1$ produces a string of length 512, we know that this position is produced by the left
A\textsubscript{1} in the A-production. Thus, we derive \( A \Rightarrow A\textsubscript{2}A\textsubscript{2}A\textsubscript{1} \). Since \( A\textsubscript{2} \) produces a string of length 256, we now need to expand the second \( A\textsubscript{2} \), to obtain \( A\textsubscript{2}A\textsubscript{3}A\textsubscript{3}A\textsubscript{1} \). After a few more steps, we obtain
\[
A\textsubscript{2}A\textsubscript{3}A\textsubscript{4}A\textsubscript{7}A\textsubscript{8}aA\textsubscript{1}0A\textsubscript{9}A\textsubscript{2}A\textsubscript{6}A\textsubscript{5}A\textsubscript{3}A\textsubscript{1}
\]
and as we may verify: 256 + 64 + 8 + 4 = 332, thus, the \( a \) above is at the correct position.

Let \( G \) be a grammar and \( u \) a node in \( \text{val}_G(S) \). In one pass through \( G \), we can precompute for every nonterminal \( A \) of \( G \) of rank \( k \), the numbers \( \text{size}(A, 0), \text{size}(A, 1), \ldots, \text{size}(A, k) \). These are the numbers of nodes in \( \text{val}_G(A) \) that appear before \( y_1 \) (in pre-order), between \( y_1 \) and \( y_2 \), . . . , after \( y_k \). For instance, if \( \text{val}_G(A) = f(y_1, g(h(a, y_2), g(a, y_3))) \) then size\((A, 0) = 1,\)
\[
\text{size}(A, 1) = 3, \text{size}(A, 2) = 2, \text{and size}(A, 3) = 0.
\]

Using these numbers, it is straightforward to determine a shortest derivation \( S \Rightarrow * \xi \), such that a terminal symbol occurs in \( \xi \) that uniquely represents \( u \): we start at the root node of the right-hand side of \( S \). Assume that it is a nonterminal \( X \) of rank \( k \). Using the size-information, we can determine whether \( u \) is produced by \( X \), and if not, determine whether the \( j \) in \( \{1, \ldots, k\} \) such that \( u \) is produced in the \( j \)-th subtree. The tree \( \xi \) is the path isolation of \( G \) for \( u \), denoted by \( \text{iso}(G, u) \). Path isolation is used already in [6]. The following lemma is straightforward because each production is applied at most once during path isolation.

**Lemma 1.** Given a grammar \( G \) and a node \( u \) in \( \text{val}_G(S) \), \( |\text{iso}(G, u)| \leq 2 \cdot |G| \).

**B. Updating the Path-Isolated Grammar**

As we have seen, even performing a single rename operation over a grammar can, due to path-isolation, cause a blow up of (almost) a factor two. Clearly, if we repeatedly apply more and more updates, then we can blow-up again and again, eventually loosing all compression of the grammar. In general this cannot be avoided: consider renaming every single node by a new unique symbol. Then the output cannot be compressed, thus, after \( n \) steps (where \( n \) is the size of the uncompressed tree) we obtain a grammar of size \( n \), which does not exhibit any compression.

Typically, however, the grammars after update are much larger then they need to be. Consider again the grammar \( G\textsubscript{8} \) of before. We now insert a \( b \) symbol before the first symbol, and an \( a \) symbol after the last symbol. These positions are available immediately in the start rule’s right-hand side, without any further path isolation needed. We obtain this grammar:
\[
\{A \rightarrow bBbA, B \rightarrow CC, C \rightarrow DD, D \rightarrow ab\}.
\]
The represented string is \( b(ab)^8a \). How would RePair compress this string? Clearly, the most frequent digram now is \( ba \) and not \( ab \) as previously. Therefore, we want to execute repair compression directly on the grammar shown above, causing as little as possible decompression.

**C. Grammar Recompression**

Let us outline our algorithm on this example. At first, we need to count the number of occurrences of digrams. Let us defer this step until later, and assume that we have determined that \( ba \) is the most frequent digram and hence shall be replaced by the new nonterminal \( X \). We traverse the productions bottom-up. Obviously, the right-hand side \( ab \) of \( D \) does not contain any occurrence of \( ba \). We record that the first letter of \( D \) is an \( a \), and that the last letter of \( D \) is a \( b \). Consider now the production \( C \rightarrow DD \). None of the \( D \)'s contains \( ba \), however, the string of \( DD \) does contain an occurrence, stemming from the last letter of the first \( D \) and the first letter of the second \( D \). Hence, this production is changed into
\[
C \rightarrow aXb.
\]

With this, the right-hand side of \( B \) rewrites to \( aXbaXb \) and hence becomes \( aXXXb \). In order to limit decompression when we need to insert this string for \( B \) in the future, we would introduce a new auxiliary nonterminal \( Y \) for \( XXX \). We refer to the memorization technique as lemma generation. Thus, we have \( \{B \rightarrow aYb, Y \rightarrow XXX\} \). Finally, we replace \( B \) by \( aYb \) in the right-hand side of \( A \) to obtain: \( baYbaYba \), which becomes \( XYXYX \). Thus, we obtain the following grammar:
\[
\{A \rightarrow XYXYX, Y \rightarrow XXX, X \rightarrow ba\}.
\]
Note that the size of this grammar is 10, while the corresponding grammar without lemma generation has size 11 (its \( A \) rule has \( X^9 \) in its right-hand side). Thus, even in this tiny example, lemma generation is beneficial.

We now need to run a next round of RePair. Again, let us defer how digram occurrences are counted in this grammar, and assume that \( XX \) is the most frequent digram. We introduce \( Z \rightarrow XX \) and change \( Y \rightarrow XXX \) into \( Y \rightarrow ZX \). With this, the right-hand side of \( A \) becomes \( XXZZXX \) and thus \( XXXZZZ \). The resulting grammar is \( \{A \rightarrow XZZZZ, Z \rightarrow XX, X \rightarrow ba\} \). Even though this will not change the size of this grammar, RePair would replace \( ZZ \) by a new nonterminal \( W \) to yield this final grammar:
\[
\{A \rightarrow XWW, W \rightarrow ZZ, Z \rightarrow XX, X \rightarrow ba\}.
\]
Observe, when \( Y \)'s right hand side \( XXX \) was changed to \( ZX \). This is a left-greedy replacement. Such a replacement need not be optimal. Consider a rule \( S \rightarrow XYYX \). Replacing \( Y \) by \( ZX \) in this rule yields \( XXZZXX \) which contains only one occurrence of \( XX \) (and thus would become \( XZZZZZ \)). Similarly, choosing the right-greedy \( ZX \) would give us \( XXZZXX \) and then \( ZZXXZ \). To get an optimal result (especially in the presence of more occurrences of \( YY \), \( X \), \( Z \)), one would need to introduce two rules: \( Y \rightarrow ZX \) and \( Y' \rightarrow ZX \). With these rules, the \( S \)-rule becomes \( XXXXXX \) and thus \( ZZ \). In the case of trees, even two rules would not be enough, e.g., the rule \( Y \rightarrow a(a(y_1, y_2), a(a(y_3, y_4), y_5)) \) would require four versions. In a tree, left-greedy matching of a digram is generalized to “top-down greedy” matching. In summary:

1. we use lemma generation in order to keep the grammar compressed during replacement of digrams, and
2. we always replace digrams in a left-greedy way, in order to only generate one version of a production.

**IV. Grammar RePair Compression**

TreeRePair [3] takes as input a labeled ordered binary tree \( T \) and produces as output an SLCF tree grammar \( G' \) with
GrammarRePair works on so called digrams, i.e., pairs of adjacent terminal nodes of the tree \( T \) generated by \( G \).

Algorithm 1, the steps of which are later described in more detail, shows an overview of GrammarRePair.

\begin{algorithm}
\caption{GRAMMARREPAIR(\( G = (F, N, P, S) \))}
\begin{algorithmic}
  \State \textit{RetrieveOccs}(\( G \))
  \While {\textit{MostFrequentDigram} exists}
    \State replace each occurrence of \( \alpha \) in \( G \) by \( X \), where \( X \) is a new nonterminal
    \State \( P = P \cup \{X \rightarrow t_X\} \) with \( t_X \) representing \( \alpha \)
    \State \( F = F \cup X \)
    \State \textit{UpdateDigramOccurrences}
  \EndWhile
  \State \textit{PruningPhase}
\end{algorithmic}
\end{algorithm}

First, GrammarRePair computes a set of digram occurrences within \( G \) for each digram and counts how often the occurrences are present within \( T \) (line 1). Afterwards, it iterates through the following steps:

1) A most frequent appropriate digram \( \alpha \) regarding \( T \) is selected (line 2).

2) Each digram occurrence of \( \alpha \) in \( G \) is replaced by \( X \), where \( X \) is a new nonterminal (line 3).

3) A rule \( X \rightarrow t_X \) is added to \( G \) where the pattern \( t_X \) represents the digram \( \alpha \) (line 4). In the following steps, \( T \) is the tree generated by the grammar that treats \( X \) as a terminal (line 5).

4) As the grammar is modified, the sets of digram occurrences are updated (line 6).

Finally, if no more appropriate digram exists, the \textit{Pruning Phase} (line 7) removes all rules that do not contribute to the compression. To be more precise, GrammarRePair consists of the following steps outlined in the sections IV-A to IV-D.

A. Counting and Locating Digram Occurrences

In this section, we first explain the steps of TreeRePair working on trees as a simple case, and then extend each step to work on arbitrary SLCF grammars.

Considering the simpler case of a tree as input, TreeRePair searches for digrams in a tree \( T \) and counts the sizes of maximal sets of non-overlapping occurrences of each digram. This is done in a bottom-up greedy way: TreeRePair traverses \( T \) in post-order and adds each occurrence of a digram \( \alpha \) to the digram list of \( \alpha \), provided it does not overlap with an occurrence that is already stored in the digram list of \( \alpha \).

While the location of each digram occurrence \((v, w)\) in \( T \) can be uniquely described by the child node \( w \), as its parent \( v \) is unique, an SLCF grammar may contain a terminal node \( w \) that belongs to different digram occurrences \((v_1, w), (v_2, w)\).

For an arbitrary SLCF grammar \( G \) as input, we define digrams, digram occurrences, their generators, their tree parent, and their tree child in \( G \) as follows.

Let \( R \rightarrow t_R \) be a rule in \( G \) and \((R, n)\) a node in \( t_R \) with \( n \neq 1 \) and \( \text{label}(t_R, n) \neq y_i, i \in \mathbb{N} \), i.e., \((R, n)\) is neither the root of \( t_R \) nor a parameter. Then, \((R, n)\) is called a digram occurrence generator of a digram occurrence \(((\alpha, k), (B, m))\) of a digram \( \alpha = (a, b) \) in grammar \( G \), where \( a = \text{label}(t_A, k) \) and \( b = \text{label}(t_B, m) \).

We call the node \((B, m) = \text{TreeChild}(R, n)\) the tree child in \( G \) of this digram occurrence, and the node \((A, k)\) with \(((\alpha, k), i) = \text{TreeParent}(R, n)\) the tree parent in \( G \) of this digram occurrence. Here, the functions \text{TreeChild} and \text{TreeParent} are defined as follows.

\begin{algorithm}
\caption{TreeChild(B, m)}
\begin{algorithmic}
  \While {\( (B, m) \) is nonterminal}
    \State \( B := \text{label}(t_B, m); \)
    \State \( m := 1; \)
  \EndWhile
  \State \textbf{return} \((B, m)\)
\end{algorithmic}
\end{algorithm}

The function \text{TreeParent} returns the tree parent of \((R, n)\) and the child index \( i \) as follows. The auxiliary method \text{parent}(t, m) computes the position of the parent node of the \( m \)-th node in preorder within \( t \). Furthermore, \( i := \text{index}(t, m) \) denotes that the \( m \)-th node within \( t \) is the \( i \)-th child node of the node at position \text{parent}(t, m). Note that \text{TreeParent} is only called for nodes \((A, k)\) that are not the root of the tree \( t_A \), i.e., nodes with \( k \neq 1 \).

\begin{algorithm}
\caption{TreeParent(A, k)}
\begin{algorithmic}
  \While {\text{label}(t_A, \text{parent}(t_A, k)) \text{ is nonterminal}}
    \State \( i := \text{index}(t_A, k); \)
    \State \( A := \text{label}(t_A, \text{parent}(t_A, k)); \)
    \State \( k := \text{index}(t_A, y_i); \)
  \EndWhile
  \State \textbf{return} \(((A, \text{parent}(A, k)), \text{index}(t_A, k))\)
\end{algorithmic}
\end{algorithm}

Consider the following fragment “Grammar 1” of an SLCF grammar

\[
\begin{align*}
C & \rightarrow A(B(\bot), \bot) \\
A & \rightarrow a(y_1, 1(B(\bot), a(\bot, y_2))) \\
B & \rightarrow b(y_1, \bot)
\end{align*}
\]

For example, consider the node \((C, 2)\) labeled with the label \( B \) in the first rule of Grammar 1 as digram occurrence generator. By calling \text{TreeChild}(C, 2), we obtain the terminal \((B, 1)\) having label \( b \) as \((C, 2)\’s\) tree child. And by calling \text{TreeParent}(C, 2), we obtain \(((A, 1), 1)\), i.e., the terminal node \((A, 1)\) having label \( a \) as \((C, 2)\’s\) tree parent and the index 1 as the child index. This together forms an occurrence of the digram \( \alpha = (a, 1, b) \).

Let \((R, n)\) be a digram occurrence generator in \( G \) with tree child \((B, m)\), tree parent \((A, k)\), and child index \( i \) and let \( T = \text{val}_G(S) \) be the tree generated by \( G \). We define the set \( \text{CDO}_{(R, n)} \) of digram occurrences in \( T \) corresponding to \((R, n)\) as

\[
\text{CDO}_{(R, n)} = \{(v_T, w_T) \mid v_T \text{ corresponds to } (A, k) \text{ and } w_T \text{ corresponds to } (B, m) \text{ and } w_T \text{ is the } i\text{-th child of } v_T \text{ in } T\}.
\]
As one digram occurrence generator within a rule \( Q \to t_Q \) of \( G \) can correspond to multiple digram occurrences within the tree \( T \) generated by \( G \), we have to compute based on \( G \), the number of digram occurrences within \( T \). Informally, this depends on how often the nonterminal \( Q \) is used within \( G \) to generate \( T \). Formally, we recursively define the function \( usage_G \) with \( usage_G(S) = 1 \), and for each \( Q \in N \setminus \{S\} : usage_G(Q) = \sum_{(R,n) \in ETREEC(Q)} usage_G(R) \). In general, for each digram occurrence generator \((C,m_c)\), the number of corresponding digram occurrences within \( T \) is given by \( usage_G(C) \).

For example, in Grammar 1, if we assume that \( usage_G(C) = 3 \) and \( A \) is called only twice in a further rule \( S \) of Grammar 1, \( usage_G(A) = 2 * usage_G(S) + usage_G(C) = 2 * 1 + 3 = 5 \), i.e., the occurrence of the digram \( \alpha = (a,1,b) \) generated by node \((A,4)\) corresponds to five occurrences of \( \alpha \) in tree \( val_G(S) \).

One major challenge when dealing with SLCF tree grammars instead of DAGs or trees is to find a maximum set of non-overlapping occurrences of a digram \((b,i,b)\).

To avoid overlapping occurrences, we use the same restriction as TreeRePair uses on DAG-compressed grammars. That is, we do not consider occurrences of equal label digrams where the tree child is the root of a rule. Note however that we do consider occurrences of digrams of the form \((b,i,b)\) that cross parameter boundaries.

### Algorithm 4: RETRIEVEOCCS (G = (F, N, P, S))

1. **foreach** \((C \to t_C) \in P, C \in N;\) in anti-SL order do
2.   **foreach** \((C, n) \in t_C\) in pre-order do
3.     if \( n \neq 1 \land label(t_C, n) \neq y_j \) then
4.       \((A, k), i :=\) TREEPARENT(C,n);
5.       \((B, m) :=\) TREECHILD(C,n);
6.       \( \alpha := (label(t_A, k), i, label(t_B, m));\)
7.       if \( label(t_A, k) \neq label(t_B, m) \) then
8.         \( label(t_C, n) \neq N \land\)
9.         \((A, k) \notin occ_G(\alpha) \) then
10.    \( occ_G(\alpha) := occ_G(\alpha) \cup \{(C, n)\};\)

Algorithm RETRIEVEOCCS computes the set \( occ_G(\alpha) \) of generators of all non-overlapping occurrences of all digrams \( \alpha = (b,i,b) \) and of all occurrences of digrams \( \alpha = (a,i,b) \) within grammar \( G \) in parallel within a single pass through grammar \( G \). Therefore, RETRIEVEOCCS assumes that the tree parents \((A,k)\), the child indices \( i \) (line 6), and tree children \((B,m)\) (line 7) of all occurrences have been precomputed in a prior single pass.

The algorithm works as follows: We traverse the set of rules in an anti-SL order. Each rule is traversed in a pre-order run which assures that we combine the digrams in a “top-down greedy” way. If the node \((C,n)\) visited is not a formal parameter and if it is not the root of \( t_C \), we check, whether a digram occurrence can be added for the digram occurrence generator \((C,n)\). There are two cases where a digram occurrence generator can be added: first (line 9), its tree parent \((A,k)\) and its tree child \((B,m)\) have different labels, second, if \( label(t_A, k) = label(t_B, m)\), the node \((C,n)\) is no nonterminal, and \((A,k), (B,m)\) does not overlap with another occurrence of the same digram. Such an overlap can only happen, if the tree parent \((A,k)\) is a tree child of another occurrence of the same digram \( \alpha \), so line 11 checks whether \((A,k)\) was added to the set of occurrences of \( \alpha \) before. The missing case \(-(C,n)\) is a nonterminal and \( label(t_A, k) = label(t_B, m)\) would create an overlapping occurrence crossing the root of a rule, i.e., we do not add this occurrence. This implies that our approach does not produce overlapping digram occurrences, and it does not add occurrences of equal label digrams that cross rule boundaries to the root of a rule. This allows us to directly replace digram occurrences where tree parent node and tree child occur within the same rule without preparing the grammar. However, as shown in the next section, occurrences covering nodes from different rules still need to be treated in a separate way.

Consider again the fragment Grammar 1 of an SLCF grammar. Table I shows the results of applying the inner loop (lines 3-12) of Algorithm RETRIEVEOCCS to the rule \( A \to t_A \) of applying it to the rule \( C \to t_C \). Note that we omitted all results having the empty tree \( \perp \) as tree child. The column \( \Delta occ_G(\alpha) \) contains the digram occurrence generators added to the set of digram occurrence generators. The nodes \((A,1)\) and \((C,1)\) calculated in the first lines of Table I and of Table II do not generate a digram occurrence, as the considered node is the root of each rule. Nodes \((A,2)\) and \((A,8)\) are parameters and therefore do not generate digram occurrences either. The digram occurrence generated by node \((A,6)\) is not added to \( occ_G(\alpha = (a,2,a)) \) as it overlaps with the previously stored occurrence generated by node \((A,3)\). Node \((C,2)\) is an example for a digram occurrence generator where the tree parent \((B,1)\), the tree child \((A,1)\), and the node generating the digram occurrence occur all within different rules.

### Table I. RETRIEVEOCCS FOR RULE A → t_A

<table>
<thead>
<tr>
<th>node</th>
<th>tree parent</th>
<th>tree child</th>
<th>( \alpha )</th>
<th>( \Delta occ_G(\alpha) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>((A,1))</td>
<td>(\perp)</td>
<td>(\perp)</td>
<td>(\perp)</td>
<td>(\perp)</td>
</tr>
<tr>
<td>((A,2))</td>
<td>(\perp)</td>
<td>(\perp)</td>
<td>(\perp)</td>
<td>(\perp)</td>
</tr>
<tr>
<td>((A,3))</td>
<td>((A,1),2)</td>
<td>((A,3))</td>
<td>((a,2,a))</td>
<td>((A,3))</td>
</tr>
<tr>
<td>((A,4))</td>
<td>((A,3),1)</td>
<td>((B,1))</td>
<td>((a,1,b))</td>
<td>((A,4))</td>
</tr>
<tr>
<td>((A,6))</td>
<td>((A,3),2)</td>
<td>((A,6))</td>
<td>((a,2,a))</td>
<td>(\perp)</td>
</tr>
<tr>
<td>((A,8))</td>
<td>(\perp)</td>
<td>(\perp)</td>
<td>(\perp)</td>
<td>(\perp)</td>
</tr>
</tbody>
</table>

### Table II. RETRIEVEOCCS FOR RULE C → t_C

<table>
<thead>
<tr>
<th>node</th>
<th>tree parent</th>
<th>tree child</th>
<th>( \alpha )</th>
<th>( \Delta occ_G(\alpha) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>((C,1))</td>
<td>(\perp)</td>
<td>(\perp)</td>
<td>(\perp)</td>
<td>(\perp)</td>
</tr>
<tr>
<td>((C,2))</td>
<td>((A,1),1)</td>
<td>((B,1))</td>
<td>((a,1,b))</td>
<td>((C,2))</td>
</tr>
</tbody>
</table>

### B. Replacing Digrams

We start again with considering the simpler case of trees. After computing all digrams and their occurrences, a most frequent digram \( \alpha = (a,i,b) \), with \( rank(\alpha) \leq k_{in} \) is selected, where \( k_{in} \) is a predefined constant limiting the maximum rank of digrams that are replaced. For \( \alpha \), TreeRePair introduces a grammar rule \( X \to t_X \) with \( t_X \) representing \( \alpha \).
Let $v.i$ be the subtree rooted in the $i$-th child node of a node $v$. For each occurrence of $\alpha$ consisting of tree parent $v$ and tree child $w$, TreeRePair applies an operation that is inverse to inlining $X$, i.e., it replaces the subtree rooted in $v$ by a subtree $X(v,1,\ldots,v.(i-1),w,1,\ldots,w.rank(b),v.(i+1),\ldots,v.rank(a))$.

In order to replace digrams $\alpha = (a,i,b)$ in arbitrary SLCF tree grammars, first, GrammarRePair generates the same rule $X \rightarrow t_X$ as TreeRePair does. Then, Algorithm 5 or an optimized version of it, Algorithm 6, is used for replacing all occurrences of the digram $\alpha$.

### Algorithm 5: REPLACINGALLOCCURENCESOF($\alpha$)

1. $DD_{\alpha} =$ DEPENDENCYDAG($G = (F, N, P, S)$)
2. foreach $(Q \rightarrow t_Q) \in P$ in anti-SL order do
   3.   foreach node $(Q, n) \in t_Q$ do
      4.     if $(Q, n) \in DD_{\alpha}$ and $(Q, n)$ is a nonterminal $R$ then
      5.        inline $t_R$ for $(Q, n)$
      6.     replace each digram occurrence of $\alpha$ in $t_Q$ by $X$ as done in TreeRePair.

Algorithm 5 consists of two parts:

1. The computation of a so-called DependencyDAG $DD_{\alpha}$ for $\alpha$ (line 1), which contains all the rules that need to be considered when replacing the occurrences of $\alpha$. This DependencyDAG contains all paths of grammar nodes $(Q, n)$ which are nonterminals that have to be expanded when computing the tree parent or the tree child of a digram occurrence generator of $\alpha$ by Algorithm 2 or Algorithm 3.
2. A loop (lines 2-6) proceeding bottom-up through $DD_{\alpha}$ that does the digram replacements (line 6), preceded in some cases by inlining steps (lines 3-5).

The replacement of $\alpha = (a,i,b)$ traverses the grammar $G$ bottom-up (lines 2-6), and for each production rule $Q \rightarrow t_Q$ of $G$, the following steps are executed:

1. For each node $(Q, n)$ that is contained in $DD_{\alpha}$ and that is a nonterminal $R$, we inline $t_R$ for $R$, thereby isolating $a$ and $b$.
2. As now $a$ and $b$ are isolated, and $b$ is the $i$-th child of $a$, we replace the digram occurrence $\alpha = (a,i,b)$ as in TreeRePair.

### C. Updating the Context

The updates of the contexts are mainly the same, no matter whether we work on trees or on grammars — considering that we know for grammars how to compute the tree child and the tree parent of a digram occurrence.

For each occurrence that is replaced, we update the digram occurrences within the neighborhood. Conceptionally, this step results in the same result as repeating the step described in Section IV-A. But in order to avoid unnecessary re-counting, only the occurrences that overlap with an occurrence of the replaced digram $\alpha = (a,i,b)$ have to be adapted. If we consider a digram occurrence of $\alpha$ having tree parent $v$ and tree child $w$, such that $v$ is the $j$-th child node of its parent node $p$, with $label(p) = q$, we have to delete the occurrence consisting of tree parent $p$ and tree child $v$ of digram $(q, j, a)$ and instead add an occurrence of $(q, j, X)$. Similarly, for each child $c$ of $w$ with $label(c) = d$ being the $k$-th child of $w$, we delete the occurrence consisting of tree parent $w$ and tree child $c$ of digram $(b, k, d)$ and instead add an occurrence of $(X,i+k-1,d)$. Finally, the complete digram list of digram $\alpha$ has to be deleted.

The replacement step and the context update are repeated as long as a digram exists that has less than $k_{in}$ parameters and occurs more than once.

### D. Pruning

Finally, the resulting grammar is pruned, i.e., so-called unproductive grammar rules are removed. This step again is similar, no matter whether we work on trees or on arbitrary SLCF tree grammars. The productiveness of a rule $R \rightarrow t_R$ can be calculated by $sav_G(R) = |ref_G(R)| * (size(t_R) - rank(R)) - size(t_R)$, where $size(t_R)$ is the number of edges of tree $t_R$ [3]. If $sav_G(R) < 0$ holds for a rule $R$, this rule is unproductive and it is removed by inlining.

Note that the order of removing unproductive rules matters, as inlining a rule $Q$ into a rule $R$ changes the size of $t_R$, and therefore $sav_G(R)$ changes. We follow the greedy strategy used by TreeRePair. We first remove all rules $Q$ with $|ref_G(Q)| = 1$ and then analyze the grammar in anti-SL order to find unproductive rules.

### E. Optimized Digram Replacement Algorithm

Before we can explain the main optimization idea, i.e., the export of rule fragments to new rules, we present an example motivating that inlining has to consider multiple rule versions.

Consider $\alpha = (a,1,b)$ and the following fragment of a grammar, “Grammar 2”, and assume that $A$ and $C$, but not $B$ are called elsewhere in Grammar 2.

\[
\begin{align*}
\text{C} & \rightarrow A(\bot, A(A(B, \bot), A(B, A(\bot, A(\bot, \bot)))))) \\
\text{A} & \rightarrow b(a(y_1, c(d(a(y_2, \bot), \bot), \bot)), \bot) \\
\text{B} & \rightarrow b(\bot, \bot)
\end{align*}
\]

The nodes $(C,3), (C,4), (C,5), (C,7), (C,8)$, and $(C,9)$ are digram occurrence generators of $\alpha$. In order to replace all digrams $\alpha$, we have to isolate the tree parent $a$ and the tree child $b$ of these digram occurrence generator by inlining.

Inlining $t_B$ into nodes $(C,5)$ and $(C,8)$ is needed for later digram replacement and makes the rule $B \rightarrow t_B$ superfluous.

As the nonterminal $A$ is called in different contexts, i.e., different parts of $t_A$ have to be isolated to isolate $a$ or $b$, we consider different versions of the rule $A \rightarrow t_A$ and inline that version of $t_A$ that matches the demands of the context.

Whenever a node with nonterminal $A$ is a digram occurrence generator (e.g., nodes $(C,3), (C,4), (C,7)$, and $(C,9)$), we have to isolate $t_A$’s root node $(A,1)$.

Similarly, we have to isolate the tree parents: e.g., the node $(C,1)$ has as second parameter a digram occurrence generator, i.e., $val_G((C,1))$ contains a tree parent. Therefore, to isolate...
the tree parent found in volG((C, 1)), we have to isolate node (A, 6) which is the parent of the second parameter within tA before inlining tA for (C, 1).

But this is not the only version of tA that needs to be inlined. Nodes (C, 3) and (C, 7) contain the tree parents of both of their parameters (and are digram occurrence generators themselves), i.e., we need a further version of tA, where we isolate the nodes (A, 1), (A, 2), and (A, 6). A third version of tA is needed for inlining the node (C, 4), which contains the tree parent of its first parameter (and is a digram occurrence generator itself). Finally, a fourth version of tA is needed for inlining into the node (C, 9) that does contain no tree parent, but is a digram occurrence generator itself.

If we want to isolate the root node of tA only, the first-child of tA’s root can be exported into a new rule to minimize the decomposition when inlining tA, i.e., instead of inlining tA for (C, 9), in an optimized version, we create a rule version A → tA and inline tA for (C, 9) with tA_r = b(A’(y_1, y_2), ⊥) and generate a new export rule A’ → (y_1, 1c(d(a(y_2, ⊥), ⊥))).

On the other hand, if we have to isolate the root node and the parent of both parameters, the only fragment that is not needed in order to isolate the tree parent or the tree child is the subtree c(d(y_2, ⊥), ⊥). So here we create a different version of tA and inline tA_{y_1, y_2} with tA_{y_1, y_2} = b(a(y_1, A'_{y_1, y_2}(a(y_2, ⊥))), ⊥) instead and generate a new export rule A’ → (y_1, 1c(d(a(y_2, ⊥), ⊥))).

We create in a similar way versions tA_{y_2} for (C, 1) and tA_{y_2} for (C, 4). Altogether, as we want to keep the grammar small, we need to consider multiple versions, in this case tA_{y_2}, tA_{y_1, y_2}, and tA. Of a rule A → tA for the nonterminal A, depending on in which context A is called in tC.

These different versions of a rule A → tA will be further optimized by exporting fragments to new rules as described at the end of this section.

To avoid a full decomposition, instead of the simple DependencyDAG (line 1 of Algorithm 5), we use the so-called ReplacementDAG RDα to collect and manage the different versions of a rule, e.g. A → tA. In order to compute the ReplacementDAG RDα, we go top-down through the grammar, and for each rule C → tC, we do the following: For each digram occurrence generator (C, n) in tC, we generate two flags: The first is an r flag attached to (C, n), the second is a y flag attached to (C, n)’s parent P in tC, such that (C, n) is the i-th child of P in tC. After having set all flags in tC, for each nonterminal D in tC that has a non-empty set FD of flags attached to it, RDα contains a node DFD and an edge from C to DFD. Repeatedly, for each new node CFC of RDα, we generate a copy CFC → tCFC of C → tC and assign flags to nodes of tCFC as follows. If r ∈ FC and (C, n, 1) is a nonterminal R, we add a flag r to R. For each yi ∈ FC, we find the parent P of yi in tC, such that yi is the k-th parameter of P, and if P is a nonterminal, we assign the flag yk to P. After having set all flags in tCFC, again for each nonterminal node R in tCFC that has a non-empty set FR of flags attached to it, we generate a node RFR in RDα if it does not exist, and we add an edge from CFC to RFR. If we consider Example 2, we obtain the following ReplacementDAG RDα:

**Algorithm 6: REPLACINGALLOCURRENCESOF(α)**

1. RDα = REPLACEMENTDAG(G = (F, N, P, S))
2. foreach node CFC of RDα in bottom-up order do
   3. APPLYLININGSTEPS(CFC)
   4. replace each digram occurrence of α in tQ by X as done in TreeRePair
   5. EXPORTFRAGMENTS TO NEWRULES(tC)
   6. Add exported rules to G

We replace Algorithm 5 by the optimized version, Algorithm 6, that uses the ReplacementDAG RDα instead of the simple DependencyDAG to keep track of all paths to nonterminals that shall be expanded.

**Algorithm 7: APPLYLININGSTEPS(QFQ)**

1. foreach node (Q, n) in occC(α) do
   2. if (Q, n) is a nonterminal R then
      3. inline tR for (Q, n)
   4. if parent(Q, n) is a nonterminal P then
      5. inline tP for parent(Q, n)
   6. if r ∈ FQ then
      7. if root(tQ) is a nonterminal R then
         8. inline tR for root(tQ)
         9. mark root(tQ)
      10. foreach parameter yi in tQ with yi ∈ FQ do
         11. if parent(tQ, yi) is a nonterminal P then
            12. inline tP for parent(tQ, yi)
            13. mark parent(tQ, yi)

The inlining steps of Algorithm 5, e.g., inlining a tree tQ for a nonterminal Q, have the purpose to isolate the a node or the b node of a given digram occurrence α = (a, i, b). However, tQ contains more nodes than the a or the b node to be isolated. In order to distinguish nodes that shall be isolated like the a or b node of a digram occurrence α = (a, i, b) from other nodes of tQ, Algorithm 7 not only applies inlining steps (lines 3, 5, 8, and 12), but also marks the nodes that shall be isolated (lines 9 and 13).

The key idea of the optimization called in line 5 of Algorithm 6 is the following. A tree tQ that is inlined for multiple occurrences of Q may contain a connected fragment f_U of nodes all of which are non-marked, i.e., do not need to be considered for replacing occurrences of α. By inlining tQ multiple times, multiple copies of f_U will occur. To reduce the grammar size, a new rule U → f_U will be inserted, and all occurrences of f_U can be replaced by U.

Assume, that we look for digrams (a, 1, b) and have a rule

```markdown
C
A_{y_1} ← A_{r,y_1,y_2} ← A_{r,y_1} ← A_r
```

```markdown
Algorithm 6: REPLACINGALLOCURRENCESOF(α)
```

1. RDα = REPLACEMENTDAG(G = (F, N, P, S))
2. foreach node CFC of RDα in bottom-up order do
   3. APPLYLININGSTEPS(CFC)
   4. replace each digram occurrence of α in tQ by X as done in TreeRePair
   5. EXPORTFRAGMENTS TO NEWRULES(tC)
   6. Add exported rules to G

We replace Algorithm 5 by the optimized version, Algorithm 6, that uses the ReplacementDAG RDα instead of the simple DependencyDAG to keep track of all paths to nonterminals that shall be expanded.

```markdown
Algorithm 7: APPLYLININGSTEPS(QFQ)
```

1. foreach node (Q, n) in occC(α) do
   2. if (Q, n) is a nonterminal R then
      3. inline tR for (Q, n)
   4. if parent(Q, n) is a nonterminal P then
      5. inline tP for parent(Q, n)
   6. if r ∈ FQ then
      7. if root(tQ) is a nonterminal R then
         8. inline tR for root(tQ)
         9. mark root(tQ)
      10. foreach parameter yi in tQ with yi ∈ FQ do
         11. if parent(tQ, yi) is a nonterminal P then
            12. inline tP for parent(tQ, yi)
            13. mark parent(tQ, yi)

The inlining steps of Algorithm 5, e.g., inlining a tree tQ for a nonterminal Q, have the purpose to isolate the a node or the b node of a given digram occurrence α = (a, i, b). However, tQ contains more nodes than the a or the b node to be isolated. In order to distinguish nodes that shall be isolated like the a or b node of a digram occurrence α = (a, i, b) from other nodes of tQ, Algorithm 7 not only applies inlining steps (lines 3, 5, 8, and 12), but also marks the nodes that shall be isolated (lines 9 and 13).

The key idea of the optimization called in line 5 of Algorithm 6 is the following. A tree tQ that is inlined for multiple occurrences of Q may contain a connected fragment f_U of nodes all of which are non-marked, i.e., do not need to be considered for replacing occurrences of α. By inlining tQ multiple times, multiple copies of f_U will occur. To reduce the grammar size, a new rule U → f_U will be inserted, and all occurrences of f_U can be replaced by U.

Assume, that we look for digrams (a, 1, b) and have a rule
Algorithm 8: ExportFRAGMENTSToNewRules(tQ)
1 if |refQ(G)| > 1 and at least one node of tQ is marked
then
2 Identify in tQ all root nodes rU of a fragment fU
containing multiple connected non-marked
non-parameter nodes.
3 foreach such rU do
4 Create a tree tU, such that tU is a copy of the
subtree of tQ rooted in rU
5 foreach tQ do
6 Substitute each subtree rooted in a marked node
with a new parameter y_j
7 Add a rule R_U → t_U
8 Replace each fragment f_U in t_Q by applying the
compression rule R_U → t_U to f_U
9 Remove all marks from t_Q

If a node is marked within the modified tree t_Q, we know that
Q → t_Q will be inlined when continuing the bottom-up traversal of RDα. However, only marked nodes need to
be inlined or are needed in order to execute replacements
of occurrences of α. Therefore, the optimization algorithm,
Algorithm 8, in line 1, checks whether |refQ(G)| > 1 and at
least one node of t_Q is marked, and if so, decreases the size
of t_Q as follows: Each marked node of t_Q partitions t_Q into
fragments of multiple connected non-marked, non-parameter
nodes (line 2). In the previous example, d and e form such
a fragment that is exported into the rule C → c, where the
parameter y_1 marks the position of the next marked node a of
t_Q. For each such fragment f_U of t_Q that consists of multiple
nodes, we generate a subtree t_U for f_U (lines 3-7), export f_U
into a new rule R_U → t_U (line 7), and replace f_U by applying
this rule (line 8), before we inline Q → t_Q itself.

Intuitively, exporting f_U of t_Q is inverse to inlining R_U →
t_U in t_Q. Exporting a fragment f_U that is rooted in a node v_U
of t_Q means that in t_Q, we replace the subtree t_v_U rooted in v_U
with the tree U(t_1, ..., t_n), such that t_1, ..., t_n are the subtrees
of t_v_U rooted in the top-most nodes of t_v_U in preorder that are
not contained in f_U. Formally, t_Q := t_Q[v_U/U(t_1, ..., t_n)].
Then, t_U is t_v_U[t_1/y_1, ..., t_n/y_n].

Finally, we unmark all marked nodes within the modified,
but possibly smaller tree t_Q (line 9).

F: Concluding Example

In this example, we discuss the optimized replacement of
the digram α = (a, 1, b) by a new rule X → a(b(y_1, y_2), y_3)
on the grammar fragment Grammar 1. We assume that it is
a fragment only, i.e., the rules A, B, and C are called by
further nonterminals anywhere else in the grammar. For this
fragment, the two occurrence generators of α are the nodes
(A, 4) and (C, 2). The ReplacementDAG RD_α consists of the
nodes \{C, A, B\} and the edges \{(C, A), (C, B), (A, B)\}.

We start bottom-up with rule B. The root node (B, 1) is
marked, but there is no fragment to export, so rule B stays
unchanged.

Next, we continue with rule A. First, we inline 
A, 4) yielding the new right-hand side
a(y_1, a(b(⊥, ⊥), a(⊥, y_2))). Now, we can replace
the digram occurrence by nonterminal X yielding
a(y_1, X(⊥, ⊥, a(⊥, y_2))). The root node α is marked, and as
we assume that A is called from further rules, we apply the
optimization and export the fragment X(⊥, ⊥, a(⊥, y_2))
into the new rule D → X(⊥, ⊥, a(⊥, y_2)). This leads to the new
rule A → a(y_1, D(y_2)).

Finally, we continue with rule C. We start with inlining
the right-hand sides of rules A and B yielding the new right-
hand side a(b(⊥, ⊥), D(⊥)). Now, we replace the digram
occurrence by X yielding C → X(⊥, ⊥, D(⊥)).

This leads to the new grammar version

C → X(⊥, ⊥, D(⊥))
D → X(⊥, ⊥, a(⊥, y_2))
X → a(b(y_1, y_2), y_3)

V. Experiments

We present an experimental evaluation of our prototype
implementation of GrammarRePair. There are two groups
of experiments: (1) Static Compression and (2) Dynamic
Compression. In the static part, we gauge the compression,
and memory behavior of GrammarRePair; these experiments
are over static XML document trees, without any updates
performed. In the dynamic part, we investigate the compression
and runtime behavior of GrammarRePair for XML document
trees under updates.

A. Setup

Machine. All tests are performed on an Intel Core2 Duo
CPU P8400 @3GHz with 4GB of RAM running the 64 bit
version of Linux 3.11.10-25. The tests of our prototype are
performed using Java 1.9 (64 bit) with 3400MB heap space
and a thread stack size of 4MB. The TreeRePair implementation
is Version 20100428\(^1\) and was compiled using g++ 4.8.1. All
runtimes are measured using the user runtime given through
the linux command ‘time’. When measuring running times,
we always perform four consecutive runs and then report the
average over those runs.

Datasets. We execute over XML files that consist of
element nodes only; thus, there is no text content of attributes
(or other features of XML such as comments or processing
instructions). These files have been obtained from well known
benchmark files, by stripping off all non-element content.
Almost all of our files can be found on the XMLComp-
Bench [16] site, which provides the original XML documents,
the structure-only documents (as used in the experiments),
\(^1\)Available at https://code.google.com/p/\!/trp/downloads/detail\?name=
TreeRePair_20100428.zip
and the references to where these files originated. The two additional files used are Medline\textsuperscript{2} and NCBI\textsuperscript{3}. Table III lists all the files used in our tests, together with some document statistics: the number of edges and the depth (denoted “dp”) of the original document tree. The number c-edges refers to the number of edges in the grammar obtained by GrammarRePair (calculated as defined in Section II); the last column shows the corresponding compression ratio of c-edges divided by \#edges.

<table>
<thead>
<tr>
<th>dataset</th>
<th>#edges</th>
<th>dp</th>
<th>c-edges</th>
<th>compression ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXI-Weblog</td>
<td>93434</td>
<td>2</td>
<td>42</td>
<td>0.04</td>
</tr>
<tr>
<td>XMark</td>
<td>167864</td>
<td>11</td>
<td>22105</td>
<td>13.17</td>
</tr>
<tr>
<td>EXI-Telecomp</td>
<td>177633</td>
<td>6</td>
<td>107</td>
<td>0.06</td>
</tr>
<tr>
<td>Treebank</td>
<td>2437665</td>
<td>35</td>
<td>503830</td>
<td>20.67</td>
</tr>
<tr>
<td>Medline</td>
<td>2866079</td>
<td>6</td>
<td>118067</td>
<td>4.12</td>
</tr>
<tr>
<td>NCBI</td>
<td>3642224</td>
<td>3</td>
<td>59</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

TABLE III. DOCUMENT STATISTICS AND GRAMMARREPAIR-COMPRESSION RESULTS

B. Static Compression

Recall that GrammarRePair takes an SLT grammar $G$ as input, and produces as output a new SLT grammar obtained by running RePair-compression over the tree represented by the grammar $G$. Thus, since the original grammar $G$ may consist of a single rule with a large tree as right-hand side, GrammarRePair can be seen as a tree compressor itself. It is therefore reasonable to compare the tree compression behavior of GrammarRePair with existing compressors. This is done in two parts: first the compression ratios, then the running times are compared. In a third part, we investigate the memory consumption of GrammarRePair. This is measured in terms of the sizes of the intermediate grammars.

Compression Ratio Comparison. We compare the compression ratios obtained by TreeRePair, and our new GrammarRePair applied to trees as well as to grammars. In nearly all cases, our algorithm GrammarRePair produces similar or better results than TreeRePair. Better results are obtained on documents with very strong compression, viz. very small ratios; for instance, on EXI-Weblog (which compresses to only 0.1\% by TreeRePair), our compressor GrammarRePair is better than TreeRePair by almost 20\%. But in general, all three approaches compress as good as the others. That is, there is hardly a difference in the absolute compression ratio reached by TreeRePair, or our new GrammarRePair applied to trees as well as to grammars.

Memory Consumption. During compression, GrammarRePair produces a number of intermediate grammars, each of them being the outcome of a digram replacement. We measure the memory consumption in terms of the sizes of these intermediate grammars. We call the fraction (max. size of intermediate grammar) divided by (size of final grammar) the compression ratio, and the compression ratio at maximum blow-up are shown below the name. The worst blow-up is just over 2, while for many files, it is only around a few percent above 1. This is easily explained: files with extreme compression ratios such as NCBI and EXI-Weblog contain long lists that are compressed exponentially. Breaking such a list “open” during the run of GrammarRePair can quickly double the size of the intermediate grammar. However, that intermediate grammar is still minuscule compared to the original tree; this is seen by comparing the compression ratios of the grammars against the original tree. As we can see, for NCBI and EXI-Weblog, these values coincide, i.e., the difference in intermediate grammar size can hardly be measured when compared against the original tree size. The biggest difference here is on XMark, where the intermediate grammar has around 12\% compression, while the compression by GrammarRePair is around 8\%.

Effect of the Optimization. In this experiment we compare the optimized version – i.e., exporting fragments $t_U$ not needed in order to replace the digram into new rules $R_U \rightarrow t_U$ – to the non-optimized version.

For simplicity, we show string grammars here (to obtain a similar tree grammar, the reader may consider one additional root symbol, under which these grammars generate long children lists). For a natural number $n$, we consider the grammar $G_n$ with these rules:

$S \rightarrow aA_n a b$

$A_i \rightarrow A_{i-1} A_{i-1}$, for $1 \leq i \leq n$

$A_0 \rightarrow b a.$

The grammar $G_n$ produces a string consisting of a list of $n+1$ pairs of siblings $a$, $b$. This grammar is recompressed to the grammar with these rules:

$S \rightarrow B_0 B_i B_1$

$B_i \rightarrow B_{i-1} B_{i-1}$, for $1 \leq i \leq n$

$B_0 \rightarrow a b.$

We choose values for $n$ in the range from 64 to 4096. As this example compresses exponentially, i.e., the size of the tree $val(G_n)$ is exponential in the size of the compressed grammars, the obtained grammars have at most 44 edges.

Figure 3 shows the results of our experiments. As we can see, if we use the optimization of Algorithm 8, the size of the grammar as well as the runtime seems to stay linear in the size of the compressed grammar. As expected, we reach a blow-up of less than 2 (the green curve varies from 1.2 $\rightarrow$ 1.7). Due to the exponential compression of our files, we have several...
input XML documents of different sizes and thus several data points for each resulting grammar size.

If we deactivate the optimization, the blow-up seems to grow rather in the size of the original XML tree. For our sample data, we reach a blow-up of more than 110. This can also be observed for the runtime, as the runtime of the non-optimized version scales much worse than the runtime of the optimized version. Only for very small grammars, the runtime overhead caused by the optimization is greater than the yielded benefit, resulting in runtimes including the optimization that are slightly worse than the runtimes without the optimization.

C. Dynamic Compression

In this section, we investigate the use of GrammarRePair as a tool generating a mutable compressed tree data structure. For this purpose, we perform updates on the grammar-compressed tree and then execute GrammarRePair on the resulting grammar. We compare this grammar with the one obtained by first performing the update, then decompressing the grammar, and then compressing the obtained tree.

We consider three update operations: (1) to rename the label of a given node, (2) to insert a tree before a given node (i.e., as previous sibling of that node), and (3) to delete the subtree rooted at a given node.

Since we work on binary tree representations of the original unranked XML trees, we can execute an insert on a null pointer, thus executing an “insert after” the last element (or an insert into an empty child sequence).

More formally, if \( t \) and \( s \) are binary trees representing XML structures, \( u \) is a node of \( t \), and \( v \) is the right-most leaf of \( s \) (which is necessarily a null pointer), then insert\((t,u,s)\) is defined as \( t[u/s] \) if \( u \) is a null pointer node, and otherwise as \( t[u/s'] \) where \( s' = s[v/tu] \) and \( tu \) is the subtree of \( t \) rooted at \( u \).

**Compression Ratio Comparison.** Here, we consider sequences of random insert and delete operations (10% deletes and 90% inserts). The sequences are obtained by starting from a given document, and then applying the inverse of the operations until a seed document is derived. In this way, each update sequence starts with a seed document and ends up with an original document from our datasets. This is a well-known technique for approximating realistic update workloads.

As was observed previously, repeated updates to a grammar compressed tree causes rapid degradation of the compression ratio [6]. The reason is that the path isolation process causes decompression of the grammar. We show this effect on a subset of six files: three with very strong compression (NCBI (NC), EXI-Telecomp (ET), and EXI-Weblog (EW)) and three with moderate compression (XMark (XM), Treebank (TB), and Medline (MD)). The results are shown in the top plots of Figures 4 and 5. In a second measurement, we perform the recompression from scratch, i.e., decompression to XML and the compression of the XML to remove the overhead caused by the update after every 100 updates. We compute the overhead as (size of grammar after update) divided by (size of grammar after recompression from scratch). For the files with moderate compression, the overhead caused by updates is up to 0.4, while for the files with very strong compression, there is a blow-up of up to 400. The latter is because exponentially compressed lists are broken down. The “spikes” in Figure 5 are caused by the decrease of the grammar size after each recompression from scratch.

In contrast, we now run GrammarRePair on the grammar obtained after every 100 updates. Again we compute the overhead as (size of grammar after recompression by GrammarRePair) divided by (size of grammar after recompression from scratch). The results are shown in the bottom plots of Figures 4 and 5. For the three moderate files, we obtain an overhead of less than 0.008. On the extreme files, we obtain a maximal overhead of around 5; recall that the latter grammars are still minuscule compared to the original. Thus, the result shows that GrammarRePair can perform incremental updates and obtains results that are comparable to update-decompression.

**Runtime Comparison.** In this experiment, we randomly rename 300 nodes of the document to fresh labels (not used in the document). The resulting documents are recompressed using either decompression followed by compression of TreeRePair (gray line), decompression followed by compression of
GrammarRePair (green boxes) or the recompression of GrammarRePair (red box). Figure 6 shows the results. The number of edges within each XML file and the runtime of 1 are shown below the name. Note that we consider the minimum decomposition runtime of TreeRePair and GrammarRePair as decomposition time. Although the sequence of decompression and compression of TreeRePair performs better for the smallest file in our tests, it is outperformed by the recompression of GrammarRePair for the larger files. The recompression of GrammarRePair outperforms the sequence of decompression and compression of GrammarRePair for all files. For the largest files, already the compression time of TreeRePair is worse than the recompression time of GrammarRePair. When both applied to trees, TreeRePair only takes 0.1 – 0.4 times the time needed by GrammarRePair. This it not astonishing: GrammarRePair is a prototype implemented in Java operating on complicated data structures (as it takes grammars as input), while TreeRePair is a C implementation and optimized for tree input.

In terms of space (not shown), we observe that in the worst case, GrammarRePair uses 23% of the space needed by update-decompress-compress (udc). On average, the space needed by GrammarRePair is much less: it uses only 6% of udc. Thus, GrammarRePair is not only faster, but is also much more space efficient than udc.

VI. Conclusions

Straight-line context-free (SLCF) tree grammars offer a compact in-memory representation of XML document trees. Many operations can be supported efficiently over the representation, such as traversals [2, 4], XPath queries [4, 17], equivalence [2], and unification and matching [18]. One major drawback of SLCF tree grammars has been the lack of efficient updates. One way to tackle this problem is to recompress the grammar after an update. In this paper, we present an algorithm that recompresses a given SLCF tree grammar, using the RePair compression scheme. It works surprisingly well in practice: The resulting grammars are as small as compressing from scratch from an uncompressed document. A possible application are compressed DOM representations with updates, as DOM is a memory-hungry ingredient of any web browser.

REFERENCES