We define a compilation scheme for a constructor-based, strongly-sequential, graph rewriting system which shortcuts some needed steps. The object code is another constructor-based graph rewriting system. This system is normalizing for the original system when using an innermost strategy. Consequently, the object code can be easily implemented by eager functions in a variety of programming languages. We modify this object code in a way that avoids total or partial construction of the contracta of some needed steps of a computation. When computing normal forms in this way, both memory consumption and execution time are reduced compared to ordinary rewriting computations in the original system.

1 Introduction

Rewrite systems are models of computations that specify the actions, but not the control. The object of a computation is a graph referred to as an expression. The actions are encoded by rules that define how to replace (rewrite) one expression with another. The goal of a computation is to reach an expression, called a normal form, that cannot be further rewritten.

In the computation of an expression, a rewrite system does not tell which subexpression should be replaced to reach the goal.

Example 1. Consider the following rewrite system. The syntax is Curry [17].

\[
\begin{align*}
\text{loop} & = \text{loop} \\
\text{snd} (\_, y) & = y
\end{align*}
\]

(1)

A computation of \(\text{snd} (\text{loop}, 0)\) terminates with 0 if the second rule of (1) is ever applied, but goes on forever without making any progress if only the first rule is applied.

In a computation a strategy is a policy or algorithm that defines both which subexpression should be replaced and its replacement. The intended goal of a strategy is to efficiently produce a normal form of an expression when it exists. A practical strategy, called needed, is known for the class of the strongly sequential term rewriting systems [18]. This strategy relies on the fact that, in every reducible expression \(e\), there exists a redex, also called needed, that is reduced in any computation of \(e\) to a normal form.

The needed strategy is defined and implemented as follows: given an expression \(e\), while \(e\) is reducible, reduce an arbitrarily chosen, needed redex of \(e\). In the systems considered in this paper, finding a needed redex is easy without look-ahead [3]. This strategy is normalizing: if an expression \(e\) has a normal form, repeatedly reducing arbitrary needed redexes will terminate with that normal form. This strategy is also optimal in the number of reduced redexes for graph (not term) rewriting.

The above outline shows that implementing a needed strategy is a relatively straightforward task. Surprisingly, however, it is possible to shortcut some of the needed steps in the computation. This paper shows how this shortcutting can be introduced into an implementation of a needed strategy.
Terminology and background information are recalled in Sect. 2. The compilation scheme and its properties are in Sect. 3 and 4. The transformation that allows shortcutting needed redexes, and its properties, are in Sect. 5 and 6. Sect. 7 presents two benchmarks and sketches further opportunities to shortcut needed steps. Sect. 8 discusses the application of our work to the implementation of functional logic languages. Related work and conclusion are in Sect. 9 and 10 respectively.

2 Preliminaries

A rewrite system is a pair \((\Sigma \cup \mathcal{X}, \mathcal{R})\) in which \(\Sigma = \mathcal{C} \cup \mathcal{D}\) is a signature partitioned into constructors and operations (or functions), \(\mathcal{X}\) is a denumerable set of variables, and \(\mathcal{R}\) is a set of rewrite rules defined below. Without further mention, we assume that the symbols of the signature have a type, and that any expression over the signature is well typed.

An expression is a single-rooted, directed, acyclic graph defined in the customary way [11, Def. 2]. An expression \(e\) is a constructor form (or value) if, and only if, every node of \(e\) is labeled by a constructor symbol. Constructor forms are normal forms, but not vice versa. For example, \(\text{head}([[]])\) where \(\text{head}\) is the usual operation that returns the first element of a (non-empty) list, is a normal form, but not a constructor form. In a constructor-based system, such expressions are regarded as failures or exceptions rather than results of computations. Likewise, a head-constructor form is an expression whose root node is labeled by a constructor symbol.

A rule is a graph with two roots abstracting the left- and right-hand sides respectively. The rules follow the constructor discipline [23]. Each rule’s left-hand side is a pattern, i.e., an operation symbol applied to zero or more expressions consisting of only constructor symbols and variables. Rules are left-linear, i.e., the left-hand side is a tree.

The objects of a computation are graphs rather than terms. Sharing some subexpressions of an expression is a requirement of functional logic programming [13, 14, 15, 21]. Incidentally, this sharing ensures that needed redexes are never duplicated during a computation. The difference between our graphs and ordinary terms concerns only the sharing of subexpressions.

A computation of an expression \(t\) is a possibly infinite sequence

\[ t = t_0 \rightarrow t_1 \rightarrow \ldots \]

such that \(t_i \rightarrow t_{i+1}\) is a rewrite step [11, Def. 23]. For all \(i\), \(t_i\) is a state of the computation of \(t\).

Given a rewrite system \(R\), an expression of \(R\) is an expression over the signature of \(R\). When \(s\) is a signature symbol and \(n\) is a natural number, \(s/n\) denotes that \(n\) is the arity of \(s\). When \(t\) and \(u\) are expressions and \(v\) is a variable, \([u/v]\) is the substitution that maps \(v\) to \(u\), and \(t[u/v]\) is the application of \([u/v]\) to \(t\). The reflexive closure of the rewrite relation \(\rightarrow\) is denoted \(\rightarrow^*\).

Each operation in \(\mathcal{D}\) is inductively sequential; that is, its rewrite rules are organized into a hierarchical structure called a definitional tree [1] which we informally recall below. An example of a definitional tree is shown in (3). In a definitional tree of an operation \(f\), there are up to 3 kinds of nodes called branch, rule and exempt. Each kind contains a pattern of \(f\) and other items of information depending on the kind. A rule node with pattern \(\pi\) contains a rule of \(f\) whose left-hand side is equal to \(\pi\) modulo renaming nodes and variables. An exempt node with pattern \(\pi\) contains no other information. There is no rule of \(\bar{f}\) whose left-hand side is equal to \(\pi\). A branch node with a pattern \(\pi\) contains children that are subtrees of the definitional tree. At least one child is a rule node. The children are obtained by “narrowing” pattern \(\pi\).

Let \(x\) be any variable of \(\pi\), which is called inductive. For each constructor \(c/m\) of the type of \(x\), there is a child whose pattern is obtained from \(\pi\) by instantiating \(x\) with \(c(x_1, \ldots x_m)\), where \(x_i\) is a fresh variable.
An operation \( f/n \) is \textit{inductively sequential} \[1\] if there exists a definitional tree whose root has pattern \( f(x_1, \ldots, x_n) \), where \( x_j \) is a fresh variable, and whose leaves contain all, and only, the rules of \( f \). A rewrite system is \textit{inductively sequential} if all of its operations are inductively sequential.

Inductively sequential operations can be thought of as “well designed” with respect to evaluation. To compute a needed redex of an expression \( e \) rooted by an operation \( f \), match to \( e \) the pattern \( \pi \) of a maximal (deepest in the tree) node \( N \) of a definitional tree of \( f \). If \( N \) is an exempt node, \( e \) has no constructor normal form, and the computation can be aborted. If \( N \) is a rule node, \( e \) is a redex and can be reduced by the rule in \( N \). If \( N \) is a branch node, let \( x \) be the inductive variable of \( \pi \) and \( t \) the subexpression of \( e \) to which \( x \) is matched. Then, recursively compute a needed redex of \( t \).

The inductively sequential systems are the intersection \[16\] of the strongly sequential systems \[18\] and the constructor-based systems \[23\]. The following notion \[8\] for inductively sequential systems is key to our work. We abuse the word “needed” because we will show that our notion extends the classic one \[18\]. Our notion is a binary relation on nodes, or equivalently on the subexpressions rooted by these nodes, since they are in a bijection.

\textbf{Definition 1.} Let \( R \) be an inductively sequential system, \( e \) an expression of \( R \) rooted by a node \( p \), and \( n \) a node of \( e \). Node \( n \) is \textit{needed for} \( e \), and similarly is \textit{needed for} \( p \), if, and only if, in any computation of \( e \) to a head-constructor form, the subexpression of \( e \) at \( n \) is derived to a head-constructor form. A node \( n \) (and the redex rooted by \( n \), if any) of a state \( e \) of a computation in \( R \) is \textit{needed} if, and only if, it is needed for some outermost operation-rooted subexpression of \( e \).

Our “needed” relation is interesting only when both nodes are labeled by operation symbols. If \( e \) is an expression whose root node \( p \) is labeled by an operation symbol, then \( p \) is trivially needed for \( p \). This holds whether or not \( e \) is a redex and even when \( e \) is already a normal form, e.g., \texttt{head}([]). In particular, \textit{any} expression that is not a value has pairs of nodes in the needed relation. Finally, our definition is concerned with reaching a \textit{head-constructor} form, not a \textit{normal form}.

Our notion of need generalizes the classic notion \[18\]. Also, since our systems follow the constructor discipline \[23\] we are not interested in expressions that do not have a value.

\textbf{Lemma 1.} Let \( R \) be an inductively sequential system and \( e \) an expression of \( R \) derivable to a value. If \( e' \) is an outermost operation-rooted subexpression of \( e \), and \( n \) is both a node needed for \( e' \) and the root of a redex \( r \), then \( r \) is a needed redex of \( e \) in the sense of \[18\].

\textit{Proof.} Since \( e' \) is an outermost operation-rooted subexpression of \( e \), any node in any path from the root of \( e \) to the root of \( e' \), except for the root of \( e' \), is labeled by constructor symbols. Hence, \( e \) can be derived to a value only if \( e' \) is derived to a value and \( e' \) can be derived to a value only if \( e' \) is derived to a head-constructor form. By assumption, in any derivation of \( e' \) to a head-constructor form \( r \) is derived to a head-constructor form, hence it is reduced. Thus, \( r \) is a needed redex of \( e \) according to \[18\]. \( \square \)

\textbf{Lemma 2.} Let \( R \) be an inductively sequential system, \( e \) an expression of \( R \), \( e_1 \), \( e_2 \) and \( e_3 \) subexpressions of \( e \) such that \( n_1 \) is the root of \( e_1 \) and the label of \( n_1 \) is an operation, for \( i = 1, 2, 3 \). If \( n_3 \) is needed for \( n_2 \) and \( n_2 \) is needed for \( n_1 \), then \( n_3 \) is needed for \( n_1 \).

\textit{Proof.} By hypothesis, if \( e_3 \) is not derived to a constructor-rooted form, \( e_2 \) cannot be derived to a constructor-rooted form, and if \( e_2 \) is not derived to a constructor-rooted form, \( e_1 \) cannot be derived to a constructor-rooted form. Thus, if \( e_3 \) is not derived to a constructor-rooted form, \( e_1 \) cannot be derived to a constructor-rooted form. \( \square \)
3 Compilation

For simplicity and abstraction, we present the object code, $C_R$, of $R$ as a constructor-based graph rewriting system as well. $C_R$ has only two operations called head and norm, and denoted $H$ and $N$, respectively. The constructor symbols of $C_R$ are all, and only, the symbols of $R$. The rules of $C_R$ have a priority established by the textual order. A rule reduces an expression $t$ only if no other preceding rule could be applied to reduce $t$. These semantics are subtle, since $t$ could become reducible by the preceding rule only after some internal reduction. However, all our claims about computations in $C_R$ are stated for an innermost strategy. In this case, when a rule is applied, no internal reduction is possible, and the semantics of the priority are straightforward.

Operation $H$ is defined piecemeal for each operation of $R$. Each operation of $R$ contributes a number of rules dispatched by pattern matching with textual priority. The rules of $H$ contributed by an operation with definitional tree $T$ are generated by the procedure compile defined in Fig. 1. The intent of $H$ is to take an expression of $R$ rooted by an operation and derive an expression of $R$ rooted by a constructor by performing only needed steps.

The expression “$\{x\}$” embedded in a string, denotes interpolation as in modern programming languages, i.e., the argument $x$ is replaced by a string representation of its value. The notation $t[u]_p$ stands for an expression equal to $t$, in which the subexpression identified by $p$ is replaced by $u$. In procedure compile, the notation is used to “wrap” an application of $H$ around the subexpression of the pattern at $o$, the inductive node. An example is the last rule of (4). The loop at statement 12 is for collapsing rules, i.e., rules whose right-hand side is a variable. When this variable matches an expression rooted by a constructor of $R$, no further application of $H$ is required; Otherwise, $H$ is applied to the contractum. Symbol “abort” is not considered an element of the signature of $C_R$. If any redex is reduced to “abort”,

```plaintext
compile $T$
  case $T$ of  
  when branch($\pi, o, T$) then
    $\forall T_i \in T$ compile $T_i$
    output “$H(\{\pi\}) = H(\{\pi[H(\{[\pi\})], o\})$”
  when rule($\pi, l \rightarrow r$) then
    case $r$ of
      when operation-rooted then
        output “$H(\{l\}) = H(\{r\})$”
      when constructor-rooted then
        output “$H(\{l\}) = \{r\}$”
      when variable then
        for each constructor $c/n$ of the sort of $r$
          let $l' \rightarrow r' = (l \rightarrow r)[c(x_1, \ldots, x_n)/r]
          output “$H(\{l\}) = \{r\}$”
          output “$H(\{l\}) = H(\{r\})$”
      when exempt($\pi$) then
        output “$H(\{\pi\}) = \text{abort}$”
```
the computation is aborted since it can be proved that the expression object of the computation has no constructor normal form.

**Example 2.** Consider the rules defining the operation that concatenates two lists, denoted by the infix identifier “++”:

\[
\begin{align*}
  []++y &= y \\
  (x:xs)++y &= x:(xs++y)
\end{align*}
\]  

The definitional tree of operation “++” is pictorially represented below. The only branch node of this tree is the root. The inductive variable of this branch, boxed in the representation, is \( x \). The rule nodes of this tree are the two leaves. There are no exempt nodes in this tree since operation “++” is completely defined.

\[
\begin{align*}
  \text{Example 2.} \\
  \text{Consider the rules defining the operation that concatenates two lists, denoted by the infix identifier “++”:} \\
  \begin{align*}
  []++y &= y \\
  (x:xs)++y &= x:(xs++y)
  \end{align*}
\]  

The definitional tree of operation “++” is pictorially represented below. The only branch node of this tree is the root. The inductive variable of this branch, boxed in the representation, is \( x \). The rule nodes of this tree are the two leaves. There are no exempt nodes in this tree since operation “++” is completely defined.

\[
\begin{align*}
  \text{Example 2.} \\
  \text{Consider the rules defining the operation that concatenates two lists, denoted by the infix identifier “++”:} \\
  \begin{align*}
  []++y &= y \\
  (x:xs)++y &= x:(xs++y)
  \end{align*}
\]  

The definitional tree of operation “++” is pictorially represented below. The only branch node of this tree is the root. The inductive variable of this branch, boxed in the representation, is \( x \). The rule nodes of this tree are the two leaves. There are no exempt nodes in this tree since operation “++” is completely defined.

\[
\begin{align*}
  \text{Example 2.} \\
  \text{Consider the rules defining the operation that concatenates two lists, denoted by the infix identifier “++”:} \\
  \begin{align*}
  []++y &= y \\
  (x:xs)++y &= x:(xs++y)
  \end{align*}
\]  

The definitional tree of operation “++” is pictorially represented below. The only branch node of this tree is the root. The inductive variable of this branch, boxed in the representation, is \( x \). The rule nodes of this tree are the two leaves. There are no exempt nodes in this tree since operation “++” is completely defined.

\[
\begin{align*}
  \text{Example 2.} \\
  \text{Consider the rules defining the operation that concatenates two lists, denoted by the infix identifier “++”:} \\
  \begin{align*}
  []++y &= y \\
  (x:xs)++y &= x:(xs++y)
  \end{align*}
\]  

The definitional tree of operation “++” is pictorially represented below. The only branch node of this tree is the root. The inductive variable of this branch, boxed in the representation, is \( x \). The rule nodes of this tree are the two leaves. There are no exempt nodes in this tree since operation “++” is completely defined.

\[
\begin{align*}
  \text{Example 2.} \\
  \text{Consider the rules defining the operation that concatenates two lists, denoted by the infix identifier “++”:} \\
  \begin{align*}
  []++y &= y \\
  (x:xs)++y &= x:(xs++y)
  \end{align*}
\]  

The definitional tree of operation “++” is pictorially represented below. The only branch node of this tree is the root. The inductive variable of this branch, boxed in the representation, is \( x \). The rule nodes of this tree are the two leaves. There are no exempt nodes in this tree since operation “++” is completely defined.

\[
\begin{align*}
  \text{Example 2.} \\
  \text{Consider the rules defining the operation that concatenates two lists, denoted by the infix identifier “++”:} \\
  \begin{align*}
  []++y &= y \\
  (x:xs)++y &= x:(xs++y)
  \end{align*}
\]  

The definitional tree of operation “++” is pictorially represented below. The only branch node of this tree is the root. The inductive variable of this branch, boxed in the representation, is \( x \). The rule nodes of this tree are the two leaves. There are no exempt nodes in this tree since operation “++” is completely defined.

\[
\begin{align*}
  \text{Example 2.} \\
  \text{Consider the rules defining the operation that concatenates two lists, denoted by the infix identifier “++”:} \\
  \begin{align*}
  []++y &= y \\
  (x:xs)++y &= x:(xs++y)
  \end{align*}
\]  

The definitional tree of operation “++” is pictorially represented below. The only branch node of this tree is the root. The inductive variable of this branch, boxed in the representation, is \( x \). The rule nodes of this tree are the two leaves. There are no exempt nodes in this tree since operation “++” is completely defined.

\[
\begin{align*}
  \text{Example 2.} \\
  \text{Consider the rules defining the operation that concatenates two lists, denoted by the infix identifier “++”:} \\
  \begin{align*}
  []++y &= y \\
  (x:xs)++y &= x:(xs++y)
  \end{align*}
\]  

The definitional tree of operation “++” is pictorially represented below. The only branch node of this tree is the root. The inductive variable of this branch, boxed in the representation, is \( x \). The rule nodes of this tree are the two leaves. There are no exempt nodes in this tree since operation “++” is completely defined.

\[
\begin{align*}
  \text{Example 2.} \\
  \text{Consider the rules defining the operation that concatenates two lists, denoted by the infix identifier “++”:} \\
  \begin{align*}
  []++y &= y \\
  (x:xs)++y &= x:(xs++y)
  \end{align*}
\]  

The definitional tree of operation “++” is pictorially represented below. The only branch node of this tree is the root. The inductive variable of this branch, boxed in the representation, is \( x \). The rule nodes of this tree are the two leaves. There are no exempt nodes in this tree since operation “++” is completely defined.

\[
\begin{align*}
  \text{Example 2.} \\
  \text{Consider the rules defining the operation that concatenates two lists, denoted by the infix identifier “++”:} \\
  \begin{align*}
  []++y &= y \\
  (x:xs)++y &= x:(xs++y)
  \end{align*}
\]  

The definitional tree of operation “++” is pictorially represented below. The only branch node of this tree is the root. The inductive variable of this branch, boxed in the representation, is \( x \). The rule nodes of this tree are the two leaves. There are no exempt nodes in this tree since operation “++” is completely defined.

\[
\begin{align*}
  \text{Example 2.} \\
  \text{Consider the rules defining the operation that concatenates two lists, denoted by the infix identifier “++”:} \\
  \begin{align*}
  []++y &= y \\
  (x:xs)++y &= x:(xs++y)
  \end{align*}
\]  

The definitional tree of operation “++” is pictorially represented below. The only branch node of this tree is the root. The inductive variable of this branch, boxed in the representation, is \( x \). The rule nodes of this tree are the two leaves. There are no exempt nodes in this tree since operation “++” is completely defined.

\[
\begin{align*}
  \text{Example 2.} \\
  \text{Consider the rules defining the operation that concatenates two lists, denoted by the infix identifier “++”:} \\
  \begin{align*}
  []++y &= y \\
  (x:xs)++y &= x:(xs++y)
  \end{align*}
\]  

The definitional tree of operation “++” is pictorially represented below. The only branch node of this tree is the root. The inductive variable of this branch, boxed in the representation, is \( x \). The rule nodes of this tree are the two leaves. There are no exempt nodes in this tree since operation “++” is completely defined.
\[
\begin{align*}
N((1:[])++(2:[])) &\rightarrow N(H((1:[])++(2:[]))) \\
&\rightarrow N(1:([]++(2:[]))) \\
&\rightarrow 1:N(H(1:([]++(2:[])))) \\
&\rightarrow 1:N(2:') \\
&\rightarrow 1:N(2):N(1) \\
&\rightarrow 1:2:[]
\end{align*}
\] 

Computation (8) is longer than (7). If all the occurrences of \(N\) and \(H\) are "erased" from the states of (8), a concept formalized shortly, and repeated states of the computation are removed, the remaining steps are the same as in (7). The introduction and removal of occurrences of \(N\) and \(H\) in (8), which lengthen the computation, represent the control, what to rewrite and when to stop. These activities occur in (7) too, but are in the mind of the reader rather than explicitly represented in the computation.

4 Compilation Properties

\(C_R\), the object code of \(R\), correctly implements \(R\). Computations performed by \(C_R\) produce the results of corresponding computations in \(R\) as formalized below. Furthermore, \(C_R\) implements a needed strategy, because every reduction performed by \(C_R\) is a needed reduction in \(R\). In this section, we prove these properties of the object code.

Let \(Expr\) be the set of expressions over the signature of \(C_R\) output by \texttt{compile} on input a rewrite system \(R\). The erasure function \(E: Expr \rightarrow Expr\) is inductively defined by:

\[
\begin{align*}
E(H(t)) &= E(t) \\
E(N(t)) &= E(t) \\
E(s(t_1,\ldots,t_n)) &= s(E(t_1),\ldots,E(t_n)) \quad \text{for } s/n \in \Sigma_R
\end{align*}
\] 

Intuitively, the erasure of an expression \(t\) removes all the occurrences of \(H\) and \(N\) from \(t\). The result is an expression over the signature of \(R\).

\textbf{Lemma 3.} Let \(R\) be an inductively sequential system and \(H\) the head function of \(C_R\). For any operation-rooted expression \(t\) of \(R\), \(H(t)\) is a redex.

\textit{Proof.} Let \(f/n\) be the root of \(t\), and \(\mathcal{T}\) the definitional tree of \(f\) input to procedure \texttt{compile}. The pattern at the root of \(\mathcal{T}\) is \(f(x_1,\ldots,x_n)\), where each \(x_i\) is a variable. Procedure \texttt{compile} outputs a rule of \(H\) with left-hand side \(H(f(x_1,\ldots,x_n))\). Hence this rule, or a more specific one, reduces \(t\). \(\square\)

Comparing graphs modulo a renaming of nodes, as in the next proof, is a standard technique \[\text{due to the fact that any node created by a rewrite is fresh.}\]

\textbf{Lemma 4.} Let \(R\) be an inductively sequential system and \(H\) the head function of \(C_R\). Let \(t\) be an operation-rooted expression of \(R\), and \(H(t)\) be reduced by a step resulting from the application of a rule \(r\) originating from statement 04 of procedure \texttt{compile}. The argument of the inner application of \(H\) in the contractum is both operation-rooted and needed for \(t\).

\textit{Proof.} Let \(\mathcal{T}\) be a definitional tree of the root of \(t\). Let \(\pi\) be the pattern of the \textit{branch} node \(n\) of \(\mathcal{T}\) from which rule \(r\) originates and let \(o\) be the inductive node of \(\pi\). Since \(r\) rewrites \(t\) and \(\pi\) is the left-hand side of \(r\) modulo a renaming of variables and nodes, there exists a graph homomorphism \(\sigma\) such that \(t = \sigma(\pi)\). Our convention on the specificity of the rules defining \(H\) establishes that no rule textually
preceding \( r \) in the definition of \( \mathbf{H} \) rewrites \( t \). Since procedure \texttt{compile} traverses \( \mathcal{T} \) in post-order, every rule of \( \mathbf{H} \) originating from a node descendant of \( n \) in \( \mathcal{T} \) textually precedes \( r \) in the definition of \( \mathbf{H} \). Let \( q = \sigma(\pi|_o) \). For each constructor symbol \( c/n \) of \( R \) of the sort of \( \pi|_o \), there is a rule of \( \mathbf{H} \) with argument \( \pi[c(x_1, \ldots, x_n)]|_o \), where \( x_1, \ldots, x_n \) are fresh variables, and this rule textually precedes \( r \) in the definition of \( \mathbf{H} \). Therefore, the label of \( q \) is not a constructor symbol, otherwise this rule would be applied to \( t \) instead of \( r \). Since the step of \( t \) is innermost, \( q \) cannot be labeled by \( \mathbf{H} \) either. Thus, the only remaining possibility is that \( q \) is labeled by an operation. We now prove that \( q \) is needed for \( t \). If \( n_1 \) and \( n_2 \) are disjoint nodes (neither is an ancestor of the other) of \( \mathcal{T} \), then the patterns of \( n_1 \) and \( n_2 \) are not unifiable. This is because they have different constructors symbols at the node of the inductive variable of the closest (deepest) common ancestor. Thus, since \( t = \sigma(\pi) \), only a rule of \( R \) stored in a \( \text{rule} \) node of \( \mathcal{T} \) below \( n \) can rewrite (a descendant of) \( t \) at the root, if any such a rule exists. All these rules have a constructor symbol at the node matched by \( o \), whereas \( t \) has an operation symbol at \( q \), the node matched by \( o \). Therefore, \( t \) cannot be reduced (hence reduced to a head-constructor form) unless \( t|_q \) is reduced to a head-constructor form. Thus, \( q \) is needed for \( t \).

Example 5. The situation depicted by the previous lemma can be seen in the evaluation of \( t = ([1]++[2])++[3] \). According to (4), \( \mathbf{H}(t) \to \mathbf{H}(\mathbf{H}([1]++[2])++[3]) \). The argument of the inner application of \( \mathbf{H} \) is both operation-rooted and needed for \( t \).

Lemma 5. Let \( R \) be an inductively sequential system and \( \mathbf{H} \) the head function of \( C_R \). Let \( t \) be an operation-rooted expression of \( R \) and let \( A \) denote an innermost finite or infinite computation \( \mathbf{H}(t) = e_0 \to e_1 \to \ldots \) in \( C_R \).

1. For every index \( i \) in \( A \), \( \mathcal{E}(e_i) \to \mathcal{E}(e_{i+1}) \) in \( R \).

2. If \( A \) terminates (it neither aborts nor is infinite) in an expression \( u \), then \( u \) is a head-constructor form of \( R \).

Proof. Claim 1: Let \( l \to r \) be the rule of \( \mathbf{H} \) applied in the step \( e_i \to e_{i+1} \). There are 3 cases for the origin of \( l \to r \). If \( l \to r \) originates from statement 04 of \texttt{compile}, then \( \mathcal{E}(e_i) = \mathcal{E}(e_{i+1}) \) and the claim holds. Otherwise \( l \to r \) originates from one of statements 08, 10, 14 or 15. In all these cases, a subexpression of \( e_i \) of the form \( \mathbf{H}(w) \) is replaced by either \( \mathbf{H}(u) \) (statements 08 and 15) or \( u \) (statements 10 and 14), in which \( w \) is an instance of the left-hand side of a rule of \( R \) and \( u \) is the corresponding right-hand side. Thus, in this case too, the claim holds.

Claim 2: If \( A \) aborts or does not terminate, the claim is vacuously true. So, consider the last step of \( A \). This step cannot originate from the application of a rule that places \( \mathbf{H} \) at the root of the contractum, since another step would become available. Hence the rule of the last step is generated by statement 10 or 14 of procedure \texttt{compile}. In both cases, the contractum is a head constructor form.

If \( A \) denotes a computation \( \mathbf{H}(t) = e_0 \to e_1 \to \ldots \) in \( C_R \), then, by Lemma 5, we denote \( \mathcal{E}(e_0) \to \mathcal{E}(e_1) \to \ldots \) with \( \mathcal{E}(A) \) and—with a slight abuse—we regard it as a computation in \( R \). Some expression of \( \mathcal{E}(A) \) may be a repetition of the previous one, rather than the result of a rewrite step. However, it is more practical to silently ignore these duplicates than filtering them out at the expenses of a more complicated definition. We will be careful to avoid an infinite repetition of the same expression. We extend the above viewpoint to computations of \( N(t) \) in \( C_R \), where \( t \) is any expression of \( R \).

Theorem 1. Let \( R \) be an inductively sequential system and \( \mathbf{H} \) the head function of \( C_R \). Let \( t \) be an operation-rooted expression of \( R \) and let \( A \) denote an innermost finite or infinite computation \( \mathbf{H}(t) = e_0 \to e_1 \to \ldots \) in \( C_R \). Every step of \( \mathcal{E}(A) \) is needed.
Proof. We prove that for every index \( i \) such that \( e_i \) is a state of \( A \), every argument of an application of \( H \) in \( e_i \) is needed for \( \mathcal{E}(e_i) \). Preliminarily, we define a relation “\( \prec \)” on the nodes of the states of \( \mathcal{E}(A) \) as follows. Let \( p \) and \( q \) be nodes of states \( \mathcal{E}(e_i) \) and \( \mathcal{E}(e_j) \) of \( \mathcal{E}(A) \) respectively. We define \( p \prec q \) iff \( i < j \) or \( i = j \) and the expression at \( q \) is a proper subexpression of the expression at \( p \) in \( \mathcal{E}(e_i) \). Relation “\( \prec \)” is a well-founded ordering with minimum element the root of \( t \). The proof of the theorem is by induction on “\( \prec \)”. Base case: Directly from the definition of “need”, since \( t \) is rooted by an operation of \( R \). Induction case: Let \( q \) be the root of the argument of an application of \( H \) in \( e_j \) for \( j > 0 \). We distinguish whether \( q \) is the root of the argument of an application of \( H \) in \( e_{j-1} \). If it is, then the claim is a direct consequence of the induction hypothesis. If it is not, \( e_{j-1} \rightarrow e_j \) is an application of a rule \( r \) generated by one of the statements 04, 08 or 15 of procedure compile. For statement 04, there is a node \( p \) of \( \mathcal{E}(e_j) \) that by the induction hypothesis is needed for \( \mathcal{E}(e_j) \) and matches the pattern \( \pi \) of the branch node of a definitional tree from which rule \( r \) originates. Let \( q \) be the node of the subexpression of \( e_j \) rooted by \( p \) matched by \( \pi \) at \( o \). By Lemma 3 \( q \) is needed for \( p \). Since \( p \) is needed for \( \mathcal{E}(e_j) \), by Lemma 2 \( q \) is needed for \( \mathcal{E}(e_j) \) and the claim holds. For statements 08 and 15, \( q \) is the root of the contractum of the redex matched by \( r \) which by the induction hypothesis is needed for \( \mathcal{E}(e_{j-1}) \). Node \( q \) is still labeled by an operation, hence it is needed for \( \mathcal{E}(e_j) \) directly by the definition of “need”.

\[ \blacksquare \]

Corollary 1. Let \( R \) be an inductively sequential system. Let \( t \) be an expression of \( R \) and let \( A \) denote an innermost finite or infinite computation \( N(t) = e_0 \rightarrow e_1 \rightarrow \ldots \) in \( C_R \). Every step of \( \mathcal{E}(A) \) is needed.

Proof. Operation \( N \) of \( C_R \) applied to an expression \( t \) of \( R \) applies operation \( H \) to every outermost operation-rooted subexpression of \( t \). All these expressions are needed by Def. [1]. The claim is therefore a direct consequence of Th. [1].

\[ \blacksquare \]

Corollary 2. Let \( R \) be an inductively sequential system. For all expressions \( t \) and constructor forms \( u \) of \( R \), \( t \rightarrow u \) in \( R \) if, and only if, \( N(t) \rightarrow u \) in \( C_R \) modulo a renaming of nodes.

Proof. Let \( A \) denote some innermost computation of \( N(t) \). Observe that if \( A \) terminates in a constructor form \( u \) of \( R \), then every innermost computation of \( N(t) \) terminates in \( u \) because the order of the reductions is irrelevant. Therefore, we consider whether \( A \) terminates normally. Case 1: \( A \) terminates normally. If \( N(t) \rightarrow u \), then by Lemma [5] point 1, \( t \rightarrow u \). Case 2: \( A \) does not terminate normally. We consider whether \( A \) aborts. Case 2a: \( A \) aborts. Suppose \( N(t) = e_0 \rightarrow e_1 \rightarrow \ldots \rightarrow e_i \), and the step of \( e_i \) reduces a redex \( r \) to “aborts”. By Theorem [1] \( r \) is needed for \( e_i \), but there is no rule in \( R \) that reduces \( r \), hence \( t \) has no constructor form. Case 2b: \( A \) does not terminates. Every step of \( \mathcal{E}(A) \) is needed. The complete tree unraveling [9] of the rules of \( R \) and the states of \( \mathcal{E}(A) \), gives an orthogonal term rewriting system and a computation of the unravelled \( t \). Since redexes are innermost, in this computation an infinite number of needed redexes are reduced. The hypernormalization of the needed strategy [9] Sect. 9.2.2] shows that hence \( t \) has no constructor form.

\[ \blacksquare \]

The object code \( C_R \) for a rewrite system \( R \) is subjectively simple. Since innermost reductions suffice for the execution, operations \( H \) and \( N \) can be coded as functions that take their argument by-value. This is efficient in most programming languages. Corollary 2 in conjunction with Theorem 1 shows that \( C_R \) is a good object code: it produces the value of an expression \( t \) when \( t \) has such value, and it produces this value making only steps that must be made by any rewrite computation. One could infer that there cannot be a substantially better object code, but this is not true. The next section discusses why.
5 Transformation

We transform the object code to avoid totally, or partially, constructing certain contracta. The transformation consists of two phases.

The first phase replaces certain rules of $H$. Let $r$ be a rule of $H$ in which $H$ is recursively applied to a variable, say $x$, as in the third rule of (4). Rule $r$ is replaced by the set $S_r$ of rules obtained as follows. A rule $r_f$ is in $S_r$ if $r_f$ is obtained from $r$ by instantiating $x$ with $f(x_1, \ldots, x_n)$, where $f/n$ is an operation of $R$, $x_1, \ldots, x_n$ are fresh variables, and the sorts of $f(x_1, \ldots, x_n)$ and $x$ are the same. If a rule in $S_r$ still applies $H$ to another variable, it is again replaced in the same way.

Example 6. The following rule originates from instantiating $y$ for “++” in the third rule of (4).

$$H([[++]^1(u++)]) = H(u++)$$ (10)

The first phase of the transformation ensures that $H$ is always applied to an expression rooted by some operation $f$ of $R$. The second phase introduces, for each operation $f$ of $R$, a new operation, denoted $H_f$. This operation is the composition of $H$ with $f$, and then replaces every occurrence of the composition of $H$ with $f$ with $H_f$.

Example 7. The second phase transforms (10) into:

$$H_{++}([[],u++]^2) = H_{++}(u,v)$$ (11)

After the second phase, operation $H$ can be eliminated from the object code since it is no longer invoked. We denote the transformed $C_R$ with $T_R$ and the outcome of the first phase on $C_R$ with $C_R'$.

The mapping $\tau$, from expressions of $C_R$ to expressions of $T_R$, formally defines the transformation:

$$\tau(t) = \begin{cases} 
H_f(\tau(t_1), \ldots, \tau(t_n)), & \text{if } t = H(f(t_1, \ldots, t_n)); \\
s(\tau(t_1), \ldots, \tau(t_n)), & \text{if } t = s(t_1, \ldots, t_n), \text{ with } s \text{ symbol of } R; \\
v, & \text{if } t = v, \text{ with } v \text{ variable}. 
\end{cases}$$ (12)

$T_R$ is more efficient than $C_R$ because, for any operation $f$ of $R$, the application of $H_f$ avoids the allocation of a node labeled by $f$. This node is also likely to be pattern matched later.

Example 8. Consider the usual length–of–a–list operation:

$$\begin{align*}
\text{length } [] &= 0 \\
\text{length } (_:xs) &= 1+\text{length } xs
\end{align*}$$ (13)

The compilation of (13), where we omit rules irrelevant to the point we are making, produces:

$$\begin{align*}
H(\text{length}([])) &= 0 \\
H(\text{length}(_:xs)) &= H(1+\text{length}(xs))
\end{align*}$$ (14)

The transformation of (14), where again we omit rules irrelevant to the point we are making, produces:

$$\begin{align*}
H_{\text{length}}([]) &= 0 \\
H_{\text{length}}(_:xs)) &= H_+(1,\text{length}(xs))
\end{align*}$$ (15)

Below, we show the traces of a portion of the computations of $N(\text{length } [7])$ executed by $C_R$ (left) and $T_R$ (right), where the number 7 is an irrelevant value. The rules of “+” are not shown. Intuitively, they evaluate the arguments to numbers, and then perform the addition.
Corollary 3. Let R be an inductively sequential system. For every operation-rooted expression t and head-constructor form u of R, \( H(t) \rightarrow u \) in \( C_R \) if, and only if, \( \tau(H(t)) \rightarrow u \) in \( T_R \) modulo a renaming of nodes.

Proof. Preliminarily, we show that for any s, \( H(t) \rightarrow s \) in \( C_R \) iff \( \tau(H(t)) \rightarrow \tau(s) \) in \( T_R \). Assume \( H(t) \rightarrow s \) in \( C_R \). There exists a rule \( l \rightarrow r \) of \( C_R \) and a match (graph homomorphism) \( \sigma \) such that \( H(t) = \sigma(l) \) and \( s = \sigma(r) \). From the definition of phase 2 of the transformation, \( \tau(l) \rightarrow \tau(r) \) is a rule of \( T_R \). We show that this rule reduces \( \tau(H(t)) \) to \( \tau(s) \). Since \( \tau \) is the identity on variables, and \( \sigma \) is the identity on non variables, \( \sigma \circ \tau = \tau \circ \sigma \). Thus \( \tau(H(t)) = \tau(\sigma(l)) = \sigma(\tau(l)) \rightarrow \sigma(\tau(r)) = \tau(\sigma(r)) = \tau(s) \). The converse is similar because there a bijection between the steps of \( C_R \) and \( T_R \).

Now, we prove the main claim. First, the claim just proved holds also when \( H(t) \) is in a context. Then, an induction on the length of \( H(t) \rightarrow u \) in \( C_R \) shows that \( \tau(H(t)) \rightarrow \tau(u) \) in \( T_R \). Since by assumption \( u \) is an expression of \( R \), by the definition of \( \tau \), \( \tau(u) = u \). \( \square \)
Finally, we prove that object code and transformed object code execute the same computations.

**Theorem 2.** Let \( R \) be an inductively sequential system. For all expressions \( t \) and \( u \) of \( R \), \( N(t) \rightarrow u \) in \( C_R \) if, and only if, \( N(t) \rightarrow u \) in \( T_R \).

**Proof.** In the computation of \( N(t) \) in \( C_R \), by the definition of \( \tau \), each computation of \( H(s) \) in \( C_R \), for some expression \( s \), is transformed into a computation of \( \tau(H(s)) \) in \( T_R \). By Lemma 5, the former ends in a head-constructor form of \( R \). Hence, by Corollary 3, \( \tau(H(s)) \) ends in the same head-constructor form of \( R \). Thus, \( N(t) \rightarrow u \) in \( T_R \) produces the same result. The converse is similar. \( \Box \)

### 7 Benchmarking

Our benchmarks use integer values. To accommodate a built-in integer in a graph node, we define a kind of node whose label is a built-in integer rather than a signature symbol. An arithmetic operation, such as addition, retrieves the integers labeling its argument nodes, adds them together, and allocates a new node labeled by the result of the addition.

Our first benchmark evaluates \( \text{length}(l_1++l_2) \), where \( \text{length} \) is the operation defined in (13). In the table below, we compare the same rewriting computation executed by \( C_R \) and \( T_R \). We measure the number of rewrite and shortcut steps executed, the number of nodes allocated, and the number of node labels compared by pattern matching. The ratio between the execution times of \( T_R \) and \( C_R \) varies with the implementation language, the order of execution of some instructions, and other code details that would seem irrelevant to the work being performed. Therefore, we measure quantities that are language and code independent. The tabular entries are in units per 10 rewrite steps of \( C_R \), and are constant functions of this value except for very short lists. For lists of one million elements, the number of rewrite steps of \( C_R \) is two million.

<table>
<thead>
<tr>
<th>( \text{length}(l_1++l_2) )</th>
<th>( C_R )</th>
<th>( T_R )</th>
<th>( O_R )</th>
</tr>
</thead>
<tbody>
<tr>
<td>rewrite steps</td>
<td>10</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>shortcut steps</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>node allocations</td>
<td>20</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>node matches</td>
<td>40</td>
<td>26</td>
<td>18</td>
</tr>
</tbody>
</table>

The column labeled \( O_R \) refers to object code that further shortcuts needed steps using the same idea behind the transformation. For example, in the second rule of (14), both arguments of the addition in the right-hand side are needed. This information is known at compile-time, therefore the compiler can wrap an application of \( H \) around the right operand of “+” in the right-hand side of the rule.

\[
H(\text{length}(-:xs)) = H(1+H(\text{length}(xs)))
\]

(16)

The composition of \( H \) with \( \text{length} \) is replaced by \( H_{\text{length}} \) during the second phase. The resulting rule is:

\[
H_{\text{length}}(-:xs)) = H_+(1,H_{\text{length}}(xs))
\]

(17)

Of course, there is no need to allocate a node for expression 1, the first argument of the addition, every time rule (15) or (17) is applied. A single node can be shared by the entire computation. However, since the first argument of the application of \( H_+ \) is constant, this application can be specialized or partially evaluated as follows:
\[ H_{\text{length}}(-:xs)) = H_{+1}(H_{\text{length}}(xs)) \] (18)

The application of rule (18) allocates no node of the contractum. In our benchmarks, we ignore any optimization that is not directly related to shortcutting. Thus \( C_R, T_R \) and \( O_R \) needlessly allocate this node every time these rules are applied.

The number of shortcut steps of \( T_R \) and \( O_R \) remain the same because, loosely speaking, \( O_R \) shortcuts a step that was already shortcut by \( T_R \), but the number of nodes allocated and matched further decreases. The effectiveness of \( T_R \) to reduce node allocations or pattern matching with respect to \( C_R \) varies with the program and the computation.

Our second benchmark computes the \( n \)-th Fibonacci number for a relatively large value of \( n \). The program we compile is:

\[
\begin{align*}
\text{fib} \ 0 &= 0 \\
\text{fib} \ 1 &= 1 \\
\text{fib} \ n &= \text{fib} \ (n-1) + \text{fib} \ (n-2)
\end{align*}
\] (19)

To keep the example simple, we assume that pattern matching is performed by scanning the rules in textual order. Therefore, the last rule is applied only when the argument of \( \text{fib} \) is neither 0 nor 1.

<table>
<thead>
<tr>
<th>( \text{fib} \ (n) )</th>
<th>( C_R )</th>
<th>( T_R )</th>
<th>( O_R )</th>
</tr>
</thead>
<tbody>
<tr>
<td>rewrite steps</td>
<td>10</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>shortcut steps</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>node allocations</td>
<td>24</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>node matches</td>
<td>44</td>
<td>26</td>
<td>16</td>
</tr>
</tbody>
</table>

The tabular entries are in units per 10 rewrite steps of \( C_R \) and are constant functions of this value except for very small arguments of \( \text{fib} \). For \( n = 32 \), the number of steps of \( C_R \) is about 17.5 million. With respect to \( C_R \), \( T_R \) avoids the construction of the root of the right-hand side of the third rule of (19). \( O_R \) transforms the right-hand side of this rule into:

\[ H_{+}(H_{\text{fib}}(H_{-}(n,1)),H_{\text{fib}}(H_{-}(n,2))) \] (20)

since every node that is not labeled by the variable or the constants 1 and 2 is needed. In this benchmark, there is also no need to allocate a node for either 1 or 2 every time (20) is constructed/executed. With this further optimization, the step would allocate no new node for the contractum, and the relative gains of our approach would be even more striking.

8 Functional Logic Programming

Our work is motivated by the implementation of functional logic languages. The graph rewriting systems modeling functional logic programs are a superset of the inductively sequential ones. A minimal extension consists of a single binary operation, called \textit{choice}, denoted by the infix symbol “\( ? \)”. An expression \( x ? y \) reduces non-deterministically to \( x \) or \( y \). There are approaches [5, 6] for rewriting computations involving the \textit{choice} operation that produce all the values of an expression without ever making a non-deterministic choice. These approaches are ideal candidates to host our compilation scheme.

Popular functional logic languages allow variables, called \textit{extra variables}, which occur in the right-hand side of a rewrite rule, but not in the left-hand side. Computations with extra variables are executed by \textit{narrowing} instead of rewriting. Narrowing simplifies encoding certain programming problems into
Needed Computations Shortcutting Needed Steps

programs [4]. Since our object code selects rules in textual order, and instantiates some variables of the rewrite system, narrowing with our object code is not straightforward. However, there is a technique [7] that transforms a rewrite system with extra variables into an equivalent system without extra variables. Loosely speaking, “equivalent”, in this context, means a system with the same input/output relation. In conjunction with this technique, our compiler generates code suitable for narrowing computations.

9 Related Work

The redexes that we reduce are needed to obtain a constructor-rooted expression, therefore they are closely related to the notion of root-neededness of [22]. However, we are interested only in normal forms that are constructor forms. In contrast to a computation according to [22], our object code may abort the computation of an expression \( e \) if no constructor normal form of \( e \) is reachable, even if \( e \) has a needed redex. This is a very desirable property in our intended domain of application since it saves useless rewrite steps, and in some cases may lead to the termination of an infinite computation.

Machines for graph reduction have been proposed [10, 20] for the implementation of functional languages. While there is a commonality of intent, these efforts differ from ours in two fundamental aspects. Our object code is easily translated into a low-level language like \( C \) or assembly, whereas these machines have instructions that resemble those of an interpreter. There is no explicitly notion of need in the computations performed by these machines. Optimizations of these machines are directed toward their internal instructions, rather than the needed steps of a computation by rewriting, a problem less dependent on any particular mechanism used to compute a normal form.

Our compilation scheme has similarities with deforestation [24], but is complementary to it. Both anticipate rule application, to avoid the construction of expressions that would be quickly taken apart and disposed. This occurs when a function producing one of these expressions is nested within a function consuming the expression. However, our expressions are operation-rooted whereas in deforestation they are constructor-rooted. These techniques can be used independently of each other and jointly in the same program.

A compilation scheme similar to ours is described in [2]. This effort makes no claims of correctness, of executing only needed steps and of shortcutting needed steps. Transformations of rewrite systems for compilation purposes are described in [12, 19]. These efforts are more operational than ours. A compilation with the same intent as ours is described in [8]. The compilation scheme is different. This effort does not claim to execute only needed steps, though it shortcuts some of them. Shortcutting is obtained by defining ad-hoc functions whereas we present a formal systematic way through specializations of the head function.

10 Conclusion

Our work addresses rewriting computations for the implementation of functional logic languages. We presented two major results.

The first result is a compilation scheme for inductively sequential graph rewriting systems. The object code generated by our scheme has very desirable properties: it is simple consisting of only two functions that take arguments by value, it is theoretically efficient by only executing needed steps, and it is complete in that it produces the value, when it exists, of any expression. The two functions of the object code are easily generated from the signature of the rewrite system and a traversal of the definitional trees of its operations.
The second result is a transformation of the object code that shortcuts some rewrite steps. Shortcutting avoids partial or total construction of the contractum of a step by composing one function of the object code with one operation symbol of the rewrite system signature. This avoids the construction of a node and in some cases and its subsequent pattern matching. Benchmarks show that the savings in node allocation and matching can be substantial.

Future work will rigorously investigate the extension of our compilation technique to rewrite systems with the choice operation and extra variables, as discussed in Sect. 8, as well as systematic opportunities to shortcut needed steps in situations similar to that discussed in Sect. 7.

Acknowledgments

This material is based upon work partially supported by the National Science Foundation under Grant No. CCF-1317249. This work was carried out while the second author was visiting the University of Oregon. The second author wishes to thank Zena Ariola for hosting this visit. The authors wish to thank Olivier Danvy for insightful comments and the anonymous reviewers for their careful reviews.

References


