A Language for Semantic Analysis

Jiazhen Cai

Technical Report 635

May 1993

New York University
Department of Computer Science
Courant Institute of Mathematical Science
251 Mercer Street, New York, NY 10012
<table>
<thead>
<tr>
<th>1. REPORT DATE</th>
<th>2. REPORT TYPE</th>
<th>3. DATES COVERED</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAY 1993</td>
<td></td>
<td>00-00-1993 to 00-00-1993</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4. TITLE AND SUBTITLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Language for Semantic Analysis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5a. CONTRACT NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5b. GRANT NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5c. PROGRAM ELEMENT NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5d. PROJECT NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5e. TASK NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5f. WORK UNIT NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York University, Department of Computer Science, Courant Institute of Mathematical Sciences, New York, NY, 10003</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>8. PERFORMING ORGANIZATION REPORT NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>10. SPONSOR/MONITOR’S ACRONYM(S)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>11. SPONSOR/MONITOR’S REPORT NUMBER(S)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>12. DISTRIBUTION/AVAILABILITY STATEMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approved for public release; distribution unlimited</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>13. SUPPLEMENTARY NOTES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>14. ABSTRACT</th>
</tr>
</thead>
<tbody>
<tr>
<td>see report</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>15. SUBJECT TERMS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>16. SECURITY CLASSIFICATION OF:</th>
<th>17. LIMITATION OF ABSTRACT</th>
<th>18. NUMBER OF PAGES</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. REPORT</td>
<td>unclassified</td>
<td>22</td>
</tr>
<tr>
<td>b. ABSTRACT</td>
<td>unclassified</td>
<td></td>
</tr>
<tr>
<td>c. THIS PAGE</td>
<td>unclassified</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>19a. NAME OF RESPONSIBLE PERSON</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
A Language for Semantic Analysis

Jiazhen Cai *

New York University/Courant Institute
New York, NY 10012
and
Purdue University
West Lafayette, IN 47907

ABSTRACT

Semantic analysis is important for compilers. In the APTS program transformation system, semantics is specified by rules in the language RSL. The semantic rules are interpreted by APTS to generate the semantic information of the program, which is then used by the rewriting engine for program translation. This approach has proved to be convenient and powerful in our construction of a SETL-to-C compiler. In this paper, we discuss the features, applications, implementation strategy, and performance of RSL.

1. Introduction

RSL is the specification language of the APTS system, an experimental program transformation system on which Robert Paige and the author have been working for several years. Recently, a SETL-to-C translator, written in RSL, has been built in the APTS environment with some success [4]. SETL is a very-high-level language and is convenient to use, but usually much slower than C. With the SETL-to-C translator, we combine the convenience of SETL with the efficiency of C.

Fig. 1 illustrates how the SETL-to-C translator (also called the SETL Accelerator) is built on the top of APTS. Considering the importance of semantic information to the high level language translation and optimization, we designed an inference engine in APTS to perform the semantic analysis. Semantics are specified by RSL rules. By applying the semantic rules on the parse tree of the input program, the inference engine computes the semantic information and stores it as database relations, which are then used by the rewriting engine to transform the input program. The actions of different modules are coordinated by a control file through the command

* The research of this author was partially supported by National Science Foundation Grant No. CCR-9002428, MIP-9300210, and by Air Force Office of Scientific Research Grant No. AFOSR-91-0308.
RSL was first designed for specifying semantic rules, and then extended to become the specification language of the APTS. In this paper, we will only discuss the subset of RSL which is used for semantic specifications. We argue that RSL is more convenient and powerful than the attribute grammar in specifying semantics.

2. The language

2.1. Rules and Transcripts

In APTS, semantics are specified in rules, and rules are organized into transcripts, as shown in Example 1.

A transcript consists three parts: a header, a declaration list, and a transcript body. In this example, the header is the first line, giving the name of the transcript `sg`; the declaration list contains one `rel` declaration, which declares three relations: a unary relation `person` of strings and two binary relations `parent` and `same_generation` of pairs of strings; and the body contains four rules, each having the form `left-hand-side → right-hand-side`. The first rule says that `p[1]`, `p[2]`, `p[3]`, `p[4]` and `p[5]` are all persons. The second rule says that `p[1]` is the parent of `p[2]`, and so on. The last two
Example 1. A transcript

```plaintext
transcript sg;
rel person: [string];
   parent, same_generation: [string, string];
begin
  true
  \rightarrow\ person(p1) and
       person(p2) and
       person(p3) and
       person(p4) and
       person(p5);
  true
  \rightarrow\ parent(p1, p2) and
       parent(p1, p3) and
       parent(p2, p4) and
       parent(p3, p5);
  person(.x)
  \rightarrow\ same_generation(.x, .x);
  parent(.x, .y) and
  parent(.z, .w) and
  same_generation(.x, .z)
  \rightarrow\ same_generation(.y, .w);
end;
```

rules defines the relation `same_generation`: two persons are of the same generation if they are the same person (rule 3) or their parents are of the same generation (rule 4). In RSL, variables are prefixed by ".".

The syntax and semantics of the rules are more or less conventional. Each left-hand-side is either a boolean constant `true`, a `left-term`, or several left-terms connected by logical connectors `and`, `or`, or `not`.

Each left-term has the form $R(p_1, \ldots, p_k)$, where $R$ is either a system defined predicate or a user declared relation, $p_1, \ldots, p_k$ are parameters, and $k$ is the arity of $R$. A `left-term` can be evaluated only when each of the free variables contained in the parameter list is bound to some constant. In case $R$ is a user declared relation, the left-term $R(p_1, \ldots, p_k)$ is true if and only if the tuple $[p_1, \ldots, p_k]$ (after variable substitution) is in $R$.

Each right-hand-side can be one `right-term` or several right-terms connected by `and`. Each right-term specifies an action and also has the form $R(p_1, \ldots, p_k)$, where $R$ can be either a system
defined procedure or a user defined relation name. In case \( R \) is a user defined relation, the action \( R(p_1, ..., pk) \) will insert the tuple \([p_1, ..., pk]\) into \( R \), if it is not there already.

A term \( R(p_1, ..., pk) \) is called a relational term if \( R \) is a user declared relation name. Notice that a relational term in the left-hand-side of a rule is interpreted differently than it is in the right-hand-side.

During the execution of a transcript, a rule becomes active if its left-hand-side evaluates to true with respect to some variable binding. A variable binding that makes the rule’s left-hand-side true is called an activating binding of this rule. If \( r \) is an active rule, and \( B \) is its activating binding, then we call the pair \((r, B)\) an active instance of \( r \). In general, there may be more than one active instance at a time. The inference engine will choose one active instance \((r, B)\) at a time nondeterministicly and perform the right-hand-side actions of \( r \) with respect to \( B \). An active instance may become non-active and vice versa because of the actions. Each active instance will be selected by the inference engine only once before it becomes non-active. The execution of a transcript terminates when all the right-hand-side actions of the active instances are performed, and none of them can change the current values of user defined relations.

When the transcript \( sg \) in Example 1 is executed, the resulting relation same_generation will be \{ \([p_1, p_1]\), \([p_2, p_2]\), \([p_3, p_3]\), \([p_4, p_4]\), \([p_5, p_5]\), \([p_2, p_3]\), \([p_4, p_5]\), \([p_3, p_2]\), \([p_5, p_4]\) \}.

### 2.2. External relations

In a transcript \( T \), relations declared in the rel declaration are called the local relations of \( T \), or the relations defined in \( T \). One transcript can make reference to relations defined in other transcripts by declaring them to be external. For example, the transcript in Example 1 can be broken into three transcripts using the external declaration, as shown in Fig. 2.

Relations used in a transcript must be declared as either local or external. A transcript can only modify its local relations.

Each transcript can be parsed separately. Parsed transcripts can be unloaded to a file and loaded back to the system later. When a transcript is executed, the system will first compute its external relations recursively before all the local relations are computed.

Let \( R_1 \) and \( R_2 \) be two relations declared in a transcript \( T \), and let \( r \) be a rule in \( T \). If \( R_1(...) \) is a left-term of \( r \), and \( R_2(...) \) is a right-term of \( r \), then we say \( R_2 \) depends on \( R_1 \). We define the dependency graph of a set of transcripts \( V \) to be \( G = (V, E) \), where \( E = \{ [x, y] \mid x, y \in V, \text{ an external relation of transcript } x \text{ is defined in the transcript } y \} \). The current implementation requires that \( G \) be acyclic. This means that if two relations recursively depend on each other, then they must be defined in the same transcript.
2.3. Tree patterns

To specify semantic properties of a language, we need some way to make reference to syntactic objects. In principle, a parse tree is just a set of relations, and so syntactic objects can be referred to using relational terms. But this approach is both inconvenient to use and inefficient to implement. In RSL, the syntactic objects are referred to as tree patterns and accessed through the built-in predicate match. For example, we can use match(%expr,.x+.y%) to find a node in the parse tree that represents an expression with an operator "+". We use a pair of "%" to delimit the tree pattern and use a "|" to connect the match-term with the rest of the left-hand-side.

Example 2 is a small portion of the transcript type for type analysis, which shows the usage of tree patterns and the match predicate. In this example, the relation type is declared to be a binary relation between trees and strings. A tree is just a subtree of the parse tree.

The first rule says that if .x is any expression in the parse tree, and its lexical type is int, then the type of .x is int. Here, eq is a system predicate that tests the equality of its two
Example 2: Tree patterns and the match predicate

transcript type;
rel type: [tree, string];
begin
  match(%expr, .x%)
  |  eq(lextyp(.x), int)
     -> type(.x, int);

  match(%expr, .x + .y%)
  |  type(.x, .t) or type(.y, .t)
     -> type(%expr, .x + .y%, .t);

  match(%expr, .x + .y%)
  |  type(%expr, .x + .y%, .t)
     -> type(.x, .t) and type(.y, .t);

-- inference rule for other syntactic constructs
-- ...
end;

arguments, and lextyp is a system function that returns the lexical type of its argument. The information about lexical type is computed by the lexical scanner and stored with the parse tree. The other rules are self-explanatory.

The lines introduced by the "--" signs are comments and are ignored by the system.

When applied to the parse tree x+y+1, the transcript will yield the following information:

\[ type = \{[1, \text{int}], [x+y, \text{int}], [x, \text{int}], [y, \text{int}], [x+y+1, \text{int}] \} \]

2.4. Type declaration

We require that the types of the parameters of each user defined relations be declared. Right now, the system only supports a limited number of types. Some most frequently used types are: string, list, tree and node. We have seen the types string and tree in the previous examples. The type node is also used to represent subtrees in the parse tree. A node represents an occurrence, but a tree represents a common subtree. In other words, different occurrences of the same subtree are represented as different nodes, but the same tree. Internally, trees are represented as value numbers, but nodes are represented as pointers to the root of the subtree.

A list can be either an empty list [], a string, or a list of one or more lists. List can be used to encode structured information. For example, we can use [set, t] to encode the type of sets
whose members have the type $t$, where $t$ is a string encoding a type variable, and use $[set, \ [tuple, \ [t1, t2]]]$ to encode the type of sets of tuples whose first component is of type $t1$ and second component $t2$. A string is a special case of a list. Thus, it would be more convenient to declare the type of the relation type to be $[tree, list]$ in Example 2.

2.5. Type conversion

The system supports limited run-time conversion between types. Consider the following example:

Example 3. Type conversion

```plaintext
transcript type_convert;
rel node_rel: [node];
tree_rel: [tree];
string_rel: [string];
list_rel: [list];
begin
match(%expr, .x+.y%) -> tree_rel(%expr, .x+.y%);
tree_rel(.x) -> node_rel(.x);
tree_rel(.x) -> string_rel(.x);
string_rel(.x) -> tree_rel(.x);  -- run time error
string_rel(.x) -> list_rel([.x, [.x]]);
end;
```

In the first rule, the pattern $.x+.y$ is first instantiated with a node and then converted into a tree when stored into tree_rel. In the second rule, the conversion goes the other way around. In general, tree-to-node conversion is not unique and should be used with caution. In the third rule, a tree is converted to a string by unparsing. But string to tree conversion, as used in the fourth rule, is not allowed. There is no conversion in the last rule.

2.6. Incremental operations

Consider the transcript in Example 4. When executed, it will give the following results:

- `addition_rel`: $\{[a, [b, _v1]], [a, [_v2, c]]\}$
- `replace_rel`: $\{[a, [b, _v1]]\}$ or $\{[a, [_v2, c]]\}$
- `unification_rel`: $\{[a, [b, c]]\}$

The two tuples in `addition_rel` are added by the first rule. By default, the incremental operation of the relation `addition_rel` is insert, i.e., when a new tuple is added to `addition_rel`, the old tuples in it will remain unchanged. Although the same two tuples are also added to `replace_rel`, only one remains, since the incremental operation of `replace_rel` is declared to be `replace`, and the first
Example 4. Incremental operations

```rsl
transcript incremental_ops;
rel addition_rel, replace_rel, unification_rel:
    [string, list];
key replace_rel, unification_rel: [1];
incremental replace_rel: replace;
    unification_rel: unify;
begin
    true
    -> addition_rel(a, [b, _v1]) and
        addition_rel(a, [_v2, c]);
    addition_rel(.x, .y)
    -> replace_rel(.x, .y) and
        unification_rel(.x, .y);
end;
```

components of its tuples are declared to be keys. Thus, when a new tuple is added to replace_rel, the system will remove from replace - rel the old tuple with the same first component. Similarly, since the incremental operation of unification_rel is declared to be unify, and the first components of its tuples are declared to be the keys, the system will unify the the new tuple with the old one in unification_rel that has the same first component as the new one. The identifiers with a prefix "_" are considered unification variables.

3. Questions and answers

Following are the RSL solutions to some of the frequently asked questions about rule-based systems.

3.1. Safe Rules

The set of activating bindings of a rule is called the conflict set of the rule. A conflict set can be infinite, as in the rule

```
not person(x) -> non_person(x);
```

We call such rules unsafe. A rule is safe if its conflict set is always finite. We want to avoid dealing with unsafe rules and give a simple sufficient condition for safe rules.

A left-term is positive if it is in the scope of even number of negations, and negative otherwise. A variable occurrence in some parameter of the match predicate or a positive relational left-term is called binding occurrence if it is not also contained in any function applications.
Let \( r \) be a rule. Let \( \text{binding}_\text{var}(r) \) be the set of variables that have binding occurrences in \( r \). Let \( \text{all}_\text{var}(r) \) be the set of all variables in \( r \). Obviously, if \( \text{binding}_\text{var}(r) = \text{all}_\text{var}(r) \), then \( r \) is safe. Rules satisfying such condition are acceptable. Our system only accepts acceptable rules. For example, the following rule is acceptable

\[
\text{living}_\text{being}(x) \text{ and not } \text{person}(x) \rightarrow \text{non}_\text{person}(x);
\]

where \( \text{living}_\text{being} \) is a user-defined relation.

### 3.2. Termination

It is easy to write a transcript that does not stop even if all the rules are acceptable:

```plaintext
transcript non_stop;
rel list_rel: [list];
begin
  true -> list_rel(1);
  list_rel(.x) -> list_rel([.x]);
end;
```

which defines the infinite unary relation \{ [1], [[1]], [[[1]]], ... \}

However, if all the rules are acceptable, and the incremental operations replace is not used, and if all the parameters to the relational right-terms are either constants or variables, then the termination is guaranteed.

The system is not responsible for the termination.

### 3.3. Fixed point

Even if a transcript terminates, its result may not be unique:

```plaintext
transcript non_unique;
rel R1, R2: [string];
begin
  not R1(1) -> R2(1);
  not R2(1) -> R1(1);
end;
```

If the first rule is fired first, the result would be \( r1 = \{ \}, r2 = \{[[1]]\}; \) otherwise, the result would be \( r1 = \{ [[1]] \}, r2 = \{ \}. \)

Consider the execution of a transcript \( T \), which defines the local relations \( R1, ..., Rk \). Let \( R = \{ [i, x]: i = 1..k, x \in Ri \} \). Let \( \text{next}(R) \) be the new value of \( R \) resulting from a right-hand-side action of some active rule. Then \( \text{next}(R) \) is a nondeterministic function of \( R \). Let \( \text{NEXT}(R) \) be the union of all possible values of \( \text{next}(R) \). From fixed point theory [2], if \( \text{next}(R) \) is inflationary, i.e., \( R \subseteq \text{next}(R) \), and if \( \text{NEXT}(R) \) is monotone in \( R \), then \( \text{next}(R) \) has a unique least fixed point if
it has any finite fixed point at all. In this case, the values of \( R_1, \ldots, R_k \) will be unique when \( T \) terminates.

The nondeterministic function \( \text{next}(R) \) is inflationary if the incremental operations of \( R_1, \ldots, R_k \) are all \textit{insert}. This claim is still true even when we allow some or all of the incremental operations to be \textit{unify}, if we generalize the relation \subseteq in such a way that \( X \subseteq Y \) iff for each \( x \) in \( X \), there is a \( y \) in \( Y \) that is more specific\(^*\) than or equal to \( x \).

Testing monotonicity is difficult in general. But it is easy to see that \textbf{if all the relational left-terms with local relation names are positive, then NEXT}(R) is monotone if \textit{next}(R) is inflationary}. There is no constraint on external relations, since external relations are not modified.

A transcript is called \textit{positive} if all the relational left-terms with local relation names are positive, and all their incremental operations are either \textit{insert} or \textit{unify}. In the knowledge-base system theory, it has been proved that the relations defined by a system of DATALOG rules has a least fixed point if the rules are stratified \cite{13}. It is not difficult to see that a system of DATALOG rules is stratified only if it can be organized into a set of positive RSL transcripts whose dependency graph is acyclic.

### 3.4. Expressive Power

In \cite{5}, Davis and Weyuker present a small language \( L \) which has the expressive power of a Turing machine. The programs in \( L \) use only three kinds of instructions: \( v := v+1, \ v := v-1 \), and \( \textbf{if} \ v \neq 0 \ \textbf{goto} \ A \). In Appendix I, we show that any \( L \) program can be simulated by an RSL transcript. For this reason, the termination of an RSL transcript is undecidable in general.

### 4. Implementation

The main difficulty in the implementation of rule based systems is the tremendous search space. For example, to evaluate the RSL rule

\[
\text{parent}(.x,.y) \ \textbf{and} \\
\text{same\_generation}(.x,.z) \ \textbf{and} \\
\text{parent}(.z,.w) \\
\rightarrow \text{same\_generation}(.y,.w);
\]

we have to search the relation \textit{parent} for all the pairs of the tuples \([.x,.y]\) and \([.z,.w]\) such that \([.x,.z]\) is a tuple in the relation \textit{same\_generation}. This search has to be done again after each modification of the relations \textit{parent} and \textit{same\_generation}, and this repeated search has to be done

\(^*\) We say \( y \) is more specific than \( x \) if \( y \) is a unification variable, or \( x \) and \( y \) are lists of the same length, and \( y \) is more specific than \( x \) componentwise.
for each rule in the transcript.

Our solution is to maintain a set pool of partially instantiated rules, called instances. An instance I is a pair (r, B) containing a rule r and a partial variable binding B. The rules can be implemented as pointers to the locations where the rules are actually stored, and the bindings can be implemented as sets of pairs. If the domain of B contains all the pattern variables of r, then we say that the instance I is ready and the binding B is complete. For correctness, we want pool to be large enough to contain all the active instances (recall that these are ready instances whose left-hand-sides evaluate to true). For the efficiency, however, we want pool to be as small as possible.

Pool is initialized as follows. First, the rules with a leading match predicate are considered. A tree pattern matching [3] is done in linear time to determine the subset of the patterns occurring in the match predicates that match some of the subtrees of the parse tree. For each rule with a matched pattern, one instance is created for each matching. No instance is created for those rules with unsatisfied match predicates. The rules with no match predicates are added to pool with an empty binding set.

If there is any external relation, then they are used to initialize pool also. Let R be an external relation, and t be a tuple in R. Let I = (r, B) be an instance in pool. Let R(p) be a positive left-term of r with the parameter list p. If t matches p with a resulting binding B_{p,t}, and B_{p,t} is consistent with the binding B, then a new instance (r, B_{t,p} ∪ B) is inserted into pool.

After the initialization, the main loop begins:

\[ new := \text{the set of ready instances}; \]
\[ \textbf{while} new \neq \{\} \ \textbf{loop} \]
\[ \text{remove an instance } (r, B) \text{ from } new; \]
\[ \text{evaluate the left-hand-side of } r \text{ with the binding } B; \]
\[ \text{if the result is } \textbf{true} \]
\[ \text{then perform the right-hand-side actions of } r; \]
\[ \text{if any new term } t \text{ is added to a relation } R \text{ then} \]
\[ 1 \ \textbf{for} \text{ each instance } (r_i, B_i) \text{ having a positive} \]
\[ 2 \ \text{left-term } R(p) \text{ such that } t \text{ matches } p \text{ with a resulting binding} \]
\[ 3 \ B_{i,p} \text{ that is consistent with } B, \textbf{loop} \]
\[ \text{if } B_i \text{ is already complete} \]
\[ \text{then add } (r_i, B_i) \text{ to } new; \]
\[ \text{elseif } B_{i,p} \neq \{\} \text{ then} \]
\[ \text{add the new instance } (r, B_{i,p} \cup B_i) \text{ to } pool; \]
\[ \text{if } B_{i,p} \cup B_i \text{ is complete} \]
then add \((r, B_{t,p} \cup B_t)\) to new;
end if;
end if;
end loop;
end if;
end if;
end loop;
end if;
end loop;

Indices are created to speed up the search of lines 1, 2, and 3. Since the indexing method for ready instances is different than that for the non-ready instances, we store the two kinds of instances separately. For ready instances, the index is established on the instantiated positive relational left-terms. For non-ready instances, the indices are established in two stages. Let \(P\) be the set of non-ready instances in pool. If \(I = (r, B)\) is an instance, \(t\) is a term in \(r\), and \(t'\) is obtained from \(t\) by variable substitution using \(B\), then we say \(t'\) is a term of \(I\). A term of \(I\) is partial if it still contains variables. Let \(T\) be the set of partial positive relational left-terms contained in instances in \(P\). There is one first stage index \(\text{INDX}_1\_{R,i}\) for each component \(i\) of each relation \(R\):

\[
\text{INDX}_1\_{R,i} = \{ [x, t] : t \text{ is a term in } T \text{ with predicate symbol } R, \text{ the } i\text{th component of } t \text{ is } x, \text{ and } x \text{ is a constant } \}
\]

In addition, for each relation \(R\), there is an index \(\text{INDX}_1\_{R,0}\) that contains the set of terms in \(T\) with no constant components.

The second stage index \(\text{INDX}_2\) maps each term in \(T\) to the instances in \(P\) that contains it:

\[
\text{INDX}_2 = \{ [t, I] : t \in T, I \in P, t \text{ is a positive relational left-term of } I \}
\]

Now suppose a new tuple \(t\) is added to the relation \(R\) whose arity is \(k\). To find the set of instances in \(P\) that can be further instantiated with \(t\), we first use \(\text{INDX}_1\) to find the \(R\)-term in \(T\) whose arguments may match \(t\):

\[
\text{may_match} = \text{INDX}_1\_{R,0} \cup \text{INDX}_1\_{R,1}(t[1]) \cup \ldots \cup \text{INDX}_1\_{R,k}(t[k])
\]

where \(t[i]\) is the \(i\)th component of \(t\). For each \(R\)-term \(tr\) in \(\text{may_match}\), we check whether its argument list matches \(t\) with some nonempty binding \(B_t\). If so, we find the instances in pool that contain \(tr\) as a positive left-term:

\[
\text{affected} = \text{INDX}_2\{tr\}
\]

For each instance \((r, B)\) in \(\text{affected}\) we create a new instance \((r, B \cup B_t)\), and increment \(\text{INDX}_1\) and \(\text{INDX}_2\) accordingly.

The following examples give further details of the indexing method. Consider the rule from the transcript in Example 1:

\[
r. \quad \text{parent}(x, y) \text{ and}
\]
parent(z, w) and
same_generation(z, x)
-> same_generation(y, w);

Suppose the instance \( I = (r, \{[x, p 1], [z, p 1]\}) \) is in the pool already, and a new tuple \( t = [p 1, p 2] \) is added to parent. We expect that the active instance \( I_0 = (r, \{[x, p 1], [y, p 2], [z, p 1], [w, p 2]\}) \) will be generated. As described above, we use \( \text{INDX}_1 \) to find the two partial terms \( \text{parent}(p 1, y) \) and \( \text{parent}(p 1, w) \), and use \( \text{INDX}_2 \) to find the instance \( I \) for further instantiation. By matching \( \text{parent}(p 1, y) \) with \( \text{parent}(p 1, p 2) \), we create the new instance \( I_1 = (r, \{[x, p 1], [z, p 1], [y, p 2]\}) \). By matching \( \text{parent}(p 1, w) \) with \( \text{parent}(p 1, p 2) \), we create the new instance \( I_2 = (r, \{[x, p 1], [z, p 1], [w, p 2]\}) \). Indices are modified accordingly: for \( I_1 \), we add the tuple \( \text{parent}(p 1, w), I_1 \) to \( \text{INDX}_2 \); for \( I_2 \), we add the tuple \( \text{parent}(p 1, y), I_2 \) to \( \text{INDX}_2 \). If we stop here, then the ready instance \( I_0 \) is still missing. The actual algorithm continues as follows. The newly added \( \text{INDX}_2 \) entries mean that if a new tuple \( [p 1, w] \) is added to \( \text{parent} \), than \( I_1 \) can be further instantiated; and if a tuple \( [p 1, y] \) is added to \( \text{parent} \), than \( I_2 \) can be further instantiated. Since both \( [p 1, w] \) and \( [p 1, y] \) match \( t \), the two instances \( I_1 \) and \( I_2 \) are retrieved immediately and the ready instance \( I_0 \) is created.

Here is another interesting example. Consider the rule
\[
\begin{align*}
  r: & \ A(x, f(y)) \quad \text{and} \quad B(y, f(x)) \rightarrow C(x, y); \\
\end{align*}
\]
where \( A \) and \( B \) are two local relations, and \( f \) is a system defined function with \( f(b) = b \) and \( f(a) = c \). Let \( I = (r, \{\}) \) be an instance in pool. Suppose first the tuple \( t_1 = [a, b] \) is added to \( A \) and then the tuple \( t_2 = [b, c] \) added to \( B \). We show how our algorithm works in this situation. When \( t_1 \) is added to \( A \), the term \( A(x, f(y)) \) is retrieved using \( \text{INDX}_1_{A,0} \), since both \( x \) and \( f(y) \) are not instantiated. By matching \( [x, f(y)] \) with \( t_1 \), we get the binding \( \{[x, a]\} \). No binding is generated for \( y \), however, since \( f(y) \) is a function application. As a result, a new instance \( I_1 = (r, \{[x, a]\}) \) is created. Since \( f(a) = c \), the tuple \( [c, B(y, c)] \) is added to \( \text{INDX}_1_{B,2} \) and \( [B(y, c), I_1] \) is added to \( \text{INDX}_2 \). When \( t_2 \) is added to \( B \), the term \( B(y, c) \) is retrieved through \( \text{INDX}_1_{B,2} \), and the instance \( I_1 \) is retrieved through \( \text{INDX}_2 \) with the term \( B(y, c) \). By matching \( t_2 \) with \( B(y, c) \), a ready instance \( I_2 = (r, \{[x, a], [y, b]\}) \) is generated.

4.1. Rule splitting

Consider the rule
\[
\begin{align*}
  r: & \ A(x) \quad \text{and} \quad \\
  & \ B(y) \quad \text{and} \quad \\
  & \ C(x, y) \\
  & \rightarrow ABC(x, y); \\
\end{align*}
\]
If our algorithm is implemented naively, at least \( |A| \ast |B| \) instances will be generated from \( r \).
alone, which is not desirable. If we split \( r \) into the following two rules:

\[
\begin{align*}
\text{r1: } & \quad A(x) \textbf{ and } \\
& \quad BC(x, y) \\
& \quad \rightarrow ABC(x, y);
\end{align*}
\]

\[
\begin{align*}
\text{r2: } & \quad B(y) \textbf{ and } \\
& \quad C(x, y) \\
& \quad \rightarrow BC(x, y);
\end{align*}
\]

then the number instances generated from \( r1 \) and \( r2 \) will be \( O(|A| + |B| + |C|) \). Finding a good splitting is actually a problem of join optimization [13], and has been studied extensively in the relational database theory.

5. Performance

We did several experiments on the inference engine to see the effectiveness of our implementation strategy. In these experiments, we applied two typical transcripts, one for control flow analysis, the other for live code analysis, to different groups of SETL programs to see how the running time depended on the size of the output relations. Fig. 3 shows the result of applying the transcript for control flow analysis to ten real SETL programs.

![Fig. 3. Control flow analysis on ten real SETL programs](image)

This result does not give a very clear picture of the performance. To reduce the influence of the structures of different programs on the result, we constructed two groups of artificial programs. Each program in the first group contains a sequence of identical \textbf{for}-loops, as shown in
Fig. 4. Artificial programs with sequences of loops

```plaintext
program sequence;
    read(x, y);
    for z in y loop
        if z not in x then
            x with:= z;
        end if;
    end loop;
    for z in y loop
        if z not in x then
            x with:= z;
        end if;
    end loop;
...
    print(x);
end sequence;
```

In the programs of the second group, these loops are nested, as shown in Fig. 5. The results from the experiments with these artificial programs are shown in Fig. 6, Fig. 7, and Fig. 8.

In control flow analysis, the output is linear to the size of the input program. In live code analysis, the output is quadratic. In both cases, the experiments show that the running time is linear to the size of the output, which is the best we can expect.

6. Comparison with other rule based systems

There are many other implementations of rule based systems. Among those best known are RETE algorithm [6-8, 12] and TREAT algorithm [9-11]. The semi-naive algorithm [13], according to [10], is a special case of TREAT.

In RETE, the left-hand-sides of rules are compiled into a data flow network, as illustrated in Fig. 9. There are two kinds of nodes in the network. The one-input nodes match new tuples with relational left-terms, and the two-input nodes implement the join operations. The output of one-input nodes is stored in the α memories and the output of two-input nodes is stored in the β memories. When a new tuple is inserted into some relation, a sequential search is performed to locate the α memories that need to be modified. For each two-input node, when the memory at one input is modified, a sequential search is performed on the memory on the other input to modify the output β memory. In our system, the search of affected α memories and the
program nested;
    read(x1, y1);
    read(x2, y2);
    ...
    for z1 in y1 loop
        if z1 not in x1 then
            x1 with:= z1;
            for z2 in y2 loop
                if z2 not in x2 then
                    x2 with:= z2;
                    ...
                end if;
            end loop;
        end if;
    end loop;
    end if;
end loop;
...
print(x2);
print(x1);
end nested;

Fig. 5. Artificial programs with nested loops

Fig. 6. Live code analysis on sequential loops
incremental modification of the $\beta$ memories are effectively achieved by maintaining the set of instances, and are performed more efficiently with the two-step indexing.

A close look at our implementation and RETE algorithm reveals that the set of instances in our algorithm corresponds to the union of all the tuples in all the $\alpha$ memories and $\beta$ memories. Thus the space usage of our implementation is comparable to that of RETE.

The TREAT algorithm saves space by omitting all the $\beta$ memories. Portions of the $\beta$ memory elements are computed when needed. Some version of TREAT uses index to speed up
the search of the $\alpha$ memories in computing the joins, but linear search is still used to locate the $\alpha$ memories that must be modified when new tuples are generated.

Another advantage of our system is that we separate tree patterns and relational patterns. This allows us not only to write semantic rules more concisely and intuitively, but also to implement the language more efficiently using the linear-time bottom-up multi-pattern tree matching algorithm [3].

7. Comparison with attribute grammars

Like RSL, the attribute grammar [1] is also a formal tool of specifying program semantics. In an attribute grammar, semantics are specified as attributes of syntactic objects. While RSL associates semantic rules with tree patterns, the attribute grammar associates semantic rules with grammar productions. Therefore the attribute grammar is more grammar dependent than RSL.

In the attribute grammar, semantic information is strictly propagated along the parse trees. Considering the efficiency of implementation, most systems using attribute grammars restrict the order in which the attributes can be evaluated. For example, in order to evaluate the attributes in one pass through the parse tree, a compiler system usually requires that the computation of an attribute at one node of the parse tree should not use information from its right siblings and the
right siblings of its ancestors. With this restriction, it is very difficult to specify the semantic rules in our Example 2 using the attribute grammar. In RSL, however, the parse tree is traversed once only to get the variable bindings, and then the semantic information is computed as database relations. The user do not have to worry about the evaluation order.

Appendix I: Simulation of the Language $L$

We show how to use a RSL transcript to simulate a program written in the language $L$ described in [5], which can express all the partially computable functions.

Language $L$ has three kinds of statements: $v := v + 1$, $v := v - 1$, and if $v \neq 0$ goto $A$.

Consider a program $P$ in $L$ with statements $s_1, \ldots, s_k$. To simulate $P$ in RSL, we create a transcript $T_P$ with the following declarations:

```plaintext
transcript $T_P$;
rel next: [string, string];
value: [string, list];
done: [string];
key value: [1];
incremental value: replace;
begin
...
end;
```

The transcript body contains the following.

1. A rule to simulate the entrance of the program $P$:

```plaintext
true -> next($s_1$, newatom($s$));
```

where $newatom(x)$ is a system function returning a fresh new string with the prefix $x$ each time it is called.

2. Rules to initialize the variables, one rule for each variable $v$ in $P$:

```plaintext
true -> value($v$, []);
```

We use [] to encode 0. By convention, the initial value of each variable is 0.

3. Rules to simulate statements in $P$. If a statement $s_i$ has the form $v := v + 1$, then it is simulated by the rule

```plaintext
next($s_i$, $x$) and
not done($x$) and
value($v$, $y$)
-> done($x$) and
```
\(\text{next}(s_{i+1}, \text{newatom}(s)) \textbf{ and} \)
\[
\text{value}(v, \ [y]);
\]
Here \([x]\) encodes \(v+1\) if \(x\) encodes \(v\). If \(s_i\) has the form \(v := v - 1\), then it is simulated by two rules:
\[
\text{next}(s_i, x) \textbf{ and} \]
\[
\text{not done}(x) \textbf{ and} \]
\[
\text{value}(v, \ [y])
\]
\[- \rightarrow \text{done}(x) \textbf{ and} \]
\[
\text{next}(s_{i+1}, \text{newatom}(s)) \textbf{ and} \]
\[
\text{value}(v, \ .y);
\]
\[
\text{next}(s_i, x) \textbf{ and} \]
\[
\text{not done}(x) \textbf{ and} \]
\[
\text{value}(v, \ [])
\]
\[- \rightarrow \text{done}(x) \textbf{ and} \]
\[
\text{next}(s_{i+1}, \text{newatom}(s));
\]
By convention in \(L\), \(0 - 1 = 0\). If \(s_i\) has the form \(\text{if } v \neq 0 \text{ goto } s_j\), then it is simulated by two rules:
\[
\text{next}(s_i, x) \textbf{ and} \]
\[
\text{not done}(x) \textbf{ and} \]
\[
\text{value}(v, \ [y])
\]
\[- \rightarrow \text{done}(x) \textbf{ and} \]
\[
\text{next}(s_j, \text{newatom}(s));
\]
\[
\text{next}(s_i, x) \textbf{ and} \]
\[
\text{not done}(x) \textbf{ and} \]
\[
\text{value}(v, \ [])
\]
\[- \rightarrow \text{done}(x) \textbf{ and} \]
\[
\text{next}(s_{i+1}, \text{newatom}(s));
\]

The proof of the correctness should be straightforward.

\textbf{Acknowledgment} I want to thank Robert Paige for his encouragement and support, without which the design and implementation of RSL would be impossible.
References


