A HISTORY-BASED SCHEME FOR ACCELERATING PROLOG INTERPRETATION

Vishv Mohan Malhotra
Discipline of Computer Science
The Flinders University of South Australia
Bedford Park 5042
AUSTRALIA.

Tang Van To
Division of Computer Science
Asian Institute of Technology
G.P.O. Box 2754, Bangkok 10501
THAILAND.

ABSTRACT

An algorithm to improve the performance of a Prolog interpreter is introduced. The algorithm, unlike the intelligent backtracking schemes which improve the performance by avoiding redundant redos, avoids redundant calls. The algorithm identifies the redundant calls by maintaining a history of the program execution. The algorithm can be used in conjunction with an intelligent backtracking scheme for a further speed-up of the programs.

1. INTRODUCTION

Most intelligent backtracking (IB) schemes for Prolog, reported in literature [1, 4, 6, 8, 10, 12], consist of two parts: (a) an algorithm to compute the set of suspects; and (b) an algorithm to choose the backtrack-points using a data-base of these sets. Much of the work has centered on the former algorithm (see also 2, 5, 9). The emphasis has been to design algorithms to construct smaller sets of suspects at lower costs. It is the algorithm to compute the set of suspects that distinguishes one backtracking scheme from the other.

We believe that the IB schemes for choosing the backtrack-points do not make a full use of the available information. Specifically, after a backtrack, the interpreter resumes the search without avoiding the failed calls that have not yet been attended to. As a consequence, the interpreter repeats several failures in the resumed search. The algorithms reported in [8] is expected to reduce the search space for 6-queens problem (clever solution) [1] by about 10%. It is shown later in this paper that the reduction can exceed 35% through a better use of the information.

In this paper, we introduce an algorithm to improve the performance of a Prolog interpreter by reducing thrashing. The algorithm achieves this improvement by avoiding calls which begin searches that are destined to eventually fail. An IB scheme, on the other hand, controls the thrashing by selecting the appropriate goals for redo. The two schemes affect the programs differently; there may be redundant calls that an IB scheme fails to eliminate. On the other hand, the scheme described in this paper may not skip some unproductive redos. Thus, the two methods can be viewed as complimenting each other. Indeed, many data-structures and procedures are common in the two schemes and can be shared.

In the next section, we illustrate, with an example, the calls that can be avoided to reduce the search space. Section 2 also introduces additional terminology that supplements the common Prolog terminology [3, 7]. The algorithm is introduced in Section 3. Section 4 presents the statistics obtained by running certain benchmarks using the algorithm. In the final section, the paper is concluded with some remarks about the possible extensions of the scheme.

2. PRELIMINARIES: EXAMPLE AND TERMINOLOGY

Consider the following Prolog program and its execution trace under an IB scheme:

```
7 -> r(A), r(B), s(A), t(B).
R1: r(1).
R2: r(2).
S1: s(C) <- t(C).
S2: s(1).
T1: t(2).
```

Execution trace under an IB:

```
1: goal r(A) calls clause R1. A unifies with 1.
2: goal r(B) calls clause R1. B unifies with 1.
3: goal s(A) calls clause S1. C unifies with 1.
4: goal t(C) fails. REDO s(A).
5: goal s(A) calls clause S2.
6: goal t(B) fails. REDO r(B).
7: goal r(B) calls clause R2. B unifies with 2.
8: goal s(A) calls clause S3. C unifies with 1.
9: goal t(C) fails. REDO s(A).
10: goal s(A) calls clause S2.
11: goal t(B) calls clause T1.
```
Steps (8) and (9), in the above trace, are redundant as they repeat steps (3) and (4) under the same bindings. We call a search redundant if it is known, from the past behavior of the interpreter, that the search will end unsuccessfully. The scheme that we describe, in this paper, aims at maintaining the execution history of the interpreter in a form that can be used readily to identify and avoid repeated execution of the redundant calls.

To detect the redundant calls, the execution history of a program is maintained as an AND-OR tree. Indeed, we abstract the actions of a Prolog interpreter as a traversal (and construction) over this tree. In a history tree a goal is represented by an OR node. The children of a goal-node represent the clauses in the goal procedure. A clause is represented by an AND node with its children representing the goals (literals) in the body of the clause. A special goal-node -- we call it GO -- constitutes the root of the tree. A clause-node, CO, is attached as the only child of GO. The user query defines the child goal-nodes of CO. The children of the goal-nodes are ordered by the sequencing of the clauses in the program. Similarly, the children of a clause-node are ordered by the literal ordering in the body of the clause.

For a goal-node, G, we use the term parent clause to refer to its parent node in the tree; the parent goal refers to the parent of the parent clause; the sibling of a goal is its right sibling -- a sibling shares the parent clause with the goal and is defined by the literal in the defining clause that immediately follows the literal defining node G. An analogous terminology is used for the clause-nodes. For a clause-node, its sibling is the node sharing the parent goal and is defined by the next clause in the goal procedure.

### 2.1 A PROLOG INTERPRETER

Initially, a tree consists of nodes GO, CO and the goal-nodes representing the user query. The interpreter starts the execution by visiting the leftmost child goal-node of CO and invoking a call for the goal.

**CALL**

During a call, the interpreter executes the goal that it is currently visiting by unifying the goal with the head of a clause. As a clause is selected, an AND node representing the clause is inserted as a child of the goal-node if it is not already there. (It simplifies the presentation of algorithms, in the later sections, if the interpreter also inserts a dummy node representing the sibling of the clause. No dummy sibling is needed for the last clause in the goal procedure.) The goals (literals) in the body of the clause are attached to the clause-node as its child goal-nodes. After a successful call -- a call is successful if a unifying clause is found -- the interpreter traverses to the leftmost child (goal-node) of the unifying clause. A call is then invoked for the new goal.

**EXIT**

The interpreter exits a goal-node if no descendent of the goal remains to invoke a call. After an exit from a goal-node the interpreter traverses to its sibling and invokes a call.

**REDO**

If the interpreter fails to execute a goal, it traverses back, in the reverse order of call invocations, to a previously executed goal and invokes a redo. An attempt is made to execute the goal using a different clause. After a successful redo the interpreter traverses to the leftmost goal-node in the unifying clause and invokes a call for the new goal. A goal that has unified with a head, either during a call or redo, is called an executed goal. The set of executed goal-nodes and their unifying clause-nodes is called a search.

**FAIL**

A goal is said to have failed if it can not unify with the head of any clause during a call or a redo step. The interpreter traverses (backtracks) to an executed goal-node, as described earlier, and performs a redo. As the interpreter backtracks it undoes the effects of the executed calls and redos.

### THE MOST RECENT SUSPECT

When a goal G fails to unify with the head, H, of a clause an IB interpreter computes a set, S(G,H), of the executed goals that are suspected to be contributing bindings impeding the unification. Some IB schemes (e.g., [61]) may do so only when G fails. For a set S(G,H), let L(G,H) denote the most recently executed goal in the set. Clearly, it is unnecessary to try to execute goal G using H until the interpreter has backtracked to L(G,H). It is only after the interpreter has backtracked to goal L(G,H) that the bindings in G may change to let G unify with H.

In next section, we introduce an algorithm to control the execution of an interpreter to avoid the redundant calls.
3. AN ALGORITHM TO AVOID REDUNDANT CALLS

The algorithm for identifying the redundant calls has two main components. One component remembers for each clause-node that has been removed from a search, the most recently executed goal among the goals suspected to have caused the removal. A clause-node that was tried but failed to be a part of a successful search is not used again as a candidate clause for executing the parent goal until the interpreter has backtracked to this search. The extent of the backtracking done by the interpreter between two consecutive visits of a node is determined by the other component of the algorithm. The algorithms for these components are introduced in the following subsections.

3.1 REMEMBERING THE MOST RECENT SUSPECT

The algorithm remembers the most recently executed suspect for a clause-node by assigning tag, called L tag, to the node. An L tag of a node specifies the most recent goal to which the interpreter must backtrack before executing a call by including the node in the search. An algorithm to determine L tags for the nodes in the history tree, as the interpreter traverses over it, is described in the following paragraph. The algorithm also assigns L tags to the goal-nodes in the tree. The tags on the goal-nodes enable the interpreter to compute L tags for the clause-nodes as it traverses over the tree.

A clause-node may be removed from a search for three reasons:

(i) The clause fails to unify with the goal invoking the call or redo. In this case L(G,N), where G is the goal and N the head of the clause, is assigned as L tag with the clause-node.

(ii) The child goals of the clause-node fail. In this case L tag of the clause-node is computed from the L tags of the its child goal-nodes. For the clause-node its L tag is the earliest executed goal that appears as L tag on a child goal-node. The clause can not be in a successful search unless all its child goals can execute successfully. For a goal-node, L tag is determined by the L tags of its child clause-nodes. A goal can be executed using any one of the clauses in its procedure. Thus, L tag of a goal-node is the most recently executed goal appearing as a L tag on its child clause-nodes. Or,

(iii) A redo step is executed for the parent goal of the clause-node. The clause-node remains a potential candidate for inclusion in a search and is used as a candidate clause if the parent goal invokes a new call.

The L tags for the nodes are computed and assigned to the nodes as the interpreter backtracks from a goal to another goal to invoke a redo. The tag for a clause-node is computed when the interpreter backtracks from the leftmost goal in the clause to the parent goal.

3.2 DETERMINING THE EFFECTIVE BACKTRACKING

The other major component of the scheme is an algorithm to determine, as the interpreter visits the goal-nodes and invokes calls, the earliest goal to which the interpreter has backtracked since the previous visit of the node. This information is compared with the L tag on a clause-node to determine if the interpreter should try to execute the goal using the clause or not. The search involving a clause-node will end unsuccessfully if the interpreter has not backtracked to the goal specified by the L tag of the clause-node. The idea behind the algorithm is described below:

Imagine that each goal has a characteristic color. A goal executed earlier has a darker hue than the one executed later -- a darker hue can be painted over a lighter hue but not the visa versa. The interpreter picks the color of the goal when it performs a redo. It spreads this color by painting the sibling and the child nodes as it visits the nodes in its traversal over the tree. As the interpreter backtracks and invokes more redos, several balloons of these colors accumulate. It is essential that these colors be spread ahead of the interpreter over the history-tree. If a darker hue overtakes a lighter hue the interpreter may economize its efforts and discard the lighter hue. The color of a goal-node, when the interpreter invokes a call for the goal, is determined by the darkest hue that is painted on the node. Thus, if a goal-node is set colorless when it fails, the color of the node will determine the earliest goal to which the interpreter has backtracked since it failed. For certain other considerations, the goal-nodes are set colorless after the interpreter has invoked a call for the goal. The color is, however, remembered by assigning the same to one of its child clause-nodes.

3.3 THE ALGORITHM

Let G be a goal that has just failed. Let L be its L tag. As every goal in a search must execute successfully, no search that involves G can be successful unless the interpreter backtracks to L. The interpreter may, therefore, directly invoke a redo for L without exploring the intervening goals for a solution. There, however, may be an alternate search that does not involve goal G. The interpreter may begin this search by performing a redo for the parent goal of G. To systematically explore all searches the algorithm directs the
INITIALIZATION:

1. Initialize the history tree by creating a goal-node \( G_0 \) with child clause \( C_0 \). Insert goals in the user query as children of \( G_0 \). Let \( G_1 \) be the leftmost child. Set \( G_0.C = G_0 \). For each child \( G_i \) of \( G_0 \) set \( G_i.C = \text{no-color} \), and \( G_i.L = \text{nil} \).

2. Set \( \text{GOAL} = G_1 \). Set \( \text{COLOR} = G_0 \).

3. Goto CALL.

CALL:

1. Set \( \text{GOAL}.C = \text{no-color} \); set \( \text{GOAL}.L = \text{nil} \).

2. (Paint the sibling) If \( \text{GOAL} \) has a sibling and the call for \( \text{COLOR} \) was invoked before the call for \( \text{SIBLING}.C \) then set \( \text{SIBLING}.C = \text{COLOR} \). This assignment is also done if \( \text{SIBLING}.C = \text{no-color} \).

3. Goto GET_UNIFYING_CLAUSE.

REDO:

1. Set \( \text{COLOR} = \text{GOAL} \).

2. (Paint the sibling) If \( \text{GOAL} \) has a sibling and the call for \( \text{COLOR} \) was invoked before the call for \( \text{SIBLING}.C \) then set \( \text{SIBLING}.C = \text{COLOR} \).

3. Goto GET_UNIFYING_CLAUSE.

GET_UNIFYING_CLAUSE:

1. Choose the next clause in the procedure for \( \text{GOAL} \). If no clause remains then goto FAIL. Otherwise, let \( \text{CLAUSE} \) be the chosen clause. Let \( \text{HEAD} \) be its head.

2. (Insert in the tree) If \( \text{CLAUSE} \) is not a child of \( \text{GOAL} \), insert \( \text{CLAUSE} \) as a child of \( \text{GOAL} \). A dummy clause-node representing the sibling of \( \text{CLAUSE} \) is also inserted as a child of \( \text{GOAL} \) if \( \text{CLAUSE} \) is not the last clause in the goal procedure. Insert child goal-nodes of \( \text{CLAUSE} \). A newly inserted node in the tree has its C tag set to no-color and L tag set to nil.

3. If the call for \( \text{CLAUSE}.C \) was invoked before the call for \( \text{COLOR} \) then set \( \text{COLOR} = \text{CLAUSE}.C \).

4. Set \( \text{CLAUSE}.C = \text{no-color} \). (Paint the sibling) If \( \text{CLAUSE} \) has a sibling and the call for \( \text{COLOR} \) was invoked before the call for \( \text{SIBLING}.C \) then set \( \text{SIBLING}.C = \text{COLOR} \).

5. (Does \( \text{CLAUSE} \) begin a redundant search?) If the call for \( \text{CLAUSE}.L \) was invoked before the call for \( \text{COLOR} \) then go back to step 1.

6. If \( \text{GOAL} \) does not unify with \( \text{HEAD} \) then set \( \text{CLAUSE}.L = L(\text{GOAL}, \text{HEAD}) \) and go back to step 1.

7. (\( \text{CLAUSE} \) has unified with \( \text{GOAL} \)) Set \( \text{CLAUSE}.L = \text{nil} \). If \( \text{CLAUSE} \) has no child goal then goto EXIT.

8. (Traverse to the the leftmost goal in the unifying clause) Select leftmost goal in \( \text{CLAUSE} \) as the next \( \text{GOAL} \) and goto CALL.

EXIT:

1. If \( \text{GOAL} \) has a sibling then set \( \text{COLOR} = \text{SIBLING}.C \); select \( \text{SIBLING} \) as the next \( \text{GOAL} \) and goto CALL.

2. Otherwise, set parent goal of \( \text{GOAL} \) as the new \( \text{GOAL} \) and goto EXIT.

FAIL:

1. Compute and assign L tag to \( \text{GOAL} \). This tag is chosen from the L tags of the child clause-nodes of \( \text{GOAL} \) by selecting the one that has invoked its call most recently. If only \( \text{nil} \) and \( \text{GOAL} \) appear as L tags on the child clause-nodes of \( \text{GOAL} \) then L tag assigned to \( \text{GOAL} \) is the goal that was executed just before \( \text{GOAL} \).

2. If \( L \), computed in step 1, is a goal that had its call invoked after the call by the parent goal of \( \text{GOAL} \) then set \( L = \text{L}(\text{GOAL}, \text{HEAD}) \) and goto REDO.

3. Otherwise, compute and assign L tag to the parent clause of \( \text{GOAL} \). This tag is the goal that invoked its call earliest among those that appear as L tags on the children of the clause-node. After assigning the L tag to the clause-node, clear the L tags (set them to nil value) of all children of the clause-node.

4. Set \( \text{GOAL} = \text{the parent goal of GOAL} \) and goto REDO.
The interpreter, when G fails, to goal-node L, if the call for L was invoked after the call for the parent of G. Otherwise, the interpreter is directed to invoke a redo for the parent of G.

A complete description of the algorithm is given in Algorithm 1. The algorithm consists of six procedures: Initialization, Call, Redo, Get-unifying-clause, Exit and Fail. The control is passed between these procedures through goto statements. Tag C on a node specifies the color (goal) of the node. The algorithm uses a global variable, COLOR, to spread colors from the goals to their child goal-nodes. Execution trace of the algorithm when executing the example program is shown in Fig. 1 and Fig. 2.

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.</td>
<td>Assume that initially tree (Fig. 2) has nodes G0, G0, and G1 thru' G4. Let COLOR = G0; GOAL = G1; C0 = GO; G1.C = no_color, for i = 1 thru' 4; and G1.L = nil, for i = 1 thru' 4.</td>
</tr>
<tr>
<td>1.</td>
<td>CALL G1: (1) G1.C = no_color; G1.L = nil; (2) G2.C = GO. GET_UNIFYING_CLAUSE: (1) CLAUSE = C1. (2) Insert C1 and (dummy) C2 as children of G1. (4) G2.C = GO.</td>
</tr>
<tr>
<td>2.</td>
<td>EXIT G1: (1) GOAL = G2.</td>
</tr>
<tr>
<td>4.</td>
<td>EXIT G2: (1) GOAL = G3.</td>
</tr>
<tr>
<td>8.</td>
<td>REDO G3: (1) COLOR = G3. GET_UNIFYING_CLAUSE: (1) CLAUSE = C6. (3) COLOR = GO; (4) C6.C = no_color.</td>
</tr>
<tr>
<td>13.</td>
<td>EXIT G2: (1) COLOR = G2; GOAL = G3.</td>
</tr>
<tr>
<td>15.</td>
<td>EXIT G3: (1) COLOR = G2; GOAL = G4.</td>
</tr>
<tr>
<td>17.</td>
<td>EXIT G4: (1) GOAL = G0.</td>
</tr>
<tr>
<td>18.</td>
<td>EXIT GO: DONE.</td>
</tr>
</tbody>
</table>

Fig. 1: Execution trace of Algorithm 1 for the example Prolog program. The numbers in the parenthesis refer to step numbers in Algorithm 1. For brevity sake, the steps that do not change values or affect the control-flow are not shown in the trace.
4. PERFORMANCE

We have implemented the scheme on a naive interpreter. An algorithm based on the scheme introduced in \( (8) \), is used to compute \( L(G,A) \). The algorithm returns a goal, contributing some binding to \( G \), to which the interpreter must necessarily backtrack to remove a cause of the non-unifiability of goal \( G \) and head \( H \). The algorithm (see Algorithm 2), though less selective than the one described in \( (8) \), is closer to the algorithm implicit in \( (6) \) to construct the suspect sets. Thus, it will allow us to draw meaningful conclusions about the effectiveness of the scheme by comparing the execution statistics. The algorithms used in \( (11) \) and \( (4) \) for constructing the sets of suspects are, however, more selective than the one implicit in our choice.

The statistics collected during the experiments are summarized in Table 1. The columns in the table show the data about the search space, the execution time and the size of the history tree. The search space is defined as the total number of attempted unifications, successful or otherwise, between the goals and the heads of the clauses.

Table 2 compares the speed-ups and the reductions in the search space with some IB schemes. The speed-ups mentioned under scheme KL are based on the number of Harren Abstract Machine (HAM [1]) cycles \( (4) \).
5. CONCLUSIONS

Comparisons show that the scheme results in considerable reduction in the search space. The improvement in execution time, though moderate, is significant. Indeed, most of the overheads -- e.g., maintenance of the tags and construction of the set of suspects -- are common with the IB schemes. As a result, we expect a good reduction in the execution time if the algorithm is used along with an IB interpreter. The more selective is an IB scheme in choosing the backtrack points, the better is the expected improvement from the combined scheme. A better IB scheme is expected to create a larger number of paths over which the interpreter may make redundant searches. Indeed, the algorithm benefits the IB scheme too. Smaller number of paths over which the interpreter makes searches implies that the smaller will be the database of sets of suspects. This, in turn, will reduce the number of spurious backtrack-points -- the backtrack-points that do not result in a successful search.

The scheme can be tailored to trade search space for storage. A goal that has completed an exit step can be deleted, along with its descendants, to recover storage. The price, we pay, is the redundant search that may result when the interpreter returns to a goal that has been removed from the history tree. Perhaps, the compromise lies in the use of a heuristic to choose the goal-node for reclaiming the storage. We experimented with a heuristic that reclaims storage by removing the nodes with no descendant goal-nodes. Even this simple heuristic promises a graceful tradeoff between the search space and the storage for the history tree.

ACKNOWLEDGEMENTS

This work was done when the first author was at AIT. The work was also supported by AIT Project Initiation Grant, 1988-89. The Master's program of the second author was supported by an Australian government scholarship. Graeme Port and Kanchana Kanchanasut helped the authors in improving the clarity and focus of the paper.

REFERENCES

TABLE 2: A comparison with some ib schemes.

<table>
<thead>
<tr>
<th>PROGRAM</th>
<th>SEARCH SPACE (% reduction)</th>
<th>EXECUTION TIME (Speed-up)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KL (6)</td>
<td>Thin</td>
</tr>
<tr>
<td>6-queens (simple)</td>
<td>01</td>
<td>32</td>
</tr>
<tr>
<td>6-queens (clever)</td>
<td>03</td>
<td>38</td>
</tr>
<tr>
<td>Circuit design</td>
<td>73</td>
<td>67</td>
</tr>
<tr>
<td>Database query</td>
<td>64</td>
<td>46</td>
</tr>
<tr>
<td>Binary tree</td>
<td>00</td>
<td>00</td>
</tr>
<tr>
<td>Map color (bad order)</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td>Map color (good order)</td>
<td>06</td>
<td>18</td>
</tr>
</tbody>
</table>

Function unify(g, h; term; p; goal): goal;
begin
if g is an unbound variable
    then begin
        bind g with h;
        remember G as the goal that binds g;
        SPECIAL CASE: if g and h are both variables and g is in a variable in head H
        then instead of binding g with h
        we make g an alias of h;
        return null;
    end;
if h is an unbound variable
    then return unify(h, g, p);
if g is a bound variable
    then begin
        let g' be the term bound to g and
        let q be the goal that binds g to g';
        if q invoked its call after p
        then return unify(g', h, q);
        else return unify(g', h, p);
    end;
if h is a bound variable
    then return unify(h, g, p);
if functor(g) = functor(h)
    then return p;
for each pair gi, hi of corresponding arguments in g, h do
begin
    /* g and h can not unify */
    then return p;
    for each pair gi, hi of corresponding arguments in g, h do
begin
    X := unify(gi, hi, p);
    if X # null
        then return X;
    end;
end;
return null
end.

Algorithm 2: An algorithm to compute L(G, E) -- the suspect that executed the most recent call. The
algorithm is a simplified version of the algorithm given in [8]. For a goal G and head H
(both G and H are treated as terms), L(G, H) =
unify(G, H, G).