Controlling Recursive Inference

David E. Smith, Michael R. Genesereth and Matthew L. Ginsberg
Knowledge Systems Laboratory, Department of Computer Science, Stanford University, Stanford, CA 94305-2095, U.S.A.

Recommended by Eric Sandewall

ABSTRACT
Loosely speaking, recursive inference occurs when an inference procedure generates an infinite sequence of similar subgoals. In general, the control of recursive inference involves demonstrating that recursive portions of a search space will not contribute any new answers to the problem beyond a certain level. We first review a well-known syntactic method for controlling repeating inference (inference where the conjuncts processed are instances of their ancestors), provide a proof that it is correct, and discuss the conditions under which the strategy is optimal. We also derive more powerful pruning theorems for cases involving transitivity axioms and cases involving subsumed subgoals. The treatment of repeating inference is followed by consideration of the more difficult problem of recursive inference that does not repeat. Here we show how knowledge of the properties of the relations involved and knowledge about the contents of the system's database can be used to prove that portions of a search space will not contribute any new answers.

1. Introduction

1.1. Motivation

Consider a system for reasoning about circuits based on descriptions of circuit topology and the functional characteristics of circuit elements. Such a system might need to know that connection between terminals in a circuit is transitive,

$$\mathrm{Conn}(x, y) \wedge \mathrm{Conn}(y, z) \Rightarrow \mathrm{Conn}(x, z), \tag{1}$$

where the proposition $\mathrm{Conn}(x, y)$ means that the point x is electrically connected to the point y in the circuit. The problem with such facts is that they often result in infinite searches. Suppose, for instance, that we want to find *all* of the connections to some point A in a circuit. A portion of the backward AND/OR search tree for this problem is shown in Fig. 1. Applying the

Artificial Intelligence **30** (1986) 343–389

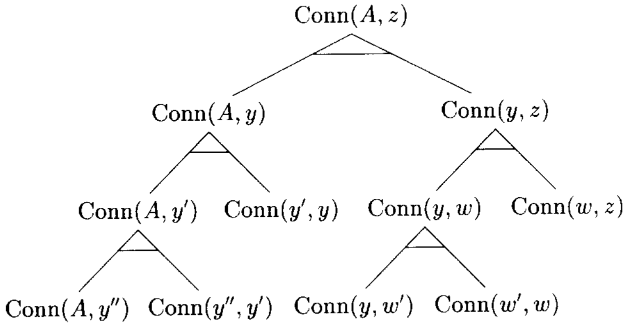

FIG. 1. A portion of the backward search space for the goal Conn(A, z).

transitivity rule to the query Conn(A, z) results in the subgoal Conn$(A, y) \wedge$ Conn(y, z). The transitivity rule can be applied again to both of these conjuncts yielding the subgoals Conn$(A, y') \wedge$ Conn(y', y) and Conn$(y, w) \wedge$ Conn(w, z) respectively. Transitivity applies again to each of these four conjuncts, and so on. For this problem a backward inference procedure would apply the transitivity rule again, and again, and again until it runs out of storage, the user runs out of patience or money, or the machine crashes. We could eliminate the recursion in the connectivity problem simply by discarding all repeating subgoals, but we might sacrifice important answers in the process. For example, suppose that the database contains the facts,

$$\begin{aligned}&\text{Conn}(A, B),\\&\text{Conn}(B, C),\\&\text{Conn}(C, D),\end{aligned}$$

and the query is to find all answers to Conn(A, z) as before. A database lookup on the goal expression yields the solution $z = B$. The remaining two solutions, $z = C$ and $z = D$ both require an application of the transitivity axiom. But after one application of the transitivity axiom, we have a subgoal containing the conjunct Conn(A, y), which matches the original goal. This repeating subgoal is the only possible subgoal for our problem. If it is eliminated, neither of the remaining two answers will be found.

For this example, if we were to allow recursion only one level deep, all of the answers could be found. A solution $y = B$ to the subgoal Conn(A, y) is found using only database lookup. Substituting this binding into the remaining conjunct gives the subgoal Conn(B, z). The solution $z = C$ is found to the original goal, Conn(A, z), by database lookup on this subgoal. Next the transitivity axiom is applied again to this subgoal yielding the subgoal Conn$(B, w) \wedge$ Conn(w, z). Again permiting a single level of recursion, the solution $w = C$ is found by database lookup on the first clause leaving us with the subgoal Conn(C, z). A database lookup on this subgoal yields the final answer $z = D$.

Unfortunately not all recursion can be eliminated this simply. In general, an arbitrary number of levels of recursion may be required to find all answers to a problem. For example, consider an axiom of the form

$$P(x) \wedge R(x, y) \Rightarrow P(y),$$

where we want to find all $P(z)$. A portion of the search space for this example is shown in Fig. 2. Suppose that the database contains the facts

$$\begin{array}{l} P(1), \\ R(1,2), \\ R(2,3), \\ R(3,4), \\ \vdots \end{array}$$

The answer $z = 1$ will be found instantly by database lookup. The answer $z = 2$ requires one level of recursion. The answer $z = 3$ requires two levels, and so on. To find the nth answer requires searching $n - 1$ levels deep in the recursive space.

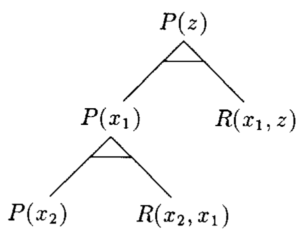

FIG. 2. A portion of the backward search space for the goal $P(z)$.

In this paper we develop methods for deciding how deep a problem solver must go in a recursive space. The amount of the space that must be examined depends, in general, upon the form of the recursion, as well as on the facts that are present in the system's database. But before proceeding, we need to consider some alternatives for dealing with recursion.

1.2. False hopes

It might seem that there are several easy solutions to this problem, such as using breadth-first search, using selective forward inference, or reformulating the troublesome axioms to make the problem go away. In fact, as we will demonstrate in this section, none of these measures work very well.

1.2.1. *Breadth-first search*

Suppose we were to use breadth-first search on a space like that in Fig. 1. Using breadth-first search we are guaranteed to find any answer in the search space (eventually). Unfortunately this does not help if we are looking for *all* answers to a problem. If there is no information about when to stop looking for answers, the entire infinite space must be explored, and breadth-first search will never halt. Even if we are only interested in a specific number of answers, breadth-first search will only halt if the space contains at least that many answers. For example, if we ask for only a single answer to a query, but it turns out there are no answers, breadth-first search will not halt. Thus, breadth-first search alone does not solve the problem of recursive inference.

1.2.2. *Selective forward inference*

In the connectivity example, suppose that forward inference is selectively performed on all facts of the form $\mathrm{Conn}(x, y)$ using the transitivity rule, and the transitivity rule is *not* used for backward inference. The recursion problem would then be eliminated, since the transitivity rule does not cause problems for forward inference. Unfortunately, there are several serious difficulties with this approach. First of all, even the selective use of forward inference can result in the computation and storage of many irrelevant facts. For the example above, forward inference would result in computation of the transitive closure of all connections in the circuit, even though we may only need to know the connections to a few. This would be unacceptable for circuits with high fan-out or fan-in, or for connections to common busses, power supplies, or grounds.

To make matters worse, it is not always possible to limit the use of forward inference to just those rules responsible for recursion.[1] Suppose that, in addition to the transitivity axiom, there is another axiom $P(x, y) \Rightarrow \mathrm{Conn}(x, y)$ in the database, along with the facts $\mathrm{Conn}(A, B)$ and $P(B, C)$. If forward inference is confined to just the transitivity axiom, incompleteness can result. Since there is only one connectivity fact, no conclusions are drawn using forward inference on the transitivity axiom. If one were then to ask for all connections to A, the answer B would be found, but not the answer C. To fix this incompleteness, the axiom $P(x, y) \Rightarrow \mathrm{Conn}(x, y)$ must also be subject to forward inference. More generally, if an axiom is restricted to forward inference, all axioms that can be used in the proof of any of its premise clauses must also be subject to forward inference. Subjecting additional axioms to forward inference further increases the number of potentially irrelevant facts that must be computed and stored.

[1] Minker and Nicolas [19] and Reiter [23] have also pointed this out.

Another problem with forward inference is that it can also lead to infinite deductive chains. Consider the rule for computing Fibonacci numbers:

$$x = \text{Fibonacci}(i-2) \wedge y = \text{Fibonacci}(i-1) \wedge z = x + y$$
$$\Rightarrow \text{Fibonacci}(i) = z. \tag{2}$$

When two Fibonacci numbers are given to a forward inference procedure it would proceed to compute Fibonacci numbers forever. This rule can cause an infinite loop in either a backward or forward inference engine.

Thus, even the selective use of forward inference is not a good solution to the problem of recursive inference.

1.2.3. *Reformulation*

The technique of reformulation is one quite familiar to PROLOG programmers. It involves rewriting the facts available to the inference procedure so that the search space for the goal is no longer infinite, or so that the inference procedure will not discover the recursive portion. As an example of what we mean by reformulation, consider the troublesome transitivity rule (1) for circuit connections, together with the database

$$\text{Conn}(A, B),$$
$$\text{Conn}(B, C),$$
$$\text{Conn}(C, D).$$

Suppose we introduce a new relation IConn, meaning "immediately connected". The database and transitivity rule can then be replaced by the facts

$$\text{IConn}(A, B),$$
$$\text{IConn}(B, C),$$
$$\text{IConn}(C, D),$$

and the two rules

$$\text{IConn}(x, y) \Rightarrow \text{Conn}(x, y),$$
$$\text{IConn}(x, y) \wedge \text{Conn}(y, z) \Rightarrow \text{Conn}(x, z).$$

For this example the answer $z = B$ can be found using the first rule. The reformulated transitivity rule can also be applied, yielding the conjunctive subgoal $\text{IConn}(A, y) \wedge \text{Conn}(y, z)$. By database lookup, the solution $y = B$ can be found to the first clause. Substituting this binding into the remaining conjunct leaves the subgoal $\text{Conn}(B, z)$. The solution $z = C$ can be found using the first rule. The reformed transitivity rule can also be applied again yielding the subgoal IConn(B, y') $\wedge$ Conn(y', z), and so on. The search space for this problem is shown in Fig. 3. With this reformulation, we have eliminated the recursion on all left-hand branches of the tree. As a result, this reformulation does not lead to an infinite search, so long as the IConn conjunct is solved before the Conn conjunct.

There are several serious problems with reformulating information in this way. First of all, reformulations generally only work well for a few of the possible forms that a query might take. For example, the above reformulation works well for the query Conn(A, z) but it does not work well for the query Conn(x, D). On applying the reformulated version, we would get the subgoal IConn(x, y) $\wedge$ Conn(y, D). If the IConn conjunct is expanded first we end up searching through all of the immediate connections in the circuit, a horribly inefficient process in a large circuit. Alternatively, if the Conn conjunct is expanded first, we again end up with an infinitely repeating search space. The dual reformulation,

$$\begin{aligned}&\text{IConn}(x, y) \Rightarrow \text{Conn}(x, y),\\ &\text{IConn}(y, z) \wedge \text{Conn}(x, y) \Rightarrow \text{Conn}(x, z),\end{aligned}$$

works well for the query Conn(x, D), but performs miserably for the query Conn(A, z). Neither of these reformulations is reasonable if both kinds of queries are expected, as might be the case for an asymmetric relation. In general, reformulations only work effectively for some subset of the possible queries covered by the original domain knowledge.

A second problem with reformulations is that they are often fragile. For the example above, adding the fact IConn(B, A) to the database would again lead to a loop. As before, the subgoal Conn(B, z) can be generated from the goal Conn(A, z), but using the new fact, the subgoal Conn(A, z) can be generated from the subgoal Conn(B, z).[2]

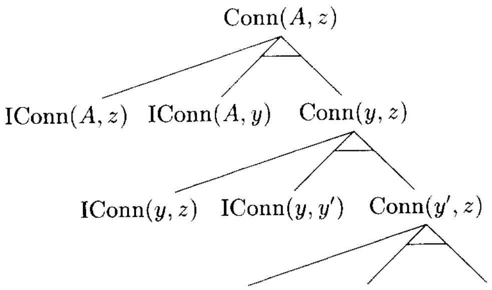

FIG. 3. Reformulated search space for the goal Conn(A, z).

[2]We are indebted to an anonymous reviewer for this observation.

A third problem with reformulation is that it can be an arbitrarily difficult programming task. Suppose that, in addition to the transitivity rule for connections, we also have the symmetry rule

$$\mathrm{Conn}(x, y) \Rightarrow \mathrm{Conn}(y, x).$$

The reformulation now requires four facts, and another intermediate relation, TConn.

$$\begin{aligned}
&\mathrm{IConn}(x, y) \Rightarrow \mathrm{TConn}(x, y),\\
&\mathrm{IConn}(x, y) \wedge \mathrm{TConn}(y, z) \Rightarrow \mathrm{TConn}(x, z),\\
&\mathrm{TConn}(x, y) \Rightarrow \mathrm{Conn}(x, y),\\
&\mathrm{TConn}(x, y) \Rightarrow \mathrm{Conn}(y, x).
\end{aligned}$$

It is not so obvious that this reformulation covers all of the possible cases.

As an even more problematic example, consider the recursive rule that states that a person will be an albino if both his parents are albinos.

$$\begin{aligned}
&\mathrm{Albino}(x) \wedge \mathrm{Parents}(z) = \{x, y\} \wedge \mathrm{Albino}(y)\\
&\quad \Rightarrow \mathrm{Albino}(z).
\end{aligned} \tag{3}$$

Suppose that the query is to find all albinos:

$$\text{find all } z\text{:}\quad \mathrm{Albino}(z).$$

Expanding the Parents conjunct first would result in an unacceptable search through all parent/child pairs. Alternatively, expanding either of the Albino conjuncts first would result in an infinite repeating search space. By indulging in *knowledge programming*, we could reformulate this rule so that depth-first backward inference results in an efficient search of the space for this query. As a first step we introduce the new predicate GivenAlbino(x) to refer to those individuals given as albinos initially. The first two rules in (4) below state that any given albino is an albino, and that any nth generation descendent of a given albino (along albino lines) is also an albino. We still need to define what it means for an individual to be of albino descent from a given albino. The third rule states that an individual is of albino descent from a given albino if the given albino is a parent of the individual, and the other individual's parent is an albino. The fourth rule simply expresses the transitive closure of this relationship, that an individual is of albino descent from a given albino if that individual is of albino descent from the given's albino children. The final two rules are for checking whether a given individual is an albino and are a simple reformulation of the original albino rule (3).

$$\begin{aligned}
&\text{GivenAlbino}(z) \Rightarrow \text{Albino}(z);\\
&\text{GivenAlbino}(x) \wedge \text{AlbinoDescent}(z, x) \Rightarrow \text{Albino}(z);\\
&\text{Parents}(z) = \{x, y\} \wedge \text{CheckAlbino}(y) \Rightarrow \text{AlbinoDescent}(z, x);\\
&\text{Parents}(w) = \{x, y\} \wedge \text{CheckAlbino}(y) \wedge \text{AlbinoDescent}(z, w)\\
&\quad \Rightarrow \text{AlbinoDescent}(z, x); \qquad (4)\\
&\text{GivenAlbino}(z) \Rightarrow \text{CheckAlbino}(z);\\
&\text{Parents}(z) = \{x, y\} \wedge \text{CheckAlbino}(x) \wedge \text{CheckAlbino}(y)\\
&\quad \Rightarrow \text{CheckAlbino}(z).
\end{aligned}$$

Performing depth-first backward inference on this reformulation effectively results in forward inference from given albinos to their progeny, and backward inference at each step to verify that the other parent of the progeny is also an albino. Note that this backward portion of the inference is accomplished using a simple reformulation of the original albino rule (3). As with the connectivity example, this reformulation only works efficiently for the query Albino(z), where one or more albinos are desired. It does not work well for checking whether a given individual is an albino.

From these examples, we can see that there are several serious disadvantages to reformulation as a method of controlling recursive inference. First, the resulting knowledge programs only work effectively for some subset of the possible queries covered by the original domain knowledge. Second, the programs may be fragile, in that additional data can reintroduce recursion. Third, it may be an arbitrarily difficult programming task to do such a reformulation. Finally, it is more difficult to understand, explain, and modify reformulations. Reformulation results in an implicit embedding of control information into the domain information. Instead of having facts about the domain and facts about control, the two are merged into knowledge-rich programs for a given interpreter. This has little advantage over building expert systems in more traditional programming languages like LISP or PASCAL. Many authors have argued against reformulation for exactly these reasons [3–5, 9, 10, 17].

1.3. Definitions

So far we have relied on the reader's intuitions and the examples to indicate what we might mean by the term recursive inference. We now give a precise definition.

Let the term *goal set* refer to the set of all conjuncts for a conjunctive goal in a search space. We say that one goal set g' is a *descendant* of another goal set g, if there is some sequence of goal sets beginning with g and ending with g', such that each goal set in the sequence is a subgoal of its predecessor. An *inference path* in a search space is a sequence of goal sets in the space such that each goal set in the sequence is an (immediate) subgoal of the preceding goal set. For example, the sequence

$$\begin{array}{l} \{\mathrm{Conn}(A, z)\}, \\ \{\mathrm{Conn}(A, y),\mathrm{Conn}(y, z)\}, \\ \{\mathrm{Conn}(A, y'),\mathrm{Conn}(y', y),\mathrm{Conn}(y, z)\}, \\ \quad\vdots \end{array} \tag{5}$$

is an inference path for the connectivity problem.

We use the notation $p|_b$ to refer to the expression formed by substituting the variable bindings b into the expression p. Using this notation, an expression c' is said to be an *instance* of an expression c if there is a substitution (a set of variable bindings) b for the variables in c such that $c' = c|_b$.

Definition 1.1. An inference path is *recursive* if there is an infinite subsequence $\langle g_1, \ldots, g_i, \ldots \rangle$ of the goal sets in the path and a distinguished clause c_i in each goal set g_i such that,

(1) c_i is an instance of c_{i-1}, and

(2) c_i is in the subset of g_i that are descendants of c_{i-1}.

An inference procedure generating any recursive inference path is said to be involved in recursive inference.

As an example, consider the infinite inference path (5) for the connectivity problem. The conjunct Conn(A, y) in the second goal set is an instance of the conjunct Conn(A, z) in the first goal set. The descendants of Conn(A, z) constitute the entire set {Conn(A, y),Conn(y, z)}, which contains Conn(A, y). Likewise, the conjunct Conn(A, y') in the third goal set is an instance of the conjunct Conn(A, y) in the second goal set. The descendants of Conn(A, y) are the subset {Conn(A, y'),Conn(y', y)}, which contains Conn(A, y'). Thus, with g_i as the ith goal set in the inference path and c_i as the first conjunct in each goal set, the inference path satisfies the definition for a recursive path.

The definition of recursive inference given above actually covers a much broader class of problems than we have considered so far. For example, the definition includes recursive paths where there are intermediate descendants in between those descendants with repeating conjuncts. The definition also includes paths where the repeating conjunct may have its variables bound before it is actually processed. Consider the simple axiom

$$y = x + 1 \wedge \mathrm{Integer}(x) \Rightarrow \mathrm{Integer}(y) \tag{6}$$

with the query Integer(2.5). One inference path for this problem is shown in

Fig. 4. When the Integer conjuncts are expanded, they are each different, since the variable x is already bound. The subsequence of goals

$$\begin{array}{l} \text{Integer}(2.5), \\ \text{Integer}(1.5), \\ \text{Integer}(0.5), \\ \quad\vdots \end{array}$$

does not repeat. However, the subsequence

$$\begin{array}{l} 2.5 = x + 1 \wedge \text{Integer}(x), \\ 1.5 = x + 1 \wedge \text{Integer}(x), \\ 0.5 = x + 1 \wedge \text{Integer}(x), \\ \qquad\vdots \end{array}$$

satisfies our definition for recursive inference. Each member contains the conjunct Integer(x), which is both an instance and a descendant of the preceding Integer(x) conjunct.

Although both the integer example and the connectivity example constitute recursive inference, there is an important difference between the two examples. Consider the sequence made up of the conjuncts actually processed by the inference engine at each step. For the connectivity example this sequence repeats:

$$\begin{array}{c} \text{Integer}(2.5) \\ | \\ 2.5 = x + 1 \ \wedge \text{Integer}(x) \\ | \\ \text{Integer}(1.5) \\ | \\ 1.5 = x + 1 \ \wedge \text{Integer}(x) \\ | \\ \text{Integer}(.5) \\ | \\ .5 = x + 1 \ \wedge \text{Integer}(x) \\ \vdots \end{array}$$

FIG. 4. Inference path for the query Integer(2.5).

$$\begin{aligned}&\mathrm{Conn}(A, z),\\ &\mathrm{Conn}(A, y),\\ &\mathrm{Conn}(A, y'),\\ &\quad\vdots\end{aligned}$$

In other words, the repeating conjuncts $\mathrm{Conn}(A, \phi)$ are instances of their predecessors at the time they are actually reduced to subgoals. We refer to such recursive inference as *repeating inference*.

In contrast, the sequence of conjuncts

$$\begin{aligned}&\mathrm{Integer}(2.5),\\ &\mathrm{Integer}(1.5),\\ &\mathrm{Integer}(0.5),\\ &\quad\vdots\end{aligned}$$

does not repeat. The argument of $\mathrm{Integer}(x)$ is always bound before the conjunct is actually reduced to subgoals, and each time the argument is bound to a different constant.[3] We refer to this nonrepeating recursive inference as *divergent* inference.

1.4. The approach

Control of recursive inference means eliminating those portions of the search space that are *superfluous* or *redundant*. We say that a goal is *superfluous* if there are no facts in the database that will satisfy it or any of its descendants. For a particular problem we say that a goal g is *redundant* with another goal g' if none of its descendants will result in any solutions to the problem not produced by descendants to the goal g'. The difficulty is to determine which branches of a search space are indeed superfluous or redundant. If all recursive inference were unproductive it would be a simple matter to provide effective control. However, as we illustrated with some of the examples in Section 1.1 there are many instances where a limited amount of recursive inference is necessary in order to arrive at desired answers. If too much of a recursive space is discarded, important answers to the problem are lost. Alternatively, if not enough of the recursive space is discarded, valuable problem solving effort is wasted.

In general, it is not decidable whether or not a given portion of a recursive search space is redundant. However, there are special cases where it is possible to prove redundancy without completely exploring the space. For repeating

[3]The latter qualification is important. It is still possible for the sequence of goals processed to be repeating even though all of the arguments are bound. We will show examples of this in Section 3.

inference, a simple syntactic solution is possible. We can decide when to cut off inference by keeping track of the answers produced with each additional level of repetition. For divergent inference the problem is much harder. Here we must generate automatic proofs that no answers exist in a portion of the search space. These proofs are similar to proofs of program termination using well-founded sets. They require information about the properties of the relations involved, and about the content of the system's database. Finally, where rule sets are commutative and each set alone cannot produce answers, it is possible to generate automatic proofs that no novel answers will appear in a portion of the search space, again by making use of knowledge about the properties of the relations involved, and about the contents of the system's database.

1.5. Organization

In the next section we consider the types of facts that make recursive inference possible, and consider the conditions under which recursive inference will actually occur. The reader more interested in a solution to the problem of recursive inference can skip ahead to Sections 3 and 4 and refer back as needed to Section 2. In Section 3, techniques for the common special case of *repeating inference* are reviewed. Although several of the algorithms presented are not novel, we consider them from the viewpoint of search control, introduced in Section 1.4. We provide a proof that the methods are correct and consider the conditions under which the pruning strategies are optimal. In addition, powerful methods for dealing with the special cases of transitivity and logical subsumption are described.

The more general class of nonrepeating recursive inference is considered in Section 4. Here we show how properties of the relations involved and knowledge about the contents of the system's database can be used to demonstrate that a portion of the search space is redundant. Finally, in Section 5 we consider the problem of detecting recursive inference so that control can be instituted only when necessary. Related work is also discussed.

2. The Conditions for Recursive Inference

2.1. Cyclic and recursive collections

Suppose that a set of facts can be arranged in the form,

$$
\begin{array}{ll}
F_1: & L_1 \wedge \phi_1 \Rightarrow L'_2, \\
F_2: & L_2 \wedge \phi_2 \Rightarrow L'_3, \\
 & \vdots \\
F_{n-1}: & L_{n-1} \wedge \phi_{n-1} \Rightarrow L'_n, \\
F_n: & L_n \wedge \phi_n \Rightarrow L'_{n+1},
\end{array}
$$

where the consequent, L'_{i+1} of each fact F_i unifies with the premise L_{i+1} in its successor F_{i+1}, and the consequent, L'_{n+1}, of the final rule F_n unifies with the premise L_1 in the first rule F_1. We say that such a set of rules forms a *cycle* and constitutes a *cyclic collection*.[4] For example the set of rules,

$$P(x) \Rightarrow Q(x),$$
$$Q(B) \Rightarrow P(A)$$

form a cycle, since $Q(x)$ unifies with $Q(B)$ and $P(x)$ unifies with $P(A)$. A rule can be involved in more than one cycle, so we also refer to the union of any two cyclic collections that share rules as a cyclic collection.

If a common set of bindings is possible for all of the unifications in a cycle, the facts are said to be *recursive* and constitute a *recursive collection*. In other words, a group of facts is recursive if there is some common set of bindings b for the variables in each of the facts F_1 through F_n such that $L_i|_b = L'_i|_b$ and $L'_{n+1}|_b = L_1|_b$. (The notation $P|_b$ refers to the clause P under the variable bindings b.) The cyclic collection given above is not a recursive collection because x cannot be bound to both A and B simultaneously. However, the cyclic collection

$$P(x) \Rightarrow Q(x),$$
$$Q(y) \Rightarrow P(y)$$

is a recursive collection since the binding $x : y$ unifies $Q(x)$ with $Q(y)$ and $P(x)$ with $P(y)$. Likewise, the transitivity and symmetry rules, the albino rule, and the Fibonacci rule given in the previous section are all recursive collections.

As with cyclic collections, it is possible for a single rule to be part of more than one recursive collection. We therefore refer to the union of any two recursive collections that share rules as a recursive collection.

2.2. Recursive search spaces

We say that a search space is *recursive* if it contains a recursive inference path (Definition 1.1). It should come as no surprise that recursive collections give rise to recursive search spaces.

Theorem 2.1. *For any recursive collection of facts there is at least one goal that will result in a recursive search space.*

Proof. By the definition of a recursive collection, there is some common set of bindings b for the variables in each of the facts F_1 through F_n such that each $L_i|_b = L'_i|_b$ and $L'_{n+1}|_b = L_1|_b$. From the goal proposition $L_1|_b$, using the rules

[4] This notation and terminology is derived from [19]. There, Minker and Nicolas express these definitions in terms of facts in conjunctive normal form. For simplicity we have expressed these definitions in terms of rules.

$R_1, \ldots, R_n$ it is possible to regenerate the subgoal $L_1|_b$. This process can be repeated arbitrarily many times, resulting in an arbitrarily long inference path. □

As an example, consider the transitivity rule for circuit connections. The search space in Fig. 1 shows that the goal Conn(A, z) has a recursive space since each subgoal in the sequence

$$\langle \mathrm{Conn}(A, z), \mathrm{Conn}(A, y), \mathrm{Conn}(A, y'), \mathrm{Conn}(A, y''), \ldots \rangle$$

is an instance of the preceding goal.

Corollary 2.2. *If a goal proposition g results in a recursive search space for a given recursive collection, any generalization g′ of the goal will also result in a recursive search space.*

By a generalization of a proposition g, we mean a proposition g' such that $g'|_b = g$ for some set of bindings b. For example, since the query Conn(A, z) has a recursive search space, the query Conn(x, z) will also have a recursive search space.

It is natural to ask whether recursive collections are the only kinds of facts that can lead to infinite search spaces. Infinite search spaces can always occur if there is an an infinite database, but barring this possibility, the answer appears to be yes.

Conjecture 2.3. If an infinite path exists in the search space, and the database is finite, there must be a recursive collection of facts involved in the generation of that path.

In fact, we believe that a stronger statement holds.

Conjecture 2.4. A set of n axioms that is not a recursive collection can generate an inference path of at most length $(2^n - 2)a + 1$ where a is the maximum arity (number of arguments) of all relations in the collection.[5]

Lewis [13] has proven a weaker theorem, but we are not aware of any proof of these conjectures.

2.3. Recursive inference

As we stated in Section 1.3, recursive inference occurs when an inference procedure follows a recursive path in a recursive search space. By this definiton a recursive search space is a necessary condition for recursive inference, but it is not a sufficient condition. Thus, even though a given problem may have a

[5]We arrived at the formula $(2^n - 2)a + 1$ by empirical generalization of a set of examples, beginning with the cyclic collection, $P(y, x) \Rightarrow Q(x, y)$, $Q(A, x) \Rightarrow P(B, x)$, and progressing to higher arity, more rules, and rules involving functional expressions.

recursive search space, recursive inference will not necessarily result. The inference procedure must also happen on to a recursive path. Consider the reformulated version of the transitivity axiom for circuit connections (Section 1.2.3). Although the search space for the goal Conn(A, z) is still a recursive space, if the IConn conjunct is always solved first, recursive inference will not result for this goal and the database of connections given.

In general, whether or not recursive inference occurs depends upon
- the characteristics of the recursive collections involved,
- the facts available in the database,
- the search strategy employed by the inference procedure,
- the strategy for evaluating embedded functional expressions (whether they are evaluated or treated syntactically),
- the number of answers desired for the problem, and
- the number of answers actually available for the problem.

The first of these criteria, the characteristics of the recursive collection, influences the shape of the search space and therefore affects the chances of recursive inference. The facts in the database can also affect the likelihood of recursive inference, as we saw with the reformulated version of the transitivity axiom for connections (3). In that example, the presence of the facts IConn(A, B) and IConn(B, A), cause a loop, but either fact alone will not.

The search strategy also affects whether or not recursive inference will occur. If nonrecursive subgoals are preferred to recursive subgoals, the chances of recursive inference will be less. This is because the inference procedure might be able to find enough answers without ever exploring a recursive portion of the space. Likewise, if nonrecursive clauses are preferred to recursive clauses in conjunctive subgoals, the chances of recursive inference will be less. This is because the nonrecursive clauses may fail, stopping recursion.

Finally, recursive inference becomes more likely as the ratio of number of solutions sought to number of solutions available increases. If more answers are required, a larger percentage of the space must be searched, making it more likely that recursive paths will be explored.

Unfortunately, there is no simple precise characterization of when recursive inference will or will not occur. Any such characterization would require a classification of all the different possibilities for each factor, and a multi-dimensional table to consider all of the different combinations. When a recursive collection is present, there is always the potential for recursive inference, although, as we have seen, it can sometimes be avoided by careful search.

3. Repeating Inference

As indicated in Section 1.3, repeating inference occurs when the sequence of processed goal conjuncts actually repeats. More precisely, there is some infinite subsequence of the goal conjuncts processed, such that each successive conjunct is an instance of its predecessor. Most of the examples that we considered in Section 1 were of repeating inference. In particular, the connectivity example had this characteristic, since a goal expression of the form $\mathrm{Conn}(A, z)$ is generated and expanded repeatedly in the leftmost branch of the AND/OR search tree.

In repeating inference, a portion of the AND/OR search space is repeated over and over again. To control the search we must determine the level at which the repetition can be cut off. The search space below the cutoff point must not hold any new answers.

3.1. Finding a single answer

First consider the special case where only a single answer is needed for a query. In such cases, if an answer cannot be found without exploring a repeating portion of the space, no answer can be found at all. As a result, the search space can be pruned drastically.

Theorem 3.1. *If only a single answer is needed for a goal g, any subgoal g′ that is an instance of g can be discarded (along with the entire subspace descending from g′). Furthermore, it is* optimal *to do so, in the sense that the simplest proof of an answer for g will not involve a repeating subgoal g′.*

Some definitions are needed to prove this result. An answer for a goal expression g consists of a set of variable bindings, i.e. a substitution list, such that the steps in the search space, when reversed, would constitute a proof of the goal expression with those bindings substituted in. We refer to that portion of the search space as a proof tree for that particular answer. We will use the operator $\circ$ to refer to the composition of two binding lists, e.g.

$$\{x : z,\ y : B\} \circ \{z : A,\ w : C\} = \{x : A,\ y : B,\ w : C\}.$$

We say that one proof tree is *isomorphic* to another if they are identical up to variable bindings for the goals and subgoals. Note that if t is a proof tree for a goal $g|_b$ there is an isomorphic proof tree for the generalization g, constructable using the same axioms and rules of inference.

Proof. Let t be a proof tree for the goal g that contains a repeating subgoal g'. Let g'' be the deepest repeating subgoal in the proof tree, and let t'' be the subproof tree for g''. Since g'' is the deepest repeating subgoal, t'' contains no repeating subgoals. Since $g'' = g|_b$, t'' is also a proof tree for $g|_b$. So there is also a proof tree t^* for g that is isomorphic to t''. Since there is a nonrecursive way of finding an answer to g, g' can be discarded. Furthermore, since t^* is isomorphic to a proper subtree of t, t will never be the simplest proof. Thus, the simplest proof will never involve repeating subgoals. □

Note that in the statement and proof of this theorem we assumed nothing about the relative cost or simplicity of proofs, except that if one proof is isomorphic to a subproof of another, it is simpler.

There is a useful corollary of the above theorem.

Corollary 3.2. *Repeated ground queries and functional queries can always be pruned from a search space.*

This result holds because functional queries never have more than one answer.

Finally, note that Theorem 3.1 does not mean that all repetitions can be discarded, only those for goals that require only one solution. Consider the hypothetical search space in Fig. 5. The goal g, which has only a single solution, generates a conjunctive descendant $h \wedge j$. It might be necessary to search through several of the answers to the conjunct h in order to find a solution to the conjunction. Thus, while any reoccurrences of g can be discarded, reoccurrences of the goal h cannot be.

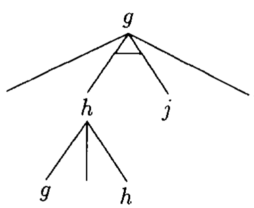

FIG. 5. Search space for a single-solution problem.

3.2. Finding multiple answers

In cases where more than one answer is needed, Theorem 3.1 does not apply. Such problems arise far more often than might be expected. Even though only a single answer is needed for a problem, some of its subproblems may be conjunctive, as in the example above. Solving a conjunction frequently requires generating more than one solution to at least one of the conjuncts.

3.2.1. *The theory*

If multiple answers are needed, in order to eliminate a portion of the repeating space we must show that that portion of the space is *redundant* (i.e. will not produce any novel answers to the original problem). What makes such a proof possible is the observation that if a search of one or more levels of repetition deeper in a recursive space does not produce any new answers, no amount of additional search will produce any new answers to the original repeated supergoal.

Some notation is needed in order to state this theorem precisely and prove that it is correct. Let $S(g)$ refer to the search space beginning with the goal g and containing all of the legal descendants of the goal g. A *frontier set* F of a search space $S(g)$ is defined to be a set of goals in the space such that no goal in the frontier set is a descendant of any other goal in the frontier set. Intuitively, a frontier set is some possibly jagged, partial slice through a search space. Let $S_F(g)$ refer to that portion of the space $S(g)$ not including any of the frontier goals $f \in F$ or their descendants $S(f)$. In other words, the restricted search space $S_F(g)$ is just $S(g)$ with all of the frontier branches pruned out. Let $A_F(g)$ refer to the set of answers to the goal g present in the restricted space $S_F(g)$.

For a recursive space let $R_n(g)$ refer to the frontier set consisting of the nth level repetitions of the goal g.

Theorem 3.3. *Let F be the frontier set $R_n(g)$ consisting of nth-level instances of the goal g. Let F' be a frontier set consisting of repeating descendants of goals in the set F. If $A_{F'}(g) = A_F(g)$, all of the frontier subspaces $S(r)$ for $r \in F$ are redundant.*[6]

Proof. The proof is by induction on the level of repetition in the search space. First we prove the theorem for the case where $F' = R_{n+1}(g)$.

Let g' be a first-level repeating descendant of g and let c be the set of bindings such that $g' = g|_c$. Let r be the subset of $R_n(g)$ that are descendants of g' as shown in Fig. 6. Thus $r = R_{n-1}(g')$. Let r' be the set $R_1(r) = R_n(g')$ (all first-level repeating descendants of r) and let r'' be the set $R_2(r) = R_{n+1}(g')$ (all second-level repeating descendants of r).

The space $S_{r'}(g')$ is an instance of a portion of the space $S_F(g)$. In fact,

$$A_{r'}(g') = \{a : c \circ a \in A_F(g)\}.$$

Likewise the space $S_{r''}(g')$ is an instance of a portion of the space $S_{F'}(g)$ so

$$A_{r''}(g') = \{a : c \circ a \in A_{F'}(g)\}.$$

Since $A_F(g) = A_{F'}(g)$, it follows that $A_{r'}(g') = A_{r''}(g')$, i.e. there are no additional answers to g' available by going a level deeper.

[6]This theorem relies on the assumption of complete indexing in the problem solver's database. In other words, the system must be able to find any fact in the database that matches a goal. Without complete indexing, answers could be found to an instance of a goal when they could not be found for the original goal. A weaker version of the theorem still holds if complete indexing of the problem solver's database is not assumed. In this case, the frontier set F must contain repetitions instead of instances of the goal g. Essentially, this means that search must be a few recursion levels deeper until a specialization of the initial goal is found for which F will contain pure repetitions.

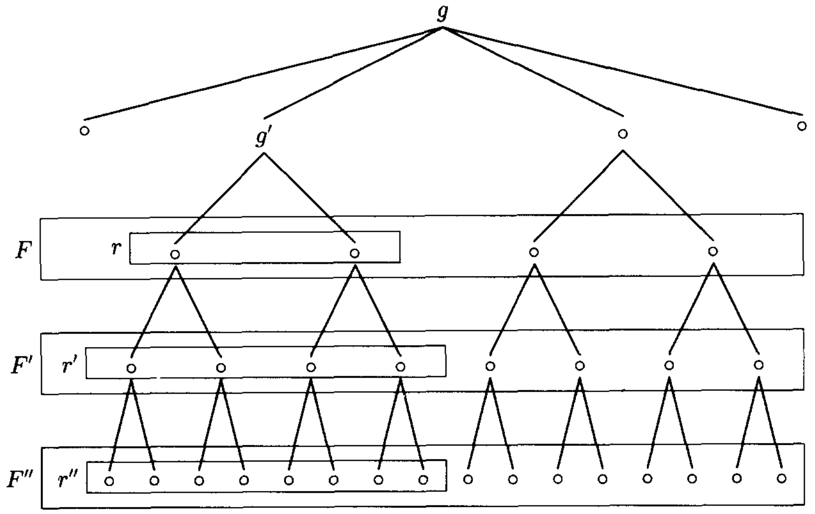

FIG. 6. Abstract repeating search space.

Let $b_{g'}$ be the set of bindings connecting answers to the subgoal $g' \in R_1(g)$ to answers to the supergoal g. In other words, if a' is an answer to g', $b_{g'} \circ a$ is an answer to g. Then,

$$A(g) = A_{g'}(g) \cup \{a : a = b_{g'} \circ a' \wedge a' \in A(g')\}.$$

Now consider the frontier $F'' = R_{n+2}(g)$. Using the two results above,

$$\begin{aligned} A_{F''}(g) &= A_{R_1(g)}(g) \cup \bigcup_{g' \in R_1(g)} \{a : a = b_{g'} \circ a' \wedge a' \in A_{r''}(g')\} \\ &= A_{R_1(g)}(g) \cup \bigcup_{g' \in R_1(g)} \{a : a = b_{g'} \circ a' \wedge a' \in A_{r'}(g')\} \\ &= A_{F'}(g). \end{aligned}$$

By induction $A_{F^{(k)}}(g) = A_F(g)$ for all k. Thus, $A(g) = A_F(g)$, which means that the repeating descendants in F are redundant.

Finally, for any set F' satisfying the requirements of the theorem, $S_{R_{n+1}(g)}(g) \subseteq S_{F'}(g)$, so $A_{F'}(g) = A_F(g)$ implies that $A_{R_{n+1}(g)}(g) = A_F(g)$. Since the theorem holds for $F' = R_{n+1}(g)$, it holds for arbitrary F'. □

Corollary 3.4. *The depth of repetition in a search space can be limited to one less than the total number of answers desired for the problem.*

Theorem 3.1 is a special case of this corollary.

Example 3.5. Consider the connectivity axiom for circuits,

$$\mathrm{Conn}(x, y) \wedge \mathrm{Conn}(y, z) \Rightarrow \mathrm{Conn}(x, z).$$

As before, suppose that the problem is to find all points in a circuit connected to a given point A,

$$\text{find all } z\colon \quad \mathrm{Conn}(A, z).$$

An initial portion of the backward AND/OR search space for this problem is reproduced in Fig. 7. If there are no answers in the system's database for Conn(A, z), there are no answers at all. In this case the frontier sets F = {"Conn(A, z)"} and F' = {"Conn(A, y)"} satisfy the conditions of Theorem 3.3. $S_F(g)$ is the null space and $S_{F'}(g)$ is the space consisting of only the goal g = "Conn(A, z)". Since there are no answers in the database for the goal g, $A_F(g) = A_{F'}(g) = \emptyset$. As a result, Theorem 3.3 states that no search is necessary for the problem.

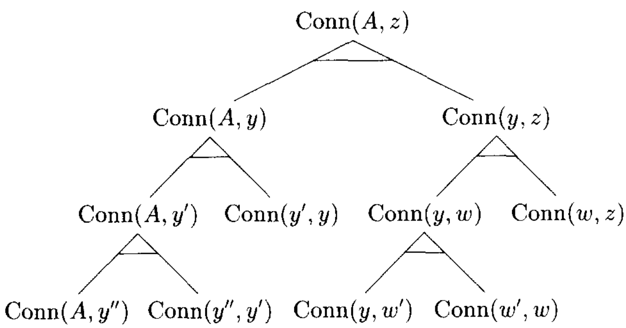

FIG. 7. A portion of the backward search space for the goal Conn(A, z).

If, instead, the database contains the fact Conn(A, B), but no facts about the connections to B, the sets

$$F = \{\text{"Conn}(A, y)\text{"}\} \quad \text{and} \quad F' = \{\text{"Conn}(A, y')\text{"}\}$$

satisfy the theorem. In this case, $S_F(g)$ and $S_{F'}(g)$ both contain the single answer $z = B$. As a result, only database answers to the initial goal Conn(A, z) need be located in this case.

Finally, suppose that the database contains the facts Conn(A, B) and Conn(B, C) but no other connections to A, B, or C. For this case, the cutoff frontiers contain two terms since the right-hand branch of the conjunction also contributes a recursive branch for the binding $z = B$:

$$F = \{\text{“Conn}(A, y')\text{”}, \text{“Conn}(B, w)\text{”}\}$$

$$F' = \{\text{“Conn}(A, y'')\text{”}, \text{“Conn}(B, w')\text{”}\}$$

For both of these frontiers, the answer set will consist of $z = B$ and $z = C$.

Optimality. Although Theorem 3.3 tells us some conditions under which a portion of the search space is redundant, it does not tell us that pruning the redundant portion of the space is necessarily the best thing to do. In some cases it can be advantageous to search part of the redundant portion of the space. As an example, consider the connectivity example of the previous section, where the available facts were Conn(A, B) and Conn(B, C). Suppose that we also have the (nonrecursive) collection of facts

$$\begin{aligned} H(x, y) &\Rightarrow \text{Conn}(x, y), \\ G(x, y) &\Rightarrow H(x, y), \\ F(x, y) &\Rightarrow G(x, y), \\ E(x, y) &\Rightarrow F(x, y), \end{aligned}$$

together with the facts $E(A, B)$ and $E(A, C)$. In this case, we could find all the answers to the query Conn(A, z) by exploring this nonrecursive path. Theorem 3.3, therefore, allows us to conclude that the goal Conn(A, y) is redundant. However, the nonrecursive way of finding the answer $z = C$ is longer and probably more costly than finding the same answer by exploring a level deeper in the repeating space. As a result, pruning the subgoal Conn(A, z) would not be the best thing to do in this case.

In the case where all of the solutions are needed to a problem, we can show that pruning the redundant portion is a good idea.

Theorem 3.6. *For recursive problems where all of the solutions are sought (and the number of solutions is not known), if there exists a frontier F that obeys the conditions of Theorem* 3.3 *it is* optimal *to prune the frontier goals F, in the sense that the amount of search required to find all answers can only be reduced by this pruning.*

Proof. In order to find all answers in a space, all portions that may contain novel answers must be searched. Assuming that we do not know which portions of $S_F(g)$ are redundant with $S(r)$ for each $r \in F$ then, $S_F(g)$ must be searched in any case. If $S_F(g)$ must be searched, each of the $S(r)$ contain only redundant answers, so there is no advantage to searching any of them. As a result, the amount of search necessary to find all answers can only be reduced by pruning the goals in F. $\square$

As we demonstrated in Example 3.5, this result does not hold for problems where some specific number of solutions is sought.

3.2.2. *Repetition cutoff algorithms*[7]

In order to make use of Theorem 3.3 we need a mechanism for finding the repetition level that satisfies the conditions of the theorem. Finding such a frontier set requires preserving the answers to any goal with repeating descendants.

Algorithm 3.7.
(1) If a goal g_i is an instance of one of its supergoals g, the goal g_i is suspended until all other alternatives for solving g have been exhausted.
(2) If any new answers are found to the goal g, all repeated instances g_i of g are enabled for another level of expansion. If not, the inference is terminated.

A flowchart of a problem solver incorporating this procedure appears in Fig. 8. One major efficiency improvement can be made on this procedure. The answers produced by expanding the search space an additional level of repetition will be a subset of those produced in the first level since each repeated descendant g_i is an instance of the goal g. Therefore, it is not necessary to reproduce the space at each level. It is sufficient to cache all of the answers to the supergoal and use them as the answers to any repeated descendants. Thus, a more efficient procedure would be:

Algorithm 3.8.
(1) Each time a solution is found to a query (or subquery) the solution is cached.
(2) When a repeated descendant is encountered, only instances found in the system's database (including cached answers) are used as solutions to the repeated descendant. No additional inference is performed on this repeated descendant.
(3) The solution of a repeated descendant is not complete until no additional solutions can be found to the goal that it is a repeat of. In other words, new answers to a goal must continually be plugged into all repeated descendants until quiescence occurs and no new answers appear.

As an illustration of this method, consider the search tree shown in Fig. 9. This tree is a snapshot of the goal stack for the inference engine at some point in the computation. There are two repetitions of the original goal expression g, both of which are suspended awaiting answers to g. If the answer α is found to g, this answer is cached and consequently plugged in as an answer for g' and g''. If these branches generate additional answers α' and α'' to g, then these answers, in turn, are cached and must be tried in the two repeated descendants. When no new answers to g can be produced the process is complete.

[7]These algorithms were first discovered by Black [1] and were later rediscovered by McKay and Shapiro [18] and by the authors.

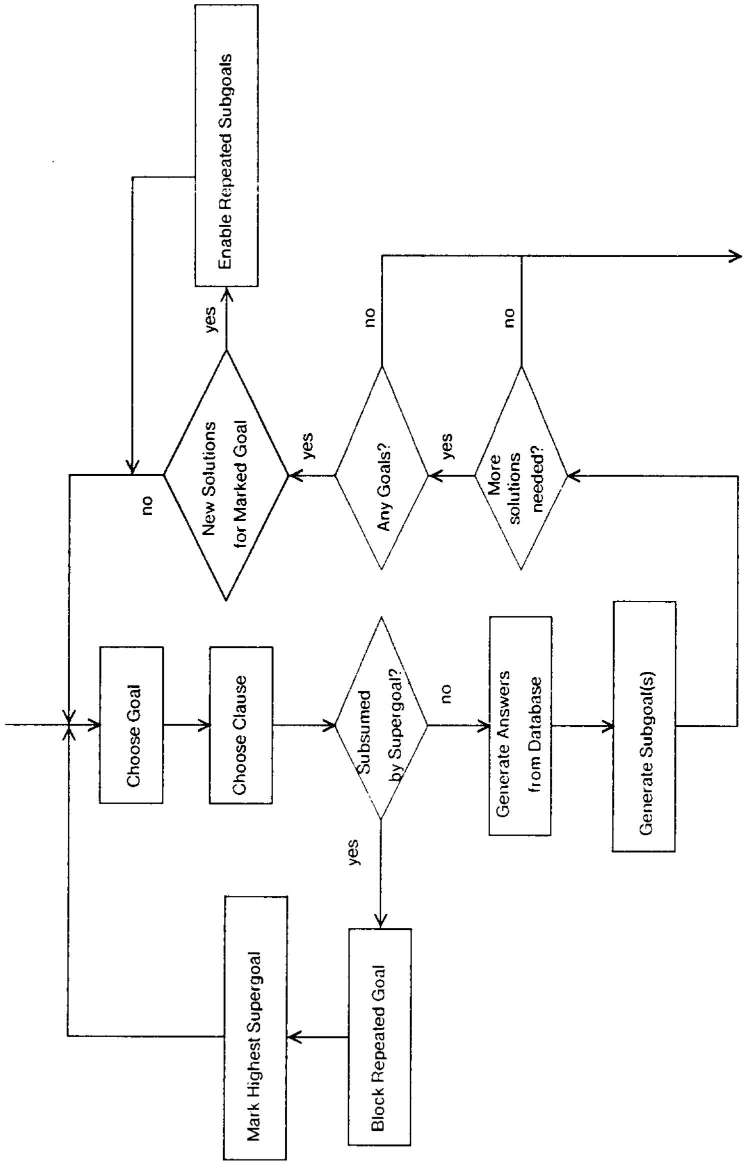

FIG. 8. Backward inference procedure with repetition control.

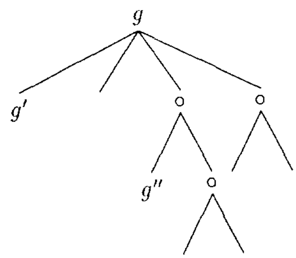

FIG. 9. An idealized AND/OR tree containing repetition.

Example 3.9 (*Circuit connections*). Consider Example 3.5, where the goal is to find all points in a circuit connected to a given point A and the database contains the facts $\mathrm{Conn}(A, B)$ and $\mathrm{Conn}(B, C)$. First the answer $z : B$ is found. The transitivity rule is then applied to the initial goal yielding $\mathrm{Conn}(A, y) \wedge \mathrm{Conn}(y, z)$. Since the clause $\mathrm{Conn}(A, y)$ is an instance of the initial goal, $\mathrm{Conn}(A, z)$, no inference is performed on this clause. However, since there is already a cached solution to the original goal $\mathrm{Conn}(A, z)$, the solution $y : B$ is found for the repeated descendant. Substituting this binding into the other conjunct yields the subgoal $\mathrm{Conn}(B, z)$. The answer $z : C$ is found in the system's database and is therefore cached as a solution to the original goal. The descendant $\mathrm{Conn}(B, z)$ is then expanded using the transitivity rule, yielding the conjunction $\mathrm{Conn}(B, y') \wedge \mathrm{Conn}(y', z)$. As with the first expansion, the clause $\mathrm{Conn}(B, y')$ is an instance of the subgoal $\mathrm{Conn}(B, z)$ so no inference is performed on the repeated clause $\mathrm{Conn}(B, y')$. As before there

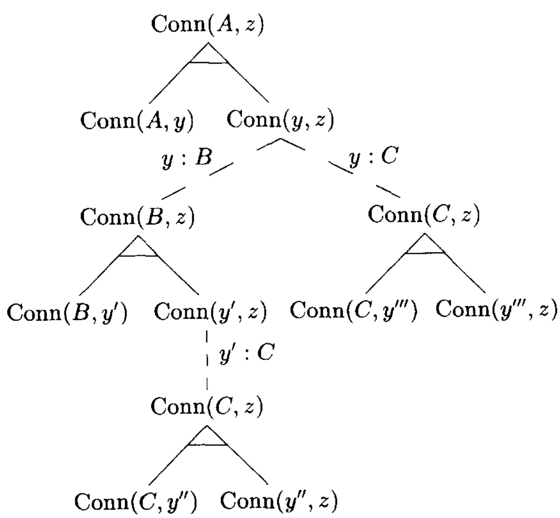

FIG. 10. Search for the query $\mathrm{Conn}(A, z)$.

is already a solution, $y' : C$, in the system's database, so this solution is substituted into the other conjunct yielding the subgoal $\text{Conn}(C, z)$. There are no solutions to this clause in the system's database. The expansion to $\text{Conn}(C, y'') \wedge \text{Conn}(y'', z)$ again contains the repeating subgoal $\text{Conn}(C, y'')$, so no further inference is performed and no answers are found to $\text{Conn}(C, z)$. This leaves no further alternatives for the supergoal $\text{Conn}(B, z)$. However, this subgoal did generate an additional answer $z : C$ to the initial goal $\text{Conn}(A, z)$, so the cached fact must be used in the first repeating descendant $\text{Conn}(A, y)$. Substituting the binding $y : C$ into the other conjunct yields the subgoal $\text{Conn}(C, z)$. As before, this subgoal yields no solutions, and the inference process terminates.

Example 3.10 (*Ancestry*). As a second example, consider the problem of finding all albinos, given the rule

$$\begin{aligned}&\text{Albino}(x) \wedge \text{Parents}(z) = \{x, y\} \wedge \text{Albino}(y)\\ &\quad \Rightarrow \text{Albino}(z).\end{aligned}$$

Assume that our database contains the facts:

$$\begin{aligned}&\text{Parents}(ABCD) = \{AB, CD\},\\ &\text{Parents}(AB) = \{A, B\},\\ &\text{Parents}(CD) = \{C, D\},\end{aligned}$$

$$\begin{aligned}&\text{Albino}(A),\\ &\text{Albino}(B),\\ &\text{Albino}(C),\\ &\text{Albino}(D).\end{aligned}$$

Beginning with the conjunct $\text{Albino}(z)$ the system would first discover the four answers in its database. It would then apply the recursive rule resulting in the conjunction $\text{Albino}(x) \wedge \text{Parents}(z) = \{x, y\} \wedge \text{Albino}(y)$. The first of these conjuncts is identical to its parent so the algorithm would halt further inference on this branch. However, since there are already four cached solutions to the original problem, these are substituted in as solutions to the repeated descendant. We are left with the conjunction $\text{Parents}(z) = \{x, y\} \wedge \text{Albino}(y)$ for the cases of $x = A$, $x = B$, $x = C$, and $x = D$. For these different bindings, the parents conjunct yields values for y and z:

x	y	z
A	B	AB
B	A	AB
C	D	CD
D	C	CD

For each of these solutions for y, the final conjunct Albino(y) is verified by reference to the database. Thus, the answers AB and CD are produced and cached for the original query. These, in turn, are substituted into the repeated descendant, again yielding the conjunction Parents(z) = $\{x, y\} \wedge$ Albino(y) for the cases $x = AB$ and $x = CD$. The parents conjunct yields new values for y and z:

x	y	z
AB	CD	$ABCD$
CD	AB	$ABCD$

Again the conjuncts Albino(CD) and Albino(AB) are verified by reference to the database so the answer $ABCD$ is produced and cached for the original query.

Finally, substitution of $ABCD$ into the repeated descendant yields no additional answers (since $ABCD$ has no progeny) and the search terminates. A sketch of this search space appears in Fig. 11.

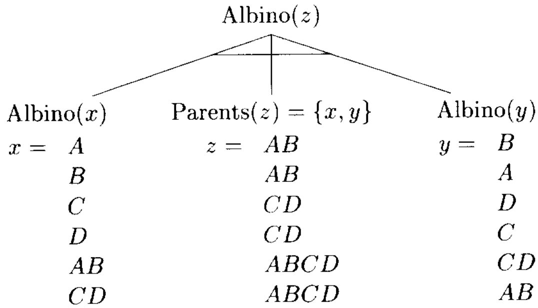

FIG. 11. Search space for the query Albino(z).

Soundness, completeness, and optimality. Since Algorithms 3.7 and 3.8 are merely finding frontier sets that satisfy Theorem 3.3, they do not adversely affect the logical soundness or completeness of an inference procedure. Although use of these algorithms will result in a drastic reduction of the size of repeating search spaces, their use does not guarantee termination of search. This is because the restricted search space $S_R(g)$ can still be infinite. Theorem 3.3 and the algorithms do not detect or eliminate divergent inference paths. As a result, an inference procedure making use of the cutoff algorithms might still encounter a divergent path and might therefore never terminate or find all of the answers in the space.

However, barring divergent paths, inference procedures based on Algorithms 3.7 and 3.8 are guaranteed to terminate, and any solution in the search space will be found. If the subgoal generator is logically complete, such an inference procedure will also be logically complete.

The algorithms limit search to one recursion level beyond the minimal level that satisfies the conditions of Theorem 3.3. In Section 3.2.1 we pointed out that cutting off redundant recursive paths at the earliest possible level may not be optimal. This observation therefore extends to the algorithms as well.

3.3. Special types of repetition

The theorems of the previous section are general, but weak. There are special cases of repeating inference where stronger results are possible. For example, in Section 3.1 we developed a much stronger result for the case where only a single answer was needed. There are two other special cases that merit particular attention, descendant subsumption and transitivity. Both of these cases rely on *goal subsumption* for their power. We say that a goal set g *subsumes* another goal set g' if there is a set of bindings b such that $g|_b \subseteq g'$. If only a single answer is needed for a problem any goal set g' subsumed by another goal set g will be redundant with that descendant [22]. The situation is somewhat more complicated when more than one answer is needed.

Theorem 3.11. *Let h' and h'' be arbitrary descendants of g and let b' and b'' be the binding sets that relate solutions to h' and h'' to solutions to g (i.e. $h' \Rightarrow g|_{b'}$ and $h'' \Rightarrow g|_{b''}$) as illustrated in Fig.* 12. *Suppose that h' subsumes h'' with the bindings c. If the bindings for the output variables in $b' \circ c$ are a subset of the bindings b'', h'' is redundant with h'.*

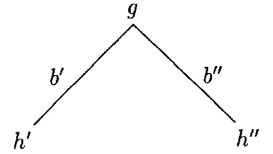

FIG. 12. Illustration for the subsumption theorem.

Proof. Let a'' be a solution found to h''. Then $b'' \circ a''$ is a solution to g that will be found by exploring h''. We must show that the same solution (or a generalization of it) can be found by exploring h'. Since $h'|_c \subseteq h''$ we know that there is some solution $a' \subseteq c \circ a''$ that will be found for h' (assuming complete database indexing). Thus, $b' \circ a'$ will also be a solution to g. But since $a' \subseteq c \circ a''$ and $b' \circ c \subseteq b''$ we have

$$b' \circ a' \subseteq b' \circ c \circ a'' \subseteq b'' \circ a''.$$

So the solution $b' \circ a'$ found by the descendant h' is a generalization of the solution $b'' \circ a''$ found by the descendant h''. h'' is therefore redundant with h'. □

As an example, consider the simple search space of Fig. 13 generated from the goal $Q(x)$ using the rules

$$\begin{aligned}
R1:&\quad P(x) \Rightarrow Q(x),\\
R2:&\quad P(A) \Rightarrow Q(A),\\
R3:&\quad P(A) \Rightarrow Q(B),\\
R4:&\quad P(A) \Rightarrow Q(x).
\end{aligned}$$

The subgoal $P(x)$ subsumes the three subgoals $P(A)$ with the bindings $c = \{x : A\}$, but not all of these subgoals are redundant with $P(x)$. The leftmost $P(A)$ in Fig. 13 is redundant because its binding set $\{x : A\}$ is identical to c. However, the second $P(A)$ has the binding set $\{x : B\}$, which does not contain $x : A$. Therefore, this subgoal is not redundant with the subgoal $P(x)$, unless x is not an output variable. The third instance of $P(A)$ is also not redundant with the subgoal $P(x)$ (assuming x is an output variable), since it has an empty binding set. Finally, it is worth noting that the first and second instances of $P(A)$ are redundant with the third instance of $P(A)$. For these cases c is empty and therefore, is contained in any binding set.

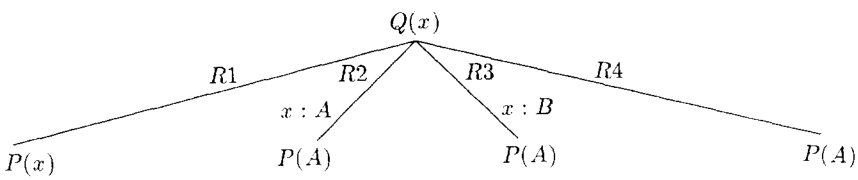

FIG. 13. Subsumption example.

3.3.1. *Descendant subsumption*

We can apply Theorem 3.11 to cases of repeating inference. In this case, h'' is a descendant of h', and the root goal g can be taken to be h'. Thus, the binding list b' is empty.

Corollary 3.12. *Let g' be a repeating descendant of g and let b be the set of bindings that relate solutions to g' to solutions to g ($g' \Rightarrow g|_b$). Let c be the set of bindings such that $g' = g|_c$. If the bindings for the output variables in c are contained in b, the goal g' is redundant with the goal g. Furthermore, it is* optimal *to discard g', since for every proof of an answer to g involving g', there is a simpler proof of the same answer not involving g'.*

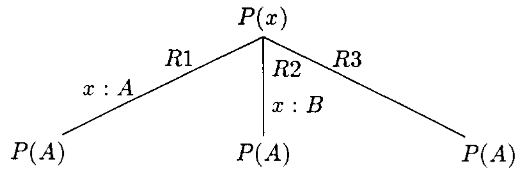

FIG. 14. Repetition subsumption example.

As an example, consider the simple search space of Fig. 14 generated from the goal $P(x)$ and the rules

$$\begin{aligned} &R1\colon\ P(A) \Rightarrow P(A), \\ &R2\colon\ P(A) \Rightarrow P(B), \\ &R3\colon\ P(A) \Rightarrow P(x). \end{aligned}$$

The three subgoals, $P(A)$, are subsumed by the root goal $P(x)$ with the binding set $c = \{x : A\}$. The first subgoal has the binding set $b = \{x : A\}$, which is identical to the binding set c. According to Corollary 3.12 the first subgoal can therefore be eliminated. This agrees with our intuitions, since any proof of $P(A)$ could be used directly to get the solution $x : A$ for the original goal $P(x)$. The second subgoal has the set of bindings $b = \{x : B\}$, which does not contain c. Therefore, it cannot be eliminated if x is an output variable. Intuitively, a proof of the second subgoal, $P(A)$, would allow us to conclude $P(B)$, and there may be no direct proof of $P(B)$. Likewise, the third subgoal cannot be eliminated since its binding set b is empty and therefore does not contain c. In this case, a proof of $P(A)$ would allow us to conclude $\forall x P(x)$, and there may be no direct proof of this statement.

The most common cases of descendant subsumption are when a descendant is identical to an ancestor in every respect, including variables. For this case, the set c is empty and the descendant can be eliminated. These situations arise from if-and-only-if rules expressing definitions and from rules expressing properties like symmetry, associativity, and commutivity. For example, in a circuit analysis system we might need the information that electrical connections are symmetric:

$$\mathrm{Conn}(x, y) \Rightarrow \mathrm{Conn}(y, x).$$

Suppose we were to apply this rule to the problem of finding all the points in a circuit connected to a given point A.

$$\text{find all } z\colon\ \mathrm{Conn}(A, z).$$

$$\begin{array}{c} \mathrm{Conn}(A, z) \\ | \\ \mathrm{Conn}(z, A) \\ | \\ \mathrm{Conn}(A, z) \\ \vdots \end{array}$$

FIG. 15. Search path for a symmetry rule.

We first get the subgoal Conn(z, A). Applying the rule to this subgoal gives the subgoal Conn(A, z), as shown in Fig. 15. But this subgoal is identical to the original goal, so we can prune the repetition using the subsumption theorem.

3.3.2. *Transitivity*

The subsumption theorem also has a direct application to repeating inference resulting from transitivity rules. Consider the connectivity example used in the previous sections, with the query Conn(A, z) and a database containing the facts

$$\begin{array}{l} \mathrm{Conn}(A, B), \\ \mathrm{Conn}(B, C), \\ \mathrm{Conn}(C, D). \end{array}$$

Figure 16 shows the portion of the space that would be generated for this problem using Algorithm 3.8. First, the answer $z = B$ is found in the database. Then, the conjunctive subgoal Conn(A, y) $\wedge$ Conn(y, z) is generated. The first of these conjuncts is repeated, so we plug in the answer $y = B$ that has already been found. Continuing on the remaining conjunct Conn(B, z) we find one answer in the database, $z = C$, and use the transitivity rule to generate the subgoal Conn(B, u) $\wedge$ Conn(u, z). The first of these conjuncts is a repeat of its parent so we again plug in the solution already found, $u = C$. The remaining conjunct becomes Conn(C, z). Again, there is a single answer in the database, $z = D$. We apply the transitivity rule one more time yielding the conjunction Conn(C, v) $\wedge$ Conn(v, z). The first of these is again a repeat of its parent and we plug in the available solution, $v = D$. The remaining conjunct, Conn(D, z), yields no solutions, so we begin to unwind. Note that we have already found all of the solutions to the problem, $z = B$, C, and D, yet we have not substituted the newly generated answers into the two remaining repeating descendants. According to the algorithm, we must substitute $u = D$ into the conjunction Conn(B, u) $\wedge$ Conn(u, z). Following this, we must substitute the answers $y = C$ and $y = D$ into the first subgoal Conn(A, y) $\wedge$ Conn(y, z). Each of these

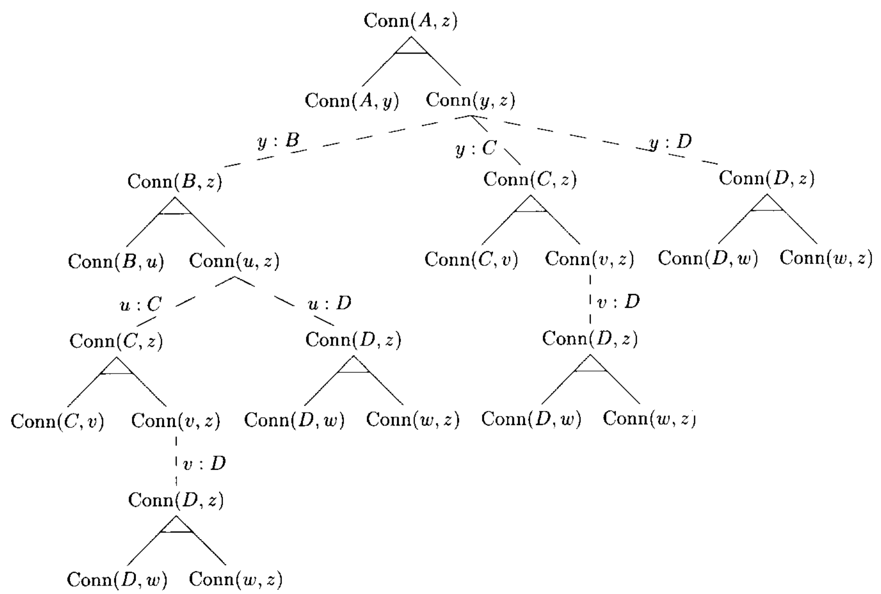

FIG. 16. Search space for the query Conn(A, z).

substitutions causes more redundant inference. In effect the procedure produces each of the answers twice. A similar situation occurs with the dual query Conn(x, D) (assuming the conjuncts are processed in a reasonable order) and with the general query, Conn(x, z)

Much of the duplication can be eliminated by recognizing and pruning subsumed goals. For example, the two instances of the goal Conn(C, z) are mutually redundant according to the subsumption theorem. Likewise, the four instances of the goal Conn(D, z) are mutually redundant. If all but one of each are eliminated the remaining search space does not contain any redundant portions. Using the subsumption theorem, together with Algorithm 3.8 therefore solves the problem. However, the two results can be combined into a more succinct reduction theorem.

Theorem 3.13. *Let $g' \wedge g''$ be the conjunctive subgoal produced by applying a transitivity rule*

$$R(x, y) \wedge R(y, z) \Rightarrow R(x, z)$$

to the goal g, as illustrated in Fig. 17. Let h' and h'' be the conjunctive subgoals produced by the application of the transitivity rule to the conjuncts g' and g'' respectively. Then, h' and h'' are mutually redundant. In other words, $A(g) = A_{h'}(g) = A_{h''}(g)$.

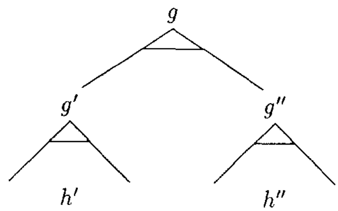

FIG. 17. Transitivity search space.

Proof. For all possible g that match $R(x, z)$, the conjunction $g' \wedge h''$ subsumes $h' \wedge g''$ and vice versa. For example, if

$$g = \text{“}R(x, z)\text{”}, \qquad g' = \text{“}R(x, y)\text{”}, \qquad g'' = \text{“}R(y, z)\text{”},$$
$$h' = \text{“}R(x, v) \wedge R(v, y)\text{”} \quad \text{and} \quad h'' = \text{“}R(y, w) \wedge R(w, z)\text{”},$$

then $R(x, y) \wedge R(y, w) \wedge R(w, z)$ subsumes $R(x, v) \wedge R(v, y) \wedge R(y, z)$ and vice versa for any subset of $\{x, z\}$ as output variables. The theorem therefore follows immediately from the subsumption theorem. $\square$

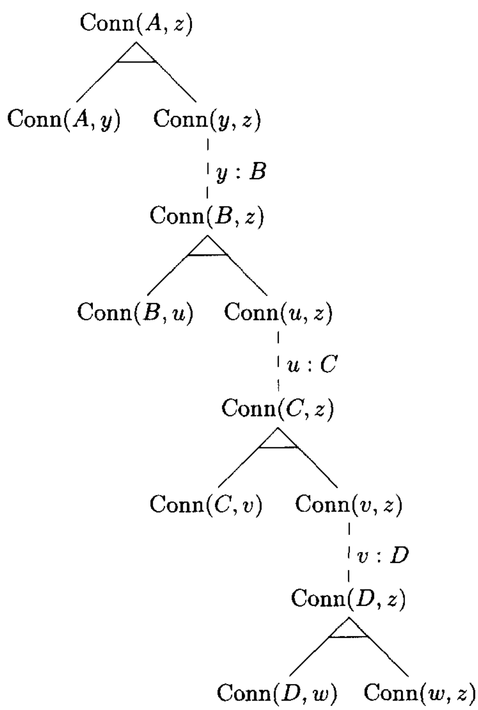

FIG. 18. Left-pruned search space for the goal Conn(A, z).

This result is easily implemented. When a transitivity rule is applied to a goal, the rule should not be applied to one of the two conjunctive subgoals generated. For the connectivity example the two possibilities are shown in Figs. 18 and 19. If the transitivity rule is not reapplied to the left-hand branches the result is a simple, but lopsided search space. If it is not reapplied to the right-hand branch, repeating inference occurs in the left-hand branch, and the methods of Section 3.2 must be applied. Using Algorithm 3.8, inference on the left-hand branch would stop after one level. All answers are generated merely by caching solutions and substituting them into the left branch.

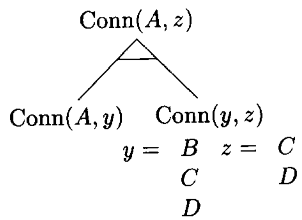

FIG. 19. Right-pruned search space for the goal Conn(A, z).

4. Divergent Inference

The most troublesome form of inference loops are those that do not repeat. Consider again the simple rule describing the integers:

$$y = x + 1 \wedge \text{Integer}(x) \Rightarrow \text{Integer}(y). \tag{7}$$

A query such as Integer(2.5) generates an infinite sequence of subgoals like that shown in Fig. 20. If we were to list the sequence of goal conjuncts reduced at each step in the inference process, no specific conjunct would appear more than once in this sequence. There are an infinite number of Integer conjuncts

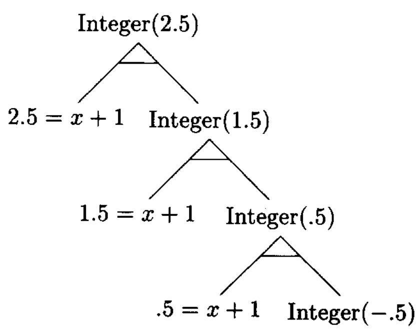

FIG. 20. Search space for the query Integer(2.5). The first conjunct has been evaluated for each subgoal.

in this sequence, but each one has a different argument. As we indicated in Section 1.3 we refer to such nonrepeating recursive inference as *divergent* inference.

How do we go about cutting off inference in such cases? In general it is only semidecidable whether or not the space below a given subgoal will contain novel answers to the problem. Yet for a case like the one above we can supply a fairly simple argument for pruning the infinite recursion from the search space. Suppose that the smallest integer in the database is 2. The sequence of descendants from Integer(2.5) is monotonically decreasing. As a result, once we have passed Integer(2) all further descendants are superfluous; they will never be able to match any fact in the database. This argument is not unlike the sort of arguments used in proving program correctness or program termination. Here we have used, as an invariant assertion, the fact that every descendant of Integer(2.5) will be of the form $\phi(x) \wedge \text{Integer}(x)$ and that x will always be less than 2.

This kind of argument can be generalized to arbitrary recursive collections. What is necessary is to find an invariant assertion for each goal form in the loop that implies that there will be no answers in the database for the corresponding goal. In addition we must show that all other rules that apply to goals in the loop (rules not in the recursive collection) will not produce any answers.

We can make this kind of argument more precise. Let the relation No(p) mean that there are no facts in the database that unify with the proposition p.

Theorem 4.1. *Let $\{R_1, \ldots R_m\}$ be the relations occurring in the consequents of the rules in a cyclic collection (recursive collections included). Let*

$$F_{j,k,n} = \text{“}\phi_{j,k,n}(y, z) \wedge R_j(y) \Rightarrow R_k(z)\text{”}$$

designate the nth rule in the collection having a relation R_j in its premise, and R_k in its consequent. (The $\phi_{j,k,n}$ may contain additional R from the set.) Suppose that there is a predicate β_k on the domain of each relation R_k such that

(1) $\beta_k(y) \Rightarrow \text{No(“}R_k(y)\text{”)}$,

(2) $\beta_k(z) \wedge \phi_{j,k,n}(y, z) \Rightarrow \beta_j(y)$, *and*

(3) $\beta_k(z) \Rightarrow \text{Superfluous}(\gamma)$, *for all other facts*, $(\gamma \Rightarrow R_k(z))$, *not in the recursive collection.*

Then, if $\beta_k(A)$ holds, the goal $R_k(A)$ is superfluous.

Here $\beta_k(z)$ is the invariant assertion for those goals with the relation R_k. The first condition states that the invariant assertion assures that no answers will be found. The second condition states that the invariant assertions are preserved from a goal to its immediate subgoals, and the third condition assures that none of the exit points of the loop will lead to any answers.

Proof. First we consider just those descendants of $R_k(A)$ that can be generated using rules in the cyclic collection. We want to show that there are no answers in the database for any of these descendants. We know that each of these descendants will contain a clause $R_j(y)$ for some R_j in the set of relations described in the theorem. If we can show that the invariant assertion $\beta_j(y)$ holds for that descendant, condition (1) in the theorem tells us that there will not be any answers in the database. Thus, we want to show that, for each descendant g' generated using only the cyclic collection, there is some expression $\psi(y, z)$ and some relation R_j in the set so that g' takes the form

$$g' = \text{“}\psi(y, z) \wedge R_j(y)\text{”} \tag{8}$$

and that

$$\psi(y, z) \Rightarrow \beta_j(y). \tag{9}$$

We prove this by induction on descendant depth. For the initial goal $R_k(A)$, the induction hypothesis (8) holds if we let ψ be the empty clause, $y = A$ and let $j = k$. Likewise, (9) follows from the given $\beta_k(A)$.

Now assume (8) and (9) for every lth-level descendant of the goal $R_k(A)$. Any $(l + 1)$st-level descendant will be a subgoal of some lth-level descendant. There are two possible ways of obtaining subgoals from an lth-level descendant $\psi(y, z) \wedge R_j(y)$.

(1) Apply some rule to a clause of ψ. In this case the new subgoal will be of the form $\psi'(y, z) \wedge R_j(y)$, which satisfies (8). Furthermore, since

$$\psi'(y, z) \Rightarrow \psi(y, z) \quad \text{and} \quad \psi(y, z) \Rightarrow \beta_j(y)$$

we get $\psi'(y, z) \Rightarrow \beta_j(y)$, which proves the second half of our induction hypothesis (9).

(2) Alternatively, we could apply some rule $F_{i,j,n}$ to the clause $R_j(y)$. In this case the new subgoal will be

$$\psi(y, z) \wedge \phi_{i,j,n}(x, y) \wedge R_i(x).$$

If we let

$$\psi'(x, z) = \phi_{i,j,n}(x, y) \wedge \psi(y, z)$$

our subgoal becomes $\psi'(x, z) \wedge R_i(x)$, which again satisfies (8). Furthermore, since

$$\psi(y, z) \Rightarrow \beta_j(y) \quad \text{and} \quad \beta_j(y) \wedge \phi_{i,j,n}(x, y) \Rightarrow \beta_i(x).$$

(condition (2) of the theorem) we get that

$$\phi_{i,j,n}(x, y) \wedge \psi(y, z) \Rightarrow \beta_i(x) \quad \text{or} \quad \psi'(x, z) \Rightarrow \beta_i(x).$$

Thus the second part of the induction hypothesis also holds.

The hypotheses (8) and (9) therefore hold for all $(l+1)$st-level descendants of $R_k(A)$ and by induction, for all descendants of $R_k(A)$ produced using only rules in the cyclic collection. It follows from condition (1) in the theorem that there will not be any answers in the database for any of these descendants.

What remains is to consider those descendants produced using rules not in the cyclic collection. Every such descendant will involve either applying such a rule directly to the goal $R_k(A)$, or to one of the goals $\psi(y, z) \wedge R_j(y)$ generated using the cyclic collection. Again there are two ways of producing subgoals to a goal of the form $\psi(y, z) \wedge R_j(y)$.

(1) Apply some rule to a clause of ψ. As before, such a subgoal will still satisfy the induction hypothesis and the previous argument holds.

(2) Apply a rule, $\gamma \Rightarrow R_j(y)$, to the clause $R_j(y)$ to yield the subgoal $\psi(y, z) \wedge \gamma$. But by the induction hypothesis we know that $\psi(y, z) \Rightarrow \beta_j(y)$. By condition (3) of the theorem we conclude that γ is superfluous. Thus there are no answers in the database to any of the descendants of this subgoal.

Therefore, there are no answers in the database for any of the descendants of $R_k(z)$, which means it is superfluous. □

4.1. Example

To see how this theorem applies, consider the simple integer example introduced earlier. For this example there is only one rule in the recursive collection. The relation in its consequent R_1 is the Integer relation, and its ϕ will be $\phi_{1,1,1}(x, y) =$ "$y = x + 1$". If we choose $\beta_1(x) =$ "$x < 2$", the condition $\beta_1(y) \wedge \phi_{1,1,1}(x, y) \Rightarrow \beta_1(x)$ will be satisfied. If the smallest integer in the database is 2, $\beta_1(x) \Rightarrow \text{No}(R_1(x))$ is true. The final condition, that all rules not in the recursive collection result in superfluous goals, is true since there are no other rules. Then, according to the theorem $\beta_1(x) \Rightarrow$ Superfluous("$R_1(x)$"), or $x < 2 \Rightarrow$ Superfluous("Integer(x)"). Therefore, we conclude that the subgoal Integer(1.5) is superfluous.

4.2. Application of the theorem

In general, mechanizing the application of Theorem 4.1 is not a simple matter. First we choose an applicable recursive collection to apply the theorem to. It may be that all recursive collections are already known and marked in the database. In this case finding an applicable recursive collection is a straightforward lookup operation. If not, we must recursively enumerate the set of applicable rules looking for recursive collections. This is done by mapping through each rule that applies to a goal and doing the same for each subgoal. If the same rule is used again in any path, a cycle and possible recursive collection has been located. If several independent recursive collections are found we must choose one. In satisfying the final criterion of the theorem (that all other applicable rules do not result in any answers) the others will be considered. Note that the theorem may need to be applied recursively to prove these cases.

Ambiguity in choosing the recursive collection also arises when two or more recursive collections share a common rule. In this case we have nested or interwoven loops. According to the definition of a recursive collection, their union also constitutes a recursive collection. We could therefore choose to apply the theorem to one of the individual recursive collections, or to the composite collection. If we choose to apply it only to an individual collection, in the final step it will be necessary to prove that none of the other interwoven loops can yield an answer. This is usually more difficult. It is therefore probably wise to consider the maximal recursive collection first.

The second step is to collect the set of consequent relations in the recursive collection. This is straightforward.

The third step is to find a set of invariants β_i that satisfy the characteristics

(1) $\quad \beta_k(z) \Rightarrow \text{No}(``R_k(z)")$,

(2) $\quad \beta_k(z) \wedge \phi_{j,k,n}(y, z) \Rightarrow \beta_j(y)$.

This task involves generating possible predicates β_i and testing them to see if they satisfy the above axioms. The most efficient way to do this is to start at one place in the loop, and proceed around the loop in an orderly fashion, generating the β at each step. Thus we start by generating a possibility β_k for some k and check to see that it satisfies the first axiom above. Then, choose j so that there is some rule $F_{j,k,n}$. Now generate β_j and check to see that it satisfies both the first and second axioms. Then, choose i so that there is some rule $F_{i,j,n}$ and so forth.[8]

The real problem is in generating good candidates for any individual β_i, particularly since the desired β_i might be a conjunction of known relations. We could start by considering all known predicates on the domain D_i of R_i. If none of these work, we could try all known relations from D_i to a new domain D' and conjoin these with known predicates on D'. If none of these work, we consider conjunctions containing three relations, and so on. In general this may be necessary. However, it seems that β often takes the form of an integer bound,

$$\beta_i(x) = ``\gamma(x, l) \wedge l \sim N"$$

[8]Manna and Waldinger [16] discuss more sophisticated ways of generating loop invariants for the purposes of program verification. Much of this work appears to be applicable here.

where $\sim$ is one of $<$, $>$ or $=$ and N is a fixed integer. In our simple integer case, γ was the identity relation, $\sim$ was $<$ and N was 2. However, if we were dealing with lists of ever increasing length, the Length function might be appropriate. Similarly, if we were dealing with human ancestry a function such as Birthdate might be appropriate. The strategy for generating β_i therefore involves first considering an empty γ if the domain of R_i is the integers. If this fails, known relations mapping the domain D_i onto the integers are considered. In effect, this is a way of generating possible ordering relations for domains where an ordering relation is not already known. Although it may, in theory, be necessary to consider conjunctions of known relations for γ, if the number of known relations is large, the space of possibilities quickly becomes intractable.

The final step involves verifying that none of the other relevant facts will generate any answers to the problem. This may be trivial as in the integer example, or it can be arbitrarily difficult, if there are other recursive collections involved. In the latter case, this final step may well involve application of Theorem 4.1 or any one of the cutoff theorems developed in Section 3.

4.3. Functional embedding: A special case

A common cause of divergent inference are rules that contain functional expressions on their left-hand sides. By this we mean rules of the form,

$$P(F(x)) \wedge \phi \Rightarrow P(x).$$

For example, the rules,

$$\text{Jewish}(\text{Mother}(x)) \Rightarrow \text{Jewish}(x),$$
$$\text{Integer}(\text{Successor}(x)) \Rightarrow \text{Integer}(x),$$

are both of this sort. Such rules will always lead to divergent inference if the inference engine does not evaluate the embedded functional expressions. For example, a query such as Jewish(Job) would lead to the subgoals Jewish(Mother(Job)), Jewish(Mother(Mother(Job))), etc.

For such cases we can often choose β to be a lower bound on the level of functional embedding in a subgoal. If there are no rules relevant to a problem that can shrink the amount of functional embedding (e.g. $Q(x) \Rightarrow P(f(f(x)))$), it is possible to stop the inference process whenever the level of functional embedding exceeds the largest embedding available in the database. For example, in the jewish ancestry problem, if the fact with the largest functional embedding is Jewish(Mother(Mother(Job))), any subgoal having a functional embedding deeper than two could be discarded. Notice that this strategy refers to total functional embedding independent of the actual functions involved.

The reason is that there may be rules available such as $P(G(x)) \Rightarrow P(F(x))$ that can result in new subgoals having different embedded functions.

4.4. Commutivity of inference steps

In the previous section we considered only cases where it was possible to prove that no answers existed in a portion of the search space. In fact we can generalize Theorem 4.1 by only insisting that subgoals contribute no novel answers to the overall problem (as opposed to no answers at all). It is usually quite difficult to prove that answers generated somewhere in a loop will not result in novel answers to the overall goal. However, there are special cases where redundancy can be proven, and in such cases Theorem 4.1 can be applied. In this section we develop such a special case result for situations where the inference steps are *commutative*.

Consider the pair of axioms:

$$R_1: \quad y = x + 1 \wedge \text{Integer}(x) \Rightarrow \text{Integer}(y),$$
$$R_2: \quad y = x - 2 \wedge \text{Integer}(x) \Rightarrow \text{Integer}(y).$$

As before, suppose that the problem is to determine whether or not 2.5 is an integer, and the smallest integer in the database is 2.

When only the first of the above two rules was available, we argued that the sequence of subgoals from Integer(1.5) was monotonically decreasing, and therefore the subgoal Integer(1.5) was superfluous. Given both rules, this argument no longer holds, since Integer(3.5), Integer(5.5), ... are now descendants of the subgoal Integer(1.5). In fact, imagine that the fact Integer(5.5) happened to be in the database. Then, the subgoal Integer(1.5) would not be superfluous, since the problem could be solved by exploring one of its descendants.

Even though the goal Integer(1.5) may not be superfluous, we can argue that it is redundant with its supergoal, Integer(2.5). The argument depends on the observation that any application of the two rules above is *commutative*. In other words, if a subgoal can be produced by applying one rule, then the other, it can also be produced by applying the rules in the reverse order. For example, the subgoal Integer(3.5) can be produced from the goal Integer(2.5) by applying R_1 followed by R_2. Alternatively, it can be produced by applying R_2 followed by R_1.

Using this observation, we separate the descandants of Integer(1.5) into two groups, those generated by applying only the first of the two rules, and those that involve at least one application of the second rule. From our earlier argument we know that the first class will not result in any answers. For the second class, since the two rules are commutative, the same subgoal can be produced by first applying the second rule to the goal Integer(2.5). Thus the subgoal Integer(1.5) is redundant.

We now make this kind of argument precise. We say that two sets of rules are commutative if any subgoal produced using a rule from one set followed by a rule from the other set could also be produced by using the rules in the opposite order. More formally,

$$\begin{aligned}&\text{Commutative}(s, t) \Leftrightarrow \forall q \in s,\ r \in t,\ g,\ g',\ h\\&\quad(\text{Subgoal}_q(g, g') \wedge \text{Subgoal}_r(g', h)\\&\qquad\qquad\Rightarrow \exists g''(\text{Subgoal}_r(g, g'') \wedge \text{Subgoal}_q(g'', h))),\end{aligned}$$

where the notation $\text{Subgoal}_r(g, h)$ means that the goal h can be derived as a subgoal of the goal g using the rule r.

Theorem 4.2. *Suppose that the set of applicable facts for a goal g can be broken up into two commutative subsets s and t. Suppose that all descendants of g, generated using only rules in s, produce no novel answers (i.e. if only rules in s are used, the subgoal g would be redundant). Then, all (immediate) subgoals of g produced using rules in s are redundant for the entire collection $s \cup t$.*

Proof. Let g' be an (immediate) subgoal of g generated using a rule from the set s. Consider an arbitrary descendant, d, of g'. Suppose d is produced using only the rules in the set s. Then, by the premises of the theorem there are no answers in the database for d that result in novel answers to the overall problem. Alternatively, suppose that d is produced using, at least one rule r from the set t. Since the rules in s and t are commutative, the descendant d will also be a descendant of the subgoal g'' generated by applying r to the goal g. All descendants of g' (including g' itself) are therefore redundant with immediate subgoals of g produced using rules in t. Since our choice of g' was arbitrary all immediate subgoals of g produced using rules in s are redundant. □

4.5. Example

Using Theorem 4.2, together with the results of previous sections, the two-rule integer example can now be handled.

There are two immediate subgoals of the goal Integer(2.5). The first, Integer(1.5), is redundant according to Theorems 4.2 and 4.1. Therefore, it is possible to eliminate it without sacrificing any answers. The other subgoal, Integer(4.5), has the two immediate subgoals Integer(3.5) and Integer(6.5). If the maximum integer in the database is 3, we can again apply Theorems 4.2 and 4.1 to show that Integer(6.5) is redundant. The other subgoal, Integer(3.5), has the two immediate subgoals Integer(2.5) and Integer(5.5). The latter is again redundant, by application of Theorems 4.2 and 4.1. The remaining subgoal Integer(2.5) is identical to the original goal, and so by

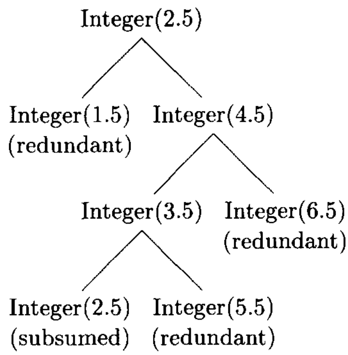

FIG. 21. Search space for the query Integer(2.5). The first conjunct has been evaluated and removed from each subgoal.

Corollary 3.12 it can also be eliminated. The abbreviated subgoal tree is shown in Fig. 21.

It is interesting to note that if we changed the constant in either of the two rules to an irrational number, the inference could not be completely stopped. We could still apply Theorems 4.2 and 4.1 at each level of the search space, but we would never return to the original goal, and therefore could never apply Corollary 3.12. In this case, the chain of subgoals would continue to bounce back and forth between the two extreme integers in the database.

4.6. Remarks

In this section, we first developed a general theorem for terminating divergent inference, and used it to solve a simple problem involving a single loop. We also applied the theorem to the special case of functional embedding.

It is generally more difficult to apply the theorem to cases of redundancy, as we can see from the two-rule integer example above. Powerful special case methods like Theorem 4.2 seem essential for dealing with such complex cases. We have investigated only one such result here. Additional work is needed to build up a library of theorems, like 4.2, that can be used for various difficult cases of divergent inference.

5. Discussion

5.1. Detecting recursive inference

There is always a price to pay for controlling inference and control of recursive inference is no exception. For repeating inference the cost is relatively low. It involves suspension of repeating goals and the caching of answers to goals with repeating subgoals. However, for divergent inference, control involves explicit proofs that subspaces are superfluous or redundant. When the alternative is an infinite loop, any finite control cost is justifiable. The problem is, we usually cannot be certain whether or not an infinite loop will result. As we pointed out in Section 2, even when a problem has an infinite recursive search space, recursive inference will not necessarily occur. The necessary answers might be found before an infinite path is explored by the inference engine.

In general, it is undecidable whether or not a given inference procedure will terminate when searching a recursive space.[9] The best that can be done is to institute control only when it is considered likely that it will be necessary or cost-effective. This general issue is discussed further in [27, 28]. For recursive inference there are several interesting strategies. The simplest is to monitor search depth or total search space size and institute control when it exceeds some threshold. A more elaborate, but more costly scheme is to preserve the subgoal and justification trees and institute control when a given fact has been used more than some fixed number of times in the derivation of a particular subgoal. A third alternative is to limit control to those cases where a recursive collection is involved in the deduction. (Recursive collections could be recognized either when rules are entered into the system, or during the problem solving process.)

Each of the strategies has certain advantages and disadvantages. In general, they trade accuracy for expense. For example, the recognition of rule reuse is usually a more accurate predictor of recursive inference than overall search depth, but it is also more expensive since it requires keeping the goal and justification stacks and searching them for each new subgoal.[10] As a result, the best strategy for a given application will depend upon such things as the average depth of inference, the frequency of recursive inference, and the density of recursive collections in the system's database.

There is also no reason why these strategies cannot be combined. For example, we could use search space depth or complexity to determine whether or not to initiate the strategy of checking for recursive collections or repeated rules. Likewise, the strategy of checking for repeated rules could be used as a filter for the strategy of looking for recursive collections. These combined strategies allow a less expensive but less accurate detection criteria to serve as a filter for a more accurate and more costly one. Such combinations may, in fact, prove to be the most cost-effective for many applications.

5.2. History and related work

5.2.1. *Recursive inference*

Black [1] and McKay and Shapiro [18] describe algorithms for stopping

[9]The problem is equivalent to the halting problem for Turing machines since backward inference over a set of axioms is Turing equivalent.

[10]Associating a marker or counter with each rule doesn't work in general. The marker would have to be path-dependent since we do not wish to count the repeated usage of rules in independent inference paths.

repeating inference similar to those developed in Section 3.2.2. However, they do not provide any proof that the pruning strategy is correct and do not consider the question of optimality. They also do not consider any of the special cases (like transitivity, subsumed subgoals, or single-answer queries) where more powerful strategies can be used.

The special case of eliminating identical subgoals appears to have been first used by Gelernter in his geometry theorem proving program:

> Subgoals . . . are rejected . . . that appear as higher subgoals on the [subgoal] graph (or are syntactically symmetric to some higher subgoal). [7, p. 142]

In Gelernter's application, since all goals and subgoals are geometry theorems requiring only a yes or no answer, both Theorem 3.1 and Corollary 3.12 apply. As a result, all repeated subgoals can be eliminated for this particular application. Loveland and Reddy [14] have shown that the technique of eliminating identical subgoals can be extended to backward inference mechanisms that recognize and make use of "contradiction constructs".

Special cases of repeating inference were also used in building the MYCIN system [25]. One cause of repeating subgoals in MYCIN was the use of *self-referencing* rules. A self-referencing rule states that if there is already evidence for a condition ψ and some other condition ϕ holds, there is additional evidence for the condition ψ. These rules therefore include the proposition ψ in both the premise and conclusion. MYCIN handles a self-referencing rule by postponing it until all other rules for concluding ψ have been used. Then, the self-referencing rule is applied exactly once.

This strategy involves pruning all repeated applications of a rule, a much stronger pruning strategy than is indicated by Theorem 3.3. This strategy works for self-referencing rules because they are actually quite different from recursive rules. Consider how we would translate a self-referencing rule into a precise declarative statement. We might be inclined to write something like

$$\psi \wedge \phi \Rightarrow_p \psi$$

where $\Rightarrow_p$ means that we have p additional evidence for the conclusion (the actual calculus for combining certainties is unimportant). If this statement were true, given some small amount of evidence for ψ, and the fact that ϕ is true, we could use this axiom over and over again to derive greater and greater belief in ψ.[11] This is not the intended meaning of the self-referencing rule. In a

[11]It is interesting to note that, Theorem 3.1, as stated, will not hold in the case of uncertain reasoning. Even if ψ is a ground clause, recursion could continue to increase the belief in ψ. However, from an evidential point of view we would never want to allow this, since arguments should not be circular.

self-referencing rule the recursive premise is a screen to prevent exploration of ϕ unless there is already some evidence of ψ. In other words, the recursive premise is control knowledge about when to apply the simpler rule

$$\phi \Rightarrow_p \psi .$$

Thus a logical translation of a self-referencing rule would consist of two rules; the simple rule given above, and a control rule indicating that the above rule should not be tried unless there is already evidence for ψ. Neither of these rules are recursive, so the theorems developed in Section 3 do not apply. This translation also shows why MYCIN's pruning strategy for self-referencing rules is appropriate. Since the above rule is not recursive, it need be applied only once.

Repeating inference also occurs in MYCIN as a result of rule loops. For example, a rule might allow a parameter B to be inferred from a parameter A, while another rule might allow A to be inferred from B. MYCIN's strategy in such cases is to never use a rule more than once in a single reasoning chain. This is equivalent to pruning all repeating subgoals. This works because, once the context is bound, the premises and conclusions of such rules are ground clauses. Thus, as in Gelernter's application, the powerful pruning strategy of Corollary 3.12 applies.

More recently, Minker and Nicolas [19] have developed a special case of Corollary 3.12 and have shown that for the class of "singular" recursive rules all repeating subgoals will be subsumed and can therefore be eliminated.

Other approaches to controlling repeating inference have also received some attention, although the results have been limited to special cases. Reiter [23] and Minker and Nicolas [19] have shown conditions where it is possible to use only forward inference on recursive collections. Automatic reformulation of recursive collections has been explored by Chang [2] and Naqvi and Henschen [11, 20]. They describe methods of automatically generating efficient procedures for the special class of "regular" recursive collections. Minker and Nicholas [19] have shown that this method applies to a slightly broader class of recursive collections. Recently, Ullman [29], Van Gelder, and Naughton [21] have obtained results for several special cases of recursive inference. In particular, they have proposed techniques similar to those suggested in Section 4.3 for handling the special case of functional embedding. They have also isolated classes of recursive collections where the recursion can be eliminated by reformulating the rules involved.

The more difficult problem of divergent inference has received little attention in the literature. Fischer Black noted the problem in developing his natural deduction system [1] but provides no solution other than depth-limited search.

Problems of recursive inference have also arisen outside of the artificial intelligence community. Both repeating and divergent inference are constant obstacles in the construction of PROLOG programs. Users of PROLOG become well versed in manual reformulation of rules to eliminate infinite loops. As we illustrated, this is not always an easy task and the resulting programs can be quite opaque.

Recently, at Stanford we have encountered repeating and divergent inference in the construction of systems for reasoning about digital circuits [8, 12, 26]. The techniques described here are being implemented in an experimental version of the MRS system [24].

5.2.2. *Program verification*

There is a strong similarity between recursive inference and recursive programs that do not terminate. In essence, an inference procedure together with a goal and recursive collection of facts is a recursive program. Thus, it should come as no surprise that the techniques for deciding whether a given inference path terminates, loops, or diverges, bear a striking resemblance to the method of well-founded sets used to prove program termination [6, 15].

But the similarities end here. In general, a recursive program that does not terminate is of little use. It must be modified so that it does terminate. Doing this requires some understanding of the intention behind the program. In contrast, a recursive collection of facts has *meaning*, independent of the particular inference engine being used. In this case, the inference engine must be modified so that it will perform the proper deductions. The intent of the inference procedure is known. Thus, the change takes the form of information about how to prune the search space. In short, a program that does not terminate may be incorrect in an arbitrary manner, while an inference engine that loops (for a given goal and recursive collection) is incorrect in a very specific way; it explores too much of the search space.

5.3. Summary and final remarks

The control of recursive inference involves demonstrating that portions of a search space are either superfluous or redundant. When either of these properties has been demonstrated, the offending portion of the search space can be discarded. Although this will always be logically correct, it may not be optimal in every case.

Proofs of redundancy and superfluity involve knowledge about the content of the system's database, and about properties of the relations involved in the inference, such as ordering relations on the domains, monotonicity, boundedness and commutivity. This kind of information is commonly available but has rarely been needed or used in AI systems. In contrast to the general domain-dependent character of the control problem, the special case of repeating inference admits control that is domain-independent. The method of suspending and reenabling repeated subgoals does not depend upon the meaning of the symbols involved. It is not entirely clear why this fortuitous result should hold.

It is true however, that in some cases the more general domain-dependent techniques of proof can lead to more severe pruning for repeating inference than is possible with the syntactic method of Section 3.

Determining whether or not recursive inference will occur for a given problem is in general undecidable. We have suggested three possible criteria for determining when control of recursive inference should be instituted; inference depth or complexity, repeated rule usage, and the use of recursive collections. Combinations of these approaches also appear promising, although the decision will almost certainly prove dependent upon the mix of problems encountered in any particular application.

Finally, there is no a priori reason why the techniques of proving redundancy and superfluity could not be applied to nonrecursive inference. The limiting factor is cost. When an infinite search is avoided, a high cost is justifiable. However, for nonrecursive inference the problem would have to be a difficult one for expensive analysis like that of Section 4 to be cost-effective. For such cases, complex monitoring strategies like those proposed in Section 5.1 would be indispensable.

ACKNOWLEDGMENT

We are indebted to Narinder Singh, Glenn Kramer, and to an anonymous reviewer for bringing interesting examples of repeating and divergent inference to our attention. Thanks also to Russ Greiner, Jock Mackinlay, Vineet Singh, Richard Treitel and the other members of the Logic Group at Stanford for general discussion on this subject over the last several years. Jim Bennett, Jan Clayton and Ted Shortliffe provided information on self-referencing rules and loops in MYCIN and EMYCIN. Jeff Finger, Pat Hayes, and Richard Waldinger provided useful pointers to related work on program verification and Richard Treitel and Jeff Ullman provided pointers to related work on databases and logic programming. Thanks to Jan Clayton and Jock Mackinlay for proofreading assistance and comments on presentation.

This work was supported by ONR contrast N00014-81-K-0004.

REFERENCES

1. Black, F., A deductive question-answering system, in: M. Minsky (Ed.), *Semantic Information Processing* (MIT Press, Cambridge, MA, 1968) 354–402.
2. Chang, C.L., On evaluation of queries containing derived relations in a relational database, in: H. Gallaire, J. Minker and J.M. Nicolas (Eds.), *Advances in Data Base Theory* (Plenum Press, New York, 1980) 235–260.
3. Clancey, W.J., The epistemology of a rule-based expert system, *Artificial Intelligence* **20** (1983) 215–251.
4. Davis, R., Meta-rules: reasoning about control, *Artificial Intelligence* **15** (1980) 179–222.
5. Doyle, J., A model for deliberation, action, and introspection, Artificial Intelligence Laboratory Memo AI-TR-581, MIT, Cambridge, MA, 1980.
6. Floyd, R.W., Assigning meaning to programs, in: J.T. Schwartz (Ed.), *Proceedings Symposium on Applied Mathematics* (American Mathematical Society, Providence, RI, 1967) 19–32.
7. Gelernter, H., Realization of a geometry-theorem proving machine, in: E. Feigenbaum and J. Feldman (Eds.), *Computers and Thought* (McGraw-Hill, New York, 1963) 134–152.
8. Genesereth, M.R., The use of design descriptions in automated diagnosis, *Artificial Intelligence* **24** (1985) 411–436.

9. Genesereth, M.R. and Smith, D.E., Procedural hints in the control of reasoning, Knowledge Systems Laboratory Rept. KSL-84-11, Stanford University, Stanford, CA, 1985.
10. Hayes, P.J., Computation and deduction, in: *Proceedings Symposium on Mathematical Foundations of Computer Science* (Czechoslovakian Academy of Sciences, Prague, 1973) 105–117.
11. Henschen, L. and Naqvi, S., Compiling queries in recursive first order databases, *J. ACM* **31** (1984) 47–85.
12. Kramer, G., Personal communication, 1984.
13. Lewis, H.R., Cycles of unifiability and decidability by resolution, Tech. Rept., Aiken Computation Laboratory, Harvard University, Cambridge, MA, 1975.
14. Loveland, D. and Reddy, C., Deleting repeated goals in the problem reduction format, *J. ACM* **28** (1981) 646–661.
15. Manna, Z., *Mathematical Theory of Computation* (McGraw-Hill, New York, 1974).
16. Manna, Z. and Waldinger, R., *Studies in Automatic Programming Logic* (North-Holland, New York, 1977).
17. McCarthy, J., Programs with common sense, in: M. Minsky (Ed.), *Semantic Information Processing* (MIT Press, Cambridge, MA, 1968) 403–418.
18. McKay, D.P., and Shapiro, S., Using active connection graphs for reasoning with recursive rules, in: *Proceedings Seventh International Joint Conference on Artificial Intelligence*, Vancouver, BC (1981) 368–374.
19. Minker, J. and Nicolas, J., On recursive axioms in deductive databases, *Information Systems* **8**(1) (1983) 1–13.
20. Naqvi, S.A. and Henschen, L.J., Performing inferences over recursive data bases, in: *Proceedings First National Conference on Artificial Intelligence*, Stanford, CA (1980) 263–265.
21. Naughton, J., Data independent recursion in deductive databases, in: *Proceedings ACM/SIGMOD Symposium on Principles of Database Systems* (ACM, New York, 1986).
22. Nilsson, N.J., *Principles of Artificial Intelligence* (Tioga, Palo Alto, CA, 1980).
23. Reiter, R., On structuring a first order data base, in: R. Perrault (Ed.), *Proceedings Second National Conference* (Canadian Society Computational Studies of Intelligence, Toronto, Ont., 1978) 19–21.
24. Russell, S., The compleat guide to MRS, Tech. Rept. STAN-CS-85-1080, Stanford University, Stanford, CA, 1985.
25. Shortliffe, E.H., Details of the consultation system, in: B.G. Buchanan and E.H. Shortliffe (Eds.), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project* (Addison-Wesley, Reading, MA, 1984) 78–132.
26. Singh, N., Exploiting design morphology to manage complexity, Ph.D. Thesis, Stanford University, Stanford, CA, 1985.
27. Smith, D.E., Controlling backward inference, Knowledge Systems Laboratory Rept., Stanford University, Stanford, CA, 1986.
28. Smith, D.E., Controlling inference, Ph.D. Thesis, Stanford University, Stanford, CA, 1985.
29. Ullman, J.D., Implementation of logical query languages for databases, *ACM Trans. Database Syst.* **10**(4) (1985).

Received July 1985; revised version received April 1986