POSTER ABSTRACT

A New Algorithm for Computing Transitive Closures

Yangjun Chen*
Dept. of Business Computing
University of Winnipeg, Manitoba, Canada R3B 2E9
ychen2@uwinnipeg.ca

ABSTRACT

In this paper, we propose a new algorithm for computing transitive closures. It needs only O(e·b) time and O(n·b) space, where n represents the number of the nodes of a DAG (directed acyclic graph), e the numbers of the edges, and b the DAG's breadth.

Categories & Subject Decriptors: E.1. Data Structure

General Terms: Algorithms, Theory

Key Words: DAGs, Transitive Closure, Branching, Topological Order

1. INTRODUCTION

Let $G = (V, E)$ be a directed graph (digraph for short). Digraph $G^* = (V, E^*)$ is the reflexive, transitive closure of G if $(v, w) \in E^*$ iff there is a path from v to w in G. This problem exists in a variety of applications, such as web and document databases, CAD/CAM, CASE, office systems and software management, as well as in object-oriented databases. In these kinds of systems, information is organised as a graph to represent object/sub-object (e.g., in object-oriented databases), design/sub-design (e.g., in CAD databases), and reference relationships in the citation indexes of scientific literature. In this paper, we present a new algorithm for computing efficiently the transitive closure of a digraph.

In this paper, we propose a new algorithm for this problem, which has a better computational complexity than any existing method.

2. A TREE LABELING METHOD

In this section, we discuss how to label a tree to speed up the computation of recursion in a relational environment.

Consider a tree T. By traversing T in *preorder*, each node v will obtain a number (it can be integer or a real number) $pre(v)$ to record the order in which the nodes of the tree are visited. In a similar way, by traversing T in *postorder*, each node v will get another number $post(v)$. These two numbers can be used to characterize the ancestor-descendant relationships as follows.

Proposition 1. Let v and v' be two nodes of a tree T. Then, v' is a descendant of v iff $pre(v') > pre(v)$ and $post(v') < post(v)$.

Proof. See Exercise 2.3.2-20 in [1]. □

If v' is a descendant of v, then we know that $pre(v') > pre(v)$ according to the preorder search. Now we assume that $post(v') > post(v)$. Then, according to the postorder search, either v' is in some subtree on the right side of v, or v is in the subtree rooted at v', which contradicts the fact that v' is a descendant of v. Therefore, $post(v')$ must be less than $post(v)$. The following example helps for illustration.

Example 1. See the pairs associated with the nodes of the graph shown in Fig. 1(a). The first element of each pair is the preorder number of the corresponding node and the second is its postorder number. With such labels, the ancestor-descendant relationships can be easily checked.

For instance, by checking the label associated with b against the label for f, we see that b is an ancestor of f in terms of Proposition 1. Note that b's label is (2, 4) and f's label is (5, 2), and we have $2 < 5$ and $4 > 2$. We also see that since the pairs associated with g and c do not satisfy the condition given in Proposition 1, g must not be an ancestor of c and *vice versa*. □

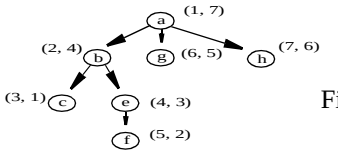

Fig. 1. Labeling a tree

3. TRANSITIVE CLOSURE OF DAGS

Now we discuss how to recognize the ancestor-descendant relationships w.r.t. a general structure: a DAG or a graph containing cycles.

What we want is to apply the technique discussed above to a DAG. To this end, we establish a *branching* of the DAG as follows.

Definition 1. (*branching* [2]) A subgraph $B = (V, E')$ of a digraph $G = (V, E)$ is called a branching if it is cycle-free and $d_{indegree}(v) \leq 1$ for every $v \in V$.

Clearly, if for only one node r, $d_{indegree}(r) = 0$, and for all the rest of the nodes, v, $d_{indegree}(v) = 1$, then the branching is a directed tree with root r. Normally, a branching is a set of directed trees. Now, we assign every edge e a same cost (e.g., let cost $c(e) = 1$ for every edge e). We will find a branching for which the sum of the edge costs, $\sum_{e \in E'} c(e)$, is maximum.

For example, the trees shown in Fig. 2(b) are a maximal branching of the graph shown in Fig. 2(a) if each edge has a same cost.

Assume that the maximal branching for $G = (V, E)$ is a set of trees T_i with root r_i ($i = 1, ..., m$). We introduce a *virtual root* r for the branching and an edge $r \rightarrow r_i$ for each T_i, obtaining a tree G_r, called the representation of G. For instance, the tree shown in Fig. 2(c) is the representation of the graph shown in Fig. 2(a). Using Tarjan's algorithm for finding optimum branchings [2], we can always find a maximal branching for a directed graph in $O(|E|)$ time if the cost for every edge is equal to each other. Therefore, the representative tree for a DAG can be constructed in linear time.

By traversing G_r in *preorder*, each node v will obtain a number $pre(v)$; and by traversing G_r in *postorder*, each node v will get another number $post(v)$. These two numbers can be used to recognize the ancestor-descendant relationships of all G_r's nodes as discussed in Section 2.

In a G_r (for some G), a node v can be considered as a representation of the subtree rooted at v, denoted $T_{sub}(v)$; and the pair (pre, $post$)

SAC'04, March 14-17, 2004, Nicosia, Cyprus.

* The author is supported by NSERC 239074-01 (242523) (Natural Sciences and Engineering Council of Canada).

associated with v can be considered as a pointer to v, and thus to $T_{sub}(v)$. (In practice, we can associate a pointer with such a pair to point to the corresponding node in G_r.) In the following, what we want is to construct a pair sequence: (pre_1, $post_1$), ..., (pre_k, $post_k$) for each node v in G, representing the union of the subtrees (in G_r) rooted respectively at (pre_j, $post_j$) (j = 1, ..., k), which contains all the descendants of v. In this way, the space overhead for storing the descendants of a node is dramatically reduced. Later we will shown that a pair sequence contains at most O(b) pairs, where b is the breadth of G. (The breadth of a digraph is defined to be the least number of the disjoint paths that cover all the nodes of the graph.)

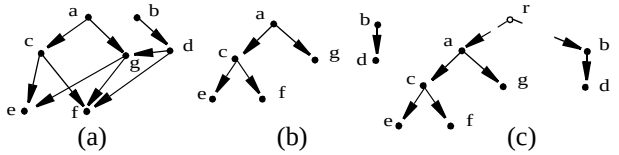

Fig. 2. A DAG and its branching

Example 2. The representative tree G_r of the DAG G shown in Fig. 2(a) can be labeled as shown in Fig. 3(a). Then, each of the generated pairs can be considered as a representation of some subtree in G_r. For instance, pair (3, 3) represents the subtree rooted at c in Fig. 3(a).

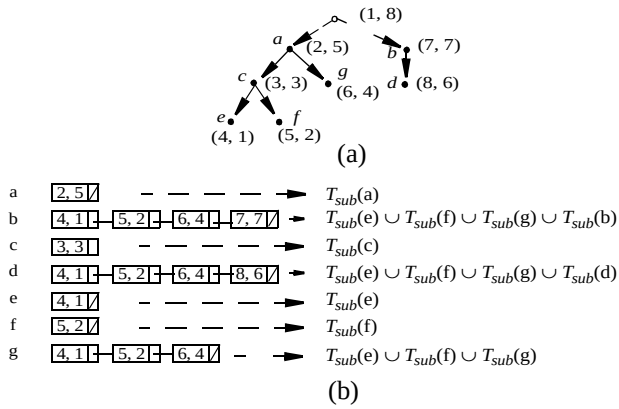

Fig. 3. Tree labeling and illustration for transitive closure representation

If we can construct, for each node v, a pair sequence as shown in Fig. 3(b), where it is stored as a linked list, the descendants of the nodes can be represented in an economical way. Let L = (pre_1, $post_1$), ..., (pre_k, $post_k$) be a pair sequence and each (pre_i, $post_i$) is a pair labeling v_i (i = 1, ..., k). Then, L corresponds to the union of the subtrees $T_{sub}(v_1)$, ..., and $T_{sub}(v_k)$. For example, the pair sequence (4, 1)(5, 2)(6, 4)(8, 6) associated with d in Fig. 3(b) represents a union of 4 subtrees: T_{sub}(e), T_{sub}(f), T_{sub}(g) and T_{sub}(d), which contains all the descendants of d in G. □

The question is how to construct such a pair sequence for each node v so that it corresponds to a union of some subtrees in G_r, which contains all the descendants of v in G.

First, we notice that by labeling G_r, each node in G = (V, E) will be initially associated with a pair as illustrated in Fig. 4. That is, if a node v is labeled with (pre, $post$) in G_r, it will be initially labeled with the same pair (pre, $post$) in G.

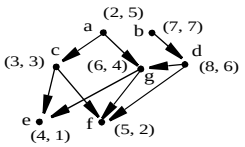

Fig. 4. Graph labeling

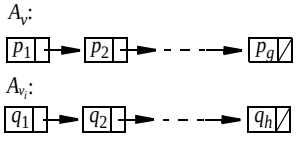

Fig. 5. linked lists associated with nodes in G

To compute the pair sequence for each node, we sort the nodes of G topologically, i.e., (v_i, v_j) ∈ E implies that v_j appears before v_i in the sequence of the nodes. The pairs to be generated for a node v are simply stored in a linked list A_v. Initially, each A_v contains only one pair produced by labeling G_r.

We scan the topological sequence of the nodes from the beginning to the end and at each step we do the following:

Let v be the node being considered. Let v_1, ..., v_k be the children of v. Merge A_v with each A_{v_l} for the child node v_l (l = 1, ..., k) as follows. Assume $A_v = p_1 \rightarrow p_2 \rightarrow ... \rightarrow p_g$ and $A_{v_l} = q_1 \rightarrow q_2 \rightarrow ... \rightarrow q_h$, as shown in Fig. 5. Assume that both A_v and A_{v_l} are increasingly ordered. (We say a pair p is larger than another pair p', denoted $p > p$' if $p.pre > p'.pre$ and $p.post > p'.post$.)

We step through both A_v and A_{v_l} from left to right. Let p_i and q_j be the pairs encountered. We'll make the following checkings.

(1) If $p_i.pre > q_j.pre$ and $p_i.post > q_j.post$, insert q_j into A_v after p_{i-1} and before p_i and move to q_{j+1}.
(2) If $p_i.pre > q_j.pre$ and $p_i.post < q_j.post$, remove p_i from A_v and move to p_{i+1}. (*p_i is subsumed by q_j.*)
(3) If $p_i.pre < q_j.pre$ and $p_i.post > q_j.post$, ignore q_j and move to q_{j+1}. (*q_j is subsumed by p_i; but it should not be removed from A_{v_i}.*)
(4) If $p_i.pre < q_j.pre$ and $p_i.post < q_j.post$, ignore p_i and move to p_{i+1}.
(5) If $p_i = p_j$' and $q_i = q_j$', ignore both (p_i, q_i) and (p_j', q_j'), and move to (p_{i+1}, q_{i+1}) and (p_{j+1}', q_{j+1}'), respectively.
(6) If $p_i = nil$ and $q_j \neq nil$, attach the rest of A_{v_i} to the end of A_v.

The following example helps for illustration.

Example 3. Assume that A_1 = (7, 7)(11, 8) and A_2 = (4, 3)(8, 5)(10, 11). Then, the result of merging A_1 and A_2 is (4, 3)(7, 7)(10, 11). Fig. 6 shows the entire merging process.

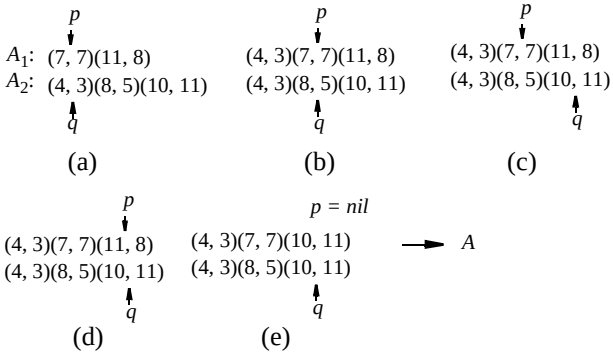

Fig. 6. An entire merging process

In each step, the A_1-pair pointed to by p and the A_2-pair pointed to by q are compared. In the first step, (7, 7) in A_1 will be checked against (4, 3) in A_2 (see Fig. 6(a)). Since (4, 3) is smaller than (7, 7), it will be inserted into A_1 before (7, 7) (see Fig. 6(b)). In the second step, (7, 7) in A_1 will be checked against (8, 5) in A_2. Since (8, 5) is subsumed by (7, 7), we move to (10, 11) in A_2 (see Fig. 6(c)). In the third step, (7, 7) is smaller than (10, 11) and we move to (11, 8) in A_1 (see Fig. 6(d)). In the fourth step, (11, 8) in A_1 is checked against (10, 11) in A_2. Since (11, 8) is subsumed by (10, 11), it will be removed from A_1 and p becomes nil (see Fig. 6(e)). In this case, (10, 11) will be attached to A_1, forming the result A = (4, 3)(7, 7)(10, 11) (see Fig. 6(e)). □

REFERENCES

[1] D.E. Knuth, *The Art of Computer Programming, Vol.1,* Addison-Wesley, Reading, 1969.
[2] J. Tarjan: Finding Optimum Branching, *Networks*, 7. 1977, pp. 25 -35.