EFFICIENT REPRESENTATION OF TAXONOMIES

G. Gambosi^(*), J. Nešetřil^(**), M. Talamo^(*)

(*) Istituto di Analisi dei Sistemi ed Informatica
del C.N.R., Viale Manzoni 30, 00185, Roma, Italy

(**) Charles University Malostranské Náměstí 25,
11600 Prague, Czechoslovakia

ABSTRACT

In this paper, the use of the concepts of dimension and boolean dimension of partial orders is studied for what concerns the efficient representation of taxonomies. Under this approach, labelings of the nodes (linear extensions of partial orders) are used to answer efficiently to such queries as "Is element a related to element b?".

INTRODUCTION

The efficient representation of taxonomies, i.e. free-structured sets of concepts, represents a rather interesting problem, as for example in some sectors of Artificial Intelligence, where it turns out to be able to answer quickly to such queries as "Is element a included (derived) from element b?".

In order to tackle this problem in the case of a static structure, the use of suitable labelings of the nodes (elements) seems to be worth of investigation. If we consider a taxonomy as a partial order, the approach (extensively studied in poset theory) of using the concept of linear extensions and of dimension of posets [KT] makes it possible to determine the number of such labelings and, in some cases, to obtain the labeling themselves.

Moreover, a generalization of the concept of dimension, denoted as boolean dimension, introduced and extensively studied in the full paper [GNT], seems to be convenient to use in order to obtain a better performance both in terms of query time and of space.

In section 1, some definitions are given together with the introduction of the concept of boolean dimension.

This work was partially supported by the ESPRIT Project ALPES.

In section 2, a labeling of simple taxonomies (rooted trees) is presented which allows an efficient treatment of inclusion queries.

In section 3, such approach is extended to two types of generalized taxonomies.

1. BASIC DEFINITIONS AND TERMINOLOGY

Given a partially ordered set $P = (N, <_P)$, where N is an order relation on set N , a *linear extension* L (topological sorting) on P [R] is a total ordering $<_L$ on N which embeds P itself, i.e. such that for each $n_1, n_2 \in N$, if $n_1 <_P n_2$ then $n_1 <_L n_2$.

The (order) *dimension* of P ($\dim P$) is defined as the minimum number of different linear extensions of P whose intersection is necessary to determine P itself. Hence, $\dim P$ is the minimum k for which there exists a set $L = \{L_1, L_2, \dots, L_k\}$ of linear extensions of P such that $n_1 <_P n_2$ iff $n_1 <_{L_i} n_2$ for each $i = 1, \dots, k$ (fig. 1).

In their seminal paper on the dimension theory of posets, Dushnik and Miller [DM] have shown, among others, that the dimension of an arbitrary partial order can be unbounded as $n = |N|$ tends to infinity and can grow as fast as $O(n)$ (*).

Note that, if $\dim P = k$, it is possible to represent P introducing a k -labeling on set N : hence, for each node $n \in N$ there exist k labels $\ell_1(n), \ell_2(n), \dots, \ell_k(n)$ and, given $n_1, n_2 \in N$, $n_1 <_P n_2$ iff the following boolean formula is verified:

$$F = x_1 \wedge x_2 \wedge \dots \wedge x_k$$

where x_i is true iff $\ell_i(n_1) < \ell_i(n_2)$.

(*) As usual, we will say that, given two functions $f(n), g(n), f(n) = O(g(n))$ iff there exist c, n_0 such that $f(n) \leq c \cdot g(n)$ for $n > n_0$.

Moreover, given a poset $P = (N, <_P)$, the *Hasse diagram* of P is defined as a dag $HD(P) = (N, A)$ such that:

1. A is not transitive, i.e. if $\langle n_1, n_2 \rangle \in A$, then there is no path of length greater than two from n_1 to n_2 in $HD(P)$ ($HD(P)$ is a transitive reduction).
2. The transitive closure $(*) HD(P)^* = (N, A^*)$ of $HD(P)$ is isomorphic to $P = (N, <_P)$.

In the following, we will refer to posets only in terms of the associated Hasse diagrams and of their characteristics.

Given a poset $P = (N, <_P)$ and a boolean formula F defined on a set $\{x_1, \dots, x_k\}$ of boolean variables, we will say that P is F -representable iff there exists a set of k linear extension $L = \{L_1, \dots, L_k\}$ such that, given any $n_1, n_2 \in N$, $n_1 <_P n_2$ iff $F(x_1, \dots, x_k) = \text{true}$, where $x_i = \text{true}$ iff $l_i(n_1) < l_i(n_2)$. In fig. 2 an example of a boolean formula F and a

Fig. 2

poset P which is F -representable is given.

Given a poset P , let us denote as *boolean dimension* of P [GNT] ($\dim_B P$) the minimum k for which there exists a formula $F(x_1, \dots, x_k)$ such that P is F -representable.

Observe that such definition includes the usual definition of order dimension given above as the particular case where all formulas are of the form

(*) Given a directed graph $G = (N, A)$ the transitive closure of G is defined as the directed graph $G^* = (N, A^*)$ such that, given $n_1, n_2 \in N$, $\langle n_1, n_2 \rangle \in A^*$ iff there exists a path from n_1 to n_2 in G .

$$F(x_1, \dots, x_k) = \bigwedge_{1 \leq i \leq k} x_i.$$

This makes it possible to derive immediately the following result:

FACT

For any poset $P = (N, <_P)$, $\dim_B(P) \leq \dim(P)$.

This easy inequality is complemented by the following theorem

Theorem 1. For every n there exists P_n such that

$$\dim_B(P_n) = 4 \text{ and } \dim(P_n) = n$$

Theorem 1 is based on the standard example on Fig. 2.

2. ROOTED TREES

Let us denote a poset $T = (N, <_T)$ as a rooted tree if its Hasse diagram $HD(T)$ is a rooted tree.

Hence, T is a rooted tree if:

1. There exists $r \in N$ such that $r <_T n$ for any $n \in N - \{r\}$
2. There exists a partition of $N - \{r\} = \bigcup_{i=1}^k N_i$, $N_i \cap N_j = \emptyset$ if $i \neq j$, such that, denoting by $T_i = (N_i, <_{T_i})$ the subposet induced by set N_i , $<_T = (\bigcup_{i=1}^k <_{T_i}) \cup \{(r, n) \in <_T | n \in N\}$, i.e. nodes in different subsets N_i, N_j of $N - \{r\}$ are not related in T .
3. Each $T_i = (N_i, <_{T_i})$ is a rooted tree.

From [TM] it is trivial to derive that a rooted tree T has dimension $\dim T \leq 2$.

As well known, simple taxonomies can be modeled as rooted trees: hence, by providing two suitable labelings L_1, L_2 of the entities considered in the taxonomy, it will be possible to answer queries of the type "Is entity a included in entity b ?" in time $T(n) = O(1)$ and space $S(n) = O(1)$ for each entity.

In order to identify such two labelings, let us denote as O_i , $i \in N$, the total order defined by an arbitrary permutation on the set S_i of sons of node i in $HD(T) = (N, A)$, $S_i = \{n \in N | \langle i, n \rangle \in A\}$.

Then, the labels $\ell_1(n), \ell_2(n)$ of a node $n \in N$ can be derived as:

1. $\ell_1(n)$ is the rank of node n in the total order determined by a Depth-First traversal of T , provided that, for each $i \in N$, the order of application of the DFS to the set of subtrees induced by the set

of i is exactly O_i .

2. $\ell_2(n)$ is the rank of node n in the total order derived by a DFS traversal of T , provided that, for each $i \in N$, the order of application of the DFS to the set of subtrees induced by S_i is O_i' , the ordering obtained by reversing O_i (fig. 3).

Fig. 3

The following theorem holds:

Theorem 2. Given a rooted tree $T = (N, <_T)$ and the two labelings L_1, L_2 over N defined above, the following holds: for each $n_1, n_2 \in N$, n_1 is an ancestor of n_2 (i.e. $n_1 <_T n_2$) iff $\ell_1(n_1) < \ell_1(n_2)$ and $\ell_2(n_1) < \ell_2(n_2)$.

Proof (if). Let us assume w.l.o.g. that node n_1 is not an ancestor of node n_2 . Let us denote as n_3 the Lowest Common Ancestor of n_1 and n_2 in T and let $r(n_1)$ ($r(n_2)$) be the son of n_3 which is root of the subtree containing n_1 (n_2).

Two cases are possible:

1. $r(n_1) < r(n_2)$ in O_{n_3} : hence, since $r(n_2) < r(n_1)$ in O_{n_3}' ,
 - a. $\ell_2(n_2) < \ell_2(r(n_1))$ by the DFS properties
 - b. $\ell_2(r(n_1)) < \ell_2(n_1)$
 then, $\ell_2(n_2) < \ell_2(n_1)$ and the condition is not verified.
2. $r(n_2) < r(n_1)$ in O_{n_3} , then:
 - a. $\ell_1(n_2) < \ell_1(r(n_1))$

$$b. \ell_1(r(n_1)) < \ell_1(n_1)$$

hence, $\ell_1(n_2) < \ell_1(n_1)$ and the condition is not verified again.

(Only if). By the properties of DFS traversal, both $\ell_1(i) < \ell_1(j)$ and $\ell_2(i) < \ell_2(j)$, where $i, j \in N$ and $i <_T j$ (i is the root of a subtree containing j). $\square$

3. GENERALIZED TAXONOMIES

In order to represent sets of taxonomies defined on the same set of elements in order to answer efficiently to queries of the type "Is class a included by class b ?" we need to represent the corresponding structures (generalizations of trees) in such a way to answer efficiently to such queries as "Is there any path from node a to node b ?".

In the following we will consider two types of generalized taxonomies: local rooted trees and leaf-analgamated forests, and will show how, under general condition, it is possible to represent such taxonomies efficiently.

Local rooted trees

Local rooted trees are an extension of the definition of rooted trees introduced, among others, to represent more complex taxonomies, where a new taxonomy is defined among the elements sons of the same element.

It is possible to define a k -local rooted tree recursively as:

- A rooted tree is a 1-local rooted tree
- A k -local rooted tree $T = (N, <_T)$ is a poset such that the corresponding Hasse Diagram $HD(T) = (N, A)$ is a digraph such that $A = A_1 \cup A_1'$, $A_1 \cap A_1' = \emptyset$, where:

1. The digraph $RT = (N, A_1)$ is the H.D. of a 1-local rooted tree.
2. Let us denote as S_i the set of sons of node $i \in N$ in RT . Moreover, let us denote as $A_1'(i) \subseteq A_1'$ the subset of A_1' induced by S_i on A_1' . Then:

$$- A_1' = \bigcup_{i \in N} A_1'(i)$$

$$- A_1'(i) \cap A_1'(j) = \emptyset \text{ if } i \neq j, i, j \in N.$$

- $(S_i, E_1'(i))$ is a t -local rooted tree with $t < k$ and there exists at least one node $n \in N$ such that $(S_n, E_1'(n))$ is a $(k-1)$ -local

rooted tree (fig. 4).

Fig. 4

It is easy to note that the transitive closure of a local rooted tree T is equal to the transitive closure of a rooted tree: that immediately implies that, if T is a local rooted tree, the transitive closure of T can be represented by a couple of labelings. Such labelings can be immediately derived applying the labeling procedure for rooted trees presented above to the rooted tree T' obtained from T deleting all transitive arcs, i.e. all arcs $\langle i, j \rangle \in A$ such that there exists a path from i to j of length greater than two.

Leaf amalgamated forests

Definition. Let $P = (N, <_P)$ be a poset: given a minimal element $x \in N$ let $I(x) = \{y \in N \mid x <_P y\}$.

We say that P is a leaf amalgamated forest if:

1. Every $I(x)$ is a rooted tree
2. Denoting as $L(x) \subseteq I(x)$ the set of leaves (maximal elements) in $I(x)$, $I(x) \cap I(y) \equiv L(x) \cap L(y)$ ($x, y \in N$), i.e. for any two nodes x, y , $I(x)$ and $I(y)$ intersects only in a set of leaves (fig. 5).

Let us now prove the following theorem

Theorem. Let $P = (N, <_P)$ be a tree amalgamation, then $\dim_B(P) \leq 2 \cdot d$, providing that all in-degrees of nodes in N are bounded by d .

Proof. Let us first note that, given a maximal element $x \in N$, for each $y <_P x$ there exists a unique maximal chain containing both x and y .

Furthermore, if C and C' are maximal chains which intersect in a

Fig. 5

non maximal element y , then $C \cap C' \supseteq \{r \in N \mid r <_P y\}$.

Let us consider a d -coloring of the maximal chains in such a way that two chains with the same color do not intersect in a leaf. Then, the union of chains with the same color turns out to be a forest, i.e. a poset whose order dimension is 2.

Let us denote as $\ell_{1,k}, \ell_{2,k}$ the two labels relative to color k : then it is easy to note that the boolean function $\phi(x_{1,1}, x_{2,2}, \dots, x_{1,d}, x_{2,d}) = (x_{1,1} \wedge x_{1,2}) \vee (x_{2,1} \wedge x_{2,2}) \vee \dots \vee (x_{d,1} \wedge x_{d,2})$ (where $x_{i,j} = \text{true}$ iff $\ell_{i,j}(n_1) < \ell_{i,j}(n_2)$, $n_1, n_2 \in N$) describes poset P , hence $\dim_B(P) \leq \leq 2 \cdot d$. $\square$

Corollary. It is possible to answer to inclusion queries on leaf-amalgamated forests in time $O(k)$ and space $O(k \cdot n)$, where k is the number of leaf-amalgamated trees.

Proof. Derives immediately from the theorem above. $\square$

Let us denote as L the set of leaves and as $R = \{x \in N \mid \exists (x, y) \in A, y \in L\}$.

In the case where also the following hypothesis is verified:

3. $\forall x \in R$, $\text{outdegree}(x)$ is a constant it is moreover possible to state the following theorem

Theorem. Given a leaf-amalgamated forest F which verifies condition 3 above, then $\dim_B F = O(h)$, where $h = \max_{x \in N} \text{height}(x)$.

Corollary. Given a leaf-amalgamated forest F which verifies condition 3, there exists a data structure which makes it possible to answer to

inclusion queries on F in time $T(n) = O(h)$, where $h = \max_{x \in N} \text{height}(x)$, and space $S(n) = O(n \cdot h)$.

From the results above it follows that, given a leaf-amalgamated forest F , it is possible to manage the problem of answering to inclusion queries in time $\min(h, k)$ and space $\min(h \cdot n, k \cdot n)$. This result is strongly dependent from the use of boolean dimension; it is possible in fact to state the following theorem.

Theorem. There exists a leaf-amalgamated forest F which verifies condition 3 above such that $\dim F = \Omega(n)$.

CONCLUSION

The concepts of order dimension and boolean dimension of posets have been applied to the representation of (generalization of) taxonomies, in order to efficiently manage inclusion queries. This work is part of a more general study of the concept of boolean dimension of posets developed in the full paper.

REFERENCES

- [DM] B. Dushnik, E. Miller: *Partially ordered sets*, Amer. J. Math. 63 (1941), 600-610.
- [GNT] G. Gambosi, J. Nešetřil, M. Talamo: *On locally presented posets*, submitted to "Theoretical Computer Science".
- [KT] D. Kelly, W.T. Trotter: *Dimension theory for ordered sets in Ordered sets*, (I. Rival ed.), D. Reidel, Dordrecht, 171-211.
- [R] I. Rival: *Linear extensions of finite ordered sets*, Annals of Discrete Math. 23 (1984), 355-370.
- [TM] W.T. Trotter, J.I. Moore: *Some theorems on graphs and posets*, Discrete Math. 15 (1976), 79-84.