AU SUJET DE LA REPRÉSENTATION DE PROFILS AU MOYEN D'ARBRES SUPERPOSÉS

Par: E. García Camarero
SABINI-AUXOC
Don Ramón de la Cruz, 39
28001 MADRID

ABSTRACTS

The aim of this paper is to describe a method of how to represent profiles of the related references with the minimum use of memory and time resources.

So, we use a global representation of all the profiles through arborescence or superposed trees in such a way that each node will be represented only once, independently of the number of trees in which it appears, minimizing in that way the space in memory.

Besides, it will be minimized the time of references elaboration associated with the set of profiles, because for each node it will be only necessary to calculate its references once.

We will start by making some informal regards on profiles representation in an intuitive way. Afterwards, we'll give some formal definitions about the different files used in the representation; also, the operations carried out for file's origination and update, and lastly the elaboration of the reference sets.

We shall take a very simple command language for the profile definition to make easier the description of the method but wide enough so that it could spread out without difficulties to other general cases.

AU SUJET DE LA REPRÉSENTATION DE PROFILS AU MOYEN D'ARBRES SUPERPOSÉS

Par: E. García Camarero
SABINI-AUXOC
Don Ramón de la Cruz, 39
28001 MADRID

1. INTRODUCTION

L'objet de la présente communication est de décrire une méthode pour représenter des profils, ainsi que de rechercher les références correspondantes en employant un minimum de ressources de mémoire et de temps.

Pour en venir à bout, on se sert d'une représentation globale de tous les profils; c'est-à-dire, au lieu de considérer l'ensemble des arbres correspondant al tous les profils, on utilise des arborescences ou arbres superposés, de telle façon que chaque sommet soit représenté une seule fois, indépendamment du nombre d'arbres dans lesquels il apparaît et, en conséquence, nous minimisons l'espace en mémoire. De même, on minimise le temps d'élaboration des références associées à l'ensemble des profils, puisque si un sommet apparaît dans plusieurs profils (ou arbres) on ne devra calculer ses références qu'une seule fois.

Nous allons faire, préalablement, quelques considérations informelles pour décrire de forme intuitive la représentation. Ensuite, nous donnerons aussi quelques définitions formelles des différents fichiers employés dans la représentation, ainsi que des opérations à réaliser pour la création et l'actualisation des fichiers et pour élaborer les ensembles des références.

On a pris un langage de commandes très simple pour la définition des profils, afin de faciliter la description de la méthode, mais qui soit suffisamment ample pour s'en servir sans difficulté dans d'autres cas plus généraux.

Cette méthode a été appliquée dans le développement et l’implémentation du système automatisé de Bibliothèques de l’INI (SABINI) et elle est en train d'être exploitée avec efficience.

2. CONSIDÉRATIONS INFORMELLES SUR DES PROFILS ET LEURS ARBRES ASSOCIÉS

On sait qu'un profil personnel est la définition d'un domaine spécifique de connaissance pour lequel une personne montre son intérêt, ou sur lequel elle réalise son activité professionnelle. Dans un Centre de Documentation, les profils sont utilisés pour obtenir et distribuer périodiquement, parmi les usagers, la description des documents entrés dans le Centre pendant cette période et qui sont en rapport avec les th~mes spécifiés au profil de chaque usager.

Dans un système automatique il faudra donc considérer deux taches;

- La construction des nouveaux profils.
- La sélection des références par rapport aux domaines de chaque profil.

La première tâche est faite une seule fois lors de l'entrée du profil. La seconde se réalise une fois à chaque période de temps.

Afin de montrer par des exemples notre méthode, nous allons employer, comme langage de description de profils, un langage de commande simple pour en sélectionner des termes, réaliser des opérations booléennes (dans notre cas l'union (U), l'intersection (I) et la différence (D) parmi des ensembles de références, ou en trouver des sous-ensembles par des opérations du type limiter (dans notre cas, limiter par des années d'édition du document (A), par le pays d'édition (P), ou par la langue dans la quelle il a été écrit (L)).

La structure des commandes serait la suivante:

i	S	t	sélectionner
i	U	j, k	union
i	I	j, k	intersection
i	D	j, k	difference
i	A	j, (a_1-a_2)	limitation années
i	P	j, p	limitation pays
i	L	j, l	limitation langue.

En règles générales, on peut procéder de la même façon pour des commandes autrement définies (par exemple, pour les rendre plus compact es et pour faciliter la construction de profils, ou employer d'autres opérateurs logiques, ou d'autres limitateurs) puisque les commandes font l'une des suivantes opérations:

Type 1.- Former des ensembles primitifs de références (en règle générale à partir de termes primitifs, en employant des commandes type SELECT).

Type 2.- Obtenir de nouveaux ensembles de références à partir de ceux qui existent déjà en employant des opérateurs logiques (les plus fréquents sont les opérateurs union, intersection, différence, employés dans des commandes type COMBINE).

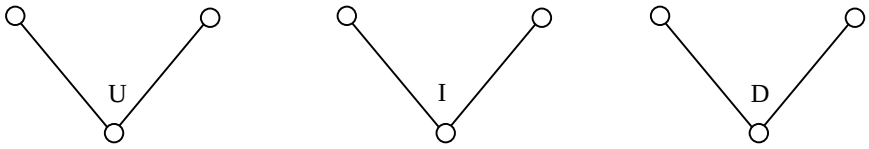
Type 4.- Obtenir des sous-ensembles des ensembles de référence qui existent, en appliquant quelque restriction (ces restrictions sont faites fréquemment par des commandes type LIMIT et se rappor-

tent à des périodes de temps (a_1-a_2), à des pays (p), ou à des langues (l).

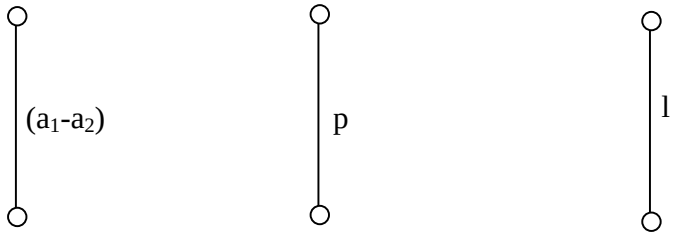
On peut représenter graphiquement ces opérations de la façon suivante, en déterminant un arbre pour chaque profil.

– Par l'opération du type 1, on ajoute de nouveaux points au diagramme, indépendamment de ceux qui y sont déjà.

– Par les opérations du type 2, on ajoute de nouveaux points au diagramme en fonction des deux points qui y existent, en connexion avec eux par des arcs, dont la confluence est marquée par le symbole de l'opération logique correspondante. On a, par exemple:



– Par les opérations du type 3, on ajoute un nouveau point en fonction d'un point déjà existant; il se met en connexion avec ce point moyennant un arc; l'arc est marqué avec la limitation correspondante. Ainsi:



Voyons quelques exemples de profils qui ont été construit avec le langage de commande donné plus haut, et de leurs représentations en arbre correspondantes.

Exemple 1.- Soit le profil défini par la séquence de commandes suivante:

1	S	t_1
2	S	t_2
3	S	t_3
4	I	1,2
5	U	4,3
6	L	5, français
7	P	6, Canada

L'arbre correspondant serait celui qui est donne par la figure 1, dans lequel les sommets sont numérotés selon le numéro de la commande dans le profil

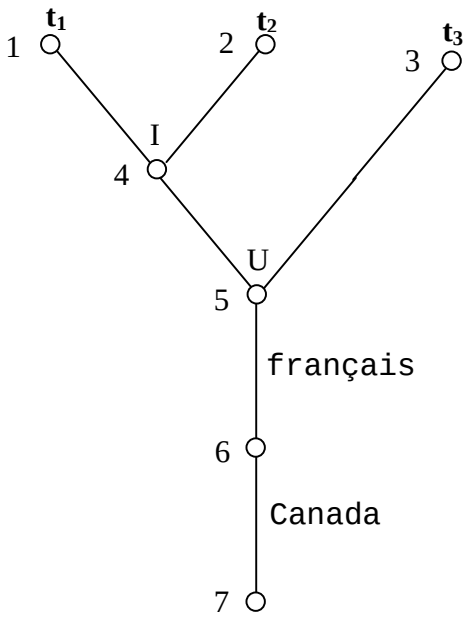


Figure 1.

Exemple 2.- À la suite, nous définissons un nouveau profil avec quelques éléments communs au précédent, moyennant la séquence de commandes suivante

1	S	t_1
2	S	t_2
3	S	t_3
4	I	1,2
5	D	4,3
6	A	(1975-1980)

L'arbre correspondant serait celui qui est donné dans la figure 2.

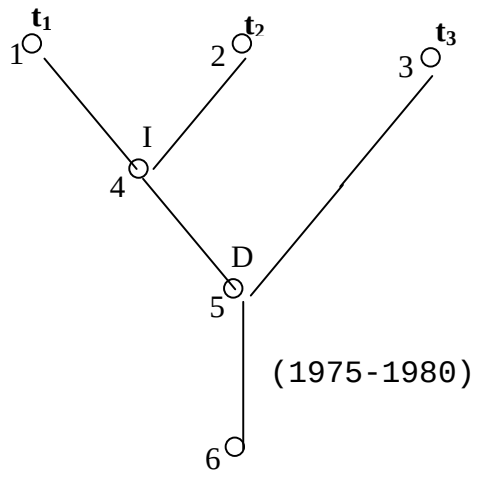


Figure 2

3. ARBRES SUPERPOSÉS CORRESPONDANT À DES ENSEMBLES DE PROFILS

Au paragraphe précédent nous avons vu la forme d'associer un arbre à chaque profil, afin que les sommets représentent les ensembles de références partielles qui se construisent par l'application de chaque commande, jusqu'à obtenir l'ensemble des références de la racine de l'arbre, pour obtenir la réponse au profil.

Un centre de documentation qui s'occupe d'un service de diffusion sélective de l'information n'accueille pas des profils isolés, mais des ensembles de nombreux profils qui présentent des éléments communs dans une certaine proportion. En conséquence, nous croyons qu'il faudra considérer une représentation globale pour des collections de profils, au lieu de considérer et de traiter isolément des profils indépendants.

Ainsi, nous représentons l'ensemble de profils, non par une collection d'arbres, mais par l'ensemble des sommets qui apparaissent dans tous les arbres, et qui sont mis en rapport par les arcs correspondants.

De cette façon, nous obtenons un ensemble d'arbres superposés (avec des éléments communs) qui consiste en une arborescence non planaire avec (éventuellement) de multiples racines. Nous assignons à chaque sommet de l'arborescence un ordre de multiplicité qui nous indique le nombre de profils auxquels ce sommet est commun.

C'est ainsi que chaque profil deviendra un arbre submergé dans l'arborescence, et il sera déterminé univoquement par le sommet que nous assignerons comme racine de l'arbre du profil.

On calculera les ensembles de références qui correspondent à tous les sommets, pour la construction de la solution des profils. Les ensembles associés aux racines des arbres, correspondant à chaque profil contiennent les références à ce profil.

Un ordre, obtenu en fonction de la position de chaque commande dans le profil, et de l'ordre d'inclusion de chaque profil dans l'arborescence, peut être assigné à l'ensemble de sommets de l'arborescence. Cet ordre est utilisé dans la création des fichiers de termes et de sommets correspondant à l'arborescence, et dans la construction des ensembles de références associés.

Pour montrer par des exemples la notion d'arbres superposés et des fichiers ordonnés correspondants, il faut prendre comme base les deux profils donnés aux exemples du paragraphe précédent. Dans la figure 3, les sommets de l'arborescence apparaissent avec une numérotation (ordre) que nous appelons absolue, pour la différencier de la numérotation (ordre) qui apparaît dans les arbres, que nous appelons relative.

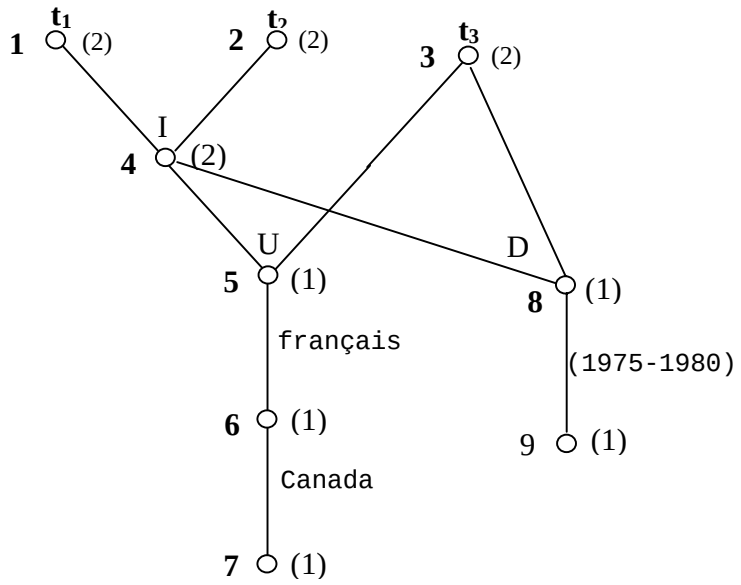


Figure 3

I1 y a 1ieu de remarquer que cette représentation n’emp1oie que neuf sommets, au lieu des treize qui son nécessaires pour sa représentation par des arbres indépendants ou non superposés.

Les fichiers de termes et sommets correspondant à l'arborescence de la figure 3 sont donnés par les tables des figures 4 et 5.

x	**t**	**n**
1	$\mathbf{t_1}$	**2**
2	$\mathbf{t_2}$	**2**
3	$\mathbf{t_3}$	**2**

Figure 4.

x	**o**	**y**	**α**	**n**
4	I	1	2	2
5	U	3	4	1
6	L	5	français	1
7	P	6	Canada	1
8	D	4	3	1
9	A	8	1975-1980	1

Figure 5.

4. DIFINITIONS

4.1. Pour la définition externe des profi1s (c'est-à-dire, pour sa définition par écran), nous emploierons les commandes définies au paragraphe 2 et que nous appellerons commandes d'écran.

4.2. Dans la représentation globale interne des profils, nous emploierons les arbres superposés décrits au paragraphe précédent, moyennant les fichiers suivants:

(F1).- On appelle fichiers de termes un fichier construit à partir des commandes i S t, et qui a la forme suivante:

x [t,n] {r}

où x es un identificateur de sommet, t indique un terme, n un nombre naturel qui indique la multiplicité du sommet et {r} es un ensemble de références.

(F2).- On appelle fichier inverse de termes le fichier construit a partir du précédent, à structure

t [x]

où t et x on la même signification que dans le cas précédent.

(F3).- On appelle fichier de sommets le fichier construit à partir des commandes différentes des i S t et qui a la forme

x [o(y,α) n] {r}

où x, n, r ont la même signification que dans les cas précédents; o indique un des opérateurs U, I, D, A, P, L; y es un identificateur de sommet; et α prend les valeurs suivantes

$\alpha = x$	si	o = U, L, D
$\alpha = (a_1\text{-}a_2)$	si	o = A
$\alpha = p$	si	o = P
$\alpha = 1$	si	o = L

(F4).- On appelle fichier de commandes le fichier qu'on obtient a partir du précédent, a structure

o (y, α) [x]

où o, y, α ont la même signification que dans les cas précédents.

(T1).- On appelle table de conversion un fichier qui a la structure

i [x]

qu'on emploie pour assigner à la numérotation relative i (identificateur du sommet dans l'arbre particulier) sa correspondante numérotation absolue x (identificateur du sommet dans l'arborescen-

ce d'arbres superposés). Cette table est transitoire et relative à l'élaboration de chaque profil personnel.

4.3. Dans la construction des ensembles de références pour satisfaire les profils, nous employons les fichiers suivants:

(F5).- Nous appelons catalogue un fichier a structure

r [b]

où r est une référence et b est la description bibliographique correspondante.

(F6).- Nous appelons inverse-Thesaurus un fichier a structure

t {r}

où t est un terme et {r} un ensemble de références.

5. OPÉRATIONS

5.1. Inclusions de nouveaux profils. Pour inclure de nouveaux profils dans l'arborescence, ou dans les fichiers correspondants on y ajoutera les sommets qui se dérivent de chaque commande d'écran (utilisé pour définir le nouveau profil) et qui n'existent pas préalablement.

Ensuite, nous allons donner, pour chaque type de commande d'écran, les opérations à réaliser pour l'inclusion de nouveaux profils.

i S t	On doit vérifier si **t [x]** se trouve dans le fichier F2; si la réponse est affirmative, on augmente d'une unité l'ordre de multiplicité n du sommet **x [t,n]** du fichier Fl; si la réponse est négative, on ajoute **x[t, 1]** a F1, **t [x]** à F2, et **i [x]** à Tl.
i o j, k	On remplace les valeurs relatives j, k par les valeurs absolues y, z au moyen de la table Tl. On met y , z en ordre croissant si l'opération est commutative. On consulte le fichier F4 pour voir si la commande **o(y,z) [x)** y est; si elle y est, on augmente d'une unité l'ordre de multiplicité n du sommet **x [o(y,z) n]** de F3 et on ajoute **i [x)** à Tl; si elle n'y est pas, on ajoute le sommet **x [o(y,z) 1]** à F3, **i[x]** à Tl et **o (y, z) [x]** à F4.

i A j,(a_1-a_2)	On remplace la valeur relative de j par la valeur absolue y en se servant de la table Tl. On vérifie si **A y,(a_1-a_2) [x]** se trouve dans F4; si la réponse est affirmative, on augmente d'une unité l'ordre de multiplicité n du sommet **x[A(y,(a_1-a_2),n]** de F3, et **i [x]** à T1; si ce sommet n'y est pas, on ajoute **x[A(y,(a_l-a_2), l]** à F3, et **A(y,(a_1-a_2)[x]** à F4, et **i [x]** à Tl.
i P j, p i L j,1	On procède mutatis-mutandis de façon analogue a celle du cas précédent.

5.2. Elaboration des réponses aux profils. L'élaboration de réponses se fait d'une façon globale. On procède ainsi:

1º) On ajoute les références {r} pertinentes à chaque terme du fichier Fl, au fur et à mesure de l'entrée des descriptions des documents dans F5.

2º) Lors de la diffusion de l'information, on exécute la recherche globale des ré-

férences relatives à chaque sommet d'accord a la commande indiquée au sommet et dans l'ordre d'apparition en F3. Par ce procédé on assigne les ensembles de références {r} à chaque sommet.

3º) On éditera les descriptions des documents (à partir de l'information enregistrée a F5) dont les références figurent dans le sommet-racine associé à chaque profil.

6. CONCLUSION

On a donné une méthode pour traiter globalement et simultanément un système de profils, lequel, en général, améliore la méthode habituelle de traiter les profils séparément. Cette méthode a été implémentée dans le système SABINI de l'Instituto Nacional de Industria, et elle est opérative dans ses centres de documentation.

Pour déterminer l'efficacité de la méthode, on peut introduire une mesure du degré de superposition des profils intégrés dans l'arborescence. Pour cela, nous employons le degré de multiplicité des sommets; supposons que l'arborescence a m sommets (x_1, x_2,...x_m) et que chacun suit l'ordre de multiplicité n1, n2,...nm ; alors, nous définissons comme coefficient de superposition ou d'efficacité un nombre ω obtenu par l'expression suivante

$$\omega = \frac{\sum_{i=1}^{m} n_i}{m}$$

Comme n est toujours ≥ 1, on obtient le moindre d'efficacité lorsque la multiplicité de tous les sommets est 1, étant alors ω = 1, dans le cas ou tous les profils sont disjoints (dont les sommets ne sont pas communs). Le maximum d'efficacité s'obtient quand tous les profils son égaux, c'est-al-dire, si on a k profi1s et tous sont égaux, l'efficacité est donnée par

$$\omega = \frac{\sum_{i=1}^{m} k}{m} = k$$

Donc, en général, on a

$$1 \leq \omega \leq k$$

oú k est le nombre total de profils qui entrent dans le système.

7. REMERCIEMENTS

Nous remercions Luis Angel García Melero et Carmen López de Sosoaga Torija de la lecture de ces pages, ainsi que de leurs suggérences et commentaires au cours de leur élaboration, et à Felisa Casaseca d'avoir traduit au Français ces lignes.

8. REFERENCES

DEWEZE, André
Diffusion sélective de 1'information
Profil documentaire. Dans:
L'accès en ligue aux bases documentaires
Paris, Masson, 1983
pp. 130-136

DOLAN, D. R. et KREMIN, M. C.
The quality control of search analysts
On-line, 3, NO. 2,8-16 (1979)

GARCIA MELERO, Luis Ángel
SABINI: Sistema de Organización automatizada de bibliotecas, Dans: Jornadas Españolas de Documentaci6n Automatizada (1.1984. Madrid) Primeras Jornadas Españolas de Documenta ci6n Automatizada; Madrid, 20-21 de Noviembre de 1984.- Madrid: ICYT, 1984.- P. 291-303.

HALL, J. L. et DEWE, A.
Online information retrieval 1976-1979 on international bibliography
London, Aslib, 1980 - 230 pp.

LANCASTER, F.W.
On-line Information Retrieval. Dans:
Information Retrieval Systems:
Characteristics, Testing and Evaluation
New York, John Wiley 1979

MEADOW, Charles T. et COCHRANE, Pauline
Storing Searches and SDI. Dans:
Basic of online Searching
New York, John Wiley, 1981
pp 118-126