Ann. sc. math. Québec, **13** (1), 1989, 19–38.

COMPLEXITÉ DE STRUCTURES DE TREILLIS

ROBERT GODIN

RÉSUMÉ. Certains résultats sont présentés au sujet de la complexité d'une structure particulière de treillis fini qui est utile au problème de dépistage de l'information. Cette structure est appelée un double-treillis parce qu'elle est constituée de deux treillis isomorphes et duaux l'un par rapport à l'autre. Une borne supérieure sur la cardinalité du double-treillis est obtenue. Des formules sont dérivées dans un cadre aléatoire uniforme pour la cardinalité moyenne du double-treillis et d'une généralisation possible de la structure.

ABSTRACT. Results are presented on the complexity of a finite lattice structure that is of interest in information retrieval. Such a structure is called a double-lattice because it is formed of two isomorphic dual lattices. An upper bound on the number of elements of the double-lattice is obtained. Formulas are derived in a random uniform context for the mean number of elements of the double-lattice and of a possible generalisation.

1. Introduction. La théorie des treillis joue un rôle important dans le cadre du problème de dépistage de l'information [Salton, 1968]. En particulier, les treillis ont servi de cadre conceptuel pour la construction automatique de thesaurus [Sparck Jones, 1971]. Récemment, une nouvelle approche au problème de dépistage de l'information basée sur la navigation dans le diagramme de Hasse d'une structure particulière de treillis a été proposée [Godin et al., 1986; Godin et Gecsei, 1986]. La structure a été baptisée "double-treillis" parce qu'elle est constituée de deux treillis isomorphes et duaux l'un par rapport à l'autre. Cet article présente certains résultats au sujet de la complexité du double-treillis qui ont des implications importantes pour la mise en oeuvre informatique de la nouvelle approche.

La section deux définit la structure de double-treillis. La section trois traite de la complexité du double-treillis. La section quatre porte sur la complexité d'une extention possible de la structure.

2. Définition de la structure du double-treillis. La structure de double-treillis est totalement déterminée par une relation, notée R, entre deux ensembles finis, notés $\mathcal{O}$ et $\mathcal{D}$, tel qu'illustré dans l'exemple du Tableau 1. Dans le cadre du problème de dépistage d'information, l'ensemble $\mathcal{O}$ correspond à un ensemble d'objets à accéder et l'ensemble $\mathcal{D}$ correspond à un ensemble de descripteurs. La représentation matricielle de R, notée M_R, sera utile par la suite (voir la Figure 1).

Le *double-treillis*, noté $\mathcal{H}^*$, sera défini dans ce qui suit à partir de la relation R. La Figure 2 montre le diagramme de Hasse du double-treillis correspondant aux données du Tableau 1.

Définissons d'abord une fonction, notée f, qui associe à un descripteur d, l'ensemble des objets en relation R avec lui (voir le Tableau 2).

DÉFINITION 1.

$$f : \mathcal{D} \to 2^{\mathcal{O}}$$
$$f(d) = \{o \in \mathcal{O} : (o, d) \in R\}.$$

À partir de f sont définis les ensembles d'objets, $F(D)$, qui sont en relation avec tous les descripteurs de l'ensemble D en prenant l'intersection des ensembles $f(d)$ correspondant à chaque descripteur d dans D.

Reçu le 8 septembre 1987 et, sous forme révisée, le 19 février 1988.
Ce travail a été subventionné en partie par le CRSNG, subvention A-9184.

TABLEAU 1.
Exemple de relation R entre un ensemble d'objets et un ensemble de descripteurs.

$o \in \mathcal{O}$	$d \in \mathcal{D}$
1	A
1	F
1	83
1	4M
2	A
2	F
2	84
2	6M
3	A
3	R
3	84
3	6M
4	C
4	F
4	83
4	4M
5	C
5	T
5	84

Abréviations pour les descripteurs

A	:	auto
C	:	camion
F	:	Ford
R	:	Renault
T	:	Toyota
4M	:	$ 4000.00
6M	:	$ 6000.00
83	:	1983
84	:	1984

Descripteurs	Objets 1	2	3	4	5
A	1	1	1	0	0
C	0	0	0	1	1
F	1	1	0	1	1
R	0	0	1	0	0
T	0	0	0	0	1
83	1	0	0	1	0
84	0	1	1	0	1
4M	1	0	0	1	0
6M	0	1	1	0	0

Figure 1. La matrice M_R pour les données du Tableau 1.

DÉFINITION 2.

$$F : 2^{\mathcal{D}} \to 2^{\mathcal{O}}$$

$$F(D) = \bigcap_{d \in D} f(d).$$

En particulier, nous adoptons la convention $F(\varnothing) = \mathcal{O}$. F correspond donc à la fonction d'extraction d'un système de recherche documentaire traditionnel où les requêtes et les descriptions des objets sont des ensembles de descripteurs. On peut interpréter les requêtes comme la conjonction des descripteurs de l'ensemble, la fonction d'extraction correspondant alors à l'implication logique [Dabrowski, 1975]. Symétriquement sont définies

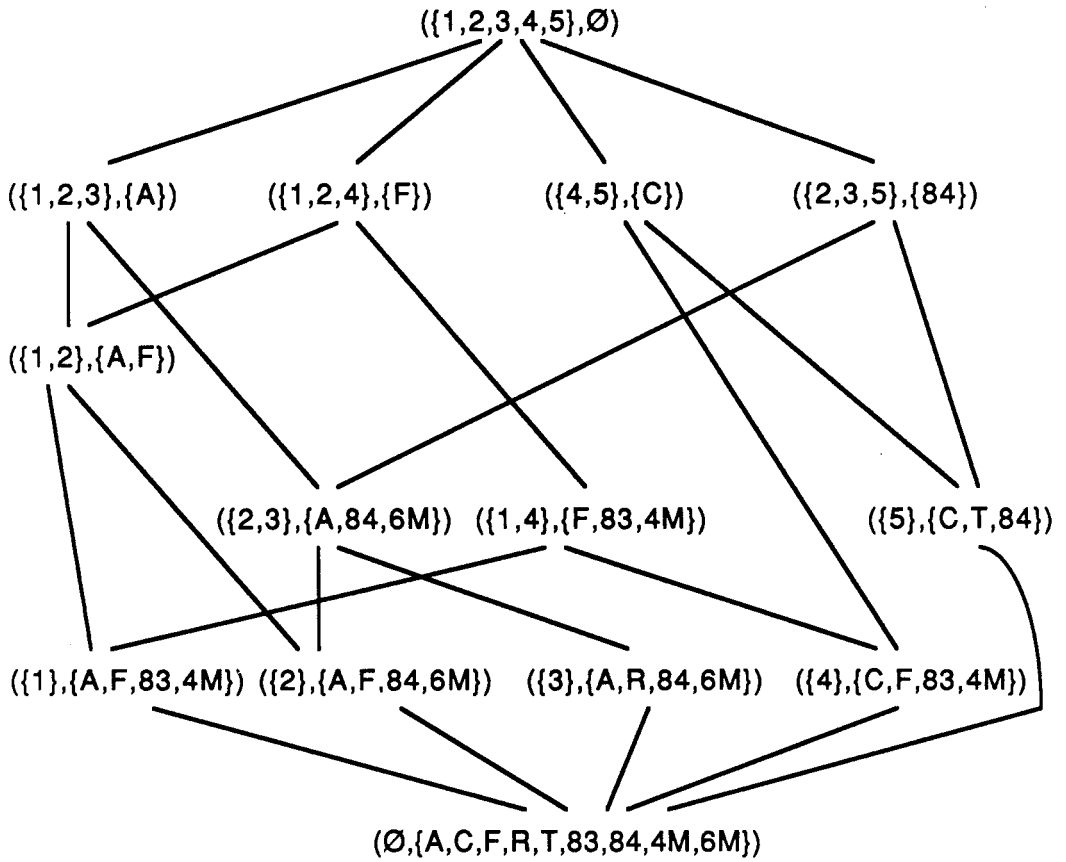

Figure 2. Diagramme de Hasse de $\mathcal{H}^*$.

TABLEAU 2.
Fonctions f et g pour l'exemple du Tableau 1.

f

d	$f(d)$
A	{1, 2, 3}
C	{4, 5}
F	{1, 2, 4}
R	{3}
T	{5}
83	{1, 4}
84	{2, 3, 5}
4M	{1, 4}
6M	{2,3}

g

o	$g(o)$
1	{A, F, 83, 4M}
2	{A, F, 84, 6M}
3	{A, R, 84, 6M}
4	{C, F, 83, 4M}
5	{C, T, 84}

les transposés $f^t = g$ (voir Tableau 2) et $F^t = G$.

DÉFINITION 3.

$$g : \mathcal{O} \to 2^{\mathcal{D}}$$
$$g(o) = \{d : (o, d) \in R\}.$$

DÉFINITION 4.

$$F^t = G : 2^{\mathcal{O}} \rightarrow 2^{\mathcal{D}}$$
$$G(O) = \bigcap_{o \in O} g(o).$$

En particulier, $G(\varnothing) = \mathcal{D}$. Le double-treillis $\mathcal{H}^*$ est défini à partir du concept de couple complet.

DÉFINITION 5. Un couple d'ensembles (O, D), $O \subseteq \mathcal{O}$ et $D \subseteq \mathcal{D}$ est complet pour la relation R s'il satisfait aux propriétés suivantes:

a) $O = F(D)$
b) $D = G(O)$.

DÉFINITION 6. $\mathcal{H}^*$ est l'ensemble de tous les couples complets de la relation R.

Il est important de noter la symétrie dans les définitions précédentes ainsi que la symétrie dans la définition de $\mathcal{H}^*$ par rapport aux descripteurs et aux objets. Il est facile de démontrer [Godin et al., 1986] que les axiomes d'un treillis sont satisfaits par $\mathcal{H}^*$.

Une autre façon de générer le double-treillis est de considérer les familles d'ensembles suivantes:

DÉFINITION 7.

$$\mathcal{F} = \{f(d) : d \in \mathcal{D}\} \quad \text{et}$$
$$\mathcal{F}^* = \{F(D) : D \in 2^{\mathcal{D}}\}.$$

Similairement, les transposés sont définies:

DÉFINITION 8.

$$\mathcal{F}^t = \mathcal{G} \ = \{g(o) : o \in \mathcal{O}\} \quad \text{et}$$
$$\mathcal{F}^{*t} = \mathcal{G}^* = \{G(O) : O \in 2^{\mathcal{O}}\}.$$

$\mathcal{F}^*$ est donc la fermeture de $\mathcal{F}$ sous l'intersection de même que $\mathcal{G}^*$ pour $\mathcal{G}$. Il est facile de démontrer [Godin, 1986] que $\mathcal{F}^*$ et $\mathcal{G}^*$ sont deux treillis isomorphes et duaux l'un par rapport à l'autre en identifiant chaque ensemble d'objet $F(D)$ dans $\mathcal{F}^*$ avec l'ensemble de descripteurs $G(F(D))$ qui apparaissent avec tous les objets de $F(D)$. Cet isomorphisme permet de ne considérer qu'une seule structure de treillis formée de l'ensemble des couples d'éléments correspondants de $\mathcal{F}^*$ et $\mathcal{G}^*$ qui sont précisément les couples complets de $\mathcal{H}^*$ [Godin, 1986].

3. Complexité du double-treillis. Cette section traite de la complexité du double-treillis, ou plus précisément, de sa cardinalité, notée $\|\mathcal{H}^*\|$. Ce nombre dépend de plusieurs paramètres dont, entre autres, le nombre moyen de descripteurs par objet, noté k, le nombre d'objets, noté n, et le nombre de descripteurs, noté m. Après avoir déterminé une borne supérieure générale par rapport à ces paramètres, des formules de complexité dans un cadre aléatoire uniforme de distribution des descripteurs seront dérivées.

3.1. Borne supérieure pour la cardinalité du double-treillis. Dans cette sous-section, une borne supérieure pour $\|\mathcal{H}^*\|$ sera dérivée par une récurrence simple basée sur la proposition 1 qui caractérise l'ensemble des couples du double-treillis après l'adjonction d'un nouvel objet par rapport à l'ensemble des couples avant l'adjonction.

En supposant que $\|\mathcal{O}\| = n$, le nombre maximal de sous-ensembles est 2^n mais heureusement la taille du double-treillis ne croît pas exponentiellement avec n. En effet, la proposition 1 garantit que l'ajout d'un nouvel objet, noté o', où $\|g(o')\| = k$ ne peut générer qu'au maximum 2^k nouveaux noeuds étant donné que leurs ensembles de descripteurs sont nécessairement des sous-ensembles de $g(o')$.

Dans ce qui suit, la notation X' servira à représenter l'ensemble ou la fonction X, selon le cas, après adjonction de l'objet o', lorsque la distinction est nécessaire. Par exemple, dans le théorème suivant, $\mathcal{G}^*$ représente le treillis formé des ensembles de descripteurs (voir définition 8) et $\mathcal{G}^{*\prime}$ le treillis résultant de l'adjonction du nouvel objet o'.

PROPOSITION 1. *Soit $\mathcal{G}^{*\prime}$ le treillis mis à jour et $\mathcal{D}'$ l'ensemble universel de descripteurs après adjonction de $g(o')$ au treillis $\mathcal{G}^*$. Alors*

$$D' \in \mathcal{G}^{*\prime} - (\mathcal{G}^* \cup \{g(o'), \mathcal{D}'\}) \rightarrow \exists D \in \mathcal{G}^*,\ D' = D \cap g(o').$$

PREUVE: Soit $D' = \bigcap_{o \in O'} g'(o)$ un élément de $\mathcal{G}^{*\prime}$.

Si $o' \notin O'$ alors $D' = \mathcal{D}'$ ou $D' \in \mathcal{G}^*$. Dans le cas où $o' \in O'$, à moins que $\|O'\| = 1$, auquel cas $D' = g(o')$, on a $D' = g(o') \cap G(O)$ où $O \subseteq O$, donc $G(O) \in \mathcal{G}^*$. □

PROPOSITION 2.

$$\|\mathcal{G}^*\| \leq n\ 2^K \text{ pour } n \geq 1,\ k = \max \|g(o)\|.$$

PREUVE: La preuve se fera par récurrence sur n, le nombre d'objets. Pour $n = 1$, $\mathcal{G}^*$ ne contient qu'un seul élément soit $\mathcal{D}$, l'ensemble de tous les descripteurs qui correspond aussi à l'ensemble des descripteurs du seul objet.

Supposons l'hypothèse vérifiée pour n. En omettant $\mathcal{D}$ qui est toujours présent car $G(\varnothing) = \mathcal{D}$, la proposition 1 garantit que, tout nouvel ensemble de descripteurs sauf peut-être $g(o')$, produit par l'adjonction d'un nouvel objet, est formé par l'intersection de l'ensemble des descripteurs du nouvel objet et d'un ensemble de descripteurs du treillis avant mise à jour. Il s'ensuit qu'un nouvel ensemble de descripteurs, D', est nécessairement inclus dans l'ensemble de descripteurs du nouvel objet, $o' : D' \subseteq g(o')$. Si $\|g(o')\| = k$, le nombre de nouveaux sous-ensembles est donc limité par 2^k. Donc pour le treillis $\mathcal{G}^{*\prime}$, $\|\mathcal{G}^{*\prime}\| \leq \|\mathcal{G}^*\| + 2^k \leq (n+1)2^K$, où $K = \max \|g(o)\|$. □

Comme chaque noeud de $\mathcal{H}^*$ est identifiable à l'ensemble de descripteurs correspondant de $\mathcal{G}^*$, on déduit que $\|\mathcal{H}^*\| <= 2^{K*}n = O(n)$ où K est le nombre maximum de descripteurs par objet. Par conséquent, le double-treillis croît linéairement par rapport à n dans le pire cas. Cette borne supérieure est cependant très large à cause du facteur exponentiel en K. Pour une valeur plausible en pratique de $K = 10$ par exemple, cette borne supérieure est considérable. On peut appliquer un raisonnement symétrique pour les ensembles d'objets et obtenir une borne supérieure linéaire en $m = \|\mathcal{D}\|$.

Des calculs de complexité dans un cadre aléatoire uniforme de distribution des descripteurs, des simulations et des applications expérimentales montrent que dans un contexte réel d'application, le rapport $\|\mathcal{H}^*\|/n$ est assez stable et de beaucoup inférieur à 2^K [Godin, 1986]. La section suivante montre comment dériver les formules de complexité pour $\|\mathcal{H}^*\|$ dans un cadre aléatoire uniforme de distribution des descripteurs.

3.2. Complexité dans un cadre aléatoire uniforme. Le problème traité dans cette sous-section consiste à calculer la cardinalité moyenne de $\mathcal{H}^*$, notée $\overline{\|\mathcal{H}^*\|}$, par rapport aux trois paramètres $n = \|\mathcal{O}\|$, $m = \|\mathcal{D}\|$ et k, le nombre moyen de descripteurs par objet. L'hypothèse d'une distribution uniforme des descripteurs sera adoptée. Des formules de

complexité seront dérivées pour $\overline{\|\mathcal{H}^*\|}$ sous deux hypothèses différentes. Sous la première qualifiée de *variable* étant donné que la cardinalité de $g(o)$ est supposée variable, on adopte le modèle suivant pour la distribution des descripteurs:

$$\text{Prob}[(o,d) \in R] = p = k/m,$$

des probabilités indépendantes. La cardinalité de $g(o)$ suit donc une distribution binomiale de moyenne k. La notation suivante sera utilisée pour représenter un coefficient binomial:

$$C(a,b) = b!/a!(b-a)!.$$

Sous la seconde hypothèse qualifiée de *fixe*, $\|g(o)\| = k$ pour tout objet o. Cette hypothèse donne des résultats comparables et comporte certains avantages au point de vue calculs. On suppose encore une distribution uniforme sauf qu'une combinaison d'*exactement* k parmi m descripteurs choisis aléatoirement est assignée à chaque objet o. La section 3.2.1 montre comment obtenir les formules de complexité sous les deux hypothèses précédentes.

3.2.1. Dérivation des formules de complexité. Avant d'aborder chacune des deux hypothèses, un cadre commun permettant de faire ces calculs est présenté dans cette sous-section. Considérons la probabilité $PN(i,j)$ qu'un couple particulier $H = (O,D)$ de j objets, $\|O\| = j$, et i descripteurs, $\|D\| = i$, soit dans $\mathcal{H}^*$. Dans la matrice M_R, si l'on regroupe des j objets dans les colonnes 1 à j de la matrice et les i descripteurs dans les lignes 1 à i en permutant les lignes et colonnes à cet effet, on obtient une matrice $M_R[i,j]$ découpée en 4 sous-matrices (Figure 3).

Objets

		1	2	.	.	.	j	j+1	j+2	.	.	.	n
D	1	1	1	.	.	.	1	?	?	.	.	.	?
e	2	1	1	.	.	.	1	?	0	.	.	.	?
s	.	.	.				.	0	?				?
c	.	.	.	R1			.	?	?	R2			?
r	.	.	.				.	?	?				0
i	i	1	1	.	.	.	1	?	?	.	.	.	?
p	i+1	?	?	0	?	?	?	?	?	.	.	.	?
t	i+2	?	?	?	?	0	?	?	?	.	.	.	?
e	.	.					.	.	.				.
u	.	.		R3			.	.	.	R4			.
r	.	.					.	.	.				.
s	m	?	0	?	?	?	?	?	?	.	.	.	?

$M_R[i,j]$

Figure 3. Caractérisation d'un couple complet à partir de la matrice M_R. N.B. ? signifie que la valeur n'importe pas, c'est-à-dire qu'il peut s'agir soit d'un 0 ou d'un 1.

Par définition d'un couple complet, le noeud $H = (O, D)$ appartiendra au double-treillis $\mathcal{H}^*$ si les propriétés suivantes sont vérifiées pour les 4 sous-matrices.

$R1$: il y a des 1 partout car $(d, o) \in R$, pour tout $d \in D$, $o \in O$.

$R2$: chaque colonne doit contenir au moins un zéro, sinon l'objet correspondant appartiendrait aussi à l'ensemble d'objets du couple complet, O.

$R3$: chaque ligne doit contenir au moins un zéro, sinon le descripteur correspondant appartiendrait à l'ensemble de descripteurs du couple complet, D.

$R4$: sans importance.

Cette observation servira à calculer PN en le représentant par un produit de probabilités correspondant aux propriétés de chaque sous-matrice $R1$, $R2$, $R3$. Ayant trouvé PN, la cardinalité moyenne sera obtenue en faisant la somme sur tous les couples H possibles.

$$\overline{\|\mathcal{H}^*\|} = \sum_{i=0}^{m} \sum_{j=0}^{n} C(i,m)\ C(j,n)\ PN(i,j). \tag{1}$$

En particulier, la somme des termes de (1) correspondant à $j = 0$ donne la probabilité qu'un ensemble D soit l'intersection d'aucun ensemble d'objets, Prob$[(\varnothing, D) \in \mathcal{H}^*]$. Pour $i = 0$, c'est Prob$[(O, \varnothing) \in \mathcal{H}^*]$.

3.2.1.1. Cardinalité moyenne de $\mathcal{H}^*$ dans le cas uniforme variable. Soit $p =$ Prob$[(o, d) \in R]$, une constante fixe indépendante de o et d. On en déduit que $k = mp$ sera le nombre moyen de descripteurs associés à chaque objet. Comme les probabilités sont indépendantes pour $R1$, $R2$, $R3$, on obtient directement:

$$P1 = p^{ij}$$
$$P2 = (1 - p^i)^{(n-j)}$$
$$P3 = (1 - p^j)^{(m-i)}$$

et en conséquence,

$$PN(i,j) = p^{ij}(1 - p^i)^{(n-j)}(1 - p^j)^{(m-i)}. \tag{2}$$

En remplaçant (2) dans (1), on obtient la formule suivante:

$$\overline{\|\mathcal{H}^*\|} = \sum_{i=0}^{m} \sum_{j=0}^{n} C(i,m)\ C(j,n)\ p^{ij}(1 - p^i)^{(n-j)}(1 - p^j)^{(m-i)}. \tag{3}$$

En effectuant successivement les deux substitutions suivantes qui sont des conséquences directes de la formule du binôme de Newton,

$$(1 - p^j)^{(m-i)} = \sum_{q=0}^{m-i} C(q, m-i)(-1)^q p^{jq}$$

$$\sum_{j=0}^{n} C(j,n)[(p^{i+q})/(1 - p^i)]^j = (1 + [(p^{i+q})/(1 - p^i)])^n,$$

on obtient la formule équivalente suivante:

$$\overline{\|\mathcal{H}^*\|} = \sum_{i=0}^{m} \sum_{q=0}^{m-i} C(i,m)\ C(q, m-i)(-1)^q [1 - p^i + p^{i+q}]^n. \tag{4}$$

3.2.1.2. Cardinalité moyenne de $\mathcal{H}^*$ dans le cas uniforme fixe. La situation est plus complexe si l'on fixe le nombre de descripteurs associés à chaque objet, $\|g(o)\| = k$:

$$P1 = [C(k-i, m-i)/C(k,m)]^j.$$

$C(k-i, m-i)$ est le nombre de choix qui restent lorsqu'on fixe i éléments parmi k.

$$P2 = [1 - C(k-i, m-i)/C(k,m)]^{(n-j)}.$$

Lorsque $i > k$, la convention suivante s'applique [Riordan, 68]:

$$C(-p, q) = 0, \text{ pour tout entier } q \text{ et pout tout entier positif } p.$$

Cette situation se présente ici car i varie de 0 à $m > k$, en général. Plusieurs termes de la sommation seront nuls parce qu'il ne peut y avoir d'ensembles de cardinalité supérieure à k sauf pour l'ensemble $\mathcal{D}$ $(i = m)$.

Pour la sous-matrice $R3$, la probabilité $P3$ n'est pas indépendante de l'événement $R2$. On doit tenir compte du fait que i descripteurs sont déjà attribués à chaque objet de O. $P3(j, i, m, k)$ peut s'interpréter comme la probabilité que l'intersection de j ensembles de $k - i$ parmi $m - i$ descripteurs soit vide (étant donné qu'il y aura au moins un zéro dans chaque ligne). $P3$ peut se formuler récursivement de la façon suivante:

$$P3(j, i, m, k) = 1 - \sum_{q=1}^{m-i} C(q, m-i)\ P(q). \tag{5}$$

$P(q)$ est la probabilité qu'il y ait q descripteurs particuliers dans l'intersection des j ensembles. Pour voir réapparaître $P3$ dans le calcul de $P(q)$, on n'a qu'à découper $R3$ en deux sous-matrices en regroupant les q descripteurs dans les lignes $i+1$ à $i+q$ (Figure 4).

Les propriétés pour $R31$ et $R32$ sont analogues à $R1$ et $R3$. On obtient donc:

$$P(q) = P31\, P32$$

où

$$P31 = [C(k-i-q, m-i-q)/C(k-i, m-i)]^j, \qquad 1 \leq k$$
$$P32 = P3(j, q, m-i, k-i).$$

D'où l'équation récursive pour $P3$:

$$P3(j, i, m, k) = 1 - \sum_{q=1}^{m-i} C(q, m-i) \left[\frac{C(k-i-q, m-i-q)}{C(k-i, m-i)}\right]^j P3(j, q, m-i, k-i). \tag{6}$$

La condition terminale de la récurrence est

$$P3(j, i, m, k) = 1 \quad \text{pour } i \geq m \text{ (la sommation devient dégénérée).}$$

La solution suivante a été obtenue par une méthode ad hoc et peut être vérifiée facilement en la substituant dans l'équation (6):

$$P3(j, i, m, k) = \begin{cases} \sum_{r=0}^{m-i} (-1)^r\ C(r, m-i) \left[\frac{C(k-i-r, m-i-r)}{C(k-i, m-i)}\right]^j & \text{pour } i < m \\ 1 & \text{pour } i = m \end{cases}. \tag{7}$$

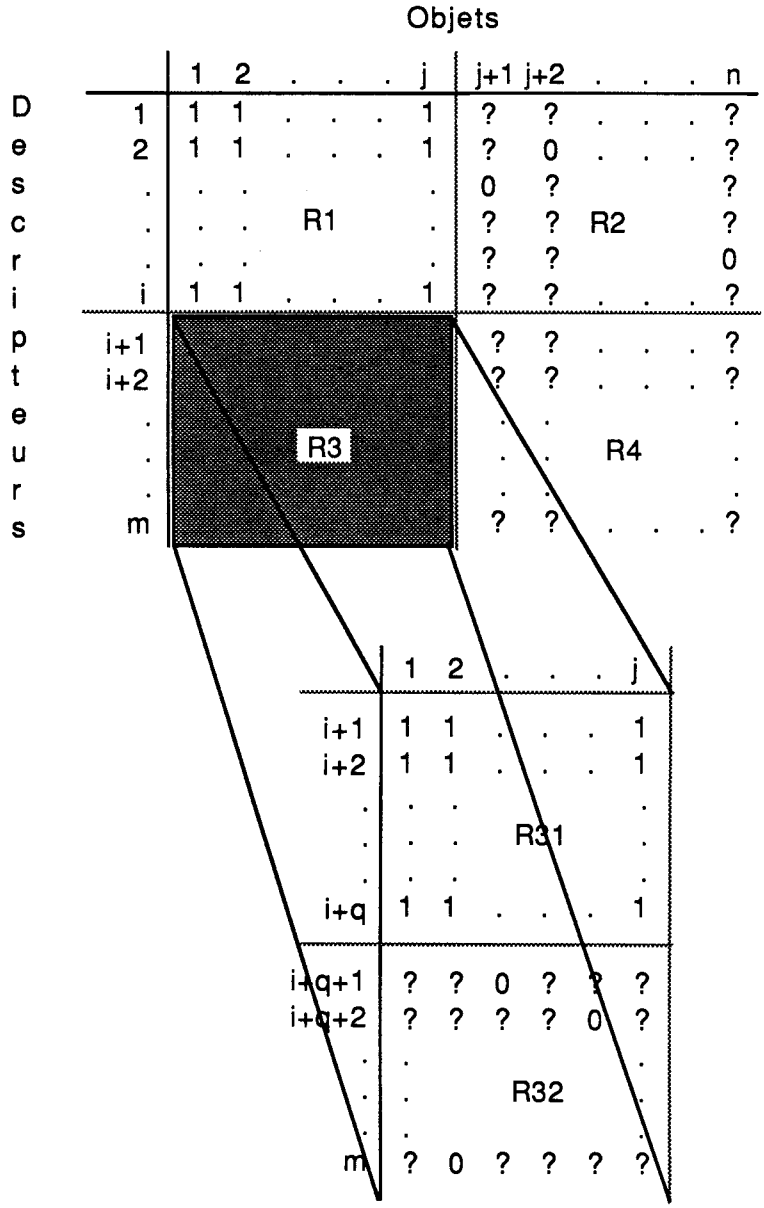

Figure 4. Découpage de $R3$ en deux sous-matrices $R31$ et $R32$ dans la formulation récursive de $P3$. N.B. ? signifie que la valeur n'importe pas, c'est-à-dire qu'il peut s'agir soit d'un 0 ou d'un 1.

On obtient en remplaçant dans (1):

$$\overline{\|\mathcal{H}^*\|} = \sum_{i=0}^{m}\sum_{j=0}^{n} C(i,m)\ C(j,n)\ \{[C(k,m) - C(k-i,m-i)]^{(n-j)}/C(k,m)^n\}$$

$$\times \sum_{r=0}^{m-i}(-1)^r\ C(r,m-i)\ C(k-i-r,m-i-r)^j. \quad (8)$$

En effectuant la substitution suivante, conséquence directe de la formule du binôme

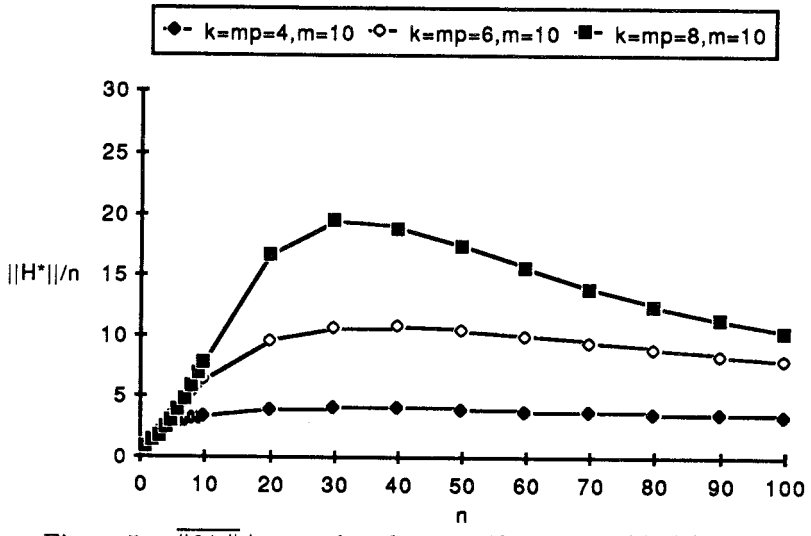

Figure 5-a. $\overline{\|\mathcal{H}^*\|}/n$ vs n dans le cas uniforme variable (3), $m = 10$.

de Newton,

$$\sum_{j=0}^{n} C(j,n)\ \{C(k-i-r,m-i-r)/[C(k,m)-C(k-i,m-i)]\}^j$$
$$= (1+\{C(k-i-r,m-i-r)/[C(k,m)-C(k-i,m-i)]\})^n,$$

on obtient une formulation plus compacte:

$$\overline{\|\mathcal{H}^*\|} = \sum_{i=0}^{m}\sum_{r=0}^{m-i}(-1)^r\ C(i,m)\ C(r,m-i)$$
$$\{[C(k,m)-C(k-i,m-i)+C(k-i-r,m-i-r)]/C(k,m)\}^n. \quad (9)$$

3.2.2. Analyse des formules. Les Figures 5-a-c et 6-a-d montrent des valeurs comparables de $\overline{\|\mathcal{H}^*\|}/n$ pour les deux formules de complexité (3) et (9) avec m fixé et $k = 4$, 6, 8. L'envergure des calculs à effectuer nous a contraint à limiter la taille des paramètres, le nombre de valeurs calculées et parfois à introduire des approximations. Les courbes de ces figures sont formées par interpolation linéaire entre les valeurs calculées. Afin de permettre des valeurs plus élevées pour les paramètres m et n sans débordement pour les calculs dans le cas uniforme fixe, ceux-ci ont été effectués à partir d'une formule de récurrence semblable à (6) et sans tenir compte du terme $i = m$ correspondant à l'ensemble universel de descripteurs. Par ailleurs, ce terme vaudra toujours 1. Ceci permet de laisser tomber tous les termes pour i plus grand que k. Les Figures 5-a et 6-a permettent de constater que les formules de complexité sous les deux hypothèses se comportent de façon similaire et que le maximum atteint est de beaucoup inférieur à la borne supérieure dérivée plut tôt, 2^K. Lorsque n tend vers l'infini, $\overline{|\,|\mathcal{H}^*|\,|}/n$ tend vers 0, le numérateur étant borné par 2^m:

$$\lim_{n\to\infty} \overline{\|\mathcal{H}^*\|}/n = 0.$$

Ce qui est plus significatif en pratique est que pour $k \ll m$, lorsque m croît, le maximum diminue et est atteint à des valeurs beaucoup plus grandes de n.

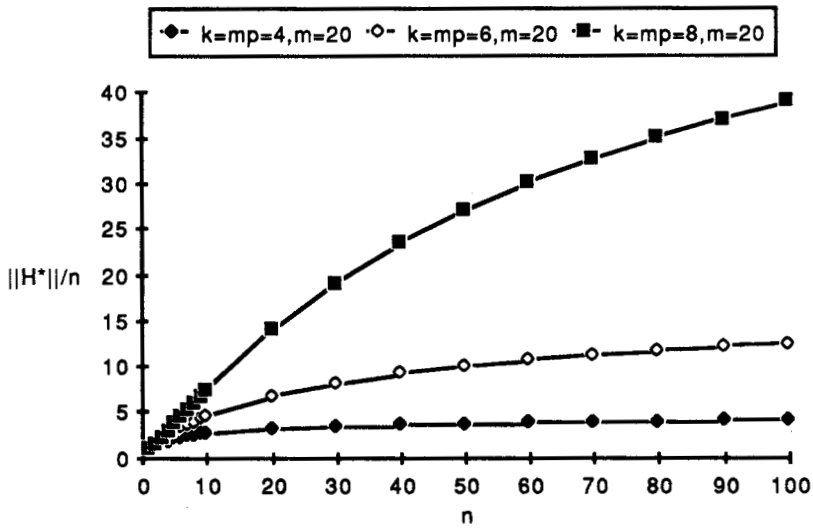

Figure 5-b. $\overline{\|\mathcal{H}^*\|}/n$ vs n dans le cas uniforme variable (3), $m = 20$.

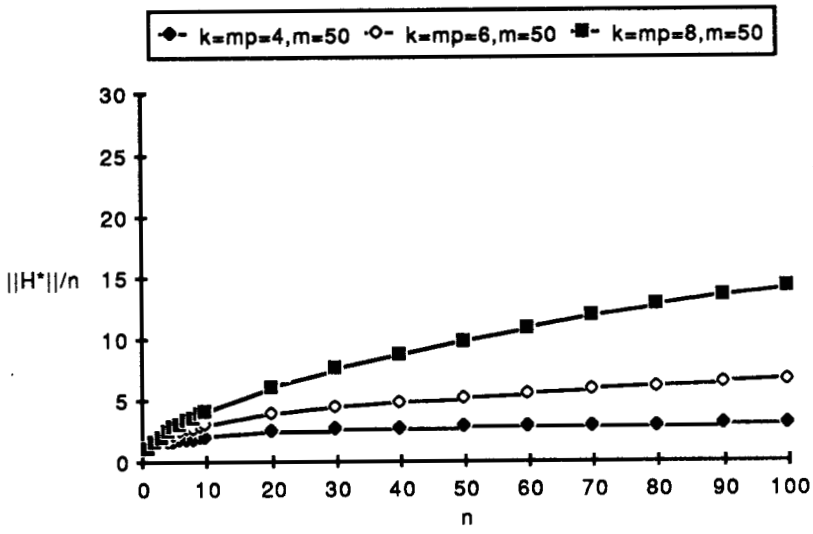

Figure 5-c. $\overline{\|\mathcal{H}^*\|}/n$ vs n dans le cas uniforme variable (3), $m = 50$.

Si l'on fixe n et fait varier m, les courbes ont une allure générale semblable. Les Figures 7-a-d montrent les courbes obtenues dans le cas fixe seulement. Lorsque m tend vers l'infini, la probabilité d'intersection non vide entre ensembles de descripteurs devient nulle et chaque objet génère un et un seul élément. On peut donc démontrer

$$\lim_{m \to \infty} \overline{\|\mathcal{H}^*\|}/n = 1.$$

Ces calculs nous donnent un premier aperçu au sujet de la complexité du double-treillis. En réalité, la distribution des descripteurs est plus complexe et difficilement modélisable [Tague et al., 1981]. Des comparaisons empiriques et des simulations sont faites dans [Godin, 1986] afin d'évaluer la robustesse des hypothèses simplificatrices au sujet de la distribution des descripteurs. On observe que les approximations obtenues pour $\overline{\|\mathcal{H}^*\|}/n$

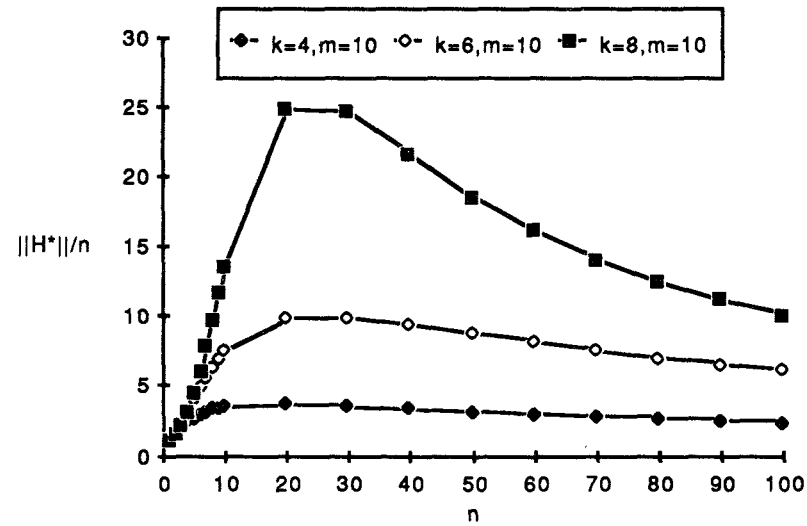

Figure 6-a. $\overline{\|\mathcal{H}^*\|}/n$ vs n dans le cas uniforme fixe (9), $m = 10$.

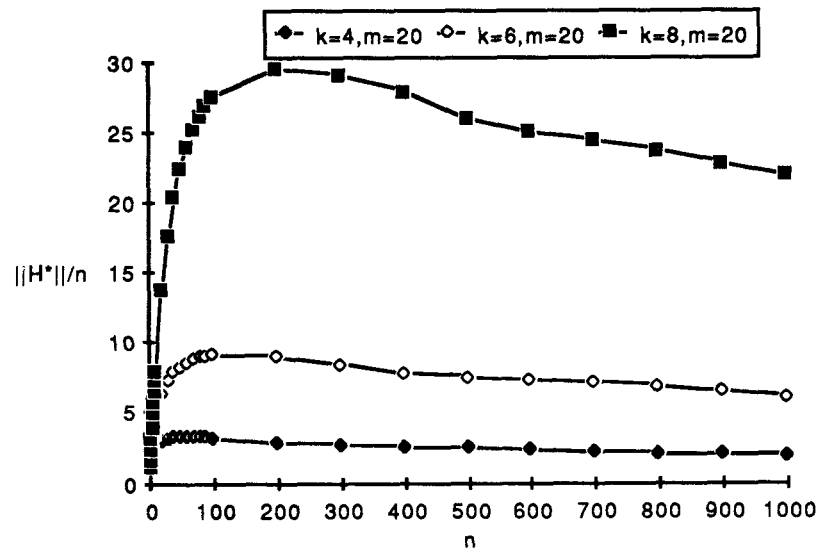

Figure 6-b. $\overline{\|\mathcal{H}^*\|}/n$ vs n dans le cas uniforme fixe (9), $m = 20$.

sont assez bonnes. Par exemple, dans une application expérimentale de recherche documentaire [Godin, 1986], les prédictions calculées pour $\overline{\|\mathcal{H}^*\|}/n$ avec des valeurs réelles pour les paramètres n, m et k, dans le cas uniforme fixe varient approximativement de 3 à 5 alors que les valeurs observées varient d'environ 4 à 8 pour différentes valeurs de n. Dans cette application, les objets correspondent à des rapports techniques tirés d'une bibliothèque d'informatique. Au total, 3042 rapports techniques ont été inclus dans le double-treillis.

4. Complexité du double-treillis avec compléments. Le double-treillis considéré jusqu'à maintenant exprime l'équivalent de requêtes formées d'ensembles de descripteurs que l'on interprète comme la conjonction des descripteurs. En théorie, il serait possible de considérer une structure plus expressive en incluant, par exemple, la pleine généralité des expressions booléennes. Dans cette section, un premier pas dans cette direction sera

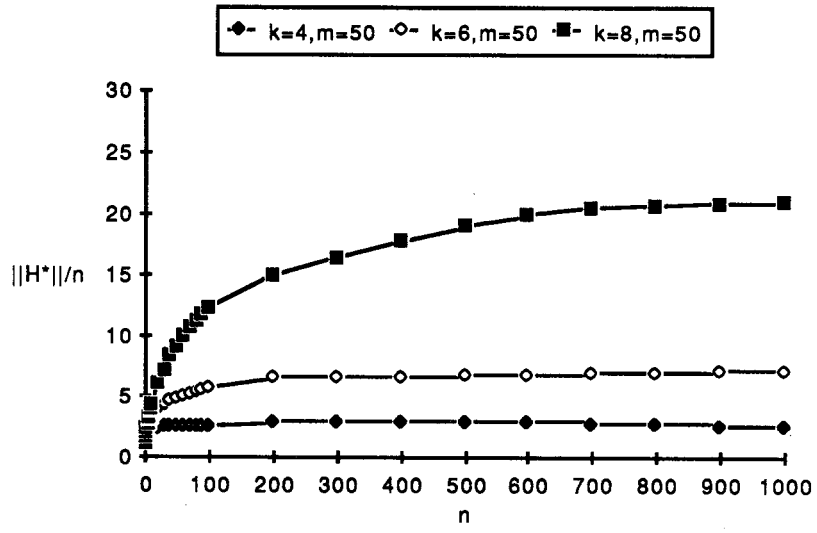

Figure 6-c. $\overline{\|\mathcal{H}^*\|}/n$ vs n dans le cas uniforme fixe (9), $m = 50$.

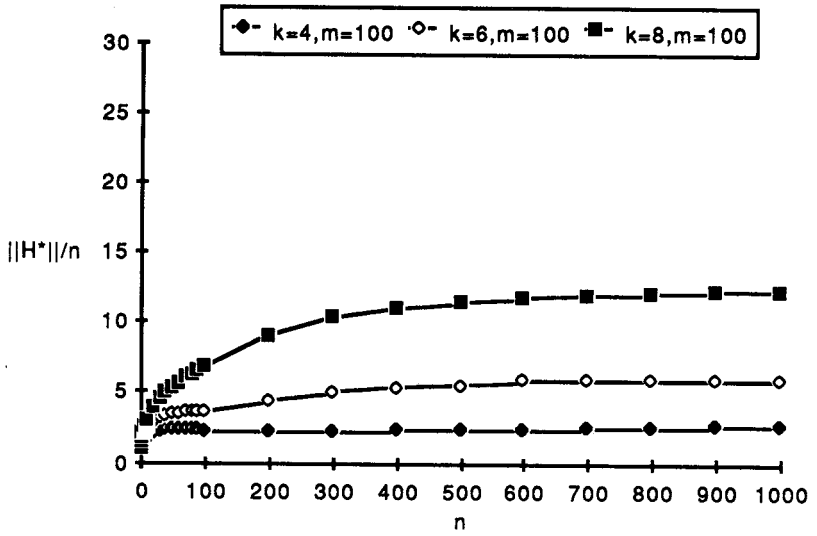

Figure 6-d. $\overline{\|\mathcal{H}^*\|}/n$ vs n dans le cas uniforme fixe (9), $m = 100$.

envisagé en incluant la possibilité de complémenter individuellement les descripteurs dans l'ensemble. Pour un descripteur donné, l'interprétation est:

$$f(d^c) = f(d)^c,$$

d^c dénote la négation du descripteur d et $f(d)^c$ est le complément ensembliste de $f(d)$ dans $\mathcal{O}$.

L'objectif visé dans l'introduction de ces compléments est de permettre à l'usager de spécifier qu'il veut les objets qui sont décrits par un ensemble de descripteurs dont certains sont complémentés, l'interprétation étant que les objets ne doivent pas être décrits par les descripteurs sous forme complémentée. Par exemple, l'usager pourrait formuler une requête du genre,

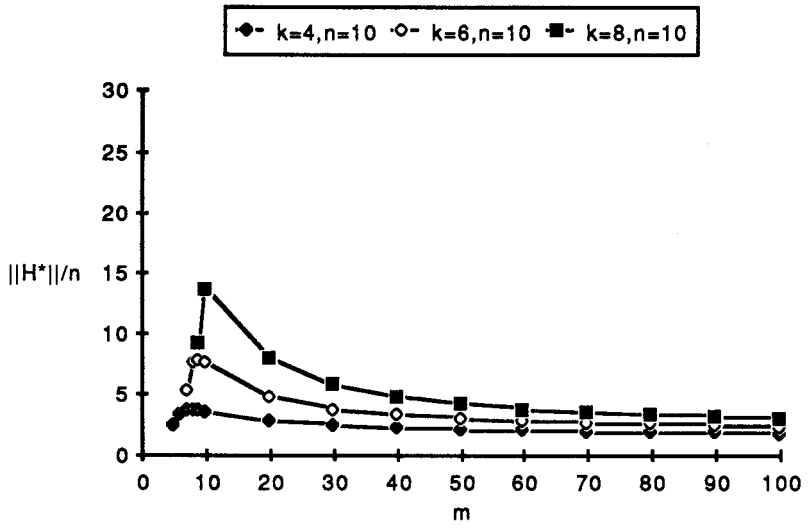

Figure 7-a. $\overline{\|\mathcal{H}^*\|}/n$ vs m dans le cas uniforme fixe (9), $n = 10$.

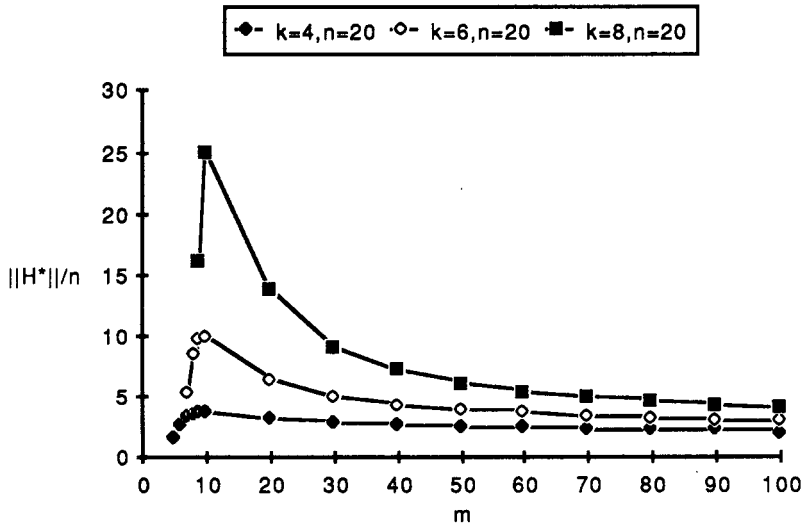

Figure 7-b. $\overline{\|\mathcal{H}^*\|}/n$ vs m dans le cas uniforme fixe (9), $n = 20$.

"les autos de l'année 1985 qui ne sont pas de marque Ford",

qui correspond à l'ensemble,

{Auto, 1985, Ford^c}.

Du point de vue génération de la nouvelle structure de double-treillis, ceci correspond à considérer les descripteurs complémentés comme des descripteurs distincts et à modifier la relation R de manière à associer à chaque objet o, les descripteurs qui ne sont pas en relation avec lui sous forme complémentée en plus de ses descripteurs usuels. Notons cette nouvelle relation R'.

$$R' = R \cup \{(o, d^c) \mid (o, d) \notin R\}.$$

En interprétant la relation R' de cette façon, on utilise exactement la même définition de couple complet que celle introduite à la section 2 en remplaçant R par R'. Lors de

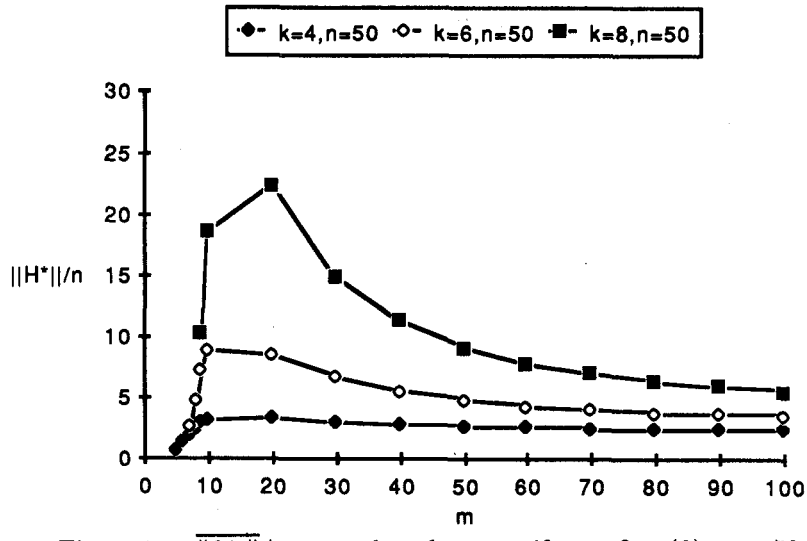

Figure 7-c. $\overline{\|\mathcal{H}^*\|}/n$ vs m dans le cas uniforme fixe (9), $n = 50$.

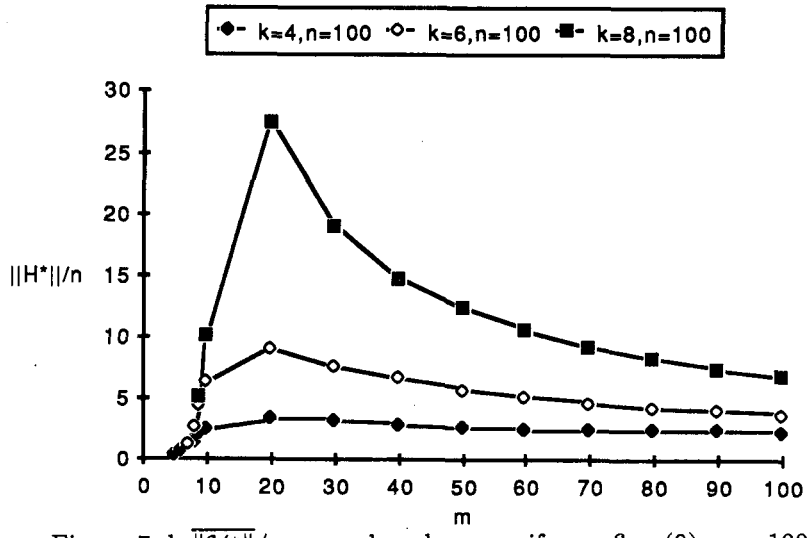

Figure 7-d. $\overline{\|\mathcal{H}^*\|}/n$ vs m dans le cas uniforme fixe (9), $n = 100$.

la génération du nouveau double-treillis, les descripteurs complémentés sont considérés comme des éléments distincts. En conservant les mêmes conventions pour les paramètres, le nombre de descripteurs sera donc $2m$. La Figure 8 montre un double-treillis sans compléments et la Figure 9 le nouveau double-treillis avec compléments correspondant. Pour le premier cas, la relation R est:

$$R = \{(1,A),(1,B),(1,C),(2,A),(2,C),(3,A),(3,B),(3,D),(4,C),(4,E)\}.$$

Pour le second cas, la relation R' devient:

$$R' = \left\{ \begin{array}{l} (1,A),(1,B),(1,C),(1,D^c),(1,E^c),(2,A),(2,B^c),(2,C),(2,D^c),(2,E^c), \\ (3,A),(3,B),(3,C^c),(3,D),(3,E^c),(4,A^c),(4,B^c),(4,C),(4,D^c),(4,E) \end{array} \right\}$$

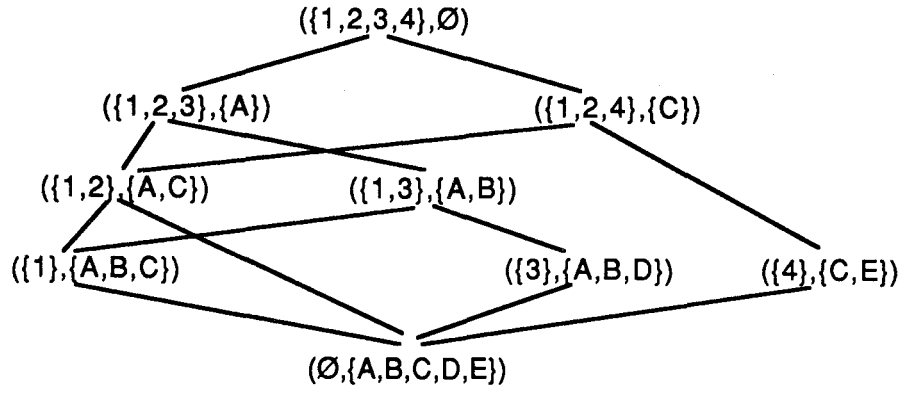

Figure 8. Exemple de double-treillis sans compléments.

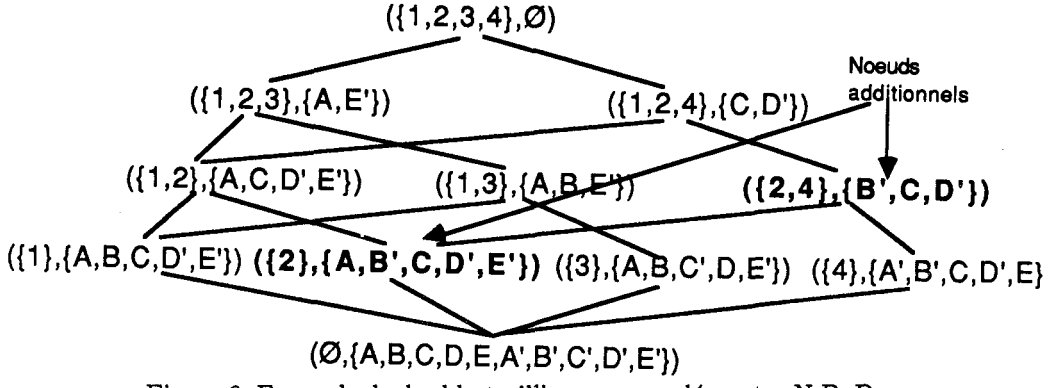

Figure 9. Exemple de double-treillis avec compléments. N.B. Dans cette figure, la notation X' représente X^c.

Dans cet exemple, il n'y a que deux noeuds additionnels dans le double-treillis avec compléments. En général cependant, la complexité additionnelle engendrée par cette généralisation est considérable.

La cardinalité du double-treillis avec compléments dans le cadre aléatoire uniforme peut se déduire de manière analogue au double-treillis sans compléments en considérant la nouvelle matrice $M_R[i, j, q]$ découpée en quatre sous-matrices semblables (Figure 10).

Par définition, un élément $H = (O, D)$, possédant j objets et i descripteurs dont q sont non complémentés, appartiendra au double-treillis avec compléments si les propriétés suivantes sont vérifiées:

$R1$: il n'y a que des 1 car $(o, d) \in R'$, pour tout $d \in D$, $o \in O$.

$R2$: chaque colonne doit contenir au moins un zéro, sinon l'objet correspondant appartiendrait aussi à l'ensemble d'objets O du couple complet.

$R3$: chaque ligne $i + 1$ à m correspondant à un descripteur qui n'apparaît pas dans D sous forme complémentée ou pas doit contenir au mois un 0 et un 1; sinon le descripteur ou son complément appartiendrait à D. On laisse tomber les autres lignes ($m+1$ à $2m$) de cette sous-matrice dans les calculs car elles sont redondantes, un descripteur et son complément ne pouvant être en relation avec le même objet (sauf pour l'ensemble vide d'objet $j = 0$ que nous omettons pour simplifier les calculs).

$R4$: sans importance.

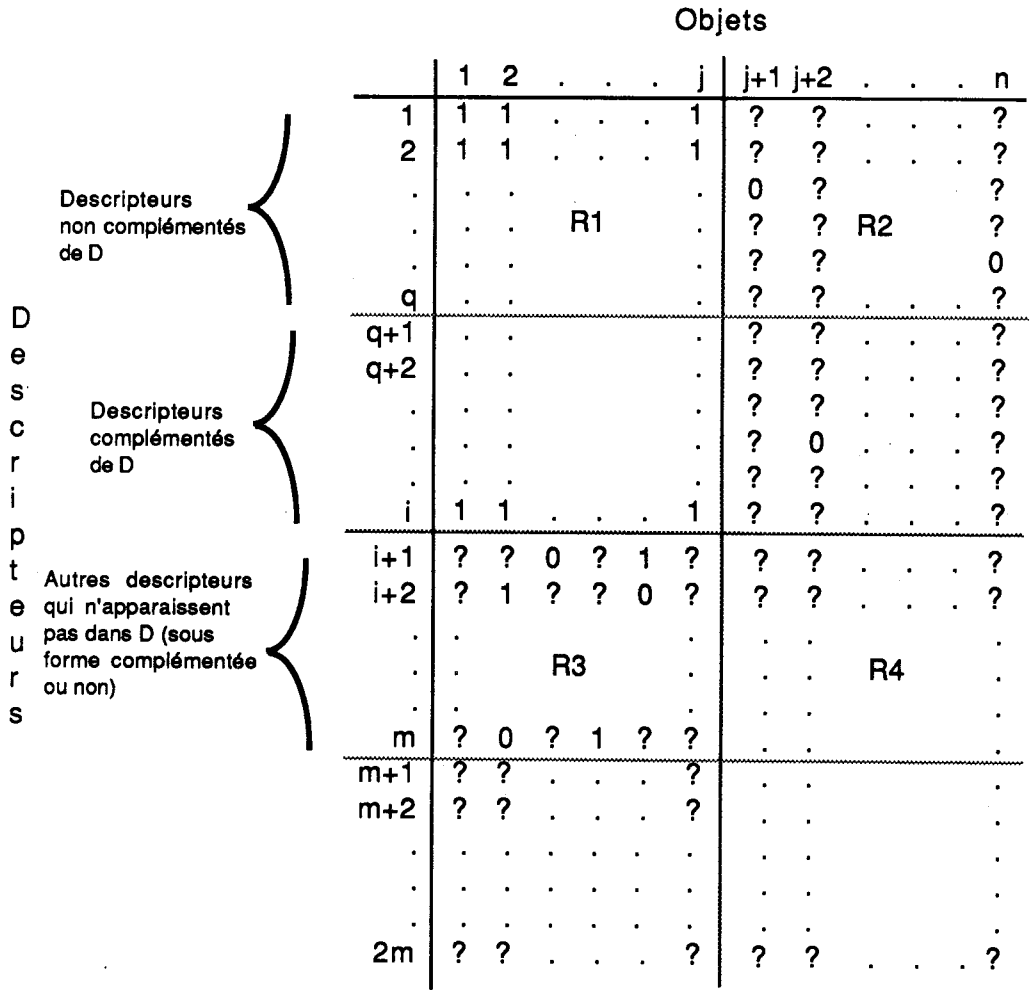

Figure 10. La matrice $M_R[i,j,q]$ découpée en quatre sous-matrices pour le double-treillis avec compléments.

On obtient donc en omettant le cas $j = 0$ pour l'hypothèse uniforme variable:

$$P1 = p^{qj}(1-p)^{(i-q)j}$$
$$P2 = [1 - p^q(1-p)^{(i-q)}]^{(n-j)}$$
$$P3 = [1 - p^j - (1-p)^j]^{(m-i)} \quad \text{où} \quad p = k/m.$$

En faisant la somme sur tous les couples possibles de j objets et i descripteurs dont q particuliers sont non complémentés, on obtient:

$$\overline{\|\mathcal{H}^*\|} = \sum_{i=0}^{m} C(i,m) \sum_{j=1}^{n} C(j,n) \sum_{q=0}^{i} C(q,i)\, P1\, P2\, P3. \tag{10}$$

Parmi les i descripteurs choisis dans les m disponibles, on en choisit q qui ne sont pas complémentés, d'où la nouvelle sommation par rapport à (1). Les Figures 11-a-c comparent la complexité moyenne des double-treillis avec et sans compléments dans l'hypothèse

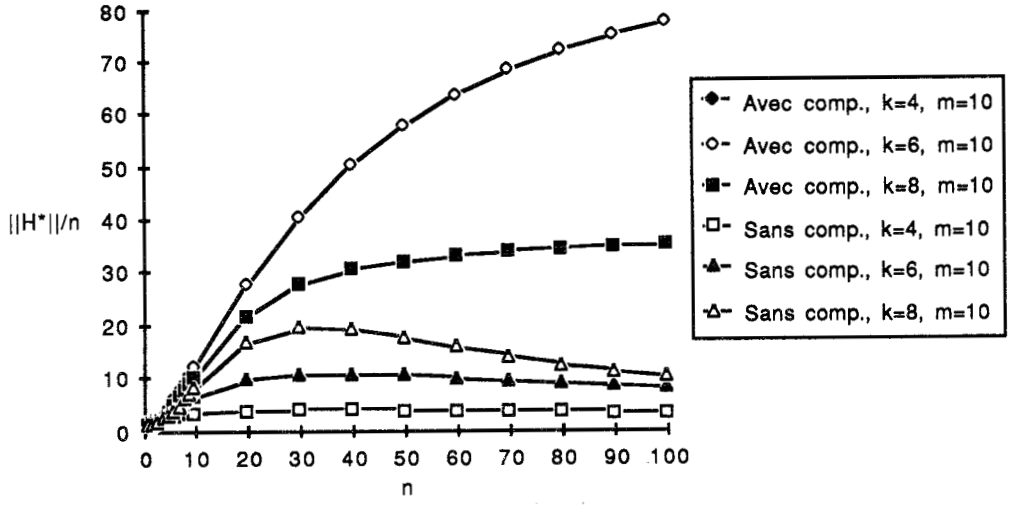

Figure 11-a. Comparaison de $\overline{\|\mathcal{H}^*\|}/n$ avec et sans compléments dans le cas uniforme variable, $m = 10$.

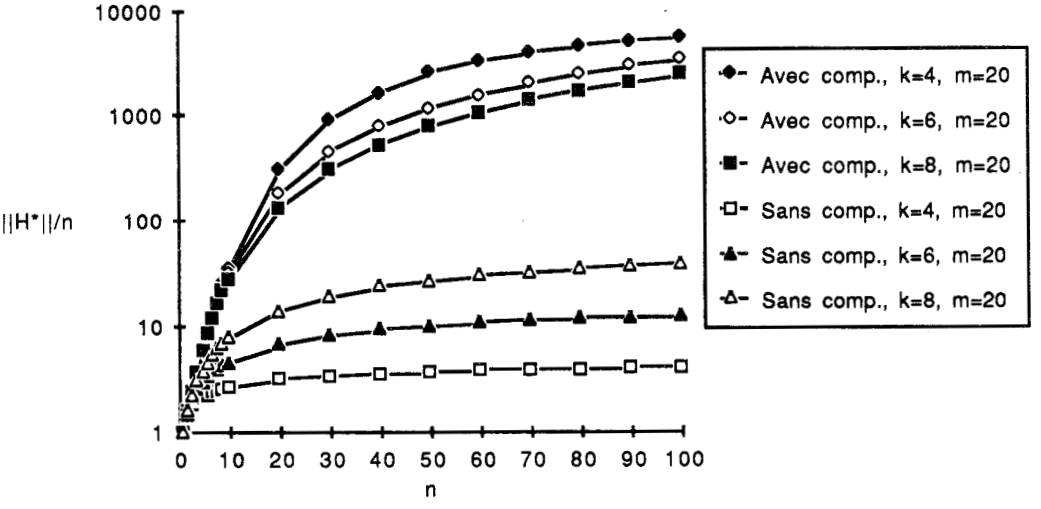

Figure 11-b. Comparaison de $\overline{\|\mathcal{H}^*\|}/n$ avec et sans compléments dans le cas uniforme variable, $m = 20$.

uniforme variable. Notons que sur la Figure 11-a, les courbes avec compléments pour $k = 4$ et $k = 6$ sont confondues à cause de la symétrie résultante dans la formule (10). Le double-treillis avec compléments exhibe une complexité de beaucoup supérieure qui s'accentue rapidement avec l'accroissement de m et qui rend prohibitive la mise en oeuvre de cette généralisation. À noter l'échelle logarithmique des abcisses qui diminue graphiquement l'ampleur des différences dans les Figures 11-b-c.

5. Conclusion. Dans cet article, nous avons traité de la complexité d'une structure particulière de treillis fini, appelée double-treillis, qui a été proposée comme support au problème de dépistage de l'information. Les résultats obtenus ont des implications importantes pour la mise en oeuvre informatique de la navigation dans le double-treillis par rapport à la complexité en espace de la structure, à la complexité des algorithmes de mise à jour et à

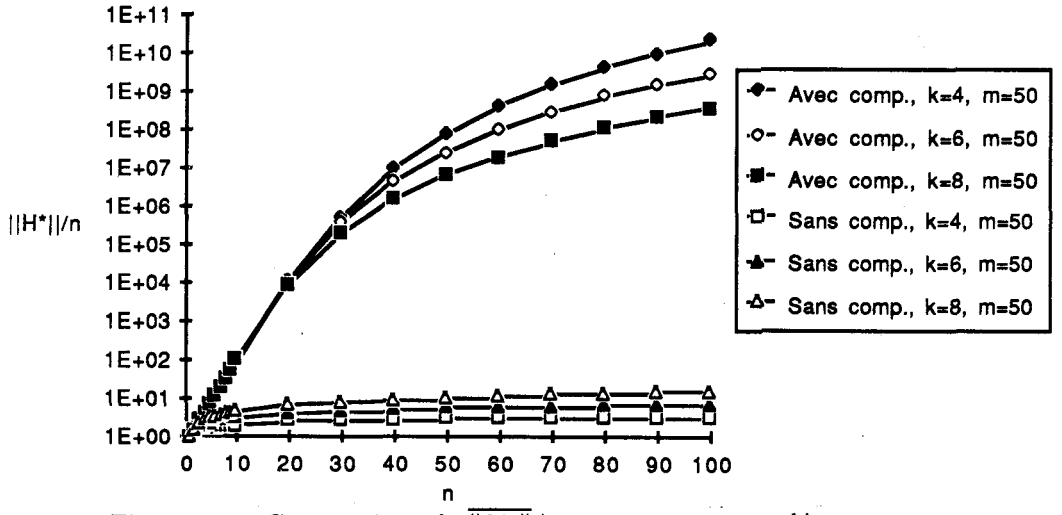

Figure 11-c. Comparaison de $\overline{\|\mathcal{H}^*\|}/n$ avec et sans compléments dans le cas uniforme variable, $m = 50$.

l'interface à l'usager.

À partir d'une caractérisation simple des modifications du double-treillis suite à l'adjonction d'un nouvel objet, une borne supérieure sur la cardinalité du double-treillis qui est linéaire en n, le nombre d'objet, et exponentielle en K, le nombre maximum de descripteurs par objet a été dérivée. Des formules de complexité obtenues sous l'hypothèse d'une distribution uniforme des descripteurs confirment un accroissement de la cardinalité moyenne qui est approximativement linéaire en n mais dont le coefficient est de beaucoup inférieur à la borne supérieure exponentielle en K.

Enfin, nous avons considéré un premier pas vers une généralisation de la structure qui consiste à incorporer la possibilité de complémenter individuellement chaque descripteur. La formule obtenue pour la cardinalité moyenne montre une explosion trop rapide du nombre d'éléments dans une perspective pratique.

D'autres aspects importants de la complexité du double-treillis restent à analyser. En ce qui a trait à la cardinalité moyenne du double-treillis, des calculs effectués dans un cadre réaliste pour les paramètres, suggère que l'accroissement soit linéaire en k, le nombre moyen de descripteurs par objet, contrairement à ce que laisserait supposer la borne supérieure exponentielle en K, le nombre maximum de descripteurs [Godin, 1986]. Ceci reste cependant à confirmer analytiquement.

La complexité de deux sous-ensembles du double-treillis, importants pour la complexité des algorithmes de mise à jour, a aussi été dérivée à partir des calculs présentés dans cet article pour le cadre aléatoire uniforme [Godin, 1986]. Cependant, certaines observations empiriques au sujet du comportement des formules restent à vérifier analytiquement.

Enfin, des analyses empiriques au sujet du nombre moyen de voisins dans le diagramme de Hasse du double-treillis montrent un accroissement logarithmique en n, le nombre d'objets [Godin, 1986]. Ceci est très important du point de vue pratique et reste à confirmer au moins dans le cadre aléatoire uniforme.

Remerciements. L'auteur désire remercier l'arbitre pour ses commentaires judicieux qui ont permis de corriger certaines inexactitudes et de clarifier certains points. L'auteur désire aussi remercier Jan Gecsei pour son aide dans les travaux de recherche qui ont mené

à cet article.

Bibliographie

Drabowski, M., *A General Model of Distribution of Objects in Information Retrieval Systems*, Information Systems **1** (1975), 147-151.

Godin, R., Saunders, E. et Gecsei, J., *Lattice Model of Browsable Data Spaces*, Information Sciences **40** (1986), 89-116.

Godin, R. et Gecsei, J., *Navigation dans les bases de données hétérogènes à l'aide de treillis*, Actes Information Communication '86, AFCET-CESTA, Paris (juin 1986), 313-323.

Godin, R., "L'utilisation de treillis pour l'accès aux systèmes d'information," Thèse de doctorat, Université de Montréal, 1986.

Riordan, J., "Combinatorial Identities," Wiley, New York, 1968.

Salton, G., "Automatic Information Organisation and Retrieval," McGraw-Hill, New York, 1968.

Sparck Jones, K., "Automatic Keyword Classification for Information Retrieval," Archon Books, Connecticut, 1971.

Tague, J., Nelson, M. et Wu, H., *Problems in the Simulation of Bibliographical Retrieval Systems*, in "Oddy et al. (Ed.): Information Retrieval Research, London: Butterworths," 1981, pp. 237-324.

Département de mathématiques et d'informatique
Université du Québec à Montréal
C.P. 8888, Succ. A
Montréal, Québec, H3C 3P8