ACCES DIRECT A DES FICHIERS MULTICLES

Gérard LEVY*, Witold LITWIN*, Samba NDIAYE**, Mouhamed T. SECK***

ABSTRACT

The access performance to records in secondary memory depends on their organization. The access time is smaller if records are close. More, the retrieval is faster if the records are ordered sequentially. We generalize for p-algebras the results obtained by Faloutsos in the case of binary algebra.

RESUME

La manière dont les enregistrements d'un fichier sont rangés en mémoire a évidemment une influence sur les temps de recherche : plus les enregistrements qui répondent à une question sont voisins, plus vite ils sont retrouvés. La recherche est encore accélérée s'ils sont situés séquentiellement, car on peut y accéder sans effectuer de saut. Notre intention est de généraliser à des algèbres à p éléments les résultats obtenus par Faloutsos pour l'algèbre binaire, et d'étudier le gain qu'on obtient à l'occasion de ces généralisations.

Mots-clés : fichiers multi-clés. méthodes d'accès. produits cartésiens. ordres totaux sur des produits cartésiens.

1. INTRODUCTION

Dans les méthodes d'accès aux fichiers multiclés, chaque enregistrement est rangé en mémoire en fonction des valeurs de ces différentes clés, et sa recherche est également fondée sur ces valeurs. L'article [1] de Faloutsos fournit une présentation très claire de ce problème qu'il étudie dans le cas d'une algèbre à deux éléments. Nous renvoyons à ce travail pour les aspects "sémantiques" du sujet, et allons nous attacher aux aspects mathématiques sous-jacents en vue de les résoudre dans un cadre assez général. On peut en effet ramener le problème à la forme suivante : il y a n clés ou attributs $x_1,...,x_n$, chaque x_i prenant ses valeurs dans un ensemble E_i qui a au moins deux éléments. A chaque article est associé la suite de ses valeurs $(x_1,...,x_n)$ qui permet de le ranger en mémoire M. Il s'agit ensuite de retrouver les articles dont toutes ou partie des valeurs des attributs sont précisées. Cette recherche aboutira d'autant plus vite que les différents articles qui répondent à une question seront bien regroupés. Plus précisément, après avoir défini un ordre total sur $E_1 \times ... \times E_n$ et avoir rangé tous ses éléments selon cet ordre en mémoire, on dira que des éléments de ce produit sont bien regroupés, ou qu'ils constituent un *bloc*, s'ils sont consécutifs, car ayant trouvé le premier les autres s'obtiennent en séquence. Ainsi, la réponse à une question s'obtiendra d'autant plus vite que l'ensemble des éléments du produit qui répondent à cette question sera partitionné en un plus petit nombre de blocs. Les performances dépendant bien sûr de l'ordre total choisi, nous allons nous intéresser à deux relations d'ordre : l'*ordre*

* Gérard LEVY, Witold LITWIN, Université Paris 9 Dauphine (France)
** Samba NDIAYE, Université Cheikh Anta Diop, Faculté des Sciences, Dpt Math-Informatique (Dakar-Sénégal)
*** Mouhamed Tidiane SECK, Université Cheikh Anta Diop, Ecole Supérieure Polytechnique, Dpt Génie Informatique (Dakar-Sénégal)

lexicographique, parce qu'il est le plus usité, et l'*ordre de Gray* où deux éléments consécutifs ne diffèrent que d'une coordonnée et d'une seule unité. Le plan de l'article est le suivant :

2. NOTATIONS, DEFINITIONS

2.1 Notations de base

Tous les ensembles dont il sera question ici sont totalement ordonnés et finis. Chacun de ces ensembles E a sa relation d'ordre total propre notée $\leq_E$, un plus petit élément noté $t(E)=t_E$, un plus grand élément noté $q(E)=q_E$, une fonction "suivant" s_E: $E\setminus\{q(E)\} \rightarrow E$ qui fait correspondre à chaque x de E autre que q(E), l'élément $s_E(x)$ qui lui est immédiatement supérieur, enfin une fonction de rang r_E : $E \rightarrow N$ telle que $r_E(t(E))=0$, $r_E(s_E(x))=r_E(x)+1$, et $r_E(q(E))=card(E)-1=|E|-1$.
L'ensemble E muni de cet ordre peut être considéré comme une suite finie. E muni de l'ordre inverse est noté $\underline{E}$, avec $t(\underline{E})=q(E)$, $q(\underline{E})=t(E)$, $s_{\underline{E}}(x)=s_E^{-1}(x)$ si x = t(E), et x inférieur à y dans $\underline{E}$ ssi y est inférieur à x dans E. Pour tout entier naturel a, on conviendra que $E^{|a|}= E$ si a est pair, et $E^{|a|}= \underline{E}$ si a est impair.
Pour éviter d'alourdir les notations, on supprimera l'indice E s'il n'y a pas de risque d'ambiguïté, et on utilisera les mêmes signes ≤, s, r, t, q pour des ensembles différents.
E étant totalement ordonné, on appelle *bloc* de E, tout ensemble $B = \{x_0,...,x_k\}$ d'éléments de E tel que $x_i = s(x_{i-1})$, pour i = 1,...,k. On a donc $t(B)=x_0$, $q(B)=x_k$, et B est un intervalle de E.

2.2. Ordre lexicographique

Soient F, G deux ensembles totalement ordonnés finis, ayant chacun au moins deux éléments. On définit un ordre total sur F×G comme suit:
t(FxG)=(t(F),t(G)), q(FxG)=(q(F),q(G)), et s(y,z)=(y,s(z)) si z≠q(G), s(y,z)=(s(y),t(G)) si z=q(G) et y≠q(F).
On pose (y,z)≤ (y',z') si il existe un entier naturel k tel que $(y',z')=s^k(x, y)$.
On voit qu'il s'agit bien d'un ordre total sur FxG, et que (y,z)≤(y',z') ssi (y<y') ou (y=y' et z≤z'). C'est cet ordre que nous appellerons lexicographique. Il peut se généraliser au produit d'un nombre fini d'ensembles finis totalement ordonnés. Ainsi ExFxG peut être muni de cet ordre si on le considère comme E×(F×G). On a alors t(E×F×G)=(t(E),t(F×(G))=(t(E),t(F),t(G)); de même q(ExFxG)=(q(E),q(F),q(G)), et s(x,y,z)=(x,s(y,z))=(x,y,s(z))
si z≠q(G), =(x,s(y),t(G)) si z=q(G), et y≠q(F), =(s(x),t(F),t(G)) si x≠q(E) et y = q(F)
et z=q(G). De plus (x,y,z) ≤(x', y', z') ssi (x<x' ou x=x' et (y,z)≤(y',z')).
On remarque que l'ordre lexicographique sur F×G revient à écrire

$$FxG = \sum_{y=t(F)}^{q(F)} \{y\} \times G \text{ soit } F\times G = \{t(F)\}\times G + \{s(t(F))\}\times G +...+ \{q(F)\}\times G.$$

ou le signe "+" exprime la concaténation des ensembles totalement ordonnés {y}×G considérés comme des suites.

2.3. Ordre de Gray

Avec les mêmes hypothèses que ci-dessus, on prend maintenant $t(F\times G)=(t(F),t(G))$;
$s(y,z)=(y,s(z))$ si $r(y)$ est pair et $z<q(G)$
$=(s(y),q(G))$ si $r(y)$ est pair et $y<q(F)$ et $z=q(G)$
$=(y,s^{-1}(z))$ si $r(y)$ est impair et $t(G)<z$
$=(s(y),t(G))$ si $r(y)$ est impair et $y<q(F)$ et $z=t(G)$
il s'ensuit que $q(F\times G)=(q(F),t(G))$ si $|F|$ est pair,
$=(q(F),q(G))$ sinon.

On s'assurera qu'on définit bien ainsi sur $F\times G$ un ordre total qu'on appellera ordre de Gray car il généralise celui que Gray avait défini pour $\{0,1\}^n$. Cet ordre peut se définir également pour le produit d'un nombre fini d'ensembles totalement ordonnés finis, comme précédemment, en traitant $E\times F\times G$ comme $E\times(F\times G)$ et ainsi de suite.

On constatera que l'ordre de Gray sur $F\times G$ se ramène à écrire que $FxG = \sum_{y=t(F)}^{q(F)} \{y\} \times G^{|r(y)|}$

soit $F\times G=\{t(F)\}\times G + \{s(t(F))\}\times\underline{G} + \{s^2(t(F))\}\times G + \{s^3(t(F))\}\times\underline{G} +...$,
où le signe "+" exprime encore la concaténation des ensembles ordonnés considérés comme des suites.
A titre d'exemple prenons $F=(0,1,2,3)$ et $G=(0,1,2)$ munis de l'ordre des entiers naturels.
L'ordre lexicographique sur $F\times G$ donne :
$F\times G=\{0\}\times\{0,1,2\} + \{1\}\times\{0,1,2\} + \{2\}\times\{0,1,2\} + \{3\}\times\{0,1,2\}$
$=\{(0,0),(0,1),(0,2)\}+\{(1,0),(1,1),(1,2)\}+\{(2,0),(2,1),(2,2)\}+\{(3,0),(3,1),(3,2)\}$
$=\{(0,0),(0,1),(0,2),(1,0),(1,1),(1,2),(2,0),(2,1),(2,2),(3,0),(3,1),(3,2)\}$.
L'ordre de Gray sur $F\times G$ donne :
$F\times G=\{0\}\times\{0,1,2\} + \{1\}\times\{\underline{0,1,2}\} + \{2\}\times\{0,1,2\} + \{3\}\times\{\underline{0,1,2}\}$
$=\{0\}\times\{0,1,2\} + \{1\}\times\{2,1,0\} + \{2\}\times\{0,1,2\} + \{3\}\times\{2,1,0\}$
$=\{(0,0),(0,1),(0,2),(1,2),(1,1),(1,0),(2,0),(2,1),(2,2),(3,2),(3,1),(3,0)\}$
Si de plus on prend $E=\{0,1\}$, alors l'ordre de Gray sur $E\times F\times G$ est :
$E\times(F\times G)=\{0\}\times(F\times G) + \{1\}\times(\underline{F\times G})$, où $\underline{F\times G}$ est la liste des éléments de $F\times G$ munie de l'ordre de Gray inverse, soit $\{(3,0),(3,1)...(0,1),(0,0)\}$.

2.4 Produits généralisés

Soient $I=\{1,...,n\}$, $n>1$, et F_i, $i=1,...,n$, une famille finie d'ensembles finis totalement ordonnés.

2.4.2 Ordre de Gray sur les F_i

Ici encore on le définit par récurrence sur les F_i. Sur $F_n=E_n$ c'est l'ordre de E_n et donc $T_n=t_n$, $Q_n=q_n$; puis, pour i décroissant de n-1 à 1, on a $F_i=E_i\times F_{i+1}$, $T_i=(t_i,T_{i+1})$, et la fonction suivante de F_i est telle que pour tout (x_i,y) de $E_i\times F_{i+1}$, $s(x_i,y)$ vaut $(x_i,s(y))$ si $r(x_i)$ est pair et $y<Q_{i+1}$, $(s(x_i),Q_{i+1})$ si $r(x_i)$ est pair et $y=Q_{i+1}$, $(x_i,s^{-1}(y))$ si $r(x_i)$ est impair et $T_{i+1}<y$, $(s(x_i),T_{i+1})$ si $r(x_i)$ est impair et $y=T_{i+1}$, à condition que $x_i<t_i$.

Il en résulte que $Q_i=(q_i,T_{i+1})$ si $|E_i|$ est pair, et que $Q_i=(q_i,Q_{i+1})$ si $|E_i|$ est impair.

Enfin la suite des éléments de F_i peut s'écrire symboliquement $F_i=\sum_{x_i=t_i}^{q_i}\{x_i\}\times F_{i+1}^{|r(x_i)|}$

2.5 Projections, extensions

Toute partie non vide J = {i,j.....,k} de l'ensemble de I des indices, est supposée telle que i<j<...<k. On dit que J est un intervalle de I si ses éléments sont des entiers consécutifs.

On note $E_J=E_i\times E_j\times\ldots\times E_k$, on a donc $E_I=E_1\times E_2\times\ldots\times E_n$, et pour tout $x=(x_1,x_2,\ldots x_n)$ de E_J, on note $x_J=(x_i,x_j,\ldots,x_k)$ la *projection* de $x=x_I$ sur E_J.

Inversement pour tout élèment fixé a_J de E_J, on appelle *extension* de a_J et on note $Ext(a_J)$, l'ensemble des élèments x de E_I qui ont a_J pour projection sur E_J. On a donc :

$Ext(a_J)=\{x\in E$ tel que $x_J=a_J\}$.

A titre d'exemple, pour n=7 et J={2,6,7}, on a $a_J=\{a_2,a_6,a_7\}$ et $Ext(a_J)$ est l'ensemble des élèments $x=(x_1,x_2,\ldots,x_7)$ de $E_1\times E_2\times\ldots\times E_7$ tels que $x_2=a_2$ et $x_6=a_6$ et $x_7=a_7$. Autrement dit, $Ext(a_J)$ est la réponse à la question ou requête partiellement définie : quel est l'ensemble des x tels que $x_J=a_J$?

Si on désigne par K le complément de J par rapport à I, et si on fait abstraction de tout ordre sur E_I, on peut écrire que $Ext(a_J)=\{a_J\}\times E_K$. (1)

Par contre si on veut tenir compte de l'ordre défini sur E , il est nécessaire d'avoir une approche plus fine. Il existe en effet deux suites de parties de I, $(K_0,K_1,\ldots K_m)$ et $(J_1,\ldots,J_m)$, deux à deux disjointes, non vides, à l'exception éventuellement de K_0 et/ou K_m, telles que J soit égale à la réunion des J_i et K à celle des K_i, et que K_0, J_1, $K_1,\ldots K_{m-1}$, J_m,K_m soient des intervalles consécutifs de I. Il s'ensuit que tout élément de E_I peut s'écrire d'une façon unique sous la forme $x=(x_{K0},x_{J1},x_{K1},\ldots,x_{Km-1},x_{Jm},x_{Km})$ et donc que

$Ext(a_J)=E_{K0}\times\{a_{J1}\}\times E_{K1}\times\ldots\times E_{Km-1}\times\{a_{Jm}\}\times E_{km}$. (2)

Suivant l'ordre qui est défini sur E_I, $Ext(a_J)$ qui est un sous-ensemble de E_I, va être partitionné en un ensemble minimal de blocs. C'est ce problème que nous nous proposons d'étudier pour l'ordre lexicographique et pour l'ordre de Gray.

3. ETUDE DU NOMBRE DE BLOCS EN ORDRE LEXICOGRAPHIQUE

Tous les résultats relatifs à l'ordre lexicographique sont fondés sur le résultat suivant.

Lemme 1 : en ordre lexicographique, quels que soient les ensembles E, F, G, totalement ordonnés, finis, de cardinalité ≥ 2. On a:

a) G est un bloc de G ;

b) pour tout élément b de F et tout bloc B de G, $\{b\}\times B$ est un bloc de $F\times G$;

c) $E\times(\{b\}\times G)$ peut être partitionné en $|E|$ blocs.

Preuve :

a) est triviale ;

b) soit $B=(t_B,\ldots,z,s(z),\ldots,q_B)$ le dit bloc.

On a $\{b\} \times B=((b,t_B),\ldots,(b,s),(b,s(z)),\ldots,(b,q_B)$; pour tout z de B autre que q_B, le suivant de (b,z) dans $\{b\} \times B$ aussi bien que dans $F \times G$ est (b,s(z)). Donc $\{b\} \times G$ est bien un bloc de $F \times G$.

c) Compte tenu de a) et b), pour chaque x de E, $\{x\} \times (\{b\} \times G)$ est un bloc de $E \times (F \times G)$. De

4. ETUDE DU NOMBRE DE BLOCS EN ORDRE DE GRAY

Avec les mêmes conventions que précédemment, appelons C(n,p,k) le nombre total de blocs relatifs, en ordre de Gray, aux extensions de tous les a_J, tels que $|J|=k$, et $a_J \in L^k$. Nous allons commencer par établir une relation de récurrence entre ces coefficients, puis nous l'utiliserons pour les calculer explicitement.

Lemme 2 :
Les coefficients C(n,p,k) sont dans la relation
$C(n,p,k) = p(C(n-1,p,k) + C(n-1,p,k-1)) - (p-1)C^k_{n-1}$, pour $p \geq 2$ et $0<k<n$.

Preuve :
Posons $Q'_{n,k}=\{J \in Q_{n,k} \mid j_1=1\}$ et $Q''_{n,k}=\{J \in Q_{n,k} \mid j_1>1\}$
Tout J de $Q'_{n,k}$ est de la forme $J=\{1\}\cup J'$, avec $J'=\{j_2,...j_k\} \in Q_{n-1,k-1}$, et $a_J = (a_1,a_{J'})$ avec $a_1 \in L$ et $a_{J'} \in E_{J'}=L_{J'}$.
A tout bloc B de $Ext(a_J)$ correspond un bloc B' de $Ext(a_{J'})$ tel que $B=\{a_1\} \times B'_\varepsilon$, cette dernière suite valant B' ou $\underline{B'}$ suivant que a_1 est pair ou impair. Par conséquent,

$$\sum_{J \in Q'_{n,k}} \sum_{a_J \in L_J} nb(a_J) = \sum_{J' \in Q_{n-k-1}} p \sum_{a_{J'} \in L_{J'}} nb(a_{J'}) = pC(n-1,p,k-1)$$

Par contre si $J \in Q''_{n,k}$, certains blocs peuvent se regrouper par contiguïté. En effet ici $J \subset \{2,...,n\}$ et si a_J est telle que $T_{n-1} \in Ext(a_J)$ ou $Q_{n-1} \in Ext(a_J)$ dans $L^{n-1}_n = L_{\{2,...,n\}}$, alors, si l'élément a_1 de L est pair, (a_1,Q_{n-1}) et $(s(a_1),Q_{n-1})$ sont voisins dans $L_1 = L$, et si a_1 est impair, ce sont (a_1,T_{n-1}) et $(s(a_1),T_{n-1})$ qui le sont.
On est donc ramené au problème : quels sont les J et a_J tels que $J \in \{2,...,n\}$, $|J|=k$, et que
a) $T_{n-1} \in Ext(a_J)$ dans $L_{\{2,...,n\}} = L^{n-1}$?
b) $Q_{n-1} \in Ext(a_J)$ dans le même ensemble ?
Or, il est aisé de voir que $T_{n-1} = 0^{n-1}$, et que $Q_{n-1} = (p-1).0^{n-2}$ ou $(p-1)^{n-1}$ suivant que p est pair ou non.

a) Pour que $T_{n-1} \in Ext(a_J)$, il est nécessaire que les k chiffres de a_J soient égaux à 0. Comme J est un sous-ensemble de $\{2,...,n\}$, cela peut se faire de C^k_{n-1} manières différentes.
b) Si p est pair, on a $Q_{n-1} = (p-1).0^{n-2}$. Si $J = \{j_1 j_2,...j_k\}$, et $j_1=2$, pour que $Q_{n-1} \in Ext(a_J)$, il faut que $a_2 = p-1$, et que les k-1 autres chiffres de a_J soient égaux à 0. Cela peut se faire de C^{k-1}_{n-2} façons différentes. Si $j_1>2$, il faut que les k chiffres de a_J soient égaux à 0 : ce qui peut se faire de C^k_{n-2} façons différentes. Donc le nombre total de a_J tels que $|J| = k$ et que
$Q_{n-1} \in Ext(a_J)$ est $C^{k-1}_{n-2} + C^k_{n-2} = C^k_{n-1}$

Si p est impair, on a $Q_{n-1} = (p-1)^{n-1}$; pour que $Q_{n-1} \in Ext(a_J)$, il est nécessaire que les k chiffres de a_J soient égaux à p-1, ce qui peut se faire de C^k_{n-1} manières différentes.

Appelons $U = \{a_J / J \in \{2,.....,n\} et \quad |J| = k et \quad a_J \in L_J et \quad T_{n-1} \in Ext(a_J)\}$
et $V = \{a_J / J \in \{2,.....,n\} et \quad |J| = k et \quad a_J \in L_J et \quad Q_{n-1} \in Ext(a_J)\}$

Quelle que soit la parité de p, ces ensembles sont donc de même cardinalité $|U| = |V| = C_{n-1}^{k}$

Pour tout $J \in Q''_{n,k}$ et tout $a_j \in L_p$ tout bloc B' de l'extension de a_I dans $L\{2,...,n\} = L^{n-1}$, engendre les p blocs $\{a_1\}\times B'_{\varepsilon}$, $a_1 \in L$, de l'extension de a_J dans $L_I = L^n$. Au total, cela fait pC(n-1,p,k) blocs de L^n. Si a_j n'appartient ni à U ni à V, les p blocs créés sont distincts car sans contiguité. En revanche, pour tout a_J de U, il existe un bloc B et un seul de l'extension de a_J dans L^{n-1} qui contient $t = T_{n-1}$. Ce bloc B, de plus grand élément q, donne naissance aux p blocs $\{0\}\times B=\{(0,t),...,(0,q)\}_n$ $(1)\times\underline{B}=\{(1,q),...,(1,t)=$, $(2)\times B=\{(2,t),...,(2,q)\}$ $,...,(p-1)\times B^{|p-1|}$ de l'extension de a_J dans L^n. Or $s(2i+1,t) = (2i+2,t)$ dans L^n, car $t = T_{n-1}$, et par conséquent les blocs $\{2i+1\}\times\underline{B}$ et $\{2i+2\}\times B$ qui sont contigus doivent être regroupés, pour $i\geq 0$ et $2i+2\leq p-1$. Pour chaque a_J de U, le nombre de ces regroupements est égal à $|p/2|-1$ ou à $|p/2|$ suivant que p est pair ou impair. Le même raisonnement montre que pour chaque a_J de V, il y a, suivant que p est pair ou impair, $|p/2|$ ou $|p/2|-1$ regroupements de blocs contigus $\{2i\}\times B$ et $\{2i+1\}\times\underline{B}$. Donc, si p est pair, il faut retrancher aux pC(n-1,p,k) blocs un nombre de blocs égal à $\left(|p/2|-1\right)\cdot|U|+\left(|p/2|\right)\cdot|V| = (p-1)\cdot C_{n-1}^{k}$

Un raisonnement similaire conduit au même résultat pour p impair. Ce qui prouve le lemme.

Théorème 2 :

En ordre de Gray, le nombre total de blocs relatifs aux extensions de tous les a_J tels que $|J|=k$ est $C(n,p,k) = C_n^k + \left(p^n - 1\right)\cdot C_{n-1}^{k-1}$

Le nombre moyen de blocks relatifs à l'extention d'un a_J telque $|J|=k$, est $\underline{C}(n,p,k) = C(n,p,k) / (p^k C_n^k)$

Preuve :

Elle se fait par récurrence sur n, p étant fixé, en utilisant la relation établie dans le lemme. En effet, la propriété est vraie pour n = 1. Supposons qu'elle le soit jusqu'à n-1, et pour tout k compris entre 0 et n-1. On a donc :

$C(n-1,p,k) = C_{n-1}^k + (p^{n-1}-1).C_{n-2}^{k-1}$ $\qquad$ $C(n-1,p,k-1) = C_{n-1}^{k-1} + (p^{n-1}-1).C_{n-1}^{k-2}$

d'où, du fait de la relation de récurrence,

$$\begin{aligned} C(n,p,k) &= p.(C_{n-1}^k + C_{n-1}^{k-1} + (p^{n-1}-1).(C_{n-2}^{k-1} + C_{n-2}^{k-2})) - (p-1).C_{n-1}^k \\ &= p.(C_n^k + (p^{n-1}-1).C_{n-1}^{k-1}) - (p-1).C_{n-1}^k \\ &= C_n^k + (p^n-1).C_{n-1}^{k-1} + (p-1).(C_n^k - C_{n-1}^{k-1} - C_{n-1}^k) \\ &= C_n^k + (p^n-1).C_{n-1}^{k-1} \end{aligned}$$

5. ECONOMIE DUE A L'UTILISATION DES CODES DE GRAY

5.1 Economie absolue

Soit $\Delta = D(n,p,k) - C(n,p,k)$, l'économie globale qui résulte de l'utilisation des codes de Gray au lieu des codes lexicographiques. On voit que :

$$D = p^n C_{n-1}^{k-1} + p^{n-1}.C_{n-2}^{k-1} + \sum_{i=k}^{i=n-2} C_{i-1}^{k-1} p^i \quad et \quad que\, \Delta = (p^{n-1} - (n-1)/k).C_{n-2}^{k-1} + \sum \ldots$$

ce qui prouve que $\Delta \geq 0$, puisque $p \geq 2$.

5.2 Economie relative

Soit $\underline{\Delta}(n,p,k) = (\underline{D}(n,p,k) - \underline{C}(n,p,k)) / \underline{D}(n,p,k)$, l'économie relative en nombre moyen de blocs qui découle de l'utilisation des codes de Gray au lieu des codes lexicographiques. Intéressons-nous à son comportement pour les grandes valeurs de n. On a :

$$C(n,p,k) = p^n C_{n-1}^{k-1}\left(1+\frac{n-k}{kp^n}\right) \approx p^n C_{n-1}^{k-1} \qquad d'où$$

$$\begin{aligned}\Delta &\approx p^{n-1}C_{n-2}^{k-1} + p^{n-2}C_{n-3}^{k-1}+\ldots \\ &\approx p^{n-1}C_{n-2}^{k-1}\left(1+\frac{1}{p}\frac{n-k-1}{n-2}+\frac{1}{p^2}\frac{(n-k-1)(n-k-2)}{(n-2)(n-3)}+\ldots\right)\end{aligned}$$

de plus on a

$$D \approx p^{n-1}C_{n-2}^{k-1}\left(1+\frac{1}{p}\frac{n-k-1}{n-2}+\frac{1}{p^2}\frac{(n-k-1)(n-k-2)}{(n-2)(n-3)}+\ldots\right).$$

Les deux développements entre parenthèses, ci-dessus, convergent d'autant plus vite que p est plus grand. On a donc, quand n tend vers l'infini,

$$\underline{\Delta} \approx \frac{1}{p}\frac{n-k}{n-1}\left(1+\frac{1}{p}\left(\frac{n-k-1}{n-2}-\frac{(n-k)}{(n-1)}+\ldots\right).= \frac{1}{p}\frac{n-k}{n-1}\left(1-\frac{k-1}{p(n-1)(n-2)}+\ldots\right)\right.$$

$$\underline{\Delta} \approx \frac{1}{p}\frac{n-k}{n-1}. \qquad D'où \quad le$$

Théorème 3 :

Asymptotiquement en n, le gain relatif moyen de nombres de blocs qu'on réalise en recourant aux codes de Gray plutôt qu'aux codes lexicographiques, pour des requêtes dont k des n valeurs sont fixées est $\underline{\Delta} \approx \frac{1}{p}\frac{n-k}{n-1}$.

Commentaires :

a) Les résultats que nous avons obtenus généralisent ceux de Faloutsos ; a posteriori on se rend compte que pour passer des résultats relatifs à p = 2 aux résultats relatifs à un entier $p \geq 2$, il suffit de remplacer 2 par p dans les formules de Faloutsos.

b) Pour p fixé, on constate que l'économie est d'autant plus faible que les questions sont bien précisées, c'est-à-dire que k est grand.

c) Pour p fixé, lorsque n tend vers l'infini, il semble plus judicieux de voir ce qui se passe quand k tend vers l'infini plutôt que quand k est fixé. Si n et k tendent simultanément vers l'infini en restant dans le rapport $k/n = \tau$, on voit que $\underline{\Delta} \approx \frac{1}{p}(1 - \tau)$, où τ est le pourcentage de valeurs fixées dans une question.

d) Au vu du Théorème 3, quand n et k sont fixés, le gain relatif moyen $\underline{\Delta}$ est fonction décroissante de p. On peut se demander si ce résultat négatif n'est pas dû au fait qu'on ne tient pas compte de la masse m d'informations à traiter. Si on se place en base 2, pour exprimer toute

cette masse, il faut choisir n tel que $m = 2^n$, et en base p choisir n' tel que $m = p^{n'}$. De même une question, dont k valeurs des n valeurs sont fixées en base 2, recouvre un ensemble de 2^{n-k} informations, tandis qu'une question dont k' des n' valeurs sont fixées en base p en recouvre $p^{n'-k'}$. Par conséquent, pour conserver les masses d'informations traitées quand on passe de 2 à p, il faut remplacer n et k par n' et k' tels que $p^{n'} = 2^n$ et $p^{n'-k'} = 2^{n-k}$.
Soient $\underline{\Delta} = \underline{\Delta}(2,n,k)$ et $\underline{\Delta}' = \underline{\Delta}(p,n',k')$. On a :

$$\underline{\Delta} \approx \frac{n-k}{2(n-1)}.$$

$$\underline{\Delta}' \approx \frac{n'-k'}{p(n'-1)} = \frac{2}{p} \times \underline{\Delta} \times \frac{1-\frac{1}{n}}{1-\frac{Ln(p)}{nLn(2)}} \approx \frac{2}{p}\underline{\Delta}$$

On voit donc que le fait de recourir à la conservation des masses d'informations ne change rien au rapport de $\underline{\Delta}$ à $\underline{\Delta}'$. Le résultat est le même : $\underline{\Delta}(n,p,k)$ est fonction décroissante de p. En conséquence, pour ce problème tout au moins, il est inutile de se placer dans une algèbre multivaluée, et le mieux est de rester en algèbre binaire.

6. CONCLUSION

Tous les résultats de Faloutsos peuvent se généraliser à des algèbres multivaluées. Il suffit d'y remplacer 2 par p. Cependant, cette généralisation est infructueuse dans la mesure où le gain relatif résultant de l'emploi des codes de Gray à la place des codes lexicographiques décroît quand p augmente.

REFERENCES

[1] Christos FALOUTSOS : Multiattribute Hashing Using Gray Codes,