

一連検索性をもつファイル構成について

近畿大・理工 田澤新成 (Shinsei Tazawa)

1. はじめに

情報収納検索システムにおいて質問に該当するレコードの検索に要する時間とレコードを収納するのに要する記憶容量との間には一般にトレードオフが存在する。しかし Ghosh [2] は検索時間を短縮しかつ要する記憶容量の減少を実現するファイル方式を考案した。これはレコードを適当な順序に配列し各質問に該当するレコードを一連に検索することる可能にするような方式にある。このようなレコードを一連に検索することる可能性を一連検索可能性 (CRP) といい、このときレコードの配列を CR-配列とよぶ。

2 値ベクトルの全体 $V_n = \{ (v_1, v_2, \dots, v_n) \mid v_i = 1, 0; i=1, \dots, n \}$ を考へる。 n は 2 値項目数, V_n のベクトルは 2 値レコードに対応する。 S_i は i 番目について値 1 をもつ, すなわち, $v_i = 1$ なる 2^{n-1} 個のベクトルの列を表わし, $\bar{S}_i$ は $v_i = 0$ なるベクトルの列を表わす。 S_i ($\bar{S}_i$) は i 番目の項目に該当

する (該当しな...) レコードの検索を要求する質問に対応する二つの二重 S_i および $\bar{S}_i$ は度数 1 の質問とよばれる。Which と Lipacki [1] は質問集合 $\{S_1, \dots, S_n\}$ に対して CRP をもつ V_n のベクトル (レコード) の CR-配列を構成するアルゴリズムを与えた。その長さ l_n は

$$(1) \quad l_n = \left(\frac{2}{3}n + \frac{2}{9}\right) 2^{n-1} - \frac{1}{9}(-1)^n$$

である。Luccio と Preparata [2] は l_n が質問集合 $\{S_1, \dots, S_n\}$ に対する CR-配列にあって最小であることを示した。これは、度数 1 の質問集合 $\{S_1, \dots, S_n, \bar{S}_1, \dots, \bar{S}_n\}$ に対して CRP をもつ V_n のベクトルの配列の長さ L_n にあって

$$L_n \geq \left(\frac{2}{3}n - \frac{1}{36}\right) 2^n - \frac{2}{9}(-1)^n$$

が成り立ち、かつこの等号を満たすような CR-配列を構成するこゝろができることを示す。

2. CR-配列の長さの下限

質問集合 $\{S_1, \dots, S_n, \bar{S}_1, \dots, \bar{S}_n\}$ に対して V_n のベクトルの CR-配列にあって長さが最小のもの A_n を記述する。 $T_i = S_i$, $T_{i+n} = \bar{S}_i$ ($1 \leq i \leq n$) とおき、 A_n において T_i の最初の位置を $A_n[T_i]$ とおく。こゝで $A_n[T_{i_1}] < A_n[T_{i_2}] < \dots < A_n[T_{i_{2n}}]$ と仮定する。このとき A_n は次の性質をもつ。

性質 1 T_i に属するベクトルの位置してゐる連続した位置

の集合 $P[T_i]$ と書くと、 $|j-k| \geq 2$ ならば $j, k = 1, \dots, 2n$

$$(2) \quad P[T_j] \cap P[T_k] = \emptyset$$

が成り立つ。

これは次のようにしてわかる。 $Q_k \in T_k$ ($k=1, \dots, 2n$) に属するベクトルの集合とする。 T_j, T_{j+1}, T_{j+2} を考え、 $P[T_j] \cap P[T_{j+2}] \neq \emptyset$ とする。このとき $P[T_{j+1}] \subseteq P[T_j] \cup P[T_{j+2}]$ 。この結果、 $Q_{j+1} \in Q_j \cup Q_{j+2}$ である。このことから、ある k に対して $T_j = S_k, T_{j+2} = \bar{S}_k$ あるいは $T_j = \bar{S}_k, T_{j+2} = S_k$ である。従って $Q_j \cap Q_{j+2} = \emptyset$ であるから $P[T_j] \cap P[T_{j+2}] = \emptyset$ である。したがって (2) が成り立つ。

性質 2 A_n の中で最初の n 個の連続の空間 $T_1, T_2, \dots, T_n$ を部分列 $A_n^{(1)}$ とし、残りの n 個の連続の空間 $T_{n+1}, T_{n+2}, \dots, T_{2n}$ を部分列 $A_n^{(2)}$ とする。このとき、 $A_n^{(k)}$ の長さは l_n 以上である ($k=1, 2$)。

$A_n^{(k)}$ が $\bar{S}_{j_1}, \bar{S}_{j_2}, \dots, \bar{S}_{j_k}$ を含むとする。 $A_n^{(k)}$ によって、各 $k=1, 2, \dots, k$ に対して $v_k=1$ のときは 0 に、 $v_k=0$ のときは 1 に置きかえてえらる新しい列を B とかく。このとき B は $\{s_1, \dots, s_n\}$ に対する CR-配列であり、その長さは $A_n^{(k)}$ のそれと同じである。 B の長さは $|B|$ から l_n 以上であるから、求める結果を得る。

定理 A_n の長さ L_n は $L_n \geq 2^{n-2}$ である。このとき

$$(3) \quad L_n \geq \left(\frac{2}{3}n - \frac{1}{3}\right) 2^{n-2} - \frac{2}{9}(-1)^n$$

証明 異なった項目を指定した空間に該当するレコード (ベクトル) の個数は 2^{n-2} であるから, $|P[T_{i_n}] \cap P[T_{i_{n+1}}]| \leq 2^{n-2}$ である。それ故, 性質 1, 2 から

$$L_n \geq 2L_{n-1} - |P[T_{i_n}] \cap P[T_{i_{n+1}}]| \geq 2L_{n-1} - 2^{n-2}$$

が成り立つ。従って, (1) を適用して (3) をうる。

3. 最小な長さをもつ CR-配列の構成

この節で, (3) の等式に到達する CR-配列 A_n を構成する手続きを述べる。この構成方法について $n=3$ の場合の構成例を図 1 に与える。与えられた空間集合に対する CR-配列 B において, B の j 番目のベクトルを $B[j]$ と書き, B におけるセグメント T の最初の位置を $B[T]$ とかく。

ステップ 1. 空間集合 $\{S_1, S_2, \dots, S_n\}$ に対して, $A_n^{(1)}[S_{i+1}] < A_n^{(1)}[S_i]$, $i=1, 2, \dots, n-1$ および $|P_1[S_1] - P_1[S_2]| = 2^{n-2}$ を満たす CR-配列 $A_n^{(1)}$ を Ehrlich-Lipshitz アルゴリズム [1] を用いて構成する。ここで $P_1[T]$ はセグメント T が位置している ($A_n^{(1)}$ において) 連続した位置の集合である。(図 1 では $P_1[S_1] - P_1[S_2]$ の代りに $S_1 - S_2$ を記してある)

ステップ 2. $A_n^{(1)}$ の各ベクトルについて $1 \leq 0 \leq 1$, $0 \leq 1$

に置きかえ, つぎに才1行と才2行を交換する。この新しい列を A^* とかく。ル - ル $A_n^{(2)}[j] \leftarrow A^*[l_n - j + 1]$ によって $A_n^{(2)}$ を形作る。

注釈 $A_n^{(d)}$ の長さは $d=1, 2, \dots, l_n$ である ($[1], [3]$ を見よ)。 $A_n^{(2)}$ は空間集合 $\{\bar{s}_1, \bar{s}_2, \dots, \bar{s}_n\}$ に対する $\mathbb{C}R$ -配列である。さしして, $A_n^{(2)}[\bar{s}_2] < A_n^{(2)}[\bar{s}_1] < A_n^{(2)}[s_2] < \dots < A_n^{(2)}[\bar{s}_n]$ かつ $|P_2[\bar{s}_2] - P_2[\bar{s}_1]| = 2^{n-2}$ が成り立つ。 $\therefore P_2[T]$ は τ

$$\begin{array}{c}
 \begin{array}{cc} s_2 & s_1 - s_2 \end{array} \\
 A_3^{(1)} = \begin{array}{cccccccc} 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \end{array} \\
 \begin{array}{cc} s_3 & s_1 \end{array}
 \end{array}
 \qquad
 \begin{array}{c}
 A^* = \begin{array}{cccccccc} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{array}
 \end{array}$$

$$\begin{array}{c}
 \begin{array}{cc} \bar{s}_2 & \bar{s}_3 \end{array} \\
 A_3^{(2)} = \begin{array}{cccccccc} 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \end{array} \\
 \begin{array}{c} \bar{s}_2 - \bar{s}_1 \end{array} \left| \begin{array}{c} \bar{s}_1 \\ \hline \xleftarrow{A^{**}} \end{array} \right.
 \end{array}
 \qquad
 \begin{array}{c}
 \text{in reverse order of } A^*
 \end{array}$$

$$\begin{array}{c}
 \begin{array}{ccc} s_2 & \bar{s}_2 & \bar{s}_3 \end{array} \\
 A_3 = \begin{array}{cccccccccccc} 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \end{array} \\
 \begin{array}{cc} s_3 & s_1 \end{array} \left| \begin{array}{c} \bar{s}_1 \\ \hline \xleftarrow{A^{**}} \end{array} \right. \\
 \xleftarrow{A_3^{(1)}} \qquad \qquad \qquad \xleftarrow{A^{**}}
 \end{array}$$

図1 $\mathbb{C}R$ -配列 A_3 の構成

$\times$ と Γ が位置する位置の集合である。(図1では $P_2[\bar{S}_2] - P_2[\bar{S}_1]$ の代りに $\bar{S}_2 - \bar{S}_1$ と記してある。)

ステップ3. $P_2[\bar{S}_2] - P_2[\bar{S}_1]$ に属する位置にあるすべてのベクトルを $A_n^{(2)}$ から削除する。そしてこの新しい列を A^{**} と書く。

ステップ4. $A_n^{(1)}$ と A^{**} を連接して $A_n = A_n^{(1)} A^{**}$ と得る。

注釈 $A_n^{(1)}$ の長さは l_n , A^{**} の長さは $l_n - 2^{n-2}$ であるから A_n の長さは $2l_n - 2^{n-2}$, すなわち (3) の右辺に等しい長さである

参考文献

- [1] H.D.Ehrich and W.Lipski, Jr., On the storage space requirement of consecutive retrieval with redundancy, Inform. Proc. Lett. 4(1976)101-104.
- [2] S.P.Ghosh, File organization: The consecutive retrieval property, Comm. ACM 15(1972)802-808.
- [3] F.Luccio and F.P.Preparata, Storage for consecutive retrieval, Inform. Proc. Lett. 5(1976)68-71.