数理解析研究所講究録
第82巻 1970年 78-88

Combinatorial Information Retrievalについて

広大 教養　岡本雅典

§1 はしがき

近時情報処理能力の飛躍的増大に伴い，社会科学諸分野全般にまたがる基礎的データの処理および利用を，その情報特性をも考慮しつゝ一つの総合的観点から行なおうという動きがあるが（日本では文献[1]，米国でのCouncil of social science data archievesの活動については文献[2]），そこでは事実検索（Fact Retrieval = F.R.）のみならず文献検索（Document Retrieval = D.R.）が重要課題となる。（情報検索（I.R.）一般およびD.R.については文献[3], [4], [5]）。組合せ数学の立場から 1° 自動索引化に際しての分類問題 ── これは純粋に組合せ数学だけの問題ではないが，グラフ論におけるtree structureの構成と関係する。 2° ファイル構成に関してCombinatorial configurationの利用。 3° I.R.に際して質問から得られた文献 ── 索引間のIncidence

matrix が蓄積情報の Incidence matrix の一部分であるかどうかの判定，グラフ論的に云えば与えられた Directed graph が他の Directed graph の部分であるかどうかの判定検出の問題等が研究の対象としてあげられる。ここでは第二の問題点を Ray-chaudhuri (1968)〔文献[6]〕に従ってその内容を紹介すると共に PG(m, Δ) が minimum cycle θ(くび) をもつ場合についてのファイル構成の一例を示す。

§2 ファイル構成

ファイルを作る場合ファイルの対象となるべき個体あるいは文献があり，これらは正の整数値で番号付けられている。各個体には記号化された カテゴリー あるいは インデックス が付けられる。ファイル方式には Document Filing と Aspect (あるいは Inverted) Filing がある。以下計算機を利用した場合のファイル F，その蓄積方式 S，検索方式 R を定式化すると次のようになる。

<u>ファイル F</u>；a triplet F(I, A, f) であって，I = 検索の対象となるべき個体の集合（文献集合），A = インデックス集合 = {0，正整数}（文献より抽出された主題インデックス集合に対応），f = I から A の部分集合への many to many mapping（索引付け），$I \ni i$ に対して f(i) = 個体 i のも

フインデックス（一般に一つとは限らない）。

Fの蓄積方式S；a triplet S(I, M, Δ)であって，M=正整数の集合=Computer addressの集合で永久保存記録(ディスク，マイクロフィルム等)のaccession numberを含むもの。Δ=IからMの部分集合への many to many mapping (番地付け)，I∋iに対してΔ(i)=個体iの計算機上の番地群，a(i)=accession number=個体iの永久記録が蓄積されている記憶装置の番地(見出し)。mapping Δにより a(i)はΔ(i)内に必ず蓄積される。　番地指定がなされるために(定義)ある $M_1 \subset M$, $A_1 \subset A$に対して $f(i) \supset A_1 \rightleftarrows |M_1 \cap \Delta(i)| = 1$ (for $\forall i \in I$) ならばM_1はA_1に関してvalidであるという。　任意の有限集合Aについて|A|=Aの要素の数とする。

検索方式R；a triplet R($\mathcal{Q}$, M, γ)であって，$\mathcal{Q}$=Aの部分集合のある族，$\gamma = \gamma(A_1)$がA_1に関して validであるような$\mathcal{Q}$からMの部分集合への mapping．(定義) すべての$A_1 \in \mathcal{Q}$に対して$|A_1| \leqq t$ならばRは order tであるという。

order tのファイル構成とは order tのRとSを合せたものを云う。　検索手続は次のようになる。

質問Q ⟶ $A_1 \in \mathcal{Q}$ ⟶ 記憶番地 $\gamma(A_1) \subset M$
↓
$\Delta(i)$ + a(i) ⟶ 永久記録 ⟶ i

転置ファイル構成(Inverted file organization)．インデッ

クス集合 $A=\{c_1, c_2, \cdots, c_m\}$ とし M の分割 $M=\bigcup_{i=1}^{n} M_i$ ($M_i \cap M_j = 0$ for $i \neq j$) を考える。$\forall i$ に対して $f(i) \ni c_j$ ならば $|\Omega(i) \cap M_j| = 1$, $f(i) \not\ni c_j$ ならば $|\Omega(i) \cap M_j| = 0$ とする。検索は質問 Q に対してインデックス部分集合 $A_1=\{c_{i_1}, c_{i_2}, c_{i_3}, \cdots c_{i_t}\} \subset A$ が定まる ($i_1 < i_2 < \cdots < i_t$ とす)。これに対して $\gamma(A_1) \subset M_{i_1}$ なる $\gamma(A_1)$ を求め，そこに含まれる $\Omega(i)$, したがって $a(i)$ を得て永久記録から x を得る。$\gamma(A_1)$ を具体的に求めるには M_{i_1} のある記憶場所から $a(i)$ を定め，$a(i)$ を $M_{i_2}, M_{i_3}, \cdots M_{i_t}$ 内の $a(i)$ と比較し，もしマッチしていれば $\gamma(A_1)$ に入れるという手順を繰返し，M_{i_1} のすべての記憶場所についてこの手順を行う。これはいはゆる Table-lookup 方式で，検索時間は主にファイルの大きさ $|I| = v$ に依存す。

拡張された転置ファイル構成 (e, i, f, Ω), $\mathcal{A}=\{A_1 \mid A_1 \subset A\}$ = 検索の対象となる索引の部分集合族。$\mathcal{A}$ の A_1 に対して部分集合 $M_{A_1} \subset M$ をえらぶ。$f(i) \supset A_1$ ならば $|\Omega(i) \cap M_{A_1}| = 1$ であり，$\gamma(A_1) = M_{A_1}$ とする。「拡張された」という意味は部分集合 A_1 の一要素 c_i だけでなく A_1 全体を含むかどうかにより bucket M_{A_1} を作り，個々の address に当らないで直ちに M_{A_1} をとり，その中の address のみを探す点（バケツ法）にある。質問 Q に対する検索時間は主として $\mathcal{A}$ の構成要素（A_1 の数）に依存し，ファイルの大きさにはあまり影響されない。

§3 Combinatorial configuration

(1) (v, k, t, b)- configuration

要素の数 v の有限集合 A, $\mathcal{A}=\{A' \mid A' \subset A, |A'| \geqq t\}$ および b 個の部分集合 $A_1, A_2, \cdots A_b$ such that (i) $|A_j| \leqq k$ $(j=1, 2, \cdots b)$ (ii) すべての $A' \in \mathcal{A}$ に対して $A' \subseteq A_j$ なる整数 j が存在する。$(j=1, 2, \cdots b)$, で構成される configuration を Combinatorial configuration (v, k, t, b) と呼び, 以後これを $\underline{C(v, k, t, b)}$ と書く。とくに $\mathcal{A}$ に関し $|A'|$ に制限がないときこの configuration を $(A, k, \mathcal{A}, b)$- configuration と呼ぶ。

次のような configuration $C(v, k, \mu, \lambda_\mu)$ (S. Vajda, 1967, 文献[7]), すなわち A (たゞし $|A|=v$), A の部分集合 $A_1, A_2, \cdots, A_b$ such that (i) $|A_j| \leqq k$ $(j=1, 2, \cdots, b)$ (ii) μ 個の要素は block にともに λ_μ 回現われる (μ-wise balanced, $k \geqq \mu > 2$) を考える。こゝに $b=\lambda_\mu \binom{v}{\mu} / \binom{k}{\mu}$, $\lambda_\mu = \lambda_\tau \binom{k-\tau}{\mu-\tau} / \binom{v-\tau}{\mu-\tau}$ for $\tau=1, 2, \cdots \mu-1$。とくに $\lambda_0=b$, $\lambda_1=r$, $\lambda_2=\lambda$。この configuration は $C(v, k, t, b)$ と比較すると, μ と t とは同じものであり, b が v, k, μ を与えれば λ_μ のみで定まる ある特別の場合となる。$\mu=2$ のとき $C(v, k, 2, \lambda)$ は BIBD をなす。このとき $P(m, s)$ の点を A の要素, d-flat を部分集合 $A_1, A_2, \cdots, A_b$ (block) にとれば次の配置が得ら

れることは知られている。 $v=\phi(m,0;s)$, $b=\phi(m,d;s)$, $\lambda_1=r=\phi(m-1,d-1;s)$, $k=\phi(d,0;s)$, $\lambda_2=\lambda=\phi(m-2,d-s,s)$ if $d\neq 1$, $\lambda=1$ if $d=1$。 したがってこの $PG(m,s)$ を $\mathcal{C}(v,k,2,b)$ の一つの表現と見做せる。組合せ数学の上から $t>2$ の場合の $\mathcal{C}(v,k,t,b)$ が存在する最小な b を求めることが一つの問題であるが，ここでは $t>2$ に対して $\mathcal{C}(v,k,t,b)$ の幾何学的構成をすることが必要である。この点について RAY-Chaudhuri (1968) は $PG(m-1,s)$ を考え，$m>t'\geq t$ なる制限された t' と $PG(m-1,s)$ 内ですべての $(t-1)$ 次元部分空間が少なくとも一つの $(t'-1)$ 次元部分空間に含まれる最小個数の Block を取った場合に $\mathcal{C}(v,k,t,b)$ が存在することを示し，また $PG(m-1,s)$ 内で n 個の点を含む order d の cap を考えることにより，制限された条件内ではあるが $\mathcal{C}(v,k,t,b)$ の存在を示した。

(2) Multistage combinatorial configuration $(A,k,\mathcal{Q},b,p)$

A, $\mathcal{Q}$ は (1) と同じ。$b_0, b_{j_1}, b_{j_1j_2}, \cdots b_{j_1j_2\cdots j_c}$ を整数とし，部分集合 $A_{j_1,j_2,\cdots,j_d}$ $(1\leq c\leq p-1,\ 1\leq d\leq p,\ j_1=1,2,\cdots b_0,\ j_i=1,2,\cdots b_{j_1,j_2,\cdots,j_{i-1}},\ i=2,3,\cdots p)$ such that (i) $d=2,3,\cdots p$ に対し $A_{j_1,j_2,\cdots j_d}\subseteq A_{j_1,j_2,\cdots,j_{d-1}}$, (ii) すべての $A'\in\mathcal{Q}$ に対し $A_{j_1,j_2,\cdots,j_p}\supseteq A'$ なる整数 $j_1,j_2,\cdots j_p$ が存在する。(iii) $|A_{j_1,j_2,\cdots j_p}|\leq k$, (iv) $b=b_0+\sum'_{(j_1,j_2,\cdots,j_c)} b_{j_1,j_2,\cdots j_c}$

($c=1, 2, \cdots, p-1$) で構成する。とくに $\mathfrak{A}=\{A' \mid A' \subset A, |A'| \leqq t\}$ のときこの configuration を (v, k, t, b, p)-multistage configuration という。もしある整数 i ($i=1, 2, \cdots, d$) に対して $j_1' \geqq j_1, \cdots j_{i-1}' \geqq j_{i-1}, j_i > j_i$ なる d-tuple $(j_1, j_2, \cdots, j_d)$ は d-tuple $(j_1', j_2', \cdots, j_d')$ に先行するという。対応する d th stage block が A' を含む d-tuple $(j_1, j_2, \cdots, j_d)$ を A' の covering d-tuple というが，A' の d-tuple $(j_1, j_2, \cdots, j_d)$ が他のすべての A' の covering d-tuple に先行するときこの $(j_1, j_2, \cdots, j_d) = C_d(A')$ を d th stage covering index という。$d=1, 2, \cdots, p$ とすべての $A' \in \mathfrak{A}$ に対して d th stage covering index が $(d-1)$ th stage covering index を含むならば simple multistage configuration という。unistage configuration $\mathring{C}(k_{i-1}, k_i, t, b_{i-1})$, $i=1, 2, \cdots, p$, $k_0=n$, $k_p=k$ が存在すれば simple multistage configuration $\mathring{C}(v, k, t, b, p)$ が存在することが証明されている。ただし $b = b_0 + b_0 b_1 + \cdots + b_0 b_1 \cdots b_{p-1}$.

§4 Combinatorial filing system (c.f. 5)

(1) $\mathring{C}(v, k, t, b)$ にもとずく場合

Block $A_1, A_2, \cdots, A_b$ と $A' \subset A$ に対し $A' \subset A_j$ なる整数 j があれば A' は covered されているという。$C(A') = A'$ が cover

されている最小整数 j, $\mathcal{A}_j = \{A' \mid A' \subset A,\ c(A') = j \text{ for any } j\}$, 蓄積方式を与える mapping σ ; $f(i) \cap A_j = A' \rightleftarrows |\sigma(i) \cap M_{j,A'}| = 1$ が成立するように各 (j, A') に対して十分大きく disjoint な $M_{j,A'}$ をえらぶ ($j = 1, 2, \cdots$, 各 $A' \in \mathcal{A}_j$)。$M_j = \bigcup_{A' \in \mathcal{A}_j} M_{j,A'}$ を bucket, $M_{j,A'}$ を subbucket という。検索方式 R は

$$\text{質問 } Q \longrightarrow A' \subset A,\ |A'| \leqq t \longrightarrow \gamma(A') = \bigcup_{A_{j_0} \supset A'' \supset A'} M_{j,A''}$$

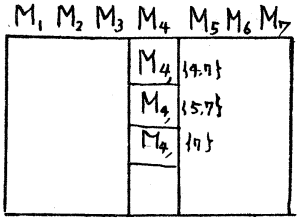
$$\downarrow$$

$$\sigma(i) + a(i) \longrightarrow i$$

こゝに $c(A'') = j_0$, $f(i) \cap A_{j_0} = A''$, i.e. $A'' \in \mathcal{A}_{j_0}$。

(例11) $C(7, 3, 2, 7)$ の場合について蓄積方式 S, 検索方式 R の実例を示す。PG(2,2) を考えると 7 個の点と 7 つの直線があり, 1 直線上に 3 点がのっている。7 個の点を A の要素 ($|A| = v = 7$), 7 つの直線を Block $A_1, A_2, \cdots, A_7$ ($b = 7$), $k = 3$。Incidence matrix は図のようになる。

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
1	1				1		1
2	1	1				1	
3		1	1				1
4	1		1	1			
5		1		1	1		
6			1		1	1	
7				1		1	1



蓄積方式 S ; $\mathcal{A}_4 = \{A' \mid A' \subset A,\ c(A') = 4\}$ を考えると $A' = \{4,7\}, \{5,7\}, \{7\}$ がある。したがって $(j, A') = (4, \{4,7\})$ に対して $M_{j,A'} = M_{4,\{4,7\}}$, $(j, A') = (4, \{5,7\})$ に対して $M_{4,\{5,7\}}$, $(j, A') = (4, \{7\})$ に対して $M_{4,\{7\}}$ なる disjoint な $M_{4,A'}$ をえらぶ。1 つのバケツ $M_4 = \bigcup_{A' \in \mathcal{A}_4} M_{4,A'}$ が定まる。mapping σ ; もし $f(7) = \{5,6,7\}$ ならば, $f(7) \cap A_4$

$= A' = \{5,7\} \in \mathcal{A}_4 \rightarrow |\Delta(7) \cap M_{4,\{5,7\}}| = 1$ となるように番地指定が必ずある。 検索方式 R ; $f(7) = \{5, 6, 7\}$ として $f(7) \supset A'$ ($|A'| \leqq t = 2$) は $A' = \{5,6\}, \{5,7\}, \{6,7\}, \{5\}, \{6\}, \{7\}$. 次に $A_{j_0} \supset A'' \supset A'$ (こゝに $j_0 = c(A'')$) なる A'' をえらび $M_{j,A''}$ を求めると $A_6 \supset \{6,7\} \supset \{6\}, \{7\} \rightarrow M_{6,\{6,7\}}$, $A_5 \supset \{5,6\} \supset \{5\}, \{6\} \rightarrow M_{5,\{5,6\}}$ $A_4 \supset \{5,7\} \supset \{5\}, \{7\} \rightarrow M_{4,\{5,7\}}$, $A_3 \supset \{3,6\}, \{4,6\} \supset \{6\} \rightarrow M_{3,\{6\}}$ $A_2 \supset \{3,5\}, \{2,5\} \supset \{5\} \rightarrow M_{2,\{5\}}$ となる。 したがって $\gamma(A') = \bigcup_{A_{j_0} \supset A'' \supset A'} M_{j_0,A''} = M_{2,\{5\}} \cup M_{3,\{6\}} \cup M_{4,\{5,7\}} \cup M_{5,\{5,6\}} \cup M_{6,\{6,7\}}$. この $\gamma(A')$ 内にあるすべての $\Delta(i)$, したがって $a(i)$ を取り出し, 更に永久記録より i を得る。

(2) $\zeta(v, k, t, b, p)$ にもとずく場合

蓄積方式 S ; $\mathcal{J}_p = \{ \mathbf{j} \mid p\text{th stage block が存在する } \mathbf{j} = (j_1, j_2, \cdots j_p)$ $j_1, j_2, \cdots j_p$ は正整数 $\}$, $\mathcal{A}_{\mathbf{j}} = \{A' \mid A' \subset A, c_p(A') = \mathbf{j}\}$, mapping Δ は十分大きく $M_{\mathbf{j},A'}$ をそらんでおいて各 $(\mathbf{j}, A')$ に対して ($A' \in \mathcal{A}_{\mathbf{j}}$), $f(i) \cap A_j = A'$, $c_p(A') = \mathbf{j}$ となるとき $|\Delta(i) \cap M_{\mathbf{j},A'}| = 1$ が成り立つように定める。 なお bucket $M_{\mathbf{j}} = \bigcup_{A' \in \mathcal{A}_{\mathbf{j}}} M_{\mathbf{j},A'}$ で $M_{\mathbf{j},A'}$ は subbucket である。 検索方式 R; $c_p(A'') = \mathbf{j}$ として, $\gamma(A') = \bigcup_{A_j \supseteq A'' \supseteq A'} M_{\mathbf{j},A''}$ を定める。

(例) $\zeta(v, k, t, b, p) = \zeta(40, 4, 2, 40 + 40 \times 13, 2)$

$PG(3,3)$ を考えると $v = \phi(3,0,3) = 40$, 2-flat の数 $= \phi(3,2,3)$ $= 40 = 1$st stage block の数 b_0, 2-flat 中の 1-flat の数 $= \phi(2,1,3)$

= 13 = 2nd stage block の数 b_1^0, 1-flat中の 0-flat の数 = $\phi(1,0,3)$ = 4 = 2nd stage block 内の点の数。

(例1) unistage configuration with minimum cycle $\theta(<v)$

$PG(m,s)=PG(3,3)$ で 1-flat 上の点の数 $\phi(1,0,3)=4$, ($\phi(3,0,3)$, $\phi(1,0,3)$)≠1 であるから m.c. $\theta<v$ をもった 1-flat がある。$(m+1, d+1)=(4,2)=2$, $\therefore j+1=2$, $\therefore j=1$, $\theta=\frac{s^{m+1}-1}{s^{j+1}-1}=\frac{3^4-1}{3^2-1}=10$ しかも 40 より小さい m.c. θ は 10 のみ。m.c. 10 の 1-flat 上の点; x^{c_0}, x^{c_0+10}, $x^{c_0+2\times10}$, $x^{c_0+3\times10}$ の 4 点で $v/\theta=4=r$, $m[1,0]=\frac{3+1}{2}-1=1$, $s[1,0]=3^2=9$, $d[1,10]=\frac{2}{2}-1=0$ であるから $v^*=\phi(1,0,9)=10$, $b^*=\phi(1,0,9)=10$, $k^*=\phi(0,0,9)=1$ となり, $A'\subset A_j$ を考慮して $C(v^*,k^*,t,b^*)=C(10,1,1,10)$ が存在する。(定理4, Yamamoto et al, 1966, 文献[8] および定理1の系1, Ray-Chaudhuri, 1968). この Configuration にもとづいてファイル構成することもできるが, trivial case ($k^*=1$) で実質的効果は薄い。

参考文献

[1] 伊大知良太郎・水田洋・藤川正信 (1968)「社会科学ドキュメンテーション」pp. 464, 丸善株式会社.

[2] Social science data archives in the United States (1967), pp 45. Council of social science data archives,

New York, U.S.A.

[3] 高橋達郎（1968）「情報検索」pp.203. 東洋経済新報社.

[4] F.W. Lancaster (1968)「Information Retrieval System. (characteristics, testing and evaluation). pp. 222. John Wiley & Sons, Inc., New York.

[5] 中村幸雄（1968）「情報処理」I－言語と概念－（情報科学講座 C・11・1）pp.115. 共立出版株式会社.

[6] D.K. RAY-Chaudhuri (1968): Combinatorial Information retrieval systems for files. SIAM J. appl. Math. vol 16. p. 973－992.

[7] S. Vajda (1967)「Patterns and configurations in finite spaces」pp.120. Charles Griffin, London.

[8] S. Yamamoto, T. Fukuda and N. Hamada (1966): On finite geometries and cyclically generated incomplete block designs. J. science of the Hiroshima Univ. Ser A-1, vol. 30, p. 137－149.