

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 archives の活動については文献[2]）、ここでは事実検索（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 — これは純粹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° ファイル構成に關して 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。こゝ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 $PG(m, \Delta)$ が minimum cycle $\theta(k, v)$ 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 Invented) 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。

ファイル F ; a triplet $F(I, A, f)$ であつて, 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 = $\{0, \text{正整数}\}$ (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 $S(I, M, \Delta)$ 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。 $\Delta = I$ から M の部分集合への many to many mapping (番地付け), $I \ni i$ に対して $\Delta(i) =$ 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 $\Delta(i)$ 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 $M, CM, A_1 \subset A$ に対して $f(i) \supset A_1 \iff |M_1 \cap \Delta(i)| = 1$ (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 $R(\mathcal{A}, M, r)$ であって, $\mathcal{A} = A$ の部分集合のある族, $r = r(A_1)$ が A_1 に関して valid であるような $\mathcal{A}$ から M の部分集合への mapping. (定義) すべての $A_1 \in \mathcal{A}$ に対して $|A_1| \leq t$ 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 $Q \rightarrow A_1 \in \mathcal{A} \rightarrow$ 記憶番地 $r(A_1) \subset M$

↓
 $\Delta(i) + a(i) \rightarrow$ 永久記録 $\rightarrow i$

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

クス集合 $A = \{c_1, c_2, \dots, c_m\}$ とし M の分割 $M = \bigsqcup_{i=1}^n M_i$ ($M_i \cap M_j = \emptyset$ for $i \neq j$) を考える。 I の i に対して $f(i) \in c_j$ ならば $|\Delta(i) \cap M_j| = 1$, $f(i) \notin c_j$ ならば $|\Delta(i) \cap M_j| = 0$ とする。 検索は質問 Q に対して インデックス部分集合 $A_1 = \{c_{i_1}, c_{i_2}, c_{i_3}, \dots, c_{i_t}\} \subset A$ が定まる ($i_1 < i_2 < \dots < i_t$ 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 $\Delta(i)$, したがって $a(i)$ を得て 永久記録から i 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 $M_{i_2}, M_{i_3}, \dots, M_{i_t}$ 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。 これはいわゆる Table-lookup 方式で、検索時間は主にファイルの大きさ $|I| = n$ に依存す。

拡張された転置ファイル構成 (e. i. f. Δ), $\mathcal{A} = \{A_1 \mid A_1 \subset A\}$ = 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。 $f(i) \in A_1$ ならば $|\Delta(i) \cap M_{A_1}| = 1$ 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_k 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 (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 , $\mathcal{A} = \{A' \mid A' \subset A, |A'| \leq t\}$ および b 個の部分集合 $A_1, A_2, \dots, A_b$ such that (i) $|A_j| \leq k$ ($j=1, 2, \dots, b$) (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 ($j=1, 2, \dots, b$), で構成される configuration を Combinatorial configuration (v, k, t, b) と呼び、以後これを $\mathcal{C}(v, k, t, b)$ 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 $\mathcal{C}(v, k, \mu, \lambda_\mu)$ (S. Vajda, 1967, 文献[7]), すなわち A (ただし $|A|=v$), A の部分集合 $A_1, A_2, \dots, A_b$ such that (i) $|A_j| \leq k$ ($j=1, 2, \dots, b$) (ii) μ 個の要素は block にともに λ_μ 回現われる (μ -wise balanced, $k \geq \mu > 2$) 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 $\tau=1, 2, \dots, \mu-1$, $\lambda_0 = b$, $\lambda_1 = \tau$, $\lambda_2 = \lambda$ 。この configuration は $\mathcal{C}(v, k, t, b)$ 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 $\mathcal{C}(v, k, 2, \lambda)$ は BIBD をなす。このとき $P(m, d)$ の真を A の要素, d -flat を部分集合 $A_1, A_2, \dots, A_b$ (block) にとれば次の配置が得ら

れることは知られている。 $v = \phi(m, 0; \Delta)$, $b = \phi(m, d; \Delta)$, $\lambda_1 = \gamma = \phi(m-1, d-1; \Delta)$, $k = \phi(d, 0; \Delta)$, $\lambda_2 = \lambda = \phi(m-2, d-\Delta, \Delta)$ if $d \neq 1$, $\lambda = 1$ if $d = 1$. (したがってこの $PG(m, \Delta)$ を $C(v, k, 2, b)$ 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 $C(v, k, t, b)$ 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 $C(v, k, t, b)$ 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 $PG(m-1, \Delta)$ を考え、 $m > t' \geq t$ なる制限された t' と $PG(m-1, \Delta)$ 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 $C(v, k, t, b)$ が存在することを示し、また $PG(m-1, \Delta)$ 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 $C(v, k, t, b)$ 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。 $b_0, b_{j_1}, b_{j_1 j_2}, \dots, b_{j_1 j_2 \dots j_c}$ を整数とし、部分集合 $A_{j_1 j_2 \dots j_d}$ ($1 \leq c \leq p-1$, $1 \leq d \leq p$, $j_1 = 1, 2, \dots, b_0$, $j_i = 1, 2, \dots, b_{j_1 j_2 \dots j_{i-1}}$, $i = 2, 3, \dots, p$) such that (i) $d = 2, 3, \dots, p$ に対し $A_{j_1 j_2 \dots j_d} \subseteq A_{j_1 j_2 \dots j_{d-1}}$, (ii) すべての $A' \in \mathcal{Q}$ に対し $A_{j_1 j_2 \dots j_p} \supseteq A'$ なる整数 $j_1, j_2, \dots, j_p$ が存在する。(iii) $|A_{j_1 j_2 \dots j_p}| \leq k$, (iv) $b = b_0 + \sum_{(j_1, j_2, \dots, j_c)} b_{j_1 j_2 \dots j_c}$

$(c=1, 2, \dots, p-1)$ で構成する。とくに $\mathcal{Q} = \{A' \mid A' \subset A, |A'| \leq t\}$ のときこの configuration を (v, k, t, b, p) -multi-stage configuration と言う。もしある整数 i ($i=1, 2, \dots, d$) に対して $j'_1 \geq j_1, \dots, j'_{i-1} \geq j_{i-1}, j_i > j_i$ なる d -tuple $(j_1, j_2, \dots, j_d)$ は d -tuple $(j'_1, j'_2, \dots, j'_d)$ に先行すると言う。対応する d th stage block が A' を含む d -tuple $(j_1, j_2, \dots, j_d)$ を A' の covering d -tuple と言うが、 A' の d -tuple $(j_1, j_2, \dots, j_d)$ が他のすべての A' の covering d -tuple に先行するときこの $(j_1, j_2, \dots, j_d) = c_d(A')$ を d th stage covering index と言う。 $d=1, 2, \dots, p$ とすべての $A' \in \mathcal{Q}$ 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 と言う。unistage configuration $C(k_{i-1}, k_i, t, b_{i-1}), i=1, 2, \dots, p, k_0=n, k_p=k$ が存在すれば simple multistage configuration $C(v, k, t, b, p)$ が存在することが証明されている。ただし $b = b_0 + b_0 b_1 + \dots + b_0 b_1 \dots b_{p-1}$.

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

(1) $C(v, k, t, b)$ にもとずく場合

Block $A_1, A_2, \dots, A_b$ と $A' \subset A$ に対し $A' \subset A_{j_i}$ なる整数 j_i 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 , $\mathcal{U}_j = \{A' \mid A' \subset A, c(A') = j \text{ for any } j\}$,
蓄積方式を与える mapping $\alpha; f(i) \cap A_j = A' \Leftrightarrow |\alpha(i) \cap M_{j,A'}| = 1$
が成立するように各 (j, A') に対して十分大きく disjoint
な $M_{j,A'}$ を与える ($j=1, 2, \dots, \forall A' \in \mathcal{U}_j$). $M_j = \bigcup_{A' \in \mathcal{U}_j} M_{j,A'}$
を bucket, $M_{j,A'}$ を subbucket という。検索方式 R は

$$\text{質問 } Q \rightarrow A' \subset A, |A'| \leq t \rightarrow r(A') = \bigcup_{A_j \supset A' \supset A'} M_{j,A''}$$

$$\downarrow$$

$$\alpha(i) + \alpha(i) \rightarrow i$$

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

(例) $C(7, 3, 2, 7)$ 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 本の直
線があり, 1 直線上に 3 点がついている。 7 個の点を A の要
素 ($|A| = v = 7$), 7 本の直線を Block $A_1, A_2, \dots, A_7$ ($b = 7$),
 $k = 3$. 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

M_1	M_2	M_3	M_4	M_5	M_6	M_7
			$M_{4,\{4,7\}}$			
			$M_{4,\{5,7\}}$			
			$M_{4,\{7\}}$			

蓄積方式 S ; $\mathcal{U}_4 = \{A' \mid A' \subset A, c(A') = 4\}$ 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 $M_{j,A'}$ を与える。 1 つのバ

ケツ $M_4 = \bigcup_{A' \in \mathcal{U}_4} M_{4,A'}$ が定まる。 mapping

α ; もし $f(7) = \{5, 6, 7\}$ ならば, $f(7) \cap A_4$

$= A' = \{5, 7\} \in \mathcal{U}_4 \rightarrow |\Delta(7) \cap M_{4, \{5, 7\}}| = 1$ となるように番地指定
 が必要である。 検索方式 R ; $f(7) = \{5, 6, 7\}$ として $f(7) \supset A'$
 $(|A'| \leq t = 2)$ 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 $M_{j_0, A''}$ 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 $\gamma(A') =$
 $\bigcup_{A_{j_0} \supset A'} M_{j_0, A'} = M_{2, \{5\}} \cup M_{3, \{6\}} \cup M_{4, \{5, 7\}} \cup M_{5, \{5, 6\}} \cup M_{6, \{6, 7\}}$.
 この $\gamma(A')$ 内にあるすべての $\Delta(i)$, (したがって $\alpha(i)$) を取り出
 し, 更に永久記録より i を得る。

(2) $\hat{C}(v, r, t, b, p)$ にもとづく場合

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

(例) $\hat{C}(v, r, t, b, p) = \hat{C}(40, 4, 2, 40 + 40 \times 13, 2)$

$PG(3, 3)$ を考えたと $v = \phi(3, 0, 3) = 40$, 2-flat の数 $= \phi(3, 2, 3)$
 $= 40 = 1\text{st stage block の数 } b_0$, 2-flat 中の 1-flat の数 $= \phi(2, 1, 3)$

$= 13 = 2\text{nd stage block の数 } b_i^0$, 1-flat中の 0-flatの数 $= \phi(1, 0, 3) = 4 = 2\text{nd stage block 内の桌の数}$.

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

$PG(m, \Delta) = PG(3, 3)$ で 1-flat上の桌の数 $\phi(1, 0, 3) = 4$, $(\phi(3, 0, 3), \phi(1, 0, 3)) \neq 1$ であるから m.c. $\theta < v$ をもった 1-flatがある。

$$(m+1, d+1) = (4, 2) = 2, \therefore j+1=2, \therefore j=1, \theta = \frac{\Delta^{m+1}-1}{\Delta^{j+1}-1} = \frac{3^4-1}{3^2-1} = 10$$

しかも 40 より小さい m.c. θ は 10 のみ。m.c. 10 の 1-flat 上の桌 $x^{c_0}, x^{c_0+10}, x^{c_0+2 \times 10}, x^{c_0+3 \times 10}$ の 4 桌で $v/\theta = 4 = r, m[1, 0] = \frac{3+1}{2} - 1 = 1, \Delta[1, 0] = 3^2 = 9, d[1, 10] = \frac{2}{2} - 1 = 0$ 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 ($k^* = 1$) で実質的效果は薄い。

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.