数理解析研究所講究録
第211巻 1974年 47-59

情報検索に関連した組合せ論の話題

広大 理　山本純恭
新居浜工専　潮和彦
広大 理　重枝新成
広大 計算センター　池田秀人
広大 理　玉利文和
広大 教育　浜田昇

§0. はじめに

広島大学計算センターでは，1969年以来，情報検索についてグループ研究をつづけている．時の流れとともに消長があり，またメンバーの異動もあったが，細々ながら若干の成果をあげて今日に至っている．

ここでわれわれが提供しようとする話題は，最近1年余の研究の中間結果や成果のうち，組合せ論に関係する部分が中心である．その1つは，IBMのGhoch[4][5][6]の提案しているConsecutive retrieval property（一連検索可能性－CR性）をもつファイル・システムの研究から洗い出された問題で

あり，いま1つは，Chow[10]の提案しているNew balanced filing schemes（NBFS）の研究にヒントを得たHUBFS（Hiroshima University BFS）の構成にまつわるグラフ理論の問題である．

問題の性質上，形式化以前の話題の背景について簡単に記すことにする．

§1. 話題の背景

情報検索に電子計算機が用いられるようになってすでに久しいが，レコードそのものの特性を検索目的との対比の上で抽出し，フォーマット化する問題，検索要求すなわち質問の集合の構造解明と体系化の問題，フォーマット化されたレコードを収納するファイル方式あるいは検索方式の問題など，たがいに無関係に考察することのできない問題があって，今後の進展にまつべきものが多い．

われわれのとりあげてきた問題は，計算機とのかゝわり合いの大きい側面であって，フォーマット化されたレコードを収納し，定式化された質問の集合に効率よく応答するファイル方式の研究である．

もっとも原始的なファイル方式は，レコードの発生順に一連にファイルする積み上げ方式（方式というに値しない）と

もいうべき単純な収納方式であろう．検索要求すなわち質問の発生ごとにファイル全体を通じて検索し該当するレコードを検索しなければならないから，収納の手順こそ最小であり，ファイルの容量もまた最小であるが，検索の手数は最大で，ファイルの容量が大きくなればいうまでもなく実時間的でなくなる．

これに対して，主要な質問項目に対して1対1に該当するレコード（その固有番号等）を収納するバケツを用意するいわゆる転置ファイル（Inverted file）とよばれる方式がある．主要な質問に関する限り，該当するバケツの内容が該当するレコードのすべてであるから，検索時間は最小である．しかしこの方式によれば，同一レコードを重複して多数のバケツに収納する必要が生じ，いわゆる冗長度の大きいファイルとなり，当然ファイルの容量も大きくなる．したがってレコードの収納に手数と時間がかかるようになる．また重要な質問の“and”または“or”からなる質問に対しては，関係するバケツの内容のマッチングが必要になり，容量の増大とともに無視できない負担となる．

後者の欠陥を除去するためには，予想される質問のすべてを対象として転置ファイルを作ればよいが，明らかに前者の欠陥がさらに助長される．

この両者の欠陥を調和する方式の模索から生まれた研究が1960年代の後半IBMワトソンのAbraham, Ghosh, Ray-Chaudhuri等と, ノースカロライナ大学のBose教授等の協同研究である.

§2. BFS, BMFSをめぐる話題

Abraham等[1]は, 有限射影幾何を用いて, 2項目の質問に対するBFS$_2$(Balanced filing schemes)を構成した. l項目について2値のレコードの全体をファイルするにあたり, 項目を点に, バケツを直線に対応させると, 結合行列は$\lambda=1$のBIBD(l, k, λ)を作るから, 2点を指定する2項目質問はすべてただ1つのバケツに対応するのみならず, 1つのバケツは$\binom{k}{2}$個の2項目質問に対応する. バケツ内にサブバケツを設け, 同一バケツ内でのレコードの重複収納を許さないことにすると, 2項目の質問全体に対して$\binom{l}{2}$個のバケツを作成する転置ファイルに比べて, 重複収納の回数が減少するというアイデアである. もちろん$\lambda=1$のBIBDを用いればBFS$_2$を構成することができる.

各項目が多値(A値$=p^n$)の場合についてGhosh等[2]はBMFS$_2$を有限幾何を用いて構成している. われわれもこの研究会の成果として有限射影幾何におけるフラットの巡回性を積

極的に利用するファイル方式を発表した[3].

この種の問題を3項目以上の質問に対して拡張することは容易でなく，一応棚上げになっている．

§3. CR性をもつファイルの話題

いま1つ興味ある研究方向は，最近Ghosh[4]の提案した一連検索可能性（Consecutive retrieval property，CR性）をもつファイルという概念である．たとえばA，B 2つの質問に該当するレコードを配列する場合，個別に転置ファイルを作ることなく，1つのバケツ内に $A\cap B^c$，$A\cap B$，$A^c\cap B$ の順に配列しておけば，質問A，Bのみならず $A\cap B$，$A\cup B$ に対してもレコードが一連に配列されているから，先頭番地とレコードの個数が与えられていると，最小の手数で該当するレコードの検索ができるという発想である．もちろんこのような配列がすべての質問に対して可能なわけではない．事実A，B，Cの3質問に対してすら $A\cap B\cap C^c$，$A\cap B^c\cap C$，$A^c\cap B\cap C$ がいずれも空でないとき，CR性をもつファイルは作れない．また $A\cap B^c\cap C^c$，$A^c\cap B\cap C^c$，$A^c\cap B^c\cap C$，$A\cap B\cap C$ が空でない場合も同様である．

Ghoshは，レコードの集合から質問の集合に対してCR性をもつための十分条件を自明なものまで含めて数多く提示し

ている．ここのCR性の問題は質問Xレコードの結合行列である0−1行列についていえばそのC1性とよばれるものであり，グラフ理論の問題にいいかえることもできる．

われわれは，CR性をもつための必要十分条件の追及という方向よりはむしろ，Ghoshのあげている興味ある例に注目した．それは4項目について2値のレコードの全体

$$R^{(4)}=\{(\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4);\ \varepsilon_i=1\text{ or }0\}$$

を4項目に対する質問の全体

$$Q^{(4)}=\{(\delta_1,\delta_2,\delta_3,\delta_4);\ \delta_i=1\text{ or }X\}$$

（ただし $\delta_i=1$ は該当するレコードが第i項目に該当することを要求し，$\delta_i=X$ は第i項目に該当するか否かを問わないことを示す．また$(X,X,\cdots,X)$を除く）

を考えるとき，

$$Q^{(4)}=Q_1^{(4)}\cup Q_2^{(4)}\cup Q_3^{(4)},\quad Q_i^{(4)}\cap Q_j^{(4)}=\phi$$

の3成分にうまく分解することにより，$R^{(4)}$はどの$Q_i^{(4)}$に対してもCR性をもつようにできることを例示したものである．

われわれは，$Q^{(l)}$（$=(X,X,\cdots,X)$）を分割して$R^{(l)}$に対してCR性をもつようにする問題（CR分割と名づけた）をとりあげ，いくつかの結果を得ている[8]．すなわち

定理1　CR性をもつバケツに含まれる質問の最大数は

$2l-1$である．

定理2　最小のCR分割数の下界は

$$t_l = \left\lceil \frac{\binom{l}{\lfloor l/2 \rfloor} + 1}{2} \right\rceil$$

である.

定理2について下界を与える最小CR分割の存在と定理1で得た最大バケツを軸として最小CR分割を生成するアルゴリズムを追及しているが，いま一歩完全な解決に達していない.

§4.　NBFS$_2$からHUBFS$_2$へ

Chow[10]はBFS$_2$を構成する場合，そのパラメーターに制約があること，有限体上の演算が必要であることの制約に着目し，$\binom{l}{2}$個の2項目質問(i, j) $(0 \leq i < j \leq l-1)$に$il+j$を対応させることにより順序をつけ，この順序に従って先頭からC個づつあつめてバケツを構成する方式を提案し，NBFS$_2$（New BFS）と名づけた（一般にK項目の場合でも順序づけは同様）．質問をまとめてバケツを作ることが冗長度を軽減するという点でBFS$_2$と同等であるが，パラメーターの制約がないこと，バケツの番地の算出が簡単なことが特色であるとしている．またBFS$_2$と同じパラメーターをもつNBFS$_2$のいくつかについて両者の冗長度を数値的に比較し，NBFS$_2$がつねに冗長度が小さいことを指摘して

いる.

われわれは, BFS_2 より $NBFS_2$ の冗長度が小さくなる原因の追及に興味をもち, バケツのグラフ的構造(項目を点, 2項目質問を線に対応)とそのバケツの冗長率の関係について調べた. その結果, グラフ構造がいわゆる claw 型($K_{1,c}$)である場合かつその場合に限り最小の冗長率を与えることがわかった[定理3]. 冗長率を, レコードの分布の一様性の仮定のもとに求める通常の方式のみならず, 各項目に該当する確率がそれぞれ一定値 p ($0<p<1$) かつたがいに独立という仮定のもとで考えても同様の結論が得られた(定理3).

この観点から Chow の結果をみると, BFS_2 のバケツ構造はすべて完全グラフ K_k であるのに対して $NBFS_2$ のそれは冗長率最小の claw 型バケツが大半を占めていることがわかった. そこで $Q^{(l)} = E(K_l)$ とみなすことができるから, $\{K_{1,c}^{(\alpha)}\}$ を決定して

$$Q^{(l)} = E(K_l) = \bigcup_{\alpha} E(K_{1,c}^{(\alpha)}) = \bigcup_{\alpha} Q_{\alpha}^{(l)}$$

$$E(K_{1,c}^{(\alpha)}) \cap E(K_{1,c}^{(\beta)}) = \phi \quad (\alpha \neq \beta)$$

とすることができないかという問題が生じてきた. この問題は, 完全グラフの claw 分解定理(定理5)を要求した. またその証明はバイグラフの存在定理(定理4)を必要とした

以下バケツの冗長率について得た結果と，完全グラフの分解定理，バイグラフの存在定理についてのべる．

<u>定理3</u>　ℓ項目について2値のレコード$(X_1, X_2, \cdots, X_\ell)$の分布が

$$P_r\{(X_1, X_2, \cdots, X_\ell) = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_\ell)\} = p^{\Sigma\varepsilon_i}(1-p)^{\ell-\Sigma\varepsilon_i}$$

であるとき，c $(1 \leq c \leq \ell-1)$個の2項目質問からなるバケツBのうち冗長率$P_r\{(X_1, X_2, \cdots, X_\ell) \in B\}$が最小となるもののグラフ型は$claw\ (K_{1,c})$であってそれに限る．claw型バケツの冗長率は

$$R_H^{(2)} = p\{1-(1-p)^c\}$$

である．

(注)　$p=\frac{1}{2}$のときが通常冗長率とよばれているものである．また，この定理は，完全グラフK_{k-1}を共有するc個の完全グラフからなるグラフを*Extended claw*とよぶことにすると，c個のk項目質問からなるバケツに拡張される．

$m+n$個の数の組$\Pi = \{(r_1, r_2, \cdots, r_m), (s_1, s_2, \cdots, s_n)\}$が非負の整数Nの$(m, n)$分割というのは

$$\sum_{i=1}^{m} r_i = \sum_{j=1}^{n} s_j = N, \quad r_i \geq 0, \ s_j \geq 0$$

のときをいう．またNの(m, n)分割$\Pi = \{(r_1, r_2, \cdots, r_m), (s_1, s_2, \cdots, s_n)\}$がバイグラフ的であるということを，バイグラフ$G_{m,n}$が存在して，頂点の集合を$V(G_{m,n}) = \{a_1, \cdots, a_m\}$

$\cup \{b_1, \cdots, b_n\}$ とするとき（但し $a_i \neq b_j$）

$$deg(a_i) = r_i$$
$$deg(b_j) = s_j$$

がすべての i, j について成りたつときと定義すると，次の定理が成りたつ．

定理4 （バイグラフの存在定理）（Ryser, H.J. 1957）

非負の整数Nの (m,n) 分割の組 $\Pi = \{(r_1, r_2, \cdots, r_m), (s_1, s_2, \cdots, s_n)\}$ がバイグラフ的であるための必要十分条件は，$r_1 \geq r_2 \geq \cdots \geq r_m$ とするとき，m個の不等式

$$\sum_{i=1}^{l} r_i \leq \sum_{j=1}^{n} \min(l, s_j) \quad l = 1, 2, \cdots, m$$

が成りたつことである．

またこの定理4を応用して次の定理が証明できる．

定理5 完全グラフ K_l が c-claw分解（すなわち完全バイグラフ $K_{1,c}$ への edge-disjoint 分解）可能であるための必要十分条件は

(i) $\binom{l}{2} \equiv 0 \pmod{c}$

(ii) $l \geq 2c$

が成りたつことである．

次の定理は定理4に基づくバイグラフの構成アルゴリズムを与える．

定理6 （アルゴリズム）

非負の整数Nの (m,n) 分割の組 $\Pi = \{(r_1, r_2, \cdots, r_m), (s_1, s_2, \cdots, s_n)\}$ がバイグラフ的であるための必要十分条件は $s_1 \geq s_2 \geq \cdots \geq s_n$ とするとき

(i) $r_1 \leq \sum_{j=1}^{n} \min(1, s_j)$

(ii) $\Pi' = \{(r_2, r_3, \cdots, r_m), (s_1-1, s_2-1, \cdots, s_{r_1}-1, s_{r_1+1} \cdots s_n)\}$ がバイグラフ的である.

参考文献

BFS, BMFSに関するもの

[1] Abraham, C.T., Ghosh, S.P. and Ray-Chaudhuri, D.K. (1968). File organization schemes based on finite geometries. Information and Control 12 143-163.

[2] Ghosh, S.P. and Abraham, C.T. (1968). Application of finite geometry in file organization for records with multiple-valued attributes. IBM J. Res. Develop. 12 180-187.

[3] Yamamoto, S., Teramoto, T. and Futagami, K. (1972). Design of a balanced multiple-valued filing scheme of order two based on cyclically generated spread in finite projective geometry. Information and Control 21 72-91.

CR性に関するもの

[4] Ghosh, S. P. (1970). On the theory of consecutive storage of relevant records. IBM Res. Rep. No. RJ 708.

[5] Ghosh, S. P. (1971). File organization: Consecutive storage of relevant records on a drum type storage. IBM Res. Rep. No. RJ 895.

[6] Ghosh, S. P. (1972). File organization: The consecutive retrieval property. Comm. ACM 15 802-808.

[7] 山本純恭, 潮和彦 (1973). Consecutive retrieval property をもつファイルに関するS. P. Ghoshの研究について. 対話型情報処理に関する研究 24-30.

[8] 山本純恭, 潮和彦, 池田秀人, 玉利文和 (1973). Consecutive retrieval property をもつファイルについて-II. 対話型情報処理に関する研究 89-105.

[9] 山本純恭, 潮和彦, 池田秀人, 玉利文和 (1973). Consecutive retrieval property をもつ file について. 情報処理学会第14回大会予稿集 59-60.

NBFSに関するもの

[10] Chow, D. K. (1969). New balanced-file organization schemes. Information and Control 15 377-396.

[11] 山本純恭, 潮 和彦, 重枝新成 (1973). Combinatorial filing system をめぐって. 特定研究集会(昭和48年8月東北大学)資料.

[12] 山本純恭, 重枝新成, 潮和彦, 池田秀人, 浜田昇 (1973). Balanced filing system について－I 最小の冗長率をもつバケツの構造. 情報処理学会第14回大会予稿集 55-56.

[13] 山本純恭, 潮和彦, 重枝新成, 池田秀人, 浜田昇 (1973). Balanced filing system について－II HUBFS$_2$の構成. 情報処理学会第14回大会予稿集 57-58.

[14] 山本純恭, 重枝新成, 潮 和彦, 池田秀人 (1973). 新型の Balanced filing system (HUBFS) について－II. 特定研究集会(昭和48年12月大阪大学)資料.