

分 類 (2)*

淵 一 博**

D. 磁気テープによる分類

ファイル媒体として磁気テープを用いるようになると、オンラインによる集中一貫データ処理が可能になってくる。ところが磁気テープにはファイル媒体として特異な性格——すなわち逐次式ファイル (sequential file) であるということ——があるので、それに伴って多くの問題が生じてくる。この逐次性——ファイル上の項目を次々に処理していかなければならないこと、あるいは、ファイル上のある項目だけを処理するにしても、その位置を探し出すために、それに至るすべての項目を逐一読んでいかなければならない（そのための時間は、それに至るすべての項目に何か意味のある処理をほどこすことにしても、ただ単に読んでいくだけでも、実質的にあまり差はないのである）ということ、がデータ処理のある局面では大きな弱点となることがある。磁気テープを使うデータ処理組織ではその欠点をできるだけ回避しながら処理を進めるようにするのが磁気テープによるデータ処理技術の眼目であるといえるであろう。

その努力を装置そのものに転嫁しようとするれば、いわゆるランダム・アクセス・ファイル (random access file) への要請となる。それも目下開発されつつあるとはいえ、現在のところ磁気テープの方が技術的に安定しており、使用者側からみれば、ここ当分の間は磁気テープによるデータ処理技術の開発が課題であるといえるであろう。

磁気テープが逐次的であるということは、項目を逐一処理できるように順に並べておくことが必要であるということを含め、分類操作が特殊な重要性をもつことになる（これは序論に述べたような分類の便宜の故にパンチカードから磁気テープへの転換へのためらいがあったとすれば一種の歴史の皮肉である）。

一方磁気テープの逐次性によって、磁気テープによる分類で使える方法に制限が生じてくる。B章で述べた基本操作の中で基数法 (radix sorting) と併合法

(sorting by merging) にはほぼ限定されてしまうことになる。

外部装置が関連するときには一般に深刻になることであるが、機械組織での諸便宜が磁気テープによる分類を考察するさいの大きな要因になってくる。すなわち、テープ台数、バッファの有無、同時制御できるテープ台数、逆方向読み込みの便宜などが問題となってくる。

ファイルが磁気テープ一巻に納まっているときは、以下に述べるような方法で、分類操作は自動的に（計算機の制御によって）完結するのであるが、ファイルの規模が大きくなるとテープ一巻に納まりきれなくなり、その場合には、リールのかえかけという人手による操作が必要になってきて、自動一貫処理というたてまえがくずれてくる。

D-1 基数法による分類

B-2 で述べた基数法の手順は磁気テープの場合に適用することができる。一般的な得失についてはそこに述べたことがこの場合にもあてはまる。

p 進法のキーであれば、 $p+1$ 台のテープ装置を使う。最初の分類されていないテープを1台にかけ、残りの p 台に“ふり分け”をする。ふり分けが終れば全部のテープを巻き戻し、 p 個の類に分けられたものを順次最初のテープに“回収”する。回収が終れば再び全テープを巻き戻し、次の桁について上のサイクルを繰り返す。このやり方だと1サイクルに全ファイル長の巻き戻しを2回する必要がある。

2 p 台のテープ装置があれば、回収とふり分けを同時に行なうことができる。まず1台に最初のテープをかけ、 p 個のテープにふり分ける。この p 個のテープを巻き戻し、順に回収していくと同時に残りの p 台（最初のテープも含めて）に、ふり分けをする。このやり方だと情報の移しかえの時間が $1/2$ になり、しかも巻き戻しは1サイクルに1回、約 $1/p$ ファイル長（項目の分布が一様だとして）のものを行なえばよいことになる。

実際には、キーの基数 (base) によらず、使えるテープ装置の数は制限されている。そこで一般には、キ

* Sorting 2, by Kazuhiro Fuchi (Electrotechnical Laboratory, Tokyo)

** 電気試験所電子部

一をもっと小さい基数に変換して使うことになるであろう。オンライン分類のときは、この変換を分類プログラムに組み込むことは容易であるし、磁気テープの読み書き時間、すなわち情報の転送時間より、計算機内での処理時間の方が短いのが普通であるから、この問題はあまり由々しくはないであろう。

巻き戻しが必要なことは、磁気テープの欠点の一つである。巻き戻しの場合は、テープ前進の場合の約2倍の速度になるようになっている装置が多いが、それでも、 $p+1$ 台を使う方法では分類時間の約1/3は全然無駄な巻き戻し時間に使われることになる。

巻き戻しのさいにも情報を読み込めるようになっていれば（この時は前進の速度と同じである）、巻き戻し時間の無駄はなくなる。

逆読みが可能な場合の MSD 法として次のようなカスケード法 (cascading radix sorting) が考えられている³⁾。 $p+1$ 台を使うとして、最初のテープを p 台にふり分ける。ふり分けられたものの最初の1台を読み返しながらか、残りの p 台（最初のものを含めて）にふり分ける。この2回目のふり分けは、最初のふり分けの後に書き加えるのである。これは停止している場所から始めれば自然にそうなるわけである。この2回目のふり分けの最初のものをまたふり分けて、続書き加える。徐々に細かくしていき十分細分されたとき（1項目になってしまうか、または、計算機内で内部分類できるくらいになったとき）、それらを順に回収して、上の分割過程を逆に一つさかのぼる。 p 個の細分について徐々にふり分け、回収を続けて、終れば、その p 個の細分を順に回収し、また一段さかのぼる。その段階での隣りのブロックについて同じようなことを繰返し、その段階での分類が終ればまた一段さかのぼる。徐々に下降、上昇を繰返し、完全にさかのぼって1本のテープに戻れば、そこで全分類が完了したことになる。

D-2 併合法による分類

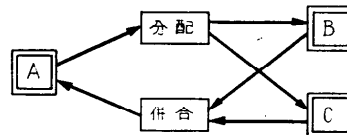
B-3 で述べた併合法 (sorting by merging) は磁気テープによる分類に使える。この方法は磁気テープ分類に最も適していると考えられている。その理由を挙げると、併合の操作が逐次式ファイルの性質に合っている、すなわち逐次性が併合操作に何ら不都合をもたらさないこと、また、併合法による分類時間はキーの大きさに関係せず、全項目数にだけであること、さらに、自然な連系 (string) を使うことにより、す

でに部分的になされている分類を生かすことができること、という諸点になる。

この節では2重併合 (2-way merge) に話を限って、その手順を考えることにする。

(a) テープ装置3台の場合

まず最初のテープはテープ装置Aにあるものとする(第1図)。それを



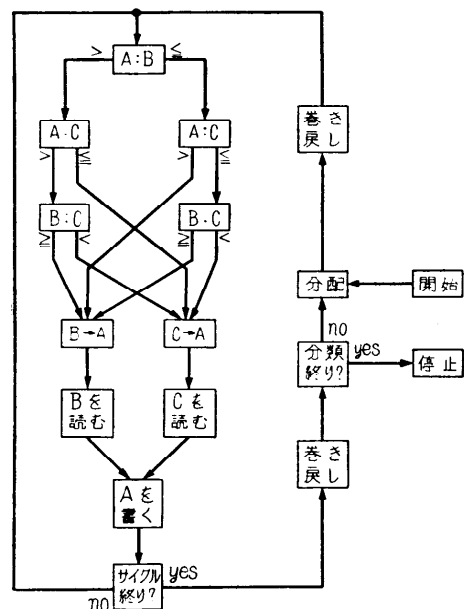
第1図

(1) 分配 (distribute)

の過程で B, C に移す。すなわち、Aテープから最初の項目を読みBテープに移す。それから始まる連系はBテープに次々に入れる。連系の切れ目で出力テープをBからCに切り換え、次の連系をCテープに入れる。徐々に連系をBとCに交代に入れていく。この分配が終れば、

(2) 併合 (merge)

に移る。すなわちBとCの各一連系からそれらを併合した長い一連系を作り、Aテープに移す。これはBとCから一項目ずつ読み、Aに一項目ずつ出していくだ



第2図

けで実現できる。その手順を第2図の流れ図に示す。

テープ装置、そのバッファ、そのキーを A, B, C の同じ記号で表わすことにする。

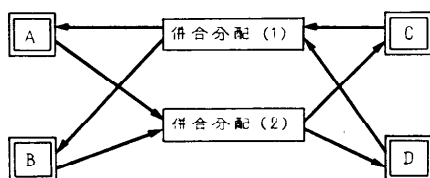
出力テープCに書いた項目のキーCと、A, B から読んだ項目のキー B, C を比較すると次の6通りになる。その比較から次の出力項目が決定される。

	大小関係	次の出力
(i)	$A > C > B$	B 連系の切れ目
(ii)	$C > B \geq A$	B
(iii)	$B \geq A > C$	B
(iv)	$A > B \geq C$	C 連系の切れ目
(v)	$B \geq C \geq A$	C
(vi)	$C \geq A > B$	C

このように各項目の比較だけで次の出力項目が決定され、項目を逐一処理できることが磁気テープのような逐次式ファイルに適しているのである。

(b) テープ装置4台の場合

テープ装置4台を用いれば、上に述べた併合と分配を同時に行なうことができる。すなわち、最初のテープをAにかけ、C, D に分配する。次に C, D を併合した最初の連系をAに入れる。C, D から生じた第2の連系をBに入れる。このように C, D を併合してできた連系を A, B に分配する。これがすめば次は A, B を併合し、その結果できた連系を C, D に分配する(第3図)。この方式で、前のものに比べ、情報の転送時間を約1/2に、巻き戻し時間を(連系の長さが一様であるとすると)約1/4に減らすことができる。



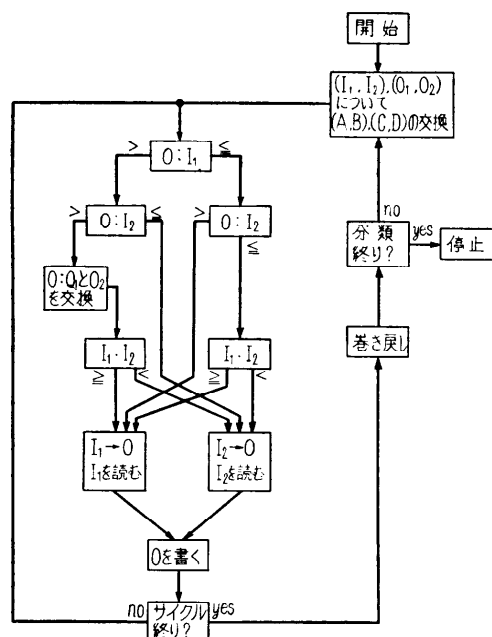
第3図

出力テープの切換えは連系の切れる点であり、これは前のキーの比較表の(i)と(iv)の場合である。それ故この場合の手順の流れ図は第4図のようになろう。テープ切換えの所は略記してある。

テープの逆読みが可能なときは、巻き戻しながら項目を読み、併合分配ができて、巻き戻し時間の無駄がなくなるのは、基数法の場合と同様である¹⁷⁾。

D-3 内部分類との関連

内部分類は、項目が内部記憶装置に納まる限りでは



第4図

高速であり、プログラムも容易である。この内部分類を利用することによって、磁気テープによる分類のテープ通読(pass)の回数を減らすことが可能である。

併合法の最初の分配パスの時、できるだけ多くの項目を主記憶装置に読み込み、それを内部分類し(これによってそれらの項目は一つの連系を形成する)、それを出力テープに分配する。前に述べたように、項目が一様確率分布していれば連系の数は平均 $N/2$ であり必要なパスの回数は $n = \lceil \log_2 N/2 \rceil$ であった。もし F 項目を内部分類しておくならば、連系の数は、 N/F となり、パスの回数は、

$$n' = \lceil \log_2 N/F \rceil$$

である。その差

$$\log_2 F/2 \quad [\text{回}]$$

のパスが節約できることになる。すなわち16項目ずつを内部分類するとして約3回の節約になる。

内部分類の効果は、しかし最初の1回だけである。というのは、それ以上連系が長くなると、二つの連系がまるまる主記憶装置に入るとことはありえないわけで、内部分類によってさらに長い連系を作るとことはあまり望めない*。

* 連系と連系の境目だけに乱れがあるときならば、それらがつながるということも考えられるが確率は小さい。

また、はじめから連系の長いデータ（すでに部分的な分類がなされている）の場合にも、内部分類の効果はうすい。

D-4 グルーピング

磁気テープの問題点の一つは、テープの起動停止の時間である。1項目を読んで止まるということになっているとすれば、テープ1本に10万項目入っていると、テープの起動停止時間を5msと仮定すると、1回のパスで8分以上がそのために使われることになる。停止なしに通読すると約6分ぐらいであるからこれは無視できない。磁気テープ読み書きの単位は普通ブロックまたは記録(block or record)と呼ばれるが、そのブロック内になるだけ多くの項目をまとめること、すなわちグループにまとめること(グルーピング-grouping)によってテープの起動停止を減らすことが考えられる。ブロックが固定長のときは、その中にできるだけ多くの項目を、可変長のときはできるだけ長いブロックに項目をまとめるのである。

D-5 p 重併合法による分類と p の選択

D-2節で説明したのは2重併合法であった。一般に p 個の連系から1個の連系を形成する p 重併合を使うのももちろん可能である。 p 重併合でのパスの回数は、連系の数を S として

$$n = \lceil \log_p S \rceil$$

$p=2$ が2重併合であるが、それと比べると一般の p 重併合では $1/\log_2 p$ の比率でパスの回数が減る。 p を大にすれば、それだけパスの回数が減るので、 p は大きいほどよいように見えるが、一概にそうもいえない事情がある。

分類法の良し悪しは結局分類時間によって決まる。磁気テープ分類で特に無駄な時間とみなされるのは、巻き戻し時間と起動停止時間である。

併合のベース p を大きくすると悪い方に働くのは、前節の内部分類とグルーピングである。内部分類についていえば、その効果は

$$\log_p F/2$$

で表わされるから p が大きくなるほど効果はうすれる。

グルーピングについては、 p ブロックを主記憶装置内で取扱うので、 p が大きくなると1ブロック当りの割当が少なくなる。すなわちブロックを小さくしなければならぬので、結局ブロック数がふえ、そのため起動停止時間が多くなる。

このように分類時間について、併合のベース、内部分類の規模、グルーピングの大きさがお互に関連し合ってくる。それらの最適値は装置および応用の具体的な数値によって決まるものである*。

p の選択についての議論にはその他いろいろな立場がある。一つの計算は、経済的観点からするもので、時間をパスの回数に比例するとし、経費をテープ台数に比例すると考えるものである^{14),16)}。これによると2重または3重の併合法が最もよいということになる。しかしこの計算は計算機本体の価格や、その利用率を考慮していないからあまり妥当ではない。

また、情報の転送時間と計算機内での処理時間の和の最小値を求めようという計算もある¹⁸⁾。この結論は、事実上テープは多いほどよいということになる。これは分類において取扱うデータの量は龐大だが、実質的な処理は比較的少ないという事実を反映している。

D-6 時分割方式と分類

磁気テープによるオンライン分類になげかけられる疑問の一つに、計算機そのものがあまり有効に使われていない、すなわち、分類に必要な磁気テープの読み書き時間と、計算機の実質的な活動時間がはなはだしく不均衡であり、計算機がもったいないということがある。それは事実で、分類などはオフラインで行ないたいということにもなるのである。このような外部装置と計算機との速度のアンバランスは、データ量が多く計算が比較的単純な事務データ処理でのオフライン機械の存在理由になる。

最近実用化されてきた時分割による入出力機械の同時制御、並列プログラム方式は、この辺の事情に大きな影響を与えるであろう。すなわち、紙テープ $\longleftrightarrow$ 磁気テープの変換や、タイプライタ $\longleftrightarrow$ 磁気テープの通信、それに分類などの操作は、計算機が必要のつどそこに立ち戻って働き、その他の（今までは待ち時間であった）ところでは何かルーチンの計算をしているということになる。これによって、たとえば、分類時間が特に短くなるということ（磁気テープ数台が同時に読み書きできるようになったりすることで短くなることを除けば）はないであろうが、計算機の遊び時間をなくすことによって、分類などのコストが下ることになるだろう**。

* この計算については〔3〕の附録参照。

** このような考えは比較的以前から一部の人たちによって主張されてきた^{19),20)}。時分割方式の計算機が普及するにつれて一般に受け入れられるようになるだろう。

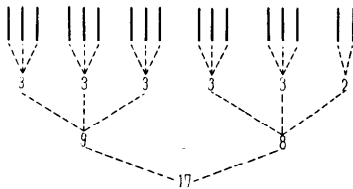
D-7 多重リールの処理³⁾

ファイルが大きくなって磁気テープ一巻に入りきれなくなると、今まで述べたような一貫作業は不可能になる。分類を完結するためには人手の介入が必要となり、テープ・リールを何回も取扱わなければならないことになる。

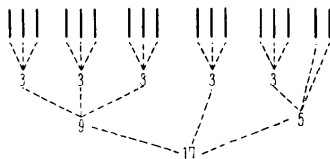
(a) 多重リールの併合

第5図のような17巻のテープからなるファイルを考える。各巻はそれぞれの中では分類が済んでいるものとする。各巻を併合していくわけであるが、3本のテープを併合すると分類された3本のテープができ、分類された3本のテープ、3組を併合すると分類された9本のテープができるというように進むのである。

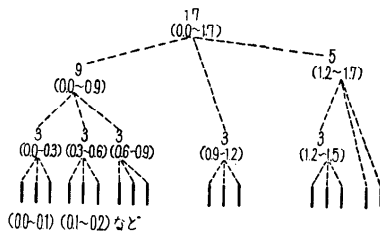
第5図は一案であるが第6図のようにすると併合の回数が減る。



第 5 図



第 6 図



第 7 図

(b) 多重リールのブロック法。

いまの併合法とは逆に進むものである。表7図で、キーが、 $0.0 \times 10^6 \sim 1.7 \times 10^6$ の範囲で一様に分布しているとする。それを図のように $0.0 \sim 0.9$ のものを9本のテープに、 $0.9 \sim 1.2$ のものを3本のテープに、

$1.2 \sim 1.7$ のものを5本のテープにというようにふり分けていく。最後にできた各一巻のテープはそれぞれ、それだけで分類する。この場合キーの分布があらかじめわかっていなくてはならない。

E. ランダム・アクセス・ファイルでの分類

磁気ディスク型などの大容量ランダム・アクセス・ファイル (random access file) では、ファイル内のどの項目も大体似たような時間で呼び出せる。項目はアドレスで指定することができる。容量と速度の点では主記憶装置と極端に違うが、性格が似ているので、内部分類法のところで述べた方法を使うことができる。アドレス計算法などは大容量ランダム・アクセス・ファイルに適するであろう。

E-1 分離されたキー

ランダム・アクセス・ファイルで使われる一つの手法として分離されたキー (detached key) の説明をしておこう。

ランダム・アクセス・ファイルでは磁気テープなどのような逐次ファイルと異って、分類のさい、ファイルの項目全体を動かさなければならないということはない。項目はアドレスを指定することにより、いつでも呼び出せるということから、項目のキーと項目の場所を指すインデックスとから非常に短い記録を作る。分類をしたいときはこの記録を分類しておけばよいのである。項目に必要なときは、キーからインデックスを調べ、それによって求める項目を呼び出す。

この分離されたキーは短いから、内部分類でもかなりの数の分類ができる。その上で、この大容量ランダム・アクセス記憶装置の片すみで、B章で述べた基本操作のどれかによって分類を完結すればよい。

F. む す び

以上でデータ処理における分類手法の解説を終るが、筆者の所でも実際の経験がまだあまり蓄積されておらず、具体的な資料を盛り込めなかったのは残念である。今後、各所で磁気テープによる分類の経験が積み重ねられて、それによってもっとつこんだ議論がなされることを希望したい。

最後に分類の重要性を示唆され、この解説の執筆にあたって種々助言をたまわった当所、高橋茂課長、西野博二主任に深く感謝の意を表したい。

参考文献

- 3) E.H. Friend, Sorting on Electronic Computer Systems, J.A.C.M. (July 1956)
- 10) A.S. Douglas, Techniques for Recording of and Reference to Data in a Computer, Computer Journal (Apr. 1959)
- 14) 飯島泰蔵: テープによるならべかえの理論, 電試集報 (1960年4月)
- 16) D.A. Bell, The Principles of Sorting, Computer Journal (July 1958)
- 17) H. Nagler, Amphisbaenic Sorting, J.A.C.M (Oct. 1959)
- 18) P.F. Windley, The Influence of Storage Access Time on Merging Processes in a Computer, Computer Journal (July 1959)
- 19) 淵, 西野: 入出力計算機 ETL MK 4B の方式, 情報処理, Vol. 1 No. 1 (1960)
- 20) 石井善昭: 時分割演算による計算機に関する研究, 日本電気株式会社 (1961年2月)

(昭和36年6月6日受付)