情報処理学会第37回(昭和63年後期)全国大会

6X-5

# 書誌情報ファイルの処理に対する効率評価

渡辺 豊英　駱 琴　吉田 雄二　稲垣 康善　柘植 利之
(名古屋大学 工学部)　(名古屋大学附属図書館)

## 1. はじめに

計算機システム技術の発展とともに、大量データを格納するディスク装置や、膨大なデータを効率良く管理し、迅速に検索するデータベース管理システムが開発され、種々のデータ・リソースを構築・利用できる環境が着実に整備されてきた。磁気ディスク装置の高密度化に加えて、光ディスク装置の出現は画像データを安価に記録することを可能にした。過去何十年にも渡って蓄積された文書類を短時間に、簡易な作業によって計算機ファイルに格納し、効率良く利用できる手段を提供する。

書誌情報も図書目録カードとして記録・保管・利用されてきたが、利用方法が限定され、保管に大きな場所を占め、管理が大変であるなどの問題があり、計算機ファイルとして蓄積することが重要な課題となっている。大量の書誌情報を直接端末などからコード・データとして作成することは、膨大な作業量・経費を必要とし、効率的なデータ作成方法が求められている。一方、近年発展が著しい画像入力装置により画像データとして作成することは、データ入力の作業量・経費を軽減するが、ファイル容量を膨大にし、また計算機で直接処理するには有効でなく、処理効率もあまり良くない。

本稿では、書誌情報をコード・データ、画像データとして計算機ファイルに格納する場合の効果を議論する。特に、近年利用可能になった追記型光ディスク装置と、今日計算機ファイルとして代表的な磁気ディスク装置に対してそれぞれコード・データ、画像データとして格納した場合の処理効率を算定する。

## 2. 書誌情報のファイル化

大量の書誌情報を計算機ファイルに格納して利用するには、(1) データ作成、(2) データ格納、(3) データ処理の視点から検討する必要がある。データ形態としてコード・データ(キーボードから入力されるバイト・データ)、画像データ(画像入力装置から読み込まれるデジタル画素データ)があり、どちらの形態がよいかを各視点で評価しなければならない。それに対応して、光ディスク装置や磁気ディスク装置などを格納媒体として検討する必要がある。光ディスク装置は現在コード・データを記録できる品質を保証していないので、画像データ用の格納媒体と見做される。もちろん、画像データを磁気ディスク装置に格納できるが、一情報当たりの価格が高く、得策とは言えない。表1に、追記型光ディスク装置と磁気ディスク装置の特性を整理した。光ディスク装置はアクセス時間で5倍以上、データ転送速度で4倍程劣るものの、価格は$1/10^4$である。

同一内容の書誌情報をコード・データと、画像データで格納した場合、これらのデータ量はかなり異なる。書誌情報をコード・データ、画像データとして和書・洋書それぞれ1000件を計算機ファイルに格納した場合に、画像データはコード・データに対して和書・洋書ともに概略500倍のデータ量であり、画像データの格納には大容量のファイルが必要になる[2)]。しかし、

表1 磁気ディスク装置と光ディスク装置

| | 磁気ディスク装置 | 光ディスク装置 |
|---|---|---|
| | IBM-3340 | Philips-DOR |
| ファイル容量 | 70 Mバイト (5Gバイト/4) | 1 Gバイト (2.7Gバイト) |
| アクセス時間 | 35ms (25.0ms) | 100-225ms (250ms) |
| データ転送速度 | 7Mビット/s (24Mビット/s) | 2Mビット/s(6.3Mビット/s) |
| 価格(セント/ビット) | $10^{-4}$ | $10^{-8}$ |

注意) 各値は参考文献1)より引用した。ただし、括弧内の値は富士通社の代表的な製品の性能である。

画像データの格納に光ディスク装置を、コード・データの格納に磁気ディスク装置を用いれば、経済的に充分対処できる。光ディスク装置では価格的に20倍の画像データを収容可能である。

書誌情報の計算機ファイル格納は、単に価格比だけでその優劣を議論できない。データ処理の視点では、データ量が少ない程良く、また処理し易い形態で構成されている必要がある。データ量が少ないことは、CPU(主記憶)と計算機ファイル(補助記憶)間の転送量を減少させ、処理効率を向上させる。また、コード・データは画像データよりもはるかに処理し易く、の処理時間を軽減し、かつ様々な処理要求にも対処可能である。

## 3. 処理効率の算定

書誌情報を計算機ファイルとして構築し、管理・利用する場合に、コード・データと画像データで扱うことの利点、欠点を評価しなければならない。その視点は、データの作成効率、データの格納効率、データの処理効率である。

(1) データの作成効率

画像データは図書目録カード原寸大をイメージ・スキャナから二値画像1168*704画素として作成され、コード・データは端末から入力されると仮定する。画像データの入力は図書目録カードの設定などの人間の操作を除けばほとんど機械的に処理され、作業時間に比例する。一方、コード・データの作成は人間の操作に依存するために作業時間に比例しない。すなわち、コード・データでは単位時間当たりの入力文字数に対して、人間の生理的な反応を考慮しなければならない。

たとえば、一画像データの作成時間を人間の操作時間も含めて2分とすれば、和書・洋書それぞれ1000件の図書目録カードでは4000分を要する。一方、コード・データでは1分間に洋書で60字、和書で20字を入力できるとすれば、一書誌情報の作成に3-4分を要する。それに、人間の生理的な反応を考慮して6分とすると、和書・洋書それぞれ1000件の書誌情報では12000分となる。コード・データの場合には、画像デ

---

A Study of Management Efficiency for Library Catalog Files

Toyohide WATANABE[1], Qin LUO[1], Yuuji YOSHIDA[1], Yasuyoshi INAGAKI[1] and Toshiyuki TSUGE[2]
1: Faculty of Engineering, Nagoya University, 2: University Library, Nagoya University

ータに比べて3倍以上のデータ作成時間が必要である。

(2) データの格納効率[2)]

コード・データと画像データでは500倍以上のデータ量の相違があった。ファイル容量の軽減にはデータ圧縮技法の適用が有効である。参考文献2)は画像データの圧縮比を13倍以上、コード・データの圧縮比を3.5-4.8倍と報告している。従って、データ圧縮技法を適用すれば、画像データはコード・データに対して約200倍のデータ量になる。データ圧縮技法はそれぞれのデータ形態に対して格納効率を向上させるが、画像データとコード・データのデータ量は大きく違う。

(3) データの処理効率

データの処理効率は書誌情報の利用に際して重要である。それには、処理要求に対する効率的な対処(CPU処理)、主記憶と計算機ファイル間のデータ転送(I/O処理)が問題である。前者はデータの構成・構造を計算機が特定し易く、処理し易いことを、後者は入出力のデータ量が少ないことを要求する。画像データはコード・データに対して何れも劣っている。

ここでは、主記憶と計算機ファイル間のデータ転送時間を算定する。データ処理に付帯する入出力動作は、CPU時間よりも全体の処理時間に占める割合が大きいためである。表2のようにパラメタを設定すると、

$n*a+(\alpha/\omega)/b$

としてデータ転送時間の算定式を導出できる。パラメタに表1の値や参考文献2)で示す値を設定して計算する。ただし、n=1とする(磁気ディスク装置ではレコード・サイズ、装置の種類で異なり、光ディスク装置では常に1である)。

i) データ圧縮技法を適用していないデータ・リソース:

・コード・データ(磁気ディスク装置に格納)

0.035+(188410*8)/7000000=0.250 ---------- 和書
0.035+(199754*8)/7000000=0.263 ---------- 洋書

・画像データ(光ディスク装置に格納)

0.100+(10278400*8)/2000000=411.236 ------- a=0.100
0.225+(10278400*8)/2000000=411.361 ------- a=0.225

ii) データ圧縮技法を適用したデータ・リソース:

・コード・データ(磁気ディスク装置に格納)

0.035+(188410*8/4.82)/7000000=0.079 ------- 和書
0.035+(199754*8/3.51)/7000000=0.100 ------- 洋書

・画像データ(光ディスク装置に格納):和書に対して

0.100+(10278400*8/13.17)/2000000=31.317 -- a=0.100
0.225+(10278400*8/13.17)/2000000=31.442 -- a=0.225

これらの値を基に画像データとコード・データの転送時間の割合を整理したのが、表3である。コード・データではデータ圧縮技法を適用すれば約1/3の転送時間で、画像データでは約1/13の転送時間で済む。一方、画像データはコード・データに比べてデータ圧縮技法を適用しなければ1560-1640倍の転送時間が必要となる。

表2 データ転送時間の決定パラメタ

| 記号 | 意味 |
|---|---|
| n | アクセス回数 (格納媒体) |
| a | アクセス時間 (格納媒体) |
| b | データ転送速度 (格納媒体) |
| $\alpha$ | 転送データの量 (データ形態) |
| $\omega$ | 圧縮比 (データ形態) |

表3 データ転送時間の割合

| | 和書 | 洋書 |
|---|---|---|
| コード・データ | 3.16 | 2.63 |
| 画像データ | 13.08-13.13 | --- |

(a)コード・データ、画像データにおけるデータ圧縮技法の適用による転送時間の割合

| | 和書 | 洋書 |
|---|---|---|
| 圧縮技法非適用 | 1641.9-1645.4 | 1563.6-1564.1 |
| 圧縮技法適用 | 396.4- 398.0 | ------- |

(b)コード・データの転送時間に対する画像データの転送時間の割合

## 4. 考察

大量データの扱いを議論するとき、データ入力、データ格納、データ処理の視点で検討しなければならない。第3節の検討から、データ入力時には画像データの形態が好ましく、データ格納とデータ処理時にはコード・データの形態がはるかに効率的である。従って、データの作成には画像データとして、データの格納・処理にはコード・データとして扱い得るアプローチが必要でである。しかし、画像データとしての書誌情報をコード・データの文字列に変換することは非常に難しい。

一方、光ディスク・ファイルと磁気ディスク・ファイルのビット当たりの価格は光ディスク・ファイルが極端に安く、データ圧縮技法を適用しても画像データではコード・データの100倍のデータを同一価格で収容できる。もちろん、画像データだけを格納しても個々の画像やその内容を探索・同定することが困難であり、手掛かりとしてコード・データを付帯させなければならない。このコード・データは磁気ディスク・ファイルに確保される。図書目録カードの場合、この付帯データを著者・書名などとして、約3割のデータ量と見積もると、画像データ入力時の作業量は増大する。このデータ量はデータの格納・処理に対してほとんど無視できるが、データの作成に対して3600分を要し、画像データ入力時に必要な作成時間は合計7600分となる。コード・データとして扱う場合に比べてそれ程差がなくなり、書誌情報では画像データとして扱うことにあまり意味がない。

## 5. おわりに

本稿では、書誌情報をコード・データ、画像データとして扱うことに対してデータ作成、データ格納、データ処理の視点より検討した。もちろん、データ整理や付随する人的な処理を無視したが、データ作成を除けば、現時点で画像データの扱いにそれ程効果がない。効率的に利用・運用するにはコード・データとして格納することが必要不可欠である。

効率的な書誌情報処理システムは画像処理・認識技術の下に入力を画像データで、格納・処理をコード・データで扱う必要があり、これは大きな課題である。

参考文献

1) S.Christodulakis: "Issues in the Architecture of a Document Archiver using Optical Technology", Proc.of SIGMOD'85, pp.34-50.
2) 駱他: "書誌情報のファイル格納に関する効率評価", 情報処理学会第37回(昭和63年後期)全国大会講演論文集.