図書目録イメージデータの検索システム

1Q－4

松本 徹也　栗田 英和　柴田 裕介　竹田 正幸　有川 節夫
九州大学大学院システム情報科学研究科

1. はじめに

現在，全国の国立大学図書館には約2億冊の蔵書があるといわれている．このうち，比較的新しいものについてはOPACなどの蔵書検索システムにより検索可能である．しかし，古い蔵書に関しては，経費と人手の問題があり，データ入力作業がはかばかしくなく，その一部しか検索できない．このため，図書目録カードを調べるためだけに，図書館に足を運ばねばならない状況が続いている．この状況を打破するために，著者らは図書目録カードをイメージデータ化し，そのイメージデータを対象とした蔵書検索システムをWWW上に構築した．現在のところ，九州大学理学部と教育学部所蔵の約17万冊分の図書目録カードに対して検索が可能である．

2. 図書目録カードの電子化

蔵書検索システムを構築する際に，文字認識処理によってテキストデータ化する方法が考えられる．しかし，図書目録カードは，手書きのものが多く，タイプライタやワープロによる印刷のものであってもそのフォントも様々であり，また汚れや文字のかすれなどが目立つものもあるなどの問題があり，認識精度は極めて低い．また，たとえ文字認識に成功したとしても，認識された個々の文字列が書名，著者名の書誌的項目のうちいずれであるのかを判断しなければならないという問題がある．実際の図書目録カードは，必ずしも一定の書式に従っていないために，この判断は非常に困難である．従って，文字認識処理技術を用いたとしても，結局は人手の介在を必要とし，入力作業の完全な自動化は望めない．

以上の理由から，古い蔵書の書誌的情報の機械可読化作業は，人手により進められているが，この作業は多大な労力を要する．このような入力作業を支援するシステムとして，学術情報センターによるCATシステムがある．CATシステムでは，全国の大学図書館と総合目録データベースをネットワークで接続し，全国の大学図書館の共同分担方式によって総合目録データベースを作成する．入力したデータは直ちにデータベースに登録されるため，目録作成の重複を防ぐことができ，目録作成業務の負担を軽減できる．しかし，他大学によって未入力のデータや，他大学の所蔵しない図書については，新たに入力を行う必要がある．九州大学附属図書館の場合，遡及入力すべきデータは約150万冊分にも上るが，経費や人手の問題から最大でも年間約6万冊分が限度である．このペースで進めていくとすべての入力作業を終えるのに約25年もの歳月を必要とし，このことが電子図書館化を推進するための大きな障壁となっている．

そこで著者らは，これらの図書目録カードをテキストデータ化するのではなく，高速イメージスキャナを用いてイメージデータ化し，これを対象とした蔵書検索システムをWWW上に構築した．実際の図書目録カードは，書名や著者名の辞書式順序で整列され格納されている．そこで，この順序の情報を保持したまま，これらのカードをイメージスキャナを用いて取り込んだ．また，引き出しのラベルや引き出し内の仕切りのラベル等の情報も取り込んだ．人手によるキーワードの付与等は行ってないが，利用者が図書館で目録カードを捲って検索する場合と同じことが可能である．

この手法によって，経費や人手を大幅に削減した蔵書検索システムが可能となり，電子図書館実現のために大きく寄与するものと考えられる．

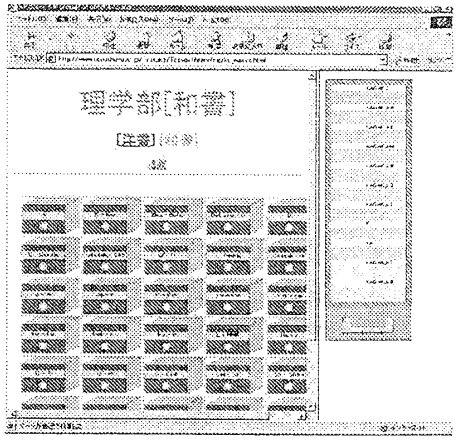
3. イメージデータのための検索システム

システムはWWW上に構築した．利用者は以下の手順で検索を行う．

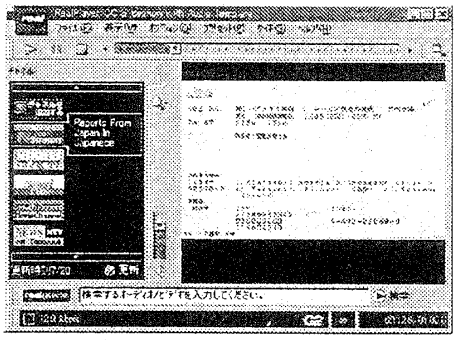
1) 図1(a)の左フレームは実際の図書目録カードが格納されている引き出しと同様に配列している．利用者は，各引き出しのラベルを目安にして，探したいカードのある引き出しをクリックする
2) 図1(a)の右フレームに，引き出しの内部が表示される．利用者は，しきりのラベルを目安として目的のカードの位置を判断し，その位置をクリックする．

A retreival system for image data of book catalog in libraries

Tetsuya Matsumoto, Hidekazu Kurita, Yusuke Shibata, Masayuki Takeda, and Setsuo Arikawa
Department of Informatics, Kyushu University, Fukuoka, 812-8581 Japan



(a)



(b)

図 1 図書目録検索システム

3) サーバ側は，クリックされた位置のカード及びその前後の数十枚分のカードイメージを表示するためのSMIL(Synchronized Multimedia Integration Language)文書を動的に作成する．クライアント側では，図(b)のようにリアルプレーヤーが起動し，カードが数秒おきに順次表示されていく．利用者は，リアルプレーヤーの機能を用いることにより，インタラクティブに目的のカードを探すことができる．

4. 文字列検索

現在のところ3節で述べたような方法での検索が可能であるが，書名や著者名などの文字列による検索も利用者の要望があり現在開発中である．そこで，文字画像の図形的特徴量を算出し，それと値の近い特徴量をもった文字画像を含むカードを検索結果として返す方法を考えている．図形的特徴量の抽出には，北海道大学の田中ら[1)]によって開発されたトランスメディア技術を用いることを検討している．

トランスメディア技術とは，文字領域の切り出しと特徴量の抽出・コード作成の2つの行程からなる．それら2つの作業が画像文書を入力する際の前処理として行われ，こうして生成された各情報を各画像文書ごとに補助データとして付加し，画像と共に保存する．そして，この画像文書の補助データ中のコードとキーワードのコードを比較することにより文字列検索を実現する．

ただし，このトランスメディア技術を図書目録カードに適用した場合，大きく分けて二つの問題点がある．

まず一つ目の問題は，図書目録カードの紙質である．紙質が古かったり，汚れていたりすると，イメージスキャナから画像を取り込む際にノイズが混入してしまい，文字の切り出しと特徴量の抽出が正確にできなくなる．もう一つの問題は，カードに印刷されている文字の印字品質である．カードは手書きのものも多く，文字がかすんでいたり，歪んでいるものが少なくなく，特徴量の計算方法から考えると，検索性能に影響がでると考えられる．

現在は，タイプライタで印字された洋書カードに限って研究を行っている．

5. まとめ

本稿では，イメージデータ化された図書目録カードを対象とする蔵書検索システムについて述べた．このシステムにより，目録カード検索のためだけに図書館に足を運ばなければならないという問題は解消された．

すべての目録カードをテキストデータ化する作業は今後も継続して行うため，本稿で述べたようなシステムは過度的なものである．しかし，2節で述べたように，人手と経費の問題から，この作業には数十年を要する．したがって，本手法はテキストデータ化の作業が完了するまでの長きにわたって有効であると考えられる．

また，図書目録カードは図書館から持ち出すことが難しいが，イメージデータ化したことによって，図書館外でもテキスト入力作業を行うことができるなど，入力作業の効率化が期待できる．現在，そのための入力支援システムの開発を進めている．

参考文献

1) Y. Tanaka, K. Takahashi, and M. Mozaffari. Transmedia machine. *J. Inf. Process.*, 12(2):139–146, 1989.