人文科学とコンピュータ　43－1
（1999．7．16）

イメージデータ化された図書目録カードの検索システム

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

要旨．現在，全国の大学図書館には約2億冊の蔵書があるといわれている．このうち，図書館電算化以後に収集された図書に関しては，OPACなどにより検索可能である．しかし，電算化以前の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．著者らは，このような状況を打破すべく，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検索システムを開発した．

A Retrieval System for Image Data of Book Catalog in Libraries

Hidekazu Kurita　Yusuke Shibata　Masayuki Takeda　Setsuo Arikawa
† Department of Informatics, Kyushu University

Abstract. There are about two hundred million books in the university libiraries in Japan. Retrieval systems, such as OPAC, deal with only relatively new books of which catalog is digitized. Catalog of old books remain not digitized. These are retrieved only by using a book catalog in the libraries. Therefore the users must go to libraries just for this purpose. We digitized the image of book catalog in the library of Kyushu University, and then developed a new retrieval system for the image data. The proposed system makes it possible to retrieve them without going to the library.

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．

この入力作業のために，文字認識処理によって図書目録カードを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システムの特徴としては，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することで，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ことにより，入力したデータは直ちにデータベースに登録され目録作成の重複を防ぐことができる．これにより，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，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，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，いったんイメージデータとして取り込むことを考えた．図書目録カードは図書館から持ち出すことが難しいが，このイメージデータ化により入力作業が図書館以外の場所で可能となるなど，入力作業の効率化が期待できる．

著者らは，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を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[2]．

このシステムは，書名順あるいは著者名順に整列された図書目録カードの中から，実際に図書館に行ってカードを探す場合と同じように検索できる．この場合，検索結果がテキストデータではなくイメージデータであるため，データサイズをできる限り小さくする必要がある．データサイズの大きなカードは，カードの汚れなどによるノイズを多く含んでいる場合が多い．そこで，圧縮効率のよいデータ圧縮法とともに，ノイズ除去法を開発する必要が生じる．

また，もう一つの検索方法として，署名や著者名などの文字列による検索が考えられる．検索対象となるデータがイメージデータであるため，著者ら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システムで用いられている技術を利用し，文字列検索システムを開発している．トランスメディアとは，文書画像中の文字の図形的特徴量に基づくものである．この技術を利用することにより，利用者が書名や著者名などの文字列を入力すると，システムはその図形的特徴量を算出し，その値と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．

本稿で提案する手法によって，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．

2 イメージデータのための検索システム

多くの利用者の要望に応えるため，実際に図書館に行って図書目録カードを探す場合と同じように，WWW上から検索できるシステムを開発した．

2.1 目録カード検索

図書目録カードをイメージスキャナに取り込む際に，カードを書名の辞書式順序で整列されたままイメージデータ化することにより，検索できるシステムを開発した．

ここで，目録カード検索のインタフェイスとして，次の2種類用意した．

- 引き出し&しきり型
  今まで図書館に自分で足を運んで，図書目録カードを頻繁に探していた利用者の要望によるもので，図書館にある引き出しとその中のしきりを全く同じ位置に配置することにより実現した．

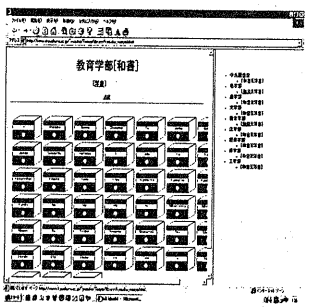
  以下の三つの行程で行われる．

  1. 探したい書名のある引き出しをクリックする．(図1(a))

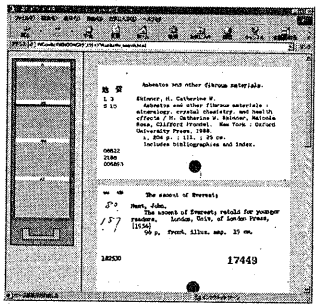
2. 引き出しの中から，しきりのカードを目安として，探したい書名の任意の位置をクリックする．(図1(b)の左フレーム)
3. 表示された数枚のカードの中から，目的のカードを見つけだす．(図1(b)の右フレーム)

・スライダー型

目録カード全体を辞書式順序に整列し，それを画面に表示された直線上に配置したものと考える．利用者が直線上の任意の位置にスライダーを移動させると，システムはその位置に対応したイメージデータをその前後も含め表示する．(図2)



(a) トップページ



(b) 次ページ

図 1: 引き出し&しきり型のインタフェイス

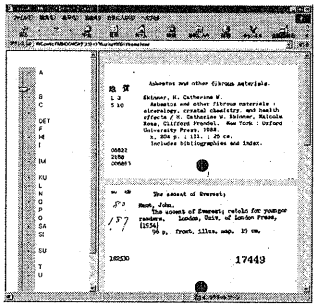


図 2: スライダー型のインタフェイス

3 文字列検索

2節で述べたようなシステムを開発したが，書名や著者名などの文字列による検索も多くの利用者の要望があることから文字列を入力とした文字列検索システムの開発も行っている．

その方法として，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．

このような方法には，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利用している．まず，3.1節でトランスメディアについて述べ，3.2節で図書目録カードへ適用した場合の問題点について述べる．

3.1 トランスメディア技術

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・コード生成の二つの行程からなり，それらのアルゴリズムを以下に示す．

3.1.1 文字画像の切り出し

文字画像の切り出しとは，まず図3のように画像文書中の画素を水平方向及び垂直方向に射影し，黒画素の存在しない部分を空白と見なし文字画像を切り出す．図書目録カードは横書きであるため，横書きの場合について説明する．横書き文書の場合，文字が横に連なる「行」が存在し，その行が縦に並ぶことにより文書が構成される．文書領域の黒画素を水平軸に対して

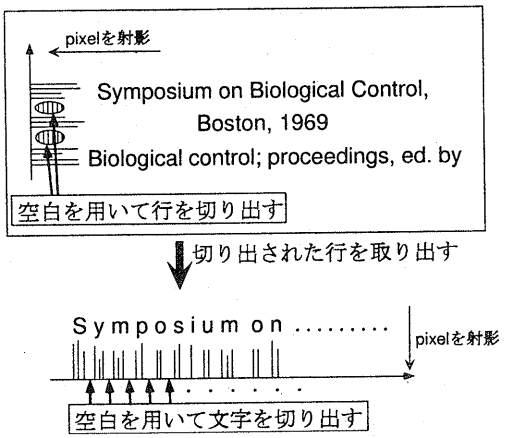


図 3: 文字の切り出し

射影した場合，そのヒストグラムのグラフは，行の領域に対応した山が，空白に囲まれている行数と同じ数だけ並んでいる形になる．この空白で囲まれた黒画素領域の両端の座標を調べることにより行の上下の座標が決定される．次に，各々の行画像の領域に注目し，その黒画素を水平軸に射影した場合，そのヒストグラムは文字の領域に対応した山が，空白に囲まれて文字数と同じ数だけ並ぶグラフとなる．この空白で囲まれた黒画素領域の両端の座標を調べることにより一つ一つの文字画像を切り出していく．

3.1.2 図形的特徴量の抽出方法

次に，文字の切り出しにより切り出された文字の図形的な特徴から特徴量を抽出し，コードを生成する．

切り出された文字画像を囲う矩形領域を文字領域と呼ぶ．この文字領域を複数に分割し，各々の分割領域についてその領域面積に対する黒画素の密度を計算する．図 4 のように，これらの密度のうちの 2 つの比を取ったものを 1 つの特徴量とする．そして，各々の特徴量について多数の文字に対する統計を取ることにより得られる密度比の分布をグラフにすると，図 5 のようなグラフが得られる．このグラフの縦軸は各々の密度比における文字の出現頻度，つまり文字数を表しているので，グラフの全面積が統計調査時に入力した総文字数となる．このグラフを複数の領域に分割する．分割を行う際，各々の分割領域に含まれる文字数は全て同じになるようにする．そして，各々の文字についてその文字中の注目している特徴量から得られた密度比がグラフ上のどの領域に含まれるかによってコードを生成する．

日本語文書画像に対しては，図 6 に示されている分割領域のうち互いに隣り合っていて，組み合わせると正方形を形作るような 2 つの分割領域の黒画素の密度間で比を取り 1 特徴量としている．また，図 5 のように 4 領域に分け，2bit のコードを生成することにより，1 特徴量から 2bit のコードを得ている．つまり日本語に対しては，18 特徴量× 2bit で 1 文字あたり 36bit のコードを生成することにより文字列検索を行っている．

ノイズの影響により同じ形の文字から生成されるコードが異なることがある．そこで 1 文字あたりの情報のうち，ある bit 数までは違っていてもそれはノイズの影響と見なし，残りのコードでマッチングを行っている．

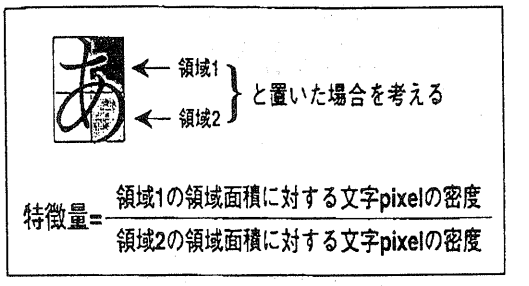


図 4: 特徴量の計算

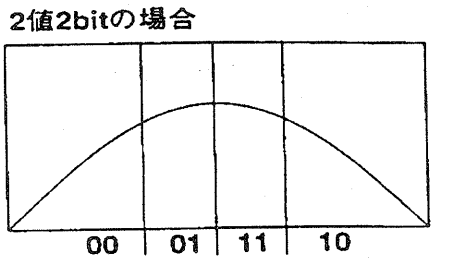


図 5: ヒストグラムによるコーディング

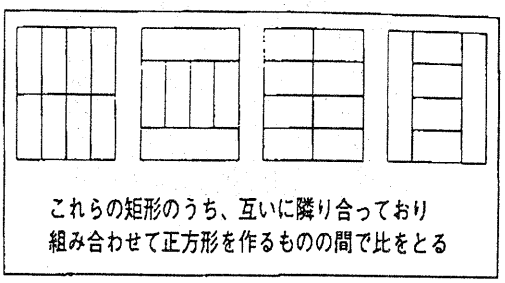


図 6: 特徴量を計算する領域

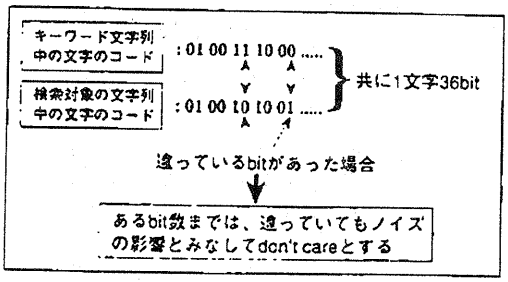


図 7: 曖昧検索

3.2 図書目録カードに適用した場合の問題点

- 印刷されている紙自体に問題がある場合
  純白でない紙に印刷されている文書の場合，イメージスキャナを用いて入力された際，画像文書の背景に多くの黒画素が含まれることになる．

- 印字品質の問題
  これは印刷されている文字の印字品質に問題がある場合，つまり文字中のシミ，かすれ等による文字の変形，傾き，歪み等があった場合である．

このようにノイズが含まれる場合，黒画素数を文字列検索の情報として用いているので，検索性能に影響がでる．

また，文字のフォントが異なる場合や手書きの文字の場合も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．

4 ノイズ除去

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，3.2 節で述べた問題を解決しなければならない．そこで，前処理としてノイズ除去を行うことが必要となる．

図書目録カードは，古いものや汚れのついているものなど，イメージスキャナから取り込んだ際に，カードに書かれている文字とは無関係なデータ，つまりノイズが混入する．このため，システムを構築する際に二つの問題が生じる．一つは，ノイズが混入することにより，データ量が増加し，WWW 上から検索する際の通信コストが余計にかかってしまうという点である．二つ目は，トランスメディア技術は，3.2 節で述べたように切り出し，特徴量の抽出の際にノイズに弱いという点である．そこで，前処理としてノイズの除去が必要となる．

ノイズは普通，周辺の画素とは関係なく，孤立的に存在するという性質を持っている．これらの性質を用いてノイズを除去する各種の方法が提案されている．そこで，いくつかの方法 [3] を試してみた．

選択的局所平滑化 着目画素の周囲を複数個の局所領域に分割し，それぞれの領域における画素値の分散を計算し，分散が最小の領域の平均画素値を着目画素の値に変更する方法である．この方法は，平坦部分同士が接する部分にエッジが存在するという性質を利用し，できるだけ平坦部分を保存することにより，エッジをぼかすことなくノイズを除去するようにした方法である．

しかし，この方法では，文字の表示ははっきりするが，背景領域のノイズは，ほとんど除去されないままである．つまり，通信コストの問題が解決されないままである．そこで，次の方法を試してみた．

2 値画像用に改良したメディアンフィルタ メディアンフィルタは，着目画素の近傍領域内の画素値を昇順にソートし，その中央値を着目画素の値に変更する方法である．この方法では，

取り扱っている画像が2値画像であるためエッジがぼける．そのため，着目画素が黒画素の場合についてのみ行うように変更した．今回は，近傍領域は5×5画素とした．

この方法によると，背景領域のノイズをかなり除去でき，見た目にもみやすくなったが，特徴量の抽出の際にはまだうまくいかなかった．さらに，図書目録カード特有の罫線等の文字以外のものも除去しなければならない．

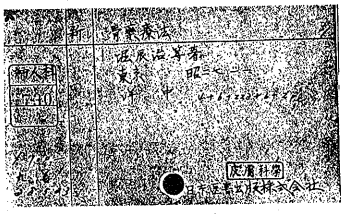
そこで，白画素を黒画素に変更することを禁止しエッジをぼかさないようにし，ノイズの図形的特徴を利用することによりノイズを除去する方法を考えた．以下にそのアルゴリズムを示す．

アルゴリズム

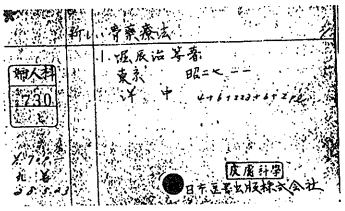
1. 粒状ノイズの除去: 画像のすべての黒画素について以下の操作を行う
   i. 着目画素から，8方向に連続する黒画素数をそれぞれ数える．
   ii. それらがある閾値以下の長さである場合，着目画素は粒状ノイズの一部であるとみなす．
2. 1の操作でノイズとみなした画素を白画素にする．
3. 線状ノイズの除去: 画像のすべての黒画素について以下の操作を行う
   i. 着目画素から，8方向に連続する黒画素数をそれぞれ数える．
   ii. それらから，ある方向にのみ長い直線で他の方向には連続する黒画素がない場合，着目画素は線状ノイズの一部であるとみなす．
4. 3の操作でノイズとみなした画素を白画素にする．

これらの方法によるノイズ除去の実験結果を図4に示す．

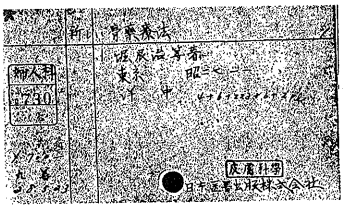
図4より，今回用いた方法が最もエッジを保存しつつノイズを除去することができていることがわかる．



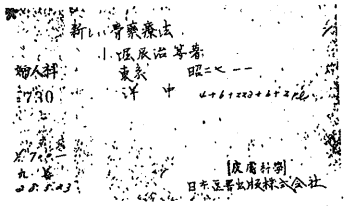
(a) 原画像



(b) 2値画像用メディアンフィルタ



(c) 選択的局所平滑化



(d) 今回用いた方法

図 8: ノイズ除去の比較

5 まとめ

本稿では，図書目録カードのイメージデータを対象とした検索システムについて述べた．このシステムは，経費や人手の面で，比較的容易に構築することができるため，大学図書館の電子化推進に寄与するものと考えられる．また，文字画像の図形的特徴量を利用した文字列検索機能について述べ，この特徴量抽出の前処理として必要であり，データ圧縮の効果もあるノイズ除去法を新たに提案した．既存のノイズ除去

法との比較実験により本手法の有効性を示した．

今後は，手書き文字に対する有効な図形的特徴量の抽出法の開発を行う予定である．

参考文献

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

[2] 栗田英和, 柴田裕介, 竹田正幸, and 有川節夫. **図書目録カードのイメージ化とその検索システム**. 情報処理学会九州支部研究会報告, 3 1999.

[3] 村上伸一. **画像処理工学**. 理工学講座. 東京電気大学出版局, 1996.