全文検索可能な文書画像データベースシステムの構築

仙田修司
日本電気(株) 情報メディア研究所
〒216 川崎市宮前区宮崎4-1-1
Tel: 044-856-8487, Fax: 044-856-2236, E-Mail: senda@pat.cl.nec.co.jp
美濃導彦
京都大学工学部
〒606-01 京都市左京区吉田本町
Tel: 075-753-5995, E-Mail: minoh@kuis.kyoto-u.ac.jp
池田克夫
京都大学工学部
〒606-01 京都市左京区吉田本町
Tel: 075-753-5371, E-Mail: ikeda@kuis.kyoto-u.ac.jp

概要

ディジタル図書館の重要な課題の一つとして、膨大な数の図書・雑誌の効率的なディジタル化を挙げることができるが、それには人手を要しない無修正全文OCRによる方式が有効である。本稿では、文字切り出し／文字認識の候補を文字ラティスとして表現し、文字ラティスに対する高速な検索手法を採用することで、無修正全文OCRによる方式よりも検索の成功率を高くする手法を提案する。この手法を実装した文書画像データベースシステムのプロトタイプ上で評価を行った結果、文字認識率が93.2%の場合に、通常のテキストの4.02倍のサイズの文字ラティスによって文字認識率99.2%のOCRと同等の性能を得ることができた。また、この場合の検索の成功率は97.8%であった。

キーワード

文字認識、文字ラティス、全文検索、文書画像データベース

A Document Image Database System with Full-text Search Functions

SENDA Shuji
Information Technology Research Labs., NEC Corp.
4-1-1, Miyazaki, Miyamae-ku, Kawasaki, 216, JAPAN
Phone: +81 44-856-8487, Fax: +81 44-856-2236, E-Mail:senda@pat.cl.nec.co.jp
MINOH Michihiko
Faculty of Engineering, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-01, JAPAN
Phone: +81 75-753-5995, E-Mail: minoh@kuis.kyoto-u.ac.jp
IKEDA Katsuo
Faculty of Engineering, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-01, JAPAN
Tel: +81 75-753-5371, E-Mail: ikeda@kuis.kyoto-u.ac.jp

Abstract

Digital Libraries need easy and low-price methods of digitizing large number of books and magazines, and thus, raw (not-modified) OCR texts are used for full-text search functions. In this paper, we propose to use hypotheses of characters (character lattices) instead of a raw OCR text to achieve higher rate of successful search. The experimental results tested on the prototype of the document image database system show that the accuracy of the character lattice (of which size was 4.02 times larger than that of the true text) was 99.2% even if the character recognition rate was 93.2%, and the rate of successful search was 97.8%.

Keywords

Character Recognition, Character Lattice, Full-text Search, Document Image Database

1. はじめに

近年、計算機の処理能力の向上、記憶媒体の低価格化、情報インフラの整備などが急速に進んだ結果、ディジタル図書館が現実のものとなりつつある。ディジタル図書館とは、文書、画像、映像、音響などを含むマルチメディア図書の収集、蓄積、配布をディジタル信号の形態で統合して扱う図書館を指す [1]。図書をディジタルデータとして蓄積しているがゆえに、世界中のあらゆる場所からネットワークを介して蔵書を検索・閲覧することができるのがディジタル図書館の最大の利点である。

ディジタル図書館を実現するための重要な課題の一つに、過去から現在にいたる間に出版された膨大な数の図書・雑誌をいかにして効率よくディジタル化するか、が挙げられる。文書のディジタル化には、ページイメージによるものと全文テキストによるものとがある [2]。ページイメージは、スキャナによって文書をビットマップイメージとして取り込んだものであり、作成コストが安く、閲覧に向くという特徴を持つ。一方、全文テキストは、OCR を利用するなどして文書をテキストとして表現したものであり、データサイズが小さく、検索に向いている。しかし、現状の OCR の読み取り精度は十分なものではなく、その誤りを人手で修正して完全な全文テキストを得るにはコストがかかり過ぎる。そのため、全文テキストのみを提供するといった形態は一般的ではない。

それに対して、TULIP[3] で実現されているように、閲覧にはページイメージを用いながら、OCR で読み取ったままの (修正を加えない) 全文テキストを検索用として利用する形態が考えられる。この方式によれば、ページイメージ方式の利点 (低コスト、容易な閲覧) はそのままに、わずかなデータ量の増加 (1/20 程度) だけで、一次データによる全文検索が可能になる。この方式の問題点は、OCR の認識率が高くなければ検索結果が役に立たないことである。20 年来の文字認識研究の結果、市販の低価格日本語 OCR ソフトでさえ、高品質の印刷文書に対しては 92%-98% 程度の認識率を達成しており [4]、これは十分実用に耐え得る値であると思われる。しかし、実際には、印刷品質が低い文書、フォントが特殊な文書、プロポーショナルフォントが混在する文書などでは認識率の低下は避けられず、特に古書などでは、現状の OCR では実用に耐えない。

そこで、本稿では、文字認識率が低くても (人手によって修正することなく) 全文検索が可能な手法を提案する。通常の OCR では、ページイメージ (文書画像) をテキストに変換するが、本稿で提案する手法では、「文字ラティス」と呼ばれる形式 [5] に変換する。文字ラティスの例を図 1 に示す。文字ラティスは、文字切り出し／文字認識の仮説をラティス構造によって表現するもので、従来は、言語知識を利用した文字認識後処理 [6] に用いられていた。筆者らは、与えられたキーワードを文字ラティス中から高速に検索する手法を提案し [7]、これによって、全文検索可能な文書画像データベースシステムの構築を行った。このシステムによって「分散」というキーワードを検索した結果を図 2 に示す。

以下、2 章では文書画像からの文字ラティスの生成法について述べ、3 章では文字ラティス中の文字列検索法について説明する。4 章では提案手法の評価を行う。5 章はまとめである。

2. 文書画像からの文字ラティスの生成

スキャナで取り込まれた文書画像から文字ラティスを生成する手順を以下に記す。なお、ここで採用した手法は文書画像処理としては特に目新しいものではなく、既存の手法 [8] の組み合わせである。

- 画像から文章以外の領域を除外する。ここでは、黒画素連結成分のうち、大き過ぎるもの、小さ過ぎるものを除外した。
- 文書画像を行に分割する。ここでは、空白によって文書画像を分割していく方法 [9] をとった。

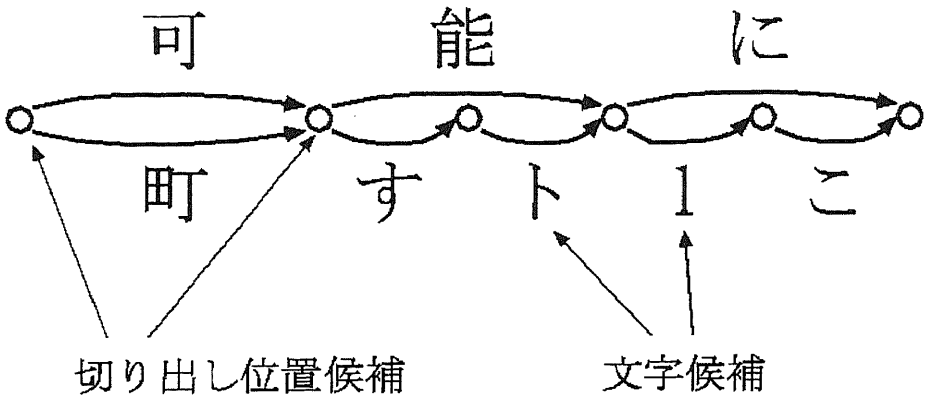

図 1 文字ラティスの例

- 行から文字の候補を切り出す。ここでは、横幅が一定値以下のものは全て文字の候補とした [5,9]。
- 切り出された候補を文字認識する。ここでは、方向線素特徴量 [10] を用い、9 書体の印刷活字フォントを平均したものを辞書として、ユークリッド距離によって認識を行い、一定個数の文字候補を出力した。
- 距離値の分布に従い文字候補数を削減する。切り出し文字ごとに、出力された文字候補の (辞書との) 距離値の平均、標準偏差を求め、偏差値が一定値以上のものを削除する。偏差値による評価は距離値の性質によらずに適用できるので採用した。

上記で述べた手順は 2 値画像を前提としているが、筆者らが提案している文献 [11] および文献 [12] の手法を適用することで、カラー文書画像に対しても適用可能である。

3. 文字ラティス中の文字列検索法

図 1 の例のように、文字ラティスは、文字切り出し／文字認識の仮説を候補として表現している。このようにラティス構造をしたテキストに対して、一次元的なテキストを対象とした従来の検索アルゴリズム [13,14,15] は適用できない。そこで、筆者らは、文字ラティス中から文字列を高速に検索するアルゴリズムを提案した [7]。本章では、このアルゴリズムの概要について述べるとともに、全文検索に有用と思われる挿入、削除、置換を許す検索についても述べる。

3.1 文字ラティス検索アルゴリズム

提案した文字ラティス検索アルゴリズムでは、まず、全ての切り出し位置候補に対して、M-state と名付けた (状態の集合を表す) 変数を用意し、それらを空に初期化する。M-state は、検索すべき文字列 (キーワード) のどの部分まで照合が済んでいるかを表すものである。文字ラティスは、各文字候補を (左端の切り出し位置、右端の切り出し位置、文字コード) の三つ組で表すことができ、この三つ組によって、M-state を更新していくことによって照合は進む。M-state の更新は、左端の切り出し位置の M-state と文字コードによって、右端の切り出し位置の M-state に新たな状態を加えることによって行う。アルゴリズムの詳細については、文献 [7] で述べている。

上記のアルゴリズムを実装する上で問題になるのは、M-state が集合であり、M-state の更新が集合演算になることである。M-state をどのように実装するかについては、文献 [7] では 2 つの方法を提案した。一つ

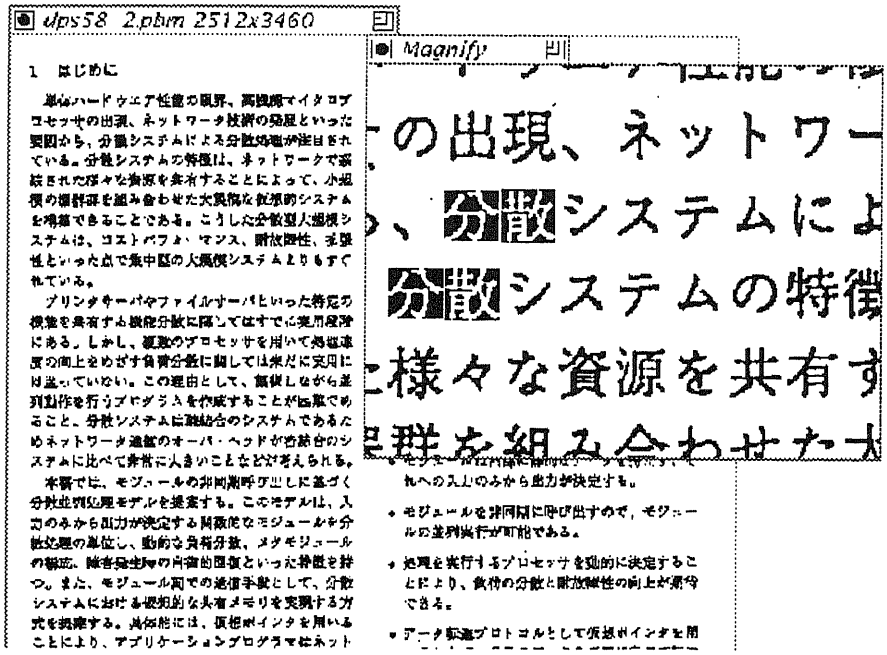

図 2「分散」を検索した例

は、トライ [16] によるものであり、もう一つは、ビット配列 [17] によるものである。トライは複数の文字列を 1 つの木構造で表現したものであり、ビット配列は有限要素の集合をビット (0 または 1) の配列で表現したものである。

トライを用いた実装では、多数の文字列を同時にかつ効率的に検索することが可能であり、文献 [7] での実験によれば、1 単語を検索する時間の 3 倍弱の時間で 10000 単語を同時に検索することができた。よって、文字ラティス中から辞書単語を抽出するといった用途に向いている。筆者らは、抽出した単語の共起関係を利用した文書解析 [18,19] などに利用した。

それに対して、ビット配列を用いた実装は以下の点で文書画像の検索に向いている。

- 小数のキーワードを検索する場合には、トライを用いる場合よりも速い。
- 誤りを許す検索、正規表現による検索などが実現できる。

後者については次節で詳しく述べる。

3.2 ビット配列による誤りを許す検索の実現

ビット配列は、有限要素の集合をビットの配列で表現したものであり、例えば、32 ビット CPU では 1 ワードで 32 要素の集合を表現できる。ビット配列を用いれば、状態の集合である M-state を容易に表現でき、かつ、M-state の更新がビットシフト演算とビット和演算とで実現できるという点で、状態数が少ない (キーワードの文字数が少ない) 場合に有利である。

文献 [20] では、通常のテキストに対して、ビット配列を用いて誤りを許す検索と正規表現による検索を実現している。ここでいう誤りを許す検索とは、k 文字以下の挿入、削除、置換があっても正解とみなす検索である。この場合、検索時間は k+1 倍になる。

文字ラティス中の文字列検索についても、文献 [20] と同様の工夫で誤りを許す検索を実現することができた。すなわち、M-state を k+1 個用意し、挿入、削除、置換を許す演算を順番に k 回行えば、k 文字以下の挿入、削除、置換が検出できるというものである。また、挿入、削除、置換の重みを変えて検索することもできる。例えば、挿入と削除を 2、置換を 1 として全部で 2 以下の誤りを許すと、1 文字の挿入または削除、もしくは 2 文字の置換を許す検索が可能になる。

正規表現による検索はやや複雑になるので実装はしていないが、同様に実現可能である。このように複雑な検索が可能になるのは、ビット配列が複数の状態を保持しながらも演算量は変わらないことによる。

4. 実験による評価

4.1 文字ラティスの生成

ある日本語の論文 1 ページ (1711 文字) から生成された文字ラティスについて評価を行った。

- 文字切り出し。接触しておらず、閾値以下の横幅を持つものは全て切り出し候補としているので、この条件を満たしていれば必ず切り出せる。実験に用いたページでは、横幅の閾値を高さの 1.1 倍とした場合、文字の接触が 1 箇所あったのでこの部分のみ切り出せなかったが、通常の OCR が不得意とするプロポーショナルピッチの英字部分を含めてその他は問題無く切り出せた。切り出された文字の候補は 2428 文字、切り出し多候補率 (切り出された文字数／真の文字数) は 1.42 倍であった。
- 文字認識。実験ページの文字認識率は 93.2% であった。8 候補中から距離値が平均以下のもの (偏差値 50 以下) だけを残した結果、多候補率 2.83 倍で 99.2% の累積認識率を得た。なお、8 候補全てを残した場合の累積認識率は 99.5% であった。

以上の結果から、通常のテキストの 4.02 倍の文字候補ラティスによって、文字認識率 99.2% の OCR と同等の性能を得ることができると言える。

4.2 文字ラティス中からのキーワード検索

文字ラティスを用いることで、検索したいキーワードを見つることができる確率は高くなるが、実際には書かれていないキーワードを見つけてしまう確率も高くなる。そこで、これらの関係を調べてみた。

まず、情報検索の評価によく使われる、再現率と適合率を定義する。

- 再現率 = (見つけたもののうち見つけたかったもの)/(見つけたかったもの)
- 適合率 = (見つけたもののうち見つけたかったもの)/(見つけたもの)

再現率が高いことは、それだけ見つけたかったものが見つかったことを意味し、適合率が高いことは、見つけたものの中にゴミが少ないことを意味する。

前節の実験で用いた文書を対象として、1 位候補のみの場合 (通常の OCR と同様だが切り出しは複数の候補を用いている)、本手法 (平均 2.83 候補/文字)、8 候補全て用いた場合 (8 候補/文字) のそれぞれについて再現率と適合率を求めた。検索すべきキーワードは、かな漢字変換用の辞書から、16234 語の名詞 (1 文字のもの、平仮名のものを除く) を与えた。結果を以下に示す。

	再現率 (見逃し)	適合率 (ごみ)	文字認識率 (誤り)
1 候補	81.6% (68)	99.0% (3)	93.2% (116)
本手法	97.8% (8)	87.0% (54)	99.2% (13)
8 候補	98.9% (4)	54.1% (311)	99.5% (8)

なお、() 内は、見つけようとしたが見逃したキーワード数、存在しないが見つけてしまったキーワード数、文字認識誤りの文字数をそれぞれ表している。

この結果から分かるように、見逃したキーワード数は文字認識誤りの個数にほぼ比例しており、文字認識率の向上がそのまま再現率の向上につながっている。また、候補数を増やすことによって適合率は急激に下がることも分かる。どの程度の再現率、適合率が必要かは状況次第であるので、状況に応じてこれらの値を変えられるような枠組みも今後は必要になるだろう。

また、3.2 節で述べた誤りを許す検索を行った場合、上記の結果よりも再現率が向上し適合率が低下するが、特に短いキーワードに対しては適合率の低下が激しく実用的ではなかった。ある程度長いキーワードに限定する、通常の検索で見つからなかった場合のみ実行する、などの工夫が必要であろう。

5. まとめ

本稿では、文字切り出し／文字認識の候補を文字ラティスとして表現し、文字ラティスに対する検索手法を提案することで、文字認識率が低くても検索の再現率が高くなる手法を提案した。この手法を採用した全文検索可能な文書画像データベースシステムのプロトタイプを構築し、その上で評価を行った。その結果、文字認識率が 93.2% の場合でも、通常のテキストの 4.02 倍の文字ラティスによって、文字認識率 99.2% の OCR と同等の性能を得ることができ、検索の再現率 97.8%、適合率 87.0% を達成した。また、誤りを許す検索手法についても述べた。

今後の課題としては、再現率と適合率を必要に応じて自由に設定する手法の考案、誤りを許す検索手法による再現率の向上などが挙げられる。

参考文献

[1] 田畑: “ディジタル図書館とは,” 情報処理, vol.37, no.9, pp.814-819, 1996.

[2] 杉本: “ディジタル図書館実現のための要素技術と環境要素,” vol.37, no.9, pp.820-825, 1996.

[3] Elsevier Science: “TULIP - The University Licensing Program,” http://www.elsevier.nl/homepage/about/resproj/tulip.shtml.

[4] ソフトバンク技研, 大原: “徹底比較! 日本語 OCR ソフトウエア,” DOS/V Magazine, Technical Test Labs, 5.1, 1996.

[5] 村瀬, 若原, 梅田: “候補文字ラティス法による枠無し筆記文字列のオンライン認識,” 信学論, vol.J68-D, no.4, pp.765-772, 1985.

[6] 村瀬, 新谷, 若原, 小高: “言語情報を利用した手書き文字列からの文字切り出しと認識,” 信学論, vol.J69-D, no.9, pp.1292-1301, 1986.

[7] Senda, S., Minoh, M. and Ikeda, K.: “Fast String Searching in Character Lattice,” Trans. on IEICE, vol.E77-D, no.7, 1994.

[8] 美濃: “文書画像処理の現状と動向,” 信学会誌, pp.502-509, vol. 76, no. 5, 1993.

[9] Senda, S., Minoh, M. and Ikeda, K.: "Document Image Retrieval System Using Charcater Candidates Generated by Charcater Recognition Process," Proc. of 2nd ICDAR, pp.541-546, 1993.

[10] 孫, 田原, 阿曽, 木村: "方向線素特徴量を用いた高精度文字認識," 信学論, vol.J74-D-II, no.3, pp.330-339, 1991.

[11] 久保, 美濃, 池田: "カラー文書画像からの写真領域抽出手法," 第 48 回情処全大, 2-41, 1994.

[12] 仙田, 美濃, 池田: "文字列の単色性に着目したカラー画像からの文字パタン抽出法," 信学技報, PRU94-29, 1994.

[13] Boyer, R. S. and Moore, J. S.: "A Fast String Searching Algorithm," Comm. of the ACM, vol.20, no.10, pp.762-772, 1977.

[14] Aho, A. V. and Corasick, M. J.: "Efficient String Matching: An Aid to Bibliographic Search," Comm. of the ACM, vol.18, no.6, pp.333-340, 1975.

[15] Knuth, D. E., Morris, J. H. and Pratt, V. R.: "Fast Pattern Matching in Strings," SIAM J. Comput., vol.6 , no.2, pp. 323-350, 1977.

[16] Aoe, J.: "Computer Algorithms – Key Search Strategies," IEEE Computer Society Press, 1991.

[17] Baeza-Yates, R. and Gonnet, G. H.: "A New Approach to Text Searching," Comm. of the ACM, vol.35, no.10, pp.74-82, 1992.

[18] 津田, 仙田, 美濃, 池田: "自動作成された単語間リンクによる検索質問作成支援," 第 48 回情処全大, 1993.

[19] 仙田, 美濃, 池田: "文書画像を対象とした未知単語の抽出法," 第 48 回情処全大, 2-15, 1994.

[20] Wu, S. and Manber, U.: "Fast Text Searching Allowing Errors," Comm. of the ACM, vol.35, no.10, pp.83-91, 1992.