情 報 学 基 礎 39－9
（１９９５．９．１４）

文書認識と全文検索の融合技術に関する実験的検討

丸川 勝美　藤澤 浩道　嶋 好博

marukawa@crl.hitachi.co.jp
(株) 日立製作所　中央研究所　知能システム研究部
〒185　東京都国分寺市東恋ケ窪 1-280

紙の世界と電子的な世界との掛け橋となる文書認識と，全文検索を融合する技術の一つとして，「文書認識の認識誤りを考慮して検索できる全文検索」が挙げられる．取り挙げる手法は，誤り特性「正解を複数候補中に含む率は高い」に着目し．必要な候補文字を絞り込むことで不要な検索ノイズを低減させ，複数候補の曖昧性を利用して検索する．本報告では，1,083文書(約40万字)のテキストに対し印字品質の異なる二種類の認識結果を生成し検索精度を測定することで，本手法が認識と検索との融合を可能にする技術であることを示す．実験の結果，通常印字品質に対する再現率を97.9％から99.5％，低印字品質に対するそれを91.4％から98.7％に向上させることを確認した．

Evaluation of Information Retrieval Method based on 'non-deterministic text' of Character Recognition

Katsumi MARUKAWA, Hiromichi FUJISAWA and Yoshihiro SHIMA

Central Research Laboratory, Hitachi, Ltd.
Kokubunji-shi, Tokyo, 185 Japan

This paper presents an information retrieval method that uses the output of an optical character reader (OCR). The method is to combine full-text search and document recognition effectively. The OCR outputs 'non-deterministic text' which keeps multiple recognition candidates for ambiguous classification. The search algorithm is extended to search through the non-deterministic text. Experiments for 1,083 Japanese news articles have shown that the method is effective to improve the recall rate.

1. まえがき

これまで，人手で紙のメディアから電子的なデータであるテキストデータが作成され，膨大な時間とコストを要してきた．また，電子的に蓄積された文書を検索する場合には，特に，「サーチャー」と呼ばれる専門家が所望の文書を適切に検索していた．そのため，ユーザが，容易にかつ効果的に，電子的データを蓄積し所望の文書を検索することが重要となってきており，紙のメディアの世界と電子的な世界とのギャップの掛け橋となる文書認識と，全文検索を効果的に融合する技術が期待される．ここで，この期待に答える一つの解として，自由語を用いて検索[1],[2]でき，かつ，文書認識の認識誤りを考慮して検索できる全文検索[3]が挙げられる．

文書認識の出力である認識結果は，類似文字や(卜(漢字)，ト(片仮名))等の同形文字そして入力文書の品質の低下等により，100%の認識精度を期待することは困難である．W. B. Croftらはこの不完全な認識結果を利用して全文検索を行い，認識誤りが検索精度に与える影響について報告している[4]．

認識誤りを克服し検索を行うアプローチとして，以下に示す三つのアプローチがある．一つは，認識誤りを修正し誤りのないテキストを生成するアプローチである．これには，英文に対するスペルチェッカーや，言語知識を利用して認識誤りの発見およびその修正を行う方法[5]-[7]がある．二つ目のアプローチは，認識機能を利用せず，直接文書イメージを検索する方法である．この一つとして，"Transmedia"と呼ばれるシステム[8]がある．これは各文字を2-3 bitの曖昧不完全記号で表現し，これを利用して文字列マッチングにより検索を実現している．同システムにおいて，日本語文書に対応した方法[9]が開発されている．また，単語単位でイメージを扱い，文書イメージを検索する方法[10]が提案されている．これは英文対応であり，英文の場合には単語単位の高い切出し精度が期待できるからである．そして，文章の節単位で文書イメージを検索する方法[11]が提案されている．最後のアプローチは認識結果を考慮することで検索精度を向上させる方法[12]である．これは，検索の過程で，複数の認識候補を利用し，文字切出しの曖昧性も考慮している．本文献では検索キーと検索用データとのマッチング方法を提案しており，1枚の文書画像に対する処理速度等について報告がなされている．従って，第三のアプローチが有効であるのか，その有効性はどの程度であるのか，また，このアプローチで実現する手法が文書認識と全文検索を効果的に融合する技術となりうる可能性があるのかどうか明らかではなかった．

本報告は，第三のアプローチに関し，認識誤りを考慮した検索手法を示すとともに，1,083文書の日本語文書を認識した結果を利用して検索精度に関する評価ついて述べ，第三のアプローチにおける手法の有効性を始めて定量的に示し，本手法が文書認識と全文検索を効果的に融合する技術となることを示す．利用する文字認識の誤り特性は「類似文字が存在する文字は類似した文字が原因で誤り易く，また，第1位候補に正解が含まれていない場合でも複数の候補中に正解文字が含まれる率は高い」ということである．本手法(複数認識候補型検索)は適切と思われる候補文字を絞り込むことで少ない範囲で累積正解率を高め，精選した複数の候補文字を検索用データとして利用し，検索キーとのマッチングを行う．すなわち，認識の曖昧性を積極的に利用し，検索用データ内に少ないデータで正解の文字列が含まれる可能性を高め，検索キーとの不要なマッチングを回避しマッチング精度を向上させる．また，本報告では，1,083文書の新聞記事テキストを印字品質を変えてLBP出力し，品質の異なる二種類の評価用文書認識データ(計2,166文書)を生成し，それぞれのデータに対し，上述した手法を文書検索および単語検索の観点から評価した．

本報告では，2.で全文検索システムと，認識出力を検索用データとした時の検索手法の設計方針について説明する．3.で複数認識候補型検索手法に関して述べる．4.で1,083文書を検索対象としたときの本手法の精度評価実験の結果を示し，その有効性を確認する．最後に，5.で結論と今後の課題を示す．

2. システム構成と設計方針

2.1 全文検索システムの構成

本報告で扱う全文検索システムは，図1に示すように，ユーザが検索キーを入力し，これに対する出力として検索した文書の一次/二次情報を表示もしくは印刷する．本システムは大きく三つの部分から構成される．一つは，検索のための文書イメージを蓄積・管理するモジュールである．これは文書をイメージ入力する際に蓄積機能が起動され，検索結果に応じて一次/二次情報を出力する際に管理機能が起動される．二つ目の文書認識モジュー

ルは，先のモジュールにより蓄積された文書イメージから検索のための情報を抽出するモジュールである．これは検索用データの生成時にのみ起動され，認識した候補文字等を出力する．最後の全文検索モジュールは，ユーザが入力した検索キーにより，検索用データから文書を検索し，検索結果をイメージ蓄積・管理モジュールに転送し，検索結果である文書イメージ等を得るモジュールである．

図2(a)のテキストを文書認識により認識した際の出力を図2(b)に示す．括弧内は曖昧な候補文字群である．これから分かるように文字認識の出力は単一もしくは複数の候補文字である．出力は候補文字と，入力文字パターンとの類似らしさを表す距離情報(類似度)である．本実験で利用した類似度は0から256で表現されている．

2.2 設計方針

一般に，図1の文書認識モジュールで扱う認識対象はマルチフォントで記載された文字であるため，正解文字が必ずしも第1位に存在せず，図2(b)の括弧で示した文字のように類似したカテゴリが妨害となり，認識精度が低下する．また，(ト(漢字)，ト(片仮名))や(へ(平仮名)，へ(片仮名))等の同形文字に対し，文字認識としての機能は正しいが文字コードは正しくないという現象が生じる．この場合，完全一致で実行される全文検索では正しい解が得られない．また，低解像度スキャナやコピー等により品質が低下したイメージの場合，文字認識の精度が低下する．さらに，日本語文書に存在する漢字や平仮名等は一文字が偏や旁の複数の部分から構成されるものがあるため，文字切出しにおいて，一文字が二文字に認識されたり，二文字が一文字に認識される等の誤りが生じる．このように，文書認識で100%の認識精度を期待することは難しい．

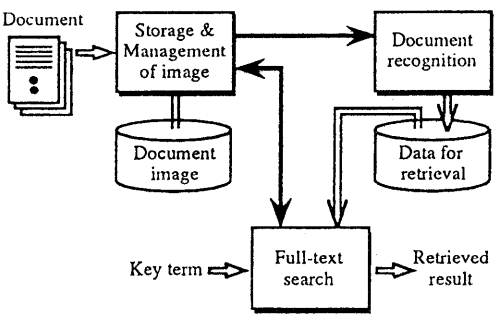


図1　文書認識を利用した検索システムの構成
Fig. 1　Architecture of document retrieval system using document recognition.

認識誤りを考慮しない第1位認識候補文字からなるテキストを検索用データとした全文検索では，認識誤りが検索されるべき単語に含まれる場合に検索漏れが生じ，また，認識誤りにより不適切な単語が文字列変形を起こし，それが検索キーと照合することで検索ノイズが生じる．

本方針では，文字認識の誤り特性「第1位候補が正解でない場合でも，第2位以下の認識候補中に正解が含まれる率は高い」を考慮する．複数認識候補型手法は，不要な検索ノイズを低減するため，候補文字を絞り込むことで少ない範囲で累積正解率を高める．そして，認識の曖昧性を積極的に利用し，絞り込んだ複数の候補文字を検索用データとして利用することで，認識誤りにより生じる検索漏れを低減させ，検索精度を向上させる．

3. 複数認識候補型検索手法

本手法は，上記方針に基づき，複数の候補文字を検索用データとして用いて全文検索を行なうものである．これは次の二つのステップからなる．まず，距離情報を用いて各文書の候補文字を絞り込む．そして，これを有限オートマトン(以下，オートマトンと略)で表現し，オートマトンと検索キーとのマッチングを行う．逆に，検索キーをオートマトンで表現し，検索用データとマッチングする方法[12]も可能である．以下，二つの各処理について述べる．

(1)　候補文字の絞り込み処理

本処理は，オートマトンを生成するための候補文字を絞り込むため，オフラインで実行される。候補文字の個数が増大するに従い，不必要な文字が検索対象となるので，検索ノイズが増加する。そ

日本にも意義深い 宇宙時代の再出発（社説）米国のスペースシャトル「ディスカバリー」が二十九日（現地時間）に打上げられた。その数時間後、ワシントンで「宇宙基地協力協定」に日、米、欧、加の十二カ国が署名した。

(a) Original

日本にも意義深い[字宇]宙時代の再出発（[社 杜]説）米国のス[ペペペ]ースシ[ャヤ][トト]ル「デ[イイ]ス[カ力][バハ]リ[ー一]」が二十九日（現地時間）に打上げられた[。．]その数時間後[、。]ワシン[トト]ンで「宇宙[基碁碁基]地協カ協定」に[日同][、。]米[、。]欧[、。]加の[ナ千]二カ国[かがカ]署名した[。．]

(b) Recognized result

図2　文書認識により出力された認識結果例
Fig. 2　Example of recognized result.

のため，本処理は，候補文字の個数がより少ない範囲で累積正解率を高めるために行なう。各文字を認識した時の候補文字の集合をT0,

$T0 = \{(E_1, F_1), (E_2, F_2), \cdots, (E_n, F_n), \}$ (1)

第i位の候補文字をE_iそしてその距離情報をF_iとする時，本処理により生成される集合T1は，第1位の候補文字E_1の距離情報との差分が閾値2以内の候補文字からなり，次式のように記述される。

$T1 = \{(E_i, F_i) | (E_i, F_i) \in T0, F_1 - F_i \leqq 閾値2\}$ (2)

(2) 検索キーとのマッチング処理

本処理で利用するオートマトンは，図3に示すように，ノードm1とノードm2間のパスにm2文字目の候補文字を順次割り当てることで生成する．ここで，本報告で示すオートマトンは文字切出しの曖昧性を許容していないので，ノード間の遷移は必ず隣のノードに遷移するものである．

本手法は，初期設定時に各文書のオートマトンを生成し，検索時に各文書のオートマトンに検索キーを入力しマッチングを行う．そして，検索キーが受理された文書を検索結果とする．

オートマトン上での検索キーのマッチングは，図3に示すように，マッチング位置制御により，検索キーをオートマトンに入力するノード番号を設定する．そして，設定したノードに検索キーの第1文字目を入力し，遷移するパスが存在すれば遷移を行い，遷移可能なパスが存在する間，検索キーの各文字に対してこの操作を繰り返す．そして，検索キーが受理された場合，処理を行っている文書を検索結果の一つとする．また，遷移するパスが存在しない場合は設定したノードからのサーチは不適切とし，マッチング位置制御によりマッチング位置を右に1ノード分シフトし，同様な処理を繰り返す．

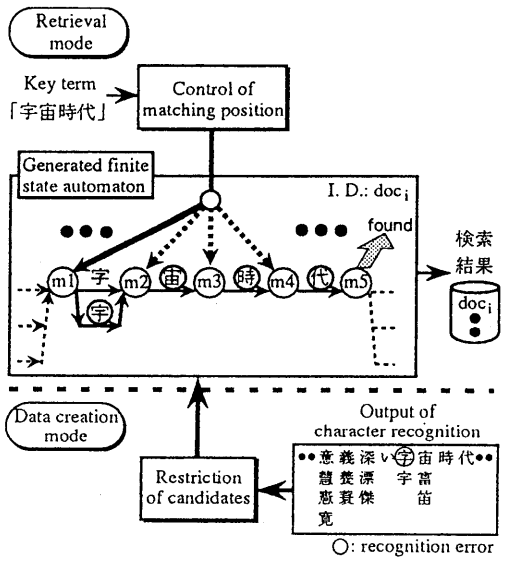


図3 複数認識候補型検索手法

Fig. 3 Retrieval method of multiple recognition candidate type.

本手法において，文字認識の出力候補数を常に1とした時が認識誤りを含む検索用データを用いた従来の全文検索である．

本手法を図3を用いて具体的に示す．紙のメディアの内容「．．深い宇宙時代の．．」を認識し，候補文字の絞り込み処理により，不要と見なされた候補文字を削除し候補文字を選択する．そして，マッチング位置制御により設定されたノードm1から検索キー「宇宙時代」を入力し，各文字に従い遷移することでノードm5に到達し，その結果，受理される。

本手法の特徴は，候補文字中に正解文字が存在すれば，オートマトン上のパスを遷移することで認識誤りが存在しても検索することができ，検索漏れを低減できる．一方，本手法の欠点は，複数の候補文字からオートマトンを生成するため，認識結果の正否に関わらず，正解と異なる領域において検索キーが受理され，検索ノイズが生じる．この具体例を4.3.2 にて示す。

4. 評価実験

4.1 評価用文書認識データ

新聞記事を評価用文書として実験に用いた．本実験では，新聞記事テキストを文字サイズ10ポイントでLBP出力し，これをスキャナ入力することで文書イメージを作成した．以下，実験に用いた文書に関するデータを示す．

(1) 文書数：1,083
(2) 文書当たりの平均文字数：369字
(3) 文書中の最大文字数：690字
(4) 文書中の最小文字数：21字
(5) 文書中に現われるカテゴリ数：2,890カテゴリ

本報告は，検索用データ中の認識誤りに対する本手法の効果を実験的に評価することから，品質の低いイメージを採取し，これを文書認識の入力とし実験を行った．日本語文書の認識を行う場合，通常400dpi(dpi: dot per inch)の解像度のスキャナが用いられるが，ここでは解像度が200dpiのスキャナを用いた．また，以上の条件下で，通常の印字品質

の文書と印字の薄い低印字品質の文書それぞれ1,083文書により，文書認識データを二種類作成した．これらの累積文字認識率を図4に示す．日本語の場合，2.2で述べたように，パターンには同形異コードが存在するため，図4にはこれを考慮した場合とそうでない場合を示す．同形異コードを異コードとして不正解とする場合の第1位認識率はデータ1の場合98.2%，データ2の場合94.3%であり，これが検索用データとなる．また，同形異コードを正解とした第1位文字認識率はデータ1の場合98.6%，データ2の場合95.0%であった．

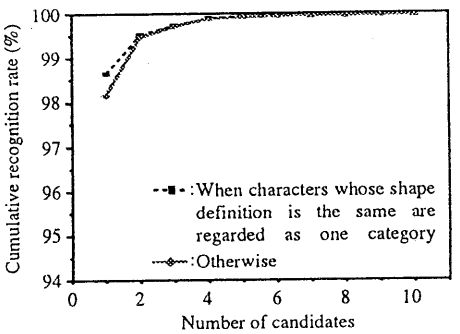
4.2 評価用検索キー

評価用検索キーを選択するため，空間密度を用いた自動キーワード付けによる方法[13)-16)]を利用した．本方法を以下に説明する．

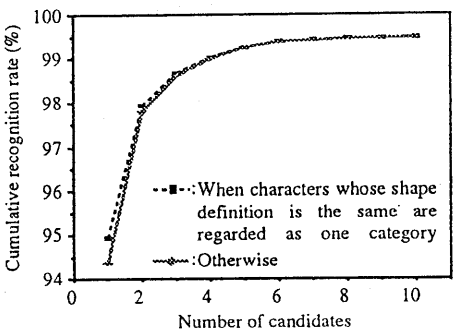
まず，文書の作る文書空間について説明する．文書の集合をP(|P| = N)，キーワードの候補となる語の集合をR(|R| = M)とした時，

$$p_i : v_i = (w_{i1}, w_{i2}, w_{i3}, \ldots\ldots, w_{iM}) \quad (3)$$

を文書p_iのキーワードベクトル表現とする．ただし，



(a) Data set 1



(b) Data set 2

図4　評価用文字認識データ
Fig. 4　Character recognition performance.

w_{ij}を語r_iに対する文書p_iの重要さを示す実数とし，重みと呼ぶ．そして，任意の二つの文書間の類似測度rsを定義することで空間を作る．

自動キーワード付けの方法であるが，空間上での各文書の分散具合いの平均を大きくするような語をキーワードとして選び出し，各キーワードの文書に対する重みを定める．これを次のように形式化する．中心となる仮りの文書p_cのベクトル表現として，

$$p_c : v_c = (w_{c1}, w_{c2}, w_{c3}, \ldots\ldots, w_{cM}) \quad (4)$$

を考える．ただし，w_{ci}は各文書に対するi番目の要素の平均．そして，各文書とこの中心との類似測度の総和を次式のように空間密度SDと定義し，これができるだけ小さい値となる語をキーワードとして選ぶ．

$$SD = \sum_{i=1}^{M} rs(p_c, p_i) \quad (5)$$

G. Saltonらは，いくつかの文書集合を対象とした結果から，以下の手順を提案した．

手順1：各語r_iが生じる文書数B_iを求める．
手順2：B_iの値がN/100からN/10に入る語r_iをキーワード候補として選ぶ．
手順3：空間密度を算出し，密度の高い順に，キーワードとする．

また，G. Saltonらの実験から，文書集合全体に存在する語r_iの頻度をf_iとして，重みw_{ci}を次式のように定義した時に良い結果が得られたと述べている．

$$w_{ci} = f_i / B_i \quad (6)$$

本報告では，上述した方法に基づき，約16万語を含む単語辞書を用いて，4.1の文書群から表1に示す評価用検索キーを50個求めた．

4.3 複数認識候補型検索手法の精度評価

4.3.1 評価基準と評価項目

評価基準として，recall(再現率)とprecision(適合率)の二つの指標を用いる．また，検索手法を文書検索そして単語検索の二つの観点から評価する．二つの指標recallとprecisionを次のように定義する．ここで，認識誤りのない検索用データを用いて全文

表1　評価用検索キー
Table 1　Key term for evaluation.

1	局員	2	物価
3	日産	4	財団
5	援助	6	イラク
7	ホテル	8	投信
9	参院	10	野党
11	日立		

検索した結果を正しい結果と呼ぶ.

recall = B / A (7)

precision = B / (B + C) (8)

ただし,

A：誤りのないデータを用い，全文検索で求めた検索結果の数

B：誤りを含むデータを用いて得られた，正しい結果の数

C：誤りを含むデータを用いて得られた，正しくない結果の数

この時，検索漏れと検索ノイズは以下のように表せる.

検索漏れ = A - B (9)

検索ノイズ = C (10)

4.3.2 評価結果

4.2で求めた50種類の検索キーに対し，文書検索と単語検索の精度を測定した.

評価を行うに当たり，パラメーターとして，文書認識が出力する候補数を変化させた．変化させたパラメーターは四種類あり，3.で述べた手順で，候補文字を絞り込み，オートマトンを生成した．扱った閾値は10/256，20/256，30/256，40/265である．また，各閾値で作成した検索用データの一例を図5に示す．これから分かるように，閾値が大きくなるに従い，曖昧性が大きくなるとともに検索用データが増加する.

上記のデータを利用して求めた文書検索のdocument recallとdocument precisionを図6(a)に示す．文字認識の第1位候補文字のみを扱った場合，認識データ1とデータ2それぞれのrecallは97.9%，91.4%であった．そして，本手法により，認識データ1とデータ2のrecallはそれぞれ99.7%，99.3%と向上した．四種類の検索用データにおいて，ベストのrecall-precisionはレベル1(閾値10/256)の場合であり，認識データ1とデータ2それぞれに対し，(recall, precision)=(99.5%，99.0%)，(98.7%，98.8%)であった．また，recallのエラーは認識データ1の場合には2.1%から0.5%，そして，データ2の場合には8.6%

日本にも意義深い 宇宙時代の再出発（社説）米国のスペースシャトル「ディスカバリー」が二十九日（現地時間）に打上げられた。その数時間後、ワシントンで「宇宙基地協力協定」に日、米、欧、加の十二カ国が署名した。

(a) Original

日本にも意義深い[字宇]宙時代の再出発（[社杜]説）米国のス[ベベベ]ースシ[ャヤ][トト]ル「デ[ィイ]ス[カ力][バハ]リ[ー一]」が二十九日（現地時間）に打上げられた[。．]その数時間後[、。]米[、。]欧[、。]加の[ナ千]二カ国[かが力]署名した[。．]

(b) Level 1

日本に[もら]意義深い[字宇]宙時代の再出発[（くくくく][社杜][説謙税]（米木]国のス[ベベベベ][ー一]スシ[ャヤ][トト]ル「デ[ィイ]ス[カ力][バハ]リ[ー一]」が[二三]十九日（現地時[間問閒]）に打ち上げられた[。．]その数時間後[、。][フフ]シン[トト]ンで「[宇宇]宙[基碁碁]地協カ協[定走]」に[日同Ｂ目固][、。．]米[、。]欧[、。．]加の[ナ千][二ニ][カ力][国図圃固][かが力][署著]名した[。．]

(c) Level 2

[日目][本森]に[もら]意[義養]深い[字宇]宙時代の再出発[（くくくく 力][社杜祉礼][説謙税誰読課課][）ノ][米木禾未氷][国圃図]のス[ベベベベ][ー一]スシ[ャヤ][トト]ル「[デテ][イイ][ス久][カ力][バハ][リリ][ー一][丿 丿丿丿]」が[二三]十[九丸][日目Ｂ６ロ][（くく]現[地池]時[間問閒閒開閒闇闇閒問]）に打ち[上七]げられた[。．、]その[数歎戴致敬欣教]時[問閒問闇闇開閒閒閒][後復][、。．，][ワフリ]シ[ンシ][トト][ンシ]で「[宇宇][宙富寓重][碁碁基塞暮幕募]地協カ[協楊][定走][丿 丿]に[日同Ｂ目固][、。．，][米未禾][、。．][欧殴][、。．，]加の[ナ千予][二ニ三][カ力][国図圃固圃因圏囚][かがカ][署著著暑署]名[しじ]た[。．、]

(d) Level 3

[日目召Ｂ][本森木]に[もら毛][意急意寛惹息][義養義][深深傑]い[字宇][宙富笛]時代の再出発[（くくくく 力][社杜祉礼][説謙税誰読課課悦調難綾鉄][）ノ][米木禾未氷ポボ太][国圃図圏匿因圃固]の[ス久][ベベベベ][ー一←]スシ[ャヤ][トト]ル「[デテ][イイ][ス久][カ力][バハ][リリ][ー一←][丿 丿丿丿]」[がカ][二三工ニ]十[九丸][日目Ｂ６ロ][（くくくく〈〈[現玩硯規視][地池他]時[間問閒閒開閒闇闇閒問閒閒]）に打ち[上七][げザばば][られ]れた[。．、][そそ][の〇。００〇][数歎戴致敬欣教激救敗殺][時晴暗][問閒問闇闇開閒閒問閒閒閒閒][後復援稜彼][、。．，]ワフリ]シ[ンシ][トトF][ンシ]で「[宇宇][宙富寓重害苗寅雷][碁碁基塞暮幕募][地池強][協幡][カ力ヵ][協楊][定走][丿 丿丿]に[日同Ｂ目固][、。．．][米未禾木氷][、。．．，][欧殴][、。．．，]加の[十千予壬][二ニ三工][カ力][国図圃固圃因圏囚困][かがカ][署著著暑署][名各][しじ][ただ][。．、]

(e) Level 4

図5 実験に用いたレベルの異なる四種類の複数認識候補例

Fig. 5 Four examples of non-dterministic text draw from the experimental data.

から1.3%に低下しており，それぞれのエラーの76.1%，84.9%が本手法により低減した．また，単語検索のword recallとword precisionを図6(b)に示す．文字認識の第1位候補文字のみを扱った単語検索のrecallは，認識データ1とデータ2それぞれに対し，96.2%，87.3%であった．そして，本手法により，認識データ1とデータ2それぞれに対するrecallは99.7%，99.5%と向上した．

本手法によるマッチングの失敗例を示す．検索キー「日立」が入力され，文書認識用テキストに「．．目立った．．」があり，この「目」の候補文字に「日」と「目」が含まれていた．この時，「目」の認識結果の正否に関わらず，検索キー「日立」はオートマトン上で受理されたため，不適切な領域で検索されノイズが発生した．

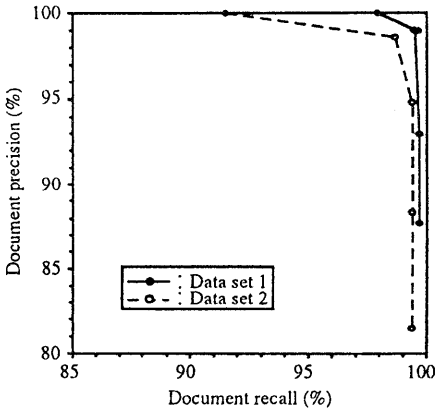
4.4 データ量について

要するデータ量は表2のように表せ，認識データ1で要したデータ量は，従来の全文検索を比率1.0とすると，閾値10/256での複数認識候補型手法が1.28と増加した．

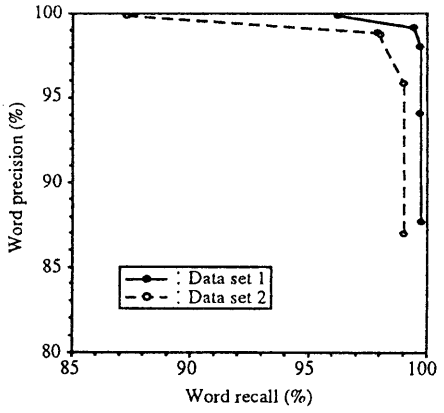
5. むすび

本報告では，文書認識と全文検索との効果的な融合技術に関し，文字認識が出力する曖昧な出力を許容した検索手法を示した．そして，印字品質の異なる文書をそれぞれ1,083文書(約40万字)認識し，その結果を用いて評価実験を行い，手法の有効性を示した．実験の結果から，通常印字品質の文書に対して，recall-precisionの最適点である閾値10/256で，通常印字品質の文書に対する(recall, precision)は(99.5%, 99.0%)で，recallのエラーを76.1%低減し，低印字品質の場合には(recall, precision)は(98.7%, 98.8%)で，recallのエラーを84.9%低減することができた．以上のことから，本手法が文書認識と全文検索とを融合する技術であることが明らかになった．

今後の課題は，precisionの向上を行うとともに，実用化を目指して文字切出しの曖昧性も考慮した手法を検討する．そして，複数の検索キーおよび検索条件でこれら手法の評価をシステマティクに行うことである．



(a) Document level recall and precision



(b) Word level recall and precision

図6　複数認識候補型によるrecall-precision
Fig. 6 Recall-precision relation given by multiple recognition candidate type.

表2　必要なデータ量
Table 2 Data size.

Method	Data size (byte)	Rate
Conventional full-text search	S×2	1.0
Multiple recognition candidate type	S × N1 × 2	N1

S : number of character into retrieval data.
N_1: average of candidates in multiple recognition candidate type.

文献

(1) Aho A. V. and Corasick M. J. : "Efficient String Matching", Commun. of ACM, 18, 6, pp. 333-340 (1975)

(2) Kato K., Fujisawa H., Kawaguchi H., Hatakeyama A., Murakami T., Ohyama M., Kaneko N. and Akizawa M. : "An Index-Free Full-Text Search for Large Japanese Text Bases", Proc. of Advance Database System Symposium '89, pp. 75-82 (1989)

(3) Fujisawa H. and Marukawa K. : "Full-Text Search and Document Recognition of Japanese Text", Proc.

Symp. of Document Analysis and Information Retrieval, pp. 55-80 (1995)

(4) Croft W. B., Harding S. M., Taghva K. and Borsack J. : "An Evalution of Information Retrieval Accuracy with Simulated OCR Output", Proc. of 3rd Sympo. on Document Analysis and Information Retrieval, pp. 115-126 (1993)

(5) Dahl D. A., Norton L. M. and Taylor S. L.: "Improving OCR Accuracy with Linguistic Knowledge", Proc. of 2nd Sympo. on Document Analysis and Information Retrieval, pp. 169-177 (1993)

(6) Itoh N. and Maruyama H.: "A Method of Detecting and Correcting Errors in the Results of Japanese OCR", Trans. of Information Processing Society of Japan, 33, 5, pp. 664-670 (1992)

(7) Niwa N., Kayashima K. and Shimeki Y.: "Postprocessing for Character Recognition Using Keyword Information", Proc. of IAPR Workshop Machine Vision Applications, pp.519-522 (1992)

(8) Tanaka Y. and Torii H. : "Transmedia Machine and Its Keyword Search over Image Texts", Proc. of RIAO 88 , pp. 248-258 (1988)

(9) Yusa M. and Tanaka Y. : "The Extension of Transmedia System for Japanese Text", 49th National Convention Record of IEICE, 2H-9, pp. 217-218 (1994)

(10) Trenkle J. M. and Vogt R. C. : "Word Recognition for Information Retrieval in the Image Domain", Proc. of 2nd Sympo. on Document Analysis and Information Retrieval, pp. 105-122 (1993)

(11) Hull J. J. : "Document Image Matching and Retrieval with Multiple Distortion - Invariant Descriptors", Proc. of 1st Int. Workshop on Document Analysis Systems , pp. 383-399 (1994)

(12) Senda S., Minoh M. and Ikeda K. : "Document Image Retrieval System Using Character Candidates Generated by Character Recognition Process", Proc. of 2nd Int. Conf. on Document Analysis and Recognition, pp. 541-546 (1993)

(13) Salton G. and McGill M. : "Introduction to Modern Information Retrieval", McGraw-Hill (1983)

(14) Salton G., Yang C. S. and Yu C. T. : "A Theory of Term Importance in Automatic Text Analysis", Journal of American Society Information Science, 26, 1 (1975)

(15) Salton G., Wu H. and Yu C. T. : "The Measurement of Term Importance in Automatic Indexing", Journal of American Society Information Science, 32, 2 (1981)

(16) Itoh T.: "Information Retrieval", Shoukoudo, Japan (1986)