ヒューマンインタフェース 29－3
（1990. 3. 9）

文書情報の蓄積検索システムに関する検討

宮原末治　鈴木章　多田俊吉　壁谷喜義
ＮＴＴ　ヒューマンインタフェース研究所

多量に発生する文書情報を効率よく運用する目的で、文書イメージ情報を文字認識してコード情報に変換しておき、情報が必要になった時点で自然言語によって検索するフルテキスト形の文書情報蓄積検索システムの開発を進めている。

このシステムの入力部は、汎用のパソコンにイメージスキャナと試作の小形並列プロセッサを接続して構築している。また、検索部はＡＩワークステーション上にソフトウェアで構築しており、検索は自然言語で入力された文と類似な文をデータベースの中から抽出する方法を取っている。

本資料では、システムにおける文書の入力精度と入力時間、および検索方法とシステムの操作性について検討した結果を述べる。

A STUDY of DOCUMENT RECOGNITION and RETRIEVAL SYSTEM

Sueharu MIYAHARA, Akira SUZUKI, Syunkichi TADA, Kiyosi KABEYA
NTT Human Interface Laboratories
1-2356 Take Yokosuka-Shi Kanagawa-Ken 238-03 Japan

A prototype for a document information processing system is discussed that efficiently perform information searchs of large amounts of documents.

This system can be realized using a document recognition unit and a natural language retrieval program on an AI workstation.

The document recognition unit for this document information system consists of a personal computer, an image scanner, and a special parallel processor for character recognition.

The retrieval program collects sentences in the database that contain any information similar to the input sentence of a natural language.

This paper discusses the accuracy and speed of document recognition, the retrieval method, and the performance of the prototype system.

1　まえがき

オフィス等では増大する文書情報を効率よく運用する目的で，光ディスクをベースにしたイメージ蓄積検索形のファイリングシステムや，文字コード情報をベースにしたデータベースシステムが活用されている．しかし，前者のシステムではデータの入力作業は比較的容易であるが，イメージ情報を直接検索する技術が確立されていない現状では，文書のファイリングに際し，統一的な文書の分類や索引付けなどの事前作業を行う必要があり，検索に際しても内容を確認するために多大な時間を要している．一方，後者の場合はデータベースの作成の際に，要約文の作成やキーワード付与などの事前作業や，データの投入など経費のかかる作業が必要になるために，多大な労力と処理時間とを要しているのが現状である．このようにデータベースの作成に，事前作業とそれに費やす膨大な経費が，従来形の文書蓄積検索システムの普及を阻害する大きな要因になっている．

これらの問題を解決する方法として，筆者らはデータの入力に際しては文書イメージ情報を文字認識して直接コード情報に変換し，検索においては情報が必要になった時点で，日常使用する自然言語を用いて検索・整理するフルテキスト形の文書情報蓄積検索システムについて検討している．

文書情報のコード化に文字読取装置（OCR：Optical Character Reader）を用いる場合，比較されるのが入手による入力であり，主な項目として，

（イ）入力コスト
（ロ）入力精度
（ハ）入力速度(データ入力の終了までの期間)
（ニ）操作性

などが上げられる．この他にも装置の大きさや他システムとの接続の可否などの項目があるが，上記の４項目が満足されれば，多くの分野で利用されるものと思われる．このような観点に立って，我々はワープロと同等の単純な操作で，かつ短時間に文書データを得ることができるＯＣＲの開発を目標にした．また，

フルテキスト形のデータベース検索の現状は，

（ａ）事前加工されていないために検索精度が低い．
（ｂ）問い合わせ文をどのように組み立てて検索すればよいのかが明確でない．
（ｃ）全データを探索するため処理時間がかかる．
（ｄ）検索結果の確認に時間がかかる．

などの問題がある[1]．本検討では，解決の第１段階としてフルテキスト検索の利用形態を整理し，そこから出てきた問題について検討した．以下，システム構成と処理概要，ＯＣＲにおける操作性の向上の工夫と性能概要，文書データの検索手順と検索性能について述べる．

2　システム構成と処理概要

実験システムは，図２．１に示すように文書情報の入力部と検索部に分離し，個々の問題点を実験的に抽出することとした．入力部と検索部とは文書データをファイル転送によって引き渡している．

入力部はパソコン（ＰＣ９８）と市販のイメージスキャナ，試作の文字認識ユニット（ＣＲＵ：Character Recognition Unit）から成り，パソコンを読取制御部とするＯＣＲシステムとなっている．スキャナとＣＲＵとはパソコンのオプションユニットとしてパソコン上のプログラムから制御され，１ページ単位の文書イメージ入力とその認識などがオペレータによるファンクションキーとマウスによる会話的操作で実現できるようにしている．検索部ではコード化された文字情報をフル

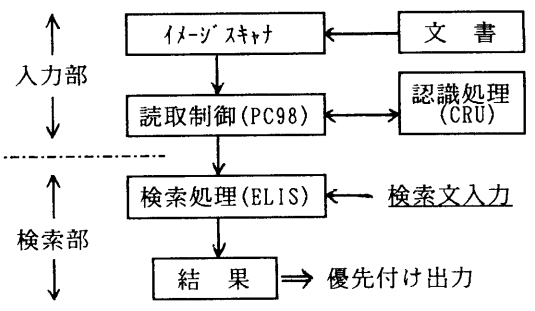

図２．１　実験システムの構成

テキストの形式を保ちつつ検索の単位となる文章，あるいは文の単位（以下，被検索文と呼び，この単位をＴとし，単位文と呼ぶ）でリスト形式に変換して検索に備える．このときどの文字列がどの被検索文に出現したのかを文字列索引として抽出しておき，検索の際に参照することによって，処理を高速化している．検索は利用者が知りたいと思っている情報の核となる単語，あるいは文（被検索文の中に記述されていると予想される単語や文）を想定し，それを問い合わせ文（以下，ここでは検索文と呼ぶ）として入力することにより，その入力文と類似な文を抽出するようにしている．検索結果は類似性の評価スコアの順に被検索文を出力して確認の量が少なくてすむようにしている．

3　文書データの入力

ＯＣＲを用いた日本語文書の入力においては，欧文ＯＣＲの単語照合や数字ＯＣＲの合計値照合などのように完全な確認は困難である．そのため，最終的な確認や修正の処理を人手に頼らなければならない[2]．そこでＯＣＲがイメージと認識結果の候補文字を有するという特長を生かして，入力データの確認･修正が容易な装置の開発を目指した．

表３．１　入力部の主な仕様

項　目	主　な　仕　様
スキャナ (IS-300)	走査給紙：フラットベット形 一枚づつ手でセット
	走査領域：Ａ４判以下
	分解能：１２本/mm(2値画像)
認識処理 (LISCAR)	ﾊｰﾄﾞｳｪｱ：ＡＡＰ2_LSIを4個 ２５６PEによる並列処理 ﾏｼﾝｻｲｸﾙ：１４３ nsec (処理速度：５００MOPS)
	ｿﾌﾄｳｪｱ：命令(80bit/w)使用
読取制御 (PC98VX)	ｲﾝﾀﾌｪｰｽ ・ｽｷｬﾅ：8ﾋﾞｯﾄﾊﾟﾗﾚﾙ ・PC-DMAC：最大160KW/S
	ＣＲＴ：１９ｲﾝﾁ(PC-TV471)

3．1　入力部の仕様と処理内容

文書データの入力部として開発したＯＣＲは，汎用のパソコンにイメージスキャナと試作した小形のＣＲＵ（名称：LISCAR： Line Scanable Cellullar ARray processer）とを接続して文書上の文字を認識する機能を実現したものである．ＬＩＳＣＡＲは２５６個の１ビットＰＥ(Processer Element）から成るプログラマブルなＳＩＭＤ形の並列処理装置であり，Ｂ４判の基板１枚でできている[3]．本装置ではこのＬＩＳＣＡＲに文字認識ファームウェア（認識処理プログラムと認識辞書）を登載して，ページ単位の画像処理機能と日本語の印刷文書認識機能を持たせいる．また，この入力部には入力所要時間の短縮を目的として，認識結果確認･修正の操作を容易にするためのイメージ表示の機能[4]とオペレータの認識結果に対する訂正情報を次回からの文字読取りに利用する学習機能とを設けている．表３．１に今回実現した入力部の主な仕様を示し、表３．２に主な機能を示す．この装置における日本語の印刷文書の入力手順は下記の５つの処理ステップから成っている．

① イメージ入力：パソコンからイメージスキ

表３．２　入力部の主な機能

項　目	主　な　機　能
読取対象	読取領域：全領域，あるいはマウス指定領域
	文字フォント：単一フォント、マルチフォント
	字並び：不定ピッチ文書読取り可
読取量	８０行／頁，１２８文字／字並び 以下
確認修正	規則を学習して自動訂正する
	リジェクト文字は黄色、訂正文字は青表示 リジェクト文字へのカーソル移動
	認識結果のカーソル位置のイメージを文字並び単位で表示（イメージにもカーソル付与）
	候補文字表示・選択(第６位まで)
結果出力	ＭＳ－ＤＯＳ標準ﾌｧｲﾙﾌｫｰﾏｯﾄ

ャナを起動して，１ページ単位の文書イメージを入力し，そのデータをそのままＣＲＵへ転送する．

② 読取領域指定： オペレータがディスプレイ上の入力イメージに対し，文字読取りの対象領域をマウスで指定し，その座標をＣＲＵへ通知して，文字認識を起動する．

③ 文字認識処理： 起動されたＣＲＵは，文字列，および文字を切り出し[5]て文字認識し[6]，認識結果をパソコン側に送り返す．パソコンでは過去の修正情報を参照して，リジェクト文字を自動訂正する[7]．

④ オペレータによる確認･修正： ディスプレイ上に表示されたイメージ情報と候補文字との助けを借りて，認識結果の確認･修正の処理を行う．

⑤ 文書データの保存： 読取結果をMS-DOSの標準フォーマットでファイルに格納する．

各処理ステップにおける占有時間を測定し，表３．３に示すような値を得た．

表３．３　入力部の各処理に要する時間

	処理内容	処理速度
1	イメージ入力と表示･認識部へのﾃﾞｰﾀ転送	Ａ４判の大きさを約４０秒
2	読取領域指定（1ﾎﾞｯｸｽ指定当り）	５～１０秒･
3	文字認識，判定処理等	(15-B)･･ 字/秒
4	ｵﾍﾟﾚｰﾀの確認・修正	認識精度とｵﾍﾟﾚｰﾀの能力に依存
5	データ保存	―――

＊　書式の複雑さに依存

＊＊　Ｂは領域指定数（Ｂ≦５以下のとき）

３．２　入力部の文字認識精度

単一フォントの文字を認識する識別辞書を用いて，Ａ４判の用紙に５号の明朝体活字でオフセットで印刷されたＪＩＳ第一水準漢字，ひらがな，カタカナ，英数字を含む３，１７４字種に対して読取試験を行った．具体的には，学習パターンを第一位正解率９９．９％で認識できる識別辞書を用いて，テストデータを読み取らせた結果，図３．４に示すように正解率９８％（第一位正解率９９．５％），リジェクト率１．８％，誤り率０．２％，認識速度１４字／秒の値を得ることができた．

表３．４　入力部の性能

項　目	性　能	条　件
認識速度（／秒）	平均１４文字	27行，850文字／頁 ｲﾒｰｼﾞ処理を含む
読取性能（％）	正解　：　98.05	５号明朝体活字
	ﾘｼﾞｪｸﾄ：　1.79	単一ﾌｫﾝﾄ用辞書
	誤り　：　0.16	文字種：3,174字種

３．３　操作性向上の工夫

ＯＣＲを用いたデータ入力の処理能力は，装置の処理速度と認識精度，およびオペレータによる読取領域の指定速度と読取結果に対する確認･修正速度によって決まる．そのため操作部はオペレータにとって使い易い設計にすることが重要となる．

認識結果に対する確認･修正の操作は，図３．１に示すように，①から⑥へ進むような処理手順となる．この図は右側に示した処理がより人手を要することを示しており，入力を高速化するには，この処理をいかに上段部分で，かつ左側の処理ですませられるようにするかである．そこで，オペレータの操作，および操作部を次の様に設定した．読取領域の指定にはマウスを用い，その他の業務は日本語入力のかな漢字変換との整合性からキーボードを用いる．読取領域の指定は，領域指定の開始時点から，ポインティング位置が適当か否かの判断ができるように，画面を４分割するクロスカーソルを使用する．読取結果の確認・修正では認識結果とカーソル位置に対応するイメージ情報（１行分），および候補文字（６文字）とを同一画面上に表示して処理を行う（図３．２参照）．この時，リジェクト文字や自動訂正文字はオペレータの注意を喚起するために黄色（リジェクト）と

青色（訂正）で表示し，１タッチのキー操作でこの文字へのカーソルの移動ができるように工夫した．また，日本語入力はパソコンに用意された市販の漢字変換フロントエンドプロセッサを使用した．さらに，表示されたイメージにもカーソルを付けて認識結果との対応を取り易くした．

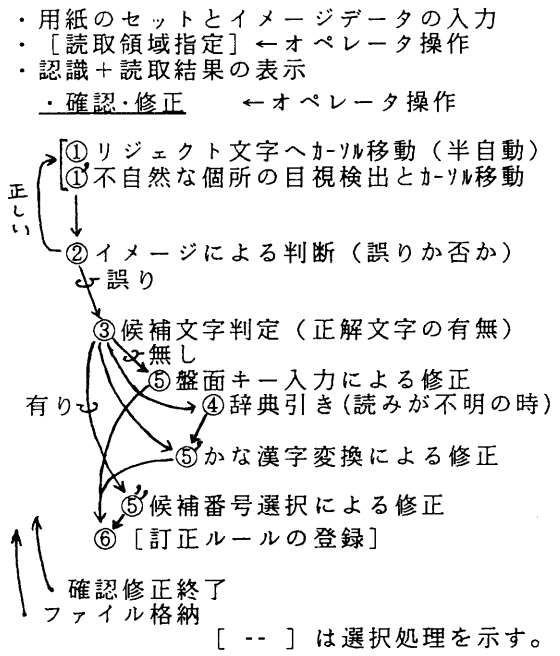

図３．１　読取結果の確認･修正の処理手順

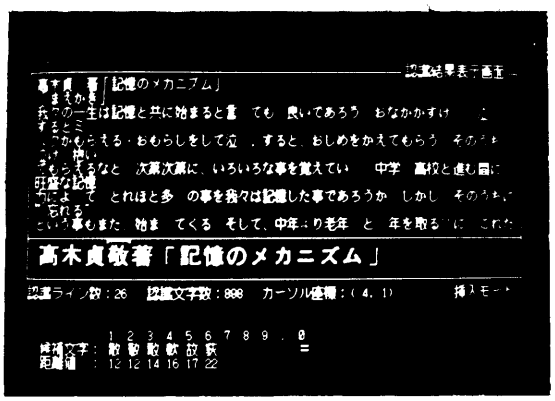

図３．２　認識結果の表示画面

３．４　訂正規則とその効果

認識結果に含まれるリジェクト文字をオペレータが修正する際に，修正した情報（カテゴリ），リジェクトの候補文字，およびその距離値の３種の情報を訂正規則として登録しておき，次回からのリジェクト文字にこの規則を適用してオペレータの操作性を改善することを試みた．この方法の特徴は文書に出現する文字種が識別辞書に存在しない文字種にも適応でき，オペレータの修正作業が少なくなることである．具体的な訂正処理は文献７の方法に従った．

実験では単一フォントの識別辞書に，異フォント文字約３，２００字種を読取らせて訂正規則を作成し，異フォント文字のテストデータの読取りを行った．結果を表３．５に示す．同表はリジェクト文字を修正した場合の訂正規則の登録個数とリジェクトの改善効果を示したものである．この方法を用いることにより，候補が第一位に出現しなかったリジェクト文字の約４４％が訂正されて確認のみの処理ですみ，また，候補中に正解が含まれなかったリジェクト文字の約５０％が候補の中に含まれるようになり，候補番号の選択で入力できるようになることが分かった．

３･５　入力精度と入力速度

確認･修正の操作をオペレータの係わり方から分類すると，下記に示す４種の形態になる．

（ａ）形態Ⅰ：認識結果をオペレータが修正することなく，読取結果をそのままファイルに格納する形態．

（ｂ）形態Ⅱ：認識結果に含まれるリジェクトや誤読文字をオペレータが容易に検出できた部分

表３．５　訂正規則と改善効果

訂正処理	規則数（個）	リジェクトの候補の内訳	
		１位正解率	含有率
前	——	７１．５％	９４．５％
後	７６３	８４．１％	９７．３％

（正解率78.8%，ﾘｼﾞｪｸﾄ率19.9%の時）

（色付きで表示されたリジェクト文字や，気付いた誤読文字）のみをイメージと候補文字とを参照して修正する形態．

（ｃ）形態Ⅲ：認識結果をオペレータが読み直し，その中で不自然と感じた部分のみ，イメージで確認し，修正を行う形態（オペレータが文書の内容を十分理解し得る場合に採用される）．

（ｄ）形態Ⅳ：認識結果とイメージとを１対１に対応づけて確認と修正の作業をする形態．

この中で形態Ⅰ，Ⅱは，校正などの事後処理が可能な場合に運用されることが多く，形態Ⅲ，Ⅳは，ＯＣＲによる入力のデータに対して，１回の校正が加わったものであり，誤りのない文書データが得たい場合の運用である．形態Ⅰは多少の誤り文字の混入を許して短時間に入力したい場合であり，文書データの入力精度と入力速度は，ＯＣＲの第一位正解率と処理速度がそのまま現われる．形態Ⅱは半自動のカーソル移動の機能を用いて比較的短時間で，しかも入力精度の高い文書データを得ることができる．形態Ⅲは誤りのない文書データが要求される場合に採用される．形態Ⅳは誤りが一個所も許されない場合に採用される．これら４つの形態について，約１，５００文字からなる日本語印刷文書（第一位正解率９９．０％）を用い，数回の練習をした未経験のオペレータ２名に対して入力速度を計測し，表３．６に示すような結果を得た．これらの値を表３．６の枠外に示したワープロ文書入力の技能検定試験の値[8]と比較すると，本装置を用いて文書データを入力する方が入力速度では１級より３～５倍，３級より７～１１倍程度有利であり，入力精度も良いことが分かった．

表３．６　利用形態と処理性能

（漢字の出現率は24.1％）

形態	平均性能（％）			処理時間（秒）（900文字）
	正解	ﾘｼﾞｪｸﾄ	誤り	
結果	95.6	4.3	0.1	112
Ⅰ	99.0	——	1.0	115～119
Ⅱ	99.9	——	0.1	172～196
Ⅲ	100.0	——	0.0	253～294
Ⅳ	100.0	——	0.0	——

＊ワープロ文書入力技能検定試験の基準
１級：入力速度900文字(10分間)，正解率98.9％
３級：入力速度400文字(10分間)，正解率97.5％

4　文書データの検索

質問文の意味を理解して検索文を組み立てる検索方法[9]は，検索者の負担を軽くしてくれるが、そのシステム作りに多くの時間と労力を必要する．また知識ベースを作成するのに多大な労力を必要とし，出来たものは用途が限定されることが多い．ここでは汎用化をめざし、検索者にも一定の負担を分担してもらうこととし，データベース内でのテキストの記述内容を想定してもらう方法を採用した．また，検索と確認の処理に計算機の支援を最大限に利用することを考え，オンラインによる対話形の検索法を採用することにした。

フルテキスト検索を行うためのハードウェアにはＥＬＩＳ[10]を用いた．検索の実験はこの装置にテキストとして新聞記事を登載し，文を単位とした検索を行うこととした．検索の処理では，入力された検索文の中から単語を切り出し，その単語をシソーラスを用いて同義語や類義語に展開し，展開した単語を用いて被検索文中の単語との照合をとり，単語の共起を利用して類似文書を抽出している．具体的な検索の処理手順は，下記のような７つの処理ステップから成っている．

① ２文字連接を用いた文字列索引の作成と，被検索文のリスト形式への変換．

② オペレータによる問い合わせ文（検索文）の入力．

③ 入力文からの単語（キー単語）抽出と，同義語シソーラスによる同義語･類語への展開（これらの単語を総称して検索単語と呼ぶ）．

④ 文字列索引を用いた検索単語と同じ文字列を持つ被検索文（候補文）の抽出．

⑤ 候補文をフルテキストサーチし，検索単語の種類数と出現数との計数．

⑥ 候補文の類似性の計算と，候補文の優先順位付け．

⑦ 優先付けされた検索結果の表示･出力．

この処理システムではシソーラス展開した検索単語群に対して，オペレータが自由に介在して検索単語の状態を変更できるようにしている．

4．1　同義語シソーラス

フルテキストデータベースを専任の検索者以外の人が利用する場合，色々と異なった表現で検索することが想定される．また，検索文に用いられた単語でそのまま検索すると，期待した結果が得られないという問題が生じる．そこで，これら表現のゆれを吸収するために，一般語用の同義語シソーラスを作成した．このシソーラスは表４．１に示すような構成になっており，これまでに約１８万語を収録している[11]．このシソーラスを新聞記事１ページに用いられる単語と，どの程度オーバラップしているかを調べたところ，約５３％をカバーしており，１つの語が平均３個の単語に広がっていることが分かった．このことは検索文においても平均３種類の表現を許容して検索ができることを意味している．

表４．１　シソーラスの関係種別と例

関　係	関係種別	例
同義語0	表記のゆれ	移り変わり／移り変り
	略語／完全語	IMF／国際通貨基金
	ｶﾀｶﾅ語／対訳語	ｺﾝﾋﾟｭｰﾀ／電子計算機
同義語1	言い替え語	お父様／おやじ
	ｶﾀｶﾅ語／対訳語	イメージ／映像
同義語2	準同義語	誤字／あて字
	ｶﾀｶﾅ語／対訳語	ﾋｯﾁﾊｲｸ／無銭旅行する
類義語	種／類	衣服／背広，ｼﾞｬﾝﾊﾟｰ
	全体／部分	動物／頭，胴，首，手，足
	属性名／属性値	色／赤い，青い，白い
	包括語／実現値	挨拶／おはよう
	使役	悩む／悩ます
	転成名詞	喜ぶ／喜び
	受動態	教える／教わる
関連語	連想的関連	現品／見本、代用品
	対語	教える／学ぶ

4．2　類似性の計算

類似性の評価は，まずデータベース内の単位文Ｔの中に検索単語が存在するキー単語の種類数（MAX:Ｎ個）の多さを優先させて，次にその優先付けされたランクの中で，算出式〔１〕の評価値Ｒが大きいものを類似性の高い文と見なした．

$$R=\sum_{i=1}^{N}\alpha(k)*W(i)+Wa \quad \cdots\cdots〔1〕$$

ここで，Ｎは検索文内のキー単語数，$\alpha(k)$はｉ番目の検索単語の品詞情報，Ｗ(i)はデータベース内の単位文中に出現した類語の中で検索文のキー単語に意味が最も近い単語の評価値（意味の近さを表わす重み係数であり，キー単語 ≡ 同義語０ ≡ 同義語１ ＞ 同義語２ ＞ 類義語の順とした）であり，Ｗａは単位文Ｔの中に出現した検索単語の総数である．

4．3　検索結果の確認

検索結果は可読性が良いように，なるべく短い単位で表示することとし，検索単位との整合を取って文単位とした（文の長さは平均５０文字であり，記事の約１／８の量になる）．また，検索単語と一致した単語には色を付けて表示し、検出･確認を容易にした。文のみで判断できない場合には，前文や記事全体を表示して、適合文か否かの確認が出来るようにした．確認の終了条件は，非適合文が連続して複数個出てきたときに確認を打ち切り，次に，確認の際に有効と思われる情報が得られた場合には，その情報を基に新たな検索条件で再検索することとした．

4．4　検索の高速化

検索は「問い合わせ」と「検索結果からの適合文の検出」とを繰り返すインタラクティブな仕事となり，これを効率よく行うために，検索の高速化が要求される．

高速化の方法としては，文字列の出現する文番

号を索引として予め用意しておき，検索の際には検索文の単語をシソーラス展開した検索単語から文字列索引を通して被検索文の中の候補文を求め，その候補文に対してのみ，フルテキストサーチの検索を行なうこととした．索引は検索洩れを少なくし，かつ索引の容量を小さくできる方法として，データベース内のテキストを言語解析し[12]，解析した単語から２文字連接の索引を作成することとした．このときの索引の容量と検索時間との関係を表４．２に示す．この表から２文字連接索引を用いた検索は，索引を用いずにフルテキストサーチを行った場合より，約２７倍速くできることが分かった．なお，このとき平仮名は索引から除外し，その対策として平仮名以外の索引から得られた候補文に対し，平仮名も含めてサーチすることにした．

表４．２　索引と検索時間の関係

（検索文５例の平均）

文字索引	索引量*(倍)	速度比(倍)
索引なし	——	１．０
２連接(解析ｱﾘ)	０．８９	２７．０
２連接(解析ﾅｼ)	１．４７	——
１文字(参考)	１．１０	１９．１

＊ DB内の文書ﾃﾞｰﾀ(0.6MB)を1.0として比較．

４．５ 検索性能

６５２件の新聞記事を対象に文単位（単位文数：５，０３９文）の検索実験を行なった．検索文の設定では，まず大きな出来事を検索テーマに選んで被験者に提示し，それに関連した内容で知りたい事柄を抽出してもらった．次に，その知りたい事柄が新聞の中でどの様に記述されていたかを想定してもらい，それを検索文として入力することとした．４名の被験者から入手した２６件の検索文の中から，シソーラスによって展開できる単語が存在する７件の検索文について検索実験を行なった．検索結果は優先付けされて出力された単位文の上位１０個と３０個の再現率によって評価した．評価結果を表４．３に示す．表４．３に示す通りシソーラスを用いた場合の再現率が高いことが分かった．また検索結果を参考に人手によって関連単語を追加した結果，再現率がさらに向上した．検索文を自然言語で入力した場合の効果を調べるために，２件の検索文を例に「が格」と「を格」について，検索文と被検索文との格がどの程度一致しているかを調べた。検索結果の上位５０文の結果は，適合文では約４０％，非適合文では約７％が一致しており，格関係の情報も評価に含めると，適合文がさらに上位に出現することが分かった．

表４．３　検索の性能

（検索文７件の平均）

手法＼性能	平均検索単語数（個）	再現率（％）	
		１０位	３０位
キー単語	４．１	１７．９	３１．０
ｼｿｰﾗｽ展開（本手法）	９．４	２４．７	４０．７
人手による単語の追加*	１４．２	３２．７	６０．１

＊：検索結果を見ながら単語を追加した場合．

５　考察

パソコンをベースにした文字読取装置とＡＩワークステーションとを組み合わせて文書情報蓄積検索システムを構築し，文書のコード化と検索実験を行なった．その結果，文書イメージ情報をコード情報に短時間で，しかも容易な操作で変換でき，蓄積した情報を計算機の支援のもとに検索者の意図で効率よく検索･整理できることが分かった．フルテキスト検索は「見出しを付けてなくてもよい」，「検索者の意図によって検索ができる」，「助詞やキーワード検索ではストップワードと見なされる語句も含めて検索できる」等の多くの利点があるが，「表現の異なりを吸収するには更に高度な知識や連想技術が必要なこと」，「定型業務はリレーショナルデータベースの方が有利であ

る」等の事が分かった．また，システムにおける各部の処理についても，

（１）文書データの入力においては，

① システム実現の容易性と汎用化を狙いに，市販のパソコンに認識処理部を付加する形でＯＣＲを実現した．これにより，柔軟性のある装置が実現できた．しかし，ＯＳにシングルタスク処理（MS-DOS）を用いたことが入力の実効速度の上がらない原因になった．

② スキャナの解像度が12本／mmと低かったため濁音や半濁音，類似文字などがリジェクトになり，オペレータの修正作業を増加させる結果となった．

③ オペレータが訂正した情報を文字読取りの性能向上に利用する方法として，文字パターンの特徴を識別辞書に登録する方法[4]があるが，認識結果とオペレータの修正情報を利用する本方式は，認識系などの前段の処理に影響されない点で汎用的な方法といえる．

④ オペレータの確認動作を調べると，類似文字の出現位置で，カーソルが停留したり，後戻りをしていることが多かった．例えば，カタカナの「ニ」，「カ」に対して漢数字の「二」，漢字の「力」などが該当する．この問題に対処するには，類似文字同士の図形的な差分を強調するようにデザインされた文字フォントをシステムに登載するなどの対策が必要と考える．

⑤ オペレータのオピニオンをまとめた結果，図３．２に示した確認･修正の画面レイアウトは，通常の日本語文書の処理には適するが，運用形態Ⅳ（認識結果とイメージとを１対１に対応づけながら確認･修正の作業をする形態）には適さず「疲れる」との意見が多かった．この原因は読取結果とイメージとが画面上で離れていたためと考えられ，これをさらに近づけて表示しなければならない事が分かった．

⑥ 読取結果の確認･修正用として，文字列単位にイメージを表示したために，分離文字や連続する数字などをまとめて確認することができるようになった．

（２） フルテキスト検索においては，

① 検索者の意図をデータベース内の記事文に変換してもらい検索する方法（データベース内の記事文を想起する方法）を採用した．しかし，検索者に対して検索文へ変換するための練習が必要になった．

② そのことから，検索者に検索システムの仕組みを理解させることによって，検索効率が向上することが分かった．

③ 検索者が抽象的な意味の単語で検索文を組立た場合，具体的に記述された記事は検索出来ないことが多い．

④ 一般の利用者に対しては，検索文とデータベースの内容とが，ほぼ等しい文構造で記述･蓄積される分野に適するものと思われる．

⑤ 「問い合わせ」文の入力と「検索結果からの適合文の検出」とを繰り返すことにより，目的とする文を効率よく検索することができる．また，本手法は関連する文が検索結果の上位に出現するため，調査業務などに適する．

⑥ 短い検索文は多義語の影響を受けるが，長文になるに従ってこの問題は少なくなる．

⑦ このシステムに複合語や指示詞の処理を組み込むことにより，さらに検索精度と検索効率が向上する．

などのことが分かった．

6　むすび

文書情報を文字読取装置によってコード情報に変換してデータベース化し，情報が必要になった時点で，関連する情報（単語）を検索文として入力し，単語の共起から知りたい情報を検索することができるフルテキスト形の文書情報蓄積検索システムを試作し，その適用性について検討した．その結果，

（１）文書データの入力においては，

（ａ）試作した文字認識ユニットを用いることにより，汎用のパソコンをベースにした小形で高速な文字読取装置が実現できる．

（ｂ）この文字認識装置は，認識結果とイメージ情報，及び候補文字とを表示する機能を有し，５

号の大きさの単一フォントの文字で印刷された日本語文書を正解率９９％以上，実効入力速度４．５～７．５字／秒（認識速度１４字／秒）でコード化できる．

（ｃ）認識結果とオペレータの修正情報とを訂正規則として利用すれば，リジェクトに正解候補を多く含ませる事ができ，オペレータの修正作業を容易にできる．

（ｄ）リジェクト文字を強調表示し，その文字へ１タッチのキー操作でカーソルを移動させることにより，オペレータの修正速度が向上する．

などのことが分かった．

（２）フルテキスト検索においては，

（ａ）試作した同義語シソーラス（１８万語）によって，検索者による検索文の異なりやデータベース内のテキスト表現の異なりなどによる検索漏れをある程度吸収できる．

（ｂ）検索は「問い合わせ」と「再検索」を繰り返すことによって類似文や関連文を検出することができる．特に，断片的な記憶から目的とする記述を探し出すような業務には有効である．

（ｃ）２連接からなる文字列索引を用意すると，被検索文の文書データの約９割の大きさの索引テーブルが必要になるが，検索の処理速度を約２７倍に高速化することができる．

（ｄ）被検索文を文単位に分割し，検索文との類似性を評価して，類似性の高いものから出力することによって，確認の労力を少なくできる．

などのことが分かった．

残された課題には，文書の入力においては，給紙の自動化やイメージ入力・転送と，認識結果の確認・修正とを並行して動作させ，オペレータの待ち時間を少なくすること．種々の印刷文書から情報の入手を可能にすることがある．また，フルテキストの検索においては，文節単位の同義語表現や言い替え表現を収集すること，意味概念を階層化した分野別シソーラスなどを作成して検索精度をさらに向上させることなどがあり，これらを解決して文書情報蓄積検索システムの適用域を拡大して行く事である．

謝辞　本研究の機会を与えていただいた当研究所川嶋功ヒューマンインタフェース方式研究部長，御意見を戴いたＮＴＴインテリジェントテクノロジー株式会社川谷隆彦部長，小橋史彦部長，当研究グループの鈴木元主幹員，協力戴いた加納英文氏に深謝します．

文献

（１）根岸：“ﾌﾙﾃｷｽﾄ･ﾃﾞｰﾀﾍﾞｰｽの実用化における諸問題”，情基研究会14-1,(1989-7).

（２）国崎，陰山：“手書き漢字OCRの入力速度に関する検討”，昭59信学総全大，1612,(1984).

（３）近藤，多田：“小形高並列ﾌﾟﾛｾｯｻ(LISCAR)の１ﾎﾞｰﾄﾞ化”，平１信学総全大，D-462,(1989).

（４）宮原，山階，山田：“低品質印刷文書読取り用漢字ＯＣＲ”，NTT研実報，Vol.34, NO8,(1985).

（５）宮原，木村，豊田，宮田：“部分ﾊﾟﾀｰﾝによる可変ﾋﾟｯﾁ文書からの文字切出しと認識”，信学論(D-Ⅱ)，J72-D-Ⅱ, 6, (平1-6).

（６）多田，近藤，宮原：“小形高並列ﾌﾟﾛｾｯｻとその文字認識への応用”，信学論(D)，J71-D,8,(63-8).

（７）鈴木，宮原，小橋：“住所認識装置の選択後処理方式”，信学技報 PRU-88 No.492,（1989）.

（８）東京商工会議所：“昭和６３年日本語文書処理（ﾜｰﾌﾟﾛ文書入力）技能検定”，(1988).

（９）杉山，秋山，伊吹，川崎，内田：“自然言語理解に基づく情報検索ｼｽﾃﾑ”，NL研究会，(1986-11).

（10）日比野，渡辺，山田：“LISPﾏｼﾝＥＬＩＳの基本設計”，情処記号処理研資12-15,(1980).

（11）加納，宮原，小橋：“情報検索用シソーラスの試み”，第39情処全大，1G-4,pp.676,(1898).

（12）福永，斎藤：“全文探索と多用な表現”，第39情処全大，1G-7,pp.682,(1898).