文書情報蓄積検索システムの検討

1G-6

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 ）

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管理するため、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が用いられている。しかし、この種のシステムはイメージ情報を直接検索する技術が確立されていないため、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、利用拡大の妨げとなっている。この問題を解決する方法として、我々は文書イメージ情報を文字認識して直接コード化し、情報が必要になった時点で検索・整理するフルテキストデータベース形の文書情報蓄積検索システムを試作したので報告する。

2．システム構成と処理概要

本システムは図1に示すように文書情報の入力蓄積部と検索部とから成る。入力蓄積部では紙面上の文書情報をイメージスキャナによって光電変換した後、文字読取り機能によって文書中の文字をコード化して蓄積する。検索部ではコード化された文字情報をフルテキストの形式を保ちつつ検索の単位となる記事あるいは文章の単位（この単位をTとする）でリスト形式に変換して検索に備える。このとき各文字がどの被検索文に出現したのかをインデクス情報として作成しておき、検索の際に参照して文字列照合の処理を高速化する。検索は利用者が日常使用する問い合わせ文（ここでは検索文と呼ぶ）を入力して、類似文書を探索する。検索結果は類似性の高い順に文書を出力する方式をとる。このような文書情報蓄積検索システムによって、有用な文章や記事をコードの形で蓄積することができるとともに、利用者の意図によって必要な文書や記事を抽出したり、分類や編集をすることができる。

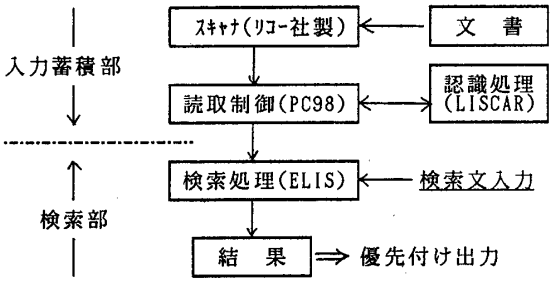

図1．システム構成

A Study of Document Recognition and Retrieval System

Sueharu MIYAHARA, Syunkichi TADA, Kiyosi KABEYA
NTT Human Interface Laboratories

3．入力蓄積部

入力蓄積部は汎用のパーソナルコンピュータ（ＰＣ）に、市販のイメージスキャナと試作した小形高並列プロセッサ（LISCAR）を接続して文書上の文字を認識する機能を実現している[1]。ＬＩＳＣＡＲは２５６個のＰＥ（Processer Element）から成るプログラマブルな並列処理装置であり、ＰＣから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）をダウンローディングするＩＰＬのみによって種々の文字認識機能を実現できる。本システムでは、この装置にページ単位の文書認識の機能を持たせるとともに、読取対象の拡大や入力所要時間の短縮を主な目的に試作した。表1に今回実現した入力蓄積部の主な仕様と機能を示し、以下に文書読取りの処理手順を示す。

① スキャナから入力された文書用紙を走査し、イメージデータを得る（Ａ４判を約２１秒で入力）。

② 画面に表示されたイメージ上で必要な読取領域を指定する（マウス指定）。

③ 読取領域内から文字列を抽出し、文字または文字の一部の組合せパターンを生成した後、正規化してから特徴を抽出し、識別する（LISCARでの処理）[2]。

④ 組合せパターンの識別結果の中から、文字らしさの高いものを文字候補として出力する。

⑤ 文字の大きさや文字列上の印刷位置により相似形の文字（カテゴリ）を区別する。

⑥ 過去にキーボードから入力された修正情報を参照してリジェクトや誤認識の文字を自動訂正する[3]。

⑦ 認識結果とイメージとを表示して、オペレータの確認修正の作業を補助する。

⑧ 読取結果をファイルに格納する。

表1．　入力蓄積部の仕様と機能

項　目	仕　様　と　機　能
読取文書	用紙はＡ４判以下（ｽｷｬﾅは12画素/mm）
	文字フォントは単一
	読取領域はﾏｳｽによるｵﾍﾟﾚｰﾀ指定
	８０行／頁，１２８文字／行以下 不定ピッチ文書の読取り可能
読取結果	MS-DOS標準ファイルフォーマット出力

4．検索部

検索部にはＥＬＩＳ[4]を用いた。検索の処理は検索文の中から単語（キー単語）を切出した後、シソーラス[5]によってキー単語に対する同義語や類義語を展開する。キー単語も含めこれらの語を総称して以下類語と称する。展開した類語とデータベース（ＤＢ）内の

語句との照合を行って、以下にその定義を示す類似性の高い単位文から出力する。

（１） 類似性の計算： 類似性の評価は、ＤＢ内の単位文Ｔの中に類語が存在するキー単語の種類数（MAX:Ｎ個）が多くて、かつ算出式〔１〕の評価値Ｒが大きいものを類似文書とした。

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

ここで、Ｎは検索文内のキー単語数、 Ｗ(i)はＤＢ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、および重み係数はキー単語 ＞ 同義語 ＞ 類義語の順とした）であり、Ｗａは単位文Ｔの中に出現した類語総数である。シソーラスの一例を表２に示す。

（２） 処理の高速化： 単語をキーにして、データベース内の全ての文字をサーチすると多くの処理時間を必要とする。たとえば本システムで新聞９００記事（約１Mﾊﾞｲﾄ）の文書データに対し、４個のキー単語から成る検索文で検索すると約１０秒を要する。そこで文字単位のインデクスを用意して候補文章を絞り、候補文章に対してのみ詳細に解析する方式を採用した。

表２． シソーラス展開の例

キー単語	同 義 語	類 義 語
災 害	災禍，被害，災難	天災，人災，…
誘 拐	誘かい，拉致，ら致，人さらい，かどわかす	営利誘拐
会 議	打ち合せ，打合せ，…… ｼﾝﾎﾟｼﾞｭｰﾑ，ｼﾝﾎﾟｼﾞｳﾑ，…	討議会，… 研究会，…

５．評価

入力蓄積部と検索部とは並行して試作した。そのため今回は各部の基本となる評価データを用いて個別に評価した。

（１） 入力蓄積部の性能： 表３に示すように５号の明朝体活字でＡ４判の用紙に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、読取速度１４字／秒、正解率９８％（第一位正解率９９．５６％）の値を得た。

表３． 入力蓄積部の性能

項 目	性 能		条 件
読取速度（／秒）	平均１４文字		27行／頁，850文字／頁 イメージ処理を含む
読取性能（％）	正 解	98.05	５号明朝体PTO活字
	ﾘｼﾞｪｸﾄ	1･79	単一フォント用辞書
	誤 り	0.16	読取対象：3,174字種

（２） 検索部の性能： 検索の一般的な例として、新聞記事を対象に文章単位（文章数：５０３９件）の検索実験を行なった。該当文章の累積出現率（人手による単語追加の場合の１００位内の出現個数を基準）について、上位１０個、および３０個の文章について調べた結果、表４に示す通りシソーラスを用いた場合の累積出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シソーラスや関連単語を追加・拡張した場合に累積出現率がさらに向上した。処理速度はＤＢ内での文字の出現の仕方に大きく依存するが、文字単位のインデックスを設けることにより、フルテキストの逐次サーチに比較して平均５～１０倍程度高速になった。

（３） 考察： ① 入力蓄積部は汎用化を狙ってＰＣに認識処理系を追加する方式を採った。そのため入力処理のスループット低下の原因となった。またスキャナの解像度が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った。 ② 本検索方式は関連する文章も上位に出現するため、調査業務などに適するものと考える。また複合語や指示詞の処理を組み入れることにより、さらに精度が向上するものと思われる。

表４． 検索実験の結果

（検索文５件の平均）

性能 / 手法	平均検索単語数（個）	累積出現率（％）	
		１０位まで	３０位まで
キー単語	４．０	２０．０	２７．７
ｼｿｰﾗｽ展開（本手法）	７．３	４３．１	５３．８
人手による単語の追加	１３．５	５８．５	８４．６

＊：検索結果を見ながら単語を追加した場合。

６．むすび

文書情報を文書読取り機能によってコード情報に変換してデータベース化し、検索者の意図によって必要な情報を自由に検索できるシステムを構築した。文書の入力蓄積と検索との実験によって、ＰＣベースのコンパクトなシステムが実現できることが分かった。今後は業務データを用いてシステムの評価行うとともに、読取り不能文字の訂正や書誌情報の認識などの処理方法について検討していく。また入力対象文書の拡大、文章解析の高度化、処理の高速化などシステム全体の高度化についても検討していく。

謝辞 本研究の機会を与えて戴いた当研究所・川嶋言語メディア研究部長、御意見を戴いた川谷主幹員、小橋主幹員、協力戴いた加納英文氏に深謝します。

文献 １）多田他：“LISCARによるﾌﾟﾛｸﾞﾗﾏ…”，平1信学総全大,D-460(1989).
２）宮原他：“部分ﾊﾟﾀｰﾝによる可変 ……”，信学論(D-Ⅱ),J72-D-Ⅱ,(平1-6).
３）鈴木他：“住所認識装置の選択後処理方式”，PRU88-154，IE88-133(1989).
４）日比野他：“LISPﾏｼﾝELISの基本設計”，情処記号処理研資,12-15（1980).
５）加納他：“情報検索用シソーラスの試み”，平１情処全大後期(第３９回).