日本語テキストリーダを利用した文書検索インタフェース

8G−1

田野崎康雄[1]　中本幸夫[2]　岩井　勇[1]

[1](株)東芝　情報処理・機器技術研究所　[2]東芝コンピュータエンジニアリング株式会社

1．はじめに

ワープロなどが普及し、文書が電子化された現在でも、流通しているほとんどのドキュメントは印刷物である。特にコード化が困難なものを多く含む場合はこの傾向が著しい。現在、このような印刷物は、そのイメージデータを電子的にファイリングすることが可能となっているが、大量のデータベースに対しては、検索をすることすら困難である。しかし、このデータにＯＣＲ処理を施すことにより、コード列を参照した検索が可能になり、さらに、イメージデータに対していくつかの情報を付加し、ユーザとインタラクティヴに反応するより活性化したデータ（ActivePaper)を自動作成することができる。

ActivePaper に持たせる諸機能は検討中であるが、本発表では、技術論文の検索システム中でのActivePaperの応用例として、検索キーワードの在処を明示する機能について紹介する。

2．システムの基本概念

◎フルドキュメントサーチ

本システムは、ＯＣＲによりテキストデータを抽出し、印刷物として流通している大量の文書資源の検索を実現することをめざしているが、同時に、従来のテキストベースの検索システムにおいて非テキスト情報の処理機能を強化したものつまりフルドキュメントサーチシステムであるという位置付けも可能である。

従来よりテキストベースの検索システムは数多く存在しているが、数式、図表などの非テキスト情報の扱いが十分ではなかった。候補の絞り込みの際のブラウジングの段階でも、これらの非テキスト情報、レイアウト構造など視覚に訴える情報は重要である。

また、技術論文などのようなテキストの含有率の高い文書でなくてもドキュメントは、その内部に、多くの場合、何らかのテキストを含む。このテキストを利用して検索が行なわれる。地図の場合はそのタイトル、写真や図などからなる文書でも、これらに付加されている説明文を手掛かりにした検索への応用が考えられる。

◎ActivePaper の概念

紙メタファを実現

ActivePaper の基本的な性質は、主に表示にあたって「紙」の持つ諸性質を保っていることである（紙メタファ）[1]。そのために、ページイメージの表示にあたって多値画像で表現するなどの工夫を行ない、高品質表示・高速表示を実現している。

イメージデータとコードデータの融合

後に述べるように、ＯＣＲを用いることにより、まず、イメージデータとコードデータの融合が行なわれる。さらに、さらに、コードデータに付随した様々のデータを利用して、新たな機能を持たせる。

ここではActivePaper を作成する段階で、できるかぎり人手を介さず自動的にデータベースを構築していくことを目指す。大量の印刷物が存在している現在、データベース作成の自動化は重要である。

3．テキストリーダによるデータ生成

印刷文書イメージから文字コードを抽出するために当社の日本語テキストリーダを用いた[2]。コードデータを入力する手段としてのみ日本語テキストリーダを用いることも有効である。この方式によりページ中の認識可能なすべての文字を抽出できるため、図表の説明文、脚注の内部の語句も検索対象となる。図や表の内部の文字も利用可能である。さらに、各ページのヘッダ部あるいはフッタ部に存在する文献情報を利用することができる。

現在の一般的な検索システムでは、タイトルや著者名、発行年月日などの文献情報は本体とは別に管理する必要があるが、本システムでは、統一的にフルテキストサーチを施すことにより、このような特別の管理を行なう必要がなくなっている。

テキストリーダでの処理により、文字コードとともに、各文字のページ中での、座標位置、文字サイズ等が得られる。そのほかページ内での領域情報なども得られる。

文字ごとに各種の情報を持つテーブルをＣＬＴ(Character Location Table)と呼ぶ。各文字データはテキスト中に並んでいる順序で格納されている。（図１）

現在、１ページのイメージデータに対してひとつのＣＬＴが対応している。ページイメージとＣＬＴのアクセスの方向には、以下の２通りのものが考えられる。

Document Retrieval Interface using Japanese-Text-Reader

Yasuo TANOSAKI[1], Yukio NAKAMOTO[2], Isamu IWAI[1]

[1] Information Systems Engineering Lab. TOSHIBA Corp.

[2] TOSHIBA Computer Engineering Corp.

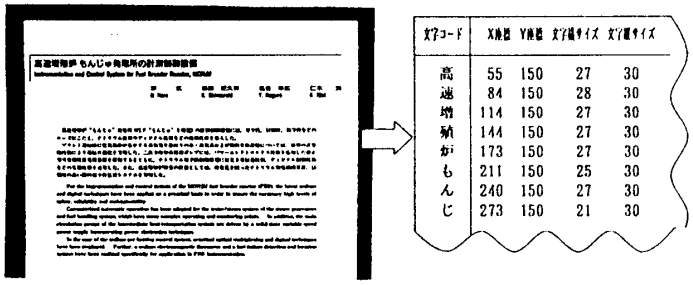

図１　ページイメージとＣＬＴとの対応

ＣＬＴ　→　イメージ

入力キーワードの文字列をＣＬＴ中よりサーチし、検索対象語句の位置にマークを付加する。文書間のマクロな検索系における検索用キーワードの表示、文書内でのミクロなレベルでの文字列サーチの際に利用できる。

イメージ　→　ＣＬＴ

ページイメージ上で文字列を指定することにより、テキストを利用するものである。

(1) ページイメージに対応する文字列をキーワードとして、再検索を行なう。ソフトリンクのハイパーテキストを実現する。

(2) イメージ上の指定した範囲に含まれるテキストを再利用する。

4．試作システム

社内技術情報紙（東芝レビュー）を対象にしたページイメージ検索システムを開発中である。現在、検索の単位は、ActivePaper ４～５ページからなる記事単位であり、この単位ごとに検索インデックスを持つ。検索結果は、高速に多階調表示される。検索システムのページめくり機能と連動しており、ページがめくられるたびにＣＬＴ内でキーワードのサーチが行なわれ、ページイメージ中のキーワードの位置が明示される。

１ページのActivePaper が表示されるまでの処理の流れを図２に示す。

キーワードをマーキングする段階では、ＣＬＴを参照し、ここで得られた各文字の矩形領域に対応するページイメージ中の対応する部分のドット情報を直接処理している。この際にも紙メタファの基本を守り、多階調表示された文字イメージを損なわないようにして「紙に描かれたラインマーカ」のメタファを実現している。

なお、検索のエンジン部では表記のゆれを吸収している。

［例］　インターフェイス　→　インタフェース
　　コンピューター　→　コンピュータ

しかし、現在、ＣＬＴ内のサーチを行なう際には、この処理を行なっていない。原文およびＣＬＴ中に、たとえば「インターフェイス」と格納されている場合でも、表記のゆれを吸収した際に用いたものと同じ異表記辞書を用いて、この語句へのマーク付けを可能にすることを検討している。

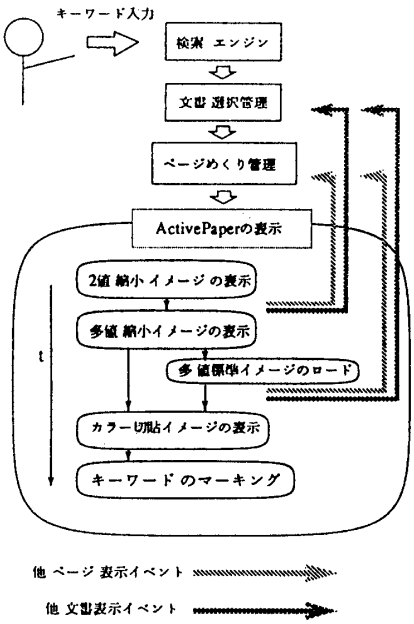

図２　ActivePaper 表示までの処理の流れ

5．まとめ

本発表では、コードデータ・イメージデータと各種の操作とが統合化されたActivePaper の応用の一段階として、入力キーワードに対応するページイメージ上の位置にマーカを付加する処理を実用的な検索システム中で実現したことについて述べた。

ActivePaper を用いることにより、現在の文書処理環境での(1) データの入力、(2) 検索、(3) 表示、(4) 利用の各段階における様々な問題が解決できる。

また、ActivePaper はデータの持ち方と操作手続きを記述したものであるため、ひとつのオブジェクトである。従来より文書の構造について、オブジェクト指向データベースの観点からモデル化することが検討されてきたが[3]、ActivePaper を設計する目的であらためてオブジェクト指向の概念を取り入れて行くことが重要である。

［参考文献］
[1]田野崎：“文書検索システムにおける紙メタファインタフェースの実現”，情処第45全大，4S-1 (1992)
[2]有吉，他：“変形パターンの自動生成によるマルチフォント印刷漢字認識”，信学全大，p1465 (1987)
[3]Wolek, D. and Kim, W: “An Object-Oriented Approach to Multimedia Databases”, ACM SIGMOD '86, pp.310-325 (1986)