4S−1　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

田野崎康雄
(株)東芝　情報処理・機器技術研究所

1. はじめに

現在、フルテキストデータベースを対象とした文書検索システムを開発中である。検索結果を表示する際に、ユーザが得られるデータ量が限られていた時代とは異なり、高速に大量のデータを扱えるようになると、新たな表示形態が必要になってきた。そこで注目されているのが、人間にとって馴染みのある紙メタファに基づく表示インタフェースである[1]。さらに単にページイメージを画面上に再現するのにとどまらず、電子化されたことによる様々なメリットを有効に生かすことも必要である。

本稿ではシステムを構築していく上での基本思想とともに、試作したシステムにおける紙メタファインタフェースのメカニズム、オブジェクト指向に基づくデータの管理・操作モデルについて述べる。

2. 表示系の基本思想

2.1　紙メディアの特徴と電子化文書の特徴

紙をベースとするメディアの特徴は、単位面積・単位時間あたりの情報の表示量が膨大であるということである。この特徴から派生する性質として以下のものがある。

(a) 瞬時に紙面全体をみわたすことが可能
(b) 多様な文字・図形表現が可能
(d) 複数ページの高速連続表示（いわゆるペラペラめくり）が可能

また、我々は長年紙メディアと慣れ親しんでいるため、操作が容易であるという特徴もある。

紙メディアにはなく、電子化された文書が持つ特徴としては、

・文書間・文書内での単語・文字などのサーチが可能
・ハイパーテキストによる多様な文書ブラウジングが実現できる。
・文書データの再利用・変形が容易

などがあげられる。

紙メタファのインタフェースを設計するにあたって上記の特徴に注目していく。

2.2 実現へのアプローチ

紙メタファの表示系を実現するひとつのアプローチとして、既存の印刷文書のページイメージデータをスキャナで入力して利用するという方式を用いた。この方式によると、コード化が困難な文字・記号の扱いが容易となる。また、豊富な文書資源をデータベース化できるという大きな利点もある。

さらに電子メディアとして扱うためには、テキストのコード情報を入力することが不可欠であるが、現在、ＯＣＲ技術が進歩しており、すでに日本語テキストリーダが商品化されている。スキャナで入力した文書のイメージデータから、テキストコードデータへの変換を実用的なレベルで行うことが可能になっている。テキストリーダによる文字認識の結果として、各文字の位置情報も得られ、より積極的なアクセスも可能になる。こうして入力したテキストデータを用いて、現在、検索用インデックステーブルを作成している。システムの構成を図1に示す。

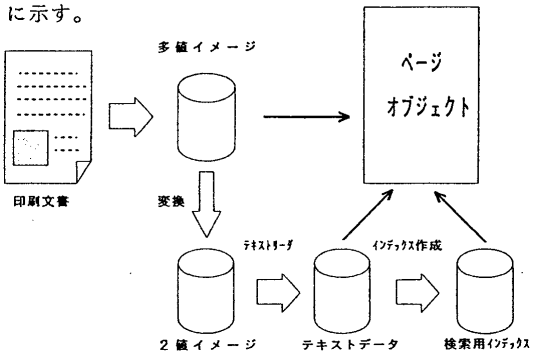

図1　システム構成

2.3 実現上の特徴

(1) ページイメージデータの階調表示

モノクロの紙イメージを２値画像としてディスプレイ上に表示した場合、紙メディアと同程度の表示品質を実現するためには、より多くのドット構成を要する。この方式によると漢字・図形を含む文書のページイメージを表示する際に、一般的な解像度のワークステーションの画面ではＡ４サイズ１ページを部分的にしか表示できない。このような場合、既存のシステムではスクロールなどの操作を行っているが、紙メディアの持つ「瞬時に全

Paper Metaphor Interface in Document Retrieval System
YASUO Tanosaki
TOSHIBA Information Systems Engineering Laboratory

体をみわたせる」という性質が失われる。そのためページイメージを多値画像として表示する方法を用いた。少ないドット数からなる文字であっても、陰影表現を行なうことにより可読性が増し、狭い面積により多くの文字・図形を高品質で表示できる。

(2) ページイメージデータの分割格納・段階的表示

上記のような多値表示を実現するためにページあたりのデータ量は増大する。このデータを操作するために多くの時間を要し、紙メディアがもつ「複数ページの高速連続表示が可能」という性質が失われる。そこでデータを以下の二つの方法で分割し、分割されたデータを順に合成しながら表示する。はじめは粗い表示であるが時間とともにより鮮明になる。各ページイメージの表示途中でもユーザの要求があれば、他ページの表示を行なうことができる。

①解像度に関する分割・・・隣接する４ドットのデータを１ドットずつ分割する。

②階調に関する分割・・・１ドットは16階調で、４ビットの情報を持つ。この各ビットの情報を分割する。

(3) データの圧縮

現在データは光磁気ディスク装置に格納しているが、ディスクからメモリにデータをロードする時間を短縮するために、またディスクスペースを有効に活用するためにもデータの圧縮を行った。分割前の150dpiのＡ４版のページイメージデータは 1.08Mバイトであるが、これをラン長法により約400Kバイトに圧縮している。

3. ハイパーテキストと紙インターフェイスの融合

3.1 ページオブジェクトの管理

システムにおいてデータの終端ノードは、先に示した紙メタファを実現するページオブジェクトや画像オブジェクトなどである。これらを管理するための非終端ノードをつかさどるオブジェクトのクラスとして Catalogというクラスを定義している。

本検索システムでは、必要な文書をアクセスする際に、キーワードを用いた検索と、階層構造を辿っていくという目次検索の方式のいずれも可能になっているが、検索によって絞り込まれた文書候補の「一覧表」と一般的な「目次」とは同じ Catalogクラスのインスタンスとしている。

「目次」オブジェクトはシステム作成時にあらかじめリンク構造が定まっているのに対し、「一覧表」オブジェクトは、実行時に新たなオブジェクトおよびリンクが自動生成されるため“アクティヴオブジェクト”[2] の性質を持っている。「一覧表」オブジェクトの生成過程を図２に示す。図中で文書オブジェクトはページオブジェクトを束ねるための非終端ノードである。

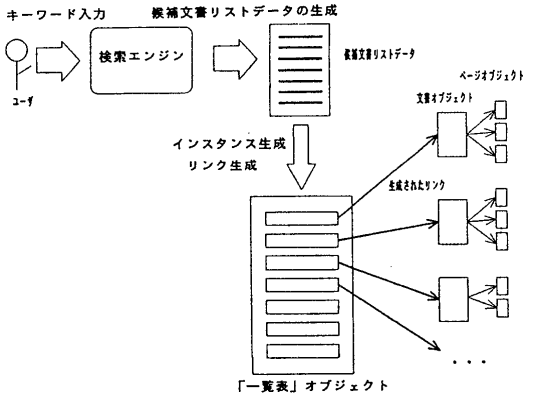

図２　一覧表での Catalogオブジェクトの自動生成

3.2 ページオブジェクトと他のメディアとの結合

カラー画像を含んでいるページについては、ページオブジェクト（グレースケール表示）から画像オブジェクトに対してリンクが張られている。画像オブジェクトは、その生成・消滅・移動に関して親のページオブジェクトに従属する。ひとつの画像オブジェクトを複数のオブジェクトから参照できるのと同時に、ページオブジェクトはグレースケール表示であり、これにカラー表示のオブジェクトを分けて管理することにより、ページ全体をカラー表示する場合に比べてデータ量を減らすことができる。

さらに、スキャナ入力したページ中から参照されている外部の画像・文書オブジェクトなどに対してもリンクづけ（一般的なハイパーリンク）を行なうための枠組みを用意してある。

4. おわりに

ハイパーテキストが提唱されて久しいが、今だ研究段階に止まっているように思われる。紙メディアの特徴を備え、これに電子メディアの長所を加味していくというアプローチが重要であろう。本稿では、その基礎となる部分についてのみ述べた。

今後、コードデータとイメージデータを融合させ、ユーザとの、よりインタラクティヴな操作が可能なインタフェースを実現していく予定である。

［参考文献］

[1] 新井、他：“ページめくり機能を持ったｳｨﾝﾄﾞｳｲﾝﾀﾌｪｰｽ：BookWindow”，情報処理学会研究報告 91-HI-36

[2] 科学技術庁科学技術振興局：“創造的開発支援のための自己組織型情報ベースシステムの構築に関する調査成果報告書”，平成３年３月