4H－2

トランスメディア·マシンにおける Index Browserの実現法

割石　浩一　　田中　譲
北海道大学　工学部

1.はじめに

トランスメディア·マシンは、あらゆる文書をそれに付随する図·写真·表も一緒にすべて画像のままの状態で蓄積し管理するシステムを目指している。蓄積された文書画像に対して、変更や再利用をするための文書編集機能と必要なデータの検索を行うためのキーワード検索機能が既に実現されている。

現在このシステムにおける文書編集機能として、スクリーンエディタが実現され、データベース機能の第一段階としてキーワード検索機能が実現されている。このキーワード検索機能は、各文字を認識することにより実現されているのではない。各文字は、前処理としてYes/Noで判定できる特徴量2つにより、2bitの不完全コードに符号化される。キーワード検索機能は、キーワードを不完全コードに変換し、不完全コード列の出現位置を前処理により得られた不完全コード列から文字列検索アルゴリズムにより検索することで実現されている。

今回は、新たなデータベース機能として、大量に蓄積された文書の中からデータ検索を行うためのガイドをおこなう、インデックス·ブラウザーの実現について報告する。

2.インデックス·ブラウザー

現在、文書を文書画像のまま蓄積し管理する電子ファイルシステムが普及しつつある。この電子ファイルシステムに記憶されている全ての文書がコード化された文字だけで構成されていれば、全ての文書のスキャンによりインデックスを自動的に与えることが可能である。しかし、図書館などで保管されているマイクロフィルムにより管理されている書籍やトランスメディア·マシンのような電子ファイルシステムに管理されている書籍はすべて画像として各ページ毎に管理されている。このような、文書画像によるデータの管理は、現在各文書に対して人間の手作業により与えられたインデックスにより管理されている。文書データの利用者は、与えられたインデックスをもとにデータを検索する。

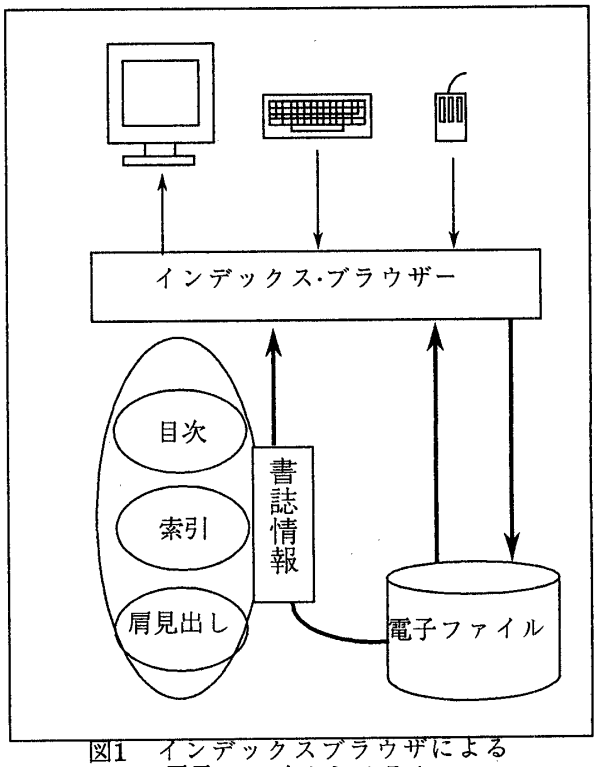

図1　インデックスブラウザによる電子ファイルシステム

書籍などの印刷物はそれ自体の中に、データベースのインデックスに対応する目次や索引の索引情報を持つ。さらに、各ページにはページ番号がふられ、章の初めにはタイトルがつけられている。これらの索引情報は、文書画像として保持されているために、現在の電子ファイルシステムでは、有効に利用されていない。文書画像のもつ書誌情報を利用するためには、目次などの項目とページ番号の対応や各ページにおけるページ番号と文書画像の電子ファイルシステム上の記憶位置の対応を取ることが必要である。目次の機能を利用するために、目次の各行を切り出し、一枚のメニューとして表示しメニュー選択可能な形で

Implementing Index Browser of Transmedia Machine.
Hirokazu WARIISHI,Yuzuru TANAKA
Hokkaido Univ.

表示を行えばよい。(図2)また、辞書や索引などにおける肩見出しを用いると、各ページ毎の肩見出しを画像として切り出し肩見出しを階層的に管理することで、順序だったキーワードの検索は容易に行える。我々は、文書画像の項目と文書の各ページの対応を目次の項目に対するページ番号の数字の認識だけに限った方法で書籍の書誌情報を利用することが出来るシステムを実現し、このシステムをインデックス･ブラウザーと呼ぶ。

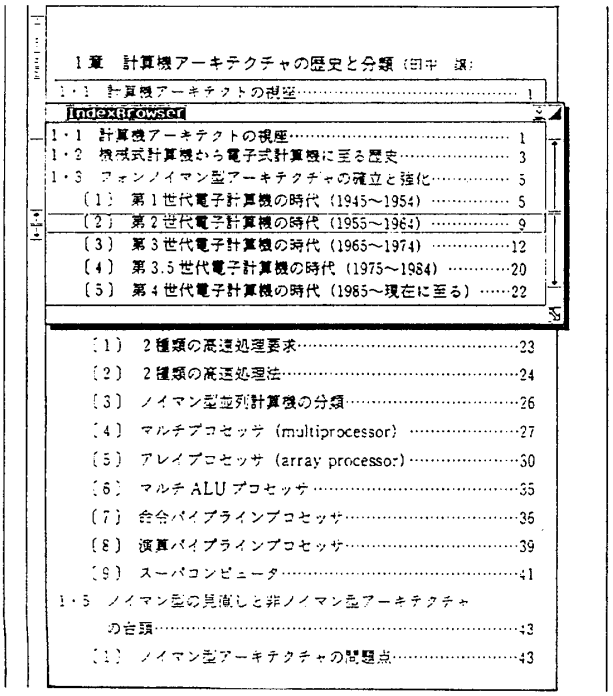

図2 インデックスブラウザーの例

3.インデックス･ブラウザーの実現

インデックス･ブラウザーを実現するには、まず処理する目次をビットイメージの画像として入力する。入力された画像は、利用者によりインデックス･ブラウザーの目次のメニューとして利用したい範囲をマウスを使って指示する。指示された範囲に対し、前処理として行の切り出しを行う。行の切り出しによって得られた、各行領域は画像データのままポップアップメニューとして表示される。利用者はマウスを使い項目の行をを選択する。システムは、選択された行の数字領域を切り出す。切り出された数字領域は、数字の認識を行うことでページ番号が得られる。

行の切り出し

行の切り出しは、目次として扱う領域の画素をを垂直軸に投射することにより得られる画素の度数を元に行の上端と下端を求める。

数字領域の切り出し

各項目のページ番号が行末に存在することから、数字領域の切り出しは、各行毎に行末から左に向かい行われる。数字領域は、行の画素を水平軸に投射し得られる画素の度数を用い、数字領域を切り出す。(図3)

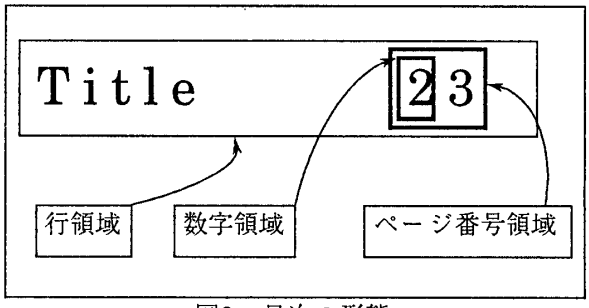

図3 目次の形態

数字の認識

数字の認識はガイドドット付ゾンデ法を元に拡張した。ガイドドット付ゾンデ法では定点が必ず文字の背景となるよう制限をつけた手書き数字に対する認識方式である。直立した数字に限定した認識では、定点を用い解析する背景は文字の中心軸上に存在する。従って、このゾンデ法の定点は、文字の中心軸をスキャンし、文字を構成する画素に囲まれた空白の中心に定点を定められる。

4.まとめ

書籍の情報としての目次と索引を画像として読み込み、行の識別と数字に限定した文字認識を行うことで、文書構造に従ったデータ検索のガイドとなるインデックス･ブラウザーが実現できることを示した。今後は、数字認識の問題点として直立した字体のみに限定している点を解消するとともに、各ページの入力時に自動的にページ番号と肩見出しの切り出しを行う文書構造の認識を行い、更に充実し、利用しやすい機能を充実して行きたい。

参考文献

鳥居、田中:「トランスメディア･マシンにおけるデータベース機能の実現法」情報処理学会第35回全国大会6Bb-7

鳥居、田中:「トランスメディア･マシンにおける英単語検索(手法の提案と実現例)」情報処理学会第37回全国大会2R-4

橋本新一郎:「文字認識概論」