イメージPDFを検索対象とする全文検索システムの運用

佐藤　郁・坂本正秀・大西みどり
（大分県農業技術センター）

Kaoru Sato, Masahide Sakamoto and Midori Onishi :
Employment of the full-text search system by Image PDF

1. はじめに

大分県農業技術センターでは，試験研究成果の情報を公表しており，その一部は，ホームページ上で公開している。

しかし，多くの情報は印刷物として存在するため，全文を対象とした内容の検索や原文の利用ができなかった。そこで，紙文書の文献についてもインターネット上で検索できる全文検索システムを構築したので報告する。

2. データファイルの作成

パソコンで作成された文書であればAdobe社のPDF（Portable Document Format）に変換することで，検索が可能な原文と同じイメージを再現できる。しかも，OSに関係なく画面表示や印刷ができるという特徴がある。

一方，紙文書の文献情報を利用するには，文献を画像情報に変換する方法と，文献をOCRソフト等で文字情報に加工する方法がある。

そこで，データファイルの作成には，それぞれの方法の特徴を組み合わせ，文献の原文イメージを表側の情報として保持し，また，文字情報を裏側の情報として保持することが可能な，イメージPDF（透明テキスト）形式を使用した。

第1図にデータファイルの作成方法を示す。イメージスキャナを使用し，解像度は400dpi，色階調は白黒2値の条件で原稿を読取り，TIFF形式の画像ファイルとして保存する。次に日本語OCRソフトを使用し，この画像ファイルの文字認識を行いテキスト形式のデータを抽出する。このTIFF形式の画像ファイルとテキスト形式のデータを融合し，イメージPDF（透明テキスト）形式として保存する。

3. 全文検索システムの改良

全文検索システムは，文書管理，文書フィルタ，インデクサ，検索エンジン，検索クライアントの各機能を持つプログラムの集合体である。日本語全文検索システムとしては，フリーソフトであるNamazuが多くのWEBサイトで使用されている。

大分県農業技術センターでは，ホームページ内の情報を検索するためにNamazu 1.3を使用してきたが，検索できるものは，HTML等のテキスト形式のファイルだけであった。しかし，Namazu 2.0以降は，文書フィルタの種類が増えており，PDF形式のファイルを検索対象とする事ができるようになったので，Namazu 2.0.12を採用した。ただしNamazuでは，事前に検索用インデックスファイルを作成する必要があり，PDF形式のファイルから検索に必要なテキスト情報を抽出するため，フリーソフトであるXpdfに含まれるpdftotext, pdfinfoというプログラムを使用した。

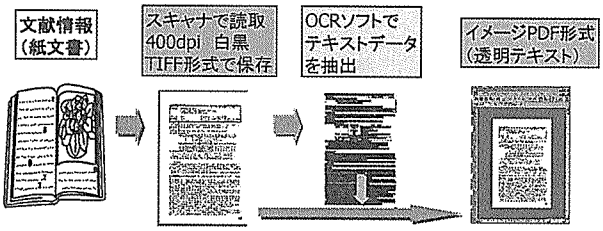
4. 全文検索システムの運用と課題

全文検索システムは，WEBサーバ上で運用しており，インターネット経由でブラウザを利用して検索することができる。ホームページ上の検索ボタンをクリックすると検索式の入力画面が表示される。検索式では複数の単語をスペースで区切って記述するとAND検索ができる。第2図に検索結果の表示例を示す。

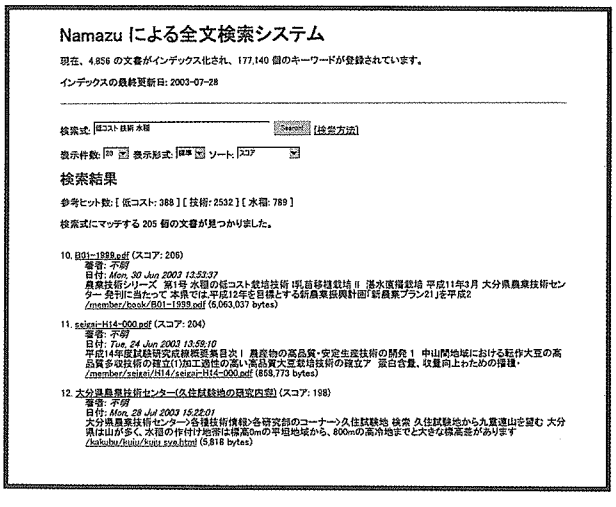
現在は，著作権や作業性の問題があり，この全文検索システムで利用できる文献は，大分県（農業技術センターおよび関係機関）が発行したものとなっている。また，利用者も，ホームページで公開したもの以外は関係機関職員に制限している。

今後は，データファイル作成手順を改善し，効率的に文献登録を行う必要がある。

また，データファイルの容量を小さくし，利用者が扱いやすくする必要がある。



第1図　データファイルの作成手順



第2図　全文検索結果表示例