解 説

学術論文誌データベース化プロトタイプシステムの構築

佐藤弘行^{*1}, 菊池俊一^{*2}, 佐藤雅之^{*3}, 小室信三^{*3}
佐藤道弘^{*4}, 山岡正輝^{*5}, 岩根和巳^{*5}, 岩城 修^{*5}

[著者抄録] 科学技術雑誌論文のページ画像をイメージスキャナで読み込んだ後, レイアウト解析モデルの知識ベースに基づき, 標題, 著者, 所属機関, 抄録, 本文等の論理属性を自動的に識別し, 図表データを含めたマルチメディア情報ベースを簡易に作成することが可能なプロトタイプシステムを構築し, 性能評価を行った。その結果, 運用上現実的な認識率および処理速度で実現可能であることが分かった。

[著者付与キーワード] 学術文献, データベース, フルテキスト, マルチメディア, SGML

Commentary article

Development of a prototype system for automatic database construction from academic journals

SATO Hiroyuki, KIKUCHI Shun'ichi, SATO Masayuki, KOMURO Shinzo
SATO Michihiro, YAMAOKA Masaki, IWANE Kazumi, IWAKI Osamu

[Author Abstract] A prototype system was developed for automatically constructing a database from scientific and technical journal articles in print. When a page of a printed technical paper is scanned with an image scanner, the system distinguishes between tables, graphs, and texts. And then in the texts, the logical attributes such as the title, the authors, the author's affiliations, the abstract, are recognized by using the page layout model. Lastly, the texts are processed with an OCR. The performance measurements show that the system has a sufficiently high recognition rate and processing speed and can come into operation.

[Keywords by Author] academic journal, database, full text, multimedia, SGML

-
- * 1 科学技術振興事業団 研究基盤情報部 Tel. 03(5214)8471
 - * 2 科学技術振興事業団 科学技術情報事業本部 情報システム開発部 Tel. 03(5214)8429
 - * 3 科学技術振興事業団 科学技術情報事業本部 文献情報部 Tel. 03(5214)8432
 - * 1~3 (〒102 千代田区四番町5-3) JST (5-3, Yonbancho, Chiyoda-ku, 102)
 - * 4 NTT データ通信㈱ 公共システム事業本部 (〒135 江東区豊洲3丁目3-3)
NTT Data Corporation (3-3, Toyosu 3-chome, Koto-ku, 135)
 - * 5 NTT データ通信㈱ 情報科学研究所 (〒210 川崎市幸区堀川町66-2興和川崎西口ビル)
Tel. 044(548)4605 NTT Data Corporation (Kowa Kawasaki Nishi-guchi Bldg., 66-2, Horikawa-cho, Saiwai-ku, Kawasaki-shi, 210)

1. はじめに

本研究は、平成3年度から平成7年度までの5年間にわたり科学技術振興調整費により行われた「創造的研究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研究プロジェクト」の一環として実施したものである。

本プロジェクトは、生物に見られる同一種類の細胞が集まり器官を作り、それらの器官が集まって生物の一個体を作り上げるといった「自己組織化」の概念を、データベースあるいは情報ベース構築・利用技術に導入しようという試みである。プロジェクトの参加機関は、大学、国立研究機関、特殊法人、民間企業で構築され、三つの研究グループ（情報の資源化方式に関する研究、自己組織型情報ベースのシステム基礎技術、情報ベースとその高度利用）に分かれて活動を行った。

我々は、このうち情報の資源化方式に関する研究グループに属し、印刷出版されている学術論文誌をスキャナで読み込み、論文中の標題や著者名、抄録、本文、参考文献といった論理属性を自動的に認識抽出してデータベース化する技術に関して取り組んだ。

本論文では、それらに関して構築したプロトタイプシステムと、その性能評価結果について報告する。

2. システム概要

2.1 ハードウェア構成

プロトタイプシステムは、図1に示すように、高速スキャナに接続している入力・蓄積サーバと、認識・変換サーバから構成されている。このようなクライアント・サーバ構成とすることにより、以下の効果が得られる。

- (1) 高速スキャナや大容量外部記憶装置といった高価なハードウェア資源を共有することにより、有効利用を図ることができる。
- (2) システム全体のスループットに対して影響の大きい認識・変換サーバの数を増設することが

できる。

2.2 ソフトウェア構成

図2に示すように、プロトタイプシステムのソフトウェアコンポーネントは、GUIレイヤー、認識レイヤー、サポートレイヤー、OSレイヤーの四つのレイヤーに分類されている。また、1) 分散処理環境に対応、2) 機能単位にモジュール化、3) 高移植性、といった特徴を有している。

3. 機能概要

プロトタイプシステムの処理の概要を図3により説明する。まず、スキャナで読み込んだ学術論文誌の画像データに対してレイアウト解析を施し、標題、著者などの論理属性のついた領域を、入力・蓄積サーバで作成する。次に、認識・変換サーバで、解析結果領域の文字認識を行い、その結果をSGML (Standard Generalized Markup Language)^{2)~7)}文書として保存する。さらに、これをHTML (Hyper Text Markup Language) 文書に変換する。

作成されたHTML文書には、論文中の図や表も画像データとして蓄積されており、図4のようにWWW (World Wide Web) ブラウザにテキストデータや画像データを表示させることもできる。

以下では、図5をもとに、各サーバが行う処理について述べる。

図1 プロトタイプシステムのハードウェア構成

図2 プロトタイプシステムのソフトウェア構成

図3 処理の概要

図4 プロトタイプシステムの利用イメージ

図5 処理の流れ

3.1 入力・蓄積サーバ

入力・蓄積サーバでは、イメージ読み込み、論文識別、入力画像の正規化といった処理を行う。

論文識別では、ユーザインタフェースを通して、ページ画像の論文種別をオペレータが指定する。本プロトタイプシステムで指定できる論文種別としては、(社)電子情報通信学会、(社)情報処理学会のご協力により電子情報通信学会論文誌、情報処理学会論文誌の2誌とした。

3.2 認識・変換サーバ

認識・変換サーバでは、入力蓄積されたページ画像のレイアウト解析、OCR処理、文字認識後処理、SGML文書化、HTML文書化等の処理を行う。対象論文誌のレイアウト構成例を、図6、図7、図8に示す。

3.2.1 レイアウト解析

(1) 基本矩形の抽出

正規化された入力画像データから、隣接する黒

画素を包含する最小矩形領域（黒連結矩形領域）を抽出する。さらに黒連結矩形領域を統合して基本矩形を抽出する（図9）。

(2) 基本矩形特徴の抽出

抽出した各基本矩形から、図10、表1に示す11種類の基本矩形特徴を抽出する。概念的には、sxは基本矩形の左マージンを、lhは文字の大きさを、nmbは基本矩形中の文字数を、dnstは文字のつまり具合を反映している。これら11種類の基本矩形特徴を要素とするベクトルを基本矩形の特徴ベクトルとする。

(3) 基本矩形の分類

算出された特徴ベクトルと項目ごとの参照ベクトルを比較することにより、基本矩形を「著者名(和)」等の論理属性に分類する。

参照ベクトルは、各項目に対応する正しい領域座標を記述した教師データをもとに作成する。これらの教師データをもとに、レイアウト解析のための「知識辞書」を作成した。

IEICEフロント頁の領域抽出結果例

IPSJフロント頁の領域抽出結果例

図6 構成要素の領域（フロントページ）

IEICE中間頁の領域抽出結果例

IPSJ中間頁の領域抽出結果例

図7 構成要素の領域 (中間ページ)

IEICE最終頁の領域抽出結果例

IPSJ最終頁の領域抽出結果例

図8 構成要素の領域 (最終ページ)

図9 基本矩形の抽出

図10 基本矩形特徴

表1 基本矩形特徴

項番	特徴名	記号
1	左上x座標	sx
2	右下x座標	ex
3	左上y座標	sy
4	右下y座標	ey
5	行の高さ	lh
6	上の行との間隔	ali
7	下の行との間隔	bli
8	黒連結矩形数	nmb
9	縦横比	sqd
10	面積	$area$
11	黒連結矩形密度	$dnst$

表2 OCR パッケージ

種別	パッケージ名称
和文	SparcReader Pro Library メディアドライブ研究所
英文	ScanWorX API Xerox Imaging Systems Inc.

3. 2. 2 OCR 処理

レイアウト解析により得られたテキスト領域を対象に、OCR パッケージソフトウェアを用いて認識し、テキストコードに変換する。使用した OCR パッケージを表 2 に示す。

3. 2. 3 文字認識後処理

ユーザインタフェースを通じ、オペレータが文字認識結果の修正を行う。

3. 2. 4 SGML 文書化

レイアウト解析結果と文字認識結果を統合して、論文構成要素名など論文の構造タグを付与した SGML 文書インスタンスを論文単位に作成する。

3. 2. 5 HTML 文書化

インターネット上の WWW による情報提供が盛んになってきていることから、作成蓄積された SGML 文書の利用イメージとして、これらの文書を、HTML 文書に変換して WWW ブラウザ可読の形態とした。

4. システム性能評価および考察

プロジェクトの第 I 期（3 年間）で構築したプロトタイプシステムの評価結果をもとに、第 II 期のプロトタイプ構築にあたっては、下記の 4 項目の技術的課題に取り組んだ。

1. 文字認識精度の改善
2. 入力対象誌と認識項目の拡大
3. 文書蓄積方式の検討
4. 処理速度の改善

さらに、実現可能と考えられる具体的な目標性能数値を表 3 のように設定した。

実験に使用したデータは、情報処理学会論文誌、電子情報通信学会論文誌（D-II 分野）のそれぞれ約 1 年分である。

4. 1 文字認識精度の改善

電子情報通信学会論文誌、情報処理学会論文誌の 2 誌について、下記の 3 項目を測定対象から除外することとして、文字認識精度の測定を行った。

表 3 重点目標課題

1. 文字認識精度（和文、英文）	97%以上
2. 読み取り速度	
・ターンアラウンド処理時間（オペレータ操作を除く）	4 分/頁
・スループット（機械処理部分のみ）	3 頁/10 分

表 4 文字認識精度測定結果

	論文誌	測定文字数	誤読数	認識精度(%)
和文	情処	5, 264	238	95. 48
	信学会	5, 757	36	99. 37
	全体	11, 021	278	97. 51
英文	情処	4, 009	6	99. 85
	信学会	981	12	98. 78
	全体	4, 990	18	99. 64

その結果を表 4 に示す。

- ・和文領域中の英単語間の空白
- ・上付き文字（参考文献番号など）
- ・英文著者領域中の著者所属指示記号（ダガーマークやアスタリスク）

レイアウト解析で得られた文字領域の種類に応じて、和文、英文 OCR パッケージを切り替えることによって、文字認識精度 97% 以上という目標を達成することができた。情報処理学会論文誌の和文認識精度が低いのは、参考文献中の英文斜体文字（本来、和文 OCR ライブラリの認識対象文字種ではない）の誤読による影響のためである。

4. 2 入力対象誌と認識項目の拡大

第 I 期で対象としたフロントページに加えて、第 II 期では、中間ページ、最終ページを新たに解析の対象とした。レイアウト解析用モジュールは、解析するページレイアウトに適した知識辞書を使用する。それぞれの知識辞書は、100 ページ程度のサンプルデータをもとに作成した。

表5 レイアウト解析用実験データ内訳

学会誌	フロント	中間	最終	合計
信学会	110	110	118	338
情処	118	108	109	335
合計	228	218	227	673

表5に示した673ページの画像データを使用して実験を行った。実験の結果、論文構成要素の分類率（正しくレイアウト解析が行われた率）は、情報処理学会論文誌の最終ページの本文領域（85.31%）を除き（本文領域を参考文献と誤認識したものが64.38%）、93%以上であった。また、電子情報通信学会論文誌フロントページの110枚の実験データのうち、レイアウト解析の結果が誤っていたのは、10ページであった。

このように、レイアウトの異なる二つの論文誌の実験結果も良好であったことから、他学会の論文誌への拡張性も高いと判断した。

項目の詳細化に関しては、レイアウト解析により得られた領域に対して、さらに解析を加えることにより、下記の4項目について取り組んだ。

- (1) 著者名領域を個々の著者名領域に分割し、その記載順序を抽出
- (2) 所属機関名領域を個々の所属機関名領域に分割
- (3) 個々の著者名と所属機関とのリンク
- (4) 本文領域から章題、パラグラフ領域の抽出

4.3 文書蓄積方法の検討

認識した文書情報は、SGML文書として蓄積することとした。そのために、処理対象学会誌2誌の文書構造を定義するDTD (Document Type Definition) を設計した。

4.4 処理速度の改善

速度性能の評価尺度としては、1画像（ページ）あたりのスループット時間と1画像あたりのターンアラウンド時間を用いた。

入力・蓄積サーバでの1画像あたりの処理時間

を T_s 、認識・変換サーバでの1画像あたりの処理時間を T_R として、両マシン間のオーバーヘッドは無視できるものと仮定した。

・システム全体としての1画像あたりのスループット時間 T_T は、

$$T_T = \max(T_s, T_R) \dots\dots\dots 1$$

・1画像あたりのターンアラウンド時間 T_A は

$$T_A = T_s + T_R \dots\dots\dots 2$$

(1) 入力・蓄積サーバの処理時間

正常に入力、正規化処理が終了した文書画像887枚について、入力・蓄積サーバの処理時間 T_s を測定した。その結果、1画像あたりの平均処理時間は約19秒であった。

(2) 認識・変換サーバの処理時間

文書画像54枚について、認識・変換サーバの処理時間を測定したところ、1画像あたりの平均処理時間 T_R は、約49秒であった。

式1、2より、システム全体としての1画像あたりのスループット時間 T_T および1画像あたりのターンアラウンド時間 T_A は、それぞれ以下のようになり、速度性能の目標値（前者200秒、後者240秒）を達成できたことを確認した。

$$T_T = \max(T_s, T_R) = 49 \text{ 秒}$$

$$T_A = T_s + T_R = 19 + 49 = 68 \text{ 秒}$$

5. お わ り に

今回のプロトタイプングを通して、印刷出版雑誌からのテキストデータの作成およびそのデータベース化に関して、運用上現実的な認識率および速度で実現可能であることが分かった。

今後、本プロトタイプシステムを実運用システムに移行させるためには、1) 文字認識率のさらなる向上、2) 処理対象誌の柔軟な拡張性、3) オペレータ操作性の向上、4) システム構成の見直し、等の技術課題に取り組む必要がある。

最後に、本プロジェクトで使用した実験用データの利用については、(社)電子情報通信学会、(社)情報処理学会の両学会からご協力いただきました。ここにお礼申し上げます。

~~~~~ 参 考 文 献 ~~~~~

- 1) 創造的研究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研究 (第 I 期)
成果報告書, 科学技術庁科学技術振興局, 1995.
- 2) ISO8879, Information processing-Text and
Office Systems-Standard Generalized Mark-
up Language (SGML).
- 3) JIS X4151, 文書記述言語 SGML.
- 4) Martin Bryan (山崎俊一監訳) SGML 入門,
アスキー.
- 5) Eric van Herwijnen (SGML 懇談会実用化
WG 監訳) 実践 SGML, 日本規格協会.
- 6) 根岸正光, 石塚英弘, SGML の活用, オーム
社.
- 7) 吉岡誠編, SGML のススメ, オーム社.