デジタル・ドキュメント 10－5
（1997. 11. 28）

多様な文書タイプに適応可能な文書構造化システム

石田 和生　　神谷 俊之　　谷 幹也　　市山 俊治
{ishidakz,kamiya,m-tani,ichiyama}@hml.cl.nec.co.jp
NECヒューマンメディア研究所
〒540 大阪市中央区城見1－4－24

近年、既存文書を電子化する手法として、文書の論理構造を抽出して構造化テキストの形式で蓄積する方法が注目されている。我々は、主にレイアウト情報も同時に蓄積する事を主な特徴とする構造化システムの開発を行なってきたが、対象とする文書タイプが変わるとシステムそのものを変更しなければならなかった。そこで、多様な文書の構造化を容易に行なうため、1) 代表的な文書タイプについてそのレイアウト的特徴を分析し、2) 構造抽出のために必要な知識を記述できる表現を定義し、3) この表現形式を用いて知識をルール化することにより、知識をシステムから分離した。本論文では、文献タイプの調査結果、ルール記述形式、及び、システムの特徴について報告する。

Structured Text Generating System Applicable to Various Types of Documents

Kazuo Ishida, Toshiyuki Kamiya, Mikiya Tani and Shunji Ichiyama
Human Media Research Laboratories, NEC Corporation
1－4－24 Shiromi, Chuo-Ku, Osaka 540, JAPAN

Storing a published document as a structured text has many advantages in terms of storage efficiency, speed and reuse. We have developed an efficient system that extracts the logical structure, digitizes contents, and store as structured texts from physically published documents. This paper describes the survey of the layout features of several document types and proposes the system that we have improved from the embedded knowledge base to the rule base in order to easily add new document types without change of the system.

1 はじめに

近年、インターネット、イントラネットの普及に伴い企業や個人での情報発信、獲得などが手軽に行えるようになってきた。その結果、ネットワークを通じてのドキュメント（より一般にはコンテンツ）の頒布や共有化が非常に注目されている。しかし、流通させるコンテンツの作成には非常に多くの手間がかかっているのが現状である。また、オフィス文書や既存の文献などのように、過去に紙ベースで流通して、なおかつ現在でも参照する必要のある文書は非常に多い。これらの文書をネットワーク上で利用するためには何らかの方法で電子化してやらなければならないが、この手間も決して少なくない。筆者らのグループは、このような既存文書の入力を簡単に、かつ一定品位で行うためのシステム「情報ファクトリ」を開発している[1]。情報ファクトリでは、既存文書をスキャナで読み込んだ後、文字認識と文書の論理構造の認識を行いデータベースに蓄積する。このとき入力としては、論文や文庫本など様々なタイプの文書が有り得るため、論理構造抽出部もそれらの多様な文書に対応したものでなければならない。そこで本研究では、多様な文書タイプに柔軟に対応できる論理構造抽出を実現するため、代表的ないくつかの文書のレイアウト的特徴を調査し、その結果に基づいて、構造化部のルールベース化を行ったので、これについて報告する。

2 構造化システムのルールベース化

既存文書の電子化については文書をスキャナで読み込んで画像ファイルとして保存する方法や、文字認識を行ってテキストファイルとして保存する方法などさまざまな形態のものが考えられるが、オフィス文書や一般図書を蓄積対象とした場合、その文書の文章（すなわち文字データ）に含まれる情報がメインであることが多いため、画像ファイルとしてではなくテキストデータとして保存しておいたほうがデータの有効利用が可能となる。さらに、文書の持つ論理構造情報（セクションのタイトルや段落といったもの）まで含めて蓄積しておけば、その論理構造情報に基づいた検索（タイトルに「電話」が含まれる文献、といったもの）や、文書の再利用などが非常に効率的に行えるようになる[2][3]。しかしこれらの作業は現在のところそのほとんどが手作業に頼っているため、非常に大きな労力が必要な上、作成したデータの品質が作業者によって大きく異なる可能性が高いという問題がある。そこで現在我々は、情報ファクトリという既存文書電子化システムの試作を行っている。これは既存の文献をスキャナで読み込み、読み込んだ画像データに対して文字認識と論理構造抽出を行い、SGML 形式と呼ばれる構造化テキストの形式にした上でデータベースに登録するというものである（図 2-1）。これにより大量にある既存文書を一定品質で電子化することが可能となる。筆者らは主に、この情報ファクトリシステムの論理構造抽出部分（文書の構造化）について研究開発を行っており、文字認識されたデータから論理構造の抽出を行い SGML 形式の構造化文書に変換するモジュールの開発を行っている。文献[4]では、その変換手法と試作した自動 SGML 変換システム（以下、「構造化システム」と呼ぶ）の概要について述べた。

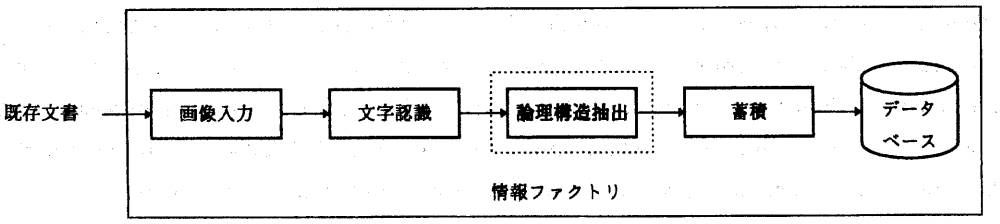


図 2-1 情報ファクトリ全体構成

開発した構造化システムは主に、文書のレイアウト情報 (文字のサイズや空白の量など) をもとに構造情報の抽出を行うのであるが、そのレイアウト的特徴と文書の論理構造との関係(構造抽出知識) を構造化システム自体にハードコーディングしているため、構造化の対象文書を変更したときにはシステムの再コンパイルが必要であり、対象文書の変更や構造抽出知識の作成に時間がかかるという問題点があった。そこで、システムから構造抽出知識を分離して、対象文書のタイプを変更した場合でも構造抽出知識だけを入れ替えればすむような構成のシステムを開発することにした。

システム開発のためには、まず、構造抽出知識の記述仕様の設定が必要である。このため、現在流通している文献の中から代表的なものをいくつかピックアップし、それらの文献についてレイアウト的特徴を調査した (結果の概要を3章で述べる)。そしてこの調査結果をもとに、論理構造抽出知識を表現する「構造抽出ルール」の記述仕様を決定し、構造化システムに組み込んだ。これにより、様々なタイプの文書の構造化に対し柔軟に対応可能な構造化システムとなった。システム概要については、5章で述べる。

3 レイアウト的特徴情報の調査

3.1 調査対象文書と調査内容

レイアウト的特徴情報を調査する文献のタイプとしては、現在流通しているメジャーな文献で、なおかつ、電子化されて有用と思われる文献の中から選定し、文庫本、論文、技術系雑誌、週刊誌、解説書、辞書の６タイプとした。調査する項目は、構造化することで有効に活用できる情報のうち

- 文書のタイトル
- 著者、所属
- セクションタイトル

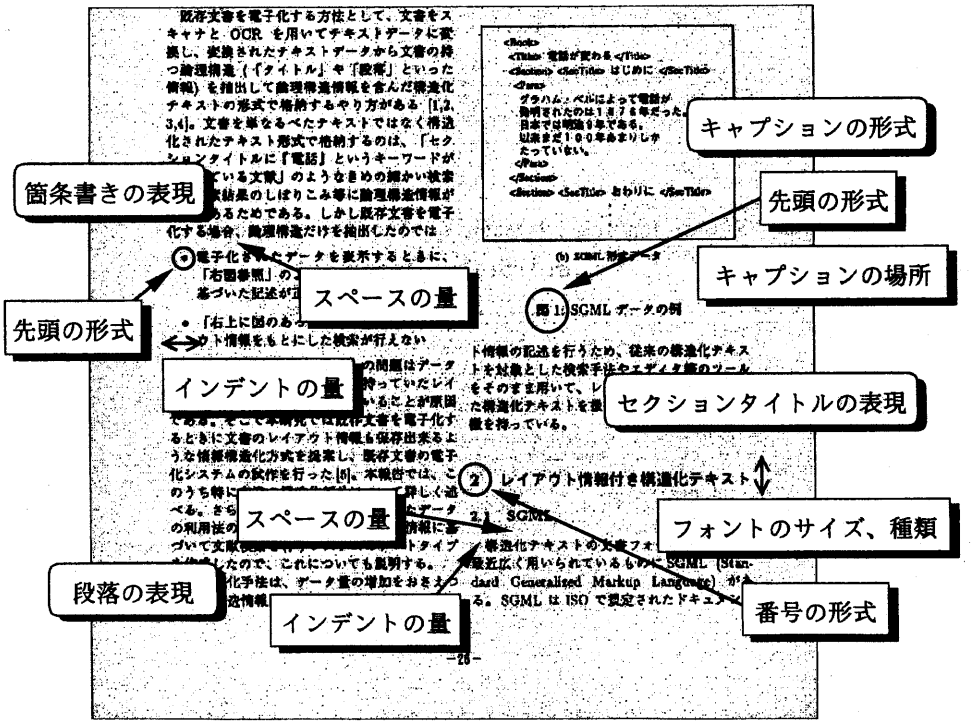


図 3-1 レイアウト的特徴情報の調査項目 (一部)

- 図表のキャプション
- 箇条書き
- 段落

とし、これらについてフォントのサイズやインデントの量などのレイアウト的特徴情報の調査を行った (図 3-1 参照)。

3.2 調査結果と考察

今回調査した結果のうち、文庫本と論文以外のものは、各タイプにおいて特徴のパターンの数が非常に多く、タイプ毎のパターンの特徴抽出をする程十分に整理が出来ていない。一方、文庫本と論文はシリーズや論文誌の種類を特定するとかなりパターンがしぼれるので、ここでは、この2種類の文献についての結果をまとめることにする。なお、対象とした文献は、文庫本は C&C 文庫、論文は情報処理学会研究報告と、電子情報通信学会技術研究報告である。現在整理できていない他のタイプに関しては、今後、各タイプにおける基本的特徴とそれらのシリーズ別の特徴抽出を、統計的手法などを用いて調査していく予定である。

文庫本および論文における特徴情報の調査結果を表 3-1に示す。表 3-1の結果を見てみると、文庫本と論文に関しては、タイトルやセクションタイトルといった構造は特にフォントサイズやページ上での位置情報に大きな特徴があることがわかる。またこの他にも、行間スペースや文字の種類、行の先頭文字などが論理構造を抽出する際に有効に利用できる情報であることが明確となった。これらの結果から、構造化システムの構造抽出ルールの表現として、特に、

表 3-1 レイアウト的特徴調査結果

	文庫本	論文
文書のタイトル	表紙の上方、約 2 倍のサイズのフォント、横書き	1ページ目上方、約2倍のサイズの強調フォント、センタリング
著者、所属	表紙で場所は様々	約 1.5 倍のフォント、センタリング、タイトルのすぐ下
セクションタイトル	「数字」で始まり、本文とほぼ同じ大きさのゴシックフォント	「数字」「数字.数字」で始まる、約 1.2〜1.5 倍のフォント
図表のキャプション	「図　数字」「表　数字」などのパターン、図表番号の部分は数字だけでなくアルファベットやセクション番号との組み合わせなどの形態もある	文庫本と同じ
箇条書き	「・」「(数字)」「数字.」などのパターン、数字の部分はギリシャ数字やアルファベットなどの場合もあり	文庫本と同じ
段落	1文字インデント	1〜1.5 文字インデント

- フォントサイズ、行間スペース
- ページ全体に対する位置情報
- 行の先頭の文字列

の関係を容易に記述できるような仕様にすることが有効であると考えられる。次章では、以上のような結果に基づいて実装した、構造化ルールを解釈するルールインタプリタについて述べる。

4 ルール記述仕様

前章での特徴抽出結果より、論理構造を抽出するのに必要なルール記述能力は以下のようなものであると考えた。

- 入力データの行単位で、レイアウト情報の関係が記述できる
- 前後の行との行間やフォントサイズの比較が可能

```
; この行はコメント

 [CandC.struct]
; 文書の構造を定義
Book=Fm,Body
Fm=Title?,Author*
Body=(Section|Bib)+

[CandC.rule]
; スコア付けのルールを定義
; ルール1
Height(0) > 20 && LineSkip(-1) > 30 {
    SetScore(0, "Title",Score(0,"Title")+10)
}
; ルール2
LineSkip(-1) < 10 {
    AddSub(0, "Continuation", 10, 10)
    Next()
}

[report.struct]
; 以下省略
```

図 4-1 ルールファイルの例

- 構造化ルールの入力データへの依存度をなるべく減らすため、ページの実際の大きさを意識せずにレイアウト情報の関係を記述できる
- 行の先頭文字列に関してパターンマッチが出来る

また、出力文書が持つ論理構造は入力文書タイプに応じて様々なパターンが存在するので

- 出力文書の論理構造の指定が可能

である必要もある。以上のことから、構造抽出ルールは、大きく分けて、構造定義部と構造抽出ルール記述部の二つのセクションからなるものとした (図 4-1参照)。

構造定義部は、SGML の DTD (Document Type Definition: SGML 文書の文書構造を定義するもの) のエレメントの構成記述部を簡略化した形式となっており (図 4-1 の CandC.struct セクション部を参照)、本システムが出力する SGML テキストの文書構造を指定する。例えば図 4-1の場合は、文書のトップは Book で、その下には Fm と Body の 2 種類の要素が存在し、さらに、Fm の下には Title が 0 か 1 回と、0 回以上の Author からなっていることなどを表している。ここに記述した Book や Fm などは SGML のタグ名を表しており、ここに挙げたものの他に、SecTitle や Figure など約 40 種類のものを用意している。

一方、構造抽出ルール記述部は構造抽出を行うために使用するルールを記述した部分である。構造抽出ルールの形式は awk などのスクリプト言語をベースにしたもので、各ルールは

```
条件部 {
    アクション部
}
```

のように、条件部と、その条件がマッチしたときにスコア付けなどを実行するアクション部のペアからなっている (図 4-1の CandC.rule セクション部を参照)。このような形態にしたのは、本システムの構造化ルールの記述仕様を、入力データの行ごとにレイアウト関係の条件

を記述し評価を行うように定義しており、この動作が awk などのスクリプト言語に非常に似たものであったためである。

条件部では入力行毎に適用されるレイアウト的特徴情報の関係を記述する。例えば、図4-1の CandC.rule セクション部で説明すると、ルール 1 の条件部は、現在の行の高さ (Height(0)) が 20 よりも大きく、かつ、一つ前の行との行間 (LineSkip(-1)) が 30 よりも大きいときにアクション部を実行することを意味している。条件部に記述できる関数は、ここで挙げた行の高さや行間を返す関数の他にもインデント量を返す関数や、行の文字列を返す関数など約 20 種類用意している。また、数値を返す関数は、入力データのページサイズに依存しないで利用できるように、ページサイズを 10000 とするように正規化したものを戻り値としている。

ルールのアクション部には、入力行がルールの条件部にマッチした時に、その行が持つ論理構造の可能性 (確からしさ) をスコアとして与えるための関数 (SetScore() など 5 種類を用意している) を記述する。システムは最終的に各行にふられたスコアの大小関係をもとに、構造定義部に記述された論理構造に従った文書を出力する。

5 ルールベース構造化システム

本章では、前章で定義した構造抽出ルールをもとに動作する、構造化システム (以下ではこのシステムを、ルールベース構造化システムと呼ぶ) の構成と動作について説明する。

5.1 システム概要

ルールベース構造化システムは、文献[4]で報告した構造化システムに構造情報抽出用の「ルールインタプリタ」を組み込んだものである。このルールインタプリタは、前章で述べたルールファイルを解釈し、入力文書の論理構造抽出と SGML テキスト出力を実行する。システムの動作の流れを図 5-1に示す。

システムの動作を簡単に説明すると以下のようになる。まず、入力文書の行毎に対しルールファイルに記述された各ルールを先頭からひとつずつ適用してスコア付けを行う。これを全ての入力行に対し実行し、その後、各行に設定されたスコアの大小関係とルールファイルの構造定義に記述された文書構造をもとに、各行の文書構造を決定し、SGML テキストを生成、出力する。

5.2 システムの入力ファイル

本構造化システムの入力ファイルは、ルールファイル、目次ファイル、文書の認識結果ファイルの 3 つである。文書認識結果ファイルは NEC 社内で開発している文字認識システム[5]が出力するファイルで、認識された文字だけでなくその位置情報なども含まれている。目次ファイルは文書の目次を記述したファイルで、これも同様に文献の目次ページを認識システムで認識した結果を用いる。この目次ファイルは主に、セクションタイトル (セクションの切れ目) を認識するために利用される。すなわち、目次ファイルに記述された文字列をもとにパタンマッチングを行い、本文中に存在するセクションタイトルの文字列を探索する。

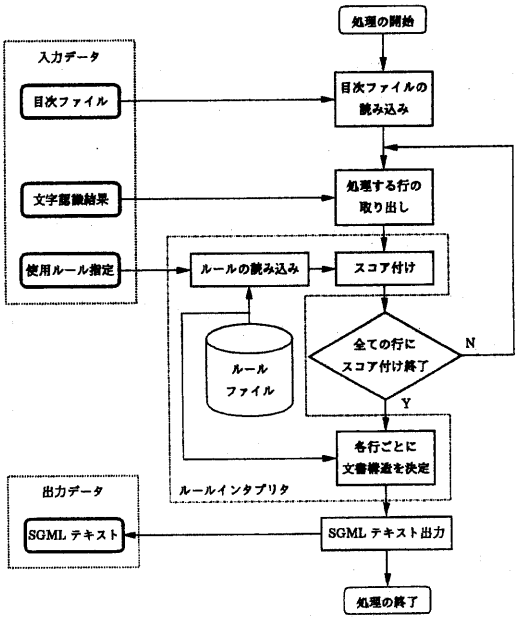


図 5-1 構造化システムの処理の流れ

ルールファイルは、前章で説明した、構造化の際に使用する構造抽出ルールと出力するSGML文書の構造定義を記述したものである。

5.3 システムの出力ファイル

本論理構造抽出ルールインタプリタを組み込んだ構造化システムの出力は、レイアウト情報付きSGMLテキスト[4]である。これは、通常の、文書の論理構造のみを含んだSGMLテキストに、文書の持つレイアウト情報（段落などのブロックがページ上で占めている矩形領域の左上の座標と幅、高さの情報など）を同時に埋め込んだ形式のテキストで、もとの文書のレイアウトをもとにした検索（右上に図があった文献、など）や、SGMLテキストからもとのページイメージ画像を再現することが可能であるといった特徴を持っている。

5.4 実行結果

今回試作した論理構造抽出ルールインタプリタを組み込んだ構造化システムの動作例を示す。対象文書としては、最も基本的な構造を持っていると考えられる文庫本を用いた。

構造化に使用した文書画像はスキャナを用いて400dpiの密度で読み込んだもので（図5-2)、この画像を文字認識システムを利用して文字認識を行い、その認識結果を構造化した結果が図5-3である。変換された結果を見るといくつかの文字認識誤りは存在するものの、文書の論理構造については正しく認識していることがわかる。また、出力結果のSGMLテキスト中に埋め込まれているLayoutタグが文書のレイアウト情報を記述したもので、Layoutタグで囲まれた部分がページのどの場所に存在していたかを表している。

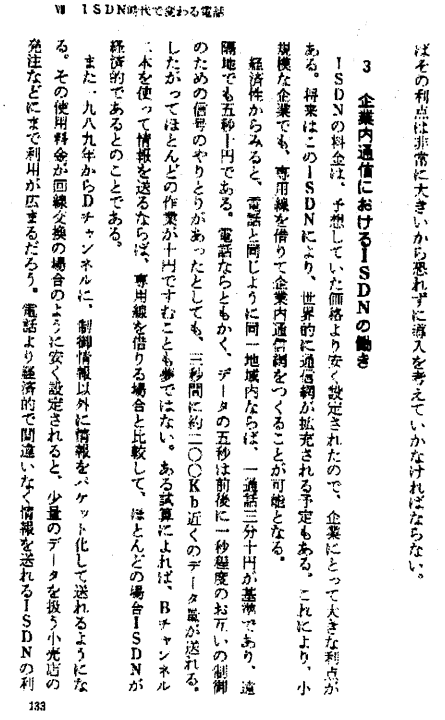
5.5 考察

現在、試作したシステムを用いて論文などの文書の構造化ルールを構築しているところであるが、従来のハードコーディングされたシステムに比べルールベース構造化システムでは、

- ルールの追加削除変更が行いやすいため、ルール作成に費やす時間が減少する
- 個々のルール自体もCのプログラムとして作成する場合に比べ、容易に記述できる

という利点がある。しかし、対象文書の構造化ルールを作成する際に、ルール同士の相互影響を考慮する必要があるという点は今回のシステムでは解決されていない。これについては今後、ルール作成の支援ツールの開発などを行う予定である。また、システムの評価、ルール記述仕様の能力に関して、

- 構造抽出精度の尺度の定義
- レイアウト的特徴情報の整理が出来ていない文書タイプのパターン整理とルール作成


VII ISDN時代で変わる電話

ばその利点は非常に大きいから恐れずに導入を考えていかなければならない。

3 企業内通信におけるISDNの働き

ISDNの料金は、予想していた価格より安く設定されたので、企業にとって大きな利点がある。将来はこのISDNにより、世界的に通信網が拡充される予定もある。これにより、小規模な企業でも、専用線を借りて企業内通信網をつくることが可能となる。

経済性からみると、電話と同じように同一地域内ならば、一通話三分十円が基準であり、遠隔地でも五秒十円である。電話ならともかく、データの五秒は前後に一秒程度のお互いの制御のための信号のやりとりがあったとしても、三秒間に約三〇〇Kb近くのデータ量が送れる。したがってほとんどの作業が十円ですむことも夢ではない。ある試算によれば、Bチャンネル二本を使って情報を送るならば、専用線を借りる場合と比較して、ほとんどの場合ISDNが経済的であるとのことである。

また、一九八九年からDチャンネルに、制御情報以外に情報をパケット化して送れるようになる。その使用料金が回線交換の場合のように安く設定されると、少量のデータを扱う小売店の発注などにまで利用が広まるだろう。電話より経済的で間違いなく情報を送れるISDNの利

133

図 5-2 入力画像

などを行っていく予定である。

6 おわりに

本研究では、多様な既存文書に柔軟に対応できる情報構造化システムを構築するために、いくつかの既存文書に対してレイアウト的特徴情報を調査し、その結果に基づいて、ルールベース構造化システムの試作を行った。論理構造抽出ルールインタプリタが構造化システムに組み込まれたことで、多種の文書タイプに、より柔軟に対応することが可能となった。しかし、構造化ルールの作成には個々のルールの記述の容易さだけでは不十分であると考えられるため、今後はルール作成の支援などについて検討する必要がある。また、本システムを用いて実際に様々なタイプの文書に対する構造化ルールを作成し、実験、及び、構造抽出結果の評価も行う予定である。

```
<Book>
<Section>
<SecTitle>
<Layout TOP=0.12646 LEFT=0.75766
 WIDTH=0.03203 HEIGHT=0.37255>
3企業内通信におけるＩｓＤＮの働き
</Layout>
</SecTitle>
<Paragraph>
<Layout TOP=0.10571 LEFT=0.59424
 WIDTH=0.19545 HEIGHT=0.79085>
ＩＳＤＮの料金は、予想していた価格より
安く設定されたので、企業にとって大きな
利点がある。将来はこのＩＳＤＮにより、
世界的に通信網が拡充される予定もある。
これにより、小規模な企業でも、専用線を
借りて企業内通信網をつくることが可能と
なる。
</Layout>
</Paragraph>
<Paragraph>
<Layout TOP=0.10543 LEFT=0.26880
 WIDTH=0.30084 HEIGHT=0.79085>
経済性からみると、電話と同じように
                (以下省略)
```

図 5-3 SGML 変換結果

参考文献

[1] 神谷 他: ユニバーサル図書館に向けての図書入力システム「情報ファクトリ」の試作, 「ディジタル図書館」ワークショップ第 8 回, pp. 44－58, 1996.

[2] M. Yamaoka, M. Sato, K. Iwane and O. Iwaki, A Document Understanding System for Converting Printed Documents to SGML Instances, Proc. ISDL'95, pp. 287－288, 1995.

[3] 山田 満: 文書画像の ODA 論理構造化文書への変換方式, 信学論 D－II, Vol. J76－D－II, No. 11, pp. 2274－2284, 1993.

[4] 石田 他: 既存文書のレイアウト情報付き構造化とその利用, 情処研報 96－FI－44, pp. 25－32, 1996.

[5] 中島 他: 文書の構成要素抽出に基づく領域分割, 第 55 回情処全国大会, 1K－03, 1997.