情報学基礎 13－5
（1989. 6. 2）

プレーンテキスト／ハイパーテキスト間の変換

土井美和子、福井美佳、山口浩司、竹林洋一、岩井勇
東芝　総合研究所

プレーンテキスト（構造化されていない文書）から章や節などの階層構造や図表などの参照構造を抽出する技術を開発した。この文書構造の自動抽出技術により、抽出された文書構造を用いて文書のレイアウトを行う文書自動レイアウト機能を開発した。

両者の技術により、プレーンテキストをハイパーテキストに、逆にハイパーテキストをプレーンテキストに変換することが可能となった。

文書自動レイアウト機能は、東芝の最上位機種の日本語ワードプロセッサ（JW－1000AI）に搭載されている。

Conversion between plain-text and hypertext

Miwako Doi, Mika Fukui, Kouji Yamaguchi, Youichi Takebayashi, Isamu Iwai
Toshiba Research and Developement Center
1, Toshiba-komukaicho, Saiwai-ku, Kawasaki-shi, Kanagawa210, Japan

A document layout system based on automatic extraction of document architecture including logical and reference structures has been developed for reducing users' effort in document preparation, and has been implemented in a practical Japanese word prosessor. The extracted document architecture is used for both automatic text formatting and layout of text, figures and tables. Automatic text element recognition is performed by morphological analysis using keywords. Through intra-line (one sentence) and inter-line (relations between sentencesa) analysis, logical and reference structures and obtained. The automatic layout system effectively lays out the document using the extracted document architecture and knowledge about the layout.

1．はじめに

近年、脚光を浴び始めているハイパーテキストと従来のプレインテキストとの相違について若干考察する。

その考察に基づき、プレインテキストからハイパーテキストへの変換に必要な文書構造の抽出技術と、ハイパーテキストからプレインテキストへの変換のための整形技術について述べる。

これらの技術は、東芝の日本語ワードプロセッサの最上位機種JW-1000AI 上で文書自動レイアウト機能として稼動している。

2．ハイパーテキストとプレインテキストの相違

ハイパーテキストの概念自体は、1945年にBush[1] より紹介されている。

通常、文書の作成者や読者が、考えたり、読んだりするときには、文書を単なる文字列の集合としてでなく、章や節などの構造のある文書として作業を行っている。つまり、プレインテキストが対象であっても、無意識のうちに、ハイパーテキストが対象であるかのように作業を行っているのである。この観点からみると、ハイパーテキストは、人間にとって、ごく自然な発想と言うことができる。

ではなぜ、ハイパーテキストが急に注目を集め出したのであろうか？

その理由の一つに、CD-ROMの普及により、出版社や印刷会社が従来の紙のメディアにこだわらず、電子メディアを用いた出版を開始したことがあげられる。つまり、コンピュータ技術の進展により、文書の構造を意識して作業できる環境が可能になってきたことが、ハイパーテキストへの関心を大いに高めたのである。

ただし、既に作成され、コンピュータ上に蓄積されているすべての文書に対して、ハイパーテキストとするための構造化を人手で行うのは困難である。また、新規に作成する文書に関しても、作成者に意図しない構造化の作業を強いるのでは、何のためにコンピュータにより支援を行っているのかわからなくなる。このような問題を解決するためには、自動的に構造化を行う処理が必要になる。

ハイパーテキストの読者は、意識的にtext間のリンクを辿らねばならず、リンクの付け方が作成者と読者の間で大きく隔たっているときには、非常に困難な作業となる。一方、プレインテキストでは、読者は、レイアウトを手掛かりとして、作成者のリンク付けの拘束されずに、リンクを辿ることができる。このような簡便性を読者が享受するには、ハイパーテキストをプレインテキストとして生成する手段が必要となる。この生成は、文書の階層構造と、参照構造に基づいた整形処理である。章見出しを太字にするような階層構造に基づいた整形処理は、Scribe[5] などの整形処理システムで既に行われている。これに対し、参照箇所の近傍に参照されている図表を配置するような参照構造に基づく整形処理は、従来行われていなかった。

以上述べてきたように、ハイパーテキストとして、文書構造を意識的に扱うためには、プレインテキストの構造化が必要であり、一方、ハイパーテキストを読者にとって読みやすいプレインテキストに変換するには、文書構造に基づいた整形処理が必要である。

著者らの開発した文書自動レイアウト機能は、この必要性に応じて開発され、文書構造抽出技術と整形技術を一体化して商品化を行った。

以降、それぞれの技術について述べる。

3．文書構造の認識

3．1　階層構造の認識

文書の基本となるのは、章や節などの論理的な骨組みを表わす階層構造である。この階層構造の一面を示すのが、目次であると考えられる。章・節、それぞれは、見出しと箇条書き、段落、さらに下位の節が入れ子になった図1 (a) のような構成になっている。

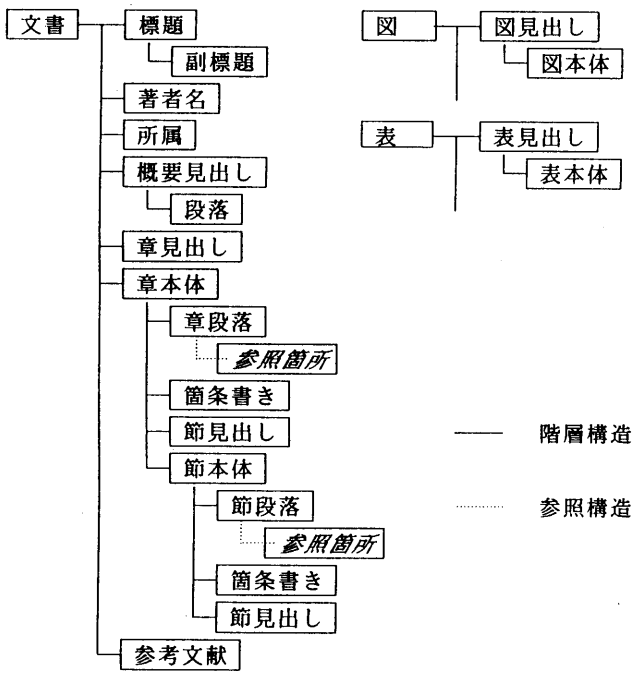


図１　文書構造（技術文書）

人間はこのような階層構造を、文書の内容に関する知識がない場合には、見出しに付加されている数字や記号などの形態的特徴を基に、認識している。

また、数字や記号などの形態的特徴のない著者名や所属などに関しては、包含されている固有名詞が決め手となっている。

このような人間がもっている階層構造を認識する知識を表現すると、図２(a) のようになる。

３．２　参照構造の認識

文書では、「*第１章で*」などのように図表や、参考文献、前出の章や節などを参照することがある。この参照する箇所（参照箇所）と参照される図表、参考文献、章・節などの間にある関係を参照構造（図１(b) ）と呼ぶ。ここでは、レイアウト上、重要な図表の参照構造の抽出を取り上げて考える。

文書の内容に関する知識がない場合には、

①「*第１図に*」「*表３に*」のように「図」や「表」などのキーワードを含んでいる。

②参照箇所「*第１図に*」と参照されている図表の見出し「*図１　システム構成*」に含まれる図表番号が等しい。

というような知識に基づいて、参照構造を認識している。レイアウトされていない文書を対象に考えているので、「*下図に*」などのように、物理的な位置を手掛かりに参照している箇所は、レイアウトにより物理的位置が変動し、確実な認識情報となり得ないので、対象としていない。

このような人間がもっている参照構造を認識する知識を表現すると、図２(b) のようになる。

４．文書構造の抽出

前章で述べた文書構造を認識する知識を備えた文書構造自動抽出システムは、後述の整形処理部と結合した形で、図３に示すように構成されている。このシステムは、階層構造抽出部と参照構造抽出部の２つに分れている。

４．１　階層構造解析部

階層構造抽出部は、さらに、形態解析部と文間構造解析部の２つに分れている。形態解析部は、改行で区切られるまで（一文と呼ぶ）を解析し、３．１節で述べた番号や記号などの形態的特徴を解析する。ここで、使われる知識は、図２(a) のＪからＮである。文間構造解析部は、形態解析部で得られた形態的特徴をもとに、章・節などの構造を決定する。まず、図２(a) のＧからＩまでの知識を使って、見出し間のつながり（見出し「*1.*

A　：［文書］→［序］，［文書本体］
B１：［序］→［標題］，［序］
B２：［序］→［著者名］，［序］
B３：［序］→［所属］，［序］
B４：［序］→［住所］，［序］
B５：［序］→［標題］
B６：［序］→［著者名］
B７：［序］→［所属］
B８：［序］→［住所］
C：［文書本体］→［章ｉ］，［文書本体］
D：［章ｉ］→［章見出しｉ］，［章本体］
E：［章見出しｉ］→N(i)，［記号］，［見出し語］
　N(i) は、番号付けを行う関数。
F１：［章本体］→［章段落］，［章本体］
F２：［章本体］→［章段落］
G：k（N(i)）＝k（N(i+1)）　(i≧1)
　kは、番号の文字種の評価関数。
H：o（N(i+1)）＝o（N(i)）＋1　(i≧1)
　oは、番号のオーダ評価関数。
I：o（N(1)）＝1
J：［標題］→名詞句
K：［著者名］→人名を含む名詞句
L：［所属］→企業名などを含む名詞句
M：［見出し語］→名詞句
N：［見出し語］→はじめに

(a) 階層構造の認識知識

a：［参照箇所］→［参照用語］，N(i)
b：［参照箇所］→N(i)，［参照用語］
c：［図表見出し］→［参照用語］,M(i)，［見出し語］
d：［図表見出し］→M(i)，［参照用語］［見出し語］
e：［参照用語］→図｜Ｆｉｇｕｒｅ
f：［参照用語］→表｜Ｔａｂｌｅ
g：k（N(i)）＝k（M(i)）(i≧1)
h：r（参照箇所の参照用語）
　＝r（図表見出しの参照用語）
　　rは、参照用語の種別評価関数。

(b) 参照構造の認識知識

図２　文書構造の認識知識

はじめに」に「*2. 文書構造の認識*」がつながる）を識別し、次に図２(a) のＡからＥまでの知識により、階層構造を決定する。

４．２　参照構造解析部

一方の参照構造解析部は、参照用語解析部と参照構造決定部の２つに分れている。それぞれ、階層構造解析部の形態解析部と文間構造解析部に対応している。参照用語解析部は、原文と図表、それぞれの一文を解析し、３．２節で述べた参照箇所および図表見出しを示す「図」「表」などの参照用語と、図表番号を抽出する。ここで使われる知識は、図２(b) のａからｈまでである。参照構造決定部は、参照箇所と図表見出し間の参照構造を決定する。参照箇所と図表見出し、それぞれに対して、参照用語解析部により抽出された参照用語と図表番号が等しいかどうかの照合を行う。このとき用いられる知識は、図２(b) のｉからｊまでである。また、階層構造の決定に用いたのと同じ知識（図２(a) のＧからＩまで）を用いて、参照箇所間のつながり（参照箇所「*図1に*」に「*図2を*」がつながる）を決定する。

４．３　評価

技術文書（情報処理学会および自動制御学会の全国大会の予稿・講演論文集より収集）100 文書、およびビジネス文書（会議通知などを収集）100 文書に対して、評価実験を行った。対象の200 文書は、見出しや段落が必ず改行で区切られるように修正を行った。

その結果、技術文書は 96%、ビジネス文書は 85%の精度[7] で文書構造が抽出できることがわかった。構造が正しく抽出できない原因としては、次のものがある。

①「左図に示す」などのように物理的な位置により参照している。

②箇条書きに続く段落が、箇条書きに続くのか、その前の段落に続くのかが、内容を読まないかぎり、判断で

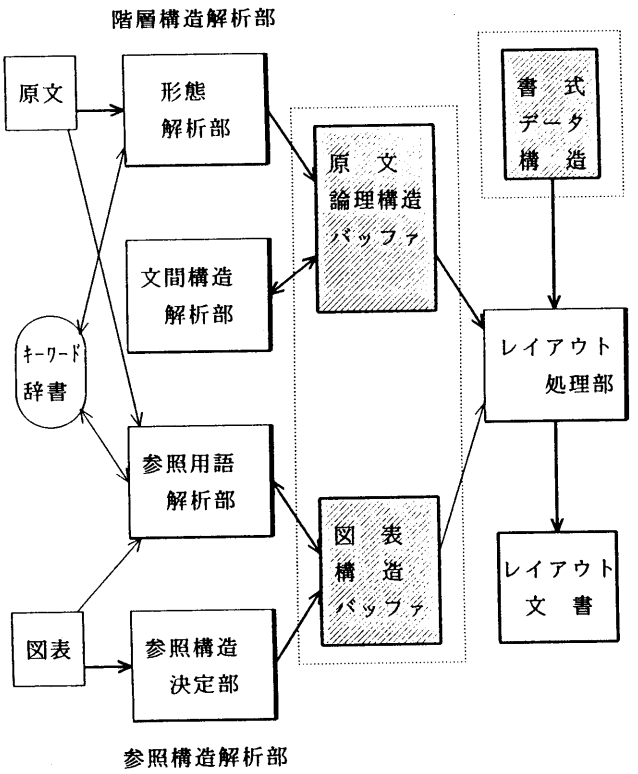


図3　システム構成

きない。

このような問題に対処する知識を殖やすことが考えられるが、知識の増加により、文書構造を一意に決定できなくなる。最終的には、人間に判断を委ね、その結果を学習していく機構が必要となってくる。

また、ビジネス文書では、技術文書に比較して精度が低くなっているが、これは、所属などに略称が使用されているため、固有名詞の抽出ができなくなっていることが原因である。略称の指示により、精度の改善が可能である。

5．整形処理の結合

抽出された文書構造の応用としては、第2章で述べたようなハイパーテキストとしての、構造に基づく編集・読解・検索など、新しい文書処理・管理など種々の応用がある。例えば、抽出された文書構造（著者名や所属、見出しなど）をキーワードに用いることにより、文書検索を容易にできる。

ここでは、構造に基づく編集の一つとして、整形処理を取り上げた[8][9][10]。

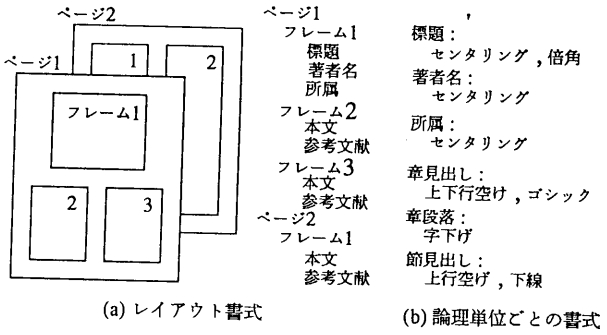
5．1　整形処理の概要

文書のページサイズや文章や図表を配置するためのフレームを定義する書式（図4(a)）と章見出しや段落などの階層構造を整形するのに用いる書式（図4(b)）は、書式ファイルとして文章データとは、別々に管理されている。

文章や図表は、書式データで定義されたページとフレームという2つの概念の下で割り付けられる。ページは整形の対象となる空間で、印刷媒体から余白などを除いた文書内容の表示（印字）領域を指す（図5(a)）。

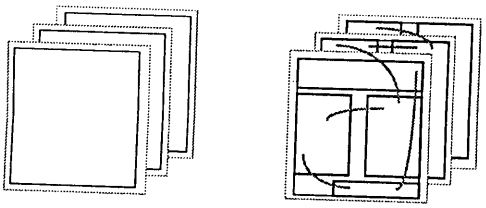
フレームは文書内容の配置領域をページ内で定義するもので、各フレームはページ内及びページを越えて全てリンクされる（図5（b））。このフレームの定義により、文書の基本的なレイアウト構造が定まる。これらのページ及びフレームは、図4の書式で定義されている。このようなレイアウト指定法は、n 段組と言うような定義方法に比べるとより自由なレイアウトが行える。

文章の整形処理はScribe等の整形処理システムと同様なので、以下に図表と関係する部分についてだけ述べる。



(a) レイアウト書式　　(b) 論理単位ごとの書式

図4　レイアウト書式と論理単位ごとの書式



(a) ページ　　(b) フレーム

図5　ページとフレーム

5．2　図表の自動レイアウト

従来の整形処理システムでは、図表はページに固定の位置、あるいは関連付けられた(anchored)段落の次の位置にしか配置できなかった。

これに対し本システムでは、図表は、自動的に出現順に参照位置の近くに配置される。そのときの割り付け位置は、ページ、フレームの概念に基づき、以下の手順で決定される（図6）。

基本的には、まずフレームに、次にページに割り付けられる。図表の大きさがページの大きさを越える場合には、それをページに入る大きさの余白が割り付ける。そのとき、図表の参照順序が逆転しないように、同一ページに複数の図表を含む場合には、全体のバランスを考慮して、配置の見直し（再配置）を行う（図7）。

文章は原則として、その流れに応じフレームに割り付けられる。但し、文書のタイトルや脚注が入るフレームでは、図表の割り付けを禁止する属性を付加しておき、図表領域が脚注やタイトルと重なったり、近接し過ぎたりするのを防ぐ。

図8に、同一文章と図表を使って、事なる書式でレイアウトした結果を示す。

6．あとがき

人間がレイアウト前の原文から文書構造を認識するときの知識を用いることにより、文書構造（階層構造と図表の参照構造）の抽出を行う技術を開発した。この技術を整形機能と結合することにより、自動的にレイアウトを行うシステムを開発した。

ここで、用いた知識は、形態的特徴のみに基づくものである。これに対し、人間が文書構造を認識するときには、形態的特徴以外に意味内容も用いている。

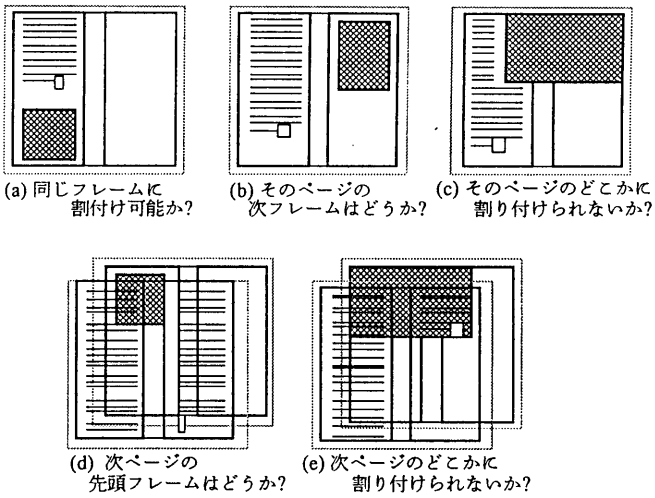


図6　図表の自動割り付け

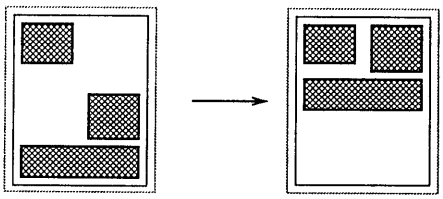


図7　再配置処理

文書構造を直接扱えるエディタと本文書自動レイアウト機能との結合を検討中である。

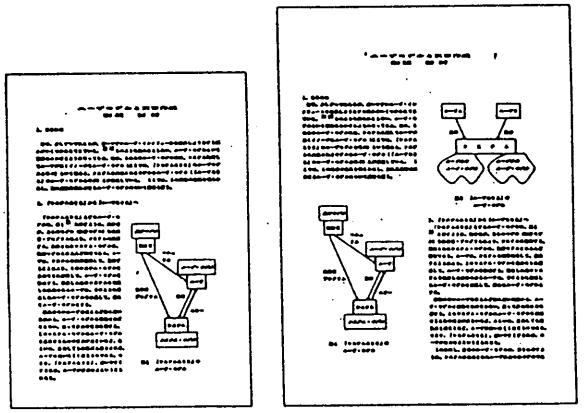
参考文献

[1] Conklin, J., 'Hypertext: An Introduction and Survey', IEEE Computer, pp. 17-41, Sept. (1987)

[2] Yankelovich, N. et. al., 'Intermedia: The Concept and the Construction of a Seamless Information Environment', IEEE Computer, pp. 81-96, Jan. (1988)

[3] Trigg, R. H. and Weiser, M., 'TEXTNET: A Network-Based Approach to Text Handling', ACM Trans. OIS, vol. 4, no. 1, pp. 1-23 (1986)

[4] Coombs, J. H., et. al., 'Markup Systems and the Future of Scholaly Text Processing', Comm. of ACM, vol. 30, no. 11, pp. 933-947 (1987)



(a) A4 一段組
A4 Single-Column

(b) B4 二段組
B4 Double-Column

図8　自動レイアウトによる整形結果

[5] Reid, B.K. and Walker, J.H., 'Scribe Introductory User's Manual', Unilogic Ltd. (1980)

[6] 'PageMaker User Manual', Aldus Corp.

[7] Doi, M., et.al., 'Research on Model Based Document Processing System DARWIN', Proc. of INTERACT'87, pp.1101-1106 (Sept. 1987)

[8] 岩井他、「知的文書処理システムにおける自動フォーマッティング機能」、情報処理学会第36会全国大会、pp.1299-1300 (1988).

[9] 山口他、「文書構造の解析機能を有する自動レイアウトシステム」、第4回ヒューマンインタフェースシンポジウム、No.1121, pp19-22 (1988).

[10] Iwai, I., et.al., 'A Document Layout System Using Automatic Document Architecture Extraction', CHI'89, pp.369-374 (1989).