

文書構造を用いた文書レイアウト機能の設計と実現

3 S - 3

遠瀬雅宏 早川栄一 並木美太郎 高橋延匡

(東京農工大学)

1. はじめに

論文など文字中心の文書作成において、見出しや図表の割付けをする手間は本質的なものではなく、削減することが望まれる。このような作業を手で行うときには、文書構造が意識されている。したがって、文書整形システムでこのような処理を行うためには、システムが文書構造を把握し、利用する必要がある。

そこで、我々はマークアップ言語で記述された文書構造をもとに、文書構造の隣接関係を利用してレイアウトする手法を考案し、文書整形システム上に実現した。本稿では、そのレイアウト機能について述べる。

2. 実装した文書整形システムの概要

レイアウト機能を実装した文書整形システムの全体構成は図1のとおりである。このシステムでは、書式設定ファイル、パラメータ設定ファイル、文書ファイルの三つの入力ファイルを必要とする。文書ファイルには、印刷する文書の他に文書構造を表す指令語が混在して記述されている。そして、書式設定ファイルとパラメータ設定ファイルの設定をもとに、文書ファイルに対して整形処理を行う。

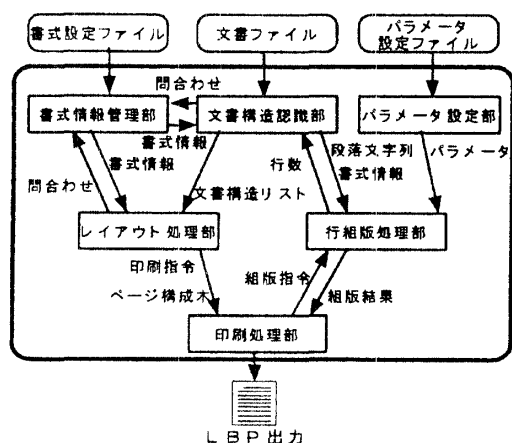


図1 文書整形システムの全体構成

図1の処理部のうち、レイアウト機能にあたる部分は、文書構造認識部とレイアウト処理部である。各処理部の機能は次のとおりである。

・文書構造認識部

文書構造をレイアウト処理で利用できるようにデータ構造に変換する。

・レイアウト処理部

文書構造認識部で作成されたデータを利用してレイアウト処理を行う。

・書式情報管理部

印刷する文書について書式全般の情報を管理している。

・行組版処理部^[1]

文章の整形処理を行なう。JIS X 4051「日本語文書の行組版方法」^[2]に準拠している。

・印刷処理部

レイアウト処理部から、1ページ分の処理結果を受け取り、実際の印刷処理を行う。

3. レイアウト処理の特徴

本機能の特徴は、次に挙げる隣接する文書構造間の性質を利用して行うことである。

・不可分性

隣接する文書構造が分離されてよいかという性質である。この性質を利用して、章や節の見出しがページ最下端に配置されるのを防いだり、図とその見出しが分離されるのを防ぐことができる。

・交換可能性

隣接する文書構造の順序を入れ替えられるかどうかという性質である。この性質を利用して、文書の配置順序を制御する。

4. 文書構造の定義

レイアウト処理で文書構造を利用するためには、文書構造を定義する必要がある。SGMLのDTD^[3]などのように動的に定義することも考えられるが、文書構造の隣接関係を利用するためには、動的な定義では困難である。そこで本システムでは、対象文書を科学技術文書とし、章>節1

Design and Implementation of a Document Layout Facility using Document Structures
 TOOSE Masahiro, HAYAKAWA Eiichi, NAMIKI Mitarou, and TAKAHASHI Nobumasa
 Tokyo University of Agriculture and Technology

>節2>節3の親子関係になる階層構造とそれらに属する次のような構造を定義する。

タイトル、著者、日付
見出し、文、図、表、図題、表題

これらのうち、タイトル、著者は、章と同じ階層で使用する特別な文書要素である。それ以外は、それぞれの階層に属する文書要素である。

5. 文書構造の抽象化

文書構造認識部で、4. で述べた文書構造を抽出し、レイアウト処理で利用できるようなデータ構造に変換する。まず、抽出した文書構造に次のような情報を付加する。

(1) 高さ

文書構造がもつ高さをシステム内部の単位系で表現したもの。

(2) 幅

文書構造がもつ幅をそれを収めるのに必要なカラムの数で表現したもの。

(3) 階層

どの階層に属しているかを表す。

(4) 種類

その文書要素の種類を表す。

上記の情報を付加した文書構造をリスト構造で表現し、3. で述べた不可分性を利用してグループ化（以下、構造群とする）する（図2参照）。

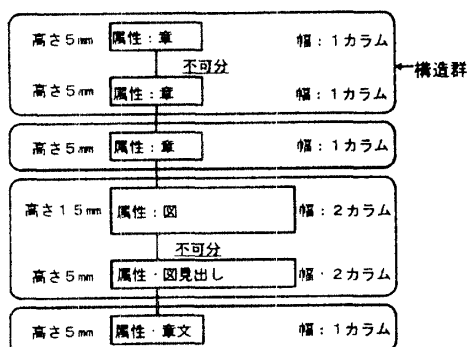


図2 文書構造リストと構造群

6. レイアウト処理

5. で述べた構造群のリストを並び換えることでレイアウトを行う。その基本アルゴリズムは次のとおりである。

このアルゴリズムでは、3. で述べた交換可能性を利用している。

- ①対象構造群がそのカラムに入るのならば、そのまま配置する。
- ②対象構造群がそのカラムに入らないのならば、次の構造群との交換可能性を見る。交換可能で、その構造群がカラムに入るのならば、構造群を交換しそのカラムに配置する。
- ③②で交換不可能な場合や交換可能な次の構造がそのカラムに入らない場合には、次のカラム（次ページ）に展開する。

①～③の処理を繰り返し行い、ページを満たしたときに印刷処理を行う。また、上記の処理中に複数カラム必要な構造群が出現した場合は、ページの空き領域を算出し、そのページ内に配置可能ならば、配置する例外処理を行う。

7. 実現

2. で述べた文書整形システムを我々の研究室で開発したプリントサーバ「浄書」上に実現した。情報処理学会論文誌の論文数編に対して本手法を適用した。そして、約75%の図表の配置について同等の結果が得られた。論文では図表をページ上端または下端に集める傾向がある。そのため、図表が参照位置より前に配置されていることがあり、そのような場合に失敗していた。

8. おわりに

本稿では、文書構造を利用したレイアウト機能について述べた。本手法によって文章と図表を統括的に扱ったレイアウト処理を行うことができた。今後の課題は、より多くの文書を処理し本手法の妥当性を検討し、改良することである。

参考文献

- [1] 遠瀬, 他: “JIS X 4051 の行組版方法に基づいた文書整形システムの設計と実現”, 情報処理学会人文科学とコンピュータ研究会, 30-1, 1996.
- [2] 日本規格協会: “JIS X 4051 「日本語文書の行組版方法」”, 日本工業規格, 1993.
- [3] 田中: “文書記述言語 SGML とその動向”, 情報処理 Vol. 32 No. 10, pp. 1118-1125, 1991.