論理構造に基づく技術文書処理

5P－7

高橋　聡子　折田　三弥彦　松本　秀和　谷藤　真也
(株)日立製作所、日立研究所

１．はじめに

文書処理には、アイディアを練る草稿段階、文書の内容を作成する編集段階、文書としての体裁を整える整形段階がある。しかしながら、従来のＤＴＰシステムでは、文書の内容とレイアウト情報が混在しており、文書全体を見通してアイディアを練る草稿段階には向いていない。また、章や節等の論理構造を単位として文書を作成するアウトラインプロセッサは、文書全体を簡単に見通せるため草稿段階に適しているが、文書のレイアウトを整える機能が不足しており、文書整形段階には適していない。そこで、私達は、既に編集・整形機能を備えた技術文書処理システムに、草稿段階を支援することができる文書構成編集機能を持たせることを検討した。

これを実現するためには、

（１）論理構造を持たない文書に対しても、章や節を単位とした編集機能を実現するための論理構造の自動抽出機能および編集機能（図１）

（２）文書整形を簡単にするため、レイアウト情報と論理構造を用いて、整形した文書を自動的に作成する自動レイアウト機能

が必要である。

以下、本稿では、論理構造の自動抽出機能について報告する。

２．論理構造の自動抽出機能

従来、論理構造を抽出するには、ユーザが論理構造を指定する方法、あるいは、テキスト情報を構文解析する方法が用いられている。しかし、ユーザ指定方式は、個々の論理構造をユーザが指定する必要があるため、操作性が悪い。また、構文解析方式は、様々な書体に対して、テキスト情報だけでは論理構造を完全に抽出しきれない。

本報告では、テキスト情報の構文解析に加えて、文字属性やレイアウト情報を利用して論理構造を自動抽出する方式を立案した。また、技術文書で特に重要な図表題も併せて自動抽出を行なう。

本方式による論理構造の抽出例を図２に示し、以下に説明する。章節題は文字属性、図表題はレイアウト情報を用いて抽出する。

（１）文字属性による抽出方式

・番号の無い章節題の場合、アンダーラインの有無や文字サイズ、フォントの種別等の文字属性を利用して題の階層関係を識別する。

（２）レイアウト情報による抽出方式

・図表内での位置と題の文字属性（文字サイズやフォントの種別）から、図表題としてふさわしいものを決定する。

・図表題が長すぎるために２行に渡っている場合、題の行間の距離と文字属性を利用して、図表題が１行か２行に渡っているかを判断する。

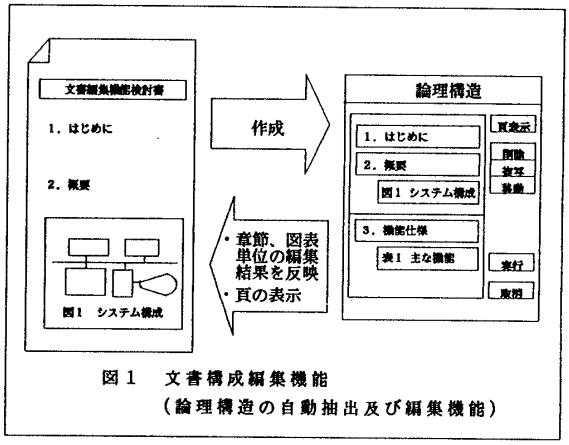

図１　文書構成編集機能
（論理構造の自動抽出及び編集機能）

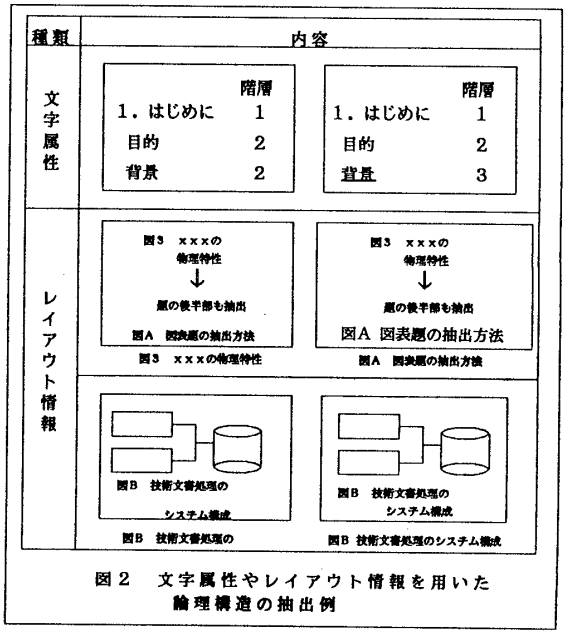

図２　文字属性やレイアウト情報を用いた
論理構造の抽出例

３．おわりに

文書の文字属性やレイアウト情報を用いた論理構造の抽出方式を提案した。現在、本方式の評価を行なっている。今後、文書の自動レイアウト機能についても検討していきたい。

参考文献

［１］福井他：文書構造を用いた自動レイアウトシステム、情報処理学会資料、文書処理とヒューマンインタフェース、２０－３、pp. １－１０（１９８８）。

A Study of Technical Document Processing Using Logical Document Structure

Satoko TAKAHASHI
HITACHI,LTD.