4W－3

文書論理構造の解析を応用したハイパーリンク自動作成支援システム

原　真男　楠元達治　大黒和夫　田村正文
東芝　情報処理・機器技術研究所

1. はじめに

近年、マルチメディアの情報（音声、テキスト、動画、イメージ等）を扱うシステムの開発が盛んになってきている。

我々はマルチメディア情報を統合的に管理するハイパーメディアシステムの開発を行っている。ハイパーメディアシステムでは、各マルチメディア情報をリンクで結び付けることにより、情報の表現を行っている。しかし、情報量の多さやデータ間のリンクの複雑さの為に、リンク情報を設定・指定するには、大変な労力を必要とする。

本報告ではべた書きの文書を文書構造解析することにより、自動的に論理構造を抽出し、この情報をもとにハイパーメディアシステムのリンクを自動的に作成するシステムの開発を行ったので、これを報告する。

2. システム構成

マルチメディア情報では、そのオブジェクトの情報量の多さや、各オブジェクト同士がリンクとしてネットワーク状に複雑に結び付いているため、それらのデータを管理するデータベースが必要となってくる。本システムではオブジェクト指向データベース（OODB）を利用することにより、これらのオブジェクトの管理を行っており、ハイパーエディタと構造化エディタはデータベースを仲介としてデータを共有している（図1）。

2.1. ハイパーエディタ

ハイパーエディタはハイパーメディア情報に、リンクを作成・編集するシステムであり、ビュアとしてプレゼンテーション用の情報を作成するアプリケーションである。

ハイパーエディタでは表示の基本単位であるページ（カード状の台紙）の上に音声、イメージ、テキストといったメディアとリンク・ボタンをレイアウトし、ハイパーメディア情報を作成する。

2.2. 文書構造解析

本システムでは文書の論理構造を抽出し、その結果をリンク情報へ変換している。

通常、文書は「章」「節」「項」といった段落より成り立っている。文書構造解析ではべた書きの文書より、文書中の各文の見出し番号などの形態的特徴を抽出し、これをもとに文書の前後関係を解析することにより「章見出し」「節見出し」「項見出し」といった階層構造（論理構造）の決定を行っている。

文書の形態的特徴の抽出では、見出しの先頭の数字部や記号部、あるいは文末の句読点などの規則にもとづいて各文単位に解析し、論理属性の決定を行っている。また同時に図表等の参照関係の抽出も行っている。

この解析結果を構造化エディタで表示し、ハイパーエディタで表示可能なデータへ変換している。

2.3. 構造化エディタ

構造化エディタは2.2.により構造化された文書の論理構造を木構造で表現することより、文書構造の編集を容易に行うことができる。木構造で表現しているため、大量の文書もその全体が把握で

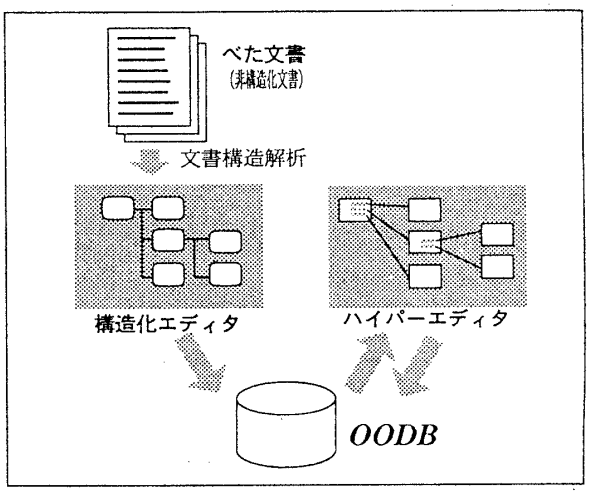

図1 システムの流れ

Automatic Linking Operation System with Document Structuer Analysis.
Masao HARA Tatsuji KUSUMOTO
Kazuo OOGURO Masafumi TAMURA
TOSHIBA Corp.Information Systems Engineering Laboratory

き文書作成・編集の効率化を支援する。

またアウトラインプロセッサ的に文書の枠組みから文書作成を行うことも可能である。

3. 自動リンク付方式

文書における「章」と「節」、「節」と「項」といった階層関係や図表の参照関係をハイパーメディアシステムにおけるリンク情報に置き換えることにより、通常の文書をハイパーエディタで利用することが可能である。ここで構造化エディタで表示されている一つの「見出し」(ノード)に注目すると、

- 内容部 (見出しに対応する段落の本文部分)
- 子供の見出し (一つ下の階層の見出し属性)
- 参照リンク情報 (内容部中の図表等の参照箇所を指し示している部分)

の3つの属性が従属体となる。注目した「見出し」を親として、これらの従属体をハイパーエディタにおける同一ページにレイアウトすることにより、ページを構成する。

3.1. 階層リンク

一つの「見出し」と、直下の従属体をハイパーエディタの同一ページにレイアウトすることにより、「見出し」+1(root のページ) のページが作成される。ここで「見出し」は、親となっているページと、従属体としてレイアウトされているページの2枚に存在する。その2枚のページの間にリンク (階層リンク) を張り、従属体の「見出し」をリンクボタンとする。このようなリンクを設定することにより親から順にカードをたどっていき、文書を読むことが可能となる。

文書の各見出しは通常の本の見出しと同様に目次として活用できる。これにより、見出しの属性を1枚のカードに配置し直接それぞれのカードにリンクを張る目次機能もある。

3.2. 参照リンク

文書構造解析では論理属性の解析以外にも参照リンク情報の抽出もおこなっている。これは文書中に「図1を参照」「写真1に示すように」といった様に明示的に他のメディアを指し示すような場合、これらの情報もリンク情報として得るものであり、文書中の参照形式とメディア本体 (図表本体) の参照形式を照合することにより参照リンク情報を抽出している。

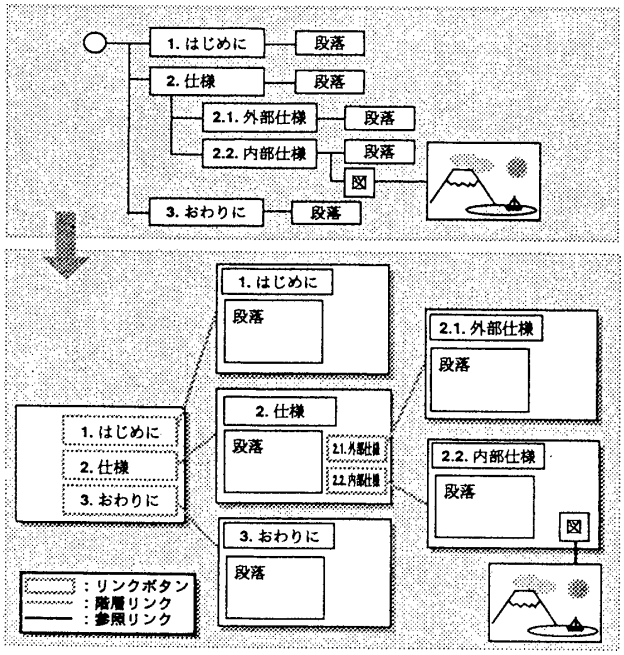

図2　構造化文書からハイパーエディタへの変換例

4. おわりに

このシステムでは、他のワードプロセッサ等で作成した既存文書を入力として簡単にハイパーメディア情報へ変換することが可能である。特に文書量が多量であり、かつ図表等の参照箇所の多いマニュアル文書等のビューアーとして利用することは有効であると考えられる。また構造化エディタによりアウトラインプロセッサのような文書の枠組みから文書作成を行い、その結果をハイパーエディタへ自動的に変換することも可能である。

通常の文書では「図」や「写真」「イメージ」といったような時間軸の流れがないオブジェクトが参照されているが、「音声」や「動画」といった時間軸の流れがあるものの参照抽出も行う予定である。

参考文献

[1] 土田他、「マルチメディア文書の論理構造を編集するための構造化エディタ」、情報処理学会第43回全国大会 10月 (1991)