5L−8

文書の見出し記号の解析に基づく木状論理構造生成の一手法

嶋　好博　古賀昌史　村上達也　藤澤　浩道
(株)日立製作所　中央研究所

１．まえがき

紙に印刷されている科学技術文献は膨大な量に及び、これをワークステーション等の計算機上に移しデータベース化するには多くの時間と人手を必要とする。このため、文献データベースの構築においては、紙面に印刷されている科学技術文献の文字を認識し、文字列の内容を解析し、所定の形式で自動的に登録することが期待されている[１]。特に、科学技術文献では、文書の章、節、項といった論理的な構成も技術情報においては重要である。そのため、紙上の要素のレイアウトを基に、章、節の文書の論理的な構成を自動的に抽出することを行なっている[２]。また、文書の構成要素を関係リンクで結合したハイパーテキストがテキストデータベースの知的システムとして注目されている[３]。

対象とする科学技術文献の論理構造では、章、節を表わす見出し記号に記載上のばらつきが文献ごとに存在するという問題があり、表記の上でのあいまい性、多様性を解消する必要がある。本研究の目的は、このような表記上のゆれがある章、節等の見出し文字列から文書の論理構造を自動的に生成することである。ここでは、本文中の章、節、項に対応した見出しを解析し、ツリー状の論理構造を自動的に生成するものであり、一次元に並んだ文字列で表現されているテキストファイルを階層的な木構造に変換することができ、文献データベースやハイパーテキストの自動構築に有効な技術である。

２．見出し記号の特徴と論理構造生成の課題

２．１　見出し記号の特徴

技術文書は、標題、著者名、抄録、緒言、本文、結言、謝辞、引用文献、付録などに区分して書かれている。また、これらの区分は分量に応じてさらに細分される。区分の名称は、章、節、項、目、の順に用いられる。このような区分は見出しとして表現されており、これを階層的に記述したものが、文書の論理構造である。

一般に、これらの文書の区分を示す見出しには、見出し名称と見出し記号が付けられる。ここで、見出し記号とは、数字記号だけ、または、文字記号と数字記号を組合せた文字列である。この内、数字記号としては、アラビア数字、ローマ数字、丸付き数字などがある。また、文字記号としては、英大文字、英小文字などがある。さらに、その他の記号として、丸括弧やピリオドが用いられている。これらの記号を組み合わせて、見出し記号にはポイントシステム「１．」「１．１」ように階層型の見出し記号と、「（ａ）」「（ｂ）」のような単独型の見出し記号がある。

２．２　論理構造生成の課題

（ａ）見出し記号の表記のばらつき

見出し記号の文字列の終端には、通常、区切り記号として空白文字が用いられる。しかしながら、数字とピリオドで構成される見出し記号では、空白文字による区切りがない見出しもある。また、数字や文字が括弧で囲まれている見出し記号では、空白文字による区切りがある見出しと、空白文字の区切りが無い見出しとがある。

見出し記号に使用する文字としては、数字記号、文字記号、ピリオド、括弧および、これらの組合せがある。数字記号としては、ローマ数字の大文字や小文字、アラビア数字が用いられる。文字記号としては、ローマ字の大文字や小文字が用いられる。括弧としては、通常、丸括弧（ ）が用いられるが、括弧が両側にある見出し記号や、片側だけにある見出し記号がある。この丸括弧は、数字記号や文字記号と組み合わされて用いられ、ピリオドを挟んだ階層型見出し記号とも組み合わされ、例えば、「（ａ．１．１）」のように記載されることがある。

（ｂ）論理構造の表記のばらつき

一般に、章、節、などには、階層型見出し記号が使用される。また、更に細かい構成要素には単独型の見出し記号が用いられるが、その使いわけは技術文書によりまちまちである。例えば、章、節を階層型見出し記号「１．」「１．１」を用いて表現し、それ以下の同じレベルの項を「（ａ）」「（ｂ）」で表現する文書に対して、章、節、項を「１．」「１．１．」「１．１．１」で表現し、それ以下を「（ａ）」「（ｂ）」で表現している文書もあり、文書によりまちまちである。

単独型見出し記号を使用した場合では、文献により、その階層関係を表現するための記号の選択はまちまちで

AN EXTRACTION METHOD OF DOCUMENT LOGICAL STRUCTURES BY ANALYZING HEADING MARKS
Yoshihiro SHIMA, Masashi KOGA, Tatsuya MURAKAMI, Hiromichi FUJISAWA
Central Research Laboratory, Hitachi, Ltd.

ある。例えば、見出し記号「ａ）」の下位に「ｉ）」「ii）」「iii）」の見出し記号が用いられている文書もあれば、見出し記号「（１）」の下位に「ａ）」「ｂ）」「ｃ）」の見出し記号を用いている文書もある。

３．　見出し記号の解析手法

対象とする技術文書は文字列の一次元的な集合である。したがって、文書の構成要素である見出しと本文は一連の文字列の集合であり、要素ごとの文字列の区分はなく一次元的に連続している。このような文書に対して、文字列を解析し、文書の論理構造を自動的に生成する。見出し記号の解析は、先ず、文書要素に対応した見出しとその本文を抽出する。抽出した見出しは、見出し記号と見出し名称で構成されている。この抽出した見出しから、次いで、見出し記号を抽出する。そして、これら見出し記号を手掛かりにして、ツリー構造を有する文書の論理構造を生成する。

見出し記号の抽出手法は、見出しの文字列から見出し記号としての条件を満足する部分文字列を求め、その桁位置を登録する。図１は見出し記号の抽出システムを示している。このシステムは、見出しの文字列を入力する機構、文字列が見出し記号としての条件を満足するかどうかを判定する機構、見出し記号として桁位置等を登録する機構からなる。文書の論理構造の生成手法は、見出し記号を順次、文書の読み順に従って入力し、注目する見出し記号の上位（親）の見出し記号を求め、ツリー構造を生成する。本手法は、見出し記号を順次、文書の読み順に従って入力し、縦型探索により注目する見出しの親見出し及び、注目見出しの木構造内での深さを求め、これによって、論理構造を生成する。

４．　木状論理構造生成の実験結果

文書を構成している見出しの文字列から、見出し記号を抽出し、その文書の論理構造を表す階層的なツリー構造を生成するとうい実験を行なった。この実験はワークステーション(1.2MIPS)上のソフトウェアで実行している。対象とした見出しは、英語の技術文献から採取したものであり、目視検査により文字入力結果を修正した誤りの無い文字列である。

実験に用いた見出しの文字列は図２に示すように情報関係の学術論文から採取したものであり、文献数は合計８４件である。文献当りの平均の見出しの数は、１２個である。見出し記号の抽出と論理構造の生成に要した処理時間は、文献当り平均８４．９ミリ秒であった。

５．　むすび

科学技術文献の章、節等の見出し文字列から文書の論理構造を自動的に生成する方式を開発した。文書に記載されている見出し記号を手掛かりに見出しの親子関係を求め、階層的な木構造として表示する。本方式は、記載の仕方がまちまちである見出し記号に対しても、論理構造を自動的に生成することができ、大量の印刷文書を読み取り、章、節等の要素の関係リンクを有するハイパーテキストの機能を有する将来型のマルチメディアシステムの構築処理に有用である。

なお、本研究は、通商産業省工業技術院大型プロジェクト「電子計算機相互運用データベースシステムの研究開発」の一環として、ＩＮＴＡＰ（(財)情報処理相互運用技術協会）がＮＥＤＯ（新エネルギー・産業技術総合開発機構）から委託を受けて、実施したものである。

参考文献

[1] Y.Shima,et al. :"A Segmentation Method of Color Document Images for Multimedia Contents Retrieval",Proc. RIAO 88,User-oriented Content-based Text and Image Handling (AFIPS), Cambridge,vol.2,pp.1001-1008 (March 1988)

[2] H.Yashiro,et al. :"A New Method of Document Structure Extraction Using Generic Layout Knowledge",Proc. Int. Workshop on Industrial Applications of Machine Intelligence and Vision (MIV-89),pp.282-287 (April 1989)

[3] 嶋好博ほか:"文献データ自動登録のための参考文献欄理解の一手法",情報処理学会第41回(平成２年後期)全国大会,2F-10,pp.4-168−4-169(1990年9月)

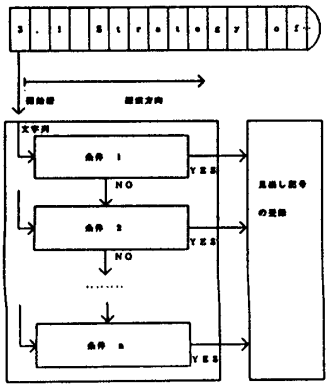
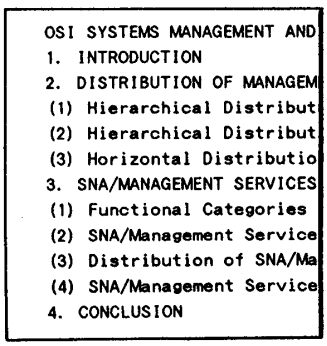
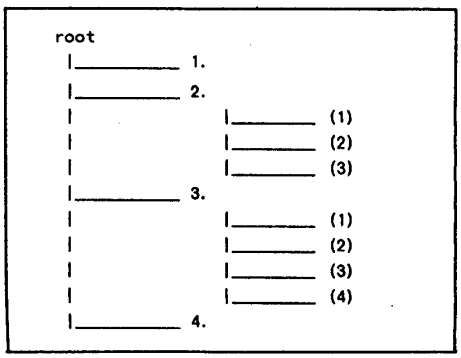


図１　見出し記号の解析システム



(a)見出し



(b)生成した木状論理構造

図２　論理構造生成結果の一例