ハイパーテキストにおけるリンク自動作成

3S－5

嶺岸則宏　大樫仁司

三菱電機(株)　情報システム研究所

1. はじめに

テキスト内で関連する項目を関連付けるハイパーテキストは、現在多方面で注目されている。しかし、関連付けを行なうハイパーリンクを作成する作業は、テキストの内容を理解した上で、人間が手作業で行なわなければならない。

個人利用の文書レベル、または小量の文書に対してハイパーリンクを作成する場合は、手作業で充分であり、逆に良いものが出来る。しかし、企業で使用するような、大量の文書や技術文書については人間の手作業の限界を越えてしまうため、リンク作成の自動化という要求が発生する。

このニーズに対し、従来、第○章・節など特徴ある文字列をもとにタイトル間のリンクを自動作成する報告があった [1]。我々は、SGML [2] 文書について、章・節などのタイトル間、および単語間のハイパーリンクを自動作成するシステムを試作した。本稿では、この試作システムについて報告する。

2. 概要

2.1 システム構成

図1にシステムの構成を示す。

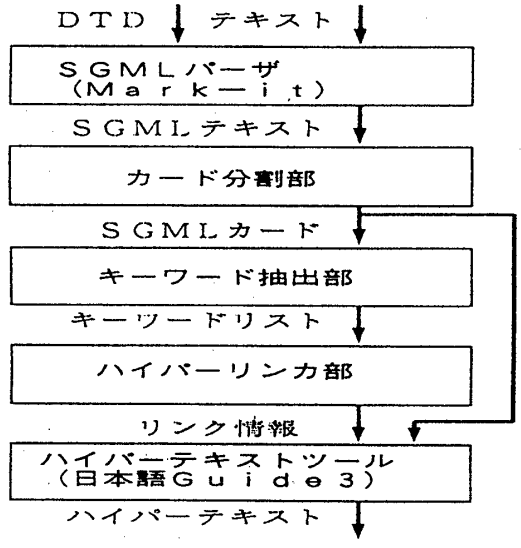

図 1: システム構成図

Automatic Link Generator on Hypertext
N.MINEGISHI, H.OHGASHI.
Computer & Information Systems Laboratory,
MITSUBISHI Electric Corporation,
5-1-1 OFUNA,KAMAKURA,KANAGAWA 247,JAPAN.

システム全体は5つの部分からなる。このうち、第1段の、テキストとDTD[1]を入力としSGMLテキストを出力するSGMLパーザ、および、最終段の、関連付ける箇所を指定したリンク情報と、SGML文書を入力すると、ハイパーテキストとして出力するハイパーテキストツールは、それぞれ市販品（Mark-It[2]、日本語 Guide3[3]）を用いた。

2.2 カード分割部

ハイパーテキストの表示形態におけるウインドゥ1枚分のファイルを以下ではカードと呼ぶ。カード分割部では、1個のファイルになっているSGML文書をSGMLのタグを基に、章、節、などで分割する。タイトル間のリンク作成は、ここで行なう。

2.3 キーワード抽出部

キーワード抽出部では、SGMLカードを入力し、形態素解析、不要語処理、を行ない抽出キーワードを出力する。システムでは、実際にリンクをはるキーワードの最終決定はせず、ユーザに判断を委ねる。キーワード抽出の結果を出現頻度順に候補として表示し、ユーザが適当に選択し、確定させる。

2.4 ハイパーリンカ部

リンク作成を行なうキーワードが確定した後、そのキーワード間でリンク作成を行ない、ハイパーテキストツールの入力となるリンク情報を出力する。

3. 実現方式

次に各モジュールの実現方式を説明する。

3.1 カード分割部

カード分割部では、1個のファイルになっているSGML文書をSGMLのタグを基に、章、節、などで分割し複数のファイルを出力する。章のレベルで分けるのか、節のレベルで分けるのか、文書の量によっても違うので、分割単位は別に指定することにした。

単純に文書の一部分を切り出すだけではなく、どのカードとして出力されたかの情報を、その文書の本来あるべき位置にタグ形式で埋め込む。すなわち原文書の階層構造が表現され、タイトル間のリンクは、ここで作成される。

[1] Document Type Definition 文書型定義

[2] Mark-It は SEMA SOFTWARE TECHNOLOGY 社の登録商標です。

[3] 日本語 Guide3 は OWL International Inc 社の商標です。

3.2 キーワード抽出部

キーワード抽出部では、形態素解析、不要語処理、を用いてキーワード抽出を行なう。形態素解析では、字種切り法を採用し、辞書を用いていないため（記号の一部のみ収録の辞書は使用）高速な処理が可能となった [3]。形態素解析後、各キーワードについて位置情報（カード、文字位置）を把握する。これは、後のハイパーリンカで使用するためである。

不要語処理では、品詞や経験則による不要語削除を行ない、残ったものをキーワード候補として出力する。ここで採用した経験則は、１文字単語の削除、出現頻度１回の単語の削除、等である。

キーワード抽出された全てのキーワードに対しリンクを作成すると膨大な量になってしまうため、キーワードを出現頻度順に表示しユーザに選択させる。ここで、システムで推薦する単語を反転表示させ、選択の参考にしてもらう。推薦の基準は、内部で重要度のパラメータを導入し、経験則で増減させる。ここで用いた経験則は、カタカナの文字列優先、出現位置別優先（タイトルか、本文中か）、である。

3.3 ハイパーリンカ部

確定したキーワードに対して、リンク情報を作成しハイパーテキストツールの入力仕様に沿った形式で出力する。

各キーワード毎に把握している出現位置を基に、１）同じカード内へのリンクは行なわない、２）もっとも単語が集中しているカードへリンクを作成する、３）集中がない場合は次のカードへリンクを作成する、というような方針によりリンク作成を行なう。

リンク作成の方針は固定であるが、それぞれの方針をモジュール化して実装しており、起動時に適当に組合せて実行する。

4. 実験結果

本システムを EWS(ME/R 7150 63MIPS) 上に実装し、実験評価を行なった。文書データには、一般の書籍（機械のメンテナンスマニュアル）を用いた。

4.1 時間性能

約 1MB のデータを処理した際の処理時間評価結果を、表１に示す。

表１：処理時間評価結果

	min	sec
カード分割	2	7
キーワード抽出	5	34
ハイパーリンカ		42

ファイルアクセスの頻度が多くなるキーワード抽出部が、やや遅い印象を受ける。この場合でも、処理速度は１５００文字／秒（すべて２バイトコードの場合）程度であり、充分高速な処理を行なっているといえる。

4.2 適正度

キーワード抽出部について、予め正解と思われる単語を選択しておきその正解に対する適合率、再現率を求めてみた。表中、キーワード候補は不要語削除等で残ったもの、キーワード推薦はさらに重要度により選択したものである。

表２：キーワード抽出の適合率・再現率

	適合率 (%)	再現率 (%)
キーワード候補	14.4	81.0
キーワード推薦	66.7	19.0

キーワード候補の場合で再現率が高く、キーワードの選択リストの中での洩れは少ないのがわかる。また、キーワード推薦の場合で適合率が高く正解に対して的を得た結果がでているのがわかる。それぞれ抽出できなかった単語を調べ、問題点を分析すると、ひらがなの単語に弱いことがわかる。これは、形態素解析で採用している字種切り法により、正確に解析なされていない場合であった。

ハイパーリンカ部の適正度の評価は、リンク作成方針の評価であり、現状では妥当であるかどうか、という程度でしかない。定量的な評価方法の検討から行なわなければならない。

5. おわりに

以上、ハイパーリンク自動作成システムの試作システムについて報告した。現状では、キーワード抽出の形態素解析、抽出経験則が練られていない等の問題点があるが、リンク自動作成およびその手法について有効性が確認できた。また、現段階では同じ文字列の単語同士のリンクのみの実現に留まっているので、今後の機能拡張が必要である。例えば同義語を考慮したリンク自動作成、単語から文章等へのリンク自動作成がある。今後は、これらの拡張機能の実現を検討していく。

参考文献

[1] 原、他: “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”, 第 47 回情報処理学会全国大会 4W-03,1993.

[2] ISO 8879: “Information Processing - Text and Office Systems - Standard Generalized Markup Language(SGML)”,1986.

[3] 大樫、他: “字種切り法による形態素解析の一改良”, 第 43 回情報処理学会全国大会 2G-04,1991.