3J-8 領域知識とマクロルールを利用した索引作成支援システム ＡＩＭＳ

空閑 茂起，勘座 浩幸，丸山 直利
シャープ(株) 情報システム研究所

１．はじめに

われわれは、分かりやすいマニュアルを短時間で効率的に制作するために、企画の段階からテクニカルライティング、索引作成、評価などマニュアル制作の過程を一貫支援するマニュアル作成支援システムの開発を行っている［１、２］。

マニュアルが分かりにくいといわれる原因は各種存在するが、重要な要因として検索したい事柄がすぐに、かつ、容易にひけないという索引に関係する問題が上げられる。

コンピュータ支援による索引作成の手法は、索引にしたい見出し語句に何らかのマーク付けを行う方法と、テキスト文書から自動的に見出し語句を抽出する方法とに大別される。

われわれは、索引見出しを重要語と解釈し、知識ベースを利用したキーワード自動抽出の手法により、自動索引作成の方法を採用したＡＩＭＳ－Ｉを試作し、マニュアル作成支援システムに組み込んで評価をしてきた。

今回、抽出ルールに加え、領域知識とマクロルールの利用により強化を図ったＡＩＭＳ－Ⅱの机上実験で、かなり良好な索引を作成できる見通しを得たので報告する。

２． ＡＩＭＳ－Ｉ

２．１ 概要

ＡＩＭＳ－Ｉは、書籍やマニュアルなどの索引を自動的に作成する目的で開発したシステムである。本システムは入力文解析、索引見出し抽出処理、ＡＩＭＳエディタ、頁割り付け、編集の５つの機能ブロックから成り立っている。

入力文解析モジュールはテキストファイルの文章を形態素、構文解析するモジュールである。この処理には、文章作成・校正支援システムＷＩＳＥ［３］の形態素、構文解析モジュールの解析結果を利用している。

索引見出し抽出処理モジュールはＷＩＳＥの出力結果から索引とする文字列の候補を抽出したり、抽出した候補を絞り込んだりして索引見出しを決定するモジュールである。

ＡＩＭＳエディタはＡＩＭＳで決定した索引見出しに対し、対話的に修正、追加、削除などの改良を行うモジュールで、自動抽出の不備を補足する役割を担っている。

頁割り付けモジュールは決定した索引見出しの出現頁情報を管理するモジュールである。ＡＩＭＳ－Ⅱで実現した頁間の優先順序の決定はこのモジュールで行う。

編集モジュールは見出しと出現頁情報をユーザの指定したフォーマットに従って出力するモジュールである。

これらのモジュールの中から見出し抽出処理モジュール、ＡＩＭＳエディタについて以下に詳しく述べる。

２．２ 索引見出し抽出処理モジュール

文書には構造的な情報と記述された内容に関する情報が存在する。分野の異なった書籍５冊とＯＡ機器のマニュアル５冊を選び、索引の見出しとこれらの情報との関係を調査した結果以下のようなことが分かった。

①各文書により索引抽出の規則は異なっている。
②タイトル文章の文字列から索引見出しが作られることが多い。
③説明文章の主語、目的語から索引見出しが作られることが多い。
④出現位置情報は索引見出しに関係する。
⑤出現頻度は索引見出しに関係する。

そこで、文書の構造情報を解析するスタイリスト

Indexing support system with domain knowledge and macro rules
Shigeki KUGA, Hiroyuki KANZA, Naotoshi MARUYAMA
Information systems labolatories SHARP Corporation

と呼ぶソフトウェアを開発した。このソフトウェアにより、テキストファイルはタイトル文（見出し文）と、本文に区分される。

索引見出しを抽出するのための抽出ルールは上の索引見出しの特徴①～⑤を反映したものである。それを表１に示す。

ルール種類	内　　容
本文ルール	特定文字列、強調記号など
見出しルール	並列名詞句、「の」名詞句など
頻度ルール	領域限定頻度、文書全体頻度

表１　索引見出し抽出ルールの枠組

たとえば、本文ルールの特定文字列ルールとは「Ｘは～Ｚと呼ぶ。Ｘは～Ｚと定義する。」などにおいて、Ｘに相当する語句を索引見出しとして抽出するルールである。

２．３　ＡＩＭＳエディタ

本エディタは、ルールで抽出した索引見出しが気にいらない場合に、索引作成者のほうで見出しを追加、変更できるようにしたエディタである。索引掲載頁などの制約条件に照らし合わせて、見出しの取捨ができるようにランク付けを行うこともできるようになっている。

また、取捨、追加の判断をより正確に行うため、エディタ中の索引見出しから、その索引見出をキーとしたＫＷＩＣを出力することができる。

３．　ＡＩＭＳ－Ⅱ

ＡＩＭＳ－Ｉをワープロマニュアルに適用し、専門家が作成した索引と比較することによりその性能を評価した。問題点は次の二つに集約できる。一つは精度の問題である。今一つは出現頁の記載方法に関するものである。精度的には、ルールベースの適用のみで８３％の見出しが抽出できている。出現頁はルールにヒットした頁をすべて拾い挙げていた。

ここでは、残りの１７％の見出しを抽出するために検討した事柄、及び出現頁記載の絞り込みの方法について述べる。

過去の当社のワープロ書院のマニュアルの索引について調査した結果、次のことが分かった。

①出現する索引見出しには偏りがある。
②新しい機能には新しい索引見出しが出現する。
③機器を特定すれば基本構造は共通部分が多い。
④表記には若干の揺らぎがある。

これらの機器に特有な知識をその分野の領域知識と呼び、それを扱う枠組をＡＩＭＳ－Ｉに付加した。

３．１　領域知識

領域知識は用語知識、構造知識、機能知識の３種類を用意した。このうち、用語知識は索引ベース、類語テーブル、揺らぎテーブルから成り立っている。索引ベースはワープロのシリーズの代表的機種のマニュアルから抽出した索引見出しに関する情報である。構造知識は機器のハードウェアの構造に関する情報から成り立ち、機能知識は機能に関する情報から成り立っている。

ここでは索引見出しを図１のように定義することで新しい機器、機能に対応するように配慮した。

```
                ┌もの＋（状態、属性）＋行動
索引見出し＝機能名

    ┌物理的もの  ┐ ┌状態┐ ┌行動１，行動２
機能＝ソフト的もの＝＝属性＝＝行動１＋処理
```

図１　索引見出しの構造

３．２　マクロルール

前に述べた抽出ルールで抽出された見出しと適用ルールの情報を蓄え、見出し出現頁を絞り込むのがこのマクロルールである。このルールは抽出ルール間の優先順位に関するルールとルールの適用に関する制約条件とから成り立っている。

４．おわりに

インデックス作成支援システムＡＩＭＳについて、概要とその実現方法の要点について述べた。ＡＩＭＳ－Ⅱはインプリメントの途中であり、まだ正式な実験と評価は行っていないが、机上実験により良好な結果が得られている。早期に実機で確認するとともに、別種の機器での本方式の有効性の検証、新規領域知識のための柔軟な追加機能、用語索引や用語集などへの適用について検討していきたい。

[参考文献]

[1]空閑茂起“マニュアル作成支援システムＲＯＭＡＮ”情報処理学会第35回全国大会7T-3 1987
[2]勘座、空閑“マニュアル作成支援システムの機能”情報処理学会第36回全国大会5U-4　1988
[3]空閑茂起“文書作成・校正支援システムWISE”電子通信学会　OS86-26 1986