3G－5

日本語全文情報の自動索引法

岩淵　保　須田彰夫　中田安昭

丸善株式会社

１．はじめに

近年、ペーパ・レスを合言葉に全文データベース・システムの構築が、盛んになってきた。構築時に問題になるのは、検索効率を高めるためのキーワード付与作業である。全文が既にコンピュータに採録されているので、その作業効率と標準化を考えると、自動索引システムを使用することになる。

実用化されている自動索引システムの問題点と、それを解決する為の方策と我々が実験したシステムの概要を報告する。

２．問題点

実用化されている自動索引システムの基本的な問題点を２点上げる。

（１）大量のキーワードが抽出される。「１，０００文字の文章から４００文字（１００ワード）程度のキーワードが抽出される。」[1]との報告もある。従って、検索結果には、多くのゴミ情報が含まれてしまう。

（２）辞書類を多用するので、その質の良否がシステムの評価に影響を与える。良質な辞書の作成と、その維持をするために、メンテナンス作業が永続的に発生する。また、ある文章ではキーワードとして不要な語句も、他の文章では必要となる場合もあり得るので、上記の大量キーワード群を不要語辞書にて単純に消し込むことも問題である。

３．解決策

我々が提案する解決策は、文章の主題を把握することで、辞書類の使用を極力排除しようとするものである。筆者らは、以前より自動抄録技術[2)-4)]を研究してきたが、その成果に基づき、自動索引法の実験システムを作成した。また、インデクサー（人間）は、文章中に存在しない語句をキーワードとして付与する。人間だからなせる技である。文章を抽象化することで似通った内容の文献を漏れなく検索することを可能にする。新聞記事では、１５～２０％、科学技術論文では、５０～６０％がこれらのキーワードである。我々は、これを想像キーワードと名付けて、その付与方法も実験した。

４．自動索引システムの概要

今回、新聞記事を対象にして実験を試みた。図１は、その概要の流れ図である。なお、以降、ＫＷは、キーワードの略号であり、単語とは、漢字・カタ仮名文字で構成される２文字以上の語句を言う。ＫＷの抽出・生成ブロックで、より上位に位置するブロックでのＫＷの方が重要度が高い。以降、図１に準じて説明する。

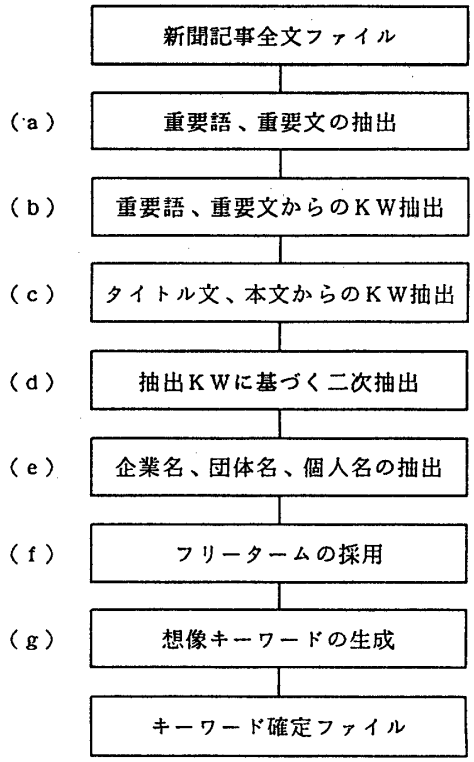

図１　自動索引システムの流れ図

（ａ）重要語、重要文の抽出
新聞記事全文を読み込み、「は」と主題化を示す複合辞[5)]の前の単語（重要語、主題を構成する語句）を抽出する。
次に、重要語の文章内での出現頻度をチェックして最重要語（最も頻度の多かった単語）を確定する。
文章を文単位で切り出し、最重要語と重要語の含有をチェックして、重要文を（主題を述べている文）抽出する。

（ｂ）重要語、重要文からのＫＷ抽出
重要語で、記事全文が１，３００文字以下であるならば、その全てをＫＷ候補単語とする。１，３００文字を越えるときは、その出現頻度４回以上の単語をＫＷ候補単語とする。
重要文で、その先頭から２文目以内で、「が」，「で」，「は」の直前の単語をＫＷ候補単語とする。但し、「では」は不採用とする。
複合辞で仕手，仲介，根拠，原因，立場，視点，相関を示す語の直前の単語をＫＷ候補単語とする。

（ｃ）タイトル文、本文からのＫＷ抽出
タイトル文（見出し文）に含まれる全ての単語をＫＷ候補単語とする。また、漢字・カタ仮名文字種類が混在する単語は、文字種で切り分けて双方ともＫＷ候補単語とする。

Automatic indexing method of Japanese language on full-text documents
Tamotu IWABUCHI , Akio SUDA , Yasuaki NAKATA
Maruzen Co.,Ltd.

本文に関しては、括弧内の単語とその直前の単語をＫＷ候補単語とする。但し、括弧内に「本社」、役職名がある時、括弧内のみ不採用とする。
「に」の直前の単語で２回以上出現の単語をＫＷ候補単語とする。
「聞き手」があるときは「インタビュー」をＫＷ候補単語とする。このような、誰でもが納得できる置き換え単語は、置き換えファイルに登録して置く。現在は、この一件だけである。
複合辞で確定，同格を示す直前の単語をＫＷ候補単語とする。
「は」と「が」にはさまれた単語をＫＷ候補単語とする。
同一文字種で３文字以上の「と」の前と後ろの単語をＫＷ候補単語とする。
鍵括弧内の単語をＫＷ候補語とする。文字種類が混在している単語であれば、文字種で切り分けた単語もＫＷ候補語とする。
（ａ）～（ｃ）で採用されたＫＷ候補語をユニーク（一意）にする。

（ｄ）抽出ＫＷに基づく二次抽出
インデクサーのＫＷ付与結果を分析すると、主要ＫＷをまず抽出し、次に文章中でそのＫＷの周辺から二次レベルのＫＷを抽出している。この癖を機械的に具現する。
全文中でＫＷ候補語の直後の「が」に続く単語も、ＫＷ候補語とする。
全文中でキーワード候補語の直前に「と」，「や」，「，」，「・」，「の」があるとき、その直前の単語もＫＷ候補語とする。
（ａ）～（ｄ）で採用されたＫＷ候補語をユニークにする。

（ｅ）企業名、団体名、個人名の抽出
（株）エレクトロニック・ライブラリで製作したＫＷ集ファイルを利用する。このファイルは、ＥＬ｜ＮＥＴを利用しているユーザに提供されるＫＷ集の機械可読版であり、検索頻度の高いＫＷ（コントロールターム、統制語）と、企業名、団体名、個人名が収録されている。ＫＷ集は、１年単位で編集し直すとのことで、ＫＷ集ファイル（統制語辞書ファイル）は、副次的に生成される。ＫＷ集ファイルにて、文章中の企業名、団体名、個人名をＫＷ候補語とする。

（ｆ）フリータームの採用
（ａ）～（ｅ）で採用されたＫＷ候補語と、ＫＷ集ファイルを突合し、ＫＷ集ファイルにある語句は、コントロールタームＫＷとして確定する。存在しない語句は、ディスプレイ画面に表示して、インデクサーがフリータームＫＷとしての採用を選択する。

（ｇ）想像キーワードの生成
インデクサーが想像ＫＷを付与する過程は、最も近い過去に生じた、内容類似の文章に付けたＫＷを想起し、その中から代表的なＫＷを付与していると考える。そこで、重要語から採用したＫＷ（最も重要度の高いＫＷ）を２個以上持つ文章を過去１年間の情報から検索し、その文章に付与されたＫＷのみを取り出す。次に、ＫＷの出現頻度をとり、一定の頻度以上のＫＷを想像ＫＷとして採用する。

５．自動索引システムの評価

以降、コントロールターム＝ＣＴ、フリーターム＝ＦＴ、想像ＫＷ＝想像と略号で記す。

（例題）日本経済新聞　９０年３月１２日朝刊　核心各論＝世界初の宇宙テーマパーク　スペースワールド社長　小池孜氏　地域活性化の先兵役に

（インデクサー付与）
リストラクチャリング（ＣＴ）・社長（ＣＴ）・観光開発（ＣＴ）・インタビュー（ＣＴ）・地域活性化（ＣＴ）・北九州市（ＣＴ）・新日本製鉄（ＣＴ）・日本興業銀行（ＣＴ）・福岡銀行（ＣＴ）・東京急行電鉄（ＣＴ）・ＪＲ九州（ＣＴ）・スペースワールド（ＣＴ）・テーマパーク（ＦＴ）・核心各論（ＦＴ）・小池孜（ＦＴ）・レジャーランド（想像）・福岡（想像）

（機械付与）
上記、ＣＴ，ＦＴは全て付与される。想像ＫＷでは、「福岡」は付与されるが、「レジャーランド」は、「レジャー施設」が採用される。また、上記以外に「宇宙」・「下関市」・「地域」が余分に切り出されてしまう。今後の改善点である。使用機器は、ＰＣ９８０１で行い、Ｃ言語で構築した。処理時間は、平均１分程度である。

６．おわりに

新聞記事を対象にして、全文情報自動索引技術の実験システムを報告した。実験で使用した全文情報は、１０件である。今後、１日分の全件数で実証をし、改善していく予定である。また、科学技術文献への適用も実験してみたい。最後に、（株）エレコトロニック・ライブラリー佐藤取締役情報部長初めインデクサーの皆様に多大なるご協力を賜りましたことを深謝いたします。

【参考文献】

１）重要キーワード抽出方式とその活用方法　内山、中村（東京電力システム研究所）情報処理学会　データベース・システム研究会報告８４－１９（１９９１．７．１８）
２）全文情報からの意味的情報の抽出と加工　岩淵、荒井藍沢（テレマティーク国際研究所）情報処理学会第３８回全国大会
３）自動抄録法　岩淵、荒井、藍沢（テレマティーク国際研究所）電気学会通信研究会ＣＭＮ－８９－２３（１９８９．７．１２）
４）全文情報からの意味的情報の抽出と加工　自動抄録法「ＴＥＬＥＭＡ－Ａ」について　岩淵、荒井　藍沢（テレマティーク国際研究所）第２６回情報科学技術研究集会発表論文
５）日本語表現文型　用例中心・複合辞の意味と用法　森田良行、松木正恵　アルク
６）キーワード集ファイル　エレクトロニック・ライブラリー　ＥＬ｜ＮＥＴ
７）新聞記事データベースにおけるキーワード自動抽出　神尾（日本経済新聞社）　情報管理　Ｖｏｌ．３２　ＮＯ．４　Ｊｕｌｙ　１９８９
８）日本語の文法（上）、（下）　国立国語研究所　大蔵省印刷局
９）自然言語の文法理論　郡司隆男　産業図書
１０）教師のための口語文法　渡辺正数　右文書院
１１）日本語概説　加藤彰彦、佐治圭三、森田良行　桜楓社
１２）いわゆる日本語助詞の研究　奥津敬一郎、沼田善子、杉本武　凡人社