自然言語処理 44－4
（1984. 7. 26）

新聞記事情報の階層構造に基づく記事分類・検索システム

藤崎博也・亀田弘之・河井恒
（東京大学　工学部）

1．はじめに

現代社会は高度情報化社会と呼ばれるように、通信・エレクトロニクスの著しい発達により大量の情報が生成され、また迅速に伝送され蓄積されるようになった。しかしながら、いかに大量の情報が流通していようとも、真に必要なものを迅速かつ的確に抽出・利用することができなければ、それは無にひとしいか、場合によっては却って有害ですらある。従って、大量の情報を収集・蓄積・管理するとともに、必要に応じて最も適切な情報を即座に供給することのできる検索システムが必要である。

このような大量情報の流通媒体であるとともにそれ故に検索需要の多いものとして、例えば新聞がある。新聞を対象とした検索システムは既に存在しており、例えば、日本経済新聞記事情報検索システム（別称：ＮＥＥＤＳ－ＩＲ）、中日新聞記事情報検索システム、ニューヨーク・タイムズ・インフォメーション・バンクなどが有名である[1～2]。しかしながら、これら従来のシステムには以下に述べるような問題点がある。

（１）データの入力が困難

（２）取扱われている情報の範囲が限られている

（３）情報の基本的構造に関する理解が必ずしも十分でない

本研究では、上記の問題（１）に関しては、朝日新聞社から新聞紙面作成・編集用データの供与を受け、（２）に関しては、究極的には新聞紙面上の様々な分野（但し、今回は新聞第１面の記事のみ）に関した記事を取扱い、（３）に関しては、従来の分類方式とキーワード方式との分析によって今回新たに得られた新聞記事情報の階層構造に着目し、これを積極的に利用することとして、この階層構造に基づく新聞記事検索システムを試作し検討した結果につき報告する。

なお、効率の良い検索を行うためには、予めデータを分類・整理することが必要なため、いわゆる情報検索システムは、一般に分類処理部と狭い意味での検索処理部とを持つ。以下では、混乱を避けるため“検索”を狭義で用いることとし、そのための前提となる分類を含まないものとする。

2．従来の分類・検索方式とその問題点

従来の分類・検索システムでは、大きく分けて２つの分類方式が用いられている。１つは分類表方式（テーマの木を利用する方式）であり、他の１つはキーワード方式（キーワードを利用する方式）である[3～7]。これに対応して、検索方式にもテーマ入力によるものとキーワード入力によるものとがある。また、入力に関しては、予め検索者に分類表（テーマの木）やキーワード集（シソーラス）を与えることにより入力を統制する方式（統制方式）と、そうではなく検索者が任意の語（文字列）を入力し検索することを許す方式（非統制方式）とがある。また、検索処理の方式としては、一括処理方式と対話処理方式とがある[8]。従来の分類・検索システムは以上のものの組合せとして記述することができる。

ここで分類方式についての長短を見ると、分類表方式の場合には、言語表現そのものではなく深層の意味に着目し、かつ、分類表という１つの体系に基づいて分類するので、

（１）分類が整然として明解である。

（２）従って、検索自体の作業は単純である。という利点があるが、

（３）分類の際には多大な人的作業を必要とするため、機械処理にはなじみ難い。

（４）１データは１テーマに分類することを基本としているため、複数のテーマに関連したデータを洩れなく検索することができない。

（５）採用した分類体系以外の観点からの検索が行い難い。
という欠点がある。

一方、キーワード方式は、文字列の表層上の特徴にのみ着目して処理するので、
（６）機械処理に向いている。
（７）既存の分類体系にとらわれず色々な観点からの検索ができ柔軟性に富んでいる。
という利点があるが、
（８）同義語や同字異義語処理が行い難いために検索洩れや不要なものの検索が生じる場合がある。
（９）検索によってどのような結果が出力されるかは、システムの作成者にも予想がつかず、全体を見通すことが難かしい。従って、検索洩れの有無やその程度を把握し難い。
という欠点がある。

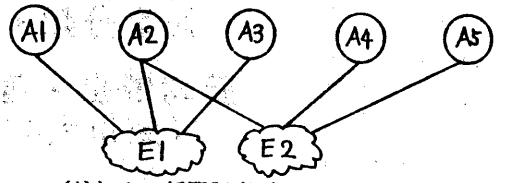
そこで本研究では、新聞記事の階層構造を積極的に利用することにより、このいずれの方式にも偏らず、これら両方の利点を取入れた分類・検索システムの試作・検討を行うとともに、上記の問題点（３）に対しては分類の自動化により、（４）に対しては関連の度合を考慮することにより、（５）と（９）に対しては、キーワード・テーマ両方式を組合せることにより、また、（８）に対しては、新たに“キー概念”の考えを導入・活用することにより対処することを試みた。

３．新聞記事の情報的側面

３－１　新聞記事情報の階層構造

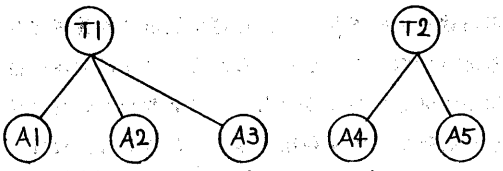
既に第２章で述べたように、従来の分類・検索システムでは、テーマの木に基づいて新聞記事を分類する分類表方式か、キーワードに着目して記事の分類を行うキーワード方式がのいずれか一方の分類方式が用いられている。ここでは、これら２方式の基本的な考えを整理して述べ、これらを包含する新しい考えを提示する。

新聞記事は、社会の新しい出来事（ＥＶＥＮＴ）に関する報道の文章であり、一般には出来事の１側面を記述したものである。図１は、この様子を



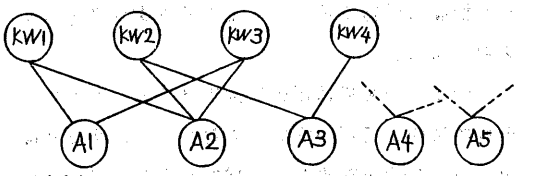
（注）Ａ：新聞記事（Ａｒｔｉｃｌｅ）

図１．記事と出来事との対応



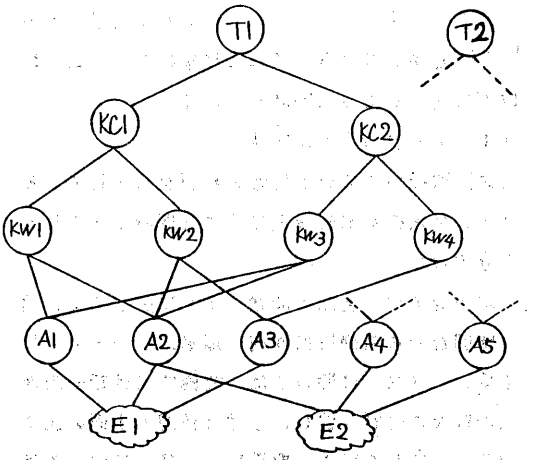
（注）Ｔ：テーマ（Ｔｈｅｍｅ）
Ａ：新聞記事（Ａｒｔｉｃｌｅ）

図２．分類表方式における情報構造



（注）ＫＷ：キーワード（Ｋｅｙ　Ｗｏｒｄ）
Ａ：新聞記事（Ａｒｔｉｃｌｅ）

図３．キーワード方式における情報構造



（注）Ｔ：（Ｔｈｅｍｅ）
ＫＣ：キー概念（Ｋｅｙ　Ｃｏｎｃｅｐｔ）
ＫＷ：キーワード（Ｋｅｙ　Ｗｏｒｄ）
Ａ：新聞記事（Ａｒｔｉｃｌｅ）
Ｅ：出来事（Ｅｖｅｎｔ）

図４．キー概念層を考慮した方式における新聞記事情報の階層構造

示したものであり、記事Ａ１、Ａ２、Ａ３は出来事Ｅ１に関連した記事であり、記事Ａ２、Ａ４、Ａ５は出来事Ｅ２に関連した記事であることを示している。図２、図３はこのような新聞記事に対して、従来の分類表方式及びキーワード方式の立場から見た情報構造をそれぞれ示したものである。

キーワード方式は図３に示すように通常複数の語の共起により記事を規定しており、これは基本的には語の表わす概念の共起により記事を規定していることに他ならない。一方、分類表方式は、実際に用いられている分類表の下位テーマ名がキーワードやキーワード列で構成されていることから、キーワード方式と同様に概念に着目した分類方式であるといえる。このように両方式は外見上異なるようにみえるが、ともにキーとなる重要な概念を媒介とする点で共通している。従って本研究では、キー概念という層を新たに設けることとする。図４はこのことを考慮した情報構造の図であり、テーマ（Ｔｈｅｍｅ、以下Ｔと記すこともある）・キー概念（Ｋｅｙ　Ｃｏｎｃｅｐｔ、以下ＫＣと記すこともある）・キーワード（Ｋｅｙ　Ｗｏｒｄ，以下ＫＷと記すこともある）・記事（Ａｒｔｉｃｌｅ、以下Ａと記すこともある）・出来事（Ｅｖｅｎｔ、以下Ｅと記すこともある）という５層から成っていると考える。このように、キー概念の層を新たに設けることにより、従来の分類表方式とキーワード方式とが密接に関係していることが明確化されるとともに、両者を統一的に取扱うことが可能となる。また、同義表現への言い換えの処理も可能となる。

以下、本論文中ではこのキー概念を表わすにも通常の語を使うので、混乱を避けるため語を「」で囲った場合はキーワードを、／／で囲った場合はキー概念を表わすものとする。例えば、１９８３年１月の時点での首相訪米に関する記事中では、キーワード「首相」はキー概念／中曽根首相／を言語表現化したものである。またキーワードとは、単語だけではなく単語列なども含み、同義表現とは、言語学的な意味のみでなく新聞用語的・時事的な意味も含むものとする。

３－２　隣接層相互の関係

ここで図４の情報構造における隣接層相互の関係について更に詳しく述べる。

（１）テーマ層とキー概念層の関係

いま、あるテーマＴ１に関する記事（複数）を考えると、これらの記事にはＴ１に関連する重要な概念（一般には複数個）が含まれている。これらの概念のうち、Ｔ１関連の記事すべてに共通しているものをそのテーマに関する“必須キー概念”、それらの記事すべてに必ずしも共起はしないが一部の記事にとって重要な概念を“選択キー概念”と呼ぶこととする。例えば、テーマ“中曽根首相訪米”に対しては、／中曽根首相／が必須キー概念であり、／レーガン大統領／、／アメリカ合衆国／、／訪問／などが選択キー概念である。

また選択キー概念の個数は、テーマＴ１がより抽象的なものである場合（例えば、外交・政治）には多く、テーマＴ１がより具体的なものである場合（例えば、衆院予算委員会）には少ないという傾向がある。

（２）キー概念層とキーワード層の関係

この両層の関係は、キー概念とそれの言語化された表現との対応関係を表わしているので、同義語の場合は１つのキー概念が複数のキーワードに対応し、逆に多義語の場合は複数のキー概念が１つのキーワードに対応する。

また一般に検索を行う場合、入力するキーワードは複数個のキーワードの組合せの形となる。例えば、“首相訪米”の記事をキーワードで検索する場合には、「首相」、「米国」、「訪問」という３つのキーワードの組合せとして検索することが考えられる。従って、これを｛首相，米国，訪問｝という１つの組合されたキーワード入力と考えることにする。しかしながら、同義表現を有するキーワードは、そのすべての同義表現に置き換えた形のものもキーワードとして保持しておかなくてはならないため、これを概念の形で保持していた方が全体の見通しが良い。従って本研究では、

語の表層上の表記よりも意味が重要であるとの立場に立ち、キーワードの意味する概念（ＫＣ）の組合せとして処理することとする。例えば、上記の例では、｛／中曽根首相／，／米国／，／訪問／｝という概念の組合せを考えることとするのである。本稿では、このようなキー概念の組合せを“組合せキー概念”と呼ぶこととする。これにより、同義表現などの見かけ上の多様性に煩わされることなく分類・検索処理することが可能となる。

（３）キーワード層と記事層の関係

この両層の対応関係は、所望のキーワードが記事文中に含まれるか否かにより定まる。

（４）記事層と出来事層の関係

記事とは出来事を記した文章のことであが、１つの記事文中に複数の出来事が述べられることもあるので、１つの記事が複数件の出来事にまたがって関連することもある。また、直接的関連・間接的関連などの関連の遠近を考慮した対応も考えられるので、一般にはその関係は複雑である。

なおここでは、隣接層間の関係について述べたが、この他に各層は内部に構造を有しており、分類・検索の際に利用可能なものもあるが、今回はその説明を省く。

４．新聞記事の言語的特徴

本研究で対象としている新聞記事文は、下記の言語的特徴を有する。

（１）用語的側面

ａ）通常の国語辞典には記載されていない固有名詞や政治・経済・科学などの専門用語が多い（例：中曽根、スト権、赤字国債、デオキシリボ核酸）。

ｂ）これらの複合語や造語・時事用語が多い（例：田中曽根）。

ｃ）略記や英文字表記されることがある（例：エンプラ、ＩＭＦなど）。

ｄ）外来語表記にはゆれが多い（例：スイートルームとスイート・ルーム）。

（２）文体的側面

ａ）１文中に大量の情報が記される事が多い。

ｂ）従って、１文字当りの文字数は、通常の文よりも長いことが多い。

（３）文章構成的側面

ａ）記事の冒頭で、「いつーどこでーだれがーどうしたーなぜ」といった概要・要点を述べる。

ｂ）それ以降でその内容を詳述する。

上記特徴（１）から、新聞記事を対象とする場合には同義表現の処理が必要である。また特徴（２）と（３）から、新聞記事検索には、記事の先頭に近い部分を重点的に利用すればよいことが推察される。

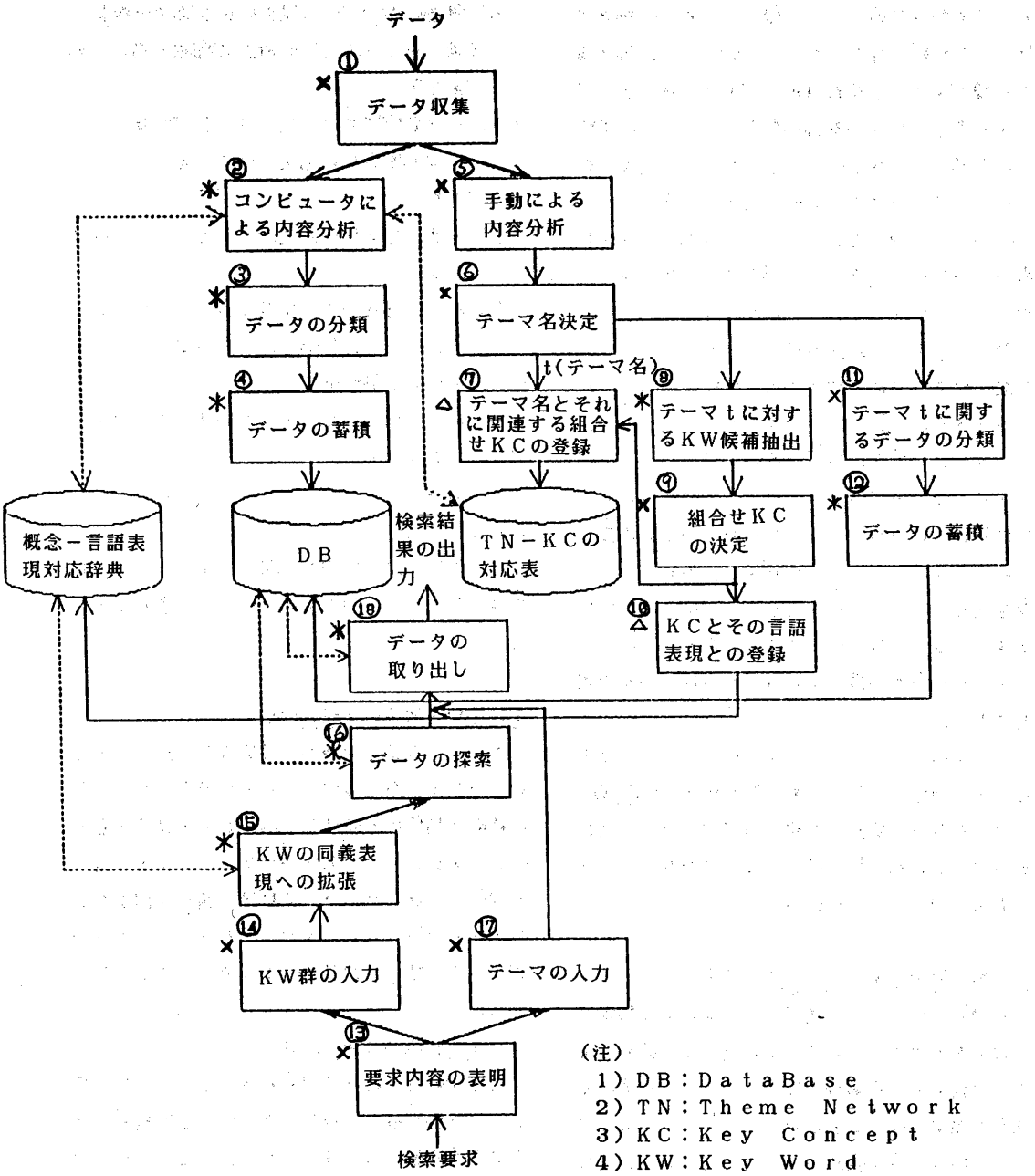
５．新聞記事情報分類・検索システムの概要

図５は前記の階層構造を利用した新聞記事情報分類・検索システムのブロック図である。四角いボックス内に処理内容を記してあり、実線の矢印 → は処理の流れを、点線の矢印 <-> はデータの参照を表わしている。また図中のＸ印は手動による処理を、＊印はコンピュータによる自動処理を、△印は１つ直前のボックス内の処理が自動化されればそのボックス内の処理も自動化することができるが、現時点では手動で行っている処理を表わす。

処理モードとしては１）自動分類を行うための知識獲得、２）自動分類、３）知識更新、４）検索という４つのモードがある。

まず１）の知識獲得モードでは、予め人間により分類されているデータに対して、各テーマ毎にキーワード候補の自動抽出を行い、次いでキーワードとキー概念を決定し、これらに基づいて“テーマ・キー概念の対応表”（以下これをＫｎｏｗｌｅｄｇｅ　Ｂａｓｅ　１＜ＫＢ１＞と呼ぶ）を作成し（図５中の①、⑤～⑦）、また“ＫＣ・ＫＷの対応表”（以下これをＫｎｏｗｌｅｄｇｅ　Ｂａｓｅ　２＜ＫＢ２＞と呼ぶ）を作成する（図５中の⑧～⑩）。

２）の自動分類モードでは、入力記事データに



（注）
1）ＤＢ：ＤａｔａＢａｓｅ
2）ＴＮ：Ｔｈｅｍｅ　Ｎｅｔｗｏｒｋ
3）ＫＣ：Ｋｅｙ　Ｃｏｎｃｅｐｔ
4）ＫＷ：Ｋｅｙ　Ｗｏｒｄ
5）＊：コンピュータによる自動処理部
6）Ｘ：手動による処理部
7）△：手動による処理部（但し、直前の処理部が自動化されればこの処理部も直ちに自動化可能）
8）⟶：処理の流れ
　⟷：知識の流れ

図５．記事分類・検索システムの概要

対してそれがどのテーマに属しているかの判定を行う。この際、各データ毎にそのテーマに関するキー概念をＫＢ１を利用して取出す。次いで、そのキー概念をＫＢ２を利用してキーワードに変更し、最後に、これらのキーワードが記事文中に含まれているか否かを文字列照合のプログラムによりチェックする。もし含まれているならばその記事はそのテーマに関連した記事であると判定し、そうでなければ無関係であると判定する。

なお、与えられた記事がこの手続きによって常に既存のテーマに分類されるとは限らない。これは分類体系が不完全であることを意味しているので、次の３）の手続きが必要となる。

３）の知識更新モードでは、２）で既存のテーマに分類されなかったデータに対し、１）の操作を施しＫＢ１、ＫＢ２を更新する。但し、この処理を既存テーマに分類されなかったデータが出現するたびごとに行うのは不経済であり、このようなデータがある個数以上になったら行うのが実際的である。

４）の検索モードでは、キーワードとテーマ名とによる検索が可能であり（図５中⑬～⑱）、また、これらを組合せた検索も可能である。特にキーワード検索の場合には、ＫＢ２を利用して同義表現への言い換え処理を行う。

６．機械支援によるキー概念の決定

適切なキーワードの抽出は、本来は文章理解を前提とすべきものであり、そのためには語彙・文法・一般常識・時事的知識などと、それに基づく推論・思考の能力とが必要である[9~11]。しかしながら、現在の計算機技術では未だこれを十分に実現することは困難である。従って本研究では、本格的な文章理解を行わず、下記の情報を利用してキーワード抽出を試みた。

Ａ）記事間における文字列の共起関係
　１）同一テーマの記事間（キーワードの抽出に適している）
　２）関連テーマの記事間
　３）関連のないテーマ同士の記事間（一般語（非キーワード語）の抽出に利用することができる）
Ｂ）同一記事内における文字列の出現位置
　１）記事の先頭から何文字以内（もしくは、何文字目）か
　２）記事の先頭から何文以内（もしくは、何文目）か
　３）記事の先頭から文字数にして（もしくは、文の数にして）何個目以内の位置かなど
Ｃ）文字列の品詞情報
Ｄ）文字種（漢字・片仮名・平仮名・ローマ字・アラビア数字など）
Ｅ）語の構文的役割（主語・目的語など）
Ｆ）情報提示の順序

ここでは上記の情報を利用してキーワード候補を抽出した後、人的作業によりキー概念を決定した。

７．分類・検索の実験と評価

本システムは、究極的には第５章で述べたように知識の更新を行うが、ここでは簡単のため、朝日新聞縮刷版の分類体系を利用して予備的実験を行い、本システムの基本的な部分に関する妥当性を調べた。評価に用いた素材は、朝日新聞１９８３年１月分第１面の記事約２７０（但し、天気予報・天声人語・素粒子などの連載ものを除く）である。

７－１　分類の実験と評価

評価実験は、朝日新聞縮刷版（１９８３年１月分）の分類体系に基づいて、まず（１）機械により記事を自動分類するとともに、（２）これとは独立に人間による記事分類を行い、次いで（２）の結果を正しいものとして（１）の分類結果を評価した。この際、（２）では従来のように与えられたテーマに関し記事が属するか否かの単純な２値的分類ではなく、Ａ）直接的に関連している（直接的関連）、Ｂ）間接的だが関連している（間接的関連）、Ｃ）ごく僅かに関連している

（希薄な関連）、D）全く関連していない（関連なし）、の４段階の関連の程度を考慮したきめの細かい分類を東京大学新聞研究所との協力で行った。図６は、このようにして行った評価結果である。実線・破線はそれぞれ再現率と精度[12]を表わしている。ここに、再現率とは分類されるべき記事のうち何％のものが正しく分類されたかの割合であり、精度とは分類されたもののうち何％のものが正しいのかを示す割合である。

この結果から、（A）この自動分類は、一般の検索システムと比べて再現率・精度共に優れている、（B）文字列照合範囲は第３文までとするのが適当、といえる。なお、この際分類洩れとなった若干の記事に関してその原因を調べてみたところ、

（１）稀にはキーワードが記事の後方で初めて出現する

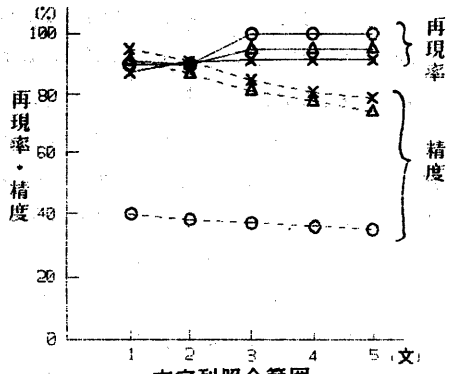
（２）記事内容が特殊（例：中曽根首相訪韓の際の“共同声明骨子”だけを取出したもの）であるなどによるものであった。（１）に対しては、必要に応じて検索範囲を変更することにより対処することができる（但し、これは経済性との兼ね合いにより決めるべきものである）。また、（２）のような特殊な記事は、それを１つの記事として独立させている点に問題があり自動分類自体の欠陥とはいえない。

７－２　検索の実験と評価

一般に、検索の良さの目安は、分類と同様再現率と精度により与えられる。

まず、テーマによる検索に関しては、これは分類の場合とまったく同じである。また、第５章で述べたＫＢ２に載っているキーワードでの検索も前節の分類と同じとなる。従ってここで問題となるのは、システムの分類体系とは無関係に任意のキーワードで検索する場合である。

表１は実際の検索要求例であるが、これらのうち、１０番台の要求例はユーザ自身も要求内容がはっきりしていないと考えられるので、本システムに備わっている対話機能により検索要求を絞ってもらう必要がある。そこで今回はこれらは評価の対象からはずし、残りのうちの２）と９）の例を実際に検索してみたところ良好な結果を得た。



文字列照合範囲

（注１）実線：再現率（＝Ａ／Ａ’）
破線：精度（＝Ａ／（Ａ＋Ｂ））
但し、Ａ：分類されるべき記事数
Ａ’：正しく分類された記事数
Ｂ：不用な記事数
（注２）横軸は、文字列照合を行う文の数
（但し、記事の先頭からのもの）
（注３）テーマは”中曽根首相訪韓”
（注４）○：直接的関連，△：間接的関連
×：希薄な関連

図６．文字列照合範囲と再現率・精度との関係

表１．実際に発生する新聞検索要求

１）金大中事件に関する記事	２）国会議員の自殺
３）小林秀雄の死去	４）西独、緑の党
５）えん罪事件	６）コウモリダコの化石
７）アメリカ合衆国議会下院に関する報道	
８）発見された発ガン物質	９）ＳＳ２０の極東移転
１０）国債	１１）憲法論争
１２）アジア諸国の実情	
１３）臨調	１４）田中角栄
１５）　核ミサイル	
１６）雪	１７）海
１８）大学入試	
１９）ビザ	２０）河川

８．おわりに

本稿では、新聞記事情報の階層構造に基づく記事分類・検索システムを提案した。このシステムは、従来のキーワード方式と分類表方式の長所を備え持つ一方、その短所を解消することをめざし

たものである。更に、キー概念に着目しこれを活用することにより、語の見かけ上の異なりに煩わされることなく同義表現の処理を行うことができる。このシステムをミニコンピュータ上に実装し、小規模のデータに適用してその基本的有効性を確かめた。しかしながら、広範囲・大量のデータを用いてのシステムの評価及び新規テーマの出現に対する知識の更新処理の有効性に関しては、今後も引き続き検討を行う予定である。

なお、この研究は、昭和５８年度科学研究補助金試験研究（２）課題番号５７８９００１７の補助によるものであり、データを供与された朝日新聞社、貴重な御意見を賜わった東京大学新聞研究所の荒瀬豊教授・山本泰助手（現在、教養学部助教授）ならびに東京大学文学部の国広哲弥教授・荻野綱男助手（現在、埼玉大学講師）をはじめ研究に協力された工学部藤崎・広瀬研究室、新聞研究所荒瀬ゼミの諸氏に謝意を表する。

９．参考文献

１）笹本　他：“オンライン情報検索，”地人書館，1981.
２）“新聞記事をデータベースとする情報検索についての研究（中間報告書），”東京大学新聞研究所内部資料，1981.
３）“新聞社説索引集－朝日・読売・毎日・サンケイ・日経－（1981年版），”東京大学新聞研究所，1983.
４）“ニュース・シソーラス，”中日新聞本社，1982.
５）“ＮＥＥＤＳ－ＩＲ　シソーラス，”日本経済新聞社，1982.
６）“ＮＥＥＤＳ－ＩＲ　補助キーワード，”日本経済新聞社，1982.
７）“新聞切抜・写真分類集　昭和48年版，”日本新聞協会，1983.
８）“ニューヨーク・タイムズ・インフォメーション・バンク　検索事例集，”日本経済新聞社.
９）堀　他：“シナリオを用いて構造化されたキーワードをアブストラクトから抽出する一手法，”計算言語学研究会資料，vol.25，情報処理学会，1981.
10）猪瀬　他：“シナリオを用いる論文抄録理解・作成援助システム，”情報処理学会論文誌，vol.24,PP.22-29，1983.
11）Lebowitz: “Generalization From Natural Language Text, “Cognitive Science,vol.7,PP.1-40, 1983.
12）Vickery: “Techniques of Information Retrieval,”Butterworth & Co., 1970.
13）藤崎　他：“新聞記事情報の階層構造を利用した記事検索システム，”情報処理学会第28回全国大会講演論文集，1984.
14）“新聞資料センターの利用実態に関する調査報告，”東京大学新聞研究所内部資料，1983.