データベースシステム 90－9
（1992. 9. 11）

統合型大規模テキストデータベースへの自動索引とその評価

木本晴夫
ＮＴＴ情報通信網研究所

内容が多種類で大規模な統合型のテキスト型データベースへの自動索引とその評価について述べる。自動索引の対象は、新聞37紙、雑誌約120紙で、新聞(社説、コラム、死亡、決算、新製品等の記事を含む)、専門紙、業界紙(化学、株式、原子力、建設、住宅、食品、繊維、その他)、雑誌(週刊誌、経営、食生活、技術雑誌、広告業界誌、調査レポート、業界雑誌、その他)で極めて多種多様である。また、データベースの規模は極めて大規模であり、現在約230万件で、毎日の追加更新は、約2000件で、これにキーワード自動付与を行うために実験システムを構築した。キーワード抽出の方式はフリーキーウード方式でキーワード重要度評価をおこなう。この実験システムの評価は、評価対象とした新聞２紙でキーワードの再現率・適合率でおのおの、70.9%、30.7%であった。自動索引システムの評価方法の問題点とその評価用データベースの必要性についても述べる。

AUTOMATIC INDEXING OF AN INTEGRATED LARGE SCALE TEXT DATABASE AND ITS EVALUATION

Haruo Kimoto
NTT Network Information Systems Laboratories
1-2356, Take, Yokosuka-shi, Kanagawa, 238-03 Japan

This paper describes automatic indexing of an integrated large scale text database and its evaluation. Included in the database are all kinds of articles from the 37 newspapers and about 120 magazines. Among 37 newspapers are daily newspapers, trade papers and other professional papers. And among 120 magazines are amusement magazines, magazines for daily life, magazines for professional engineers and others. The database contains 2.3 million articles, and about 2 thousands articles are added daily. The experimental system was made in order to index these articles automatically. The indexing scheme is free term indexing and terms are weighed according to their importance in the article. The system was evaluated, and the keyword recall rate was 70.9% and the keyword precision rate was 30.7%.

1．まえがき

近年、商用、社内用を問わずデータベース(ＤＢ)の数が急増しつつある。従来からのテキスト型ＤＢや画像ＤＢに加えて、映像ＤＢも増えている。これらのデータベースを利用するためのデータへのアクセス方法もハイパーテキストやフルテキストサーチ等、種々考案されているが、ユーザに与える満足度を含めた実用性での評価は一般的には確立されたとは言い難い。一方、古典的とも言えるキーワードによるデータベース検索には、その広い認知度と実績から根強い需要がある。

本稿では、この根強い需要のあるキーワード検索のための自動索引システムを内容が多種類で大規模な統合型のテキスト型データベースへ適用するための実験システムの構築とその評価について述べる。またシステム評価方法とその問題点について述べる。

現在、商用のオンラインデータベースの検索方法の基本方法は、キーワードを用いたＡＮＤ・ＯＲ検索が大部分である。1980年代までは、検索のために、個々の文献に付与されるキーワードは人手で選択・付与されていた。ここ数年間では、このキーワード付与の作業を自動化するケースが増えてきた。ただし、システムが付与するキーワードの中には、文献の主題には、ほとんど関係のないキーワード(以後、ノイズと呼ぶ)も数多く含まれていて、システムの出力を後で人がチェックするか、完全自動化しても、多くのノイズを含んだままであった。わずか400～500文字からなる新聞記事に対して100個以上のキーワードが自動付与されるのも希ではない。このような多くのノイズが付与されていると、当然のことながら、検索をしてみると、多くの関係のない文献まで検索してしまう。付与されたキーワードを人がチェックするにしろ、多くの関係のない文献のなかから、人が欲しい文献を選択するにしろ、人にかかる負担はかなり大きい。これらの理由から精度の高いキーワード自動抽出システムが望まれている。

今回の自動索引の対象は、新聞37紙、雑誌約120紙で、新聞では、社説、コラム、死亡、決算、新製品等の記事を含む。また専門誌・業界紙の種類は科学、株式、原子力、建設、住宅、食品、繊維、その他で、雑誌は週刊誌、経営、食生活、技術雑誌、広告業界誌、調査レポート、業界雑誌、その他、である。今回の自動インデクシングの対象は極めて多種多様であり、文書のほとんどの分野にわたっている。同じように文書を処理対象とするシステムに、機械翻訳システムがあるが、制限された文法と、技術マニュアルのような限られた分野と、システムの処理の前後に、前編集や後編集を必要とするといった例とは、比較にならない多様性を持ったものを処理の対象としている。

また、このデータベースの規模は極めて大規模で、現在約230万件で、毎日の追加更新は、約2000件である。内容が多種多様で大規模なデータベースが対象では、インデクシングが極めて困難である。

このようなデータベースにキーワード自動付与を行うために実験システムを構築した。本実験システムの特徴は以下のとおりである。精度の高い日本語解析を実現するために、全分野対応型大規模辞書を備えている。キーワード抽出の方式はフリーキーワード方式で、キーワード重要度評価をおこなう。このために重要度評価点数テーブルを作成して利用する。このテーブルはユーザによる独自配点設定が可能である。最後にキーワードの重要度を利用して、個々の文献(記事)の長さに対応したダイナミックキーワード数制限をおこなっている(フィルタ方式と呼んでいる)。

この実験システムの評価をおこなった。評価対象は新聞２紙であり、その結果はキーワードの再現率・適合率でおのおの、70.9%、30.7%であった。また、文献(記事)単位での評価も別におこなった。

上記のように、キーワード抽出の精度向上の種々の努力をして評価をおこなったが、キーワード抽出結果に対しては絶対的な評価尺度が現状では無い。なぜなら、人手キーワードを評価の基準にしているが、この基準が個人の主観が入る要素があり、客観性が十分とは言いきれないからである。

今後は、客観性のある評価用データベースを作成して、システムの評価をおこない、システム相互の比較がおこなえるようになることが望まれる。

２．内容が多種類で大規模なテキスト型データベースについて

今回の自動インデクシングの対象となるデータベースの内容は以下に示すとおり多種多様で、新聞37紙、雑誌約120紙から構成されている。企業内のデータベースでは、論文だけ、特許だけ、あるいは、技術資料だけといった単一的な内容のデータベースが多いが、一般向けの商用データベースでは、その数こそ少ないが、内容が多種類にわたっているものがほとんどである。

①新聞全国紙(社説、コラム、死亡、決算、新製品等を含む)
②新聞地方紙(上に同じ)
③専門誌(技術雑誌、業界雑誌等)
④業界紙(化学、原子力、建設、住宅、食品、繊維、株式、広告業界紙等)
⑤雑誌(週刊誌、経営、食生活、調査レポート等)

また、このデータベースの規模は現在約230万件で、毎日約2000件の追加更新がおこなわれる。この約2000件に毎日、キーワード自動付与を行う予定である。

３．自動インデクシングの必要性と困難性

現在は、このデータベースへのインデクシングは人手によっておこなっている。このために、多くの作業者が必要であり、費用がかかっている。またインデクシングの対象が、新聞が多いためにその作業は、新聞ができた後の深夜作業とならざるを得ず、早朝からのサービスに間に合わせるために時間との勝負になっている。このために、満足なインデクシングが出来る状況ではない。また作業が人の主観にたよる部分が多いので、インデクシングの結果が均質でない。などの問題点があり、主として費用の観点からインデクシングの自動化はなされねばならないものとなってきた。

一方、自動インデクシングの困難性については種々述べられている。内容が多種類で大規模なデータベースへのインデクシングが、単純なデータベースへのそれと比較して、困難な理由は次のとおりである。①インデクシングの語彙数が極めて大きくなること、またその種類が増えること、例えば、「こんにゃくの日」のような名詞句のキーワードが求められる。②インデクシングの方法・基準が、本当は単純なものが欲しいが、たとえば新聞の社説から雑誌のゴシップ・スキャンダル記事まで幅の広い多くの種類の記事・文献を処理しなければならないために、単一のインデクシング基準では適用が不可能。③自動化の処理量が極めて多きいために処理時間が多くかかってしまうこと、などである。

4．　内容が多種類・大規模データベースのキーワードの自動抽出方式

4．1　キーワード自動抽出方式の特徴

今回設計したキーワード自動抽出実験システムの特徴は次のとおり。使用するシステムはＩＮＤＥＸＥＲシステムである。このＩＮＤＥＸＥＲシステムについては、既に報告している[1]。処理の流れを図１に示す。処理は、まず、日本語解析をおこなう。日本語解析のために50万語のユーザ辞書と約43万語の日本語解析用辞書をマージした約93万語の辞書を使用する。次にフリーキーワード方式でキーワードを抽出して、それが複合語の場合は複合語分割をおこなう。その出力に対して約50万語のユーザ辞書との照合をおこなう。ユーザ辞書と照合した語を対象として、語の出現頻度などに基づいてキーワードの重要度評価をおこなう。このために重要度評価点数テーブルを作る。このテーブルの点数はユーザが自由に設定できるので、ユーザによる独自配点処理が可能である。あるいは、分野別に点数を使い分けることもできる。最後に、キーワード数制限によって重要度の順番でキーワードを決める。この場合、キーワードの数は文献(記事)の長さに応じて文献(記事)毎にダイナミックに決める。重要度の低いキーワードをふるいおとす仕組みとなっているので、この方式をフィルター方式と呼んでいる。

4．2　方式の選定

キーワード抽出の方式を選定するにあたって、３つの方式を検討した。それらは次のとおりである。これらの方式の精度比較を表１に示す。

［方式Ａ］フリーキーワード抽出によって抽出されたキーワードの中で、キーワード自動抽出実験システムで利用する日本語辞書に登録されている語をキーワードとして抽出する方式。複合語の分割はしない。

［方式Ｂ］４．１節で述べた方法。

［方式Ｃ］フリーキーワード抽出によって抽出されたキーワードをそのままキーワードとする。複合語の分割はしない。

表１　キーワード抽出の方式の精度比較

	方式Ａ	方式Ｂ	方式Ｃ
システムが抽出したキーワード数（１文献あたりの数）	２０（個）	４７（個）	２８（個）
人手抽出キーワードと同じキーワード数（１文献あたりの数）	２．５（個）	５．３（個）	４．３（個）
人手抽出キーワードの何％が再現されているか（１文献あたりの数）	３０（％）	６０（％）	５０（％）

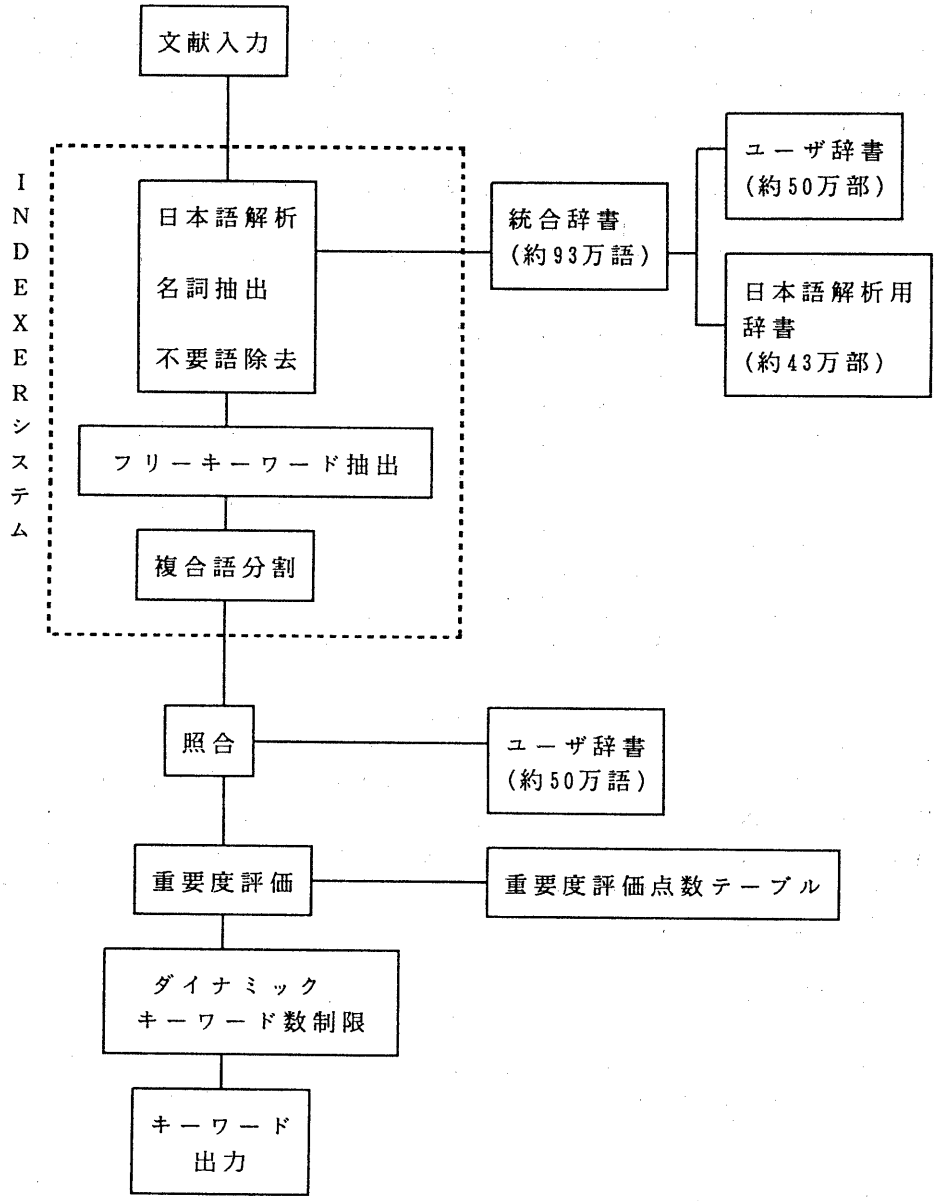

図1　キーワード自動抽出実験システムの処理の流れ

方式Aは使用する日本語辞書の登録語数が、今回の処理対象としているデータベースの大規模性と多種類性に対応するには不十分であり、適用不可能である。方式Bはデータベースの大規模性と多種類性からの要求に答えるもので、今回はこの方式の採用を検討している。方式Cは複合語の中のキーワードを抽出できないので使いものにならないが、精度的にはかなりのところまで実現できている。

4．3　独自配点処理の採用

重要度評価点数テーブルを採用して、独自配点処理を可能にした。このテーブルの点数はユーザが自由に設定できる。得点の設定は、個々の語について、語自身の性質と語の文献中での統計的性質の組み合わせ分類によっておこなう。語自身の性質とは、例えば、その語が企業名であるか、人名であるか、地名であるかなどである。語の統計的性質とは出現位置、出現頻度などのことである。重要度評価点数テーブルの一部を表2に示す。

このテーブルは、文献の分野別に配点を使い分けるためにも使用できる。

表2　重要度評価点数テーブルの一部

	企業名	人名	地名	一般名詞
基礎点	30	30	25	20
強調語加算	10	10	10	10
文頭語加算	20	20	20	20

4．4　キーワード数のダイナミックな制限

一般に、フリーキーワード方式によって文献からキーワードを抽出すると、文献の長さ(文字数)に比例した数のキーワードが抽出される。実際の人手キーワードの数を分析してみると、ある一定数に加えて文献の長さに比例してキーワード数が増えている。この特性を模擬式で近似した。そして、キーワード自動抽出の精度が100%完全でないことを考慮に入れて、キーワードの数を一定倍率ふやした。採用した模擬式は次のとおり。

$$採用キーワード制限数 = 10 + (原文文字数 / 100) \qquad (1)$$

このキーワード数制限を使用して重要度の順番にキーワードを採用する。この場合、キーワードの数は文献(記事)の長さに応じて文献(記事)毎にダイナミックに決まる。

4．5　全分野対応型大規模辞書の構築

内容が極めて多種類で大規模なデータベースを対象として、日本語解析とキーワード自動抽出をするためには、あらゆる分野をカバーする極めて大規模な辞書が必要である。このために今回用いた日本語解析用のシステム辞書は約43万語であり、キーワード抽出のためのユーザ辞書は約50万語で、日本語解析用のシステム辞書とこのユーザ辞書をマージして、合計約93万語の辞書を利用してキーワード抽出に利用する。この日本語解析用の辞書でも解析の精度には限界があり、ありとあらゆる分野に対しての精度の高い解析は不可能である。このために、今後は分野ごとの用語を補うための分野別のユーザ辞書を準備して利用する方法が有効である。

5．キーワード自動抽出の評価

5．1　評価の対象データ

第2節で述べたように今回の自動インデクシングの対象は多種多様で、大規模であり、評価においてもそれに十分対応するだけの評価対象データが必要と考えた。評価対象のデータの構成は、7グループ構成として、第1から第5のグループは、おのおののグループが、（一般新聞紙・専門紙・業界紙・地方紙）の組み合わせとした。第6グループは、（社説・コラム）の組み合わせ、第7グループは、（死亡記事・決算記事・新製品記事・その他）の組み合わせとした。評価の総記事数は約450記事である。現在は評価を終了したのは、第1のグループの2新聞紙分である。表3に評価対象データの詳細を示す。

表3　評価対象データ

	平均字数	全記事数
A新聞	900／記事	64記事
B専門紙	610／記事	49記事

なお、今回の評価でA新聞とB専門紙の間では、大きな有意差は見られなかったので、この2紙だけで判断をおこなっても評価は十分だとも考えている。

5．2　キーワードの再現率・適合率による評価

キーワード自動抽出の評価は、キーワード再現率・キーワード適合率を尺度としておこなった。これらの定義を次式に示す。

$$\text{キーワード再現率} = \frac{\text{システムが抽出したキーワードの中に含まれる人手キーワードの総数}}{\text{人手キーワードの総数}} \qquad (2)$$

$$\text{キーワード適合率} = \frac{\text{システムが抽出したキーワードの中に含まれる人手キーワードの総数}}{\text{システムが抽出したキーワードの総数}} \qquad (3)$$

なお、人手キーワードの中で原文中に出現しないキーワードは人手キーワードから除いた。これは、原文中に出現しないキーワードは現状では機械で抽出することが極めて困難だからである。表4に評価結果を示す。

5．3　記事単位での評価

再現率・適合率での評価とは別に、1記事単位で自動インデクシングの評価をおこなった。各記事ごとに自動抽出されたキーワードをみて、それらのキーワードが全体としてその記事の検索のために十分であるかどうかを判断する。すなわ

表４　キーワード自動抽出の評価結果

	キーワード再現率(%)	キーワード適合率(%)
A新聞	７０．０	２９．２
B専門紙	７２．２	３３．３
合計	７０．９	３０．７

ち、記事ごとにキーワード抽出の結果をA、B、Cランクに分類する。Aランクは自動抽出されたキーワードがそのままその記事のキーワードとして使用できる。Bランクは、やや不満は有るが使用できる。Cランクは、そのままでは使用不可、である。これらの判断はインデクシングの専門家がおこなった。この評価結果を表５に示す。今後はCランクの記事を０件にすることを第１目標とする。

表５　記事単位での自動インデクシングの評価

	Aランクの記事数	Bランクの記事数	Cランクの記事数	合計
A新聞	３９	２２	３	６４
B専門紙	２６	２２	１	４９
記事数合計	６５	４４	４	１１３
記事数の構成比率(%)	５８	３９	３	１００

６．自動インデクシングの評価方法について

第５節では、自動インデクシングの評価を、①キーワード再現率・適合率による方法と、②記事単位での評価の方法の２通りでおこなった結果を述べた。

①の方法では、評価のベースを人手キーワードに置いている。人手インデクシングでは、コントロールキーワード付与のためのシソーラスの利用や、主題に着目してキーワードを付与することなどが一般的な指針としてあげられるが、客観的・具体的な手順としては成立していない。人手インデクシングは索引を付ける人の主観に負う部分が大きく、評価方法としては確立したものと言い難い。人手キーワードには無くとも、機械が付けたもので、キーワードとして有った方が良いものも有る。

②の方法は、評価のベースを記事単位につけられたキーワードの集合としている。キーワードの集合がその記事のキーワードとして満足なものかどうかの判断

は、やはり、人がおこなうので、人の主観が入る。この方法も評価方法としては確立していない。ただし、現実の評価方法としては、説得力のあるものと考える。

7．あとがき

内容が多種類で大規模なテキスト型データベースへの自動インデクシング実験システムの構築とその評価結果について述べた。併せて、自動インデクシングの評価方法の問題点についても触れた。自動インデクシング技術が進歩・確立して、信頼性を持って社会で利用されるためには、その評価方法の確立が必須である。この評価方法についての議論と評価用データベースの構築などに向けた努力が望まれている。

参考文献

[1]木本：”日本語新聞記事からのキーワード自動抽出と重要度評価”，信学論文誌，Ｄ－１，Ｖｏｌ．Ｊ７４－Ｄ－１，Ｎｏ．８，ｐｐ．５５６－５６６，（１９９１）．