ヒューマンインタフェース 65－13
音声言語情報処理 10－13
（1996. 3. 1）

ExtractRequest
――利用者への情報開示に基づく検索要求抽出

篠原靖志
sinohara@denken.or.jp
電力中央研究所　情報研究所　情報科学部人工知能グループ
〒201　東京都狛江市岩戸北2－11－1

ExtractRequestは、利用者による適切な検索キーワードの想起を支援するシステムである。本システムは、利用者からの文書の適合性情報に基づいて、文書中に現れる用語の適合性および、用語のデータベース中での使用頻度というごく基本的な情報をを視覚的、かつ、即応的に利用者にフィードバックすることで、利用者が対象データベースでの対象テーマにかかわる用語の利用のされかたを把握することを支援する。本稿では、本システムの基本的考え方、機能、および、利用例について紹介し、その有効性について論じる。

Extract Request
-- a Query Keyword Extraction System based on visual and spontaneous information feedback to users --

Yasushi Shinohara

Communication and Information Research Laboratory
Central Research Institute of Electric Power Industry
2-11-1 Iwado-kita, Komae-shi, Tokyo, 201 Japan

ExtractRequest is a system to support users to recall approriate query keywords to retrieve documents related to a target theme. It feeds back basic information of words visually and spontaneously such as word frequency and word relevance based on the user's input on the relevance of several documents to promote the user to gasp the usage of words related to the theme in the target database. This paper describes the basic ideas of visuality and spotaneousness and functions of ExtratRequest and examples and discusses its utility.

て、その適合度を判定することで、適合した文書をより検索しやすいように、検索式を自動的に修正する方式である。このような自動化は、利用者の負担を軽減する可能性がある反面、利用者にとっては検索過程がブラックボックス化することで、検索された結果の特性を把握しづらいものとなる。特に、検索においては、検索された情報のテーマへの関連の適否の判断は、画一的なものではなく、曖昧性を持ったり、変化をするものである。利用者自身が、このような曖昧さや変化の可能性について知っている。このため、利用者からの適合度判定によって直接、検索式の変更を行うよりも、適合度判定情報を基に利用者が対象テーマについてのデータベースの特性を情報を把握することを支援することが、欲しい情報を確実に速く見つけ出すことに役立つ可能性がある。

このため、本研究では、利用者からの検索結果文書に対する適合度の判定情報を基に、利用者に検索式（キーワード加重方式の場合、キーワードと加重）を変更するために必要な基礎的情報を視覚的に、かつ、即座に提示することで、利用者が直観的に検索に適した用語を把握しくいくことを支援しようとしている。

3．用語分析におけるフィードバック方式

利用者からの対象文書の適合度（以下、用語（群）の適合度と区別するため、文書の関連性の有無、適否と呼ぶ）に基づいて、対象文書中および、データベース中での用語の利用分布を分析することを、用語分析と呼ぶ。用語分析では、検索キーとなりえる用語を抽出した後には、用語の relative term frequency による分析や、用語のクラスタ化など様々な分析が可能である。しかし、ここでは、ごく基本的な情報を視覚的、かつ、即応的にフィードバックすることにより、検索が行い易くなるかをみるために、文書の適否に基づく用語の適合度、および、用語のデータベース中での使用頻度の二つの情報のみを、視覚化して提示することとする。ただし、利用者が文書の適否を変更すると即座に、分析結果が更新されるように即応性を持たせることで、用語分析によって、データベース中の用語の特性を把握する心理的負荷を減らすこととする。

4．システムの機能と構成

ExtractRequest は、データベース検索や用語の分析を行うサーバと、利用者インタフェースであるクライアントとからなるサーバ・クライアント方式をとっている。サーバープログラムは、C で記述し、Sun IPX 上で起動している。また、クライアントは、Machintosh の HyperCard 上に構築してある。

クライアントは、検索画面と、用語分析画面の二つの画面からなる。

まず、検索画面について説明する。前述したように ExtractRequest での全文検索は、キーワード加重による検索方式を採用している。これは、文書中に現れるキーワードの加重和を優先順位として、OR 検索の結果をソートして、その上位一定数（百～数百）を利用者に提示する。キーワードの加重は、利用者が検索画面（図1）下部のキーワード配置用ボックスにキーワードを置く位置により決まる。ボックス部の中央を0として、右に置くほど大きい正数を取り、左に置くほど小さい負数を取る（重み = $sgn(x)*x^2$）。図1では、「日本語」、「言葉」の順に高い正の加重を持ち、表示されている検索結果（図1上段：文書リスト部）は、「日本語 AND 言葉」の結果が最上位に、次ぎに、「日本語」のみ、「言葉」のみの検索結果が続く。利用者は AND/OR/NOT による論理操作を使って検索結果を絞りこんだり、広げたりすることと同様のことを、キーワードの位置を変えることで上位にくる検索結果を変えることで、より容易に行える。

図1上段の文書リストは、見出し部をクリックすることで、文書（記事）全文の表示確認が行なえる。見出し表示以外にも、「用語」ボタ

1．はじめに

近年、様々なところで大量の電子情報の蓄積・出版が行われるようになっており、これらの電子データを効果的に利用することが求められている。このためには、まず、得た電子データを早く、利用可能なデータベースとすることが重要である。このため、人手による意味的タグ付けを必要としない、全文データベース検索が利用されるようになっている。次に、データベースから、目的の文書を速く、確実に見つけ出すことが重要となる。特に、従来のデータベースはサーチャー等の専門技能を有する人が利用したが、オフィスへのネットワーク端末の導入により、対象文書群の内容や用語について不案内な一般利用者が直接、必要な文書を検索する必要性が高まっているため、検索システムをより容易にする必要がある。

われわれは、一般利用者が従来のデータベース検索システムを利用し難い大きな理由として、検索の過程を通して利用者が対象データベースでの用語の利用方法などの情報を効果的に把握・学習して行く過程が支援されていないことがあると考え、利用者への情報の効果的なフィードバックに着目して、特定のテーマに関連する文書を全文検索してくるためのキーワードの選択の支援システム ExtractRequest を試作した。

本稿では、ExtractRequest における情報のフィードバックの基本的考え方、および、システムの機能、利用例について述べ、最後にその有効性について考察する。

2．テーマ検索型検索における問題点と解決のアプローチ

全文検索システムにおいて利用者が特定のテーマをカバーする文書群を広く検索したいテーマ型検索においては、次の二つが大きな問題としてある。

まず、利用者は、自分の検索したい文書群に共通に現われる用語を検索キーとして入力する必要がある。しかし、検索したい文書に現われる用語を事前に適切に想定することは多くの利用者にとっては困難である。あまりに一般的な用語をいれるとテーマに関連しない文書が多数現われるし、あまりに特定的な用語を入力すると、ごく一部の文書群しか収拾できない。また、同一テーマに関する用語でも対象とするデータベースによりその利用が異なり、適切な用語選択を行なうためには、対象データベースでの用語の使用傾向が反映される必要がある。

また、全文検索システムでは、文章中の用語レベルでの検索を行なうので、言い回しの異なる複数の関連の深い用語の OR 操作、AND 操作などを行なう必要がある場合がしばしば生じる。しかし、このような論理型操作は形式的で直観的ではなく、また、記述力も限られているので、使いこなすことが困難である。

この第２の問題については、われわれは、キーワード加重による検索方式を採用することで利用者による直観的な操作を可能にしている。これは、文書中に現れるキーワードの加重和を優先順位として、OR 検索の結果をソートして、その上位一定数（百～数百）を利用者に提示する方式である。キーワードの加重は、利用者が一つの軸のどこにキーワードを置くかにより決まる。軸の中央を０として、右に置くほど大きい正数を取り、左に置くほど小さい負数を取る。例えば、「日本語」、「言葉」の順に高い正の加重を与えると、表示されている検索結果は、「日本語 AND 言葉」の結果が最上位に、次ぎに、「日本語」のみ、「言葉」のみの検索結果が続く。利用者はキーワードの位置を変えることで、上位に来る検索結果を、AND/OR/NOT による論理操作を使って絞りこんだり、広げたりすることと同様のことを、より直観的にかつ容易に行える[堤 94]。

第一の問題については、従来型データベースについては Salton らによる適合性フィードバックシステムなど様々な研究がなされてきた[ELLIS90, Ingwer93]。Salton の適合性フィードバックは、利用者が検索結果の文書に対し

ンを選択することで、検索キーワードを部分文字列に持つ文書中の用語列を表示することもできる。文書リスト部左の「○」「？」「×」は、利用者が判定した文書の関連の有無を示す。

用語分析画面は、検索画面で「用語分析ボタン」をクリックすることで立ち上がる。用語分析画面は、検索画面の下段のキーワード配置用ボックス部が、図2に示すような用語分析結果表示に置き変わったものである。用語分析結果表示部は、文書に出現する用語の分布を可視化している。横軸方向が、用語のデータベース中での頻度であり、上下方向がその用語（群）の選択文書への適合度合である。適合度の高い順に表示している。適合度は、第1列に示され、同一文書群に出現するキーワード群をまとめて、その用語群ごとに表示している。

適合度の第1項、第2項は、その用語が出現する関連文書数－非関連文書数、その用語が出現する関連文書数＋非関連文書数である。第3項は、各関連・非関連文書への出現の有無を示すベクトルである。第一項が大きいほど関連文書に固有に現われる用語であり検索に適したキーワードである。第1項が小さい用語は、非関連文書を除外するのに有効なキーワードである。一方、第2項は大きいほど、上記のキーワードの固有性が安定していることを示す。第2項が小さいキーワードは検索キーワードにはあまり適さない。全体の適合度は、第1項、第2項、第3項の順に優先順位付けしている。なお、第2項が0の場合、すなわち、関連性が不明な文書にしか現われない用語は、関連・非関連文書に現われる用語群とは同一には扱わず、これらの用語群の後ろに表示し、用語分析の対象となっている文書全体での文書出現数、および、文書への出現ベクトルで順位づけしている。このため、利用者は特に文書の関連性を選択しなくても、検索結果の上位の文書で利用されている用語の状況について知ることもできる。

対象文書群からのキーワードとなる用語の抽出は、各文書に出現する用語列を収集し、それらの任意の組み合わせについての長さ2以上の共通部分列を抽出している。これらの抽出結果は、既知の用語列以外も生成するため、その用語のデータベース中での頻度は、2-gramによる索引、および、用語列の索引から、経験式により推定を行なっている。

このような抽出操作はサーバー側で行なうが、文書数100に対しても、Sun IPXで約10秒で済む。一旦、抽出操作を行なった後は、各文書の適否（「○」「？」「×」）を変えても、適合度の再計算とソート操作のみであり、クライアント側で瞬時に行なえる。このため、利用者が、文書の適否を変えると、瞬時に用語分析結果が更新されるので、利用者は文書の選択に神経質にならずにくり返し検討することができる。

また、用語分析と検索の間の移行を容易にするため、用語分析結果表示部の用語をクリックすると、検索画面に移行してその用語が、キーワード配置ボックス上に現れるようにしてある。

なお、即応性については、検索および用語分析とも、時間を要する場合でも10数秒程度、普通は、1秒以内から数秒以内に結果が返ってくる。このような高速化により、利用者は検索におけるキーワードの配置変更や、用語分析における文書の適否判定の変更を頻繁に行うことができる。

5.システム利用例と考察

本節では、ExtractRequestの検索例を紹介しその特性を論じる。検索対象としては、毎日新聞93年版中の東京版朝刊の記事を例題として利用した。

ここでは、利用者は「日本語の言葉の乱れに」ついての記事を知りたいとする。このため、利用者はまず、検索画面で、「日本語」「言葉」を検索キーワードとして、キーワード配置ボックスに配置して関連記事を検索する。この結果が図1である。図1では、外国人の日本語教育の問題など関心のない記事も多数ふくま

れている。このため、利用者はテーマと関連が強いと思われる記事をいくつか選択する。例では、記事のタイトルから４つの記事 93080843、930826077、930610048、93060915 を関連性のある記事とした（関連がないとして除外する記事は特に選択しなかった）。ここで用語分析を行なうと、図２に示すような結果となり、「方言」、「話し言葉」「若者言葉」「（第十九期）国語審議会」が、指定した記事にほぼ共通して現われる用語であることがわかる。４つの用語と同じ記事群に出現する用語としては、それぞれ、「日本語」「ことば」「国語」などがあるが、これらの語はデータベース中では、300～500本の記事で利用されており、一般性が高いことばである。また、「言い方」を除くと、頻度５０以下の用語と 300 以上の用語というように、用語は大きなグループに別れる。このような事例はしばしば見られる（しきい値は事例により異なる）が、一般に、適合度の高い（リストの上位）の用語群中の少頻度グループの用語から用語を組み合わせて選択することが有効である。このような用語がない場合は、テーマへの関連性が低い記事の適否を「？」から「×」として除外して行くことで、少頻度グループの用語が上位に現れてくる。そこでは、「方言」「話し言葉」「若者言葉」を新たなキーワードとして、新たに検索を行なう。この結果が図３である。各キーワードの優先順位は、データベース中の頻度に応じた優先順位づけがしてある。この結果、上位には、関連記事が非常に多くなっている。さらに、このような記事の関連記事を更に選択して用語分析を数回行うことで、関連性の高い記事を収集することができた。　本システムが利用者にフィードバックする情報は、用語群の適合度（及び、文書の適否の変化による用語群の適合度の変化）と各用語のデータベース中での利用頻度という極めて基本的な情報のみである。しかし、限られた数例ではあるが、いくつかの例ついて試行したところ、即応性により利用者は、比較的容易に適切にキーワードの選択を行えることができた。

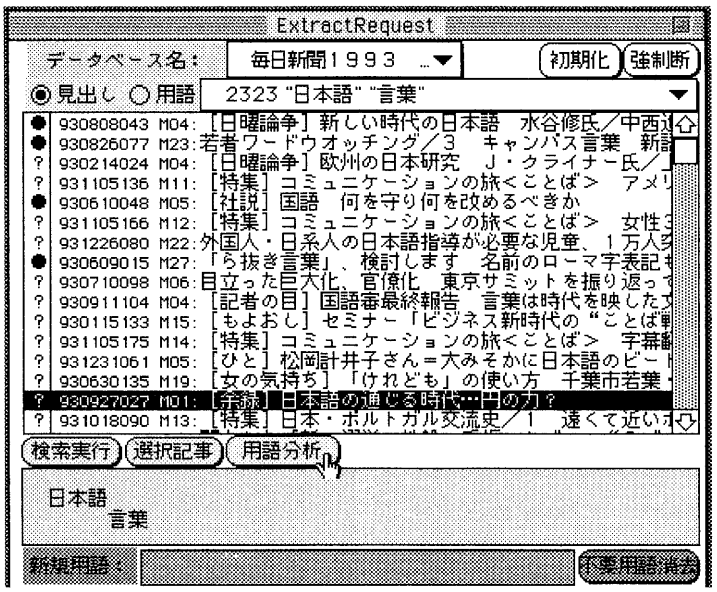

図１．ExtractRequest 検索画面（初期入力時）

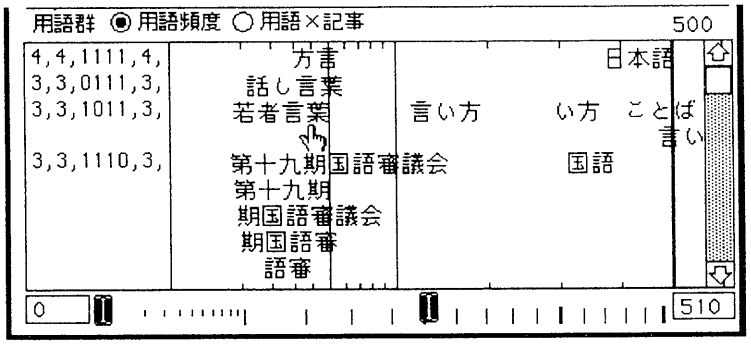

図２．ExtractRequest 用語分析画面（一部）

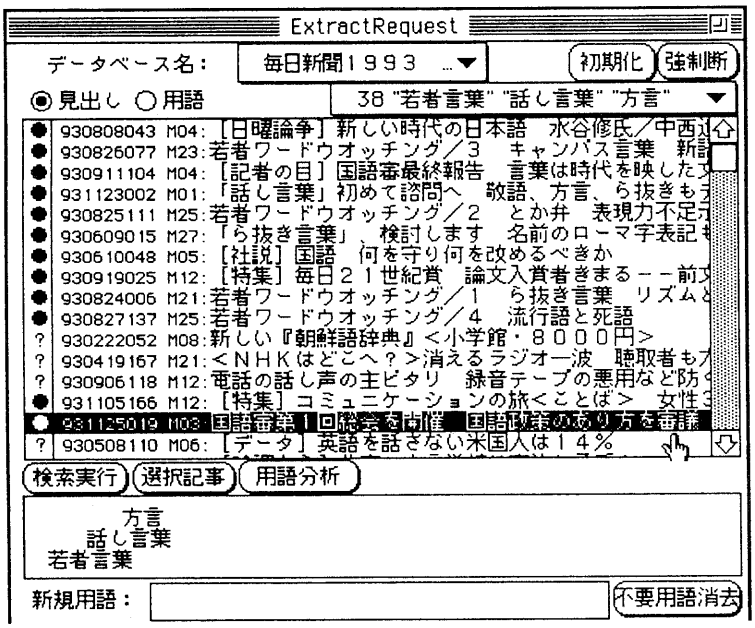

図３．新キーワードによる検索結果

6. 今後の課題

本稿では、特定テーマに関連する文書を集める場合の支援を対象として、対象文書およびデータベースでの用語の分布を視覚的、即応的にフィードバックすることにより、検索に必要な用語の選択が効果的に行える可能性を検討した。本システムの利用効果については、いくつかの事例について試行的検索を行った段階であり、より定量的評価が必要である。これについては現在検討中である。

参考文献

[Ellis90] David Ellis, New Horizons In Inforamtion Retrieval, 1990.

[Ingwer92] Peter Ingwersen, Information Retrieval Interaction, 1993.

[堤 94] 堤富士雄, “キーワードの２次元配置により検索条件をあらわす全文検索システム:Search Space” in proc of Workshop on Interactive Systems and Software, 1995.