情 報 学 基 礎 24－5
（1991. 11. 19）

情報検索理論・技法の問題点とその解決の方向

細 野 公 男
慶應義塾大学文学部 図書館・情報学科

現行の情報検索理論・技法は、主として情報の表面的な処理に基づいたアプローチに依拠しており、その面では、1950年代にこの理論・技法が開発されて以来本質的な進歩はないといえる。本稿では、データ検索と文献検索との違いに関する認識の欠如について述べ、自動索引、情報検索への推論機能・多値的関係の導入、検索方法の使い分け、画像・映像情報の索引語、情報行動、に関する現在の特徴・問題点・解決の方向を概観した。さらに、日本語抄録を対象として索引語を自動抽出する際に、主題文を同定するための手がかりと文中で索引語が出現する位置を同定するための手がかりがもつ特徴を、抽出した。

Information Retrieval Theories and Techniques
－Present Issues and Possible Solutions

Kimio Hosono
School of Library and Information Science
Keio University

Present thoeries and techniques of information retrieval are fraught with fundamental issues.These mainly came from the lack of detailed consideration of the content and user of information in developing them. This paper, after pointing the confusion of subject retrieval with data retrieval, describes characteristics and problems in the several field, such as automatic indexing, inference functions and multivalue aspects in IR, multiplicity of searching mechanisms, index terms for image data, and information seeking behavior. In addition, clues to automatically determine index terms from the abstracts written in Japanese are discussed based on the assumptions that index terms will appear in the sentences which bear subject-related concepts and in the special position of the sentence.

１．はじめに

コンピュータ技術の導入は、情報検索技術・サービスの高度化に大きく貢献したといわれ、実際多量のデータの迅速な検索を可能にした点ではみるべきものは多い。しかし、文献の効果的な検索の面では、これまで開発されてきたコンピュータ処理に基づく手法は、主題（内容）の的確な把握・表現を高度化するものであったとはいいがたい。本来、文献の検索は検索式に合致する結果を得ることだけが問題になるのではなく、情報要求を満たす内容をもつ文献がどれだけ効果的に得られるかが、問題になるからである。情報検索の理論および既存の情報検索システムは、利用者よりも主としてそれぞれの時代で利用できる情報技術を反映して構築されてきた。その結果、いかに迅速に多量のデータを効率的に蓄積・検索できるかなど情報の物理的処理（記号処理）が中心であった。一例がDBMSであるが、この進歩・発展は文献内容の検索の観点から見れば、貢献度はさほど高くないのである。

情報検索に関する基本的な考え方が生まれたのは1950年代であり、事後結合索引方式、プール検索、転置ファイルの概念が確立したのは、この時代である。その後40年経過しているが、実用にたる新たな理論・考え方は出現していないといっても過言ではない。たとえば、現在ＪＩＣＳＴやＤＩＡＬＯＧなど商用のオンライン検索サービスで使用されている各種の近接演算子はＡＮＤ演算子の拡張に過ぎず、1950年代の検索技法とは基本的に差異はない。

情報検索の概念を構成する最も主要な要素は人間である。これは蓄積・検索の対象となる情報を生産するのも利用するのも人間だからである。したがって、本来情報検索理論・システムは、情報作成者および情報要求者の特徴・行動などを十分意識して構築されることが必要なのである。その意味で今日までの情報検索技術の発展や情報検索サービスの普及は、きわめて表面的であるといえよう。

そこで本稿では、情報検索理論・技法にみられる特徴および問題点を概観し、その解決の方向の一例を示すことにする。

２．データ検索と文献検索との違い

各種の統計・数値データや業務用データ、文献の標題、特定の商標・絵・図などを求めるデータ検索・事実検索では、照合操作の結果得られた蓄積情報は、検索質問を完全に満たすものとなる。検索結果は該当する情報が存在するか否かの二通りであり、存在すればそれは適合情報となるからである。これは、検索対象とそれを表現する記号との間にずれが生じないことを示している。したがって、データ検索の場合索引語の選択は比較的容易であり、機械的な方法での自動抽出も可能である。つまりデータ検索ではデータの処理効率が重要で、蓄積・検索の問題はＤＢＭＳなどコンピュータ科学の枠内でとらえればよいのである。情報内容の分析は副次的だからである。

一方、ある特定のテーマや主題について記述されている文献を求める検索では、検索対象とそれを表現する記号との間にずれが生じることが普通である。主題検索で求めるのは、ある概念について言及している情報であり、索引語はその概念を表現する間接的な手段に過ぎないからである。さらに、概念と記号との間では一対一の対応関係は通常存在せず多対多であり、しかもこの対応関係は個人差および分野による差もある。つまり、同じ記号を使用していても、それで意図する概念は必ずしも同じではない。索引語の善し悪しは、文献検索の検索性能に大きな影響を与えるので、その決定には上述のずれをできるだけ少なくするための工夫が不可欠である。また、relevanceやpertinence、換言すれば再現率、精度、新奇性の問題も考えなければならない。

このように、文献検索では内容の処理つまり概念の処理が基本であるにもかかわらず、これまで開発されてきた多くの情報検索理論は、蓄積文献あるいは検索質問の表層的な記号処理に基づいたものであ

った。これは、データあるいは記号（用語）検索との混同が顕著であったことを示すが、この混同は情報検索理論の発展を著しく妨げたのである。

３．自動索引

適切な索引語をいかに抽出・決定するかは、情報検索システムの性能に大きな影響を与える。文献量の著しい増大およびコンピュータ処理技術の進歩と使用コストの低下によって、コンピュータ技術および各種の数量的手法を使用して文献集合を処理・分析する方法の開発が着目されるようになり、たとえば、自動索引の名称で種々の数量的な手法が開発されてきた。これらの研究の出発点になったのは、語の出現頻度の多少に基づいて重要語を抽出する試みを行ったLuhnによる一連の研究であり、この考え方を基本とするさまざまな索引語抽出法のモデルが考案された。その一例が条件付確率、ポアソン分布、ベクトル空間モデルなどであるが、そのほとんどは、蓄積情報中での語の出現頻度に基づくものであった。これは、標題、抄録、本文中の不要語以外の語は部分的にせよ文献の主題を表現しているという仮定が前提となっている。現在の商用情報検索システムで採用されている自由キーワードも、同種のアプローチに基づいているのである。

しかし、機械的に抽出されるこのような語は、索引語の本質、特徴を踏まえたものではなく、索引語として適切である保証は全くない。かつて米国で標題や抄録に語幹 "toxi" をもつ語が出現していないにもかかわらず、索引語として "toxicity" が適切であった文献が複数存在した例が指摘されたことがあるが、それは自由キーワードの不適格さを間接的に物語っている。なお、機械的に抽出した語を索引語とするアプローチは、主題検索とデータ検索の混同例の一つでもある。

文献中での用語の出現は自然現象ではなく言語活動の結果であるので、個々の用語の個別的な出現特性、あるいは複数個の用語の単なる共出現特性をとらえて処理しても、文献内容に適切にアクセスするための索引語の抽出には結びつかないのである。それよりも、文章構造上どの部分が索引語の抽出にあたって適切であるかを重視するアプローチの方が重要である。

４．情報検索機能の高度化

これまでの情報検索（文献検索、データ検索を問わず）は、すでに学術的に既知となっている知識を得る方法に過ぎず、未知の新たな知識を生み出す方法ではない。しかし蓄積情報の単なる検索だけで終わるのでは、情報処理の高度化という観点からは十分とはいえない。提示された質問に対して、蓄積情報に基づいて推論を行い答えを出すシステムとして、質問－応答（question-answering）システムが種々考案された時代があったが、現在この種のシステムは、ほとんど脚光を浴びていない。質問－応答は蓄積されたデータに基づいて推論をおこなうが、このアプローチは文献の主題を対象にしても機能すると考えられる。

"ＡならばＢ" と "ＢならばＣ" をそれぞれ扱った二種類の文献群の間に、付与された索引語集合の類似性（共通性）や書誌的つながり（共引用、書誌結合）がない場合は、これら文献群は表面的には結びつかない。しかし、両者を組み合わせれば、"ＡならばＣ" という論理的関係が導かれる可能性がある。したがって、この論理的関係を抽出するなんらか方法・手段が開発できれば、新たな知識・事実を生み出せることになる。SwansonはＡを魚油、Ｂを血液および循環系に生じる変化（血液粘度、血小板集合性、および血管反応の低下）、Ｃをレーノー病として、"ＡならばＢ" を扱った文献群と、それとは独立に "ＢならばＣ" を扱った文献群とから、魚油がレーノー病に有効であるとの現在認められている結論がえられることの意義を説明している。これは、表面的にはなんら相互関連のない文献間に、

換言すれば蓄積文献集合の中に独立に存在する知識間に、主題的に関連するものが潜在しており、それらを組み合わせることによって新たな知識が生れ得ることを示している。つまり、蓄積文献で扱われた概念間に潜在する論理的（学問的）関係を抽出することができれば、情報検索の効果が高まることが期待できるのである。異なった分野で異なった意図で研究された結果間の関連から、新たな事実・知識を明らかにする可能性があることは、相互に論理的に関連すると思われる文献群を効果的に検索し、結びつける技術の開発が、きわめて有意義であることを示している。

なお、この問題は、従来の適合性が個々の文献ごとに決定されるのに対して、複数の文献を組み合わせた結果を対象とした適合性の概念を考える必要性を暗示している。

５．多値的関係の導入

現在稼働しているシステムでは、蓄積情報あるいは検索質問中の概念とそれを表現する索引語との間の関係には、あるかないかの二値的な関係しか存在しない。したがって、検索論理は論理演算子に基づく完全照合（exact match）方式が基本になる。しかし、索引語で表現される各概念の重要度には差異があるのが普通であり、単にある概念あるいは索引語が含まれているか否かだけでは、蓄積情報や検索質問を十分表現できるとはいいがたい。つまり、重要度やあいまいさの程度を表現するために、概念とそれを表現する索引語との関係を多値的にとらえる手段の開発が、必要である。

多値的関係の導入は、部分照合（partial match）方式の採用を意味する。部分照合方式は、個々の文献が固有に持つ特質だけに基づくものと、文献集合内での文献間の相互関連に基づくものとに分けることができるが、重み(Weight)の導入がその基本である。種々の重み付き検索手法の一例として、検索式中の索引語集合と文献に付与された索引語集合との間の類似性を多値的にとらえる、コサイン関数やファジィ集合理論を使用した検索モデルがある。また、重み付き検索手法は、検索結果の適合度順出力を可能とする。

多値的関係の導入において重要なのは、個々の概念の重みおよび概念間の関係の表現方法である。従来の手法では多くの場合、重みは語の出現頻度で表現されていたに過ぎない。これは重みの定義、特徴などに関する基本的な研究がこれまでほとんどなされていなかったため、語の出現頻度が安直に使用されてきたことを物語っている。今後、蓄積情報、情報要求それぞれについて、重みとは何か、どのように表現されるべきかなど、重みの本質、特徴を明らかにすることが必要である。一方、概念間の関係については形式論理や意味ネットワークの導入例がみられるが、その方法の精緻化だけではなく、関連の種類の規定など解決すべき課題は多い。その意味で、かって種々検討されていたロールの概念を新たな視点から取り上げることは十分意味がある。

６．検索方法の使い分け法の開発

現在の情報検索システムでは、使用されている検索手法が一種類だけであるため、検索質問や蓄積情報の種類が異なっても同じ検索手法が使用されている。しかし、検索手法が検索質問や蓄積情報の種類・規模とは独立に存在すると考える根拠はなく、逆に検索質問や蓄積情報の種類・規模の違いに応じて検索手法を選択したり使い分ける方が妥当であると考えられる。

たとえば、データベースの規模が大きくなると、現在の論理演算子を基盤とする検索技法の機能では的確に対処することができなくなるといわれている。検索結果が多量になりすぎてその適合性を判断することが物理的に困難になるからである。大規模な全文データベースの検索はその一例である。それを回避するためには、検索対象を限定する演算子を多用せざるを得なくなり、結果として重要な文献の検

索もれを生じさせることになる。また、検索式が必然的に複雑にならざるを得ないことも問題である。

これはデータベースの規模に上限がありうること、データベースの規模に応じた検索技法が必要なこと、さらにどのような検索手法が最適であるかをそれぞれの環境下で明らかにする必要があることを示唆している。特に、情報量の増大が依然として顕著であることを考えれば、検索理論がデータベースの規模と独立であるか否かの検討と、もし独立でないならば規模に対応した検索理論の構築は、緊急の課題であろう。しかし、現実には規模の影響について十分考察がなされているとはいいがたい。最適な手法を同定するためには、種々の検索手法間での性能の違い、各手法が最も効果を発揮する状況、蓄積情報の種類・規模と検索結果との関連などの検討が不可欠である。

7．画像・映像を対象とした索引語の検討

ハイパーテキストなどのニューメディアは、画像・映像、音響情報を収録するデータベースの構築・利用を促進する大きな要因となっている。ハイパーテキストは、従来の印刷資料に固有な線形構造から生じる情報アクセスの限界を解消・緩和し、テキストを総合的・多面的に活用できるようにする技術であるので、ハイパーテキストを使用すれば、目次・章・節・段落・図表など情報・データの構成要素を自由に結びつけることができるからである。

しかし、この種のデータベースでは、索引付けの方法に新たな視点の導入が余儀なくされる。たとえば、画像を印象で記憶している場合には検索キーを文字情報で表現することはむずかしく、文字キーワードで検索することは効果的とはいいがたい。また、スライドや写真などの画像データベースでは、文字キーワードとして何が適切であるかを決定することはできない。画像・映像中の人物、背景の建物あるいは自然の一部などが検索キーとなり得るので、その種類には際限がないからである。したがって、画像を検索対象とするデータベースの場合は、文字キーワード以外に画像の構成要素（画像キーワード）を検索キーとするなど、新たな索引方法の開発が必要になろう。

なお、ハイパーテキストでは、種々のリンクを複合的に使用して情報の蓄積・検索をおこなうので、新たな蓄積・検索方法を開発する可能性を秘めている。しかし、そのためにはどのようなリンク付けが情報の利用にとって効果的であるかなどの検討が必要である。

8．情報行動に関する研究

既存の情報検索システムでは、利用者が情報要求を明瞭・簡潔に表現できることを、また、情報要求は検索質問の形で適切に表現されることを前提に構築されている。検索質問に基づいて検索式が作成されて検索が行われるからである。しかし、情報要求は本来あいまいであることが多く、必要とする概念を完全かつ正確に表現できるわけではない。むしろ重要な概念を落してしまうことが多いのが普通である。さらに、情報要求はこれまではその主題的側面だけで考えられていた。しかし、情報要求を適切に理解するためにはそれだけでは不十分で、情報要求を生み出した背景、目的、レベル、情報要求の緊急度、検索対象の範囲、情報探索行動などを把握しなければならない。

したがって、あいまいさをどのように処理するか、また主題的側面だけでなく情報要求を生み出した環境や情報探索行動をどのように理解するかは、適切な検索キーの決定および検索処理の遂行にあたって不可欠なのである。情報検索の分野にもエキスパート・システムの導入が試みられているが、的確に機能するエキスパート・システムを構築するためには、この種の研究を深める必要がある。

9．索引語抽出の高度化に向けて

情報検索理論の今後の発展を図るために最も必要なのは、索引語そのものに関する研究であろう。索引語の付与・抽出の問題は、従来機械的、経験的に扱われてきた。実際の索引作業を規定する索引マニュアルに索引語の本質や特徴に関する詳細な定義・記述が存在していないことは、その一例である。たとえば、文献の主題構成概念には、個々の文献に固有で個人の利用目的などとは独立にとらえられる重要概念（ABOUTNESS）と、その文献を利用する目的、理由、意図などによって、つまり各人あるいは時によって変化する概念（MEANING）とが考えられる。しかし、索引作業の対象となるのはこのどちらなのかまた両方なのかに関しては、既存の索引マニュアルではなんら具体的な指示はなされていないのである。これも主題検索とデータ検索の差異が十分認識されていない結果であると考えられる。

したがって、自動索引をはじめ情報検索理論の研究では、少なくとも以下の点を明らかにすることが不可欠である。

ａ）索引語（キーワード）とは何か

ｂ）索引語の抽出対象を抄録とすると、索引語は抄録中のどの部分から抽出されるか（抄録が一般に前提文、主題文、方法文、結果文から構成され、索引語は主題文から抽出されると仮定することができるか）

これは索引語が前提文、主題文、方法文、結果文のいずれから抽出されるべきかの問題を提起するものであるが、もしｂ）を仮定することができて、かつ主題文の自動抽出と主題文からの索引語の自動抽出が可能であれば、抽出される索引語の品質は格段に高まることになる。

一般に索引マニュアルには抄録に盛り込む内容として、前提説明、目的・主題範囲、方法論、結果・考察・結論があげられており、このうち前提説明が前提文、目的・主題範囲が主題文、方法論が方法文、結果・考察・結論が結果文と考えることができるので、主題文は索引語抽出にあたって重要な部分であることは確かであろう。したがって、主題文を人間が決定してその文がどのような特徴を持っているかを明らかにできれば、逆にその特徴を持つ文を自動的に同定することにより、主題文の自動抽出を試みることが可能になる。

そこで、ＪＩＣＳＴ電気工学ファイルのうちコンピュータ分野の論文抄録を対象に、主題文を抽出するための手がかりの抽出をおこなった。以下に示すのは、その手がかりにみられる特徴の一例である。

・主題文は過去形が多い。

・文形が受け身、否定、可能であるものは主題文でないことが多い。

・文末が形容詞、形容動詞の場合は主題文ではない。

・主題文に固有と思われる動詞群がある。例：述べ、議論、論じ、提案、記述、論述、解説、説明、紹介、研究、開発、示し、報告

・標記、標題、本（論文、文、論、報告、稿など）、について、に関して、ここでは、などの表現がある文は主題文であることが多い。

ｃ）索引語は文中のどのような部分に現れるか

主題文であっても文中のすべての語が索引語になるとはかぎらない。どのような語が索引語となるかを人間が判断しその特徴を明らかにすれば、索引語の自動抽出が可能になる。そのためには索引語と文中の他の語との係わり方（助詞、助詞相当句と動詞との組み合わせ、どの語がどの語を修飾するか、修飾は直接的か間接的かなど）の解析や名詞の種類を手がかりとして、一般的な傾向を明らかにする必要がある。

そこでｂ）の場合と同じ抄録群を対象に手がかりの抽出をおこなった。以下に示すのは、その特徴の一例である。

・否定形（「～については触れていない」の～）、「～が・・・化する」の～、「今まで～が、ここ25年間・・～が開発され・・」の～は、索引語とはならない。
・「～について」の～は索引語となりやすい。
・「最近、現在」などが直接かかる語は、索引語になりやすい。
・AのBという形の名詞句の場合、Bが「事例、効果、計画、現状、評価」などのとき、Aは索引語となりやすい。
・「新しい、最適な」がかかる語は、索引語になりやすい。

d）人間が抄録中の特定の語を索引語と同定するのは、どのような理由からなのか（索引語の抽出理由）その理由には以下が考えられるが、これに関しても調査が必要である。
・全般的理由：主題概念を表していると思うから、標題中にあるから、新奇性があるから、現在脚光を浴びている語であるから、学会名・企業名・システム名であるから、出現頻度が高いから、など。
・文内の各語の役割、語間の関係に基づく理由：文の最初に出てくるから、述べるや説明するなどの動詞の目的語となっているから、強調表現されている語であるから、「～について、～に関して、～を使って、～を用いて、～のために」の～にあたる語であるから、など。

10. おわりに

情報社会の進展は、情報量の増大をさらに促進するだけでなく、我々の知的生活がこれまで以上に情報に依存しなければならないことを意味している。そして、各情報要求者の高度で多様な情報要求に迅速・的確に対処することがこれまで以上に求められている。それにもかかわらず、情報検索の理論・方法は十年一日の感があることは否なめない。したがって、情報の内容および情報要求者の本質・特徴に基礎をおいた情報検索の考え方・理論を構築することが、時代の要請に十分応え得るために不可欠であることは確かなのである。

参考文献

Blair, D. C. Language and Representation in Information Retrieval. Elsevier, 1990. 335p.

Belkin, Nicholas J. and Croft, W. Bruce. Retrieval techniques. Annual Review of Information Science Technology. Chap. 4. Vol. 22. 1987. p.109-145.

Swanson, Don R. Two Medical Literatures that are Logically but not Bibliographically Connected. Journal of the ASIS. 38(4), p. 228-33 (1987)

*：この部分で言及しているのは、慶大文学部図書館・情報学科（細野公男、原田隆史、関根さゆり、梅田栄廣）と日本ＩＢＭ東京基礎研究所（諸橋正幸、野美山浩）との共同研究の一部である。