情　報　検　索

大阪大学基礎工学部教授　木　澤　誠

1. 情報検索とは

情報検索という言葉は英語では information retrieval 略して IR ともいわれる．その定義としては〝目的を設定し，これに対して必要にして十分な情報を，これを必要とする人がいつどこでも必要な時間で入手しうるような方式〟（喜安善市）といわれているが，要するにわれわれが何等かの目的によって情報が欲しいときに，これを適時にかなえるのにはどうすべきかということを対象にした技術である．

情報検索を行うためには，実は対象とする情報，もっと正確にいえば対象となる候補の情報はあらかじめ蓄積されていなければならない．その蓄積されている情報の中から，必要とするものをどのようにして選び出すか，そしてそのためにはどのような蓄積のしかたをしたらよいかというのがこの技術の中心点である．

蓄積ということは，物理的にいえば記録媒体に記録して集めておくこと，電子計算機的な表現をすれば記憶装置または記憶媒体に記憶させておくことである．したがって，ちょっと見ると情報検索というのは単に情報を記憶装置に記憶させておいてそれを呼出すだけのことではないか，そんなことはどの記憶装置でもやっていることだというように誤解されるおそれがある．しかし，それは全く的はずれの見方である．現在実用化されている記憶装置を見ると，そのほとんど全部が記憶とその呼出しに際してアドレスを手がかりにする機構となっており，そこで取扱われる情報の内容に対しては一切関係がない．しかしながら，情報検索においては，取扱う情報の内容が問題であり，アドレスのいかんにかかわらず，所要の内容をもった情報を呼出さねばならないのである．

情報検索は現在では電子計算機の応用の一分野であるということになっている．しかし，この技術はもともと電子計算機とは必ずしも関係なく，実際上の必要から起ったものである．具体的にわれわれの周辺でその必要な場面を探してみると，図書館において多数の蔵書の中から自分の読みたい書物をどのようにして早く入手するかという問題，毎月何十種類何百種類と刊行されている専門雑誌の集積の中から，自分の読むべき論文をどのようにして選び分け入手するかという問題などがある．これらの必要性に対して，それぞれその実務を担当する専門家達がいろいろの技術を開発しており，そのための器具や簡単な機械も現れるようになっていた．しかし，取扱う情報の量が急激に増加するにつれて，これを処理する機械にも高度の性能が要求されるようになった．一方，期を同じうして電子計算機の技術が急速に発達して来た．そしてこれが情報検索に対しても非常に有力な道具であることが判明して，必要性にうまく対応する結果となったわけである．

このような背景を考えると，情報検索は決して現存の電子計算機の応用の一分野に封じ込められているものではないことに留意を要する．電子計算機技術の分野からこの方面を手がけた人達が，機械の都合だけで安易に事を処理しようとする傾向が見られたこともあるが，これは誤った態度であるといえよう．現に情報検索を実施するには電子計算機と呼べない機械装置も使用しなければならない場合があるし，もともと電子計算機は情報検索に都合よくできているとは限らないのである．

2. 情報検索の機械化の目標

情報検索の技術を進歩させようと図るとき，

現在現実には機械化ということを抜きにしては考えられない．機械化ということは，すでに述べたように，直接には取扱わねばならない情報の量が急激に増加し，処理時間，人手などをこれに対応させねばならない必要性から起っている．

したがって，情報検索の機械化の目標の重要項目の一つが，その記憶容量（蓄積できる情報の量）と処理速度とにあることは間違いない．記憶容量はビット数，字数，語数，または情報の件数などのうち都合のよい単位で表現できるし，処理速度は単位時間当り探すためにしらべうる情報件数などでも表現できるが，実効的には情報を要求してから欲しいものを実際に入手するまでの時間すなわちいわゆるターン・アラウンド・タイムが重要なパラメータとなろう。そしてこれらのパラメータにいつもつきまとうのは価格の問題で，実際のシステムは常に投資しうる額と期待する性能と二つの相反する条件の妥協の上に成立している．

ところで，実際に情報検索の機械化システムを使ってみると，その性能の評価の考え方に挙げたものとは別種の大きな問題があることがわかって来た．それは，どういう言葉で表現したら適切であるかよくわからないが，いわば情報検索の品質に関する問題，換言すればこの情報検索システムを使って情報を要求し入手した利用者が，その入手した情報に対してどの程度満足したかという問題である．

たとえば，シリコン・トランジスタの製造法に関する情報を得たい人に対して，このシステムが該当する情報を10件蓄積しているのにもかかわらず，5～6件しか与えることができなかったら，その利用者は知ることができる筈の知識が得られないことになり不満を感ずるであろう．逆にこの利用者に対してシステムが30件も与えたとしても，その内容を見たらゲルマニウム・トランジスタに関するものであったり，トランジスタ・ラジオの製造法に関するものであったりしたら，やはり無駄なものまで読まされたことに対して不満が表明されるであろう．

このような品質または満足度をどのように数量的に表現するかは大変むずかしいことであるが，その第1段階として次のような2値が使いうると思われる．

呼出率（recall factor）

$$= \frac{\text{得られたもので要求に適合するものの件数}}{\text{要求に適合する情報の総件数}}$$

適合率（pertinency factor）

$$= \frac{\text{得られたもので要求に適合するものの件数}}{\text{得られた情報の件数}}$$

これらは，上記の例でもわかるように，いずれも1であることが理想であり，1に近いことが望ましい．また，これらとは共役の関係にある次の2値で考えることもある．

除外率（omission factor）＝1－（呼出率）

$$= \frac{\text{得られたもので要求に適合しないものの件数}}{\text{要求に適合する情報の総件数}}$$

雑音率（noise factor）＝1－（適合率）

$$= \frac{\text{得られたもので要求に適合しないものの件数}}{\text{得られた情報の件数}}$$

これらの2値は当然0であることが理想である．なお，雑音率の分子すなわち得られたもので要求に適合しないものの件数を電気通信工学の用語を援用して雑音と呼んでいる．

呼出率と適合率または除外率と雑音率とはそれぞれ互に独立であると思われるが，実際にはどうも呼出率を高めようとすると適合率は低下しやすいし，適合率を高めようとすると呼出率が低下しがちである．これらのどちらを優先させて考えるかにも実際上の考慮点がある．

なお，上の説明は1回の検索要求に対するものであるが，システムとして見た場合，数多くの検索要求に対して上記の諸量が好ましい値になっていることが望ましいことはいうまでもない．

情報検索の品質については，その表し方自身がまだ研究の対象であるから，上記ですべてを尽くしているわけではない．しかしたとえば上記のような考え方で莫然と表現される品質の向上ということが情報検索の機械化のかかえた新しい目標であり課題である．

3. 実際の手法

このように記すと，それでは呼出率や適合率

が理想的に1という値をとらないのは何故かという疑問がわいて来るであろう．そのことを述べる前に，現在情報検索のために実際に行われている手法を見渡すことにしよう．

情報検索においてはアドレスにかかわらず記憶されている情報の内容によって情報を呼出さねばならないのに，現在実用されている記憶装置のほとんど全部がアドレスを手掛りに情報を呼出す機構になっていることはすでに述べたが，この原理的な相違に対して2種の解決方法がある．その第1は情報の内容とアドレスとを対応させる方法，その第2は対象とするファイルを片端から全部順次に呼出して，それが求める内容をもっているかどうかを判定し，採否を決定する方法である．

前者によるときには情報を蓄積するときに，あらかじめ情報の内容とアドレスとの間に一義的な対応関係を設定し，情報の内容によってアドレスを判定して然るべき個所に格納することになる．分類（classification）といわれる方法はこれに属するもので，情報をその内容によっていくつかのカテゴリ（category）に区分する．カテゴリ相互の間には上位下位の関係などが生じて階層的（hierarchical）な体系を形成することが多い．その階層的な関係はそのまま10進法の数値に対応させることができるので，十進分類の形態をとることもある．図書分類などによく用いられている国際十進分類（UDC）などはその有名な例である．この方法は情報内容が比較的単純なときには探索に費される時間が少くてすむが，装置として呼出時間の小さい大容量の記憶装置を必要とする．また，情報内容のあまり複雑なものは取扱いにくい．

後者においては蓄積されている情報はいずれは全部呼出されてしらべられるわけであるから，どのアドレスに格納しておこうとも自由である．記憶装置も等速呼出しである必要がなく，逐次呼出しのものでよい．しかし，蓄積されている情報を全部しらべるのであるから，機械の高速力にものをいわせて，いわば腕力によって事を処することになる．ここでは情報の1件1件が情報利用者の要求を満足するか否かの判定は，あらかじめ各個に附しておくめじるしによって行われる．このめじるしには勿論前記の分類を用いてもよいが，多くは索引方式をとり，見出し語（keyword），記述語（descriptor）などと呼ばれる単語がその役割を果している．一般に情報1件の内容を表現するのには複数個の見出し語などが用いられ，検索の際には所要の内容を何個かの見出し語などの論理積，論理和，時には否定をも交えた関係によって表現する．

上記の2方法が機械検索による最も基本的な二つの型である．実際には索引方式を前者の方法で行ったり，両者を混合したりいろいろの変化があるが，すべてこれら2方法を基にして考えればよいと思う．

4. 機械化の問題点

さてそこで，分類や索引などの方式を利用して機械検索を行ったとき，呼出率や適合率を低下させる原因は何かと考えてみる．結論からいうと，蓄積する情報の内容を分類または索引という方法で表現するときに，および利用者が要求する情報の内容を同様の方法で表現するときに，完全に適切であったかどうかということにある．多くの場合最も適切な手段がとられるとしても，それが完全ではないために生ずる情報の損失や歪がその原因であるといえる．

これをもっと具体的にいえば，たとえば分類という方法を利用するときに，その分類の体系はあらかじめ設定された既成のものであるから，その既成のものに何とかしてあてはめようとするところに無理が生ずる．多少違っていても類似点が多ければ同じカテゴリにまとめられてしまうかも知れない．従来の体系内にピッタリあてはまるカテゴリが見当らないときには無理が一層大きくなる．その上，これを判定する人の主観も入り込む余地が大きくなって，重点の見方の相違によって，同じものでも別のカテゴリに入れられる可能性を生ずる．ことに最近の技術情報のように1件の論文が多方面の技術に関連しているときには，このあたりの悩みは大きい．これは分類体系自身のもっている大きな問題点である．

索引という方法にしても，たとえば何千語を

も使って表現してある情報をせいぜい十何語で肩代りさせようということであるから，その間に情報の損失があるのはあたりまえのことである．問題は語数，換言すればビット数の相違が本質であるわけではなく，その中に盛られている意味内容が重要であるのであるから，語数やビット数に比例して情報が少くなっているとはいえないが，それにしてもかなり情報の損失があると考えなければならない．その上，見出し語などにどのような語を選定するかということは，その作業を行う人の判断によるから，これも人によって差異を生ずることはやむを得ない．

情報を蓄積するときに生ずるこれらの損失や歪と，利用者の情報要求条件を表現するときに生ずる同様の損失や歪とが時には相重なり合って呼出率や適合率の低下を招く原因となるわけである．

呼出率や適合率は実際には測定が困難である．また情報のもつ意味内容についても，定量的な表示は出来ない．そこで話は定性的になってしまい，ともすればあいまいなものになり勝ちであるが，情報検索システムを考える上においてこのようなことを念頭に置かねばならない．

ここで指摘しておかねばならないのは，上記の範囲の現在普通に考えられている程度の機械検索においては，情報検索の品質を低下させる原因はすべて機械の外，すなわち人間の作業の部分で発生しているということである．したがって，機械は知能的な意味における情報処理にあまり関与していないことになる．つまり，高速性とそれに対する経済性にのみ依存しているのが現在の機械検索の偽らざる姿なのである．

5. 機械検索に課せられた研究対象

こう記すと，情報検索に対してわれわれが今後何を考えていかねばならないか，すなわち研究の対象として何が存在するかについて疑問をいだかれるおそれがある．しかし，この疑問の生ずる原因は，現在実施されている機械検索の姿のみに眼を注ぎ，情報検索が理想としてあるべき姿と，その理想と現実との格差とに意を致さないためであろう．実は，現在適当な方法が見当らないために，やむを得ず人手で処理している知能的な仕事を機械処理に置換えるように努力することが研究の大目標なのである．

たとえば，蓄積すべき情報の原文(一次情報)の内容を簡潔に代表する分類，索引などの情報（二次情報）を作成する作業（前処理）は従来はもっぱら人手によって行われているわけであるが，これは情報検索の品質の均一性の面からも，実行のための労力の確保と所要時間の面からも問題がある．これを放任すれば，人手を節減することを目的の一つとしているはずの機械化が，かえって多くの人手を要求するという一見矛盾した事態を生ずることになりかねない．この対策は，自動分類，および自動索引など，一次情報から分類や索引などを作成するアルゴリズムの発見である．これを行う基本的な原理は，原文の中に使用されている単語の中で頻度の高いものは（あまり意味のない常用語を除けば）その内容を表現するのに重要な役割を担っているという考え方である．勿論この考え方だけを単純に利用しても物事が必ずしもうまく行かないので，これに他のいろいろな着想を附加するわけであるが，まだあまり決定的な方法は発見されていない．

その次には，このような事前処理をしないで，つまり二次情報を作らないで，一次情報すなわち原文を直接に検索しようという考え方もある．この方法はハードウエアの記憶容量と処理速度に依存する度合が非常に高いので，現在のところは実用的に成立しないが，将来はあるいは実行可能かも知れない．この考え方によるときには，事前処理の際に発生する可能性のある情報の損失や歪は避けられるという利点がある．その代りには，情報利用者の要求が原文の表現する意味のパターンの中に含まれているかどうかをどのようにして判断するか，そのアルゴリズムの発見が重要な課題となって来る．ここには，情報利用者の要求の方をどのようにして機械で利用できる表現に変換するかという課題も生じている．

これらの課題考究には，自然言語で表現された情報の処理という問題が入って来る．自然言

語の取扱において最も厄介なのは，その意味のあいまいさにどう対処するかということであるが，総じて言語のもつ意味を取扱うことになると，現在の電子計算機はあまり威力を発揮しないので，これを現実的にどう解決して行くかが考慮の対象である．

情報が言語以外，たとえば図形や音声などで表現されている場合には，現在はこれを前処理によって言語の場合と同様な二次情報に変換しているわけであるが，これを機械化しようとするときには，図形や音声の表現するパターンを機械によっていかにして把握するかという問題が生ずる．これは文字読取装置や音声認識装置などの開発のために同様に研究の対象となっている技術であるが，これに適切な解決法が発見されれば，実用面においてはたとえば指紋や肖像写真の検索の実現に発展する可能性が期待される．

実用的な効果から見ると，検索の結果，選び出した情報の事後処理も問題になる．これは，情報の要求が，「何々に関連した情報が欲しい」といういわゆる文書検索のみでなく，情報利用者から投げかけられた質問に直接答える形の情報を与えるいわゆる事実検索の場合もあるからである．事実検索でも，解答を出す順序は恐らくは先ず関連する情報を蓄積されている情報の中から抜き出して，次に質問の形に沿って解答の形に整えるという段階を踏むことになるであろう．ここにも言語処理技術に依存せねばならない部面がある．

これらの情報検索固有の問題に加えて，その道具立て，すなわち機械装置の，情報検索に都合のよい性能の開発附与という課題がある．その最も重要なものは，記憶内容を手がかりとして呼出すことのできる機構の記憶装置の開発である．これが達成されれば，検索の所要時間は著しく短縮されよう．すでに連想記憶装置（associative memory）というものが諸種研究されているが，この種のものが廉価大容量の記憶装置として実用化されることが要望される．

ここに記した各研究課題は，勿論世界中の諸種の機関において研究が行われているし，関連する論文発表も少いわけではない．しかし現在のところは，それらの研究成果がやや非現実的な範囲を脱却できず，研究としての議論と実務実行の議論との間に大きなレベル差が存在するのが悩みである．これは或意味では頭脳ばかりが先走って実行が遅れをとっている感じがするが，逆に実行できるような研究成果を出すことの急務も感ぜられる．

6. 若干の実例

上にも述べたように，機械化された情報検索システムの現状は，その理想とする姿からはかなりの距離があるが，実用的立場からは，たとえ理想がどうであろうと，システムを利用することによる実効が上がればよいわけである．その意味で，現在実施されている若干の例を挙げて，それらがどのようなねらいどころをもち，理想と現実とをどのように妥協させているかに注目することにしよう．

この方面で先ず引合いに出されるのは，米国の国立医学図書館（National Library of Medicine）で実施されている MEDLARS（Medical Literature Analysis Retrieval System）である．これは1960年に計画を始め，1964年から実施しており，電子計算機利用の情報検索システムとして早いものに属するということとともに，現在その協力態勢を全世界にひろめようとしている典型的な文書検索システムである．使用されている電子計算機は Honeywell 800 で，対象とする情報は医学文献である．各論文を専門の文献解析者が読んで論文1件当り8個くらいの索引語をつけて，磁気テープに蓄積する．学校，研究機関，病院はもとより，一般開業医からの検索要求に無料で応じている．また，このシステムで索引誌 Index Medicus の編集もあわせ行われている．この検索システムは，初歩的で控え目な方法を採用しているが，着実なものといってよいであろう．

米国の MIT で半実験的に行われた TIP（Technical Information Project）は有名な MAC システムの応用の一つで，今日話題となっている時分割遠隔アクセス方式による情報検索システムの最初のものである．これは中央にある大型計算械を，利用者の手許にあるコンソール・

タイプライタから通信線を介して遠隔制御で使用できるので，いつでもどこでも利用したいという利用者の欲求をかなり満足させてくれる．取扱っている情報は物理学の文献で，世界の主要物理雑誌に掲載されている論文の表題（title）誌名巻数ページ（identification）著者（authors）その所属機関と所在地（locations）および引用されている文献の誌名巻数ページが蓄積される．ここに注目されることは，これらの情報を整えるのには知能的な判断を必要とせず，事務的な労力のみですむことである．これによって二次情報作成のために要する人的労力の問題と，情報変換の個人差の問題とをある程度解決している．また，論文の内容を直接に表現するものは表題であるから，表題に用いられているすべての単語が索引語として利用されている形となる．表題は勿論自然言語で記述されているから，自然言語を用いた検索の技術が必要となる．論文内容の表現が表題のみで不十分の点は，著者名や引用文献などである程度補助することができ，そのための手法も考えられている．図1に検索出力例を示すが，このように利用者からの送信は小文字で，電子計算械からの応答は大文字で印字され，利用者が送信を完了してから応答の印字が開始されるまでは数十秒という程度の時間に過ぎない．

```
tip
W 1700.6
TYPE YOUR REQUESTS.

s annals of physics 26
f t symmetric group or author cooper r. k. or author
richard cooper or location tucson, arizona
o p i t a l
go

ANNALS OF PHYSICS
VOLUME 26
J 384 V 026 P 0222
APPLICATION OF THE THEORY OFTHE SYMMETRIC
 GROUP TO THE SEVERAL-NUCLEON PROBLEM
  MAHMOUD HORMOZ
  COOPER RICHARD K.
    TUCSON, ARIZONA
    UNIVERSITY OF ARIZONA
    PHYSICS DEPARTMENT

SEARCH COMPLETED,  22 ARTICLES.
     .99 SECONDS,  22.2 ARTICLES/SEC.
      1 ARTICLES FOUND.
```

小文字は要求者側からの送信を，大文字は計算機よりの受信を示す．第1行は検索を開始するために送る信号，第2および3行は，使用許可の応答で数字は時刻．第4行から第8行までは検索指令で第4行のsは search，第5行のfは find，tは title，第7行は，output print identification, title, author, location のそれぞれ略．第8行のgoは検索指令の終了を示し，gと略してもよい．

図1　TIP の検索例

MITでは現在これをさらに発展させたIntrexというシステムを開発中である．このシステムでは文字ディスプレイ装置やマイクロフィッシュ装置も用いられ，所要の原文のマイクロフィッシュを遠隔コンソールからの制御で自動検索して，手許のディスプレイ装置で読んだり写真記録したりすることができるよう計画されている．

わが国では日本電信電話公社電気通信研究所で所内サービス用の REWDAC システムが，自然言語，記述語および分類の併用によって検索の品質を高めようとしている点が注目される．また，日本貿易振興会（JETRO）の貿易資料センターの JETAC システムは所要時間の短縮のためにハードウェア上の工夫もこらされており，わが国では数少い一般サービス用のシステムである．

7. むすび

情報検索の概略について気のつくままに記した．言い足りないことや不適確なことも多々あろうかと思うが，何等かの参考にしていただければ幸である．