学術情報の流通と情報検索技術

大阪大学基礎工学部　木　澤　　誠

1. 学術情報とは何か

学術情報の収集と検索に関する研究会ということなので，ごく一般的なことを述べて午後のむずかしいお話の序論としたい。先ず学術情報とは何かということであるが，厳密な定義はともかくとして，取敢えずは〃学問を進めて行く上に有用な情報〃のことであると考えたい。こゝで学問というのは，人文化学も自然科学もすべて包含するが，流通と検索の問題に特に関心の深いのが主としていわゆる理科系の学問であるし，私も工学を学んだ者で，その上今日の聴講者も電子計算機を扱う方々で工学に近く，いずれも理科系の学問を中心に話を進めることにする。そうすると，学術情報という呼び方のほかに科学技術情報という呼び方のものもひっくるめて考えた方がよい。学術情報と科学技術情報とは違うということを力説される方もあり，細かい比較をするとたしかに違うらしいのであるが，こゝではあまりその点は詮索しないで，大体同じものとして取扱うことにしたい。このほかに特許情報というのもあり，これは科学技術情報の一部であると考えられるが，純粋な技術的なことのほかに権利的なことも入って来ることがあるので，それだけ取扱は厄介である。

学術情報には文献情報とデータ情報とがある。文献情報は，たとえば学術雑誌に掲載された研究論文のように，いわゆる文献として存在する情報で，自然言語を主体として，時に図や写真などの助けをかりて記述されているものが多い。従来学術情報として取扱われていたものの大部分はこの文献情報である。これに対して，最近は数値的なデータそのもの，たとえば特定の物質の比重であるとか比熱であるとか，真空中を光が進む速度であるとか，このような情報が学問の進歩に重要であることが認識されて来た。そしてデータ情報として学術情報の中で重要な地位を占めつゝある。そのほか，地学などでたとえばどこそこで採取した岩石の試料というようなものをも重要な学術情報として取扱うことがある。試料そのものを学術情報と考えるのは私はあまり賛成したくないが，そのように考えている人もあることは否定できない。

いずれにせよ，これらの学術情報に共通していることは，いずれも公開された情報であるということである。世間には企業秘密などのために公開されない情報も多くあるらしいが，そのような情報はこゝで取扱う対象とはならない。

2. 学術情報流通の意義

次に，学術情報の流通とは何であるかというと，これは学術情報がこれを必要とする人の手に適時に入ること，及びそうできるような態勢を作ることと考えられる。そうすることが学問のために必要な理由は，第1に学問上の競争として最も早く発表した功績すなわちいわゆるプライオリティを得るために他の学者がどのような発表をしたかについて常に注意していなければならないこと，第2に学術上のある知識を得るためには時として実験などのために高額の経費と多大の労力とを要することがあるが，すでに発表されているものについては情報の流通によって直ちにこれが利用でき，無用の重複を排除することが可能であること，第3に学術上の情報は集積して知識となるが，これは人類共有の財産と考えられるので，この貴重な財産を最大限に利用できるように企てるのは当然であること，などである。いずれにせよ学術の進歩に関しては学術情報の流通は不可欠なものということができよう。

3. 学術情報流通のための活動

近年において学術情報の流通を盛にしようとする動きが起ったのは，それなりの動機があった。米国における動機は1975年10月4日のソ連の人工衛生の打上げである。万事が世界一でないと気のすまない米国では科学技術の面においてソ連に遅れをとったことについて反省が行われ，オークリッジ原子力研究所長のアルビン・ワインバーグ（Alvin M. Weinberg）博士を委員長とする委員会の作成した〃科学，政府，情報（Science, Government, Information）〃と題する勧告書が大統領科学諮問委員会（PSAC）から発表された。これが世にワインバーグ・レポートと呼ばれるもので，アメリカの科学界，情報関係者，図書館人などに大きな影響を与えた。その内容は大分けにして科学技術界への勧告と政府諸機関への勧告とから成っており，情報の伝達が研究・開発にとって欠くことのできない部分であることを力説している。その科学技術界への勧告として挙げている項目を読上げてみると，*

(1) 科学技術界は，技術情報の取り扱いが科学にとって価値ある，必須部分であることを認識しなければならない。

(2) すべての筆者は刊行されたものの検索について，いっそうの責任を分担しなければならない。

(3) 情報の取り扱いの技術が，広く教育されねばならない。

(4) 科学技術界は，新しい中継法を探求し，開発しなければならない。

(5) 統一性と互換性が望ましい。

＊ 〃科学と政府と情報 － 米国政府に対するワインバーグ報告 － 〃（NIPDOK）シリーズ1）日本ドクメンテーション協会発行（1966）による。

政府諸機関への勧告については一々読上げないが，重要なことはこの勧告に基づいてアメリカでは実に短期間に情報に関する諸機関が整備されたことである。たとえば情報のありかを知らせてもらうナショナル・レフェラル・センター（NRC）が議会図書館（LC）の下に，どういう研究者がどういう研究をしているかを知らせるサイエンス・インフォーメーション・エクスチェンジ（SIE）がスミスソニアン・インステイテューションの下に，文献そのもののコピーを供給するクリアリングハウス（CFSTI）が NBSの下にといった具合である。私が1965年にこれらの機関を訪問したとき，設立されてから1年半くらいにしかならないのに活発に活動しているのを眼のあたりに見て，その内容よりもむしろ重要なことに思いきって力を入れ，短時間に実現してしまうアメリカのやり方の方に驚き感心したものである。これがわが国であったら先ず予算をとるのにうまくいっても何年かかゝり，それから先に何やかやとあってさっぱり仕事が進まないであろうと思うと学術というものに対する彼我の考え方や社会の機構の差をまざまざと感じさせられたのである。

さてそのわが国であるが，わが国における情報関係の政策的な動きは昭和41年8月31日に科学技術会議議長佐藤栄作から内閣総理大臣佐藤栄作に対して提出された〃科学技術振興の総合的基本方策に関する意見〃に始まるであろう。この文書は5章から成っているが，その第3章は〃科学技術情報活動の強化に関する方策〃と題して1章を情報活動強化のための施策にあてて論じている。そこで内閣総理大臣は科学技術会議議長に対して，昭和44年3月14日諮問第4号として〃科学技術情報の流通に関する基本的方策について〃という諮問を行った。科学技術会議ではその第4部会においてこれにつき検討し，結局昭和44年10月31日に科学技術会議議長佐藤栄作から内閣総理大臣佐藤栄作に対して答申が行われた。この答申の中で特筆すべきことは，科学技術情報の全国的流通システム（NISTと呼ぶ）の構想が示されていることである。その後科学技術庁はこの答申に基いてNISTの実現を計っているが，現実には歩みは遅々たるもののようで，とてもアメリカのように急速にとはいかないようである。しかし，NISTが実現したときに当然問題になるような技術的事項に対しては地道な検討が進められている。

そのほか，世界的規模の学術情報システムの構想として，ユネスコと国際学術連合会議（ICSU）との協力の下に出されたUNISISTというシステムがあるがこゝには省略する。また，数値的データに対しては国際学術連合会議（ICSU）に設けられたCODATAという機関があり，私自身もその活動に関与しているが，これについても一応省略しておくことにしよう。

4. 情報検索技術

さてそこで情報検索とは何であるかという問題に入ろう。情報検索とは喜安善市氏の定義によれば，〃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得るような方式〃ということになっているが，要するに沢山ある情報の中から欲しい情報を早く探し出そうとする技術である。これにも情報を蓄える時点との関係によって遡及検索と情報選択提供との２種類がある。遡及検索というのは沢山の情報をあらかじめ蓄えておいて，その蓄えてあるものの中から所望のものを探すもので，一方情報選択（SDI）の方は情報をあらかじめ蓄えることをせず所望の情報の指定条件（プロフィル）を先に与えて待ちかまえているところに情報を入れて，入れながら判断して所望のところに配布するというやり方をする。両者には実用上の効果において若干の違いがあるが，これらを取扱う技術については根本的な差違はないように思う。

もう一つ別の見方として，文書検索と事実検索という分け方もある。文書検索というのは，検索の結果手に入れる情報が文書になっているもので，論文などのような文献を探すのがこれに相当する。これに対して事実検索の方は質問に対する答となる事実そのものが情報として欲しいわけである。

ともあれ，どの種類の検索を行うにせよ，技術上最もむずかしい点は，検索の対象となっている情報が所望のものであるか否かの判断をどのような方法で行うかということである。これは原理的にいえば情報の意味する内容が所望の条件の意味する内容と一致する，またはこれを包含するという判断をすればよいことになるが，これはその情報の専門分野の専門家が高度の頭脳的行為として始めてできることで，その分野の専門家でない人，いわんや機械にとっては不得手なことである。そこで，これを現実的に解決する方法としては，後刻行われる検索の際に出される要求条件をある程度予想して，蓄積する情報の１件１件に目じるしをつけておくということが行われる。その目じるしのつけ方にも，分類（classification）と索引づけ（indexing）と二つの方法が主として行われている。分類は図書館関係者などが従来から愛用している方法で，国際十進分類（UDC）などが名高く，階層的な体系を成すのが普通である。たゞ，コロン分類などいろいろな工夫はあるにせよ，分類においてはどちらかといえば情報の内容をよく表現しようというよりは書籍をどの書棚に収めるかという考え方の方が強く，たとえば１冊の書籍にはたゞ１個の分類標数を与えるという考え方で支配されているような感じがするのは気にかゝるところである。索引づけは情報の内容をいくつかの単語（キーワード）によって代表させようというもので，電子計算機によって処理するのに都合がよいので，電子計算機利用の情報検索システムではこの方法を採用することが多い。このようにしてつけられた目じるしはもとの情報を代表するもので，二次情報と呼ばれこれに対してもとの情報は一次情報と呼ばれる。検索を行うときには，二次情報をファイルしておいて，先ずこれについて所要の情報を探すことにすれば，目じるしがついているから判断は容易であるし，一次情報を直接取扱うよりも少ない記憶容量のものを対象にすればよいので所要時間その他についても有利であるから，今日現実に行われている検索はほとんどすべてこのような方法に頼っている。

二次情報が分類に基いて作られているときには，情報を検索する要求条件も同様にどの分類に属するかを判断して，これと一致した分類に属する情報を探せばよい。また，二次情報がキーワードの集合として表現されているときには，要求条件もこれを代表する若干個のキーワードで表現することができるが，複数個のキーワード間の論理的関係すなわち論理積，論理和，否定などの関係をもあわせ示すのが普通である。そしてその関係をも含めてキーワードの一致する二次情報を探せばよいことになる。

このような操作は人間が行うことができるのは勿論であるが，取扱う情報の量が多くなると人手では時間と労力がかゝりすぎて実用的でなくなる場合が生ずる。そこで従来はいろいろな形式のカードを用いたり，簡単な道具を併用したりしてその欠点を補ってきた。さらに最新の技術として電子計算機を利用すれば強力な武器となることはいうまでもない。

こういうと，情報検索が技術的にみて甚だ容易で単純なことのように見えるかも知れない。しかし，実はそれは二次情報すなわち内容を表す目じるしが作られてしまってからのことで，その二次情報を作るところに問題と困難とが存在するのである。この二次情報を作成する作業は，その情報の専門内容を理解することのできる高度の知識をもった人の頭脳に頼らざるを得ないので，現在のところは電子計算機の力もこゝには及ばず，電子計算機の高速性を活かして大量の情報を扱おうとすればする程二次情報を作成する専門家が忙しくなるという皮肉なことになっている。かの有名なMEDLARS（アメリカの医学図書館で実施している医学文献検索システム）においても，たしか二十何人かのアナリストと呼ばれる人達がこのような仕事に活躍している筈である。省力化という観点からはこれをも電子計算機で処理することが望ましいことは勿論で，いろいろの研究は行われてはいるが，実用になるような決定的な方法はまだ発見されていない。とにかく，情報検索の機械化について最も研究すべき対象はこのあたりにある。

二次情報を作成することは，情報検索の品質すなわち利用者の満足度の上からも問題を投げかけている。それは，一次情報の代表としてキーワードなどで表されるわずかの量の情報を使うために，どうしても情報の損失や時にはキーワードの選択が不適切であることによる情報の歪を生ずることで，そのために検索の判断に誤差を生じ，その結果検索もれや雑音（不要情報の混入）など利用者の不満の原因を作り易い。これを避けるには，一つの理想論としてはキーワードを多数選んだ極限として一次情報の全文のすべての単語をキーワードと考え，原文を直接に検索するという方法が考えられる。これにはこれでまた問題があるのはやむを得ないが，電子計算機の性能も年とともに向上することでもあり，将来の問題として心に留めかつ研究してよいことであると思う。

5. 電子計算機の利用

さきにも述べたように，電子計算機は情報検索を行う際の強力な武器である。それでは電子計算機を利用することの利点は何であろうか。その第１は何といっても処理の高速性にある。電子計算機は与えられたプログラムの範囲では人間よりも段違いの速度で処理をするから，人手による処理では適時性を失うおそれがある仕事を十分適時性を満足する時期に完了させる可能性が大きい。これは実際上の効果からいえば不可能を可能にしたのと同じになり，これによって用途を拡大させる見込が大きい。第２に従来情報検索に用いられていたカードその他の道具に比較すると，速度の点は別にしても，記憶容量が大きいために，たとえばカードの場合にいかにして孔の数を節約しようかといったような無用な神経を使う必要もなく，したがって情報の記録形式に課せられる拘束を比較的ゆるやかにすることができる。その結果，自然言語で表現された情報のようにやゝ冗長性のあるものでもあまり苦にしないで取扱うことができる。その他利点はいろいろあろうが，この辺で止めておいて，逆に欠点というと，さきにも述べたように，二次情報の作成，または検索操作における情報の適否の判断という頭脳的に最も肝心な部面で電子計算機は目下のところあまりなすところがないことである。これは誤解を生むおそれがあるからあまり強調したくないことであるが，電子計算機を１台おけば情報検索は万事解決するかのように安易に考えられても困るので，あえて述べておく次第である。

6. 自然言語の検索

自然言語で表現された情報を検索の対象として取扱うということは，さきにも述べたように非常に有望な方法であるが，反面問題点も残されている。その主なものは，一つの単語に多くの意味がある場合すなわち多義語と，反対に一つの意味を表すのに複数個の単語がある場合すなわち同義語とがそれぞれ存在する可能性である。もはや残りの時間も僅かになったので，自然言語の問題に深く触れる余裕がないが，問題点の解決法の一つとしてシソーラス（thesaurus）の利用が考えられる。ところがこのシソーラスは作るのが大変で，目下のところ主としては人海戦術に頼っているあり様なのが難点で，これを自動化する方法を発見することも重要な研究課題である。

7. 情報検索技術における問題点

今まで述べたことのほかに，今日の情報検索技術にはなおいくらかの問題点がある。直接電子計算機に関連した事柄では，先ず情報の入力手段として文字読取機械の能力が極めて低いために主として打鍵作業に頼らざるを得ないこと，次に記憶内容によって呼出すことのできる大容量記憶装置が実用化されていないこと，等々である。電子計算機に直接関係しない

一般的な問題としては，情報の質（すなわちたとえば論文の価値）の評価をどのようにして行うか（価値の低い情報が多量に氾濫するのは迷惑である／）とか，記述内容を簡潔にしてかつ意味のとり違えや曖昧さを避けるために記述形式を標準化するとかなどの事項がある。このような地味な事項を着々と詰めて行ってこそ真に実用的な情報検索システムが実現し得るということができよう。