情報検索の現状と動向*

笹森勝之助**

1. はじめに

Information Retrieval ということばは，1950 年に Calvin Mooers によって造られたというから，20 年の歴史をもつわけである．日本で「情報検索」ということばが使われてから，かれこれ 10 年になるであろう．使われ始めたころはなかなか理解してもらえなかったことばであったが，最近ではようやく広い範囲に定着してきたようである．

情報検索については，この 10 年ないし 20 年の間，いろいろの試みや実験や研究がなされてきたのであるが，情報検索は，元来，実用的観点の勝った仕事であるから，体系的な研究方法が取られたとはいえないようである．情報検索と隣り合う分野の機械翻訳が，言語研究という形まですすんできたことと比べて，かなり対照的であるように思われる．

しかし，情報検索が実務面で盛んになると共に，情報検索の原理面や形式化の追求も少しずつ行なわれ，また，個々の手法をなるべく定量的なベースにのせようという動きも見られるようになってきた．筆者は，情報検索のごく一部しか知らないが，幸いにも，最近は，Annual Review of Information Science and Technology などに展望記事が現われ始めたので，これらの展望記事に助けられながら，情報検索の最近 10 年間の現状を，大まかに述べてみたい．

2. 情報検索の定義と分類

Information Storage and Retrieval を略して Information Retrieval という．さらに略して IR という．IS＆R と略されることがある．これらの訳語は，「情報の蓄積と検索」，あるいは単に「情報検索」である．情報検索をどう定義するかについては，みんな苦しんでいるようである．ここでは，次の定義[1]を採用したい．

IR は目的を設定し，これに対して必要にして充分な情報を，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 (Systems) である．

IR は，データ検索・事項検索・文献検索の 3 種に大別されているが，それぞれの間に明確な区別があるわけではなく，まだ定説はない．一応の解釈をごく単純な例によって示す．

1) データ検索 (Data Retrieval)：データに関する IR.
　〔例 1〕 アマゾン川の長さはいくらか
　〔例 2〕 ひかり 13 号に空席があるか
2) 事項検索 (Fact Retrieval)：事項に関する IR. 事柄検索ともいわれる．
　〔例 3〕 世界最長の河川はなにか
　〔例 4〕 化合物Xから化合物Yが生成されるプロセスはなにか
3) 文献検索 (Document Retrieval)：文献に関する IR.
　〔例 5〕 河川の長さについて述べた文献を求む
　〔例 6〕 アインシュタインの論文 (1905 年) を引用した文献のリストが欲しい

データ検索は，どちらかというとファイル探索そのものに近い．データセンターの業務の 1 つが，データ検索である．事項検索もデータ検索に似ているようであるが，システムがもつファイルに貯えられている情報から，〔たとえば意味的推論によって〕新しい情報を作り出すことが，事項検索の特色であって，データ検索や文献検索には見られない．たとえば，あるファイルに次のような情報が貯えられているとする．

(i) トラは中国とインドに棲む
(ii) 動物園にはトラがいる
(iii) 東京に動物園がある

このファイルに対して，「東京にトラがいるか？」という問合せを出したとき，「YES」と回答を出してくれるのが事項検索である．文献検索とデータ検索では，この種の問合せ（新しい情報を作り出すことを要

---

* Present State and Development of Information Retrival, by Katsunosuke SASAMORI (The Japan Information Center of Science and Technology)

** 日本科学技術情報センター

求するような問合せ）に応じられないのが普通である．

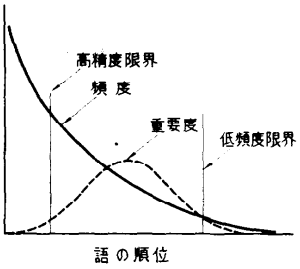
3種の IR の中で，実用化という点でもっとも遅れているのは事項検索であって，実用化された事項検索システムはまだ作られていないといってよいであろう．ただ，データ検索に事項検索的な要素を若干取り入れたシステムがいくつか作られている．たとえば，四国電力の EMERS (Easy Management and Executive Reporting System) は，巨大なデータバンクから簡単な質問手続きによって必要データを取り出し，しかも管理者の検討，判断材料として見やすいように，合計や平均や標準偏差や棒グラフの印刷などを選択することができるシステムである．

もっとも普及しているのは文献検索である．1950年に Calvin Mooers が Information Retrieval ということばを造った．1950年代の中期は，ウェスターンリザーブ大学 (WRU) の Semantic Code や米国金属学会の ASM-SLA 分類表のように，コツコツためてゆくやりかたが盛んであった．しかし，1957～1958年ころ，IBM の Luhn が統計的自動抄録法や KWIC 索引法など，キーワードの統計的処理に基づく IR の手法を相次いで発表した．Luhn のアイデアは，折からコンピュータ技術の急速な発達にも助けられて，その後の IR の発展に著しく貢献した．KWIC から SDI，SDI からオンライン検索へと検索の技術は発展し，今日，これらの技術を活用した文献検索システムが，多数作られている．ここでは，主として文献検索の立場から説明することとする．

3. 内容分析の自動化

文献検索では，文献の意味内容を分析し，分析結果に基づいて文献から重要事項を分析し，抄録や索引や分類をつくることを内容分析（または主題分析）という．内容分析の自動化は，Luhn の KWIC 索引法に始まるといってよい．それまでは，索引をつくるということは，人間がやる仕事で，しかも熟練した専門家がコツコツと根をつめてやらないとだめだというのが常識であった．Luhn のアイデアは，たかだか 1,000 語内外のストップワードをきめておけば，キーワードを自動的に拾うことができるし，また，拾ったキーワードには，いちいち，文脈（たとえば，そのキーワードを含む標題）をつけてアウトプットするから，欲しい情報かどうかはすぐにわかるというものであった．

Luhn のアイデアが出された当時は，コンピュータが一般に普及していなかったせいもあって，コツコツ派の反応はかなり冷やかであった．しかし，今日では，IR のプログラムといえば大てい KWIC もはいっているし，また，KWIC 索引誌といわれるものも，いろいろ市販されている．市販の KWIC 誌に Chemical Titles というものがあるが，これについて，“試験管の発明以来の化学における大事件”という外交辞令がとび出すほどである．もちろん，KWIC 索引法にも，長所のほかにいろいろの短所があるが，内容分析の自動化という道を切り開いた功績は大きい．



第1図 語の重要度と頻度との関係
(Luhn による)

Luhn のアイデアは，せんじつめれば，絶対頻度法ということになる．Luhn[2] によれば，一つの文献中でまんべんなく使われる語（高頻度語）も，滅多に使われない語（低頻度語）も，どちらもその文献にとってあまり重要ではない（**第1図**）．したがって，ある規準値を定めておけば，一つ一つの語をそれぞれの頻度によって重要な語であるかないか，いいかえれば，キーワードであるかストップワードであるかきめることができるというものである（このようにして，ストップワードをいくつもきめておき，このきめられたストップワードのリストを使って KWIC 索引をつくることができる）．

Edmundson[3]は，同じ語でも専門分野が違えば重要度も変わるということを理由として，キーワードとしての語の重要度のきめ方について，次のような改良案を提出した．

$S_1 = f - r$

$S_2 = f/r$

$S_3 = f/(f+r)$

$S_4 = \log f/r$

ここに，f：ある文献内でのある語の使用頻度

r：この文献の専門分野の中での，この語の使用頻度

この場合，Sがあらかじめ定めた規準値以上の語をキーワードとする．

Luhn も Edmundson も，このようなキーワードのきめ方を自動抄録法の実験に応用している．キーワードを上記の方法できめれば，次にこのようなキーワードを多数含む文を重要な文とみなして，適当な数だけ原文から抽出する．抽出した文を原文中の順序どおりに並べれば，自動抄録ができるというわけである．

Luhn や Edmundson のほかにも，いろいろ実験例があり，それぞれ，自動抄録を作ってみせて，これを読めば何とか原文の内容がわかるだろうといっている．しかし，いずれも統計的な手法であって，意味を無視して処理していることや，原文からの文の抽出にすぎない（Extracting であって，Abstracting ではない）ということに，乗り越え難い限界があるようである．その後，文法的手法や意味論的手法，自動抄録の評価法など，いくつかの研究が行なわれたが，現在では，わずかに Edmundson が研究を続けている程度である．

自動抄録法とほぼ平行して，一時，自動索引法や自動分類法の研究が盛んに行なわれた．これらの研究の中でほぼ共通に見られる手法上の特徴は，やはり統計的方法を採用しているということであった．すなわち，キーワードと〔あらかじめ設定されている〕カテゴリーとの相関性を，少数のサンプル文献から求め，この相関性に基づいた適当な計算式を作る．次に，〔サンプル文献以外の〕個々の文献中の個々のキーワードの現われ方から，この計算式を使ってその文献がどのようなカテゴリーに所属するかをきめるという方法である．ただし，相関性の測度やカテゴリーを推定する計算式にいろいろな考え方があって，そこが違うということであった．

自動索引法などが活発に研究されるようになってくると，辞書の自動作成ないし，辞書処理というものが重要視されるようになった．とくにシソーラスということが盛んにいわれるようになった（Luhn も，始めから，シソーラスという概念を提出している）．シソーラスとは，同義関係（Synonymity）や階層関係（Hierarchy）など，キーワード間の意味的関係を定義したキーワード辞書である．シソーラスを手作業で作る試みはいろいろなされており，AIChE，EJC，NASA，通産省，日本科学技術情報センター（JICST）などが，代表的な例である（手作業といっても，語の配列やインデックスの作成など，編集作業の部分は機械化されていることが多い）．また，このようにして作られたシソーラスは，IR システムにおいて，実際に活用されている．シソーラスを自動的に作るということは，“意味”に挑戦するわけで，なかなか難しい問題である．それでも，シソーラスの自動作成についていくつかの研究がみられる．これらは，ほとんどすべて，統計的関連法に基づいている．

いわゆる統計的関連法は，語の共出現性に基づくものである．この点に関し，Giuliano[4] は次のような仮説の形で共出現性の意義を主張している．

① ある文脈中に二語が見出されたとき，この二語が少なくとも近接という意味で実際に関係があるという事実に対する断片的な確率的証拠が得られたとみなしてよい．

② 大量の文脈から断片的な確率的証拠を集積すれば，その文脈群に関して意味があり，かつ有効な“関連というものの総合的な測度”に到達しうる．

まわりくどい表現であるが，要するに，二つの語が，偶然に共出現する場合より以上の頻度で共出現するとき，この二語は互いに関係があるというものである．「航空機」が「パイロット」といつも共出現するとき，「航空機」と「パイロット」とは密接な関係があるというのである．

このような仮説に基づいて，キーワード間の統計的関連を計算する公式を構成することができる．公式に必要な要素は，

C_{lk}　文脈 l とキーワード k との結合強さ
f_i　キーワード i の頻度（出現した文献数）
f_{ij}　キーワード i, j の共出現頻度
d　文献数
A_{ij}　キーワード対 (i, j) の関連係数

である．

なお，文脈とは，標題・抄録・文・節など任意に決定できる操作的な単位である．

通常は，文脈 l にキーワード i が少なくとも1個出現するとき $C_{li}=1$，そうでなければ $C_{li}=0$ の値を与える．C_{li} の値がこのように二値であるとき，次のような行列演算が成立する．

$$F={}^tCC$$

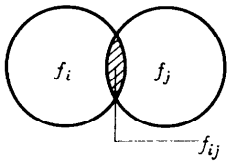
$F=(f_{ij})$　共出現行列
$C=(C_{ij})$　文脈とキーワードとの結合行列
${}^tC$　C の転置行列

これらの要素に基づいて，いろいろの関連係数の定

義が提案されているが，おもなものは次のとおりである．

① $A_{ij}=\dfrac{f_{ij}}{f_i+f_j-f_{ij}}$

重複の割合を示す．



② $A_{ij}=\log_{10}\dfrac{\left(|f_{ij}\times d-f_i\times f_i|-\dfrac{d}{2}\right)^2 d}{f_i\times f_j\times(d-f_i)(d-f_j)}$

Stiles[5] の定義．小サンプルに対するYatesの補正を行なったカイ2乗形式の関連係数．

③ $A_{ij}=\dfrac{f_{ij}\times d-f_i\times f_j}{\sqrt{f_i\times(d-f_i)\times f_j(d-f_j)}}$

Borko[6] が使った定義．

統計的関連法は，もともと自動索引法や自動分類法への適用をめざして発展してきたものであるから，文献間の関連や文献対キーワードの結合の解析に焦点がしぼられていた．

文献間の関連をみる（各文献のもつキーワードパターンを材料とする）

⟶自動分類法

文献とキーワードとの結合をみる

⟶文献にキーワードを割り当てる

⟶自動索引法

これらの試みがさらに発展して，語の関連を堀下げて，シソーラスの問題として扱おうという試みが出てきたのである．たとえば，Giuliano[7] は近接関連（シソーラスでいうRT関係）の測度と同義性関連の測度を定義し，次のようなうまい実例を示している．

Giuliano の測度

近接関連　$t_{ab}=\dfrac{Nf_{ab}}{f_a\times f_b}$

同義性関連　$S_{ab}=N\dfrac{\sum_i f_{ai}f_b/f_i}{f_a\times f_b}$

ここに

N　：語系列中の語の数

f_{ab}　：隣接する語を共出現と定義したときの共出現の頻度

実　例

The U.S. Army launches rocket missiles while the U.S. Navy launches jet missiles, however, although the Navy flies jet plane, strangely it is not the case the Army flies rocket planes.

$C=$

	launches	rocket	missiles	jet	flies	planes
Army	8	0	0	0	8	0
launches	0	8	0	8	0	0
rocket	0	0	8	0	0	8
Navy	8	0	0	0	8	0
jet	0	0	8	0	0	8
flies	0	8	0	8	0	0

$S=$

	Army	launches	rocket	Navy	jet	flies
Army	8	0	0	8	0	0
launches	0	8	0	0	0	8
rocket	0	0	8	0	8	0
Navy	8	0	0	8	0	0
jet	0	0	8	0	8	0
flies	0	8	0	0	0	8

かくて，同義性測度 S で関係づけられる語対は (Army, Navy), (launches, flies), (rocket, jet) であることがわかる．

4. 形式化

情報検索の一つの弱点は，言語学の知識に乏しいという点である．だから，言語の構造について深い知識が得られれば，情報検索〔の自動化〕はもっとうまく行くのではないかと，そちらの方に期待を寄せる傾向が一部にはある．一方，情報検索はどうせドロクサイ仕事だから，実務としてメリットがあがればそれでよく，基礎的な研究など必要ないとの声も強い．この点は，機械翻訳＝言語研究といった感じのする機械翻訳の場合とかなりきわ立った対照を示している．

そうはいいながらも，情報検索における言語の研究もぼつぼつ盛んになってきたようである．たとえば，Hillman[8] はある構文分析法を発表している．これは，文献の各文を構文カテゴリーの作りのよいストリングに変換し，次にこれをセンテンスそのものの中に存在する論理関係を示しながら，センテンシャル・ストリングに分割するといったものである．

ファイルのなかに辞書ばかりでなく，質問を作るための文法規則や意味を解釈するための規則がはいっているとすれば，自然言語で書かれた質問を人工言語の質問に翻訳して，情報を探すことができる．たとえば，Kellogg[9] が開発した CONVERSE というプログラムは，このような考え方に基づいてつくられたものである．このプログラムは，データベースから事柄 (fact) を検索するためのものである．

Burger ら[10]は，会話形式の構文分析を採用している．人間が分析した文をインプットすると，機械はこの分析された文から文法をつくり出す．人間は，オプションにより文の分析を修正したり，機械が作った文法を変更してよい．システムは核文をアウトプットする．

情報検索のプロセスを説明する場合には，ブール代数によるモデルが使われることが多い．しかし，Goffman[11]は，ブール・モデルでは不十分である（したがって，二値命題論理でも不十分である）として，Reichenbach[12]の確率論理に基づくモデルを提唱している．水谷氏[13]も，二値論理による模型と多値論理による模型とを提示し，キーワードによる文献検索の原理を追求している．Jackson[14]は，情報検索システムの基本的なオペレーションはなにかということを調べて，そこからいくつかの関数を導き出している．

5. 情報検索のオンライン化

もう一つの傾向として，情報検索のオンライン化がある．オンライン化することによって情報検索システムが根本的に変わるわけではないが，会話形式が導入されたために，1) エンドユーザが使いやすい，2) "連想" しながらファイルを探すことができる，という点に特徴があろう．

オンライン検索のはしりは，MIT の TIP (Technical Information Program) である．これは，有名な MAC プロジェクトの一環として，開発されたもので，どちらかといえばシンプルなシステムであるが，その後開発された諸システムの原型ともいうべき性格を備えている．そればかりでなく，Citation（引用文献）を使って検索できるという点で，あまり例をみない特徴がある（米国 ISI 社は，引用文献のバッチ検索システムをもっている）．

オンライン検索システムとしては，TIP のほかに，DIALOG (Lockheed), FIRST (DATA Corp.), LEADER (Lehigh University), LISTS (System Development Corp.), Ohio Bar Case and Statute Law (Data Corp.), ORBIT (System Development Corp.), OTIS Support to Acquisition and Cataloging (State of Oregon), RIMS (Northwestern University), SPIRES (Stanford University), Treaty Information Retrieval (University of Washington) がある．日本では，電気通信研究所，日本情報処理開発センター，京都大学，日本科学技術情報センターなどで，実験的に開発されつつある段階である．

これらの中で，ロッキード社の DIALOG システム[15]は，シソーラスをうまく取り入れたシステムとして面白い．このシステムは，IBM 360/30 (32 K)，2311 ディスクパック 2 台，2321 データセル 1 台を使用している．ターミナルから EXPAND キーを押してキーワードをタイプインすると，そのキーワードおよび ABC 順の近辺のキーワードが，仮語番号と〔そのキーワードが付けられている〕文献数と関連語の数と共に，2260 ディスプレイ装置に出てくる．そこからキーワードを選び，EXPAND キーを押して仮語番号をタイプインすると，そのキーワードの関連語が，やはり仮語番号と文献数と関連語数と共にディスプレイされる．このように EXPAND キーを押すことによって，いわばパラパラとシソーラスをめくっているわけである．これはと思うキーワードは，SELECT キーを押してそのキーワードの仮語番号をタイプインしておくと，計算機がおぼえておいてくれる．いいかげんキーワードを見つくろったところで，COMBINE キーを押し，キーワードの仮語番号と演算子とをタイプインすると，このコマンドにより論理式が構成される．計算機は，個々の論理式についても番号を付けてくれるから，論理式と論理式（またはキーワード）とを組み合わせて，新しい論理式を作ることができる．論理式がこれでよしとなれば，DISPLAY キーを押して論理式番号をタイプインし，該当記事を 1 件ディスプレイさせる．ENTER キーを押すと次の該当記事がディスプレイされる．プリントが欲しいときは，PRINT キーを押す．

なお，オンライン検索に関連するものとして，質問応答 (Question-Answering) があるが，これについては，坂井氏の解説[16]にゆずる．

6. 手法の比較

情報検索には，いろいろの手法が提案されたり，使用されたりしている．これらの手法のなかで，どれが最適であるかについては，たえず議論がたたかわされている．そこで，手法間の比較について，いくつかの研究が行なわれてきた．

そのなかで，ASLIB-Cranfield Research Project[17]が，最初の大がかりな研究として有名である．1957 年 7 月，アメリカの National Science Foundation が，イギリスの ASLIB（専門図書館協議会）に対し，索引システムの検索効率の比較研究についての補助金を

認可した．以来6年にわたり Cranfield の College of Aeronautics で，同大学図書館の C. W. Cleverdon を長として実験が遂行された．この実験では，まず航空工学に関する 18,000 件の文献について，(1) UDC，(2) 件名索引，(3) ファセット索引，(4) ユニターム（今日でいうキーワード）の4方法で，3名の索引作成者が索引をつくった．次にこの索引ファイルに質問を向け，探索者が該当文献を探した．検索結果を分析し，recall で測ったところ，4方法は，ほぼ同じ結果を与えるが，ユニタームが最もよい効率を示した．また，recall と relevance は逆関係にあり，一方が効率を増せば，他方が減ることがわかった（recall と relevance については，後述する）．

Cranfield Project は，研究の方法論，とくにデータの分析法が必ずしも十分に確立されていなかったために，折角，大量のデータを扱ったにもかかわらず，スッキリした結果が得られたとはいえなかったようである．しかし，情報検索の分野に“比較研究”の方法を始めて取り入れた．これは，SMART システムに受継がれていった．また，“検索効率”が本格的に論議されるようになったのも，Cranfield Project の影響が大きい．

SMART (Salton's Magical Automatic Retriever of Texts) システムは，完全自動化情報検索システムの研究，検索技術の評価を目的としている．したがって種々の手法を併用しているのが特徴である．はじめ，Harvard 大学で実施されたが，プロジェクトリーダの G. Salton の移動に伴い，1965 年から Cornell 大学に移管された．このシステムで選択的に使える各種の処理（検索）方式をまとめると，次のようになる[18]．

(1) 辞書を使わない
(2) 統計的相関による同義語
(3) シソーラスによる術語のグルーピング
(4) 概念体系参照
(5) 統計的句処理
(6) 構文分析による句の処理（構文分析法としては，Kuno-Öttinger の Multiple-palh Syntactic Analyzer を使用）
(7) 質問と文献の相関
(8) 文献の群別（クラスタリング）
(9) 適合性フィードバック（システムとの相互通信）

いろいろの興味ある結果が得られているが，おもなものは，

文献の長さ：抄録ベースの処理はタイトルベースの処理よりも効率がよい．しかし全文ベースと抄録ベースの比較では差がない．したがって，費用のことを考えると，全文ベースの処理は損である．

シソーラス：シソーラス処理は，文章中の語をそのまま使う場合より有効である．

内容分析の自動化：内容分析の完全自動化法は，人間がキーワードを付ける方法とほとんど効率が等しい．

SMART システムで処理した文献は約 1,000 件である．SMART システムの関係者は，実験条件下で得られた彼らの諸結論は，実用システムの条件下でも有効であると考えているようにみえるが，この点は，さらに実験を続けてたしかめる必要があろう．

7. システムの測度

情報検索システムの測度として重要なのは，コスト，検索効率，利益（効果）の三つであろう．

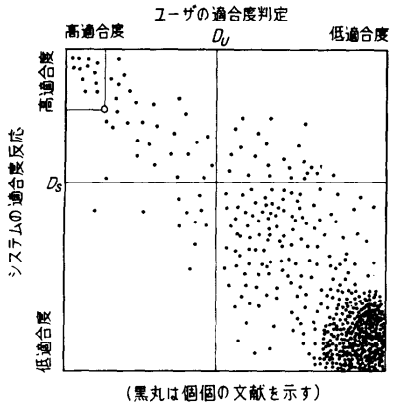
コストについてはいくつかの文献[19]が眼につくが，特定のシステムのコスト分析にとどまることが多く，むしろ，これからの問題だと思われる．

検索効率は，質問に“適合”する文献をどれだけコレクションから引出したか（逆にいえば，検索もれはどれだけあったか），また，引き出した文献群の中に適合文献はどれだけあったか（逆にいえば，ノイズはどれだけあったか）を示す測度である．ところで，質問に対する文献の“適合度”の判定は，ユーザ（すなわち質問者）と検索システムの両者によってなされる．両者の判定は必ずしも一致しない（**第2図**[20]）．システムの適合度判定をユーザの適合度判定になるべく近づけることが，個々の検索システムの目的である．

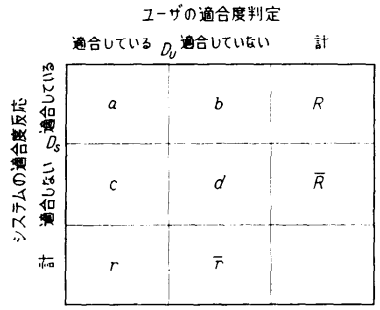
第2図にプロットした文献数をかぞえて，2×2 分割表の形式にすると，**第3図**になる．ここで，記号は次のとおりとする．

r：質問に適合している文献の数〔ユーザの判定〕
$\bar{r}$：適合していない文献の数〔ユーザの判定〕
R：検索された文献の数〔システムの判定〕
$\bar{R}$：検索されなかった文献の数〔システムの判定〕
a：検索された適合文献の数
b：検索された不適合文献の数
c：検索されなかった適合文献の数
d：検索されなかった不適合文献の数

よく使われる測度は，



第2図　ある検索質問に対するユーザの適合度判定とシステムの適合度反応



第3図　適合度判定の 2×2 分割表

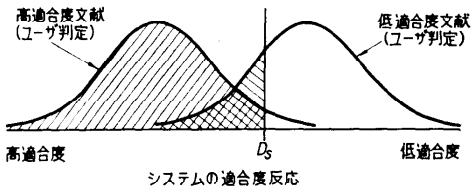
recall ratio (再現率) : a/r

relevance ratio (適合率) : a/R

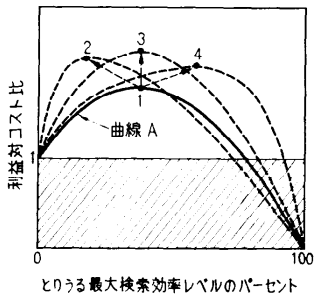
である．$1-a/r$ はもれ率，$1-a/R$ はノイズ率を示す．

Swets[21] は，高適合度文献（D_u の左側）と低適合度文献（D_u の右側）を，それぞれ，システムの適合度反応の目盛りでプロットした（**第4図**），そこで，ある臨界値 D_S を定めると，D_S の左側は，適合文献を検索する確率 $P(R|r)$ と不適合文献を検索する確率 $P(R|\bar{r})$ とを表わすことになると考えた．Swets は，D_S の値をいろいろに取って $P(R|r)$ と $P(R|\bar{r})$ の関係をプロットし，これを統計的決定理論で用いられる検査特性曲線の手法で分析している．

利益の測定によく使われる手法は，情報が提供されたためにその情報を探す時間や費用がどれだけういたかによって，情報の価値を測定しようとするものである．この方法は，情報の価値の一つの因子しか与えてくれない．むずかしいことではあるが，今後，情報の



第4図　システムの適合度反応の目盛でプロットした文献の分布 (Swets による)



第5図　利益，コスト検索効率の関係 (Murdock による)

"効用"という概念を展開すべきであろう．Emery[22] が効用の概念を導入し，Goffman と Newill[23] や Good[24] が効用関数について議論している．

まだよくわからないけれども，コストと検索効率と利益の関係は，第5図[25]のように表わせるのではないだろうか．この図で，斜線の部分は，利益/コストが1を下まわるので，こんな IR システムなんかつくらない方がましだというシステムである．あるシステムが曲線 A で表わされるものとしよう．システムを最適化するということは，利益/コストが最大になるようにすることである．たとえば，点1を点2か点3か点4かに動かすことである．点 2, 3, 4 のどれを選ぶかは，検索効率をどの程度のレベルに置くかによってきまる．

8. IR システム

米国では，企業や政府機関内部での IR システムは，数年前から，活発に活動していた．たとえば，ITIRC (IBM Technical Information Retrieval Eenter) の CIS (Current Information Selection) や NASA の SDI などがその例である．ごく最近になって，IR 用のデータをいれた磁気テープの販売や受託検索サービスなど，情報サービス機関または情報会社なるものの活動が盛んになってきた．MEDLARS, CAS, ISI,

Ringdoc などがそうである.

MEDLARS (Medical Literature Analysis and Retrieval System) というのは，米国国立医学図書館 (NLM=National Library of Medicine) が開発した医学文献の情報検索システムで，1964年1月から実際に使用されている．米国内に数箇所の地域センターがあるほか，英国，スウェーデン，オーストラリアなどにも地域センターがある．日本でも MEDLARS 導入のための実験が行なわれている．現在，約100万件の文献が NLM で蓄積されている．利用件数も，年ごとに増加し，1968年度の実績統計によれば，NLM 本部で3,000件，地域センター全体で5,000件の検索質問を受け付けたという．

CAS (Chemical Abstracts Service) は，60年来，有名な抄録誌 Chemical Abstracts を発行し，全世界の化学者に有用な情報を提供してきた．CAS は，さらにサービスの充実をはかるべく数年前から着々と機械化をすすめている．それとともに化学情報のネットワークをはることを計画しており，欧州については，OECD (欧州経済協力機構) と連繋を保ちながら計画の実現をはかろうとしている．また日本については，日本化学会と日本科学技術情報センターに対してネットワークにはいるよう呼びかけている．CAS は，Chemical Abstracts (1971年には抄録件数が40万件/年に達する予定) のほかに，各種の製品を出版物や磁気テープやマイクロフィルムの形式で提供するとともに，受託検索サービスを行なっている．CAS のファイルに登録される有機化合物 (1971年までに300万種登録される予定) は，グラフ理論を応用した Compound Registry System というシステムによって，きちんと管理されていることが特色である．

ISI (Institute for Scientific Information) というのは，米国の民間情報機関であるが，すでに1960年から情報サービスの機械化に着手している．出版物や磁気テープによる情報の提供，SDI サービスなどを行なっているが，ここの特色は，引用文献 (Citation) がファイルの中にはいっており，ダレソレのイツイツの文献を引用しているすべての文献が知りたいという形の質問にも答えられるし，引用文献による索引誌 (Science Citation Index という商品名) を発行しているということである．

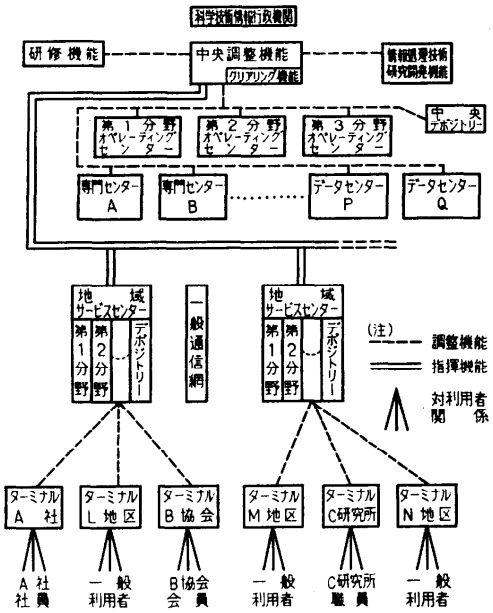
Ringdoc というのは，英国の Derwent Publication Ltd. が実施している薬学関係の情報サービスである．これは会員制度によるシステムで，所定の費用を支払い会員として登録することにより，抄録誌，索引カード，パンチカードまたは磁気テープに収録した薬学情報およびそのほかの資料の提供を受けるものである．現在，全世界で80社以上が加盟しており，そのうち，日本からは 14 社が参加している．

JICST (日本科学技術情報センター) は，1957年に法律に基づいて設立された半官半民の情報サービス機関であるが，1967年12月に FACOM 230-50 を導入し，IR 業務の機械化をはかっている．ここでの特色は，漢字プリンタをコンピュータに連動させて，日本語の文章が自由にプリントアウトできるということであって，日本の国情に適したシステムといえよう．すでに，「科学技術文献速報」という日本語抄録誌を発行しているが，このほか，英語 KWIC 索引誌の発行，SDI や個別検索のサービス，磁気テープの提供などが計画され，現在，そのテストが行なわれている．

このように，情報機関の IR システムが充実し，各種のサービスが行なわれるようになれば，各企業体の社内 IR システムも盛んになってくる．わが国でも，製薬会社や化学会社，電機メーカなど，社内 IR システムの充実に力を入れるところがふえてきた．また，通産省・日本貿易振興会・電電公社電気通信研究所など，政府関係機関でも IR システムが活発に動いている．

IR システムがあちこちで盛んになれば，当然，ネットワークづくりの運動が行なわれる．上に述べた MEDLARS や CAS は，それぞれ，医学と化学の“国際システム”の性格をもっているし，原子力の分野では，国際原子力情報システム (INIS=International Nuclear Information System) が IAEA (国際原子力機関) によってつくられつつある．

わが国でも，このような動きを反映して，科学技術情報の全国的流通システム (NIST=National Information System for Science & Technology) の構想が出された．これは，諮問に対する答申として，科学技術会議から政府へ向けて出された勧告である．これは，**第6図**にみるように，「中央調整機能」のもとに，「オペレーティング・センター」や多くの「ターミナル」を経由してエンドユーザの情報要求に応ずる「地域サービスセンター」や「専門センター」，「データセンター」などを有機的に結合した総合的な流通システムをつくって欲しいというものである．米国では，1957年のスプートニク・ショック後，NSF が科学技術情報活動の強化のためにサポートを惜しまず，さらに，ケネディ大統領時代に，ワインバーグレポート (「情報伝



第6図 科学技術情報の全国的流通システム (NIST) における機能の構造

達における科学技術界ならびに政府の責務」と題する勧告書) が出され，この勧告の具体的内容は着々と実施に移されている．

最近，JIS や ISO で，情報交換に関係する各種の規格が相次いで制定されているが，これらの規格は，IR システムのネットワークづくりに役立つであろう．また，各種のファイルが交換されるようになれば，ファイルフォーマットの互換性も問題になるが，最近の汎用ファイル処理システムの発展をみれば，大して心配することもないであろう．汎用ファイル処理システムについては，竹下氏の記事[26]に詳しく紹介されている．

9. おわりに

こうしてふりかえってみると，情報検索のおもな仕事は，1960 年代のはじめに種をまかれたものが多いようである．このうち，自動分類とか，自動抄録とかいうように，内容分析の自動化に関するもの，別のいい方をすれば意味の問題に挑戦しているものは，まだまだこれからのようである．一方，KWIC とか SDI とか，シソーラスとか (ただし，シソーラスの自動作成を除く) は，現在のコンピュータ技術で十分こなせることがわかり，盛んに実用化されるようになった．これは，あちこちで情報検索システムが作られていることをみればわかることである．これらの情報検索システムは，ほとんど例外なく，大容量ファイルとかデータバンクとかオンライン化とかに関心をもち，システムをその方向にレベルアップしようとしている．ここから，ネットワークの問題，情報交換のための標準化の問題，情報検索の目的や機能を拡大して研究とか経営とかに陽につなぐ問題などがいろいろ出てくるであろう．また基本的な問題として，形式化の問題，とくに言語に関する問題は，情報検索の実用化がすすむにつれて，ますます重要になってくるであろう．

この数年，情報検索は研究面でみるとやや停滞しているようにみえるが，上述のようにいろいろの問題が解決を迫られていることを考えれば，新しい発展を期待してもよいであろう．

参考文献

1) 喜安善市: "情報検索とは何か", 数理科学, [41], 2-7 (1966).
2) Luhn, H. P.: "Keyword-In-Context Index for Technical Literature (KWIC Index)", Amer. Documentation, **11**, 288-295 (1960).
3) Edmundson, H. P. and R. E. Wyllys: "Automatic Abstracting and Indexing-Survey and Recommendation", Comm. ACM, **4**, 226-234 (1961).
4) Giuliano, V. E.: "Analog Networks for Word Association", IEEE Trans. Milit. Elect., **MIL-7**, [283], 221-234 (1963).
5) Stiles, H. E.: "The Association Factor in Information Retrieval", J. ACM, **8**, [2], 271-279 (1961).
6) Borko, H. and M. D. Bernick: "Automatic Document Classification", J. ACM, **10**, 151-162 (1963), Ditto. Part II: Additional Experiments, **11**, [2], 138-151 (1964).
7) Giuliano, V. E.: "The Interpretation of Word Association" Statistical Association Methods for Mechanized Documentation, Symposium Proceedings, Washington, 1964, National Bureau of Standards Miscellaneous Publication 269, U. S. Dept. of Commerce, 25-32 (1965).
8) Hillman, D. J.: "Document Retrieval Theory, Relevance, and the Methodology of Evaluation", Report no. 5: Arithmetigation of Syntactic Analysis, Center for the Information Science, Lehigh Univ., Bethlehem, Pa., 29 July 1967, 19. (PB-176074).
9) Kellogg, C. H.: "Converse—a System for the On-line Description and Retrieval of Structured Data Using Natural Language", System Developm. Corp., Santa Monica, Calif. 26 May

1967, 16 (SP-2635) (AD-654622).
10) Burger, J. F.: "An Interactive System for Computing Dependencies, Phrase Structures and Kernels", System Developm. Corp., Santa Morica, Calif., 29 June 1967, 28 (SP-2454/000/01).
11) Goffman. W.: "On the Logic of Information Retrieval", Inform. Stor. Retr., **2**, 217-220 (1965).
12) Reichenbach, H.: "The Theory of Probability" University of California Press (1949).
13) 水谷静夫: "キーワードによる文献検索の論理", 計量国語学, [46], 14-31 (1968).
14) Jackson, D. M.: "A Note on a Set of Functions for Information Retrieval", Inform. Stor. Retr. **5**, 27-41 (1969).
15) Summit, R. K.: "Dialog: An Operational Online Reference Retrieval System", Proceedings-ACM, National Meeting, 51-56 (1967).
16) 坂井利之, 長尾 真: "最近の言語処理研究について", 情報処理, **10**, [1], 26-34 (1969).
17) Cleverdon, C. W.: "Report on Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems", ASLIB Cranfield Research Project, The College of Aeronautics Cranfield, England, Oct. 1962, 305.
18) "新しい検索技術と検索システムに関する基礎理論の体系化", 日本情報処理開発センター, 昭和44年3月.
19) "情報サービスのコストおよびプライス", 調査資料 No. 1, 日本科学技術情報センター, 1970年1月, 同 No. 2, 1970年3月.
20) King, D. W.: "Design and Evaluation of Information Systems", Annual Review of Information Science and Technology, **3**, 61-103 (1968).
21) Swets, J. A.: "Information Retrieval Systems", Science, **141**, 245-250 (1963).
22) Emery, J.: "Economics of Information", Presented at International Systems Meetings, Detroit, Mich., 1-4 October (1967).
23) Goffman, W. and V. A. Newill: "A Methodology for Test and Evalution of Information Retrieval Systems", Inform. Stor. Retr. **3**, 19-25 (1967).
24) Good, I. J.: "The Decision-Theory Approach to the Evaluation of Information-Retrieva Systems", Inform. Stor. Retr. **3**, 31-34 (1967).
25) Murdock, J. W.: "A General Model of Information on Transfer: Theme Paper 1968 Annual Convention", Amer. Documentation **18**, [4], 197-208 (1969).
26) 竹下 享: "汎用ファイル処理システムの性格とその発展", 情報処理, **10** [2] 84-93 (1969).