言語で表した情報の機械検索をめぐって*

木　沢　　誠**

1. 緒　言

多量の情報を収集して保存しておく場合，最も重要なことは“如何にしたら必要な情報を速かに探し（検索し）て取出し（索出し）得るか”ということである．ある要求に応じて“検索”することに困惑を感ずるような情報は，たとえ収集されていても利用価値は低く，かえって高価な空間を専有するだけである．

この“検索”を行うための通常の手段は，日録を作ること，特定の見出しの順序に配列すること，索引をつけること，分類を行うことなどである[1]．情報の量が比較的少ないか，または1個々々の情報の内容が比較的単純で，直接人力で探し出すことができるときには，これらはたしかに有効な手段である．しかし，現今のように取扱わねばならない情報の量が非常に多くなって来ると，人力によって探し出すということは，時間の点からも，所要経費の点からも実用性に難点を生じてくる．アメリカのある大きい鉄鋼会社の重役は“経費が100,000ドルを超えない範囲では，実験を実行してしまう方が，同じ実験がすでに誰かによって行われたか文献を探すよりは安上りだ”といったとさえ伝えられている．そこでここに検索を高速高能率の機械に行わせねばならない必要性が生じ，近年の電子機械の急速な発達とあいまって，機械による情報検索が注目されてきたわけである．

2. 用語の定義

ここで便宜上若干の用語（私案）を次のように定義しておく．

用　語	定　義
要求者	多量に保存されている情報の中から必要なものを取出すことを要求する人．
質　問	要求者が人間の言語をもって表現した索出の条件．
検索指令	質問を機械操作法を直接表現する形式に変換したもの．
記　事	索出の際に必ず1団として取扱われる情報の単位．たとえば文献を取扱う場合には1論文に関する情報が1記事となる．
適　格 (relevant)	与えられた質問の返答としての条件を満足すること．
不適格 (irrelevant)	与えられた質問の返答としての条件を満足しないこと．
索出記事	検索指令の条件を満足して索出された記事（必ずしも適格であるとは限らない）．

3. 機械による検索の原理

機械で情報の検索をする際の要点は，質問によって要求された条件にかなう情報を如何にして見分けさせるか，または見分けさせるように手配しておくかである．換言すれば，索出記事と適格記事とを完全に一致させるには質問より検索指令への変換を如何にして行うか，またそのためには情報をどんな形で保存しておくかということになる．

この問題に対する最も単純な解答は，人力による検索における分類もしくは索引という概念をそっくりそのまま機械にあてはめること，すなわち予想される問題に対して1対1で対応するようなめじるしを定めて，そのめじるしをあらかじめ記事に附加しておくことである．これを明瞭な形で実行している例の一つは，周囲にあけた孔に作った欠除を棒で選別する方式のカード（Hand-sorted Punched Card, HSPC と略記する）[2],[3],[4]である．

HSPC ではカードに記載されている情報のほかに，カードの周囲に附した孔に予想される質問に対する解答の意味をもたせて，その記事が適格であるような質問を代表する孔に欠除を作る方法を採っている．HSPC 以外の場合には，予想される質問を代表するもの

* On the Machine-Retrieval of Information in Natural Language, by Makoto Kizawa (Electrotechnical Laboratory, Tokyo)

** 電気試験所電子部

が孔という形をとらず，単語または整理番号（標数）で表現させる以外は，原理的に同様の方法を採ることができる．この場合，このような単語または整理番号を“見出語（key word, descriptor)”と呼ぶ．あらかじめ記事の一部で機械が判断しやすい個所にこのような見出語を附しておけば，ある質問に対して機械はその解答に相当する見出語を探せという検索指令を受けて操作を行える．

ここで look-up 方式に言及しておこう．これは同一の見出語の下にこれに対応する記事を集めておいて，見出語ごとに検索する方式であって（これに対して前記のような方式は search 方式と呼ばれている．両方式には縦割と横割との差がある），たとえば Uniterm, Peek-a-boo, Co-ordinate Indexing などという名で呼ばれる諸方式を始めとし[5]，HSPC 以外のカードによるものはほとんどすべてこの方式に属するといってよく，RAMAC, IBM 9900 などの電子機械を利用するものも同様である．筆者は今後の電子機械で取扱うものとしては look-up 方式は採るべきものではないと思う．その理由は，第1にこの方式では原則的に後に述べるような見出語のもつ欠陥をそのままもっていること，第2に同一の情報を幾とおりにも重複して蓄積せねばならないこと，第3に電子機械で取扱うものとしては鑽孔カードは過去のものであり，かつ現状では他の形の廉価で大記憶容量の random access memory が得がたいことである．

Look-up 方法の利点といわれているのは，蓄積されている情報の全部を取扱わなくても検索が行えることにあるが，それとてもカードより1桁上の処理速度をもつ最近の，および将来の電子処理機械には大した問題にはならないことであろう．さらに後に論ずるような言語による検索もしくはもっと高級な検索法は look-up 方式からは望めそうもない．

4. 検索の原理の含む問題点（1）

HSPCによって代表される方法の欠陥は主に二つある．その第1は“質問があらかじめどの程度予想できるか”換言すれば“どのような見出語を用意すべきか”である．元来見出語は情報を最初に記録するときに附さねばならず，そのときにはどんな質問が発せられるかは完全には予想できない．それは“当てものゲーム”か試験の“山をかける”ようなものである．そこで，見出語を附するときに予想されなかった質問が発せられたときには，たとえ適格記事が存在してもこの方法では検索することが不可能である．

質問が予想されなかった原因をさらによく検討すると，用意せられた見出語に対して要求者の発した質問の i) こまかさ（くわしさ）が相違する場合と，ii) 観点が相違する場合とに分けられる，前者の場合は多少の妥協により検索は可能となる．たとえば“エキサダイオード”を求めるのに，その語が見出語に用意されていないときには，これを包含している別の見出語“半導体部品”をもって代用することが考えられる．そのような代用によって要求者の質問に対する忠実度がそこなわれ，検索の有効性が低下することはもちろんで，その忠実度と有効性とを高めるにはできるだけこまかい見出語を用意せねばならない．

後者の場合には問題は深刻である．たとえば雑誌“Electrical Engineering”掲載の各論文に電気工学の観点よりの見出語のみが附せられていたとしても，“日本に関係した論文は？”という質問（“地域”という観点からの質問）に対しては施す術がない，すなわち全く予期しない種類の質問に対しては機械検索そのものが無力化してしまう．

Moore[6] が行ったように，否定の見出語によって確実に不適格のものを落す方式をとれば，この弱点が幾分救済できる場合もあるが，基本的には同じことである．これを防ぐためには，見出語を多方面の観点から用意し，前者の場合とは別の意味でこまかくしておかねばならない．

上記の2理由からかなりこまかい見出語が用意せられたとしてもまだ一つの問題がある．それは，そこに用意された見出語が，過去において発生した現象すなわち既知のことがらに対してはたとえ完全であっても，将来起るかも知れない現象すなわち未知のことがらに対しては有効性が期待できないことである．ことに日進月歩する科学技術や日夜変転する社会のニュースなどを取扱う場合にはその欠陥を遺憾なく暴露するであろう．

たとえば1948年においては“トランジスタ”という見出語を用意することは夢想もできなかったことであろうし，1961年の今日においても，人間宇宙船をはじめとして明日の世界に何が出現するか予想しにくい．その上，見出語を樹枝状に分類して標数を与えようとすると，いずれの分枝に所属させるかに迷う場合を生じ[7]，これをあえて一方に決めると，この人為的な分類とは無関係に発展した部門に著しく不自然な個所を生ずる．このことはたとえば UDC で電子計算機

(681.142) が精密機械器具 (681) として取扱われ，電気通信工学 (621.39) とは別の系統に属していることを見ても思い半ばに過ぎるものがあろう．

5. 検索の原理の含む問題点 (2)

見出語を附ける考え方の欠陥の第2は，そのような見出語は情報を“要約”して選ばねばならないことである．そのためには，自動要約機械が完成しない限り，高度の知識と判断力とをもった学者がすべての情報に目をとおさねばならない．その作業量が大きければ大きいほど，検索を行うために機械を用いようとする趣旨に反することと思われる，もっと重要なことは，要約によって，要約者の個人差や理解不足などを別にしても，原文に含まれている情報の一部は必然的に洩れてしまうことである．そして，このような“情報の損失”は測定ができないし，またしようとする努力もなされていないようである．

6. 言語による全文の機械検索

こう考えてくると，見出語をなるべく詳細に附けた極限として，原情報に用いられている単語の1語1語を見出語とみなしてはどうかということになる．すなわち原情報の全文をそのまま検索の対象とするのである．そうすれば，見出語を選ぶ手数が不要となり，また情報の損失も起らない．そして，原情報に言語が用いてあるならば，如何なる新語が出現しても，見出語の不足に悩むこともあるまい．こうして前節に述べた諸種の欠陥は一掃されてしまうことになろう．その上，言語をそのまま使用しているから，記号化または暗号化されている見出語 (UDC の標数など) を翻訳する手数も不要になる．

原文に用いられている単語を見出語に用いようとする考えは過去においても全然なかったわけではなく，表題のみについて行った例がアメリカの雑誌 Electrical Engineering の Annual Index などに見られる．昨年に至って，この考えによる索引作成操作を IBM 704 計算機に行わせた試作的出版物“Chemical Titles”が American Chemical Society より発行されている[8],[9],[10]．しかしながら，これらは検索そのものは人に行わせるためのものであるから，せいぜいが表題を取扱うに止まり，原情報の全文を取扱うことは考えられていない．

言語によって全文の機械検索を行うには，新たな疑問と問題とが発生する．その疑問や問題は用いられる言語の種類で幾分異るであろうが，ここでは簡単のために話を主として英語に限って問題を考察することにしよう．

7. 言語による検索の有効性

最も重要な疑問は，言語をそのまま用いる方法によって果して有効な検索が行い得るかであろう．この点についての確実な見通しがない限り，その他の議論はほとんど無意味になってしまう．幸にも Swanson[11] はすぐれた計画による直接的な実験によって，この疑問に対し明快な解答を与えた．彼は最近10年間の Physical Review から選んだ核物理学に関する論文100件を対象とし，これに対して50個の質問を用意した．質問はたとえば“How does charge polarization within a nucleus effect the Coulomb scattering of charged particles by that nucleus?”というように要求者の必要によって発する言語そのままの形である．

実験従事者は二組に分れており，第1組 (数人の物理学専攻者) は各論文を直接読んで各質問に対する適格性を調べ，その重要度 (適格度) を10点満点で評定し，論文と質問に対する適格度の完全な表を作成した．第2組 (物理学専攻者3人，計算機プログラマ2人，数学専攻者1人，物理学の知識のある司書1人計7人) は第1組の人々の仕事については全く知らされない状態で，与えられた50個の質問に対して次の3種の方法で1,000回にわたる検索実験を行って，その結果を第1組の人々の評定とつき合わせてみた．

(C) 核物理学に関する“事項見出索引”を作成してこれによって行った普通の検索法．ここでは機械を用いていないが，通常の見出語による機械検索法は原理的にはここに含まれる．

(U) 全文を磁気テープに記録し，単語を指定する検索指令 (たとえば前記質問では“charge polarization か charge distribution という語を含み，同時に scattering, scattered または scatter という語を含むこと”というように) によって電子計算機を用いて行った検索法．

(A) U と同様の機械検索で，類語集のような補助手段を使用した方法．類語集 (thesaurus) というのは同じ概念を表す別の語句を集めたもので，この実験のために核物理学関係のものを編集した．

実験の結果の良否は式 $R-pI$ で数値的に表現し，すべての質問について平均して**第1図**のようにまとめ

た．ここに

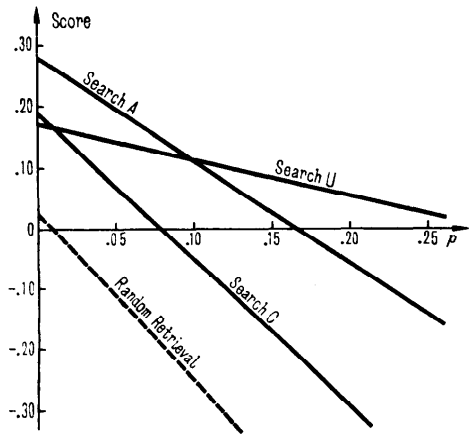
$$R=\frac{\text{索出記事の適格度の和}}{\text{その質問に対する全記事の適格度の和}}$$

I＝不適格記事の実効的量＝$N-LR$

N＝索出記事の総数

L＝適格記事の総数

p＝不適格に対する減点率



第1図　Swanson の実験における各検索法の平均評点（各個人の平均評点の平均）と減点率 p との関係．ただし源泉記事は除外されている．

索出が完全なときにはこの評点は1である．p の値は不適格記事を読んだための時間損失に比例するもので，一般には $0.05\leqq p\leqq 0.15$ くらいが適当であるとしている．この範囲でみると，言語による機械検索（U と A）が通常の見出語による検索法（C）よりすぐれている．その上，この図を作成するときには，各質問に対して直接重大な関係のある記事（適格度10）すなわち“源泉”記事はわざと除外してあるので，これらをも含めると，R の値は C 38%，U 68%，A 86% となって，その優劣の差は一層明かになっている．

8. 言語による全文検索の問題点

Swanson の実験結果をみると，C の方法では勿論，これよりすぐれているUやAの方法でもなお評点が予想に比して低いことに気がつくであろう．その原因は主として質問を検索指令に変換する際の歪にあると考えられる．したがって検索の有効性を向上させるには，この点についての考究がなされなければならない．この観点より言語による検索の問題点を考察することにしよう．

Swanson は原文の記録や機械の操作の上の単純な誤りもかなり多いことを指摘しているが，それらは別として本質的なものを集約すれば

i）単語で表される概念如何

ii）単語間の結合関係によって示される意義如何

iii）言外に表現された意義は如何

などである．もっとも概念は言語のみでは表現できず，図などを併用せねばならないという論[12]もあるが，言語以外の問題にはここでは触れないことにする．

第1の点は言語をそのまま用いることを疑問視し，内容のコード化（主題分析）が最善であると主張する人々の主たる論拠である．なるほど現在用いられている言語は同一の単語によって数種の異った概念を表すものもあれば，同一の概念を数種の異った単語で表現できる場合（この場合は類語集の利用でかなり救済できる）もある．これに対して概念と表現とが1対1で対応することを目標としたコード化は魅力的であり，Western Reserve University の Center for Documentation and Communication Research で行われている Telegraphic Abstract[13] などは現実的にすぐれたものといえよう．

しかしコード化そのものは既述（第4および5節）の諸欠陥を有するし，かりにコード化で概念が規定されるとしても，そもそもコード化というものは換言すれば機械のための言語を新しく創造することである．人間の現に使用している言語が明確な概念を規定していないとして，それと同じ程度の語彙をもち，かつ人間のすぐには理解記憶できないような新言語を作るのであるならば，われわれはその前に人間の言語が概念を明確に表現しなくてもよいものであるか否かの反省が必要であろう．そして機械のためというよりは，われわれ人間自身の生活のために，常用されている言語が概念ともっとよく結びつくための努力が払われて然るべきである．

第2の点については，言語による機械検索を行うにあたって，索引による機械検索の原理をそのまま取入れて，単に語の存在の指定のみによって検索指令を構成させたところに欠陥がある．それは英語をあまりよく話せない日本人が英語の単語を無秩序に数個ならべただけでアメリカ人に意志を通じさせようということに似ている．たとえば cat, dog, bite という3単語を無秩序に挙げたときに“犬が猫をかんだのか”“猫が犬をかんだ”のかはこれからは区別することができない．これらを区別するものは Sentence の中におけ

る単語の位置とそれにともなう文法とである．したがってここに挙げた欠陥は検索指令に単語の存在のみならず文法的問題を導入すると大いに改善される筈である．

Swanson はこれに対し，適当に（数 sentence 以内に）近接して存在している単語は構文上の関係が深いという考え方を示しているが，さらに1歩を進めて，すでに世界の各所で話題とされているような，外国語を翻訳する電子機械に用いられているのと類似の手法を導入するならば，機械の検索性能の質をかなり人間のそれに近づけることができるに違いない．最近 Green 等[14]が発表した“Baseball”と称するプログラムでは，アメリカの職業野球試合に関して日付，球場，チーム名，得点などを記憶させ，これに対して“Who did the Red Sox lose to on July 5?”というような英語のままの（ただし相当用法の制限が課せられているが）質問をそのまま機械に与えると，機械がその言語の意味を判断して検索に相当する操作を進行させて解答を出すという方法をとっているが，その考え方は上記の目標に対する前進に大いに参考となるであろう．

不幸にして現在の情報検索機械[15],[16]と称しているもののなし得ることは単語の存在の指定による検索のみに終っているが，それは古典的なドキュメンテーション（documentation）の思想を近代的な電子機械に適用したときの精一杯の努力の結果にほかならない．そして今やその努力の結果を踏台としてさらに高度の機能の実現にいどむべきときになっているのである．

第3の点については，言語を機械で取扱う際に最も苦慮するところであり，その対策はないといってよい．しかし事を自然科学および技術に関する情報に限るならば，言外に表現するというような文学的手法は極力避けるのが至当であろう．“Silent Sentinels”[17]と題して実は保護継電器に関する純粋に技術的な解説のみを収めた書籍などは例外である．“古池や蛙飛込む水の音”が“静寂”や“幽玄”につながるかどうかはかなり主観的であり，感情を伴なう判断である．そのような客観性を欠いた事柄はもともと機械に取扱わせる問題ではない筈である．

9. 結　言

情報を検索する技術はドキュメンテーションと呼ばれる専門分野において一見発達しているかに見える．しかし現在までの諸手法はそこに用いられている器具や機械によって拘束され，この枠の中で多くの成果を挙げるための努力に終始しているといってよい．ところが最近における高速の電子情報処理機械の出現と，磁気テープを始めとした大容量記憶装置の発達とは，それらの努力を原理的にあまり意味のないものにしつつある．筆者は将来の情報検索法は旧来の伝統にとらわれずに今日の工学的技術水準の上に立って考慮すべきであると思う．そして機械が人間のために働くものである以上，人間の言語の処理の問題に取組ませることは当然であろう．Swanson の実験結果はこの考えに希望を与えるものである．

機械による情報検索の問題点は要するに“如何にして質問の意味するところを忠実に表現する検索指令を作るか”ということにある．そのためには前節に述べ，かつ Swanson が前記稿および別稿[18]でくりかえし指摘しているように，構文上の検討が必要であり，大規模な言語の研究と骨の折れる実験的研究とを行わねばならない．そして，これらの研究の成果にもとづいた計画によって電子機械を駆使すべきである．機械の便宜に合わせて質問自身をまげていた過去の検索法から，人間の利益を主とし人間の都合に合わせた機械検索方式の確立へ——それが機械検索における今後の目標であろう．

参考文献

1) たとえば，B.C. Vickery: Classification and Indexing in Science (book). 2nd Ed. 235＋xix pp. Butterworths Science Publications, London. (1959)
2) 中村重男：ホールソートパンチカードの紹介．化学の領域，8巻，472—478頁．(昭29—7)
3) 平山健三：パンチカード法．化学の領域，8巻，714—725頁．(昭29—11)
4) 平山健三：南江堂式パンチカードの使い方（6回分載）．化学の領域，10巻，74—79頁，150—155頁，245—248頁，450—456頁，562—565頁，638—642頁．(昭31—1/2/3/5/6/7)
5) A.F. Glimn & R.D. Greenway: Information Storage and Retrieval——Dogs, Cats and Indexing. Electrical Engineering, Vol. 79, pp. 724-728, September, (1960)
6) Robert T. Moore: A Screening Method for Large Information Retrieval Systems. Proceedings of the Western Joint Computer Conference, Vol. 19, pp. 259-274. (1961)
7) S. Whelan: Library Retrieval. National Physical Laboratory Symposium No. 10, Mechanization of Thought Processes Vol. II, pp. 935-961 (1959)

8) Computer Uses Program of 'Key Word in Context' To Prepare Periodic Index of Scientific Literature. Electronic Design, Vol. 8, No. 10, pp. 14-15. May 11, (1960)
9) Rapid Indexing of Thousands of Chemical Articles. Computer and Automaticn, Vol. 9, No. 5, p. 18. May, (1960)
10) (日本科学技術情報センター)臨時機械検索室機械班: KWIC (Keyword-In-Context) Index について. 月刊 JICST, 3巻6号, 23—25頁. (昭35—6)
11) Don R. Swanson: Searching Natural Language Text by Computer. Science, Vol. 132, pp. 1099-1104, October 21, (1960)
12) Jacob Rabinow: Presently Available Tools for Information Retrieval. Electrical Engineering, Vol. 77, pp. 494-498. June, (1958)
13) J.W. Perry & Allen Kent: Tools for Machine Literature Searching (book). 972+xviii pp. Interscience Publishers, Inc., New York. (1958)
14) Bert F. Green, Jr., Alice K Wolf, Carol Chomsky & Kenneth Laughery: Baseball: An Automatic Question-Answerer. Proceedings of the Western Joint Computer Conference, Vol. 19, pp. 219-224. (1961)
15) C.D. Gull & P.O. Dodge: The Transistorized Information Searching Selector. International Conference for Standards on a Common Language for Machine Searching and Translation. 33 pp., General Electric, Computer Department, Phoenix, Arizona. September, (1959)
16) 磁気テープ式情報検索機 (文献処理機械)二つ, 電気学会雑誌, 81巻, 712—713頁. (昭36—4)
17) Silent Sentinels (book). Westinghouse Electric Corporation, Meter Division, New Ark, N.J. 236+vi pp. (1949)
18) Don R. Swanson: Information Retrieval: State of the Art. Proceedings of the Western Joint Computer Conference, Vol. 19, pp. 239—246. (1961)