電子計算機で処理しやすい形で標準化が行なわれることが必要である。

3. シソーラスの統一

現在各分野で不統一のまま作成されているシソーラスは統一し、調整する必要がある¹³⁾。

4. 情報検索のソフトウェアの標準化とその保護¹⁴⁾

5・4 学術誌の標準化

学術情報の流通に大きな役割を果たしている学会誌は、ほとんど従来のままの形で編集され、発行されている。

現代のように大量の情報が発生し、さらに増加が予想されるものであるから、実状に合致させるため次のように改善が図られることが重要である。

1. 不必要な情報の排除

一度流通のルートに乗れば、たとえノイズとして除外が必要があってもそのために多大な経費が浪費される¹⁵⁾。したがって、学術的に価値のない情報が否か評価し、不要のものを流通のルートに乗せないことが必要である。そのためは、レフリー制を全学術誌の編集制度に取り入れ、同じ論文を重複し別々の学術誌に取り扱わないなどの配慮がはらわれることが必要である。

2. 学術誌の国際的規格(論文の著者名、タイトル、

所属名、巻号表示、使用言語など)についての標準化を促進し、知的、経済的な無駄をはぶくことが必要である¹⁶⁾。

3. 学術誌の論文(タイトル)の改善。

論文の内容を簡潔に表現するタイトルをつけ、シソーラスのような索引語をサブタイトルとして加えることが検討される必要がある。

4. 発行形態の改善

現在では論文が受理され、発表されるまでに時間がかり過ぎるといわれている。この時間的なロス、さらに二次文献作成の時間が加えられさらに時間的ロスが加算されているのが現状である。この改善のためには論文の受理が決定されたらその“ゲラ”を資料とした専門のレターが発行され、これらの時間的ロスを補うことが必要である。

引用文献

- 1) 科学技術庁計画局：科学技術研究者などの情報利用の実態に関する基礎調査，昭和43年。
- 2) 経団連，情報処理懇談会：情報処理に関するアンケート調査，昭和44年。
- 3) 科学技術庁振興局：情報の発生と入手利用状況に関する調査，昭和44年。
- 4) 科学技術会議：科学技術情報の流通に関する基本的方策，昭和44年。
- 5) 産業構造審議会：情報処理および情報産業の発展のための施策に関する答申，昭和44年。
- 6) 経済審議会情報研究委員会：日本の情報化社会，昭和44年。
- 7) 学術体制研究会：学術研究の背景，昭和28年。
- 8) 高橋達郎：情報検索，昭和43年。
- 9) 山下直：分析情報の自動処理，昭和45年。
- 10) 情報処理学会：電子計算機ユーザー調査，昭和45年。
- 11) 科学技術庁情報処理システム研究会：情報処理システム研究会報告，昭和45年。
- 12) 日本ドクメンテーション協会：ドクメンテーションの標準化，昭和45年。
- 13) 機械振興協会経済研究所：インフォメーション・リトリバルに関する研究(6)，昭和43年。
- 14) 日本ドクメンテーション協会：シソーラス入門，昭和45年。
- 15) “ ”：科学と政府と情報。一米国政府に対するワインバーグ報告一，昭和41年。
- 16) 小谷正雄：bit, Vol. 2, No. 9, 1 (1970)

情報検索活動の考え方*

市川幸郎**

1. ドクメンテーション

情報流通の技術をドクメンテーション(documentation)という。ドクメンテーションは大きく分けて次の三つの機能をもつ。

(1) 情報の収集

(2) 再利用を想定し、どのような場合にさがし出されねば

ならないかを考慮した蓄積(storage)

(3) 要求に応じた検索(retrieval)

このなかで現在の化学工業におけるドクメンテーションで最も重要な部分は(1)であろう。しかし、収集の問題については同じ化学工業でも各分野あるいは各企業によつての差が大きい。ここで与えられた情報検索(information retrieval)の問題は化学工業というワクを離れて存在するものである。

ここで取り扱う情報は、すべて具体的にはドキュメント

* 昭和45年8月27日受理

** Yukio Ichikawa 旭硝子株式会社研究所調査室

図 1

という物の形をとっているものとする。

2. 情報検索の過程

情報検索の任務は、検索の要求に適合する情報をあらかじめ蓄積されたものからさがし出すことである。

情報の作成、蓄積の過程と、質問、検索の過程との関係をフローチャートにしたものが図 1¹⁾ である。

検索の要求のうちで最も普通のものは、ある特定の内容に関するドキュメントをさがし出す種類の質問であって、これが情報検索のうちでも最も困難を伴うものである。この種の要求の検索に応じるためには、個々のドキュメントの内容を表わす指標 (index) が必要であり、検索はこの指標を使って行なわれる。蓄積に当たっては、ドキュメントのもつ幾つかの主題から検索可能にすべき主題を選択 (E-F) し、それを検索用指標におきかえる (F-G) のである。一方、質問の内容を検討し、それを検索すべき概念 (T) に集約して指標におきかえる (U)。そこで検索指令の指標 (U) と、蓄積側の検索用指標 (G) とを、なんらかの方法で照合 (V) して求め

るものをさがし出すのである。

情報検索の効率を表わす測度に普通使われるのは、再現率 (recall) と適合率 (relevance) であり、図 2 によって次のように定義される。

$$\text{再現率 (\%)} = X / (X + Y) \times 100$$

$$\text{適合率 (\%)} = X / (X + Z) \times 100$$

一般にこの二つの間には、再現率を良くしようとする (検索洩れ Y を小さくする) と適合率が悪くなり (ノイズ Z が多くなる)、また逆に適合率を上げようすると再現率が下るというウラハラの関係がある。したがって情報検索システムを設計したり、既成の蓄積を用いて検索作業を行なったりする場合には、高度の再現率を維持しながら、いかに適合率を向上させるかに努力のポイン

図 2

トがある。

3. 情報検索に使用する指標

検索の対象が論文などのような言語情報で、検索の要求が特定の著者の書いた論文をさがし出すといった簡単なものではなく、ある内容をもったドキュメントをさがしてほしいという種類の要求であれば、指標にたよるしかない。指標として普通使われるものに、(1) 自然語と、(2) UDC (Universal Decimal Classification) のような階層分類法とがある。

3・1 自然語による指標

検索に使う蓄積情報には、自分の手で蓄積を行なったものと、既成の蓄積を利用する場合とがある。効率のよい検索を行なうためには、使われる指標のもつ意味や概念を、蓄積する側と検索する側との間で統一しておかねばならない。このことを既成の蓄積を利用して検索する場合に限っていえば、蓄積側（その索引の編集者）がどのような指標の使い方をしているか、すなわち、その蓄積の特徴を知悉した上で、それに見合った検索法をとらねばならない。自然語を指標に使用する場合はとくに注意が肝要である。

既成の蓄積情報として、化学の関係者が検索に最も多く利用する、Chemical Abstracts の Subject Index を例にとると、見出しの指標に使われるコトバには規制が行きとどいていて、*see*, *see also* などの指示で同義語、関連語との関係がはっきりしていて間違いが起きないようにになっている。1907 年の創刊以来、同義語のなかで見出し指標に使われるコトバが一貫して変わっていない点に敬服する。ところが、その一方で、ある見出しの指標のなかでの細目を表わすコトバについてはほとんど規制されていないのである。たとえば、“ある物質の精製法”についての情報を検索したいときに、見出し指標にその物質を見つけるときは、*see* や *see also* によって間違いなく限定された見出しに容易にたどり着くことができるのだが、“精製”に関する文献が、その見出しのなかの“purification of”という小見出しの所に挙げられてあるものだけだと思ったら、たいがい検索洩れを起こすのである。同時に“refining of”に注意すべきであるし、場合によっては“removal……from”を、また物質によっては“distillation of”の項も見なければならぬ。要するに“精製”という概念に対応する指標に使われるコトバの規制が、小見出しにおいてはされていないことに注意して Subject Index を使う必要があるのである。しかし、この索引の場合には見出し指標のコトバが良く規制されているので、その見出し指標のなかで十分に注意することで相当に良い検索ができるのであるか

ら、大して問題にならないのである。

3・2 規制のない自然語による指標

Chemical Abstracts の Subject Index が信頼できる索引として利用される理由の一つが、見出しの指標に使われるコトバが規制されていることであることを述べた。そこで、規制されていない自然語を指標とする情報検索が、どのような結果をもたらすかの例として、平山ら²⁾の実験の概略を紹介する。

化学関係のアメリカ特許を1年分まとめて、英語のコトバを指標として索引できるようにしてある“Uniterm Index to Chemical Patents”という市販の索引がある。この索引の主要部分であるところの、ある特定の内容をもっている特許を検索するための索引 (Chem. Abstr. における Subject Index にあたるもの) は次のようにできている。すなわち、個々の特許明細書から claim の部分を中心に明細書全体から、指標になるコトバを抽出する。そして、あるコトバを共通に指標としてもっている特許番号を、そのコトバを見出し語にしたシートに列記してある。番号の並べ方には、二つ以上のコトバで検索するときに便利なような工夫がしてある。ただし、指標に使うコトバは明細書に使われてあるコトバをそのまま使っていて、同義語、関連語、概念の上位下位などの規制は加えられていない。

平山らはこの索引によって“銀塩を使ったカラー写真法”についての1962年のアメリカ特許を調査した。

たまたまこの質問に適合する特許が1件だけわかったので、この特許から指標として抽出されているコトバをこの索引で調べてみると、silver compounds と halides と color と photography の4個が指標となることがわかった。

そこで、この4個の指標をもっている特許を1962年1年間について検索したところ、はじめからわかっていた1件を含めて3件の特許が見つかった。

この3件を調べてみると、質問に適合しているのは前記の1件だけで、残りの2件はノイズであった。適合率33%だから、まあまあの結果である。

しかし、この質問に適合するアメリカ特許が年間わずかに1件のみとは常識では考えられない。そこで、その年の化学関係アメリカ特許の全数13521件を調べたところ、適合するものが37件であることがわかった。すなわち再現率はわずか3%で、検索洩れが97%であったことになる。

このような検索結果となった原因を知るために、適合する37件の特許について指標を調べてみたところ、検索洩れの主原因が“銀塩”という概念を表わすコトバの指標の混乱、無規制な点にあることがわかった。すなわち、“銀塩”を表わすための指標としては、silver

compounds と halide との組合せが 3 件ある他に, silver halide が 34 件, silver chloride が 2 件, silver bromide が 2 件, silver bromide が 4 件, なにもついていないもの 3 件という状況であった。

結果論だが, 最初から指標に silver halide と color と photography とを選定して検索すれば, 41 件が見つかって, そのうち 34 件が適合 (適合率 83%), 検索洩れ 3 件 (再現率 92%) という検索結果が得られたはずであった。

しかし, 好結果が得られる指標の組合せがあるとしても, どの組合せが良いか事前には全然わからない。しかも, 悪い組合せで検索しても, その結果が不満足なものであることは, この例のように蓄積ファイルの全数検査をしてみないとわからないのである。

4. 指標とするための自然語の規制 (シソーラス)

以上で述べてきたような, 自然語を指標として使用することに付随する問題の解決法として, コトバの指標を最も盛んに使用しているアメリカでは, 2 つの実用的な方法を工夫してきている。第 1 は, アメリカの図書館技術の伝統である件名標目 (subject heading) の考え方で, 電算機技術者の考え方からとりつきやすいデスクリプタ (descriptor) 方式であり, 第 2 は電算機の得意な編集能力と高速性を利用して, 作表機としての機能を発揮させ, 検索の最終部分を人間の頭脳に依存する KWIC (keyword in context) インデックス方式である。

4.1 デスクリプタとシソーラス

アメリカの図書館技術においては, 図書目録に階層分類法を用いないで, 件名標目を使う習慣があった。ヨーロッパの図書館が UDC という階層分類法を図書目録にも使用するのとは対照的である。

Armed Services Technical Information Agency (ASTIA) は爆発的に増加する情報の処理を機械化するに当たって, それまで図書館で使っていた ASTIA Subject Heading をもとにして, レポートの内容を表わす指標として使うコトバを選び, これをデスクリプタと名付けてアルファベット順に並べた辞書を作り, Thesaurus of ASTIA Descriptors として出版した。1960 年のことである。

一方, 民間でもそれ以前からこれと似た考え方で, du Pont 社が自社内のレポートの整理をするために作った辞書をもとにして, American Institute of Chemical Engineers (AIChE) が化学工学用のシソーラスを作り, Chemical Engineering Thesaurus として 1961 年に発行した^{3,4)}。

ある意味 (概念) を表わすデスクリプタに, 特定のコトバを使うことを規制し, その指標に選んだコトバと他

のコトバとの関係を明らかにするものがシソーラスである。

その後シソーラスの考えはアメリカで普及し, 各分野, あるいは各団体ごとのシソーラスが発表された。しかし, それら相互にはつながりがなく作られているので, 広い範囲の情報をとり扱うには困難であった。

そこで, Defence Documentation Center (DDC, 前記の ASTIA が改組されたもの) では, Engineers Joint Council (EJC, 工学関係の学協会の連合会) と協力して, 1964 年に EJC Thesaurus of Engineering Terms を完成した。このシソーラスは広範な分野をカバーしているので好評を博し, 1969 年に Thesaurus of Engineering and Scientific Terms として改訂版⁵⁾ が出された。これが現在のところでは科学技術用語のシソーラスとして最も利用されているものであるが, これをそのままの形で使うにはやはり専門分野の部分のデスクリプタが不足するので, 部分的に補足して使われているのが実状である。シソーラスの利用についての詳細は解説書⁶⁾を参照されたい。

EJC シソーラスでは, ひとつのデスクリプタと他のデスクリプタとの関係を 4 通りの符号で示してある。

(1) UF (=used for) ……

…を使わずに (見出しのデスクリプタ) を使え。
…はデスクリプタにしない。

(2) BT (=broader term) ……

(見出しのデスクリプタ) の上位概念を表わすデスクリプタに…がある。

(3) NT (=narrower term) ……

(見出しのデスクリプタ) の下位概念を表わすデスクリプタに…がある。

(4) RT (=related term) ……

(見出しのデスクリプタと関連するデスクリプタ…がある)。

たとえば, いま “蒸留” という概念のデスクリプタを調べるとして, “distillation” の項を見ると次のようになっている。

Distillation

UF Fractional distillation
Fractionation
Rectification (distillation)

BT Separation

NT Atmospheric distillation
Azeotropic distillation
Destructive distillation
Steam distillation
Stripping (distillation)
Vacuum distillation

RT Azeotropes

—Chemical exchange isotope separation
Coking
Columns (Process engineering)

- Concentrating
- Condensing
- Deasphalting
- Demineralizing
- Desalting
- Diffusion
- Distillates
- Distillation equipment
- Evaporation
- Extractive metallurgy
- Feedwater treatment
- Flashing
- Materials recovery
- Petroleum refining
- Purging
- Purification
- Refining
- Refluxing
- Stills
- Vaporizing
- Washing
- Water softening

RT に挙げられている デスクリプタの前についている一は、そのデスクリプタには NT がある、という意味である。

このようにしてシソーラスに示されているデスクリプタ相互の関係を参考に蓄積、検索を行なうことによって、検索洩れを少なくして再現率を良くすることは可能である。情報検索のシステムのなかには、このようなシソーラスを電算機に内蔵させてデスクリプタの相互参照までを機械にさせているものもあるようである。

シソーラスを整備することにより、再現率を良くすることはできても、これだけで適合率を上げることはできない。適合率を良くするためには、デスクリプタの間に、日本語のテヌヲハに当る、links indicator と roles indicator という補助手段を使う必要がある。これらにはここでは触れない⁴⁾。

4・2 KWIC インデックス

KWIC インデックスというのは、雑誌論文の表題のように、情報の内容をコンパクトな形で表現しているコトバのつながりを利用して、そのなかの任意のコトバを中心にその前後 60 字程度までをプリントし、前後のコトバを参考にして中心のコトバ (key word) の意味する内容を判断することによって、論文そのものが自分にとって必要なものか否かを決定させるためのインデックスである。American Chemical Society 発行の Chemical Titles が KWIC インデックスである。

電算機に論文の表題とアドレスをインプットし、key word 順に並べるという、人間が行なうと大変に面倒な編集のところだけを機械にやらせて、機械にとっては苦

手とする“判断”を人間の頭脳で行なう、というある意味の合理性が KWIC インデックスにはある。

5. 階層分類法とシソーラス

情報の蓄積、検索に自然語を使う場合には、コトバの意味の規制のためにシソーラスが必要であることを述べてきたのであるが、シソーラスというものは要するに、指標 (この場合はデスクリプタ) とするコトバの間の階層関係 (同義、上位、下位など) を明確にすることで混乱を防止しようとするものと言える。ところが、このような階層関係は、たとえば UDC のような古くからある階層分類法では、はじめから明確になっているのである。

UDC は、概念を大きなものから段々と 10 個の数字を使って細分して行って小さな概念を表わし、二個以上の概念の組合せで表わす必要のある概念は、: で標数 (概念を表わす数字) を組合せて表現する分類法である⁷⁾。例として、ポリエステルを表わす標数ができる過程を示す。

6	応用科学
67	種々の工業
678	高分子材料
678.6	高縮合物
678.67	カルボン酸高縮合物
678.674	ポリエステル

3 ケタ目の点は判読の便のためのもので意味はない。

近年アメリカにおいても、自然語のデスクリプタを指標にする場合の新たな欠陥が現われたりしたこともあって、いろいろな改良法が試行されている⁸⁾。

なかでも興味をそそるのは、National Academy of Science の委託研究で、American Institute of Physics (AIP) が UDC を指標とする機械検索のかなり大がかりの実験を行ない⁸⁾、これが完全に可能適切であり、とくに英語という自然語を指標に用いない点が、国際的なシステムの場合に良い方法であると結論していることである。

もともとヨーロッパや日本では、図書分類にも、文献分類にも階層分類法が普及していた。ところがアメリカでは階層分類法の利用を嫌って、コトバだけで情報の処理をしようとしてきた結果、階層分類法では全然問題にならない種類の問題の発生に悩んでいるとも見られるのである。もちろん階層分類法にもドクメンテーションの作業の機械化による省力化のために、改良すべき点がある。しかし、文献情報を処理するのに、概念の階層性を見捨てることの非が明らかになったところで、どちらの道を選んでも行きつくべきところは同じであったというのが今の段階の結論ではないだろうか。

6. 電算機を情報検索に導入する条件

情報検索に電算機を利用することのメリットの一つは検索時間の短縮にあるといわれる。しかし、情報検索に要する時間は蓄積までに要する作業と実際の検索に要する作業とを合せて考えるべきものである。もしかりに蓄積までの作業に、労力と時間とをいかにかけても差しつかえがないのならば、たとえば検索のための各指標ごとに組合せまでを含めて、その数だけのカードを作っておき、さらにそのカードに抄録をも記入しておくことまでできていれば、どんなに優秀な電算機よりも検索時間は短くてすむであろう。なぜならば質問に対応する指標で整理されているカードの入った引出しを開けば、目の前に求める回答が現われるのだから。したがって、電算機を利用して検索時間を短くするという意味は、あくまで蓄積までに要した手間との相対で考えなければならない。

情報検索の効率化についての本質的問題点がどこにあるかについては前述の通りで、電算機を使うことではこの問題の解決にはならない。最近、情報に関するさまざまな分野を、“情報科学”といった名のもとに統合して考えることが行なわれている。システム化がどの道必要なのであるから全体をみることは結構なことである。しかし、統合して眺め、考えただけで、情報科学が誕生するよりもずっと以前から存在していたドキュメンテーション本来の問題が解決するものではない。電算機の特質を生かし効果的に働かせるためには、人間でなければならない難問の解決を含む十分な準備が必要である。

ところで、電算機を検索に導入することは、現在の日本の各企業の状況では、この他にも種々の問題をはらんでいるのである。化学工業のなかでも製薬工業はその他の分野に比べて、RINGDOC⁹⁾ など市販のすぐれた蓄積情報に恵まれている。極言すれば公開された情報の蓄積にはほとんど企業内の人手を要しないほどにすぐれた蓄積情報が、電算機による検索が可能な状態で購入できるのである。このような恵まれた製薬工業の分野においてさえ、電算機の検索への導入の条件について伴（武田薬

品工業）は次のように述べている¹⁰⁾。

(1) 電算機を利用しなければならないほど情報検索の網羅性と迅速性が強く要望されていて、迅速な検索結果にもとづいて迅速に判断と意志決定が行なわれる組織体制ができていないこと。

(2) 電算機で検索できるように、あらかじめ処理された情報が市販されていない場合は、そのような情報処理が確実に遂行できる人手が確保されていること。

(3) 電算機は、従来の機械を使わない図書館活動やドキュメンテーション活動にとって代るものではなく、人力のみでは果し得ない情報活動の飛躍的増強を目指すものであるから、ある程度の人員、スペースの増加は必然的である。人手やスペースを現在より節減することが主目的なら、電算機は導入できない。費用の増加も当然である。

(4) 経営が原価主義一本に立脚している限り電算機の導入はむずかしい。検索の網羅性、迅速性が非常に重要視される場合は、それを金銭的に評価することはむずかしくても電算機化の価値は高い。

(5) 情報検索専用の電算機をもつことができればともかく、一般事務計算用の電算機の一部を利用する場合は、時間的に著しく不自由であることを十分承知の上で、有効利用を工夫しなければならぬ。

(6) トップマネジメントが十分に電算機の価値と限界を認識していることはもちろん大切だが、実際に情報検索を担当しているミドルマネジメントが情報検索の実体をよく体得し、電算機によって解決し得る限界をよく認識することがはるかに重要である。

引用および参考文献

- 1) 中村幸雄：情報処理 I (1968) 共立出版
- 2) 平山健三，寒河 潔：情報管理，8，421 (1966)
- 3) Morse, R.: *Chem. Eng. Progr.*, 57, 55 (1961)
- 4) Holm, B. E.: *ibid.*, 57, 73 (1961)
- 5) *Thesaurus of Engineering and Scientific Terms* (1967) Engineer's Joint Council, New York
- 6) NIPDOC シリーズ 13, シソーラス入門, (1970) 日本ドキュメンテーション協会
- 7) NIPDOC シリーズ 8, UDC 入門—その基礎から応用まで, (1969) 日本ドキュメンテーション協会
- 8) Freeman, R. R., et al.: PB-168885, PB-170436, PB-173704, PB-175926, PB-176152, PB-179673, PB-178374, PB-179679, PB-182083
- 9) 武田敏一：情報管理，10，488 (1967)
- 10) 伴 彰一：ドキュメンテーション研究，18，144 (1969)