情報内容の処理*

菊池敏典** 笹森勝之助** 高橋達郎**

1. 内容分析と主題分析

情報はまず自然言語で発表されるのが普通である．特に重要な情報は紙上に書き残される．したがって，文書（文献）の処理が1次情報処理の中心となる．ここでいう情報とは言語が担う意味的な情報であることはいうまでもない．

1次文献の情報を能率良く流通・蓄積・検索・提供するために，2次資料化を行なうことが多い．これは言語記述を手がかりにして，処理目的のために必要かつ十分に詳しくその内容を把握し，その結果を適当な言語で表現する操作である．結果の表現に用いる言語は，処理目的に合致したものでさえあれば自然言語でなくてもよい．以上の操作を内容分析という．したがって，ここでいう内容分析とは，形式的にいえばある自然言語から分析目的にかなった言語への変換とみなすことができる．たとえば Edmundson はほぼ次のように表現している．

翻訳は　T: $N_1 \xrightarrow{T} N_2$　翻訳操作Tによる，ある自然言語 N_1 から他の自然言語 N_2 への等値変換

抄録は　A: $N \xrightarrow{A} N$　抄録操作Aによる，同一の自然言語における圧縮変換

索引は　I: $N \xrightarrow{I} C$　索引操作Iによる，自然言語Nから検索言語Cへの変換

内容分析法には次の2種類がある．第1は現在人間が頭脳労働として行なっていて，記述事項の理解を前提とする方法である．これを電子計算機で行なうためには人工知能の研究成果を待たなければならない．第2は言語分析を電子計算機が行なう自動化法である．この場合には，内容分析にとって有効な言語学上の法則や技法の探求が現在の課題である．両者の中間に位するもの，つまり電子計算機の助けを借りて人が行なう方法もある．

分析結果の表現の詳しさにより，内容分析を次の2種類に大別できる．第1は原文と同等に詳しく表現するもので，翻訳がこれに該当する．また文献検索や事項検索のために全文探索を行なう機械検索システムにおいて，蓄積用にコード化した2次資料もなるべく詳しい方がよい．この種の自動的内容分析法の特徴は構文分析である．1次資料の構文分析結果をそのまま機械検索用のコードとして蓄積すれば全文探索用のコードであり，分析結果を標的言語（target language）で処理すれば機械翻訳となる．第2は資料から重要事項だけを抽出する場合である．抄録・索引などがこれに該当し，資料が含んでいる意味内容の圧縮とみなすこともできる．

意味内容の圧縮を伴なう内容分析は，現在の IR において中心的な位置を占める重要課題であり，主題分析とも呼ばれている．以下主として主題分析法について述べる．

抄録：索引・分類を次のように定義する．

抄録：文献の概要を自然言語の文章で記述したもの．

索引：主題検索のために，文献に与えたタグ（手がかり）．タグ用の言語を索引言語と呼ぶ．

文献分類：文献の主題が分類体系を形成しているカテゴリーのどれか一つだけに所属している索引．

1.1 抄　録　法

1.1.1 人間抄録法

人間が抄録を作成するときには，次のように行なうとよいとされている．

主題分野別にあらかじめ重要なカテゴリーを定めておく．たとえば

医学：病原，病理，徴候，診断，経過，予後，予防，治療

データ処理分野：オペレーション，実施機関，中心機器，周辺機器，プログラム，システム

さらにこれらのカテゴリー間の相互関係も与えておくとよい．

実際の抄録作業は次の順序で行なう．文献からまず

---

* Technical Processing of Information, by Toshinori Kikuchi, Katsunosuke Sasamori and Tatsuro Takahashi (the Japan Information Center of Science and Technology)

** 日本科学技術情報センター

重要な語句を選び出す．次にそれらの語句間の文献中での意味上の相互関係を把える．最後にそのうちの重要な部分を文章にする．

したがって，抄録者には文献の記述に用いられている言語とその主題分野の両方の知識が必要である．

重要要素は抄録の読者や用途により異なるので，一つの文献から用途別に幾通りもの抄録が作られる．

1.1.2 自動抄録法

電子計算機により重要語句とその相互関係を選出することを目的として，いままでにいくつかの自動抄録法の試みがある．しかしながらいずれの方法も原文の情報の一部のみを利用する近似的な方法に過ぎない．

（1） 統計的手法

次の前提に立脚する一連の語操作法である．文献中の語句の重要度はその出現度数の函数であり，文の重要度は語句の重要度の函数である．語の出現度数が重要度を反映しているとして，キーワード抽出に出現度数を応用するアイデアは IBM の Luhn が 1957 年に提案し，次いで彼は 1958 年にこの前提に基づいた抄録法と英語の文献についての実験をはじめて発表した[1]．少し古いが彼の方法は典型的なので，次に彼の手順の概要を記す．

まず原文中の各語の出現度数を計算する．このさい語の第1字から第6字までに異なる文字がなければ同一語，そうでなければ異なり語とみなす．また前置詞，代名詞，冠詞などの機能語は不要語リストとの照合により除外する．このように計数した出現度数が，あらかじめ定めた一定値以上の語をその文献のキーワードとする．これらのキーワードを用いて次の手順で各文の重要度を計算する．各文について両端がキーワードであって，各キーワード間にはさまる不要語の数が4以上でない語系列のうち最長のものを定める．この最長語系列中のキーワードの平方をその語系列の全語数で除した値がその文の重要度である．

重要度があらかじめ定めた一定値以上の文，または抽出率があらかじめ定めた一定値になるまで重要度の順に抽出した文を原文と同じ順序で並べたのが抄録である．

Luhn の方法はこのように当該文献中での語の出現度数のみに基づいているので絶対度数法と呼ぶことができる．絶対度数法は実験が容易なので，英語・日本語・ロシヤ語・フランス語などの実験例が豊富である．

不要語リストになく，かつその文献中での出現度数が高くても重要でない語がある．Edmundson らが提案した相対度数法は次のような前提に基づいている．その分野で広く用いられている語は，特に特定の文献の主題を代表しているとはいえない．一般にはあまり用いられないが，その文献では特に頻繁に用いられる語が重要である．すなわち相対度数法は特定文献中の出現度数をあらかじめ調査した一般的な度数で修正し，その修正値をもって語の重要度とする方法である．

Edmundson は語の重要度の測度として，次の4種の相対度数を提案した．

$$s_1=f-r,\quad s_2=\frac{f}{r},\quad s_3=\frac{f}{f+r},\quad s_4=\log\frac{f}{r}$$

ここに f: 文献内の度数，r: 対象分野での度数

（2） 構文的手法

原文中の文を構文分析し，構文上の特徴を文や文の部分の重要度の判定基準とする方法である．それぞれの文から修飾要素を削除して，主語・述語・目的語だけを抽出する Climenson らの方法が代表的である．しかしながらこの方法が統計的方法よりすぐれているかどうかは疑問視されている．

（3） 意味的手法

これは人間が抄録を作成する過程を解析して，それをシミュレートしようとする方法である．そのためには，個々の語句のカテゴリーを定める分類表，同義語や関連語を管理するシソーラス，意味的な相互関係の分析を行なう分析規則，重要度の判定規準，抄録文を作り出す生成文法が完備することが前提となる．

しかも操作はカテゴリーパターンの処理を含む複雑高度のものとなる．カテゴリーパターンの処理には，文の命題への書き替え，さらには命題計算などの形式論理的な操作も必要になる．

実際の研究報告はごく初歩的なものが二，三あるに過ぎず，これの本格的な実現には意味の測定法，メタ言語の扱いを始めとして数多くの基礎的な諸問題の検討が必要である．

文献は一定の形式に従って書かれているから，それを主題抽出の手がかりとすることができる．実際，人間が抄録を作成する場合，標題，文献の小見出し，図や表の説明などに着目することによって，主題をスムースに確認することができる．同様に自動抄録においても，標題，小見出し，図や表の説明，パラグラフの冒頭文や末尾文，文献の冒頭パラグラフや末尾パラグラフ，ある種の文献常用語（conclusion, demonstra-

te, disclose, prove, show, summary など）を含む文などに重みを与えて抽出する方法がある．

以上に述べた各種の自動的方法の併用も当然考えられ，実験例もいくつかある．

1.2 主題索引法

人間検索用と機械検索用，キーワード索引と記号索引，索引語間の関係を表示したものとしないものなど，使用目的に応じて各種各様の索引が作られる．主題分析との関係で特に説明を要するのは，キーワード索引と関係記号である．

1.2.1 キーワード索引法

この場合の主題分析は人間か機械がキーワードを抽出する作業である．前述の自動抄録法の第1段階はキーワードの抽出であった．したがって，原文からのキーワード索引作成法は抄録法の一部であるともいえる．すなわち，人間が行なう場合には重要カテゴリーに属する語句の抽出であり，電子計算機で統計的，構文的，意味的のいずれかの手法でキーワードを抽出すれば，自動キーワード索引法である．

たとえば，Luhn の自動抄録法において，抽出されたキーワードをそのままリストアップすれば絶対度数法によるキーワード索引ができる．また Baxendale は前置詞句を構成するとみなされる語のみについて度数を数え，高出現度数語をもってキーワードとする方法の実験を行なった．これは統計的手法と構文的手法を併用した方法である．

重要な情報は標題や抄録に濃縮されているので，これらからキーワードを抽出する方法もある．

たとえば，Chemical Abstracts の Subject Index は抄録から人間が作成している．自動索引の代表例は標題キーワード索引（KWIC, KWOC）である．

標題や抄録からのキーワードの抽出法としては，キーワードリストとの照合によりキーワードリストに登録されている語句のみを自動的に抽出する方法も考えられ，実験も行なわれている．

1.2.2 検索語間の関係の表示法

キーワード間の相互関係を抄録では自然言語の文脈によって表現している．しかしながら自然言語の文章は多義性や冗長性などのために検索言語として最適とは限らない．そこで検索側から重要な関係を人為的に定義してそれを表現する方法がある．

第1の方法はキーワードとその相互関係をグラフで表現する．このグラフの頂点はキーワードで，辺は二つのキーワード間の関係の方向と種類である．これは機械検索用の蓄積文献の表現法である．検索質問も同様なグラフで表現し，質問のグラフを部分グラフとして含む文献のグラフをさがすことが検索操作である．抄録から自動的に文献情報のグラフを作成する試みとしては SMART, SYNTOL などがある．いずれも構文分析と簡単な意味分析を併用する方法である．

第2の方法はロールとリンクで表現する方法である．リンクは相互に結び付いているキーワードのグループを指示する指標で，ロールは個々のキーワードの文献中での役割を指示する指標である．いずれもキーワードの限定要素としてキーワードに付記する場合が多い．ロール，リンク付き索引の詳しいものは抄録と同量の情報を含む．実際，Western Reserve University の電報抄録から英文抄録への自動変換（索引言語から自然言語への機械翻訳）の試みもある．したがって，この索引の作成法は抄録の作成法に準ずるとみなしてよい．ただし原文からこれを作成するには詳しい意味分析が必要となるので，関係記号を明示した索引作成の自動化は現状ではかなりの苦労を要する．

キーワード間の相互関係にはこの他にファセット分類の phase 関係があるが，これはロールに準ずるものとみなせる．

1.3 分類法

文献分類では主題分析の結果は単独の分類記号で表示される．したがって分類はもっとも抽象度の高い主題分析である．

1.3.1 人間が行なう文献分類法

主題分析の基準は分類体系によって与えられる．すなわち，人間が分類を行なうときには文献の主題は次の手順で定められる．まずもっとも generic なレベルのカテゴリーを定め，それから分類体系で与えられた下位へと順次主題を特殊化していく．したがって，文献を分類するには分類表がまず必要である．さらに文献分類者は使用する分類体系でのカテゴリーの構成を知らなければならない．

1.3.2 自動分類法

分類カテゴリーを自動的に設定することを自動分類という．自動分類法としては，Needham らのクランプの理論，Borko らの因子分析法，Baker の潜在構造分析法，山川らの計量心理学的方法などがある．

ここで扱うのはキーワードを分類する方法である．まず，相互に関連度の高いキーワードが一つのグループを形成するように，全部のキーワードをいくつかのグループに分割する．この個々のグループに対して，

そのキーワード群の特徴を表わす語を人が与えれば，この語がそのグループを代表するカテゴリー名である．

クランプの理論が最も一般的であると思われるので，主としてDaleの実験[2]に基づいてクランプの理論を紹介する．これは，ある集合を考え，その要素が持つ性質の類似性によって集合をいくつかの部合集合に分割する方法である．この部分集合をクランプ(clump)という．いろいろのクランプが考えられているが，最もよく使われるのはGRクランプである．

GRクランプは次のように定義される．

〔定義〕

U: 要素の有限集合．要素の対の間には実数で表わせる対称関係が存在する．これを対の結合と呼ぶ．

A: U の部合集合

B: A の補集合

x: U の要素

a_i: A の要素 $(a_1, \cdots\cdots, a_t)$

b_i: B の要素 $(b_1, \cdots\cdots, b_r)$

$c(x, s)$: 要素 x と s の対の結合

$C(x, A)$: $\sum_{i=1}^{t} c(x, a_i)$

$C(x, B)$: $\sum_{r=1}^{t} c(x, b_i)$

$b(x, A)$: $C(x, A)-C(x, B)$．これをバイアスと呼ぶ．

GRクランプ A: $\{x \mid b(x, A)\geq 0, b(y, A)<0, y\epsilon B\}$ Aに属する全要素が，Aに対して非負のバイアスを持ち，Bに属する全要素がAに対して負のバイアスを持つとき，U の部合集合AはGR-クランプである．ここに $c(x, x)=0$ とする．

結合の測度 $c(x, s)$ としては，以下のような3種の測度を使用する．

1. $l(m, n)$
2. $\dfrac{l(m, n)}{l(m)+l(n)-l(m, n)}$
3. $\dfrac{l(m, n)}{\sqrt{l(m)\cdot l(n)}}$

ここに $l(m, n)$: 文献集合におけるキーワードmとnの共出現度数

$l(m), l(n)$: 文献集合におけるキーワード m, n の出現度数

GRクランプを求めるには，まず適宜の方法でUをAとBに分割し，ついで以下の手順を実行する．

(1) $b(x, A)$, $x\epsilon U$ を計算する．

(2) もしxがAの要素でない場合，$b(x, A)\geqq 0$ ならxをAへ移す．

(3) もしxがBの要素でない場合，$b(x, A)<0$ ならxをBへ移す．

(4) 移したたびにバイアスを計算し直す．

(5) U全体について完全にスキャンが終り，移される要素がなくなるまで，これを繰り返えし続ける．

(6) Aはクランプである．

もしAが無意味な集合（空集合かU）となったら，Uの別の分割を定めて上の手順を繰り返えす．

GRクランプからさらに内部結合力の強い縮小クランプを求めることもできる．

Daleは3種の測度につき，GRクランプおよび2種の縮小クランプをそれぞれいくつか求めている．また，得られた結果から求まるキーワードの関連ネットワークを図示している．

現在の自動分類法で得られるカテゴリーは，対象分野を単に区分するものでしかない．何段階もの階層構造をもつ分類体系を自動的に設定する方法は将来の課題である．GRクランプから縮小クランプを得る手続きなどからこれに対する示唆が得られる．

1.3.3 自動文献分類法

文献が所属するカテゴリーを自動的に定める方法である．これには2通りの手法が考えられる．第1は文献間の類似度に基づいて文献集合をいくつかの部分集合に分割する方法である．得られた部分集合の名前がカテゴリー名となる．したがってこの手法では文献が分類されると同時に分類カテゴリーが設定されるので，1.3.2の自動分類法とみなすこともできる．Baker, Winterは文献を標本，キーワードを項目として，潜在構造分析を文献分類に応用している．またKessler, Saltonらは引用文献のパターンの類似度により，文献分類を行なっている．

第2は既存のまたは自動的に設定した各カテゴリーに対する文献の関連度を適当な測度で表現し，最高の値を示すカテゴリーにその文献を所属させる手法である．通常次の四つの手順からなる．

(1) 学習サンプルを用いてキーワードを定める．

(2) 学習サンプルを用いて各カテゴリーに対する各キーワードの関連度を求める．

(3) キーワードの関連度からカテゴリーを推定する函数を定める．

(4) 実際に文献を分類する．すなわち，文献中のキ

ーワードからカテゴリーを計算し，最高値のカテゴリーを選ぶ．

この手法は Maron, Borko, Williams が試みている．

Maron のカテゴリー推定法

キーワード $\{w_k, w_m, \cdots\cdots, w_s\}$ を含む文献がカテゴリー C_j $(1<j<32)$ に所属する確率 $p\{C_j/w_k \cdot w_m \cdots\cdots w_s\}$ を次式で計算する．カテゴリーに関してキーワードの出現が独立であると仮定している．

$$p(C_j/w_k \cdot w_m \cdots\cdots w_s) = k \cdot p(C_j) \cdot p(w_k/C_j) \cdot p(w_m/C_j) \cdots\cdots p(w_s/C_j)$$

ここに

k: $\sum_{j=1}^{32} P(C_j/w_k \cdot w_m \cdots\cdots w_s)=1$ とするための定数

$p(C_j)$: 文献が C_j に分類される先験確率
学習サンプルで，C_j に所属する文献数を全文献数で割った値

$p(w_i/C_j)$: 学習サンプルで，C_j に所属する文献中の w_i の数を C_j に所属する文献中のキーワードの総数で割った値

Borko の推定法

因子分析で求めたカテゴリーへの分類である．次の推定式を用いる．

$$p_i = \sum_{j=1}^{90} (L_{ij} T_j)$$

ここに　p_i: 第 i カテゴリーに属する推定値

L_{ij}: カテゴリー i におけるキーワード j の正規因子負荷量で，学習サンプルについて因子分析を行なった結果得られる．

T_j: キーワード j の出現頻度

Williams の判別分析法

各カテゴリーにおけるキーワードの頻度分布を利用する方法である．学習サンプルからまず判別係数 λ を求める．第 i キーワードの判別係数 λ_i は

$$\lambda_i = \sum_{j=1}^{n} \frac{(p_{ij}-\bar{p}_{ij})^2}{\bar{p}_{ij}}$$

ここに　$p_{ij}=f_{ij}/\sum_{i=1}^{m} f_{ij}$, $\bar{p}_{ij}=\frac{1}{n}\sum_{j=1}^{n} p_{ij}$

f_{ij}: 第 j カテゴリーにおける第 i 語の出現頻度

p_{ij}: 第 j カテゴリーにおける第 i 語の相対頻度

$\bar{p}_{ij}$: 第 i 語のカテゴリー当たりの平均相対頻度

次に分類方程式を導入する．これは対象文献のキーワードの頻度分布（キーワードの観測頻度）を学習サンプルについて求めた各カテゴリーの理論頻度分布と比較して，その文献に各カテゴリーに対する関連値 (RV_j) を与える式である．

$$RV_j = 1 - \left[\frac{0.01}{m}\sum_{i=1}^{m}\left(\lambda_i \frac{(p_{io}-p_{ij}^*)^2}{p_{ij}^*}\right)\right]$$

ここに p_{io}: 対象文献中での第 i キーワードの相対観測頻度

p_{ij}^*: 第 j カテゴリーでの第 i キーワードの相対理論頻度（文献の大きさに対する変換を行なったもの）

m: グループ中の語型の数

1.4 内容分析における人間と機械の協力

電子計算機が行なった結果は完全ではないので，その結果に人間が手を加える方法が考えられる．Doyle[3] はこれを post-editing と称して，その意義を**第1表**のようにまとめている．

第1表　post-editing の意義

処理（5,000 語の記事を仮定）	推定圧縮率〔%〕	編集（入力1ビットあたりの労力）	編集作業内容
機械翻訳	100	100（任意の基準）	最終的な選択；いいかえ；辞書の修正
パラグラフ単位の索引作成	5	10	新語，エラーの管理；シソーラスの保守
自動抄録作成（4個の文の抽出）	2	5	選択した文の書き替えと圧縮
関連文献のグループ分け	0.5	3	カテゴリー構造の修正；あいまいな表現の明確化
引用索引作成	0.3	0.6	引用文献結合で得たグループを組織化し，それぞれにグループ名（小見出し）を与える．
半自動的な標題作成	0.3	0.6	キーワードとフレーズを連結して，（索引に耐える）標題を作成する．
10,000 の蓄積文献用の分類体系の表現	0.02	0.05	表現の構造と意味的特性の修正

第1表の第3カラムがそれぞれの作業に要する労働経費の相対値である．この表は圧縮率の高い処理法の方が post-editing は経済的に実用価値があることを示している．

2. コード化

主題分析の結果は，システムの目的に適した言語で表現しなければならない．この言語をコードといい，コードで表現することをコード化という．

分類に用いる個々のコード系を分類表といい，索引に用いる個々のコード系をシソーラスまたは件名標目表という．件名標目表というのは，従来，図書館の目録作業に用いられた件名標目（事項見出しともいう，

subject heading) の ABC 順リスト（または五十音順リスト）のことである．シソーラスは件名標目表の発展したものとみなされることが多いので，ここではシソーラスのみを考える．

分類表を用いて分類（コード化）された文献は，ファイル中でのアドレス（位置）が一元的に定まり，シソーラスを用いた場合は一元的には定まらない．このことは，シソーラスを用いたシステムでは多元的な検索が可能であり，分類表を用いた場合はそうでないことを示している．ただし，分類は索引につながる面があるので，狭義の分類として上記のように操作的に分類を定義した場合に，始めて分類と索引の区別が成立するというべきかもしれない．

既存のコード系の中で，Western Reserve 大学の semantic code は，やや特異な地位を占めるので，シソーラスと分類表との中間において説明したい．

2.1 シソーラス

2.1.1 シソーラスの構成

シソーラスに含まれている個々のコードは，自然語の術語である．

シソーラスの中では，個々の検索システムによって定められる術語間の関係を明示しておくのが普通である．術語間の関係としては次のものがある．

(1) 同義語の関係
(2) 上位語（下位語）の関係
(3) 関連語の関係

既存のシソーラスとして，アメリカの工学連合会 (EJC)，化学工学会 (AIChE)，国防省ドキュメンテーションセンター (DDC) 発行の各シソーラスが有名であるが，これらはいずれも上記の関係をシソーラスに導入している．ただし**第2表**のように関係を指示する記号は異なっている (ASTIA は DDC の旧称)．

第2表 術語間の関係を示す記号

シソーラス / 関係	EJC	AIChE	ASTIA	一般図書館用（件名標目表）
同義	USE	SEE	Use	See
同義	UF (Used for)	SF (Seen from)	Includes	See
上位	BT (Broader Terms)	PO (Post on)	Specific to	See also
下位	NT (Narrower Terms)	GT (Generic to)	Generic to	See also
関連	RT (Related Terms)	RT (Related Terms)	Also see	See also

EJC の記号を用いて，これらの関係の例を示してみよう．

(1) USE:

例 ACCELERATION MEASUREMENT
　　USE ACCELEROMETERS

“ACCELERATION MEASUREMENT の代わりに ACCELEROMETERS をキーワードとして用いよ”ということである．すなわち同義語の管理である．

(2) UF (Used for):

例 ACCELEROMETERS
　　UF ACCELERATION MEASUREMENT

“ACCELERATION MEASUREMENT の代わりに ACCELEROMETERS がキーワードとして用いられている”ということであって，USE 関係の逆である．このように両方向からの関係を示しておくと，シソーラスの使用上および管理（ファイルメンテナンス）上，非常に有効である．

(3) BT (Broader Terms):

例 DATA REDUCTION
　　BT DATA PROCESSING

“DATA REDUCTION の上位語には，DATA PROCESSING がある”ことを示している．主題分析や検索をより一般化したい場合に利用できる．

(4) NT (Narrower Terms):

例 DATA PROCESSING
　　NT DATA REDUCTION
　　DATA SMOOTHING

“DATA PRCCESSING の下位語には，DATA REDUCTION や DATA SMOOTHING がある”ことを示している．主題分析や検索をより特殊化したい場合に利用できる．

(5) RT (Related Terms):

例 DATA PROCESSING
　　RT ACCOUNTING
　　AUTOMATION
　　CODING THEORY
　　COMPUTERS
　　DATA TRANSMISSION
　　SYSTEMS ENGINEERING

“DATA PROCESSING に関連する語として，ACCOUNTING, AUTOMATION,……, SYSTEMS ENGINEERING がある”ということである．主題分析や検索にあたって，その範囲を変更するのに利用できる．

コード化に際しては，シソーラスを次のように利用する．まず文献を主題分析し，その文献の内容を特徴づけると思われる術語を必要な数だけ抽出する．これらの術語をシソーラスで引いて，術語間の関係を考慮しながら，実際に使用すべきキーワードを定め，これらのキーワードを文献のコードとする．さらに，多くのシステムでは，1.2.2 で述べたリンク，ロールなどを用いて，各文献内でのコード（キーワード）間の相互関係を指示しておくのが普通である．

2.1.2　各種シソーラスの比較

いろいろの分野でシソーラスが作成されているが[4]，そのほとんどはアメリカ製である．わが国では，外務省や通産省が自家用に作成している．

公刊されたシソーラスの代表として，前記 EJC, AIChE, DDC（旧称 ASTIA）の3種のシソーラスを選び，比較してみよう．

（1）　書名および出版年

EJC: Thesaurus of Engineering Terms. 1964

AIChE: Chemical Engineering Thesaurus. 1961

ASTIA: Thesaurus of ASTIA Descriptors, 2nd ed. 1962

（2）　対象分野

EJC: 工学

AIChE: 化学工学

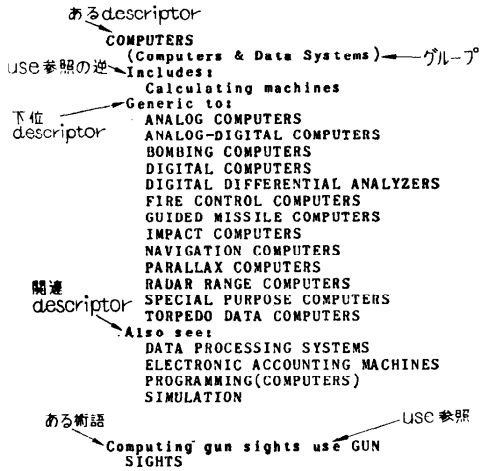
ASTIA: 軍に関係する分野

（3）　構成および収録術語数

EJC: 10,515 語．ABC 順配列

AIChE: 化学工業部門 4270 語，化学部門 2790 語に二分し，それぞれ ABC 順配列

ASTIA: 約 10,000 語．ただしキーワード（DDC では descriptor と呼んでいる）として使用可能なものは約 5500 語で，残りの語にはその代わりに使用すべき descriptor への参照が指示されている．第1部と第2部に分けられている．第1部では，約 5,500 の descriptor を 26 のフィールドの下の 170 のグループに分類している．第2部はスコープ・ノート・インデックス（相関索引）と呼ばれ，descriptor と術語（計，約 10,000 語）を ABC 順に配列したもので，術語にはその descriptor が示されている．これがいわゆるシソーラスの部分といえよう．descriptor には，それが属するグループ名や他の descriptor との関係が示されている．これをスコープ・ノートという．descriptor はすべて大文字で示されている．descriptor でない術語には，どの descriptor を代わりに使用すべきかが，Use 参照で示されている（**第1図**参照）．



第1図　スコープ・ノート・インデックスの例

2.1.3　シソーラスの自動編集

シソーラスの編集は相当の労力を要する作業である．たとえば外務省での経験[5]によれば，10,263 キーワードのシソーラスを作成するために約2年3ヵ月を要しており，その間の実質作業総人日数は延 2,200 人日（17,600 時間）に及んでいる．その他の例をみても2年くらいの日数は普通にかかるようである．したがって，労力を節約し，作業を標準化するためには，編集手順を極力機械化することが望ましい．

シソーラスの編集は，（イ）語彙の収集，（ロ）語彙の整理，の二つの作業に大別される．語彙収集（語の出現頻度の計算やソーティングなどを含む）の機械化は，コストの点を除けばあまり問題はないが，語彙の整理については簡単ではない．語彙の整理とは，収集した語彙からシソーラスに採択すべき術語を選定し，術語間の関係（同義，上位と下位，関連）を定める作業なので，シソーラスの対象とする分野についての専門知識を必要とするからである．しかし，この面でのシソーラス編集の自動化について，二，三のアイデアが発表されているので，代表的な Giuliano の方法[6]を紹介する．

Giuliano は，**3.3** で示す形式で線型関連情報検索法と呼ぶ一般化した情報検索システムを提唱している．さらに，線型関連検索の実際の型を求め，それを

シソーラスの自動編集に応用することを論じている．その要点は次のとおりである．d 文献の集合と t キーワードの集合が与えられ，各文献には1個以上の適切なキーワードが割当てられていたとする．文献 i はその文献中に含まれるキーワード j と強さ $C_{ij}>0$ のボンドで結合していたとする．この値は，たとえばキーワードの生起頻度をとってもよい．このとき $d\times t$ 行列 (C_{ij}) を，文献対キーワード結合行列と呼ぼう．次にある質問が与えられたとして，この質問を t 次の列ベクトル Q によって表わす．その要素 q_l は質問者が第 l キーワードに与えた値であって，第 l キーワードが質問に含まれる場合は正のある値をとり，そうでなければ0となる．この質問に対する検索システムの応答は，蓄積中の全文献に対して非負の値を割り当てることである．この値は質問に対する有効性の反応であり，d 次の応答ベクトル R を定める．その要素 r_m は，質問 Q に対応してシステムによって文献 m に割り当てられた値を表わす．

したがって検索プロセスは Q から R への数学的変換とみなされる．変換公式を求めるために，次のような線型性の仮定を置こう．

（1） 文献の値はそれに含まれるキーワードの値の線型函数である．C の行の和を1に正規化した $\tilde{C}$ をこの函数の係数，W をキーワードの値のベクトルとすると

$$R=\tilde{C}W \qquad (1)$$

（2） キーワードの値は，質問 Q により与えられるそのもとの値と，それを含む文献の値の線型函数 $\lambda{}^t\tilde{C}$ の和によって与えられる．ただし λ は正規化定数である．

$$W=\lambda{}^t\tilde{C}R+Q \qquad (2)$$

キーワード結合行列 $K={}^tCC$ を考え，正規化されたキーワード結合行列 $\tilde{\tilde{K}}=\lambda\tilde{K}=\lambda{}^t\tilde{C}\tilde{C}$ について

$$\lim \tilde{\tilde{K}}^n=0 \qquad (3)$$

の条件を置くと，(1)，(2) 式から，次の変換公式が得られる（I は単位行列）．

$$R=\tilde{C}[I-\lambda\tilde{K}]^{-1}Q=\tilde{C}[I+(\lambda\tilde{K})+(\lambda\tilde{K})^2+\cdots]Q \qquad (4)$$

$$W=[I-\lambda\tilde{K}]^{-1}Q=[I+(\lambda\tilde{K})+(\lambda\tilde{K})^2+\cdots]Q \qquad (5)$$

（4）式が線型関連情報検索法の検索公式を与えるものである．（4）式の収斂の早さは λ によって定まり，$\lambda=0$ のときが従来のいわゆる論理検索となる．

文献間の直接結合およびキーワード間の直接結合を導入することによって，Giuliano は (4)，(5) 式をさらに一般化しているが，ここでは簡単のために (5) 式から考えよう．(5) 式は与えられた d 文献の蓄積にとって有効な，シソーラス様のリストを作成するのに用いることができる．すなわち，Q_z はキーワード Z にのみ値を与えた単位ベクトルであるとする．このとき (5) 式から得られる W_z の値は，キーワード Z と関連する程度に従って各キーワードに順序を付ける．W_z 中の上位の数個のキーワードをリストすれば，シソーラスのリストが作れる．

次に (5) 式の右辺の係数の各項に立入って考察してみよう．第1の項 I は，キーワード Z がそれ自身に対してもつ関連を表わしている．第2項の係数は，与えられた文献の中にキーワードの対が現われる回数に関係しているので，これはいわゆる関連語の関係，あるいは概念連合の関係を反映しているものと推測される．あるキーワードに関して，他の全てのキーワードとの $\lambda\tilde{K}$ 関連の集合は，概念連合プロフィルと呼ぶべきものである．

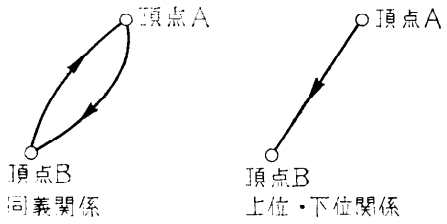
第3項 $(\lambda\tilde{K})^2$ は，概念連合プロフィルの類似性にのみ関係するキーワード間の関連の測度を生成する．すなわち，$(\lambda\tilde{K})^2{}_{ij}$ は，キーワード i と j が同じプロフィルを持つときのみ大きくなる．同義語的キーワードは，そのプロフィルの類似性によって確認されると推測されるから，この係数は主として同義語による関連の測度を与えるものとみなされる．さらに，もし上述の推測が両方とも有効であるとすれば，$(\lambda\tilde{K})$ の全ての奇数べきは，主として概念連合による関連を表わし，$(\lambda\tilde{K})$ の全ての偶数べきは，主として同義語による関連を表わしているといえよう．

以上が Giuliano の方法の要点である．これを実行するには膨大な計算を必要とするので，Giuliano は，今のところ小規模な実験例しか示していない．

Parker-Rhodes and Needham, Maron and Kuhns, Stiles, Salton, Osgood, Bennett and Spiegel 等も，それぞれ，関連の測度は異なるとはいえ，語の関連についての研究を行なっている．

2.1.4 シソーラスとグラフ理論

Abraham[7] は，シソーラスの数学的モデルとして方向性のあるグラフ（directed graph）を考えている．グラフは頂点（vertex）と辺（edge）とからなるが，術語を頂点とし，術語間の関係を辺で表わすこととする．ただし，術語の選定は既になされており，術語間の関係（同義および上位・下位）も全て確定して



第2図　術語間の関係

いるものとする．同義および上位・下位の関係は**第2図**のように表わされる．

グラフは葉（leaf）に分割でき，葉はさらに葉片（lobe）に細分割できるという性質を利用して，シソーラスを数学的に解析するアルゴリズムおよび同一の術語から構成されている異なるシソーラス間の比較を行なう方法を Abraham は示している．

Abraham は関連語の関係を考慮せず，また同義関係と上位・下位関係を始めから与えられているものとしているので，シソーラスのモデルとしては十分でない．しかし，グラフ理論が回路網理論など各方面で有力な武器となっていることを思えば，シソーラスの取扱いにおけるグラフ理論のさらにすすんだ適用を期待してもよいであろう．

2.2　分析合成関係（W.R.U. の semantic code）

主題をまずいろいろの主題アスペクトに“分析”し，続いて数種のアスペクトを“合成”して主題をコード化するシステムである．主題のアスペクトを表現するコードは semantic factor とか modulant とか呼ばれる．各 semantic factor と主題間の関係を定義するために，分析関係が用いられる．semantic factor で表現された主題の相互の関係は，合成の段階で指示される．このような合成関係を指示する記号を role, role indicator, link, operator などと呼ぶ．

コード化にあたっては，まず主題の属するクラス（複数）を定め，そのクラスの中の semantic factor を定め，semantic factor と主題との分析関係を定める．ついで各主題間の合成関係が指定される．

これは，第4図（2.3.2 参照）に示すものと同様な operator 頂点を有する木を生成することと同格である．この木の中で，nonoperator 頂点は階層的な木から選定されたものである．

この例としては，Western Reserve 大学（W.R.U.）が米国金属学会（ASM）と共同で開発した機械検索用の semantic code があげられる．

W.R.U. のコードでは，たとえば“圧延”という主題については，「動的」，「金属」，「プロセス」という三つの semantic factor が選定される．

圧延：動的　　M-CL
　　金属　　M-TL
　　プロセス　　P-SS

このコードの空位には，主題と各 semantic factor との関係を示すコード（infix code という）が挿入される．結果だけを示すと“圧延”は次のように表現される．

MQCL.　MWTL.　PASS.　001.
（動的に行なうもの）（金属に利用するもの）（プロセスの一種）（類概念の細分）

実際の文献は，“圧延によるベリリウムの機械加工”というように，いくつかの主題から構成されるので，各主題間の関係を示すために次のように role indicator が用いられる．

主　題	コ　ー　ド	role indicator
圧　延	MQCL. MWTL. PASS. 001.	KAM.（プロセス）
ベリリウム	MATL. 4. □BQE.	KEJ.（処理される材料）
加　工	CUNG. MWTL. PASS. 003.	-KAM.（プロセス）

これを，以下の形式で磁気テープに蓄積する．

KEJ. MATL. 4. □BQE., -KAM. CUNG. MWTL. PASS. 003., KAM. MQCL. MWTL. PASS. 001.

W.R.U. の semantic code は，後述するファセット分類の一種ともいえる．しかし，通常のファセット分類は人間による検索を目的としているのに対し，semantic code は機械検索用に作成されたことが，おそらく最大の相違点である．分析関係によってキーワードと術語の関係を規定してゆくという点では，これはシソーラスの構成原理と同じである．

2.3　分類表

分類表とは，同一の主題を正しく同定し，関係する主題間の望ましい関連を表示するように配列した主題のリストであって，木（tree）の構造を示している．各主題は，数字または文字のコードで表現されている．客観的な不変の主題カテゴリーは存在しないで，いろいろの分類表が存在する．ここでは，Salton[8] の考え方を参考にしながら，分類表を2種類に大別し，各々について説明を加えたい．

2.3.1　十進分類およびその他の階層分類

この種の分類は，通常，いくつかの階層配列からなる．各階層の内部では，主題は相互排他性と網羅性と

を期している．各階層は多重レベルの木で表現される．下位レベルの主題は上位レベルの主題に包含され，細分された主題のクラスを表現する．

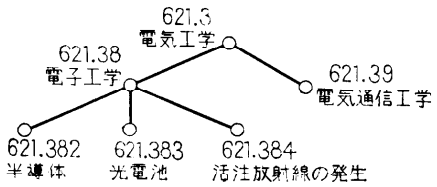
異なる階層に属する主題間の関係は決定されないのが普通である．階層間の順序付けは用意されないし，一つの階層の中でさえ一貫した特性付けが欠如していることも多い．それ故，包含関係は，木の内部についても外部についても多種多様の関連を表現している．

この種の分類の例としては，国際十進分類（UDC），日本十進分類（NDC），デューイ十進分類（DC），カッター分類などがある．国際的にはUDCが最も通用しているので，UDCの例をあげよう．

UDCでは，知識の全分野を次のようにまず10の大きな主題に分け，0，1，2，……，9の数字のどれかを割り当てる．

0 一般事項，総記
1 哲学
2 宗数，神学
3 社会科学
4 言語学，語学
5 自然科学
6 応用科学，医学，工学，農学
7 美術，演芸，娯楽，スポーツ
8 文学，純文学
9 地理，歴史，伝記

続いて各項目を詳しく展開してゆく．展開のレベルが増すごとに数字の桁数が増える．便宜上，3桁ごとに点が打たれる．かくて，たとえば，6 応用科学を展開してゆくと，5桁目に電子工学が現われる．



第3図 UDCの例

木の個々の頂点にふられている数字を標数といい，0，1，2，……，9の下に展開される標数を主標数という．主標数の他に補助標数がある．補助標数は，主標数間の関係を表わしたり，主標数に付属して主として形式に関する情報を表現する．

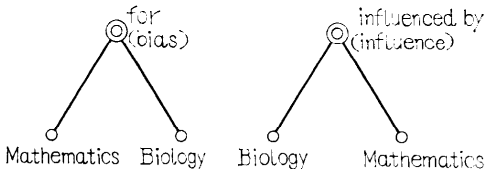
このようにUDCは，全体として1本の木から構成されているようにみえるが，実際には多数の木からなっている．たとえば，“鉄”を表わす標数は，無機化学（546），鉱物学（549），鉱床学（553），鉱山工学（622），冶金学（669）というそれぞれの木の中のある頂点として存在する．

UDCによる文献の分類とは，まずその文献が属すべき木（すなわち特定の主題分野）を定め，その木の下位レベルに文献の主題にマッチする標数を見出すことである．

2.3.2 ファセット分類

主題を表わすコード間の関係をもっと精密に規定したものがファセット分類である．ファセット分類では，予め定められた観点に基づいて一つの主題をいくつかの主題アスペクトに分析する．このアスペクトをファセット（facet）という．通常，ファセット間の順序は指定されている．また，ファセットの内部での配列は原則として階層性が保たれている．したがって，一つのファセットは一つの木に対応する．

異なる主題間の関係は，いわゆるphase関係によって与えられる．これには，bias, influene, comparison, toolの諸関係がある．これらの関係は，operator頂点によって木の中で表現することができる．“生物学者のための数学”は，“数学（bias）生物学”で表現し，“生物学に及ぼす数学の影響”は，“生物学（influence）数学”で表わすことができる（**第4図**参照）．



第4図 Phase関係

ファセット分類の例としては，知識の全分野についてのRanganathanのコロン分類の他に，特定の主題分野について幾つかのファセット分類がある．それらは，次のようなファセットを用いている．

コロン分類（Ranganathan）：

Personality, Matter, Energy, Space, Time

ASM-SLA分類（米国金属学会，米国専門図書館協会）：

プロセスと性質，製品と装置，材料

土壌科学のファセット分類（Vickery）：

土壌の種類，構造，成分，性質，土壌中で起る

諸過程，土壌に施す操作，操作に用いる物質と装置，実験技術，一般

コンテナ技術 (Foskett):

製品，部品，材料，作業，雑

Foskett のファセット分類を用いてコード化の例を示そう．“Extrusion coating of polythene on paper（ポリエチレンによる紙の押出し被覆)”という文献は次のようにコード化される．

BuDvFfj: Dgl

主　題	ファセット	分類表中の術語	コード
Extrusion	作　業	Extrusion	Ffj
Coating	材　料	Coating	Dv
Polythene	製　品	Polythene film	Bu
On	雑(関係)	Relation	:
Paper	材　料	Coated paper	Dgl

3. 検索効率

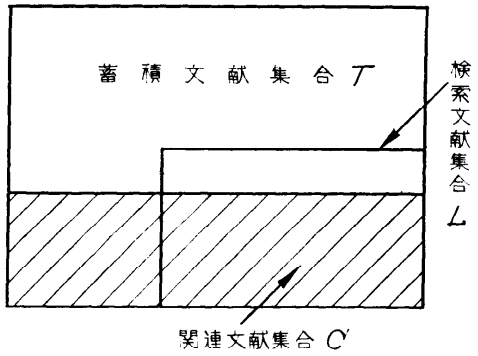
3.1 効　　率

情報検索には，事実検索 (Fact Retrieval) と文献検索 (Document Retrieval) とがあることは，よく知られていることである．前者はその事実に関しては確定的なモデルであるから，自然語のあいまいさの問題を別とすれば，検索システムに処理上のミス以外の誤差の入り込む余地は，ほとんど考えられない．しかるに後者は主題分析，コード化の段階で，文献という包括的な意味内容を，インデックスの集合という形に物理的にも内容的にも縮約した変換を行なっているので，この変換に可逆性がなく，したがってこの変換にともなう必然的な誤差が生ずる．文献検索のプロセスは下図のようになっている．

	ファイル	質問
情報レベル	情　報	情報要求
	↑	↓
言語レベル	文　献	質　問
	↑	↓
コードレベル	インデックス ←	インデックス

文献検索にともなう誤差は，上図の各プロセスにおいて生ずるが，情報要求に対して合目的文献が得られる確率をもって検索効率といっている．以下本項においてはこの文献検索の場合のみを考察することとする．前者が確定的なモデルであるのに対して，これは確率的なモデルであるといえよう．

第5図はある質問に対する検索された文献の集合と，蓄積中におけるその質問に関連ある文献集合との関係を示している．これらの関係を量的に表わすため



第5図　検索文献集合と関連文献集合の関係

の測度として，次の係数が一般に用いられている．

$\alpha=\frac{R}{C}$　　検索率 (recall factor)

$\beta=\frac{R}{L}$　　適合率 (relevance factor)

L: 検索された文献集合の大きさ

C: 関連文献集合の大きさ

R: 検索された関連文献集合の大きさ

これを誤差の観点からみると

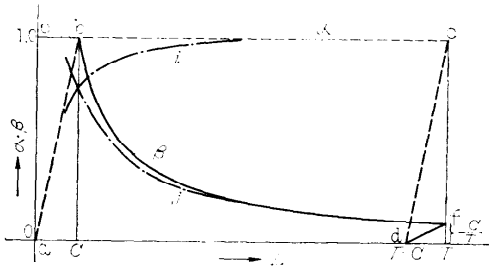
$E_1=1-\alpha$　　(omission factor)

$E_2=1-\beta$　　(noise factor)

といっている．

α, β の間には上式から

$\alpha=\frac{L}{C}\beta$



第6図　検索率，適合率，検索文献集合の関係

なる関係があるが，これを示したのが**第6図**である．検索の立場からはもちろん E_1, E_2 をともに0に近づけることが望ましいが，図から明らかなように，ともに0となるのは $L=C$ の場合のみである．図の破線で囲んだ部分は α の値の取りうる範囲，実線で囲んだ部分は β の値のとり得る範囲である．一般に実験の結

果によれば，α, β はそれぞれ鎖線で示した i, j のようになるのが普通である．

検索システムを設計するに当たっては，誤差を極力小さくするように設計しなければならないが，そのためには誤差の原因を究明する必要がある．しかし言語情報については，因子を抽出しそれを定量的に扱うことが困難であるから，普通実験によって誤差の原因を見出すとか，効率のよい方法を見出すとかの方法をとっている．

なお前述の α の式に含まれている C の大きさは，一般に未知数であるから，これを推定しなければならない．ファイルの文献集合の大きさに比し T, C の大きさは非常に小さいから推定にはランダム・サンプルを用いるのは困難であり，別途の考慮が必要である．実験の場合には全数調査によりファイルの内容を全部確定しておくとか，その他便宜的な方法を用いていることが多い．

3.2 効率測定のための実験

効率測定の実験として最も有名なのは，Cranfield Project といわれるものである[9]．英国の College of Aeronautics で米国科学財団（NSF）の援助の下に行なわれた実験で，航空工学に関する 18,000 件の文献についての結果が報告されている．これにより文献検索はインデックスによりいかにその効率が左右されるかがわかろう．

インデックスとしては，次の四つのシステムが用いられた．

(1) UDC：体系分類
(2) 件名標目：統制された語彙
(3) ファセット分類：概念のカテゴリー別に作られた体系分類
(4) ユニターム：キーワード

18,000 件の文献をそれぞれこの四つのシステムでインデックスし，蓄積ファイルを作成した．質問は 18,000 件の蓄積文献の中から 1,500 件作成した．したがって各質問については回答に相当する文献が少なくとも 1 件は存在することとなり，検索においてその質問のもととなった文献が検索されたとき，その検索は成功したと判定した．これは質問について適合しているかどうかを，全蓄積文献について調べる手間をはぶくためであった．

この報告から二，三の主な結果を紹介しよう．

（1） 総合結果

	全検索数	成 功	失 敗	成功率〔%〕
UDC	1157	875	282	75.6
件名標目	1154	941	213	81.5
ファセット	1047	773	274	73.8
ユニターム	1146	940	206	82.0

（2） インデックスを付けるのに要する時間を五つの段階に制限した場合，つまりインデックス作成時間の長短による効率への影響

作成時間〔分〕	成功率〔%〕	作成時間〔分〕	成功率〔%〕
2	72.9	12	82.7
4	80.2	16	84.3
8	76.2		

（3） インデックス作成者による効率への影響

A：その分野での作成経験者
B：他の分野での作成経験者
C：未経験者

	C	B	A
UDC	73.8%	81.0%	76.9%
件名標目	79.6	85.4	82.7
ファセット	71.4	77.9	71.1
ユニターム	84.0	83.0	86.0

（4） 検索者の違いによる効率へと影響（この 3 人は全てインデックス作成者）

	a	b	c
UDC	77.2%	75.5%	76.3%
件名標目	80.1	83.8	76.9
ファセット	73.9	71.4	70.1
ユニターム	83.2	78.5	82.4

（5） 検索者の違いによる効率への影響（インデックス作成者 A と研究スタッフ B）

	A	B
UDC	75.6%	79.6%
件名標目	81.5	73.3
ファセット	73.8	66.7
ユニターム	82.0	81.1

（6） 検索に失敗した原因（件数）

	質 問	インデックス作成	探 索	インデックスシステム
UDC	22	108	17	9
件名標目	20	69	25	7
ファセット	26	90	39	10
ユニターム	21	51	10	2

以上 Cranfield Project の報告から二，三のおもな結果を示したが，綜合的にみて，成功率の点からまた検索に失敗した原因の大部分が，インデックス作成における失敗に原因がある点からみて，ユニターム方式が最も有利と判断される．つまり最も単純な方式が最も有効であるという結果となっている．これは自然語という何人にとっても普通なコードが何人にも同等に取り扱えるという利点が，重要であることを示しているといえよう．さらにこのユニターム方式を有効に活用するには，ユニタームのままでの語彙の統制，つまり同義，上位，下位，異称，関連などの関連の把握が重要となってくる．近時情報検索の分野でシソーラスが特に注目されてきたのは，かかる応用面からである．また，このシソーラスの作成およびその更新処理は，これからの電子計算機の応用として注目に値しよう．

インデックスの個数（インデックスの深さという）もまた，検索効率に影響をおよぼす重要な因子である．これに関する実験については，日本科学技術情報センターにおける実験の詳細な報告が発表されている[10]．しかしインデックスの深さは蓄積のコストに直接影響するから，コストの面から最適な深さを決定するという問題もある．

以上述べた実験では，インデックスは個々に独立した単独な概念として扱われている．しかしインデックスを抽出した文献は一連のコンテキストからなっているから，インデックスにもそのコンテキストを表示することが可能である．これによって少なくとも適合率 β を向上させ得ることは確かである．一般にこの表示法として次の三つの記法を用いている．

ロール：概念のそのコンテキスト内におけるカテゴリー

リンク：概念の直接の結合関係

ウェイト：概念のそのコンテキスト内における重要さ

これらを用いることにより，検索効率がどのように影響されるかを，Sinnett[11]の文献により紹介しよう．

これは米国空軍材料研究所で行なわれた実験の報告で，対象とした文献は，各種材料およびその関連分野を含む 6,280 件の文献で，インデックスは 18,000 語からなるシソーラスにより管理された，組み合わせ索引（coordinate index）である．質問は蓄積中の文献をもととして作成された 22 個を用い，それぞれ次の4種類の方法により検索の実験を行なった．

A. リンクもロールも使用しない

B. リンクのみ使用

C. ロールのみ使用

D. リンクとロールを共に使用

使用したロールは 16 種類，各方式の平均検索時間は（磁気テープをファイルとし NCR 304 によるデータ）

A	B	C	D
7	14.3	17.6	34.9分

22 回の検索によって求まった文献数は

	A	B	C	D
全検索文献数	548	425	501	392
関連文献数	361	343	344	324
非関連文献数	187	82	157	68

この表の中の最も大きい関連文献数 361 をもって，全蓄積中の関連文献数（C）と仮定して，α と β を求めると

	A	B	C	D
$\alpha(R/C)$	100%	95%	95.3%	89.8%
$\beta(R/L)$	65.9	80.7	68.7	82.7

となる．さらに各方式について α と β をまとめ

$$\gamma=\alpha-E_2$$

の値を求めると

A	B	C	D
51.7	64.7%	49.4%	57.7%

となる．この結果からみると B, D, A, C の順に効率がよいことがわかるが，検索システムとして考えると，主題分析，コード化，蓄積，探索などのコストを考慮に入れると，B, A, D, C の順とするのが妥当であろう．

以上の結果からみるとロールを用いる方法は効率の向上に役立たないということになる．これはインデックスにロールを割り当てるに当たって，その用法を明確に規定することが困難であり，したがって主題分析の時点で大きな誤差が入り込むためと思われる[12]．

以上二つの実験結果から一般的にいえることは

（1） 検索率 α はインデックスの意味的正確さ，インデックスの深さ，自然語をも含めてのコードの統制にかかっており，これは実務的にはインデックス作成

手順の明文化の問題である．

（2） 適合率βはαを低下させるのと同じ原因によって低下するが，コンテキスト情報などの追加情報を導入することにより，ある程度向上させることができる．

3.3 検索効率の向上法

先に述べたように，文献検索はあくまでも確率モデルであるから，質問のインデックスと蓄積文献のインデックスとの論理的マッチングによって事足れりとするのは疑問のあるところである．この点を指摘して確率検索法の重要性を強調したのは Maron[13] である．彼によれば「検索とは質問と文献との関連度を求め，関連度の順に並べた文献のリストを求めることである」ということになる．これによりαとβをともに向上させ得ることを強調している．

いま質問Rから文献Cへの変換を

$$f(R)=C$$

とすると，もし質問Rと意味的に同等な質問R'が存在するとすれば，$f(R')$もまた関連文献であるはずである．このRからR'への変換には，一般に意味の母集団における関連パターンであるシソーラスを用いるが，シソーラスの代用として，蓄積文献内でのパターンを用いることもできる．後者の方が普通そのメンテナンスが楽である．さらにCとその情報内容が同等である文献C'が存在するとすれば，これもまた関連文献であるはずである．このCからC'への変換には，一般に書誌的関連パターン（bibliographic coupling）を用いる．これは書誌事項である著者，雑誌，引用文献などによって求めたパターンである．

上述のfの例として線型変換を用いると，以上の関係をさらにはっきり示すことができる[14]．

$$C=\Phi\theta\Omega R$$

ここにΦ：文献関連行列
（文献の著者，研究グループ，引用文献などをもとした関連行列）

θ：文献—インデックス行列
（インデックスが文献内で占めるウェイトの行列）

Ω：インデックス関連行列
（インデックスの文献内における同時出現をもとした関連行列）

なお質問と文献の論理的マッチングも，上式の特殊な場合であることは明らかであろう．インデックスおよび文献の関連をn次関連まで拡張すれば，さらに一般化されよう．

以上の趣旨に基づいて Stiles[15] が行なった実験の結果を示そう．これはインデックスの蓄積内での関連パターンを用いた実験である．対象とした文献は米国防省の 10 万件以上の文献である．まず蓄積文献中において用いられているインデックス間の相関を求めておく．相関の測度としては，2×2 分割表における Yates の補正を行なった χ^2 の値の対数を用い，この値が 1 以上を示すインデックスの組み合わせを関連語とし，あるインデックスからみて関連度が 1 以上の値をもつインデックスの集合を，そのインデックスのプロフィルという．たとえば Friction のプロフィルの一部は，次のようになっている．

インデックス	関連度
Wear	3.35
Thin	3.21
Lubrication	3.00
Belt	2.70

質問が与えられたら，そのインデックスのそれぞれのプロフィルを求め，全てのプロフィル中か，または前もって定めた数のプロフィル中に現われるインデックスを求め，これを 1 次派生語という．質問インデックスと 1 次派生語を一緒にして上述の処理を繰り返えし，新たに求まったインデックスを 2 次派生語という．関連パターンの 2 次結合まで用いた理由は，同義語を求めるためである．これらを全て集めて質問のインデックスとする．各インデックスが他のインデックスに対して有する関連度の和をもって，それぞれのウェイトとする．

検索においては，少なくとも 1 個の質問インデックスの存在する文献は抽出される．各文献についてマッチしたインデックスのウェイトの和を求め，これをその文献の質問に対するウェイトとする．このウェイトの順に文献を配列して回答とする．Thin, Film を質問インデックスとして検索した結果の一部を次に示す．

文献のウェイト	判定	Thin, Film の有無	文献のウェイト	判定	Thin, Film の有無
24.32	Yes	Thin, Film	10.05	Yes	
24.22	〃	〃 〃	9.83	〃	Film
24.22	〃	〃 〃	9.66	M	
24.22	〃	〃 〃	9.50	〃	
22.47	〃	〃 〃	9.38	〃	
19.87	〃	〃	9.38	No	
15.59	〃	〃	9.30	P	
15.30	P		9.30	〃	
14.83	Yes		9.30	〃	
12.54	No		9.30	〃	
11.81	M		9.18	〃	
11.20	〃		9.12	〃	
10.72	〃		8.66	M	
10.14	P		8.03	P	
10.08	M		7.95	M	

す.

判定欄の記号は

Yes: Thin Film の情報を含む

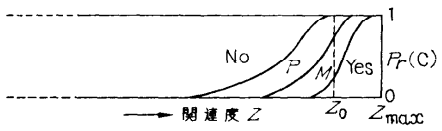
M: 関連情報として有用

P: 多分関連情報を含むであろう

No: Thin Film の情報を含まない

この結果から，Thin, Film を含まない文献から関連文献を見出したことにより α を増大させ，また Yes の文献を上位に集中させたことにより，β を増大させたということができよう.

この結果を概念的に図に示すと**第7図**のようになる．これはある質問に対して，ファイルの蓄積文献集



第7図　蓄積文献集合の分割

合を整理して配列した結果を示しているが，始めにも述べたように確率モデルであるため，ファイルを関連文献集合と非関連文献集合とに，はっきり区分する形とはなっていない．したがって Z_0 をいかなる値にとるかは効率の問題ということができよう．さらにこの曲線は質問者によって異なるという大きな問題がある．これは個々に異なる情報要求が，質問という形の表現において同一となるためである.

以上主として個々の検索の効率を論じたが，この他に検索システム全体としての，コストを導入した効率の問題があることを付記しておく.

なお，本稿の執筆分担は以下の通りである.

1. 内容分析と主題分析　　菊池敏典
2. コード化　　笹森勝之助
3. 検索効率　　高橋達郎

参考文献

おもな文献だけをあげておく．網羅的な参考文献リストとしては，たとえば以下の Stevens の報告を参照されたい.

Stevens, M.E.: Automatic Indexing: A State-of-the-Art Report. NBS Monograph 91, National Bureau of Standards, Wash. D.C., 1965, 220 pp

1) Luhn, H.P.: The Automatic Creation of Literature Abstracts. IBM J. Res. Develop. **2** (2) 159-165 (1958)
2) Dale, A.G. and Dale, N.: Some Clumping Experiments for Information Retrieval. LRC-64-WPI 1, Linguistics Research Center, University of Texas, Austin, Texas. 1964, 13 pp.
3) Doyle, L.B.: Expanding the Editing Function in Language Data Processing. Comm. ACM. **8** (4) 238-243 (1965)
4) Hyslop, M.R.: Sharing Vocabulary Control. Special Libraries, **56** (10) 708-714 (1965)
5) 鈴木幸雄：外務省におけるシソーラス編成．第2回ドキュメンテーション研究集会発表論文集，255-258 (1965)
6) Giuliano, V.E. et al.: Linear Associative Information Retrieval. Howerton, P. ed., "Vistas in Information Handling", Spartan Books, Wash. D.C., 30-54 (1963)
7) Abraham, C.T.: Techniques for Thesaurus Organization and Evaluation. Proceedings of the American Documentation Institute, 1964 Annual Meeting, 485-497 (1964)
8) Salton, G.: Manipulation of Trees in Information Retrieval. Comm. ACM. **5** (2) 103-114 (1962)
9) Cleverdon, C.W.: Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems, College of Aeronautics, Cranfield, England, 1962. 305 pp
10) 安倍浩二　ほか：金属工学文献の機械検索実験報告　V　検索効率におよぼす索引深さの影響の調査．第2回ドキュメンテーション研究集会発表論文集，259-266 (1965)
11) Sinnett, J.D.: An Evaluation of Links and Roles Used in Information Retrieval, ML TDR 64-152, AF Materials Laboratory, Wright-Patterson AFB, Ohio 1964 139 pp
12) Hyslop, M.R.: Role Indicators and their Use in Information Searching-Relationship of ASM & EJC Systems. Proceedings of the American Documentation Institute, 1964 Annual Meeting, 99-107 (1964)
13) Maron, M.E., Kuhns, J.L.: On Relevance, Probabilistic Indexing and Information Retrieval. **7** (3) 216-244, (1960)
14) Giuliano, V.E. et al.: Loc. cit.
15) Stiles, H.E.: The Association Factor in Information Retrieval. **8** (2) 271-279 (1961)

(昭和41年9月13日受付)