Library and Information Science No. 20 1982

キーワード群を対象とした自動語彙変換システム

An Automatic Vocabulary Switching System Converting from Keyword Phrases Assigned by Searchers to Descriptors

細野公男・後藤智範
Kimio Hosono *Tomonori Gotoh*
諸橋正幸・大河内正明
Masayuki Morohashi *Masaaki Ohkohchi*

Résumé

Many searchers of online information retrieval systems are unfamiliar with the structure and content of thesauri and they usually have the problem of selecting adequate descriptors for their search. Therefore, the system which converts a set of searcher-assigned keywords to the corresponding descriptors will be quite useful and helpful for them.

This paper describes the result of an experiment on developing an automatic vocabulary switching system. It is designed to switch those keywords included in the field of "System & Control Theory", "Control Technology", "Computer Hardware", "Computer Software", and "Computer Application". Switching is carried out by using a special conversion table.

1506 descriptors as well as 659 nondescriptors were extracted from the INSPEC thesaurus tape for this experiment. Since each descriptor has broader, narrower, related, and/or used-for terms, information about these relations was used to compile the conversion table. In addition, 1531 single words were extracted from the descriptors and nondescriptors to produce ingredients of the searcher-assigned keywords. This mapping relation between single words and descriptors was also kept in the same table. Thus, the table includes four kinds of relations among descriptors as well as whole-part relation between words and descriptors. Altogether it shows the extent of relatedness of individual single words to the descriptors included in the INSPEC thesaurus.

Input keywords are decomposed into single words, and they are collated with the words in

細野公男：慶應義塾大学文学部図書館・情報学科教授
Kimio Hosono, Professor, School of Library and Information Science, Keio University.
後藤智範：慶應義塾大学大学院文学研究科図書館・情報学専攻博士課程
Tomonori Gotoh, Doctor course, School of Library and Information Science, Keis University.
諸橋正幸：日本アイ・ビー・エム株式会社サイエンス・インスティチュート主任研究員
Masayuki Morohashi, Researcher, Science Institute, IBM Japan, Ltd.
大河内正明：日本アイ・ビー・エム株式会社サイエンス・インスティチュート主任研究員
Masaaki Ohkohchi, Researcher, Science Institute, IBM Japan, Ltd.

the table. Then the relation vector of the matched words are taken from the table and summed up. Finally, each element of the resultant vector, namely descriptor, is sorted by descending order of the corresponding values and those descriptors of which order is within the range specified by the searcher are outputted.

Actual switching was done successively and the performance of the system seems to be promising although there are still several problems to be solved.

I. はじめに

1970年代前半に新しい情報産業として定着した，データベース・サービスすなわちオンラインによる商用の情報検索サービスは，その後も順調に発展の一途を辿っている。提供される情報のタイプは文献情報だけでなく，新聞記事，判例，経済・経営分野を対象とした数値情報，さらには化学構造式のような画像情報など多岐にわたっており，また，データベースの数，データベース・サービスに関与する機関数は第1表に示すように共に依然として増大している。第1表は主としてヨーロッパで利用可能なものが中心であるが，少なくともこの分野の着実な発展ぶりをうかがい知ることができよう。なお，データベース・サービス関連機関には，データベース作成機関，ベンダー，ネットワーク，情報ブローカーが含まれる。

一方，データベース・サービスを利用者の観点からみると，現在でもその円滑な利用を妨げるいくつかの障壁

第1表　データベース数およびデータベース・サービス関連機関数の推移[1)]

		1975	1976	1977	1978	1979	1980	1983
データベース	書誌	335	337	422	533	565	654	762
	ファクト	51	149	268	568	715	755	1083
	合計	386	486	690	1101	1280	1409	1845
データベース・サービス関連機関		284	365	510	777	859	903	983

がある。例えば，ベンダー間でのシステム利用方法の違い，データベースの種類や構造の多様性，索引言語の違い，さらに言語の障壁などが挙げられよう。その結果，ベンダー・システム，データベースの構造に精通していない利用者は，効果的な情報サービスを受けることに多かれ少なかれ困難をきたしていることは事実である。これは，利用者の〝使い勝手〟が良いシステムの開発が進んでも依然として残る問題であろう。

現在は，コンピュータ使用コスト，データ蓄積コストの著しい低下と情報検索技術の進歩にともない簡便なアクセス・キー作成が行なわれるようになり，たとえば標題，抄録中の任意のキーワードから文献を探すことが可能になっている。また，全文データベースの数が増大するにつれ，非統制キーワードの使用がますます盛んになると思われる。しかし，これらの非統制のキーワードの使用は，重要な適合文献の検索もれと，一方では著しい検索ノイズ（不必要な文献の検索）とをもたらしている。また，包括的な検索には不向きである。したがって，シソーラスで管理されたディスクリプター(統制語)をアクセス・キーとして使用する意義は依然として極めて高いといえる。

ところで，シソーラスを使用する場合には，適切な用語を選択することが極めて重要であるが，その際困難を感じることがしばしばある。それは，主題知識を必要とするだけでなく，シソーラスの構造つまり用語間の関係を熟知する必要があるからである。さらに専門用語を充分に知っていなければならない。これは，データベースの言語が母国語でない場合，特に利用者に大きな負担を課すことになり，検索者が言語の障壁を強く感じる点である。

シソーラスを使用するアプローチの不便宜さを解消する1つの手段として，検索者の思いつくことばをシソーラス中のディスクリプターに自動的に変換することが考えられる。そのような自動語彙変換システムが存在すれば，検索者はあたかも非統制のキーワードを使用するがごとく統制キーワードを活用できることになる。

本稿はこの点に鑑みて，自動語彙変換システムの試案を提示するものである。なお，検索者が検索時に思いうかべることばをキーワード，シソーラス中の用語をディスクリプターとして，両者を区別して使用する。また，キーワード，ディスクリプターのいずれにも単語形態だけでなく，フレーズ形態のものが含まれるものとする。

II. 自動語彙変換システムの例

先に述べたように，検索者は必ずしも情報の蓄積・検索プロセスやシソーラスの構造に熟知しているわけではないので，適切なディスクリプターを常に選択することはなかなかむずかしい。このため，検索者の負担を軽減する目的でさまざまな手段・方法が開発されてきた。自動語彙変換システムは，検索シソーラス[2]や索引語のオンライン表示と共にその代表的な例である。

自動語彙変換システムの利点は，要約すると次のようになろう。

(1) データベースのより効果的，効率的な利用が可能となる。

(2) データベースの内容，構造をあまり良く知らなくても，妥当な検索語を探すことができる。

(3) 複数のデータベースにまたがる検索が容易になる。

現在までに開発されている語彙変換システムには，さまざまなタイプがある。1つは検索者の思いつくキーワードを特定のデータベースで使用されているディスクリプターに変換するものである。また，複数のシソーラス間でのディスクリプターの変換・対応付けを目的とするシステムもある。本稿では第1のタイプのみを考える。

A. Field の方法[3]

INSPEC (Information Services in Physics, Electrotechnology, Computers and Control) データベースの作成機関である Institution of Electrical Engineers の Field は以下に示すような自動変換の方法を開発した。

入力されたキーワードはまず語尾処理がなされ，それがディスクリプター辞書と照合される。たとえば electromagnetic wave propagation は，electromagnet wav propag のように処理される。辞書と一致すればプロセスは次のキーワードの処理に移ることになるが，一致しない場合には，以下の処置を行う。

（a） キーワードが単語の場合は変換作業を終了する。

（b） キーワードがn個の単語から構成されるフレーズの場合には，それから（n-1）語で構成されるフレーズを作成する。

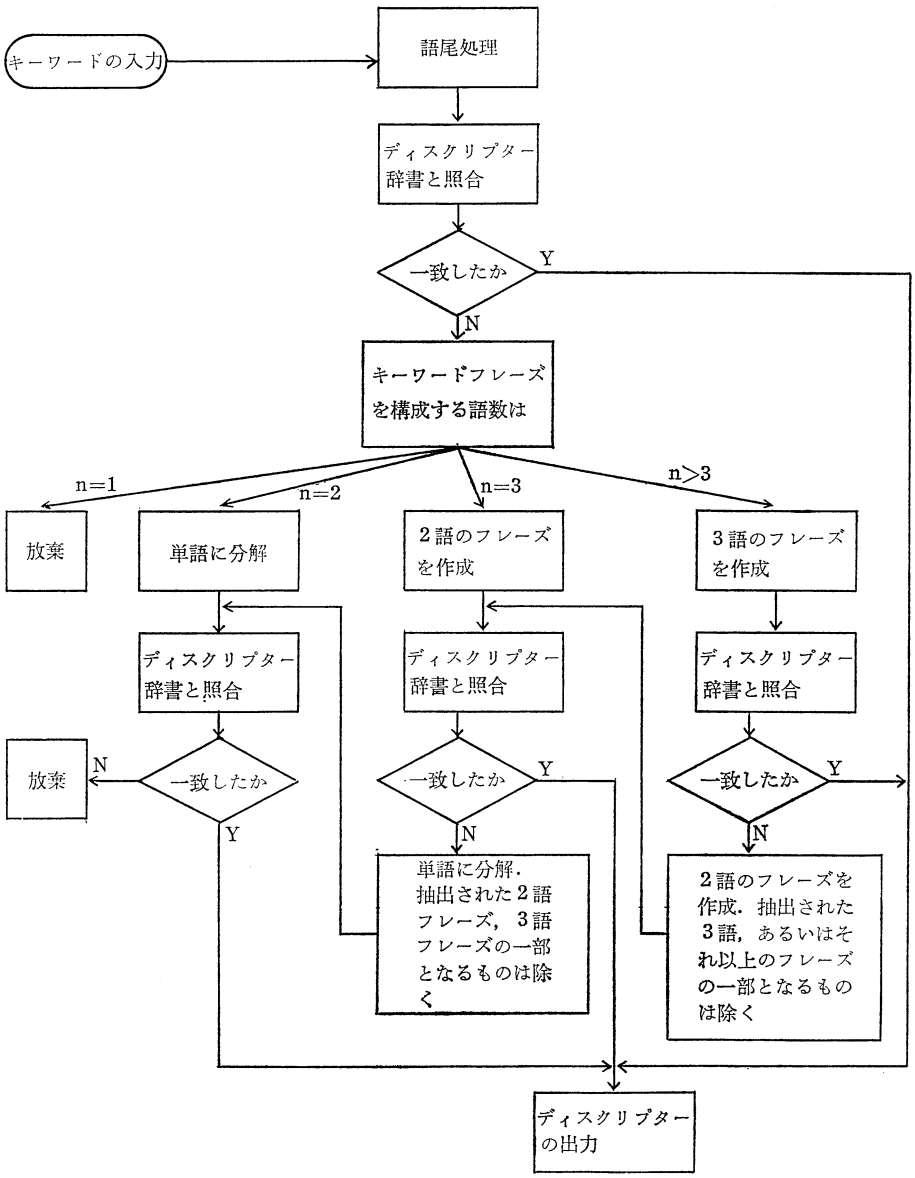
bの場合は，得られたフレーズをそれぞれ辞書と照合して，一致しないフレーズはaあるいはbの処置を再び行

うことになる。ただしその際，すでに一致したフレーズの構成要素となるフレーズは処理対象とはならない。たとえば，辞書に一致しないフレーズ $W_1W_2W_3W_4$ からは3語からなる4個のフレーズが作成される。もしそのうち $W_1W_2W_3$ のみが辞書と一致すれば残りの3個のフレーズに関して2語から構成されるフレーズが次に作成されるが，その際 W_1W_2，W_1W_3，W_2W_3 は除外されるのである。第1図は Field のアルゴリズムの概略である。

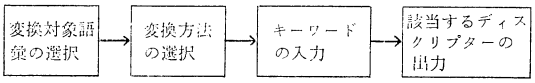
B. Battlle のシステム[4),5)]

Battlle 社は複数のデータベースを対象とした自動語彙変換システムを開発しているが，そのシステムでは，第2図のようなプロセスをとっている。

変換対象語彙の選択は，入力するキーワードに対応す



第1図　Field の自動変換アルゴリズム



第2図 Battlle システムにおける変換プロセス

るディスクリプターを必要とする語彙を指定する操作である。このシステムでは語彙として，*DOE Thesaurus*, CA Concept Edit File, *Subject Headings for Engineering* (*SHE*, Engineering Index), *INSPEC Thesaurus*, *NASA Thesaurus*, CA Keyword Frequency が用意されている。なお，エネルギー関係のことばだけを対象としたシステムもあり，そこでは上述の *SHE*, *NASA* のほか，*API Thesaurus*, CA Subject Headings, *DDC Thesaurus*, *IGT Thesaurus*, *INIS Thesaurus*, *Exploration and Production Thesaurus*, *GeoRef Index*, *Thesaurus of Engineering and Scientific Terms* が使用できる。したがって，この中から変換対象語彙を1つ以上選択することになる。

変換方法には18種類あり，そのうちから適当と思われるものを選択することになる。変換方法は，完全一致，同義語処理，語幹処理，上位・下位・関連語処理，隣接語処理，およびこれらの組合せが可能である。なお，入力されるキーワードが句形態である場合は，その構成要素である各単語に対して，単語単位の変換，語幹処理，隣接語処理が行われる。同義語・上位語・下位語・関連語の処理は，入力されたキーワードの同義語 (USE 参照)，上位語，下位語，関連語をそれぞれ出力する。隣接語処理は入力されたキーワードとアルファベット配列上隣接する1つ以上のディスクリプターを出力する。語幹処理はいわゆる前方一致処理である。

該当するディスクリプターは，第3図のような形式で

```
ENTER SEARCH TERM OR COMMAND
?ENGLAND

CODE    VOCABULARIES AND TERM
  5     B,N,R,  ENGLAND
  5     G,  GREAT BRITAIN AND IRELAND
  5     B,I,N,  UNITED KINGDOM
  5     D,N,R,  GREAT BRITAIN
  7     I,  UNITED KINGDOM ORGANIZATIONS
```

注：コードは変換方法を示す（5は完全一致と同義語処理，7は単語単位の処理）.

第3図 キーワードイングランドに対応するディスクリプターの出力[5]

出力される。選択対象となった語彙はコードで表わされている。例えば，B は *API Thesaurus*, N は *NASA Thesaurus*, R は *GeoRef Index* である。なお，このような情報を提供するものとして，*Integrated Energy Vocabulary* がある。[6]

C. Field および Battlle の方法の特徴

Field の方法ではまずシソーラスで示されているディスクリプター間の関係を表わす情報が使用されておらず，キーワードからディスクリプターへの変換が単なる用語の綴だけにもとづいている点が問題となろう。この点に関しては，Battlle の方法は同義語・上位語・下位語・関連語処理が使用できるので，シソーラスの構造が持つ情報を生かしているといえよう。なお，Field の語尾処理，Battlle の語幹処理や隣接語処理は，例えば abstract に対して abstracts, abstracting, abstracter などキーワードと概念が類縁関係にありしかも語の配列上近接するディスクリプターを得ることができる利点がある。

一般に検索では複数のキーワードで1つの概念を表わすことが多いのであるから，与えられたキーワード群全体を対象とした総合的な変換が望ましい場合が多い。しかし上述の2つの方法は，いずれも個々のキーワードを個別的に処理する形をとっている。

III. キーワード群を対象とした自動語彙変換システムの構造

A. 変換アプローチ

個々のキーワードとディスクリプターとの間に前もって何らかの関係を持たせておくことができれば，その関係を利用してキーワードからディスクリプターへの自動変換が可能になる。例えば，第2表のような数値で表現される関係を考えることができよう。

第2表で与えられている個々の値は，各キーワードとディスクリプターとの関連の度合を示し，値が大きい程

第2表 キーワードとディスクリプターの関連

	D_1	D_2	D_3	D_4	D_5
K_1	0	3	1	4	1
K_2	1	0	7	2	5
K_3	1	5	0	2	6

K：キーワード D：ディスクリプター

その関連が強く，0は両者間に何らの関係もないことを示すとする。したがって，キーワード K_3 はディスクリプター D_5 に変換されることになる。

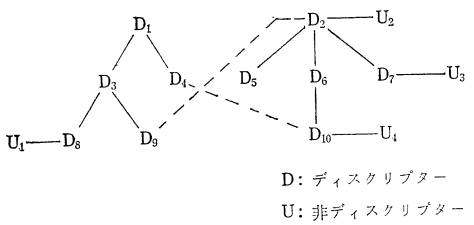
通常検索者は検索語として複数のキーワードを使用する。その際，もし K_1 と K_3 を用いた場合どのディスクリプターに変換するのが妥当であろうか。キーワード毎の対応関係のみで処理すれば，それぞれ D_4 と D_6 に変換されることになる。しかし，一般に検索者が想定するキーワード間には概念的な関係が存在し，相互に関連しあう場合の多いことを考慮すれば，各キーワードの持つ値を加算して，最も高い値を持つディスクリプターにまず変換すべきであろう。したがって，第2表の場合では，K_1, K_3 のそれぞれの値 (0, 3, 1, 4, 1) と (1, 5, 0, 2, 6) を加算すると (1, 8, 1, 6, 7) となるので，D_2 が変換すべき第1のディスクリプターとなる。

本稿で提案する変換システムは，キーワードを個別的に変換するのではなく，キーワード群全体を対象として変換を行う。変換プロセスは，まずキーワード群を単語に分解し，その単語群を上述のような変換表と参照して該当するディスクリプター群を出力する形をとっている。したっがって，変換表は，単語とディスクリプターとの対応表である。なお，シソーラスを構成するディスクリプターは，それぞれ必要に応じて上位語 (BT)，下位語 (NT)，関連語 (RT)，同義語 (UF) を持っているが，変換表にはこの情報が生かされている。

B. 変換表の作成プロセス

語彙変換に使用される変換表はキーワードに含まれることが予想される単語とディスクリプターとの対応関係を示すものである。これは，ディスクリプターとキーワードに含まれる単語に共通性が高いこと，この種の単語はディスクリプターおよびキーワードが表現する重要な概念の構成概念を示すとの前提に基づくものである。

さて，ディスクリプター間に第4図のような関係があ



第4図　ディスクリプター間の関係図

第3表　ディスクリプター間の関係

ディスクリプター	NT	BT	RT	UF
D_1	D_3, D_4, D_8, D_9			
D_2	D_5, D_6, D_7, D_{10}		D_9	U_2
D_3	D_8, D_9	D_1		
D_4		D_1	D_{10}	
D_5		D_2		
D_6	D_{10}	D_2		
D_7		D_2		U_3
D_8		D_3 D_1		U_1
D_9		D_3 D_1	D_2	
D_{10}		D_6 D_2	D_4	U_4

ると仮定する。これを表で示したのが第3表である。また，第4図のディスクリプター群は，第5図で示されるように総計20の異なり語を含んでいるとする。このようなデータから仮想的な変換表の作成プロセスを以下に示す。

まず，ディスクリプターとその関係語との関連を表わす表を作成する。上位語を例にとると第4表が得られる。第4表では D_1 が D_3 と D_4 の，D_2 が D_5, D_6, D_7 の上位語であることを数値1で示している。次にこの関係表と第5図を用いてBT対応表を作成する。第5表はそれぞれの単語を構成要素の一部とするディスクリプターを上位語として持つディスクリプターを示したものである。例えば，D_3 は単語 W_1 を含むディスクリプター (つまり D_1) を上位語として持つ。

同様な方法でNT, RT, UF, および自分自身 (DE) の対応表を作成する。変換表はこれらの表を合成し，各数

第4表　B T 関 係 表

		ディスクリプター									
		D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}
上位語	D_1	0	0	1	1	0	0	0	1	1	0
	D_2	0	0	0	0	1	1	1	0	0	1
	D_3	0	0	0	0	0	0	0	1	1	0
	D_4	0	0	0	0	0	0	0	0	0	0
	D_5	0	0	0	0	0	0	0	0	0	0
	D_6	0	0	0	0	0	0	0	0	0	1
	D_7	0	0	0	0	0	0	0	0	0	0
	D_8	0	0	0	0	0	0	0	0	0	0
	D_9	0	0	0	0	0	0	0	0	0	0
	D_{10}	0	0	0	0	0	0	0	0	0	0

$D_1 = W_1W_2$　　$D_6 = W_3W_8$
$D_2 = W_3W_4$　　$D_7 = W_3W_9$
$D_3 = W_5W_1$　　$D_8 = W_5W_{10}W_1$
$D_4 = W_6W_1$　　$D_9 = W_5W_{11}W_1$
$D_5 = W_3W_7$　　$D_{10} = W_8W_{12}W_3$
$U_1 = W_{13}W_5W_1$　　$U_3 = W_{15}W_3$
$U_2 = W_{14}W_4$　　$U_4 = W_{12}W_8W_3$

D：ディスクリプター，U：非ディスクリプター，
W：単語

第5図　ディスクリプターを構成する単語

第5表　B T 対 応 表

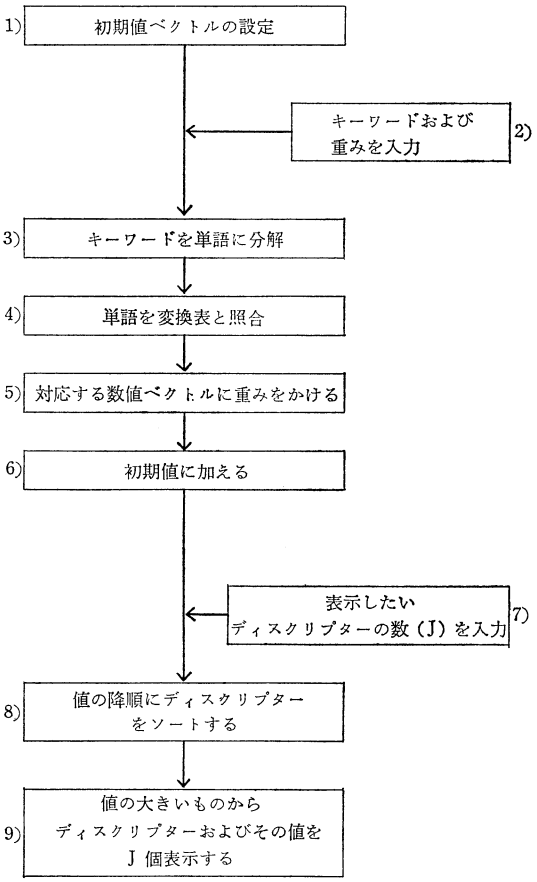
		ディスクリプター									
		D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}
単語	W_1	0	0	1	1	0	0	0	2	2	0
	W_2	0	0	1	1	0	0	0	1	1	0
	W_3	0	0	0	0	1	1	1	0	0	2
	W_4	0	0	0	0	1	1	1	0	0	1
	W_5	0	0	0	0	0	0	0	1	1	0
	W_6	0	0	0	0	0	0	0	0	0	0
	W_7	0	0	0	0	0	0	0	0	0	0
	W_8	0	0	0	0	0	0	0	0	0	1
	W_9	0	0	0	0	0	0	0	0	0	0
	W_{10}	0	0	0	0	0	0	0	0	0	0
	W_{11}	0	0	0	0	0	0	0	0	0	0
	W_{12}	0	0	0	0	0	0	0	0	0	0
	W_{13}	0	0	0	0	0	0	0	0	0	0
	W_{14}	0	0	0	0	0	0	0	0	0	0
	W_{15}	0	0	0	0	0	0	0	0	0	0

値を加算して求めたものである。なお，単純に加算すると得られる値はディスクリプター中に出現する単語の出現頻度そのものとなるので，DE表の値を重視するなど各対応表に適当な重みをつけて加算する必要がある。

C. 変換プロセス

検索者が入力したキーワード群のディスクリプター群への変換は第6図のようになる。

1) キーワードの入力に先だって初期値ベクトルを0にする。
2) キーワードとその重みを検索者が入力する。
3) キーワードを単語に分解する。
4) 各単語を変換表と照合し，対応する数値ベクトルを取り出す。
5) この数値ベクトルに入力された重みをかける。
6) この重みつきベクトルを初期値ベクトルに加算する。



第6図　変換プロセス

7) 出力したいディスクリプターの個数(30個以内)を入力する。
8) 最終的に得られた数値ベクトルの要素の降順にディスクリプターをソートする。
9) 指定された個数だけディスクリプター名とその要素値を出力する。

3）から6）の処理は，1つのキーワードに含まれる単語の数だけ繰り返して行われ，それが終了すると2）にもどり，次のキーワードについて3）から6）の処理が同様になされる。この過程はすべてのキーワードの入力が終了するまで続けられる。

IV. 自動語彙変換システムの作成

A. 対象分野

本システムではコンピュータ・サイエンス分野のキー

CD	CASCADE CONTROL
CD	CATALOGUING
CD	CATASTROPHE THEORY
CD	CATHODE-RAY TUBE DISPLAYS
CD	CELLULAR ARRAYS
CD	CEMENT INDUSTRY
CD	CERAMIC INDUSTRY
CD	CHARACTER RECOGNITION
CD	CHARACTER RECOGNITION EQUIPMENT
CD	CHARACTER SETS
CD	CHEBYSHEV APPROXIMATION
CD	CHEMICAL ENGINEERING COMPUTING
CD	CHEMICAL INDUSTRY
CD	CHEMICAL TECHNOLOGY
CD	CHEMISTRY
CD	CHEMISTRY COMPUTING
CD	CHERENKOV COUNTERS
CD	CHROMATOGRAPHY
CD	CIRCUIT ANALYSIS COMPUTING
CD	CIRCUIT CAD
CD	CIRCUIT LAYOUT CAD
CD	CIVIL ENGINEERING
CD	CIVIL ENGINEERING COMPUTING
CD	CLASSIFICATION
CD	CLIMATOLOGY
CD	CLOCKS
CD	CLOSED LOOP SYSTEMS
CD	CLOUD CHAMBERS

第7図　ディスクリプター・ファイル

ワードの変換を考える。このため INSPEC データベースで使用されているシソーラスを用いた。INSPEC は物理学，電気工学，電子工学，コンピュータ，制御工学の各分野を網羅する最大級のデータベースであるので，INSPEC シソーラスには，これらの分野に属する文献の検索に適したディスクリプターが豊富に収録されている。

使用するデータは，INSPEC が提供する Journal Classification & Thesaurus ファイルのシソーラス・レコード（ディスクリプター・データ）部分である。このレコードには5桁あるいは6桁の分類コードが付与されているが，そのうち，C1000 (System & Control Theory), C 3000 (Control Technology), C 5000 (Computer Hardware), C 6000 (Computer Software), C7000 (Computer Application) に属するものだけを対象にする。なお，シソーラス・レコードに含まれる項目として，見出し語（タグ420），非ディスクリプター（422），下位語（430），上位語（440），関連語（450），最上位語（460），分類コード(470) があるが，ここでは420，422，430，440，450を用いている。

90	AUDIO
91	AUTOMATA
92	AUTOMATIC
93	AUTOMATION
94	AUTOMOBILE
95	AUTOMOBILES
96	AVIONIC
97	BALLOONS
98	BAND
99	BANDWIDTH
100	BANG-BANG
101	BANK
102	BANKING
103	BANKS
104	BASES
105	BASIC
106	BATCH
107	BAYES
108	BEAM
109	BEAM-FOIL
110	BEAMS
111	BEAT
112	BEHAVIORAL
113	BEHAVIOURAL
114	BELTS
115	BESSEL
116	BETA
117	BETA-RAY

第8図　単語リスト

B. 変換表

1. ディスクリプター・ファイルの作成

上述のディスクリプターのうち，タグ420のものをコア・ディスクリプター（識別コードCD）とし，その他のものをマージナル・ディスクリプター（MD）とする。このファイルの作成過程は次のようになる。

1) シソーラス・テープから上述のカテゴリーに該当するレコードを抽出しシソーラス・ファイルを作成する。

2) シソーラス・ファイル中の各レコードをタグ毎に分割し，タグ420のディスクリプターにはCD，それ以外のものにはMDの識別コードを付与する。

3) 複数の識別コードを持つディスクリプターはCDコードのもののみを残し，それ以外は削除する。

NO. 26 WORD : AGRICULTURE

DESCRIPTOR NO.	RELATED DESCRIPTORS	COUNT
278	FARMING	1

NO. 27 WORD : AIDED

DESCRIPTOR NO.	RELATED DESCRIPTORS	COUNT

NO. 28 WORD : AIDS

DESCRIPTOR NO.	RELATED DESCRIPTORS	COUNT
657	SENSORY AIDS	1

NO. 29 WORD : AIR

DESCRIPTOR NO.	RELATED DESCRIPTORS	COUNT
11	AEROSPACE COMPUTER CONTROL	2
573	POLLUTION	1
574	POLLUTION DETECTION AND CONTROL	1
745	TRAFFIC	1
746	TRAFFIC COMPUTER CONTROL	2

NO. 30 WORD : AIR-TRAFFIC

DESCRIPTOR NO.	RELATED DESCRIPTORS	COUNT

NO. 31 WORD : AIRCRAFT

DESCRIPTOR NO.	RELATED DESCRIPTORS	COUNT
20	ALGOL	1
23	ALGORITHMIC LANGUAGES	1
300	FORMAL LANGUAGES	1
337	HIGH LEVEL LANGUAGES	1

第9図 B T 対応表

4) 識別コード順およびアルファベット順にディスクリプターをソートする。

第7図はこのような過程を経て作成されたディスクリプター・ファイルの一部である。本システムでは合計1506のディスクリプターが用いられている。その内訳はコア・ディスクリプターが 775, マージナル・ディスク

NO. 601 WORD : GRAMMARS

DESCRIPTOR NO.	RELATED DESCRIPTORS	COUNT
164	CONTEXT-FREE GRAMMARS	5
165	CONTEXT-FREE LANGUAGES	1
166	CONTEXT-SENSITIVE GRAMMARS	5
167	CONTEXT-SENSITIVE LANGUAGES	1
300	FORMAL LANGUAGES	1
327	GRAMMARS	6
764	VIDEO AND AUDIO DISCS	1

NO. 602 WORD : GRAMOPHONES

DESCRIPTOR NO.	RELATED DESCRIPTORS	COUNT
72	BODE DIAGRAMS	1
123	COMBINATORIAL MATHEMATICS	1
239	DIRECTED GRAPHS	1
328	GRAPH COLOURING	1
329	GRAPH THEORY	2
330	GRAPHS	1
362	INFORMATION THEORY	1
743	TOPOLOGY	1
753	TREES (MATHEMATICS)	1
764	VIDEO AND AUDIO DISCS	1

NO. 603 WORD : GRAPH

DESCRIPTOR NO.	RELATED DESCRIPTORS	COUNT
123	COMBINATORIAL MATHEMATICS	1
239	DIRECTED GRAPHS	1
328	GRAPH COLOURING	5
329	GRAPH THEORY	5
330	GRAPHS	1
362	INFORMATION THEORY	1
743	TOPOLOGY	1
753	TREES (MATHEMATICS)	1

NO. 604 WORD : GRAPHIC

DESCRIPTOR NO.	RELATED DESCRIPTORS	COUNT
140	COMPUTER GRAPHIC EQUIPMENT	7
141	COMPUTER GRAPHICS	1
146	COMPUTER PERIPHERAL EQUIPMENT	1
377	INTERACTIVE TERMINALS	1
395	LIGHT PENS	1
630	REMOTE CONSOLES	1

第10図 変 換 表

リプターが731である。上位語，下位語は必ず見出し語ともなるので，最終的に得られたマージナル・ディスクリプターは，関連語のうち見出し語とはならないディスクリプター，つまり本システムの対象外にある分野のディスクリプターである。

2. 単語リストの作成

1506のディスクリプターおよび659の非ディスクリプターを単語に分割し，単語リストを作成した。第8図は単語リストの一部である。抽出された単語の総数は1531である。なお，この中には特殊記号や数字で始まる単語は含まれていない。

3. 対応表の作成

前章で示した手続きにしたがって，DE，NT，BT，RT，UFの各対応表を作成した。第9図はBT対応表の一部である。対応表は1531×775の行列となるため，第9図では各単語について値が1以上のディスクリプターのみ出力されている（第5表参照）。

4. 変換表の作成

変換表は対応表の重みつき合成である。本システムでは，DE表，NT表，BT表，RT表，UF表にそれぞれ4，1，1，1，3の重みを与えて合成した。第10図は変換表の一部である。

C. 変換結果

第10図の変換表を用いて第11図の前半に示したキーワード群をディスクリプターに変換したのが，第11図の後

```
EXECUTION BEGINS...
===> ENTER KEYWORD(PHRASE) & WEIGHT(1 TO 9)
information retrieval 1
===> ENTER
document retrieval 2
===> ENTER
index 5
===> ENTER
indexing 5
===> ENTER

KEYWORD INPUT END
===> ENTER NUMBER OF WORDS TO BE OUTPUTTED(1 TO 30)
20
DESCRIPTOR                                  WEIGHT

VOCABULARY                                      46
INFORMATION RETRIEVAL                           46
INFORMATION RETRIEVAL SYSTEMS                   24
INFORMATION SCIENCE                             21
INDEXING                                        21
PERFORMANCE INDEX                               20
INFORMATION RETRIEVAL SYSTEM EVALUATION         20
INFORMATION ANALYSIS                            13
INFORMATION CENTRES                             15
INFORMATION SERVICES                            12
INFORMATION DISSEMINATION                       11
INFORMATION STORAGE                             10
MANAGEMENT INFORMATION SYSTEMS                   7
PHOTOCOPYING                                     6
INFORMATION USE                                  6
THESAURI                                         5
SENSITIVITY ANALYSIS                             5
MAGNETIC TAPES                                   5
INFORMATION THEORY                               5
DATA ACQUISITION                                 5
```

第11図 変 換 結 果

半部分である。

V. 考　察

本システムでは検索者が入力する個々のキーワードに対応するディスクリプターを出力するのではなく，キーワード群で表現される概念を表わすディスクリプター群を出力するものであり，Field や Battlle の方法とはアプローチがまったく異なる。

個々のキーワード単位の変換であれば，キーワードの語尾処理を行うにせよ，変換は対応するディスクリプターの有無を調べることと同義といえよう。しかし，キーワード群を総合的に変換する場合には，ディスクリプターの抽出に何らかの演算処理を施さざるを得ない。本システムでは単純な重みつきの加算によってディスクリプターの抽出を行ったが，第11図では一応妥当な結果が得られている。ただ，要素値が低くなると関係があまりないディスクリプターが出力されるので，出力すべきディスクリプターの数は入力キーワード数と同程度が妥当であろう。

この変換方法の問題点としては次のような点があげられよう。

a．変換表の作成にあたって各対応表に割り当てる重みの明確な決定基準がない。

b．同じ単語を共有する，概念上まったく異なった複数の概念が関連があるとみなされる。その典型が同形異義語であるが，その他にも Linear Programming と Programming Language のように誤った処理がなされやすいケースは少なくない。

c．変換をさらに普遍的に行うには語尾処理が必要である。しかし，変換に語幹を用いるのは，上述ので示したようなノイズが増大する可能性があるので単数・複数の処理の程度が妥当と思われる。

d．多くのディスクリプターの構成要素として出現する単語（例えば，System, Programming など）がディスクリプターの採択に大きな影響を与える。

VI. おわりに

キーワード群をディスクリプター群へ自動変換する意図は，前者を同一あるいは類似の概念を持つ後者に変換することにある。しかし，ディスクリプターの概念とその語形とは必ずしも対応がある訳ではないので，語形に基づいた変換ではある程度の無理が生じるのは仕方があるまい。この問題を解決するには，処理の段階で何らかの特殊な辞書を使用するか，シソーラスを作成する時点で後日の自動処理に適した語形のディスクリプターを考慮するかを検討する必要があろう。

しかしこの方法は，文献の検索にあたって適切なディスクリプターを選択する際きわめて有効である。検索者が思いうかべたキーワード群を入力すれば，対応するディスクリプター群が出力されるので，検索者はその中から情報要求に最も適したディスクリプターを選択することができるからである。また，出力されるディスクリプターの数を多くして，serendipity の機能をもたせることもできよう。さらに，対応表の重みを検索者が変更できるようにすれば，再現率を重視する場合はＮＴ表，関連分野の関心が高い場合はＲＴ表の重みをそれぞれ高めて変換を行うことができ，シソーラスの構造，内容に不慣れな検索者もシソーラスを最大限に活用できることになる。

一方，この方法は索引作業にも応用することが可能であり，人手による索引作業の迅速化や自動索引手法の開発にも役立ちうる。

コンピュータによる処理技術の進歩と全文データベースの着実な増加によって，情報検索におけるシソーラスの役割は影が薄くなりつつあるようにみえる。しかし，シソーラスは依然として情報検索システムの基本的な構成要素であり，将来もその重要性が失なわれることはないと思われる。また，その重要性はシソーラスを有効に活用する方法，手段の開発によってさらに高まるといえよう。

1) “EUSIDIC Database Guide”. 1983. 324 p.
2) Lancaster, F. W. “Vocabulary Control for Information Retrieval”. Washington, D. C., Information Resources, 1972. 233 p.
3) Field, B. J. “Towards Automatic Indexing: Automatic Assignment of Controlled-Language Indexing and Classification from Free Indexing”. Journal of Documentation. Vol. 31, No. 4, p. 246–65 (1975).
4) Niehoff, Robert, et al. “The Design and Evaluation of a Vocabulary Switching System for Use in Multi-base Search Environment”. Columbus, Battlle, 1980. 158 p.
5) Niehoff, R. T. and Kwasny, S. “The Role of Automated Subject Switching in a Distributed Information Network”. Online Review. Vol. 3, No. 2, p. 181～194 (1979).
6) “Integrated Energy Vocabulary”. NTIS, 1976. 447 p.