Library and Information Science No. 21 1983

索引過程における認知構造

Cognitive Structure in Human Indexing Process

後藤智範
Tomonori Gotoh

Résumé

This paper discusses the problems of information processing behavior in human indexing processes, which consist of two parts;

1. a process of recognizing of indexable concepts.
2. a process of transforming of indexable concepts into a set of descriptors.

In above each process, the problem as to what kind of information processing occurs in the human indexer's brain has not been reserched. What is known as, cognitive science is being usual taken as a new approach to this problem. Based on this approach, human indexing process is regarded as an information processing system.

An experiment on two subjects who have had experience as an indexer was conducted, protocol analysis method.

The following conclusions have been reached from the results of the experiment.

1. In the recognition process, the subject's interpretation of an abstract is affected by the structure and the level of specification of the abstract.
2. In the transformation process, more complicated information processing is carried out than that required in procedures described in indexing manuals.

I. 序
II. 研究方法
A. 認知科学的アプローチ
B. プロトコル・アナリシス

後藤智範：慶應義塾大学文学研究科博士課程，東京都港区三田2-15-45
Tomonori Gotoh, Ph. D. Course, School of Library and Infomation Science, Keio University, 2-15-45, Mita, Minato-Ku, Tokyo

III. 実　験
A. 実験計画
B. 結　果
IV. 考　察
A. 索引すべき概念の認識過程
B. 索引すべき概念のディスクリプタへの翻訳
V. 結　論
A. 索引すべき概念の認識過程
B. 索引すべき概念のディスクリプタへの翻訳

I. 序

今日，文献データベースはあらゆる学問分野にわたって作成され，その数は数百に上っている。データベース数，レコード数ともに増加の一途をたどっている。70年代，初頭にはいり，情報検索，ドクメンテーションといわれる分野と，計算機科学，通信工学の分野の成果が結集され，DIALOG, ORBIT といった大規模なオンライン情報検索システムが商用化されるに至っている。これらのシステムは文献データベースにリアルタイムでアクセスすることができる。

以前は，研究者は，自分の研究に関連する論文，レポート等を探すために，索引誌，抄録誌等の二次資料を手間暇かけて探していた。しかしながら，今日では上にあげたシステムに結ばれた端末に，質問式を入力すれば，大型計算機が瞬時に答えを出してくれる。文献情報の検索においては，20年前と今日とでは格段の差がある。

一方，文献情報の蓄積に関してはどうであろうか。文献情報の蓄積において中心的な過程は索引作業であるが，この過程に関して20年前と今日とでは検索におけるような著しい差は見られない。索引作業は，現在も過去と同様に，多くの専門の索引者によって行なわれている。

索引に関する問題は，ドクメンテーションの中心テーマとして積極的に研究されてきた。例えば，

索引作業に対する訓練の効果[1]

索引者間の consistency[2),3),4),5]

等があげられる。しかしながら，これらの研究は，索引作業の結果に関するものであって，索引作業そのもの，索引者の内的な過程ではない。索引者の内的過程，すなわち索引過程に関連する研究として，索引作業の手段，方法についての研究がある[6),7),8]。このような研究は，索引過程について，データベース作成機関が提供している索引マニュアルに示されている索引手順よりも多くの示唆を与えてくれる。しかし，実際の索引作業中に，索引者がどのような情報処理を行なっているかは，解明されない。

本論文は，このような問題に対する解明の糸口を見つけようとするものである。次章でこの問題に対する立脚点，および研究方法を述べる。

II. 研　究　方　法

A. 認知科学的アプローチ

認知科学的側面から見ると，部屋に入る，あることを思い出す等の極めて日常的な行為から，数学の問題を解く，物理の本を読む，といった知的な行為に至るまで，人間の諸々の活動の最中には，頭脳の中で何らかの情報処理が行なわれている，と考えられている。したがって，認知科学的な視点に立つと，索引過程をある種の情報処理システムとみなすことができる。このような視点から索引過程を解明することは，すなわち，この情報システムがどういう構造であるのかを明らかにすることにほかならない。

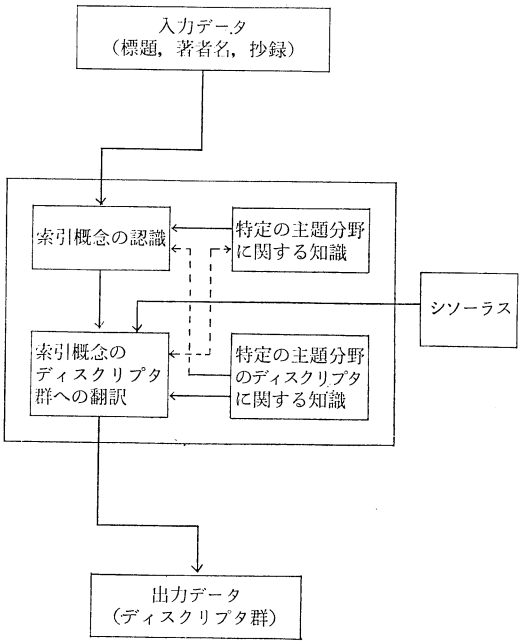
索引作業について Lancaster[9] は，次のような説明をしている。

索引作業：

主題索引の過程はつ2の明らかに独立したステップを含んでいる。すなわち文献の“概念分析”—我われはこれを内容分析と呼んでいる—，および概念分析の特定の語彙への翻訳である。個々のステップが明確に区別されるのはまれである。このことは残念なことである。というのは，それぞれのステップが異なった制

約を持ち，システムの性能に影響を与えている異なった要因をもたらしているからである。効率的な概念分析のために，索引者はその文献が何に関するものなのかを理解すること，すなわち，その主題内容の解釈，およびそのシステムの利用者の要求に関する正しい知識の，両者を必要とする。その文献が何に関するものであるか，および利用者がなぜその文献に興味を持っているか，すなわちその文献のどの側面に関心があるかの，両者を認識することは，概念分析が何を含むか，ということである。文献の概念分析の結果は紙に書かれるかもしれない。しかしながらそれは索引者の頭の中だけに存在する，というのがもっともであろう。

索引過程を1つの情報システムと仮定した時に，これを索引システムと呼ぶことにする。上記のLancaster の説明に従えば，索引システムは第1図のように表わすことができる。



第1図 索引システム

この図の中で，実線は情報が明らかに使用されることを示しており，点線はそうではないことを示している。これは，“索引すべき概念の認識プロセス”では，索引者が持っている，ある特定の分野の情報を利用することは明白であるが，このプロセスで，特定分野の情報のディスクリプターの知識を利用しているかどうかが，過去の研究では明確にされていないからである。各プロセスにおいて，どのような情報処理がなされるかを考察することが本研究のテーマである。

B. プロトコル・アナリシス

認知科学の分野においてプロトコル・アナリシス (protocol analysis) という方法が，認知過程の研究に使われている。“プロトコル”というのは，“解釈を加えないで正確に観察を報告したもの”[10] である。通常，プロトコルは被験者に課題を与え，その課題を解く過程を

プロトコール

問　題

Donald　D=5
+ Gerald
Robert

プロトコール (protcol)

1. それぞれの文字はただ一つの数値と対応する……
2. (実験者：一つの数値です。)
3. 10個のちがう文字がある。
4. そして，それらのひとつひとつが一つの数値をもっている。

5. だから，ええと，二つのDがある……
6. Dはそれぞれ5である。
7. だからTは0である。

8. そうだ，その問題をここに書いてみよう。
9. Tを書こう，Tは0だ。
10. さて，ほかにTがあるかな？
11. なし。
12. しかし，Dがもう一つある。
13. それは，反対側の上に5があるということだ。
14. さて，Aが二つある。
15. Lも二つある。
16. それぞれにだ。
17. ほかにどこか。
18. そうだ　このR。
19. Rが三つ。

(321ステップまで続く。)

ニューエルとサイモン(Newell & Simon,1972)から引用

第2図 覆面計算問題のプロトコル

出典：Mayer, R. E. (佐古順彦 訳) “新思考心理学入門” サイエンス社，1979, p. 155

声に出してもらい，それを録音することによって得られる。声を出すことによって，被験者の頭の中で行なわれる情報処理のプロセスを知ろうとするものである。

この方法を用いた研究の1つに，認知科学に大きな影響を与えた，Newell and Simon[11]の問題解決過程に関する研究がある。彼らは，1人の被験者に，第2図に示した覆面計算の問題を与え，その被験者の解を得るまでのプロトコルを録音し，分析した。分析結果から問題解決の過程をモデル化し，GPS (General Problem Solver：一般問題解決器）と呼ばれる問題解決のシミュレーション・プログラムを開発した。

プロトコル・アナリシスは問題解決の研究以外でも，テキストの構造の理解[12]や長期記憶からの情報の検索[13]の研究において，有力な研究方法として使われている。索引過程に対し，情報処理的な側面から研究する限りにおいては，プロトコル・アナリシスは有力な研究方法であると考えられる。

III. 実　　験

A. 実験計画

1. 目的

本研究は，索引過程を構成しているつのプロセスを独立のものとして扱い，それぞれのプロセスでなされる情報処理について，次に示す問題を明らかにしたい。

1) 索引すべき概念の認識過程：
   1. 被験者はフリー・キーワードを付与するのに，抄録中のどの情報を利用したか。
   2. 個々のフリー・キーワードの主題表現力は抄録の構造とどのような関係があるのか。
   3. フリー・キーワード間の関係と抄録の構造とはどのような関係があるのか。
   4. 1，2，3について被験者間でどの程度の相違があるのか。
2) 索引すべき概念のディスクリプターへの翻訳：
   1. 付与されたフリー・キーワードそれ自身がディスクリプターである場合，被験者はどのような処理をするのか。
   2. 付与されたフリー・キーワードがディスクリプターではない場合，どのような処理をするのか。

2. 被験者

過去の索引作業の研究では，被験者はすべて図書館学専攻の大学院生であった。索引作業についての経験的な知識，および索引対象とする特定の専門知識，という2つの面において，実際の索引者と，図書館学専攻の大学院生とでは大きな差があると思われる。したがって，索引過程における情報処理の詳細な分析に，被験者として大学院生を採用することは，不適当と思われる。そこで，実験では職業としての索引経験のある2名の情報専門家を，被験者とした。（多くのデータを収集できる，という点では被験者の数は多いほうがよいが，専門分野の近い情報専門家に，このような研究に参加してもらうのは，事実上不可能と思われる）被験者は2名で，ともに JICST で索引者としての経験を持っている。

Clark and Bennett[1]は，索引作業に影響を及ぼす要因をいくつかあげている。（第1表を参照）

表中の3・4は，索引作業に時間的制約がなければ，結果に大きな影響を与えないであろう。5・6・7の点について，上述したように，本研究の被験者は同じデータベース作成機関の索引経験者であり，また知識背景も近いので，被験者間で条件は等しいと思われる。年齢も索引作業に影響を与える要因としてあげられているが，年齢のどういう側面が影響を与えるのであろうか。年齢

第1表　索引作業に影響を与える要因

1. Intelligence
2. Age
3. Typing ability
4. Reading speed
5. Orientation towards a particular method of indexing
6. Orientation towards a particular library or category of library
7. Level of knowledge of topicality of documents to be indexed
8. Learning aptitude

出典：Clark, D. C. and Bennett, J. I., "An Experimental Framework for Observing the Indexing Process", J. of ASIS., vol. 24, p. 12 より

と索引経験が比例していれば，索引作業に影響を与えるであろうが，Clark らはこの点を明らかにしていない。2名の被験者は年齢は異なるが，索引経験の年数について大きな差はないので，この点についても被験者間の相違はない。8の学習態度についても，上記の理由により考慮する必要はないと思われる。

3. 使用データ

使用されたデータは抄録を含んだ2つの書誌データで，ともに原文献は英語で書かれた論文である。実際の索引作業では，必要に応じて原文献にも参照するが，ここでは抄録文中から概念をどう選ぶかに焦点を当て，原文献は与えていない。抄録1は，広い範囲の主題を扱っており，抄録2は，特定性の高い主題を扱っている。

次に，実験で使用された2つの抄録を示す。

抄録1：オンライン検索 —現状と将来

ONLINE RETRIEVAL —TODAY AND TOMORROW

WILLIAMS, M. E.

ONLINE REVIEW (USA) 2 (4), 353—366 ('78)

機械化情報検索が経済的に成り立つようになり5500万件のレコードを収録した400の公開データベースが存在し，その75%以上はオンライン利用が可能である。しかし，現在の検索は主として書誌的二次資料が対象で，一次文献へのアクセスが次の段階で必要となり，可能性を十分に発揮していない。各種のデータベースは共通性がなく，システム全体をもっと利用者に使いやすいものにする必要性がある。そのため，distributed system や transparency について研究がなされ，将来はデータとか事実，知識を直接検索する方向に向かうものと予想される；写図1 参54

抄録2：共出現データの利用による文献検索におけるフィード・バックの評価

AN EVALUATION OF FEEDBACK IN DOCUMENT RETRIEVAL USING CO-OCCURRENCE DATA

HAPPER, D. J. AND VAN RIJSBERGEN, C. J.; J DOC (GBR) 34 (3), 189—216 ('78)

関連情報の中に含まれる索引語の間には依存性があるという仮定（依存モデル）に従って，索引語の重み付け検索のモデル実験を行なった。最初に，上記の仮定を広い文献集合について検証した。重み付けによる関連文献利用は，検索効率を向上させる可能性を証明し，さらに，厳格な依存モデルは改良されうることを示して，これに基づく新しい重み付法を開発した。改良モデルに基づいて関連文献のフィード・バックを行ない，部分関連情報のみで，検索効率が有意に向上することを示した。フィード・バック実験で用いた評価法は，特に，未知文献の発見への効果が大である；写図10，表9，参18

4. 手続き

2名の被験者に対し，2つの抄録（表題，著者名を含む）を次に示す手続きに従って処理する。

1. フリー・キーワードを付与する。
2. JICST シソーラスに管理されているディスクリプターを付与する。
3. このときに，どのフリー・キーワードをもとにしたのかを明らかにするために，フリー・キーワードの前についている番号を付与する。
4. 1，2，3の処理をしているあいだに，頭の中に浮かんだことをすべて声に出す。
5. 付与されたフリー・キーワードを主題表現力の高い順に並べ換える。
6. 付与したディスクリプターどうしを関係の強弱に従って，強(＝＝＝＝＝)，中(————)，弱(－－－－－)の3種類の線で結び，フリー・キーワード間の関係を表現する。

上述のすべての課題について，時間的制約を与えていない。時間的制約を与えることによって，各過程において省略，たとえば，時間がなくなってフリー・キーワードの一部がディスクリプターに翻訳されない，などが起こりうるからである。

5. プリテスト

上記に示された実験を行なう前に，特に手続き1の過程のプロトコルを記録した。被験者Aと被験者Bとで，声に出す程度に差があり過ぎるため，プリテストが終わった後に被験者Aに被験者Bのプロトコルを見せた。そして，あるフリー・キーワードがどうして付与されるのか，その過程を声に出すように指示した。

B. 結果

1. フリー・キーワード

第2表　フリー・キーワードのリスト1（抄録1）

	被験者A	被験者B
1.	オンライン検索	オンライン検索
2.	機械化	情報検索
3.	情報検索	データベース
4.	データベース	オンライン利用
5.	オンライン利用	二次情報
6.	二次資料	オンライン・システム
7.	情報検索システム	データ
8.	共通性	事実
9.	distributed system	知識
10.	transparency	一次文献
11.	データ検索	
12.	事項検索	
13.	一次資料	
14.	利用者	

第3表　フリー・キーワードのリスト2（抄録2）

	被験者A	被験者B
1.	文献検索	文献検索
2.	フィードバック	フィードバック
3.	索引語	評価
4.	重みづけ	索引語
5.	検索効率	検索効率
6.	依存モデル	文献集合
7.	関連文献	重みづけ
8.	評価	
9.	共出現データ	
10.	文献	

2. ディスクリプター

第4表　ディスクリプターのリスト1（抄録1）

	被験者A	被験者B
1.	オンライン・システム	情報検索
2.	情報検索システム	オンライン処理
3.	データベース	オンライン・システム
4.	二次資料	二次資料
5.	一次資料	データベース
6.	オンライン処理	一次資料
7.	オンライン利用	
8.	事項検索	
9.	情報利用	
10.	情報サービス	
11.	情報検索	

第5表　ディスクリプターのリスト2（抄録2）

	被験者A	被験者B
1.	文献検索	文献検索
2.	フィードバック	フィードバック
3.	索引語	システム評価
4.	重みづけ	索引語
5.	検索効率	検索効率
6.	依存性	重みづけ
7.	模型	
8.	一次資料	
9.	評価	
10.		

3. フリー・キーワードからディスクリプターへの翻訳

第6表 フリー・キーワードからディスクリプターへの翻訳1（抄録1：被験者A）

	フリー・キーワード	ディスクリプタ
1.	オンライン検索	オンライン・システム
2.	機械化	オンライン処理
3.	情報検索	情報検索
4.	データベース	データベース
5.	オンライン利用	オンライン利用
6.	二次資料	二次資料
7.	情報検索システム	情報検索システム
8.	共通性	
9.	distributed system	情報サービス
10.	transparency	
11.	データ検索	
12.	事項検索	事項検索
13.	一次資料	一次資料
14.	利用者	情報利用

第7表 フリー・キーワードからディスクリプターへの翻訳2（抄録1：被験者B）

	フリー・キーワード	ディスクリプタ
1.	オンライン検索	オンライン処理
2.	情報検索	情報検索
3.	データベース	データベース
4.	オンライン利用	
5.	二次情報	二次資料
6.	オンライン・システム	オンライン・システム
7.	データ	
8.	事実	
9.	知識	
10.	一次文献	一次資料

第8表 フリー・キーワードからディスクリプターへの翻訳3（抄録2：被験者A）

	フリー・キーワード	ディスクリプタ
1.	文献検索	文献検索
2.	フィードバック	フィードバック
3.	索引語	索引語
4.	重みづけ	重みづけ
5.	検索効率	検索効率
6.	依存モデル	依存性
7.	関連文献	
8.	評価	評価
9.	共出現データ	
10.	文献	一次文献
		模型

第9表 フリー・キーワードからディスクリプターへの翻訳4（抄録2：被験者B）

	フリー・キーワード	ディスクリプタ
1.	文献検索	文献検索
2.	フィードバック	フィードバック
3.	評価	システム評価
4.	索引語	索引語
5.	検索効率	検索効率
6.	文献集合	
7.	重みづけ	重みづけ

4. プロトコル

＜抄録1：被験者A＞

「オンライス検索，現状と将来」

「まず，索引語として，オンライン検索ですね」

「機械化情報検索が経済的に成り立つようになり，5万，5千5万，5千5百万件のレコードを収録した400の公開データベースが存在し」

「んと，オンライン検索，から，機械化の情報検索，機械化，情報検索」

「75%以上はオンライン利用が可能である」

「それから，一次文献へのアクセスですから，一次資料」
「共通性がないということで，共通性」
「これは，情報検索システムのことがここに書いてあるので，んと，情報検索システム」
「それから，利用者に使いやすいものにする必要がある，ということで，利用者」
「distributed system，これは英文で書かれてあるので，そのまま書いておきます」
「distributed system」
「から，transparency，transparency」
「それから，データとか，事実，知識を直接検索する方向に向かうものと予想される」
「ですから，データ検索」
「それから，事項検索」
「と，これでいいかな」
「オンライン検索，機械化，情報検索，データベース，オンライン利用，二次資料，情報検索システム，共通性，distributed system，transparency，データ検索，事項検索，一次資料，利用者」
（ここまで所要時間3分3秒）
「まず，オンライン検索」
「それから，オンライン処理」
「オンライン検索，オンライン利用」
「えーと，オンライン検索っていうのは，オンラインシステムというのがありますので，オンライン・システム」
「オンライン処理」
「これは，あの，機械化というのも，オンライン検索と機械化というのが同時に含まれているので，オンライン・システムとオンライン処理」
「情報検索と情報検索システム」
「えーと，情報検索システムは，そのまま情報検索システム」
「情報検索は，情報検索は，これは情報検索とかきます」
「から，二次資料」
「これもあります，二次資料」
「から，一次資料」
「distributed system と transparency，これは，ディスクリプターがないので，後回しにしまして，データ検索と事項検索というのは，これは事項検索というのがありますので，書きます」
「それから，えー，先程言いました distributed system と transparency，適当なキーワードがないので，情報利用とか情報サービス」
「これで，ちょっと間合わせておかないと，思います」
（ここまで所要時間3分7秒，計6分10秒）

＜抄録1：被験者B＞

オンライン検索 現状と将来，ONLINE RETRIEVAL TODAY AND TOMMOROW」
「機械化情報検索が，経済的に成り立つようになり，5万，えー，5千500万件のレコードを収録した400の公開データベースが存在し，その75%以上はオンライン利用が可能である」
「しかし，現在の検索は，主として書誌的二次資料が対象で，一次文献へのアクセスが次の段階で必要となり，可能性を十分発揮していない。各種のデータベースは，共通性がなく，システム全体をもっと利用者に使いやすいものにする必要がある。そのため，distributed system と transparency について，研究がなされ，将来はデータとか事実，知識を直接検索する方向に向かうものと予想される」
「まず標題から，そのものずばりのオンライン検索」
「それから，あとは，抄録から，機械化情報検索」
「これは，情報検索」
「それから，あとは，うーん，公開データベース，データベースね」
「それから，あとは，うーん，オンライン利用。オンライン利用が可能であるの，オンライン利用」
「現在の検索。これはさっき取った」
「書誌的二次情報，これは，書誌的二次情報。二次情報でいいな」
「書誌的はいらない」
「それから，あとは，一次文献」
「うーん，あとは，アクセスが次の段階で必要となり」
「ここらへんはずっとなしで」
「データベースはさっき取った」
「そして，システム全体をもっと利用者に使いやすいものにする必要」
「そうか，事実とか知識，いわゆるファクト・リトリーバル」
「えーと，事実，知識，こういったもの」
「だいたい，こんなとこですね」
（ここまで所要時間3分）

「まず，この今のフリー・キーワードを参考にして」
「オンライン検索というのはないんだな」
「オンライン検索というのはたしかなかった」
「んー，オンラインシステムっていうのがあって，その中に情報検索システムっていうのがあるんだが」
「それだったら，オンライン検索はオンラインシステム」
「それから次に，情報検索は，まさにそのものずばりの情報検索，というのがあったと思うから，これをふると」
「それから後は，オンライン利用，これはないな」
「ない。んー，オンラインシステム，システムね」
「あと二次情報」
「二次情報というのはたしかないんだな」
「二次情報なし。それに代わって二次資料がある」
「二次資料」
「それから後は，あとは，データベース」
「データベースはそのものずばりがたしかあるから，これでデータベース」
「それから，一次文献」
「一次文献というのはないな」
「たぶん，一次資料，あ，そう」
「一次資料，そうですね。use for で一次文献。だから一次資料」
「だいたいこれでO. K.」
（これでここまでで所要時間3分53秒。計6分53秒）

＜抄録2：被験者A＞

「共出現データの利用による文献検索のフィード・バックの評価」
「えーと，これはですね，タイトルの中から索引語を選びたいと思います」
「まず，文献検索」
「から，フィード・バック，評価」
「から，共出現データ」
「抄録文中の中から選びます」
「まず，索引語，索引語ですね」
「関連文献」
「から，依存モデル」
「これは，あの，えーと，関連情報の中に含まれる索引語の間には依存性がある，という仮定にたって，索引語の重み付検索モデルの実験を行なったということで，索引語とか，重み付とか」
「関連文献というのを選びました」
「から，検索効率を向上させる可能性を証明したとあるので，検索効率」
「文献検索，フィードバック，索引語，重み付，検索効率，依存モデル，関連文献，評価，共出現データ」
「ん，あと，あと，文献ですね」
「これだけです」
（ここまで所要時間2分33秒）
「えーと，文献検索はこのまま文献検索」
「と，1番ですね」
「フィード・バック，これはそのまま」
「これもありますからそのままです」
「重み付，これは」
「重み付は4番」
「検索効率，これもあります」
「それから，依存性モデルは，これはないので」
「依存性というのがありますので，依存性，依存性と」
「モデルは模型となっているので，模型」
「文献は，use 資料と，なっているので，資料のところを見ますと，資料の中にはたくさんあるんですが，一次資料，これは10番ですか」
「関連というのは何かいいのがありますかね」
「これは，別にないので，別にないですね」
「評価，これは，評価，あります」
「評価，8番」
「こんなものかな」
「関連情報というのはうまいのがないかな」
「ないな」
「以上です」
（ここまで所要時間5分58秒。計8分31秒）

＜抄録2：被験者B＞

「共出現データの利用による文献検索におけるフィード・バックの評価」
「関連情報の中に含まれる索引語の間には依存性があるという仮定，依存モデル，に立って，索引語の重み付の検索モデル実験を行なった最初に上記の仮定を広い文献集合について検証した。重み付による関連文献の利用は，検索効率を向上させる可能性を証明し，さらに，厳格な依存モデルは改良されえることを示して，これに基づく新しい重み付を開発した改良モデルに基づいて関連文献のフィード・バックを行ない，部分関連情報のみで，検索効率が有意に向上することを示した」
「フィード・バック実験で用いた評価法は，特に，未

知文献の発見への効果が大である」
「これは，ちょっと抄録があんまりよくないんだなあ」
「つけにくいけど，まあ標題から，ん，文献検索」
「それから，フィード・バック，この二つがあげられる」
「あと，評価ね，評価ね」
「それからあとは，抄録から，まず索引語」
「それからあとは，うーん，重み付ね，重み付」
「あとは，文献集合っていうのも必要でない」
「んー，検索効率」
「んー，そんなとっかな」
「フィード・バックはさっきやった」
「表題中からとったと」
「それから，関連情報はだめ」
「未知文献，これもだめ」
（ここまで所要時間2分45秒）
「今度は，シソーラスを使って，まず，文献検索はそのものずばりがあるから，これに文献検索を採用すると」
「フィード・バック，フィードバックもあるんじゃないかな，たしか」
「フィード・バック，フィード・バック」
「フィード・バック，ある」
「それから後は，評価，あんまり好ましくないかな」
「システムの評価がいんじゃないかな」
「あとは，索引語っていうのは，そのまま索引語があるから，ここで索引語」
「検索効率もそのものずばり」
「重み付っていうのがあったかな，重み付」
「あり，あり，重み付ね」
「これで，終わり」
（ここまで所要時間2分05秒，計5分35秒）

5. フリー・キーワードの重要度順リスト

IV. 考　　察

実験により，2つの抄録に対し6種類のデータが得られたが，これらのデータは索引過程に含まれる2つの過程に関する情報として次のように整理できる。

1) 索引すべき概念の認識過程

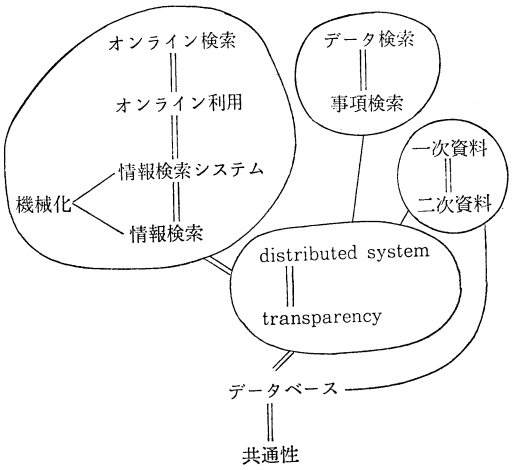
第10表　フリー・キーワードの重要度順リスト1
（抄録1）

	被験者A	被験者B
1.	オンライン検索	オンライン検索
2.	情報検索システム	情報検索
3.	情報検索	オンライン利用
4.	オンライン利用	オンライン・システム
5.	機械化	データベース
6.	データベース	二次情報
7.	distributed system	一次文献
8.	transparency	データ
9.	データ検索	事実
10.	事項検索	知識
11.	二次資料	
12.	一次資料	
13.	利用者	
14.	共通性	

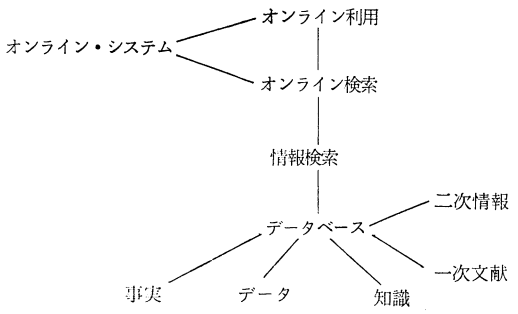
第11表　フリー・キーワードの重要度順リスト2
（抄録2）

	被験者A	被験者B
1.	文献検索	文献検索
2.	フィードバック	フィードバック
3.	索引語	評価
4.	検索効率	索引語
5.	重みづけ	重みづけ
6.	関連文献	検索効率
7.	共出現データ	文献集合
8.	文献	
9.	依存モデル	
10.	評価	

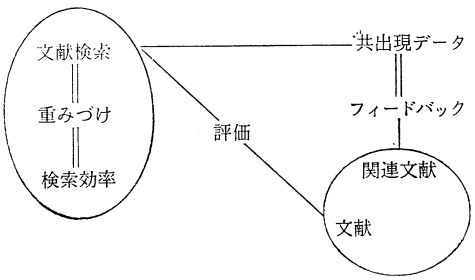
6. フリー・キーワード間の関係



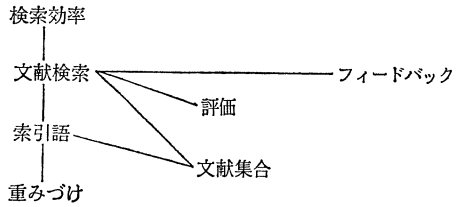
第3図 フリー・キーワード間の関係1
(抄録1：被験者A)



第4図 フリー・キーワード間の関係2
(抄録1：被験者B)



第5図 フリー・キーワード間の関係3
(抄録2：被験者A)



第6図 フリー・キーワード間の関係4
(抄録2：被験者B)

1. フリー・キーワード
2. プロトコル
3. フリー・キーワードの重要度順リスト
4. フリー・キーワード間の関係

2) 索引すべき概念のディスクリプターへの変換

1. ディスクリプター
2. プロトコル
3. フリー・キーワードからディスクリプターへの翻訳

A. 索引すべき概念の認識過程

分析をはじめる前に，ここで抄録の構造を調べる必要がある。

＜抄録1＞

抄録1は，4つの文からなり，それぞれがいくつかの単文から構成され，次のように分けられる。

T ：オンライン検索-現状と将来
S1a：機械化情報検索が成り立つようになり
S1b：5500万件のレコードを収録した400の公開データベースが存在し
S1c：その75%以上はオンライン利用が可能である
S2a：しかし，現在の検索は主として書誌的二次情報が対象で
S2b：一次文献へのアクセスが次の段階で必要となり
S2c：(現在の検索は)可能性を十分発揮していない
S3a：各種のデータベースは共通性がなく
S3b：システム全体をもっと利用者に使いやすいものにする必要がある。
S4a：そのため，distributed system や transparency について研究がなされ
S4b：将来はデータとか知識，事実を直接検索する方向に向かうものと予想される。

(Tは title，Sは sentence を示している)

＜抄録2＞

T　：共出現データによるフィードバックの評価
S１a：関連情報の中に含まれる索引語の間には依存性があるという仮定(依存モデル)にたって，
S１b：索引語の重み付け検索のモデル実験を行なった
S２　：最初に上記の仮定を広い文献集合について検証した。
S３a：重み付けによる関連文献利用は，検索効率を向上させる可能性を証明し，
S３b：さらに厳格な依存モデルは改良されうることを示して，
S３c：これに基づく新しい重み付法を開発した。
S４a：改良モデルに基づいて関連文献のフィード・バックを行ない，
S４b：部分関連情報のみで，検索効率が有意に向上することを示した。
S５　：フィード・バック実験で用いた評価法は，特に未知文献の発見への評価が大である。

1)　抄録中の情報の利用

抄録１についての両被験者のプロトコルから，抄録中の語句と付与されたフリー・キーワードが，第12表，および第13表に示される関係にあることがわかる。

この２つの表から，付与されたフリー・キーワードにおいて，両被験者間で完全に一致したフリー・キーワード，および一致していないフリー・キーワードのどちらも，ほとんど同一の情報を利用していることがわかる。抄録中の語句とフリー・キーワードが１から１に対応していないものもある。このことは，どのような判断基準によるものかはわからないが被験者の頭の中で，主題概念の分解が行なわれていることを示している。

抄録２におけるフリー・キーワードと抄録中の語句との関係は，第14表，および第15表に示される。

被験者Bのプロトコルには“重み付け”というフリー・キーワードが抄録中のどの情報を用いたのか示されていない。単に言い忘れたのだとすれば，“関連情報”，“依存モデル”の２つのフリー・キーワード以外は両被験者ともに同一の情報をもとにしている。

2)　フリー・キーワードの主題表現力と抄録の構造

２つの問題を考察するために，抄録１について第10表と第12表および第13表を合成し，抄録２については，第11表と，第14表および第15表を合成し抄録中の語句の重要度順リストを作成する。第16表は抄録１について，第17表は抄録２についてそれぞれ示している。

この表から抄録１に関しては，標題および第１文に含まれる語句は，両被験者間で同程度の重要性を持っている。第２文に含まれる語句の重要度が，被験者によって評価が異なっている。被験者Aは第４文中に含まれる語句よりも，評価が低く，一方被験者Bはその逆である。

抄録２については，抄録１よりも被験者間での，抄録中の語句に対する重要度の評価の相違が大きい。標題に含まれる語句についても相違が大きい。“評価”という語

第12表　抄録中の語句からフリー・キーワード１（抄録１：被験者A）

文章（標題および抄録）中の語		フリー・キーワード
オンライン検索（T）	→	オンライン検索
機械化情報検索（S１）	→	機械化
	→	情報検索
データベース（S１）	→	データベース
オンライン利用（S１）	→	オンライン利用
書誌的二次情報（S２）	→	二次資料
一次文献へのアクセス（S２）	→	一次資料
共通性（S３）	→	共通性
利用者（S３）	→	利用者
distributed system（S４）	→	distributed system
transparency（S４）	→	transparency
データとか事実，知識（S４）	→	データ検索
	→	事項検索

第13表　抄録中の語句からフリー・キーワード２（抄録１：被験者B）

文章（標題および抄録）中の語	フリー・キーワード
オンライン検索（T）	オンライン検索
機械化情報検索（S１）	機械化
公開データベース（S１）	データベース
オンライン利用が可能である（S１）	オンライン利用
書誌的二次情報（S２）	二次情報
一次文献（S２）	一次文献
事実，知識（S４）	事実
	知識

第14表　抄録中の語句からフリー・キーワード３（抄録２：被験者A）

文章（標題および抄録）中の語	フリー・キーワード
文献検索（T）	文献検索
フィードバック（T）	フィードバック
評価（T）	評価
共出現データ（T）	共出現データ
索引語（S１）	索引語
関連情報（S１）	関連文献
依存モデル（S１）	依存モデル
検索効率を向上（S３）	検索効率

第15表　抄録中の語句からフリー・キーワード４（抄録２：被験者B）

文章（標題および抄録）中の語	フリー・キーワード
文献検索（T）	文献検索
フィードバック（T）	フィードバック
評価（T）	評価
索引語（S１）	索引語
重みづけ（S１）	重みづけ
文献集合（S２）	文献集合
検索効率（S３）	検索効率

第16表　抄録中の語句の重要度順リスト１（抄録１）

	被験者 A	被験者 B
1.	オンライン（T）	オンライン検索（T）
2.	機械化情報検索（S 1）	機械化情報検索（S 1）
3.	オンライン利用（S 1）	オンライン利用（S 1）
4.	データベース（S 1）	データベース（S 1）
5.	distributed system	書誌的二次情報（S 2）
6.	transparency	一次文献（S 2）
7.	事実，知識（S 4）	事実，知識（S 4）
8.	書誌的二次情報（S 2）	
9.	一次文献（S 2）	
10.	共通性	

第17表　抄録中の語句の重要度順リスト１（抄録２）

	被験者 A	被験者 B
1.	文献検索（T）	文献検索（T）
2.	フィードバック（T）	フィードバック（T）
3.	索引語（S 1）	評価（T）
4.	検索効率（S 3）	索引語（S 1）
5.	関連情報（S 1）	重みづけ（S 1）
6.	共出現データ（T）	検索効率（S 3）
7.	依存モデル（S 1）	文献集合（S 2）
8.	評価（T）	

についての重要度は両被験者間で著しく異なっている。抄録２は，抄録１よりも，specificity が高いのでその処理にはより多くの専門知識を必要とする。

3)　フリー・キーワード間の関係と抄録の構造

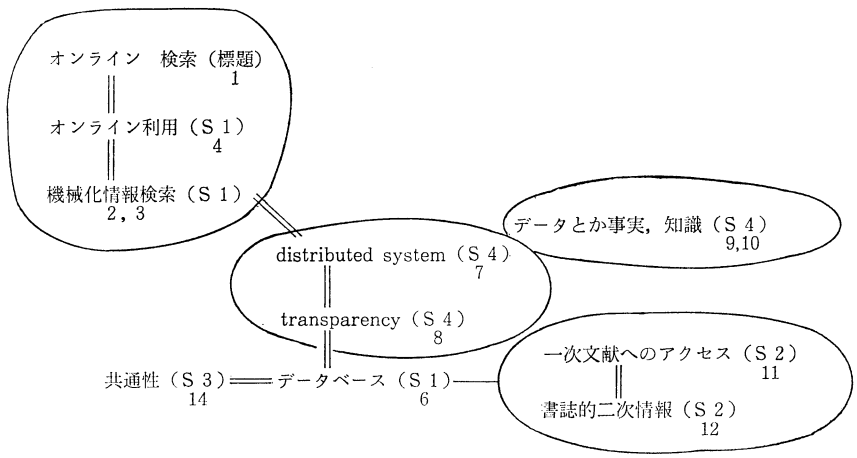
それぞれの抄録において両被験者のフリー・キーワード間の関係を示した図を比較してみたい（抄録１については第７図，および第８図，抄録２については第９図，および第10図を参照）。

＜抄録１＞

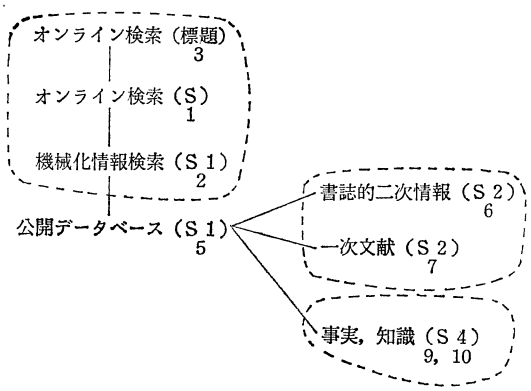
抄録１について第７図のなかの distributed system, および transparency を除くと，２つの図におけるフリー・キーワード間の関係は非常に近いことがわかる。次に第３図に対して第12表，第４図に対して第13表をそれぞれ合成してみる。第７図，および第８図は次のようになる。

筆者は，どちらの被験者に対しても，“フリー・キーワードをグループ化せよ”という指示は与えていないが，被験者Aは，フリー・キーワードをグループ化している。そこで，第８図中の，語句をグループ化してみる。グループ内の語に付与されている文番号に関しては，両図において相違が余りないことがわかる。

第５図と第６図を比較して見ると，抄録１の場合よりも両被験者間での相違が大きいことがわかる。被験者Aは，“索引語”をフリー・キーワードとして付与したのだが，第５図にはない。おそらくこれは被験者Aが書き



第７図　抄録中の語の関係と重要度１　（抄録１：被験者A）



＊：点線は筆者による

第8図　抄録中の語の関係と重要度2
（抄録1：被験者B）

忘れたと思われるが，“索引語”が，書かれていれば，“文献検索”，“索引語”，“重み付け”，というフリー・キーワード間の関係は，両被験者で相違がないと思われる。

＜抄録2＞

ここでも抄録1と同様な処理を行なってみたい。第5図と第14表，第6図と第15表から，抄録中の語の関係と重要度は，それぞれ第9図，第10図に示される。

第10図中の語をグループ化し，各グループ内の語に付与されている文番号を比較してみたい。両図においてAグループに含まれる文番号はほとんど一致している。このグループ中の語は両被験者ともに主題表現力が高いとしている語である。

以上のことから次のようなことが考えられる。

フリー・キーワード間の関係は，抄録の構造を反映していると考えられるが，抄録に対する被験者の解釈を反映している。

また，主題表現力の高いフリー・キーワード間の関係は，主題表現力の低いフリー・キーワードよりも被験者間の相違が少ない。

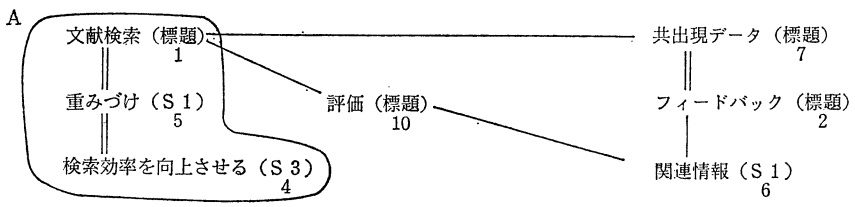
B. 索引すべき概念のディスクリプターへの翻訳。

このプロセスでは，1)および2)の問題について，それぞれの抄録での両被験者の行動を分析してゆきたい。

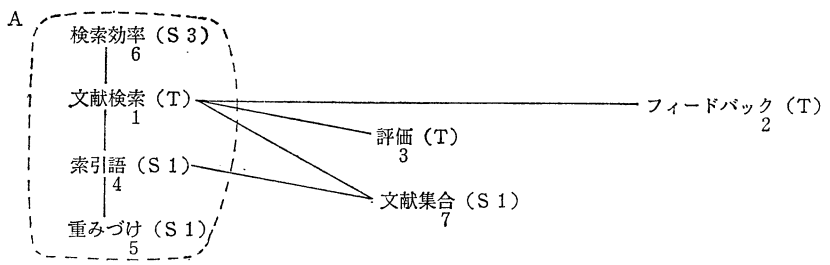
＜抄録1＞

抄録1についての2つのプロトコル，および第6表，第7表が，上述の問題に関するデータを提供している。上述のAの場合には，両被験者ともにそのままディスクリプターとして採用している。一方，Bの場合は様々な処理を行なっている。

1. ディスクリプターの付与をあきらめる。（例，共通性）



第9図　抄録中の語の関係と重要度3　（抄録2：被験者A）



＊：点線は筆者による

第10図　抄録中の語の関係と重要度4　（抄録2：被験者B）

2. 近隣概念を表現しているディスクリプターを付与する。(例，一次文献 ―→一次資料)
3. 暗黙のうちに内在している概念を表現しているディスクリプターを付与する。(例，利用者 ―→情報利用)
4. 概念を分割し，分割された概念を表現しているディスクリプターを付与する。(例，利用者 ―→情報利用，情報サービス)

＜抄録２＞

抄録２についても，１と同ような分析を行なってみる。

Aの場合でも，それ自身をディスクリプターとして付与しないことが見られる。ディスクリプターとしてより適切なものを，ワードブロック中から見つけ，付与している。(例，評価 ―→システム評価)

Bの場合については，抄録１で指摘された処理以外は見られない。

V. 結　　論

A. 索引すべき概念の認識過程

前章の考察から，この過程には次の３つのプロセスが含まれていることに気づく。

A) 当該抄録の解釈
B) 抄録を構成している個々の概念に対し，それを索引すべきかどうか，すなわち，それが主題概念であるかどうか，という評価。
C) 主題概念をどのような語句まで表現するか，すなわち主題概念に対する言語表現上の選択。

従って，索引すべき概念の選択過程，より一般的に，主題分析過程は，第11図に示されるものと考えられる。

A) 入力データの解釈

このプロセスは，索引者が，当該抄録がどういう内容のものかを，索引者の有する主題知識を用いて理解する，という処理を行なう。

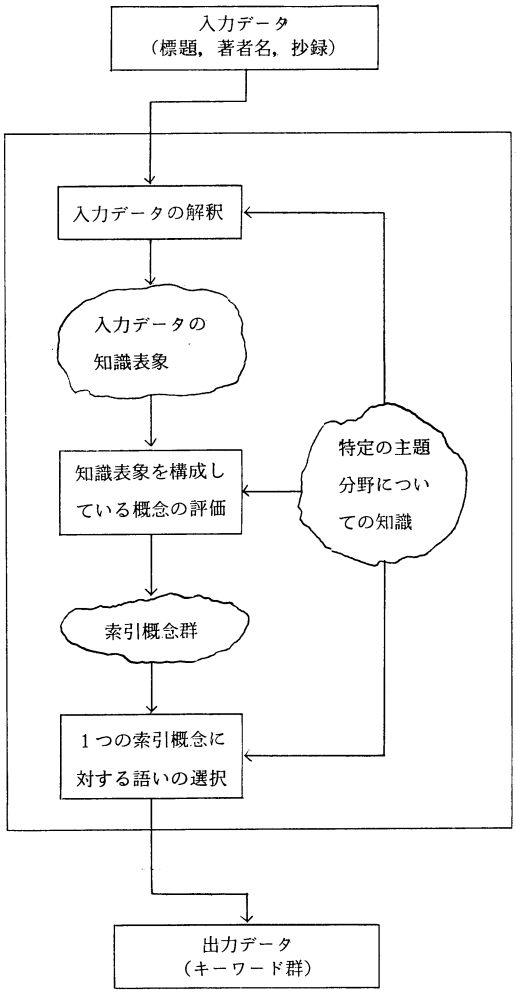
その結果，当該抄録に対する索引者の解釈結果，認知科学でいう知識表象 (knowledge representation) が形成される。

B) 知識表象を構成している概念の評価

このプロセスは，知識表象を構成する個々の概念が索引すべきものであるかどうか―個々の概念がその抄録の主題を表現しているかどうか，という評価を行なう。

C) 索引すべき概念に対する語句の選択

このプロセスで，主題概念を特定の語句に表現する１つの概念に対し，複数の語句が存在している可能性がある。したがって，その中から最も適切な語句を，索引者の持っている主題知識の中から選ばねばならない。最終的に選択された語句がフリー・キーワードとして付与される。



第11図　主頭分析過程の構造

上記の各プロセスを前章での考察を踏まえてさらに深く分析してみたい。前章の考察をまとめると次のようになる。

1) 索引者は，表題および抄録の最初の文に含まれる語句から主題概念を抽出する傾向にある，と考えられる。
2) この部分から得られたフリー・キーワードは，他

の部分に含まれる情報から得たフリー・キーワードよりも主題表現力が高い。さらに，被験者間の相違も少ない。

3) 得られたフリー・キーワードは主題表現力が高いフリー・キーワードを中心に関係づけられている。

4) 特定性の高い抄録の方が，広い範囲の主題を扱っている抄録よりも，付与されたフリー・キーワード，および，それらの関係において，被験者間での相違が大きい。

5) フリー・キーワード間の関係は，被験者の抄録に対する解釈を反映している。

A）入力データの解釈

このプロセスに最も影響を与えているのは，索引者の持っている主題知識のレベルであろう。当該分野について深く知っていればいるほど，対象としている抄録の内容を正確に理解し，当該分野の中に正確に位置付けることができるからである。上記の4）の事実はこのことを示している。個々の索引者の有する知識のレベルは様々なので，specificity の高い文献は，索引者間の解釈の相違をもたらす可能性が大きい。

このプロセスは，抄録を構成している文章の理解のしやすさにも影響されると考えられる。被験者Bは，プロトコルの中で抄録2がわかりにくいことを訴えている。抄録の理解のしやすさが，何によって判定されるのかは，本研究では明らかにされないが，この問題は研究する必要があるであろう。

B）知識表象を構成している概念の評価

索引すべき概念は，何らかの評価基準に従って，知識表象を構成している概念から選択されるのであろうか。2つの抄録に対する被験者Bのプロトコルは，何らかの評価基準が存在していることを示唆しているが，上記の2)，3)，4）は，知識表象が形成された時点で，評価がなされていることを示しているように思われる。2)，3)，4）はさらに，知識表象を構成している概念そのものよりも，概念間の構造が個々の概念の主題表現力を決定していることを示していると思われる。したがって，このプロセスにおいても主題知識が重要な要素であることがわかる。

C）索引すべき概念に対する語句の選択

このプロセスは索引経験の量が最も反映するものと考えられる。ある概念を特定の語句で表現するためには，その概念に対してどういう語句が最も頻繁に用いられているかを知っている必要がある。これは主題知識のレベルの深さだけでなく，当該分野の文献に対する，利用者の専門用語の使い方についての知識をも必要とする。また，対象分野についてのディスクリプターに精通していれば，ディスクリプターをフリー・キーワードとして付与するであろう。被験者Aの抄録1に対して付与したフリー・キーワードの結果はこのことを示唆している。

以上のことから，索引作業一般に対して，次のようなことが言えるであろう。

1. 抄録はわかりやすくかかれている必要がある。
2. 質の高い索引語を付与するには，索引者は，当該分野の知識を豊富に持っていること，さらに，専門用語の利用のされ方について精通していること，少なくともこの2つの項目を満足していることが必要である。
3. 2と関連して spcificity の高い文献は，索引経験の豊富な索引者に処理されることが望ましい。

B. 索引すべき概念のディスクリプターへの翻訳

このプロセスにおいて，次のようなことが明らかになった。

a）フリー・キーワードがそれ自身ディスクリプターである場合

1) ディスクリプターとしてそのまま付与する。

2) さらに，概念的に関係の深いディスクリプターも付与する。

3) そのディスクリプターを含んでいるワード・ブロック内に，より適切なディスクリプターがある場合には，それを付与する。

b）フリー・キーワードがディスクリプターではない場合

1) フリー・キーワードが表現している概念を分解し，分解された概念を表現しているディスクリプターを付与する。

2) フリー・キーワードに対し暗黙に関係している概念を表現しているディスクリプターがあればそれを付与する。

3) ディスクリプターの付与をあきらめる。

以上の事実は，Wellisch[8] の示した索引作業の手続きよりも，さらに複雑な処理が行なわれていることを示している。しかしながら，上記の処理を，第11図のような一連のプロセスとして示すことは，次の理由により困難である。

A．上記の処理は，個々のフリー・キーワードによって異なる。

B．上記の処理は，偶発的に行なわれている。

Aの問題は，個々のフリー・キーワードの概念の範囲と，フリー・キーワードに関連しているであろうディスクリプターの概念の範囲がどの程度一致しているか，という問題に集約していると考えられる。フリー・キーワード自身がディスクリプターでなくとも，それが意味している概念を完全に表わしているディスクリプターが存在すれば，Aの問題はほとんどなくなると考えられる。

一方，Bの問題は，索引者の持っている主題知識の量，および使用しているシソーラスに精通している度合に依存していると考えられる。通常，ディスクリプターもフリー・キーワードもともに専門用語であるので，主題知識が豊富であれば，それらの対応関係は，前もってわかっているはずである。したがって，あるフリー・キーワードに対してどのディスクリプターが関連しているか，という知識を持っているので，偶然，適切なディスクリプターをシソーラス中から見つけて付与する，という処理の仕方は減少するであろう。シソーラスに精通していることも，同様な結果をもたらすであろう。

本研究で実験対象とされた索引者は2名であり，より多くの索引者を対象とすれば，フリー・キーワードのディスクリプターへの翻訳の過程に関する限り，さらに多くの手続きが見出だせるであろうと思われる。

1) Clark, D. C. ; Bennett, J. I. "An Experimental Framework for Observing the Indexing Process". Journal of the American Society for Information Science. vol. 24, p. 9-24 (1973).
2) Bottle, R. T. ; Schwarzlander, H. "Variations in the Assesment of the Information Content of Documents". Proceedings of the American Socety for Informaiton Science, vol. 7, p. 279-281 (1970).
3) Tarr, D. ; Borko, H. "Factors Influencing Inter-Indexer Consistency". Proceedings of the Amercican Society for Information Sciece. vol. 11, p. 50-55 (1974).
4) Slamecka, V.; Jacoby, J. "Effect of Indexing Aids on the Liriability of Indexers". Bethesda, Maryland, Documentation Inc., 1963.
5) Cooper, W. S. "Is Inter-Indexer Consistency a Hobgoblin". American Documentation. vol. 20, no. 3, p. 268-278 (1969).
6) MaCalister, C. "A Study and Model of Machine-Like Indexing Behavior by Human Indexers". University of California, Berkeley, 1971. Ph. D. Thesis.
7) Oliver, L. H., et al., "An Investigation of the Basic Processes Involved in the Manual Indexing of Scientific Documents". Bethesda, Maryland, General Electric Co., 1966. (PB 169 415)
8) Wellisch, H., "A Flow Chart for Indexing with a Thesaurus". Journal of the American Society for Infomation Science. vol. 10, p. 185-194 (1972).
9) Lancaster, F. W., "Informaion Retrieval System : Characteristics, Testing and Evaluation 2ed.". New York, John Wiley & Sons, 1979. 381 p.
10) Mayer, P. E. : 佐古順彦訳. "新思考心理学入門". 東京，サイエンス社，1979，235 p.
11) Newell, A. ; Simon, H. A. "Human Problem Solving". Englewood Cliffs, N. J., Prentice-Hall, 1972.
12) Graesser, A. C., et al., "Incoporating Inferences in Narrative Repres entations ; A Study of How and Why". Cognitive Psychology. vol. 13, p. 1-26 (1981).
13) Williams, M. D., "Observations on the Process of Retrieval from Long Term Memory". Univesity of California, San Diego, 1977. Ph. D. thesis.