特定分野情報の分析と調査

Referationを用いた特定専門分野情報の分析

○浅井　勇夫*

研究活動を支援するため，特定専門分野に関するパーソナルな文献コレクションから，その特徴を分析するパソコン用ソフトを開発する。引用文献から得られるReferation関係を用いて，3種類の文献群間の関連性測度を定義する。それは，関連度数，関連係数，および関連指数である。この論文は，(1) 22種類の情報に支援された文献検索システム，(2) 文献をキーとする3種類の新しい検索方法，(3) 検索した文献群に関連の高い文献群，キーワード群，著者群，および出典群に関する情報，(4) 検索した文献群だけから構成するサブデータベースの作成と利用，などを含む。

1.はじめに

研究グループや研究者レベルのパーソナルな文献データベースの作成は，効率的な研究を進める上で非常に重要である。オンライン検索を利用すれば，関連性の高い文献群は，瞬時に収集可能である。収集した文献群を分析して，特定専門分野の情報が抽出できれば，研究管理や文献検索に大変有用である。

ここでは，文献属性のReferationを利用して，特定専門分野の書誌特性を計量的に分析するパソコン用ソフトを開発する。

2.特定専門分野データベース

種々のデータベースが提供されているが，その中で，文献データベースは最も基本的なものである。オンライン検索により，利用面の発展は著しいが，文献データベースの内容や検索方法は，ここ十数年間，あまり変化していないように思われる。

情報処理機器のめざましい進展により，大規模な文献データベースのほかに，今後，フルテキストデータベースや，研究者用パーソナルデータベースの発展が予想される。

現在利用可能な大規模なオンライン用文献データベースの特徴として，以下の①～⑤が挙げられる。

① データの収録期間は，数年から十数年である。
② 文献はすべて同じウエイトで扱う。
③ データベースの構成や利用コマンドは，文献検索用に設計されている。
④ 検索した文献群に関する書誌特性を抽出するコマンドを持たない。
⑤ 重要な文献属性である文献間の関連性を表す引用文献が入力されていない。

速報性と網羅性を目的とする二次情報の作成と提供に対して過大な要求をすれば，そのコストパフォーマンスは低くなる。

そこで，小回りのきく研究者用のデータベースを作成し，上記を補う必要がある。昨年の第21回研究集会では，文献の末尾にある引用文献を考慮した，研究者用パーソナルデータベースをパソコン上で作成し，その利用可能性を明らかにした[1]。また，Referationという新しい概念を導入して文献の引用検索を試みた。

今回も，同一のデータベースを使用する。こ

*あさい　いさお　大阪府立大学工学部

こで用いる引用文献データベースは，次のような特徴を持つ。

① 文献ファイルは，著者，標題，出典などの書誌項目から構成される。

② 引用ファイルは，登録した文献だけを対象とし，文献番号で構成される。

③ 検索や分析を高速化するために，約50種類の索引ファイルを作成する。

使用したパソコンは，NEC製の主記憶640KBを持つPC-9801E，カラー高解像度ディスプレイ，1MBのフロッピーディスク装置，およびプリンタである。パソコンの操作性を高めるため，多数の索引ファイルを主記憶上に配置する。現在，テスト用として，1160編の文献群を処理している。主記憶4MBのパソコンが登場すれば，10,000編の文献群を処理することも可能である。

プログラムは，N88-日本語BASIC(86)MS-DOS版のコンパイラを使用し，試行錯誤を繰り返しながら開発した。ソートやマージなどの，高速化の必要な個所はマシン語を用いた。同一のプログラムも，コンパイラを使用すると処理速度が速くなる。1,160編の課題から，6,730語のキーワードを切り出すのに，インタプリタでは38分かかったが，コンパイラでは12分で作成できる。

プログラムの容量は，中間言語約300KBである。プログラムはジョブごとに分割し，必要に応じて主メモリに展開する。データは共通領域に蓄積し，最初に1回読み込む。画面やカラー表示などの基本的な操作方法は前回[1]と同じで，出力の追加を行った。

表1にファンクションキーの一覧を示す。入力は，文献IN,引用IN，作成の3種類からなり，出力は，検索，関連性，リストからなる。リストは，データベースの内容をいろいろな形式に表示する。16種類のリストと10種類の分布グラフを表示する。文献の表示形式を3種類(1行，圧縮，定型）用意し，任意に選択できるようにした。検索と関連性については，次章以降で詳しく述べる。

3 .Referationの利用

文献の末尾にある引用文献は，論文を読む研究者にとっては，非常に重要な情報源である。引用文献は論文の序の背景を説明するところにあらわれ，先人の研究とその論文自身の位置づけをする。従って，引用文献は文献間の関連性を表す貴重なデータであり，文献属性の一つと見なせる。

引用文献をデータとして使うことに，疑問を感じる人がいるかもしれない。著者自身が付与する引用文献には偏りがあるとか，引用のウエ

表1　ファンクションキーの一覧表

	f･1	f･2	f･3	f･4	f･5	f･6	f･7	f･8	f･9	f･10
初期画面	検索	関連性			リスト	文献IN	引用IN	作成	交換	終了
1 検索	引用SW	索引SW	ﾗﾝｸSW	操作	表示	呼出	登録	1行	印刷SW	終了
2 関連性	式00	文献	キー	著者	出典	操作	表示		印刷SW	終了
5 リスト	引用	KWIC	索引	ランク	書誌	分布		1行	印刷SW	終了
6 文献IN	作成	索引				呼出	登録		印刷SW	終了
7 引用IN	作成	索引	リスト	削除		呼出	登録		印刷SW	終了
8 作成	文献	キー	引用	分布		全部			印刷SW	終了

イトが異なるなど，問題がないわけではない。しかし，引用文献を個々に見ると，不完全なデータのように見えるかも知れないが，それらをデータベース化し，ある程度蓄積すると，まったく新しい文献検索の世界が登場する。

ある文献Aが文献Bを引用すれば，文献Bは引用文献である。これは誰でも知っている事実である。しかし，そこで思考をストップさせると，非常に重要な事実関係を見落とすことになる。このAとBとの関係を逆にしてみると，文献Bは文献Aに引用されたという関係が明らかになる。それらを集めると，文献Bがどのように利用され，発展していったかがわかる。

従って，ある文献は2種類の文献群を持つ。一つは引用文献群（References）であり，もう一つは被引用文献群（Citations）である。文献の引用文献群は，年月が経過しても増えることなく一定であり，むしろ退化するように感じる。しかし，被引用文献群はその文献の価値に応じて増大する。

第21回研究集会で提案したReferationsは，ある文献についての引用文献群，被引用文献群，そして文献自体を一つにまとめたものであり，混乱を避けるために，新造語を使用した。図1はReferationsの概念を示したものである。Referationsは，ある文献に関連の深い文献の集まりと見なせる。被引用文献群だけを用いる場合よりも，関連データが増え，よい分析結果が得られる。

Referationsの定義に，文献自体を含めたが，

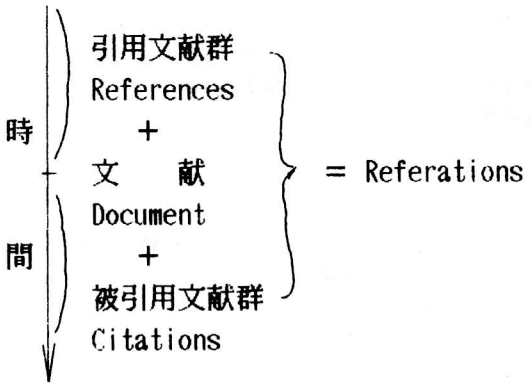


図1　Referationsの定義

それは関連性を調べる際に処理しやすいためであり，新しい考え方である。文献自体を含む場合は，文献間に直接の引用関係があれば，その度数は2となる。これに対し，文献自体を含まない場合は，直接の引用関係があっても度数はカウントされない。著者はこのReferationsの定義を使った研究を以前から行っており，高性能なパソコンの出現で，実用的なソフトが開発できた。

著者索引の支援を受けて入力した引用文献群を転置し，番号順に並べ替え，被引用文献群を作ってReferationsを作成するが，それは文献番号で構成される。番号からは，著者，標題，および出典などの書誌項目が，即時に呼び出せるため，書誌項目全部を入力したのと同じ効果がある。

計量情報学関係のデータの特徴は，① 1，2，3，と数える2～3桁のカウントデータ，② 月日ではなく，年単位のデータ，である。他の学問分野に比べると貧弱なデータ構造であるため，この分野に踏みとどまって研究を進めることは，非常に困難である。

しかし，わずかなデータでも，増幅することができればデータ量は増大する。特定専門分野の文献群に対して入力した引用文献群が，2倍強のReferationsになり，その書誌項目まで使用できれば，分析は十分可能になる。さらに，文献間のマッチングの度合いを手がかりにして，その関連性を求めれば，通常の手段では見いだせない文献間，キーワード間，著者間，および出典間の関連性が計量的に把握可能になる。

4．関連性の測度

文献Aに関連の深い文献群を求めるには，文献Aと文献Bとの関連する度合いを定義し，すべての文献ペアについて度合いを求め，それを大きい順に並べればよい。最も簡単な測度として，AとBのReferationsの中に同じ文献がないかを調べ，その数をカウントする関連度数がある。

すべての文献同士の関連度数を求めるのは，

蓄積するメモリを考えると実行不可能である。従って，関連度数は検索要求があったときだけ求めるが，この場合すべての文献ペアについての度数を調べない。ある文献に関連のある文献群は限られており，それだけを調べるようなアルゴリズムを使用する。

関連度数以外の測度として，次に示すような関連係数がある。

$$\text{関連係数}=\frac{Z}{X+Y-Z}$$

ここで，XとYは，それぞれ，文献AとBにおけるReferations数，Zは関連度数である。この測度は，XやYの大きさの影響を除いたものである。

これら2種の中間的な測度をいろいろ考察した結果，次のような方法を採用した。各度数や係数をその最大値で割って，0.0～1.0の値に規準化する。そして，規準化した度数の平方根と規準化した係数との算術平均を100倍する。これを関連指数と定義する。ある文献の関連性を求めると，これら3種の値を計算し，3回ソートし同一画面に表示する。どれを使うかは，検索者が任意に選択できる。

関連性の測度は，一つの文献とその他の文献との関連性を測るものである。しかし，検索した文献群に関連のある文献，キーワード，著者，および出典を求める場合には，文献群と文献群との関連性を測ることが必要になる。その場合には，従来の方法を拡張し，その蓄積を使用する。それはデータを行列表示して展開すれば，導き出せる。このような文献群間同士の関連性の測度はあまり定義されていない。

5．文献検索

パソコンを利用した文献検索の設計は，パソコンの特徴を生かしたものでなければならない。パソコンは従来のコンピュータにはない機能を持つ。それは，カラー，グラフィックス，そして画面制御である。情報を色別に表示すると，限られた場所へたくさんの情報を表示することが可能になる。文献は文字情報だけを扱うが，グラフィックスで画面に罫線を引けば，情報は区分して表示できる。さらに，画面制御により，情報内容に対応した位置に情報が表示できる。

パソコンを使った現在のオンライン検索においては，パソコンは単に端末機にすぎず，パソコンの特徴が十分生かされていない。ほとんどすべてのシステムは，情報をテレタイプ式に逐次送り続ける，いわゆるたれ流し式の情報提供であり，コンピュータのソフト技術としては古く，改善が必要である。

一般に検索が難しいのは，データベースにどのような内容の情報が入っているか，検索者にわからないためである。従って，検索の手がかりになるような情報はすべて画面に表示し，検索者を支援できるようにした。ここでは，13画面，22種類の情報が表示される。簡単な操作で

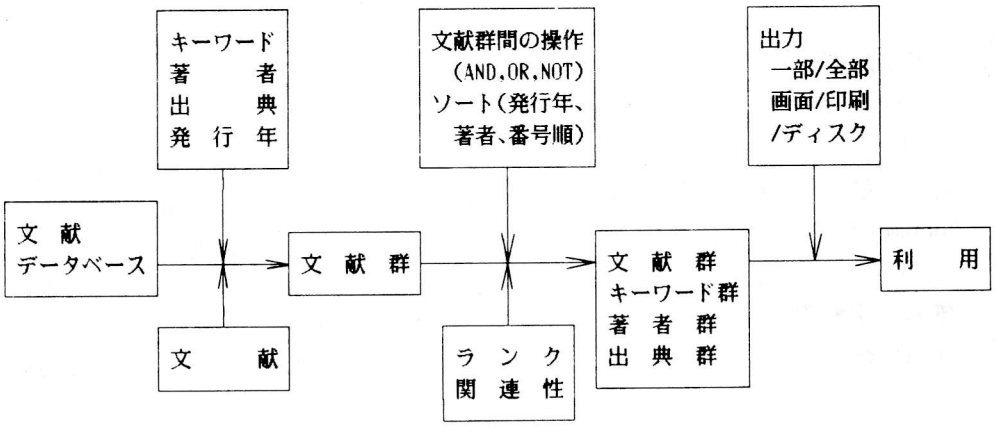


図2　文献データベースの検索から利用までの流れ

必要な画面を呼び出し，それに支援されて文献を検索する。

文献の属性にReferationsが加わったため，書誌キーによる検索以外に，文献をキーとする検索が可能になった。新しい文献データベースの検索から利用までの流れを，図2に示す。中央の線から下の部分が，新たに加わった検索の領域を示している。

文献をキーとする検索として，次の3種類の方法が考えられる。

① 文献のReferationsを検索する。

② ある文献に関連する文献を調べ，その上位n編の文献を検索する。

③ 検索した文献群に関連する文献を調べ，その上位n編の文献を検索する。

最後の③については，次節で扱う。

初期画面で，Fキー［1検索］を押すと，検索画面になり，下から2行目に新しいFキーが設定される。その[F1]～[F3]は，総計13画面の検索を支援する情報の表示スイッチである。表2に，各Fキー別の検索支援情報と種類を示した。

[F1]の引用SWは文献キーにした新しいタイプの検索，[F2]の検索SWは書誌キーによる検索，そして[F3]のランクSWはデータベースに入っている各書誌項目のランク情報と，引用ランク情報による検索である。図3に，[F1]の2番目にある文献をキーとしたReferationsの検索画面の例を示す。

検索支援情報は，画面の下側の部分に表示される。画面によって3種類の検索があり，コードm-nの検索，コードmの検索，そして1-nまでの検索がある。検索結果として，番号，文献数，検索内容が，画面の上側の部分に示される。

最下行に，各画面ごとに異なる入力指示のコメントが示される。どの画面にも共通なコマンドとして，以下のものがある。

［0］結合　：検索式のAND（＊），OR(+)，NOT(-)。

[,] 次へ　：支援情報画面を1頁分次へ。

[.] 前へ　：支援情報画面を1頁分前へ。

［=］次Fへ：Fキーが1→2→3→1へ。

表2　Fキー別の検索支援情報と種類

Fキー	検索支援情報	種類
1 引用SW	11 関連性	3
	12 Referations	1
	13 引用ランク	3
2 索引SW	21 文献番号	1
	22 キーワード	1
	23 著　者	1
	24 出　典	1
	25 発行年	1
	26 種　別	1
3 ランクSW	31 引用ランク	3
	32 キーワード	2
	33 著　者	2
	34 出　典	2

［CR］次SWへ：現在のFキー内でスクロール。

なお，ランクを除いた支援情報の表示は，自由に変更できる。例えば，CITATIONというキーワードを表示する場合，[/CIT]と省略形で入力すれば，2分探索により探し出し，即時に5個前から30個分のキーワードを表示する。

次に，他のFキーについて説明する。[F4]の操作は検索式に対する結合，ソート，および削除である。結合は検索式の論理演算をする。ソートには発行年順，第1著者順，出典順，および文献番号順がある。削除は検索式の全部削除と一部削除がある。

[F5]の表示は検索した文献群をディスプレイに表示する。この時，[F9]の印刷SWがオンになっていれば印刷する。文献の表示形式は，1行，圧縮，全部，および番号の4種類がある。[F8]のスイッチで，文献の表示形式をセットする。なお，表示数は自由にセットできる。

[F7]の登録は検索式を後で使うためにフロッピーへ保存し，[F6]の呼び出しは保存した検索式を再セットする。最後に，[F10]の終了で検索ジョブを終了する。そこから，他のジョブへ移行できる。

以上，パソコンを利用した検索の考え方，多種類の情報に支援された検索，文献をキーにした新しい検索法，そして検索ソフト全体につい

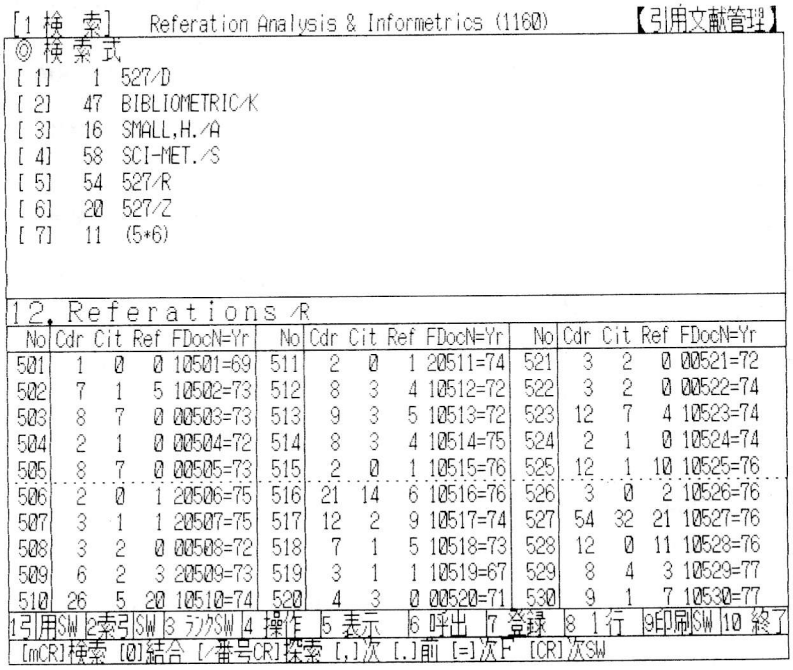


図 3　文献をキーとする Referations の検索画面例

て述べた。

6 .文献群の書誌情報

検索によって得られた検索式は，文献群から構成されている。文献群といっても，さまざまな形をしている。ただ一つだけの文献，あるキーワードを含む文献群，ある著者の書いた文献群，ある雑誌に載った文献群，新しく発行した文献群，あるいはある文献の Referations など，検索者は自由に作成することができる。

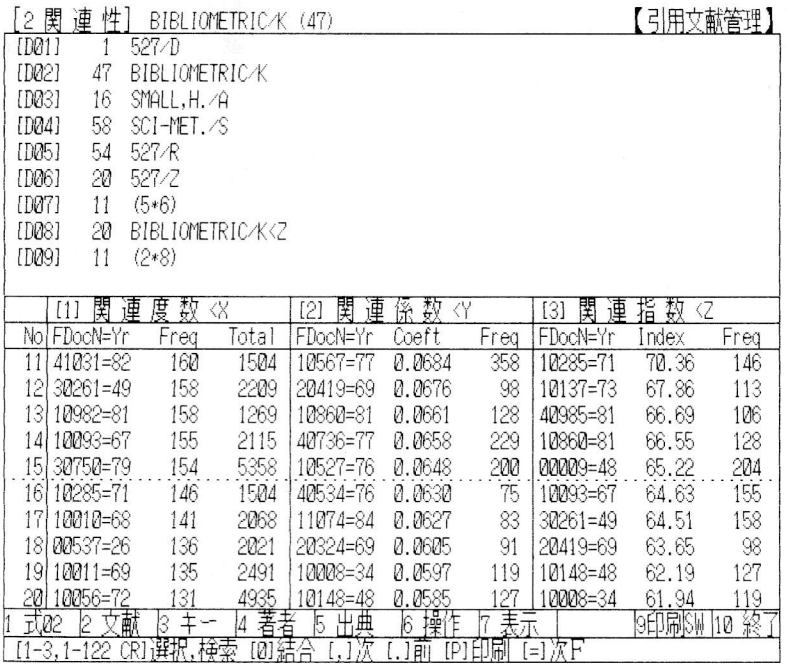
関連性情報を作成するには，まず，複数の検索式の中から一つを選択しなければならない。パソコン上では［2　関連性］ジョブを選び，F キー［1　式00］で式の選択を行う。一つの式が選択されると，その中の文献群とすべての文献との関連度数が計数される。

書誌情報は 4 種類用意されている。それは，［2　文献］，［3　キー］，［4　著者］，および［5　出典］である。その中の一つを選択すると，関連度数，関連係数，そして関連指数が計算され，大きい順に同一画面に表示される。必要なものを検索し，［6　操作］し，［7　表示］できる。続いて，他の書誌情報を求めたり，[F1] で別の検索式を選択することができる。

図 4 は，F2の文献情報とその検索画面例である。この文献情報は，検索ジョブ[11　関連性]の延長と見なせる。前章の関連性は一つの文献に対するものであったが，ここでいうのは文献群に対するものである。検索，操作，あるいは表示など，検索ジョブとまったく同じ機能が使用できる。検索結果は現在の文献検索式に追加されるばかりか，[F1]の選択対象にもなる。この機能を使えば，Referations が幾重にもループした複雑な検索が可能である。

[F3] ～[F5] における検索は，文献を検索する代わりに，それぞれ，キーワード群，著者群，および出典群を検索し，キー検索式，著者検索式，および出典検索式を作成する。各検索式内では，論理演算，ABC 順のソート検索式の削除，および表示が行える。図 5 は，著者情報とその検索画面例で，キーや出典の場合も同じ形式の画面である。

このように，Referations を使うと，今まで



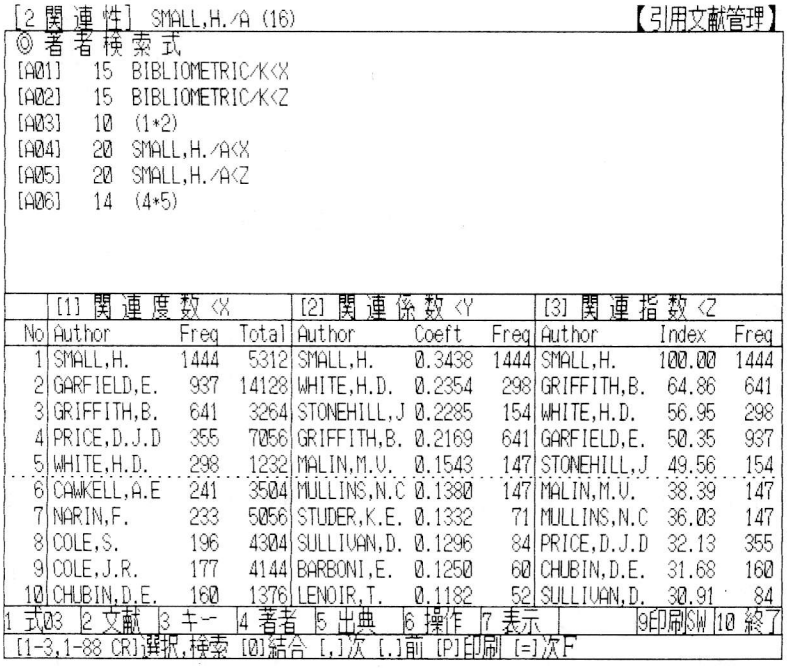
[2 関連性] BIBLIOMETRIC/K (47) 【引用文献管理】

[D01] 1 527/D
[D02] 47 BIBLIOMETRIC/K
[D03] 16 SMALL,H./A
[D04] 58 SCI-MET./S
[D05] 54 527/R
[D06] 20 527/Z
[D07] 11 (5*6)
[D08] 20 BIBLIOMETRIC/K<Z
[D09] 11 (2*8)

	[1] 関連度数 <X			[2] 関連係数 <Y			[3] 関連指数 <Z		
No	FDocN=Yr	Freq	Total	FDocN=Yr	Coeft	Freq	FDocN=Yr	Index	Freq
11	41031=82	160	1504	10567=77	0.0684	358	10285=71	70.36	146
12	30261=49	158	2209	20419=69	0.0676	98	10137=73	67.86	113
13	10982=81	158	1269	10860=81	0.0661	128	40985=81	66.69	106
14	10093=67	155	2115	40736=77	0.0658	229	10860=81	66.55	128
15	30750=79	154	5358	10527=76	0.0648	200	00009=48	65.22	204
16	10285=71	146	1504	40534=76	0.0630	75	10093=67	64.63	155
17	10010=68	141	2068	11074=84	0.0627	83	30261=49	64.51	158
18	00537=26	136	2021	20324=69	0.0605	91	20419=69	63.65	98
19	10011=69	135	2491	10008=34	0.0597	119	10148=48	62.19	127
20	10056=72	131	4935	10148=48	0.0585	127	10008=34	61.94	119

1 式02 |2 文献 |3 キー |4 著者 |5 出典 |6 操作 |7 表示 |9印刷SW |10 終了
[1-3,1-122 CR]選択,検索 [0]結合 [,]次 [.]前 [P]印刷 [=]次F

図 4　文献情報とその検索画面例



[2 関連性] SMALL,H./A (16) 【引用文献管理】

◎著者検索式
[A01] 15 BIBLIOMETRIC/K<X
[A02] 15 BIBLIOMETRIC/K<Z
[A03] 10 (1*2)
[A04] 20 SMALL,H./A<X
[A05] 20 SMALL,H./A<Z
[A06] 14 (4*5)

	[1] 関連度数 <X			[2] 関連係数 <Y			[3] 関連指数 <Z		
No	Author	Freq	Total	Author	Coeft	Freq	Author	Index	Freq
1	SMALL,H.	1444	5312	SMALL,H.	0.3438	1444	SMALL,H.	100.00	1444
2	GARFIELD,E.	937	14128	WHITE,H.D.	0.2354	298	GRIFFITH,B.	64.86	641
3	GRIFFITH,B.	641	3264	STONEHILL,J	0.2285	154	WHITE,H.D.	56.95	298
4	PRICE,D.J.D	355	7056	GRIFFITH,B.	0.2169	641	GARFIELD,E.	50.35	937
5	WHITE,H.D.	298	1232	MALIN,M.V.	0.1543	147	STONEHILL,J	49.56	154
6	CAWKELL,A.E	241	3504	MULLINS,N.C	0.1380	147	MALIN,M.V.	38.39	147
7	NARIN,F.	233	5056	STUDER,K.E.	0.1332	71	MULLINS,N.C	36.03	147
8	COLE,S.	196	4304	SULLIVAN,D.	0.1296	84	PRICE,D.J.D	32.13	355
9	COLE,J.R.	177	4144	BARBONI,E.	0.1250	60	CHUBIN,D.E.	31.68	160
10	CHUBIN,D.E.	160	1376	LENOIR,T.	0.1182	52	SULLIVAN,D.	30.91	84

1 式03 |2 文献 |3 キー |4 著者 |5 出典 |6 操作 |7 表示 |9印刷SW |10 終了
[1-3,1-88 CR]選択,検索 [0]結合 [,]次 [.]前 [P]印刷 [=]次F

図 5　著者情報とその検索画面例

はっきりしなかった関係が見えるようになる。文献を単に検索するだけでなく，文献，キーワード，著者，あるいは出典情報が作成できる。

7．サブデータベース

検索，関連性，リストなどは，データベース上のすべてのデータを対象としたものである。しかし，検索した文献群だけに関する，新しいサブデータベースを作り，検索，関連性，リストなどができれば，数倍の利用価値を持つようになる。ちょうど，オンライン検索のプライベートファイルに当たるものである。

このサブデータベースの作成は，初期画面の［9　分割］を用いる。このジョブに入ると，すべての索引ファイルが書き換えられる。もちろん，Referations データも，対象とする文献群だけから構成される。文献数にもよるが，処理時間は数分かかる。途中で，別の文献群や全データベースに戻ることもできる。なお，文献数が少ないと分布の精度が落ちるため，リストの分布分析だけクローズする。

8．おわりに

文献属性としてあまり注目されていなかった引用文献から得られる Referations を，文献データベースに組み込むことにより，文献をキーとする新しい文献検索手法と，文献群の関連性情報の抽出が可能になった。得られる書誌情報は，研究者の頭脳労働の一部を代替する一種の専門家情報であり，質の高い効率的な研究に無くてはならないものである。

人工知能の研究では，知識データベースをいかに構築するか，どのようにアルゴリズムで知識を導くかなど，まだ解決すべき問題が多い。これに対して，ここで開発したシステムは，ある種の専門家情報を非常に安価に構築でき，すぐに実行可能である。

検索支援情報の提供，検索した文献群に関する情報，そしてサブデータベースの作成などは，既存のオンライン文献データベースにも，ぜひ取り入れて欲しい機能である。

参照文献

1）浅井勇夫．パソコンによる引用文献データベースの開発．第21回情報科学技術研究集会発表論文集．21-31(1984)

質疑応答

質問　三木英生（座長　㈱タクマ）―以下2問―
広範囲の社内 DB と中小企業の小規模の特定分野の DB のどちらに適しているのか。

回答　このシステムは，研究者レベルでの文献数1,000～5,000程度のものに適している。分野の広狭より，特定分野の深い情報が処理できる。研究者個人あるいは研究グループ内での利用を念頭に置いており，Public なものではない。

質問　Reference と Citation の関係で，時代が進むと文献の重複登録が予想されるが，その対策は配慮されているのか。

回答　入力の際に Reference を全部登録しているわけではない。文献ファイルにあるものだけを対象としており，Close な使い方をしている。
ISI 社の SCI（Science Citation Index）等では，Reference と Citation がうまく関連付けられない。このシステムによって関連付けることができ，文献量の数倍増が期待できる。