The Application of Micro-Computers in Information,
Documentation and Libraries
K.-D. Lehmann and H. Strohl-Goebel (eds.)
Elsevier Science Publishers B.V. (North-Holland)
© DGD, 1987

THE STRUCTURE AND USAGE OF A "REFERATION" DATABASE ON MICROCOMPUTERS FOR PERSONAL DOCUMENT MANAGEMENT

Asai, Isao

University of Osaka Prefecture
Department of Industrial Engineering
Sakai, Osaka 591, Japan

In order to extract professional information from document collection in a specific topic, a software package on microcomputers is developed. The characteristics is to make and use "referation" composed of citations (citing documents), references (cited documents) and document itself. This paper gives a brief outline of user-friendly software designed for scientists and librarians. The process of producing the relevant information between bibliographic items suggests the possibility of referation analysis as a new field of information science.

1. INTRODUCTION

There is a trend toward more numbers of documents owned by scientists and researchers with increasing online search. It is the present state that the personal document collection is not computerized. Using today's microcomputers with lower cost, larger memory, and higher performance, document management system could be easily attained. This paper gives a brief outline of user-friendly software package for scientists and librarians.

In designing of personal document database, I think it is very important to use the reference data at the end of document. Each reference data has the same basic bibliographic items as document itself. If you notice references to be very valuable data, you would give up forming computer readable database. Garfield has been found a solution to entry the reference data about two decades ago[1]. In this paper, I will attempt a new approach to make and use reference data on microcomputers. Especially, it is important to introduce a new concept "referation" and to produce professional information about bibliographic items in a specific topic.

2. SYSTEMS REQUIREMENTS

2.1. Microcomputers

The rapid progress of micro's hardware and software has affected to all of the fields. Various microcomputers have appeared to and disappeared from our presence for a few years. Undoubtedly, new machine with higher performance will set back our present efforts. It is difficult to specify systems requirements in a micro's era. Nevertheless, I shall set to the following equipments.

(1) micro processor	: 16-bit (Intel 8086 compatible, 10 MHz).
(2) main memory	: 640 Kbytes.
(3) monitor	: 8 colors, 640 x 400 pixels.
(4) floppy disk	: 1 Mbytes x 2.
(5) printer.	

This system, NEC PC-9801VM2, is very popular in Japan. Of course, Japanese character is supported. Total hardware cost is about 3,000 dollars(March 1986).

2.2. Software

Software package runs the MS-DOS operating system. And it is programmed by BASIC compiler. A part of program, such as sorting, mergining and boolean operation, uses assembly language(Intel 8086) for rapid processing. A total of statements is about 18,000 steps(350 Kbytes).

As the important functions of micro-based software, the colored text and dot, charts and graphics, and the control of cursor's location are considered. These functions are useful to display clearly various types of information on screen with 80 x 25 characters. By the selection of function key or menu-driven command, our job proceeds user-friendly. Table 1 shows the list of function keys. Key number six, seven and eight concerns the entry of data. And one, two and five relates to the usage.

TABLE 1 List of Function Keys

	f-1	f-2	f-3	f-4	f-5	f-6	f-7	f-8	f-9	f-10
Main	Searc	Assoc			List	DocIN	RefIN	Index	Exchg	Exit
1 Searc	Refrt	Index	Rank	Opert	Disp	Load	Save	Line	Print	Exit
2 Assoc	DOO	Docmt	Keywd	Authr	Sourc	Opert	Disp		Print	Exit
5 List	Refrt	KWIC	Index	Rank	Name	Graph		Line	Print	Exit
6 DocIN	Input	Index				Load	Save		Print	Exit
7 RefIN	Input	Index	List	Delet		Load	Save		Print	Exit
8 Index	Bibli	Keywd	Refer	Distr		All				

3. STRUCTURE OF DATABASE

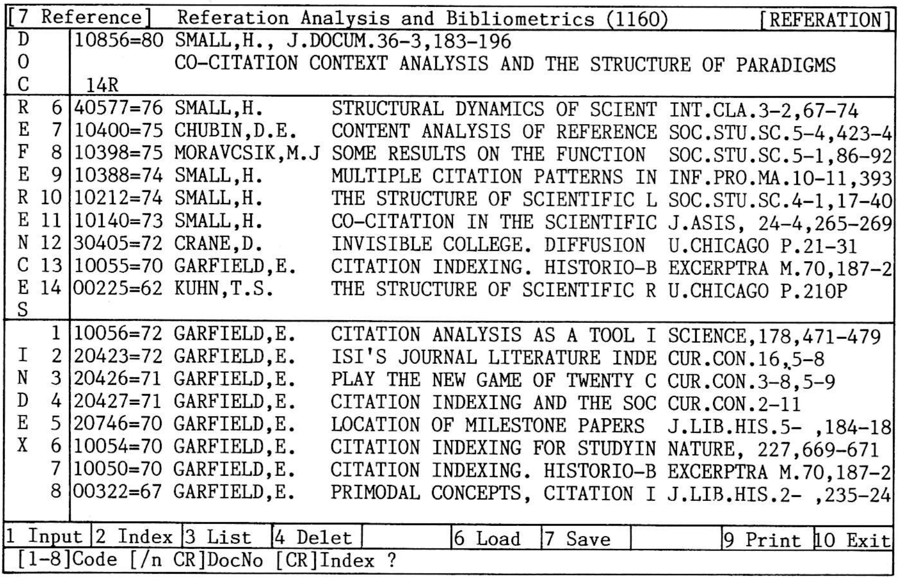
3.1. Document File

Document file describes the static features of document. One record is made of basic bibliographic items: form(1 column), document number(4), three authors(20 x 3), title(128), source(35), publication year(4), and others(24). The record with the length of 256 bytes is saved to floppy diskett as random file. The screen for the entry of document data is divided into two parts. The upper part, input screen, shows the content of record on a given document. The lower part, index screen, displays bibliographic index to assist data entry. Five indexes are ready: document number, author, title, source and publication year. By assisting of these indexes, document data can be easily entried.

3.2. Reference File

A very easy method for the entry of reference data is developed. Figure 1 illustrates an example of screen for the entry of reference data. The screen is divided into three parts. The upper part, document screen, shows the content of record on a given document. The middle part, reference screen, presents the reference data of above document. And the lower part, index screen, displays author's index to assist data entry. The entry of reference data takes two steps. First, by assisting of author's index, he and she searches the same document from document file. And second, by inputing the corresponding code, reference data is cataloged and displayed on reference screen.

It is important matter that reference file is composed of document number and only the document cataloged to document file. From document number, the bibliographic items of document can be moved quickly. This is useful to reduce the entry work and memory space. Reference data entries only the related document of references. This point differs from Science Citation Index which can survey the relationship between fields. But only the cataloged document is enough to introduce the internal relationship on a specific topic.



```
[7 Reference]  Referation Analysis and Bibliometrics (1160)         [REFERATION]
 D    |10856=80 SMALL,H., J.DOCUM.36-3,183-196
 O    |         CO-CITATION CONTEXT ANALYSIS AND THE STRUCTURE OF PARADIGMS
 C    | 14R
 R  6 |40577=76 SMALL,H.        STRUCTURAL DYNAMICS OF SCIENT INT.CLA.3-2,67-74
 E  7 |10400=75 CHUBIN,D.E.     CONTENT ANALYSIS OF REFERENCE SOC.STU.SC.5-4,423-4
 F  8 |10398=75 MORAVCSIK,M.J SOME RESULTS ON THE FUNCTION  SOC.STU.SC.5-1,86-92
 E  9 |10388=74 SMALL,H.        MULTIPLE CITATION PATTERNS IN INF.PRO.MA.10-11,393
 R 10 |10212=74 SMALL,H.        THE STRUCTURE OF SCIENTIFIC L SOC.STU.SC.4-1,17-40
 E 11 |10140=73 SMALL,H.        CO-CITATION IN THE SCIENTIFIC J.ASIS, 24-4,265-269
 N 12 |30405=72 CRANE,D.        INVISIBLE COLLEGE. DIFFUSION  U.CHICAGO P.21-31
 C 13 |10055=70 GARFIELD,E.     CITATION INDEXING. HISTORIO-B EXCERPTRA M.70,187-2
 E 14 |00225=62 KUHN,T.S.       THE STRUCTURE OF SCIENTIFIC R U.CHICAGO P.210P
 S    |
    1 |10056=72 GARFIELD,E.     CITATION ANALYSIS AS A TOOL I SCIENCE,178,471-479
 I  2 |20423=72 GARFIELD,E.     ISI'S JOURNAL LITERATURE INDE CUR.CON.16,5-8
 N  3 |20426=71 GARFIELD,E.     PLAY THE NEW GAME OF TWENTY C CUR.CON.3-8,5-9
 D  4 |20427=71 GARFIELD,E.     CITATION INDEXING AND THE SOC CUR.CON.2-11
 E  5 |20746=70 GARFIELD,E.     LOCATION OF MILESTONE PAPERS  J.LIB.HIS.5- ,184-18
 X  6 |10054=70 GARFIELD,E.     CITATION INDEXING FOR STUDYIN NATURE, 227,669-671
    7 |10050=70 GARFIELD,E.     CITATION INDEXING. HISTORIO-B EXCERPTRA M.70,187-2
    8 |00322=67 GARFIELD,E.     PRIMODAL CONCEPTS, CITATION I J.LIB.HIS.2- ,235-24

1 Input|2 Index|3 List |4 Delet|      |6 Load |7 Save |      |9 Print|10 Exit
 [1-8]Code [/n CR]DocNo [CR]Index ?
```

FIGURE 1 An Example of Screen for the Entry of Reference Data

3.3. Definition of Referations

In the field of information science, references indicate cited documents, and citations indicate citing documents. Under computer algorithm, citation file is obtained by inverting reference file.

Reference analysis uses only reference data. Kessler presents the relevant relationship between documents using bibliographic coupling[2]. In contrast to reference analysis, citation analysis uses only citation data. Small proposes co-citation as a measure of association [3]. Citation analysis used SCI/SSCI databases has been studied actively.

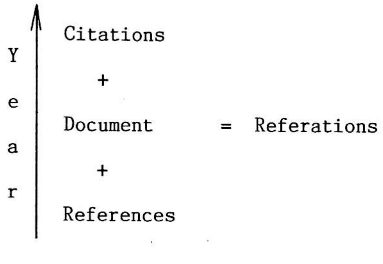


FIGURE 2
Definition of Referations

In a given document, it is possible to integrate citations, document itself, and references. This new concept is named "referations"[4]. Figure 2 shows the definition of referations. Under computer algorithm, referation file is obtained by mergining three files (citation, document and reference) made of document number.

3.4. Index Files

For the sake of faster and more efficient man-machine communication, about fifty files are produced before usage. Once we made of these files and saved to floppy diskett, end-user would be able to run the job of usage at any times. For 1,160 documents, the execution time to make index files takes about twelve minutes using RAM disk. Of this time, it takes about five minutes to extract and sort 6,730 title keywords from 1,160 titles.

4. USAGE OF DATABASE

4.1. Bibliographic Lists

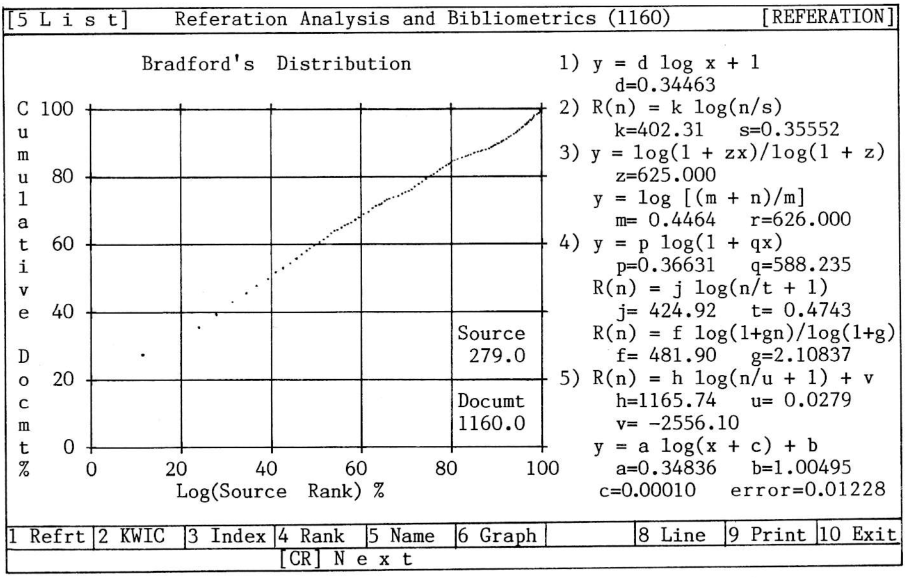
This section deals with many kinds of list on the content of database. The followings can be called by very simple command on screen or printer.

(1) Referation List : Referations to a given document are displayed. Looking at this list, the chronological aspects around document become clearly.
(2) KWIC Index : KWIC index at any title keyword which you need is obtained. There are three formats to show the corresponding document. In the case of all items, not one line, KWOC index is listed.
(3) Index List : There are six types of index list : document number, author, source, publication year, form, and title. Because of a great volumes of listing, the starting search key is possible to select freely.
(4) Rank List : This lists four types of count ranking : referation, keyword, author, and source.
(5) Name List : Keyword's name, author's name, and source's name are tabulated. The print-out sheet becomes very compact.
(6) Count List : Three types of list are ready : referation, publication year, and form.

Each list of referation, KWIC and index has three types of format to describe the content of document. A total of lists, therefore, is thirty four. As these lists contain the count information of citations and references, it can be used to an indicator of important document in closed topic.

4.2. Bibliometric Distributions

Ten kinds of bibliometric graph with theoritical models are drawn. As the parameter of each model is estimated at the job of index files in 3.4., the distribution is very quickly plotted at color. Figure 3 shows the screen of Bradford's distribution. The left side is the plot of Bradford's



```
1) y = d log x + 1
      d=0.34463
2) R(n) = k log(n/s)
      k=402.31    s=0.35552
3) y = log(1 + zx)/log(1 + z)
      z=625.000
      y = log [(m + n)/m]
      m= 0.4464    r=626.000
4) y = p log(1 + qx)
      p=0.36631    q=588.235
   R(n) = j log(n/t + 1)
      j= 424.92    t= 0.4743
   R(n) = f log(1+gn)/log(1+g)
      f= 481.90    g=2.10837
5) R(n) = h log(n/u + 1) + v
      h=1165.74    u= 0.0279
      v= -2556.10
   y = a log(x + c) + b
      a=0.34836    b=1.00495
      c=0.00010    error=0.01228
```

FIGURE 3 The Screen of Bradford's Distribution

distribution. The droop section is drawn clearly. The right side shows five types of theoritical models and the value of parameter estimated. For some details of Bradford's models, please refer to my paper[5]. It is considered three theoritical models on Lotka's distribution : negative binomial, borel-turner, and cumulative advantage distribution[6]. And the value of parameter is estimated and tested. Table 2 shows the x-axis and y-axis of each distribution.

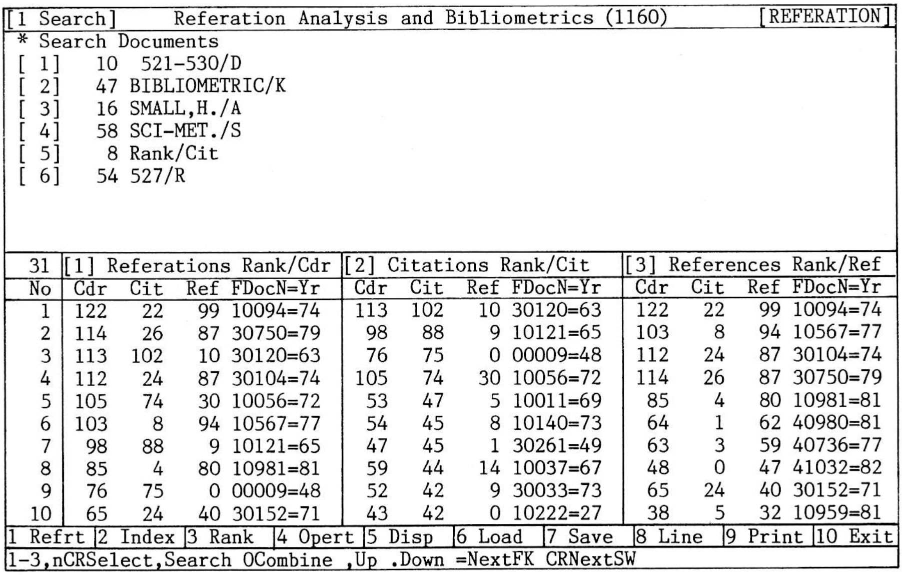
TABLE 2
Ten kinds of Bibliometric Distributions

	Distribution	x-axis	y-axis
1	Bradford	log(souce r)	cum.docmt
2	Zipf	log(keywd r)	document
3	Lotka	log(documnt)	log(freq)
4	Lotka	log(refrt)	log(freq)
5	Characteristic	source rank	cum.docmt
6	Characteristic	keywrd rank	cum.docmt
7	Characteristic	author rank	cum.docmt
8	Characteristic	refrt rank	cum.docmt
9	Growth	year	document
10	Life	age	document

4.3. Traditional Search

Using bibliographic items memorized to searcher, we usually retrieve the set of documents from databases. The difficulty of searching is due to rely on the memory of scientists and libralians. Therefore, this software is designed to display the content of database as many as possible. Six indexes (document number, keyword, author, source, publication year, and form) and four ranks (referation, author, keyword, and source) are possible to display on the screen using simple command. Through the selection of these indexes and ranks, we are able to search the set of documents.

Figure 4 shows an example of search based on referation's rank. The screen divided into two parts. The upper part displays the results of search. The lower part shows the list of referation's rank. It means that search number five consists of eight documents with citations rank higher than eight. Of course, Boolean algebra (AND, OR, and NOT) to any search documents can be operated. And also, three kinds of format to describe the items of document are ready on screen or printer.



```
[1 Search]      Referation Analysis and Bibliometrics (1160)        [REFERATION]
 * Search Documents
 [ 1]   10  521-530/D
 [ 2]   47  BIBLIOMETRIC/K
 [ 3]   16  SMALL,H./A
 [ 4]   58  SCI-MET./S
 [ 5]    8  Rank/Cit
 [ 6]   54  527/R

 31 |[1] Referations Rank/Cdr |[2] Citations Rank/Cit  |[3] References Rank/Ref
 No | Cdr  Cit  Ref FDocN=Yr | Cdr  Cit  Ref FDocN=Yr | Cdr  Cit  Ref FDocN=Yr
  1 | 122   22   99 10094=74 | 113  102   10 30120=63 | 122   22   99 10094=74
  2 | 114   26   87 30750=79 |  98   88    9 10121=65 | 103    8   94 10567=77
  3 | 113  102   10 30120=63 |  76   75    0 00009=48 | 112   24   87 30104=74
  4 | 112   24   87 30104=74 | 105   74   30 10056=72 | 114   26   87 30750=79
  5 | 105   74   30 10056=72 |  53   47    5 10011=69 |  85    4   80 10981=81
  6 | 103    8   94 10567=77 |  54   45    8 10140=73 |  64    1   62 40980=81
  7 |  98   88    9 10121=65 |  47   45    1 30261=49 |  63    3   59 40736=77
  8 |  85    4   80 10981=81 |  59   44   14 10037=67 |  48    0   47 41032=82
  9 |  76   75    0 00009=48 |  52   42    9 30033=73 |  65   24   40 30152=71
 10 |  65   24   40 30152=71 |  43   42    0 10222=27 |  38    5   32 10959=81
1 Refrt |2 Index |3 Rank |4 Opert |5 Disp |6 Load |7 Save |8 Line |9 Print |10 Exit
1-3,nCRSelect,Search 0Combine ,Up .Down =NextFK CRNextSW
```

FIGURE 4 An example of Search based on Referation's Rank

4.4. Referation Search

TABLE 3 Six Types of Referation Search

No	search key	relation	search results
1	document number	referations	set of documents
2	document number	association	set of documents
3	set of documents	association	set of documents
4	set of documents	association	set of keywords
5	set of documents	association	set of authors
6	set of documents	association	set of sources

This presents a new type of information retrieval system using referation data. Table 3 shows a concept of referation search. It devides into two categories with search key : a given document number and the set of documents. The former category has two types of search. One type is to search the referations of a given document. For example, search number six in figure 4 is made of 54 documents which is references of document number 527. The other type is to search the set of documents associated with a given document. The later category has four types of search : the set of documents, keywords, authors, and sources associated with the set of documents retrieved by user. Next section gives some details on the association measures.

4.5. Association Measures

For classifying and evaluating the item of document and/or document itself, how to quantify the relationship between items is very important. As the basic measure, the matching count between items is used. For example, reference analysis measures bibliographic coupling between two documents based on references, and then citation analysis counts co-citation based on citations. In the case of referations, the same matching method is adopted.

Figure 5 represents the association count between two documents(860 and 137). The left side is the referations of document 860 in 1981. The right side is the referations of document 137 in 1973. By observing of the figure 5 in details, document(1081) is co-citation, and documents(70, 11, 69, 260, 93, and 81) are bibliographic coupling. But documents(860, 525, 119, and 137) do not belong to any categories. Direct link is found to 860 and 137. This is indeed the reason that document itself was contained within referations. Therefore, it is defined that association count is a total of common documents between referations of two documents.

```
Cit.   1:11078=84
  ↑    2:41110=83
  |    3:11081=83 ---  1:11081=83
  └    4:10978=81
Doc.   5:10860=81 ===  2:10860=81
  ┌    6:10597=78      3:10808=78
  ↓    7:10607=78
       8:10605=78
Ref.   9:10592=77      4:10567=77
      10:10586=77
      11:10527=76                  Cit.
      12:10525=76 ---  5:10525=76   ↑
      13:10119=75 ---  6:10119=75  ─┘
      14:10090=73
      15:10137=73 ===  7:10137=73  Doc.
      16:10156=73      8:10144=71  ─┐
      17:10285=71      9:10136=71   ↓
      18:10070=70 --- 10:10070=70
      19:10011=69 --- 11:10011=69  Ref.
      20:10069=69 --- 12:10069=69
      21:20166=69
      22:10260=69 --- 13:10260=69
                      14:10010=68
      23:10093=67 --- 15:10093=67
      24:20319=67     16:10296=67
      25:10035=62     17:10066=66
      26:10081=60 --- 18:10081=60
      27:10705=55     19:30261=49
      28:10008=34     20:00009=48
```

FIGURE 5
Association Count Between 860 and 137

Bibliographic coupling is useful to latest two documents, and co-citation count is adaptive to relatively older two ones. But association count is free to publication year of document. I have developed a fasten computer algorithm for obtaining all pairs of association count between a given document and all of the other documents[7]. It takes only a few seconds to measure and sort the association count of a given document.

(1) Count (2) Coefficient (3) Ratio

$$x = c \qquad y = \frac{c}{a + b - c} \qquad z = \frac{\sqrt{x / x_{max}} + \sqrt{y / y_{max}}}{2}$$

FIGURE 6 Three Types of Association Measure

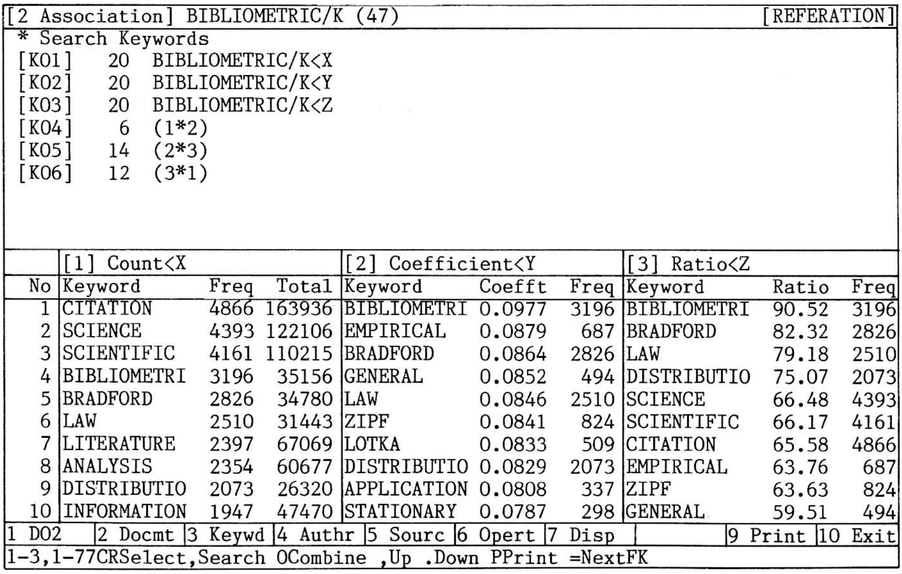
Figure 6 shows three types of association measure, which a and b is each number of references in document A and B, and c is the association count between A and B. The first type is association count itself. The second type is association coefficient normalized by a total of two referations. And the third type is association ratio adjusted by the association count and coefficient.

Many measures of association has been studied[8]. Which of measure is better may be different from every database used. Therefore, three measures of association are displayed at the same screen in descending order, that is, in relevant order. Searcher is able to select any measure and any number of documents which he need.

4.6. Association Information

The set of documents retrieved by traditional search and referation search (type 1, 2, and 3) can be utilized as search key for association information. In this case, a measure of association between two sets, not two documents, uses the cumulative counts to all documents of the set.

Figure 7 represents an example of screen on referation search. The set of documents is made of 47 documents with keyword "bibliometric". The lower side shows three sets of keywords related to "bibliometric". We find some of the well known keywords such as Bradford. It is noticeable that these keywords is automatically produced and weighted. Furthermore, the sets of documents, authors and sources related to "bibliometric" can be simply obtained.



```
[2 Association] BIBLIOMETRIC/K (47)                          [REFERATION]
 * Search Keywords
 [K01]   20  BIBLIOMETRIC/K<X
 [K02]   20  BIBLIOMETRIC/K<Y
 [K03]   20  BIBLIOMETRIC/K<Z
 [K04]    6  (1*2)
 [K05]   14  (2*3)
 [K06]   12  (3*1)
```

	[1] Count<X			[2] Coefficient<Y			[3] Ratio<Z		
No	Keyword	Freq	Total	Keyword	Coefft	Freq	Keyword	Ratio	Freq
1	CITATION	4866	163936	BIBLIOMETRI	0.0977	3196	BIBLIOMETRI	90.52	3196
2	SCIENCE	4393	122106	EMPIRICAL	0.0879	687	BRADFORD	82.32	2826
3	SCIENTIFIC	4161	110215	BRADFORD	0.0864	2826	LAW	79.18	2510
4	BIBLIOMETRI	3196	35156	GENERAL	0.0852	494	DISTRIBUTIO	75.07	2073
5	BRADFORD	2826	34780	LAW	0.0846	2510	SCIENCE	66.48	4393
6	LAW	2510	31443	ZIPF	0.0841	824	SCIENTIFIC	66.17	4161
7	LITERATURE	2397	67069	LOTKA	0.0833	509	CITATION	65.58	4866
8	ANALYSIS	2354	60677	DISTRIBUTIO	0.0829	2073	EMPIRICAL	63.76	687
9	DISTRIBUTIO	2073	26320	APPLICATION	0.0808	337	ZIPF	63.63	824
10	INFORMATION	1947	47470	STATIONARY	0.0787	298	GENERAL	59.51	494

```
1 DO2 |2 Docmt |3 Keywd |4 Authr |5 Sourc |6 Opert |7 Disp |  |9 Print |10 Exit
1-3,1-77CRSelect,Search OCombine ,Up .Down PPrint =NextFK
```

FIGURE 7 An example of Screen on Referation Search

5. CONCLUSION

I presented a new type of information retrieval system using referation data. Conventional search retrieves the set of no weighted and only documents. But referation search produces the set of weighted documents, keywords, authors and sources. It is considered that the weighted items represent some kinds of professional information. Although this approach differs from one of expert system and knowledge databases, this micro-based system is very simple, lower cost, and feasible.

Because of dealing with all pairs of documents without regard for publication year, the concept of referation will affect to the study of citation analysis using only citations. This paper indicates the possibility of referation analysis as a new field of information science. It is desirable that each scientist would manage personal referation database to promote the activity of research effectively and efficiently.

REFERENCES

[1] Garfield, E., Citation Indexes for Science, Science 122-7 (1955) 108-111.

[2] Kessler, M.M., Bibliographic Coupling Between Scientific Papers, American Documentation 14-1 (1963) 10-25.

[3] Small, H., Co-citation in the Scientific Literature : A New Measure of the Relationship Between Two Documents, Journal of the American Society for Information Science 24-4 (1973) 265-269.

[4] Asai, I., Development of a "Referation" Database Using Microcomputers, Proceedings of the 21th Annual Meeting on Information Science and Technology (October 1984) 21-31, (in Japanese).

[5] Asai, I., A General Formulation of Bradford's Distribution : The Graph-Oriented Approach, Journal of the American Society for Information Science 32-2 (1981) 113-119.

[6] Rao, I.K.R., The Distribution of Scientific Productivity and Social Change, Journal of the American Society for Information Science 31-2(1980) 111-122.

[7] Asai, I., Analysis of Bibliographic Information on a Specific Topic Based on "Referation" Relationship, Proceedings of the 22nd Annual Meeting on Information Science and Technology (October 1985) 135-142, (in Japanese).

[8] Rijsbergen, C.J.van, Information Retrieval (Butterworths, London, 1975)

REPRINTED FROM:

THE APPLICATION OF MICRO-COMPUTERS IN INFORMATION, DOCUMENTATION AND LIBRARIES

Proceedings of the Second International Conference on
the Application of Micro-Computers in
Information, Documentation and Libraries
Baden-Baden, F.R.G., 17–21 March, 1986

edited by

Klaus-Dieter LEHMANN
and
Hilde STROHL-GOEBEL

Deutsche Gesellschaft für Dokumentation e.V. (DGD)
Frankfurt am Main, F.R.G.

N·H
P C

1987

NORTH-HOLLAND
AMSTERDAM · NEW YORK · OXFORD · TOKYO