"REFERATION" DATABASE FOR DOCUMENT INFORMATION ANALYSIS

Isao Asai
University of Osaka Prefecture, Osaka, Japan

Abstract. Referations in a given document is defined as the join of citations (citing documents), references (cited documents) and document itself. It represents the relationships between a given document and the others, and produces the more valuable measures between documents than references or citations. Referation file of document collection in a specific topic can be easily obtained in microcomputers using reference data. Referation database, that is, document file with referation file, is a kind of knowledge one structured by relations. This paper gives a brief outline of some functions based on referations for the content and structure of bibliographic information.

INTRODUCTION

It is well-known that document database with references, introduced by Garfield[1], has been contributed to the progress of information science and technology. The characteristics of the database is that the attribute of document has references(set of cited documents) and citations(set of citing documents, inverted file of references) instead of descriptors and abstract. At present, we are able to access SCISEARCH and SOCIAL SCISEARCH in online.

To find out the other usage of SCI database, a decade had elapsed since the first issue of the Science Citation Index. Co-citation proposed by Small[2] is a measure for automatically finding out the subject relatedness between documents that share one or more citations. Many of citation analysis using SCI database has been done for studying the science of science[3].

For the influence of valuable citation index, references had little part in SCI database. However, an analysis was recently done using references in SCI database[4]. It is based on similarity measure of documents, that is, bibliographic coupling by Kessler[5]. The revival of bibliographic coupling is quite proper under the nature of references.

On the usage of SCI database, the following four points are summarized:

1) citation search
   cumulative citations of first author,
2) reference search
3) citation analysis
   co-citation between documents passed over five years,
4) reference analysis
   bibliographic coupling between documents less than five years.

The data structure of references in SCI database is made of first author, abbreviated journal name, and volume, page and year. The data structure of citations inverted is the same as that of references. This must be considered as an unavoidable consequences of the expanding amount of documents.

However, the usage of SCI database has some limitations for the characteristics of the data structure. There are, for example,

1) difficulty of identification of document,
2) insufficiency of document information,
3) separation of references and citations.

In particular, the realization of identifying document will enhance the value of SCI database.

This paper deals with a small scale of database with references. It mainly contains,

1) representation of reference data,
2) concept of referations,
3) new information retrieval system based on referations,
4) some functions of referation database.

Using microcomputers with higher performance, lower cost, and larger memory, I have been developed an user-friendly system designed for scientists and librarians[6-8]. This gives an outline of this system and some functions of referation database.

REPRESENTATION OF REFERENCE DATA

It is very difficult to convert the reference data at the end of document into computer readable format. Although we are able to entry all bibliographic items of references, the

redundancy of data increases, and the problem of identification becomes impossible to solve. For extracting the professional information from document database, it is very important to identify document.

SCI database think out of decreasing the entry of reference data. The length of data is 30-40 bytes. It is considered to be a special coding system. Each code is marked when new document is collected. It is noticed that the code is not known in advance. Because of long bytes of code, it is difficult to identify and operate document.

If each document has a fixed code, many problems described previously will be solved. But it is impossible to mark an code corresponding to each document in all of the world. A feasible approach of coding is to restrict within documents of narrow limits in a specific field.

The procedure for coding document in a specific topic is as follows:

1) collection of documents in a specific topic,
2) coding of document,
3) selection of documents related to the topic from references,
4) coding of undefined document,
5) representation of reference data by code.

Therefore, a reference file is made under the following conditions:

1) reference file is composed of document code,
2) reference file is only the document cataloged to document file.

As the bibliographic items corresponding to document code are stored to document file, the entry of document code equals to input all bibliographic items.

A very easy method for the entry of reference data is developed. It takes two steps of procedure. First step, we search and display the list of document from document file by assisting of author's index. And second step, we catalog reference data and display the bibliographic items by inputing the corresponding code. These are executed on multi screen.

Under computer algorithm, citation file is obtained by inverting reference file. As reference data is made of document code, this inverted procedure is very quickly.

DEFINITION OF REFERATIONS

In a given document, it is possible to integrate citations, document itself, and references. This new concept is named "referations"[6,8]. Under computer algorithm, referation file is obtained by mergining three files, that is, citation file, document file, and reference file.

Figure 1 shows an example of referations of document number 857 published in 1981 having four documents of citations and five documents of references. In this case, the number of referations is ten. It is considered that referations is an important attribute of document.

DocN	Yr		
1199	85		
1193	85	Citations	Referations
1110	83		
1097	83		= Citations
857	81	Document	+ Document
863	78		+ References
547	77		
104	74	References	
121	65		
120	63		

Figure 1 Example of Referations of Document Number 857

Now, this attempts to describe the concept of referations mathematically. Let n denote the total number of documents in a specific topic, and let R denote a matrix of reference file. Reference matrix $R=(r_{ij})$ is a square matrix of n rows and n columns. The elements r_{ij} is defined by the relation

$r_{ij} = 1$ if document i cites document j

$r_{ij} = 0$ otherwise.

Then citation matrix C is defined as follows.

$$C = (c_{ij}) = R^t$$

in where R^t denotes the transpose of the matrix R. Let E denote an unit matix which each diagonal element is set equal to 1. Referation matrix F is therefore

$$F = (f_{ij}) = C + E + R.$$

This referation matrix is a square symmetric matrix of n rows and n columns.

REFERATION DATABASE

A document file has the table type of data structure in which a record

describes a single document and the fields describes the attri butes of document. As the referations of document is one of the attributes, it is natural that referation file is taken in document file. But in this paper, referations does not take in document file. It is assumed here that referation file is an accompanying file to document file. Document file with referation file is named for referation database.

Document File(Table Data)				Referation File	
No	Author	Title		No	Referat.
1	Luhn,H	Keyword-in-con		1	1,3,7
2	Knox,D	Effective Sear		2	2,5,9
3	Cuadra	Annual Review		3	1,3,6
4	Lancas	Information Re		4	4,8,11
5	Garfie	The Permuterm		5	2,5,10,12
6	Vicker	Classification	+	6	3,6,7
7	Borko,	Indexing Conce		7	1,6,7,12
8	Meadow	The Scientific		8	4,8,9,10
9	Bradfo	Documentation		9	2,8,9
10	Garvey	Communication		10	5,8,10
11	Cawkel	The Paperless		11	4,11
12	Dewey	Dewey Decimal		12	5,7,12

Figure 2 Example of Referation Database

Figure 2 is an example of referation database. Each record in document file contains the document number, author name, and title to a single document. And each record in referation file is made of record numbers related to a single document. For example, record number three in referation file is related to record one, three, and six.

The data model in database has been developed from complicated hierarchical model and network model to simple relational model. Knowledge base in expert system is now rule model and frame model. However, a simple approach in expert system could be made efficient use of databases and softwares accumulated until the present.

Referation database described here has very simple data structure that is merely table type of file accompanied with referation file. What differs from Boolean search is that referation search makes out the weighted relationships between documents. Next, I describe about the measures.

ASSOCIATION MEASURES BETWEEN DOCUMENTS

It is possible to define an indicator of relationships between two documents using referations. As a basic measure, the matching count between items is always used. For example, reference analysis measures bibliographic coupling between two documents based on references, and citation analysis counts co-citation based on citations. In the case of referations, the same matching method is used.

Figure 3 shows the association count between document number 227 and 434. The left side is the referations of document 227 in 1968. And the right side is the referations of document 434 in 1973. In these linkages, document 959, 781, and 674 is co-citation. And document 37 and 228 is bibliographic coupling. But document 434, 32, 31,

DocN	Yr		DocN	Yr	
1122	83		1050	83	
959	81	--Co-Citation--	959	81	
			930	80	Cit.
781	78	--Co-Citation--	781	78	
674	77	--Co-Citation--	674	77	
104	74		400	75	
434	73	==Direct Link==	434	73	Doc.
			505	73	
32	72	--Cit-Ref.Link-	32	72	
Cit. 405	72				
31	71	--Cit-Ref.Link-	31	71	Ref.
88	71		764	68	
38	70		428	68	
Doc. 227	68	==Direct Link==	227	68	
406	68				
37	67	--Bib.Coupling-	37	67	
			96	67	
Ref. 228	67	--Bib.Coupling-	228	67	
440	65		122	66	
454	65		52	64	

Figure 3 Association Count Between Document Number 227 and 434

and 227 do not belong to any categories. Direct link is found to document 227 and 434. This is indeed the reason that document itself was contained within referations.

Bibliographic coupling is useful to latest documents, and co-citation count is adaptive to older documents. However, association count based on referations is free to the publication year of document.

Three types of association measure between document A and B is defined as follows:

1) association count
   $x = c$
2) association coefficient
   $y = c / (a + b - c)$
3) association ratio
   $z = 50*(SQR(x/x_{max})+SQR(y/y_{max}))$

in which a and b is each number of referations 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 normarized by a total of two referations. And the third type is association ratio adjusted by the association count and coefficient.

The merits of referation analysis is two points:
1) independence of publication year,
2) increase of the oppotunity of linkage.

It is desirable that computer readable referation database in various fields of science and technology realize.

NEW INFORMATION RETRIEVAL SYSTEM

It became clearly that the representation of reference data by code was possible to list the referations of document and to measure the relationships between documents. Using referation database, a new information retrieval system will be developed.

In traditional information retrieval system, set of documents is serched by some query key such as descriptors, author names, journal name, and publication year. And an alterative set of documents through Boolean operation and/or sorting is displayed or printed out according to user's request. It is noticed that each document searched by Boolean techniques is no weight.

Referation search is the following three approachs:
1) search of referations in a given document,
2) search of documents associated with a given document,
3) search of documents, keywords, authors, and sources associated with set of documents obtained.

Table 1 shows an example of the third type of referation search. The set of 47 documents with keyword "BIBLIOMETRIC" in title is retrieved from document file. Then the set of keywords associated with the 47 documents is searched using association measure based on referations. Table 1 is a part of screen in which three types of association measure are displayed in relevant order. Furthermore, the sets of documents, authors, and sources related to "BIBLIOMETRIC" can be searched by simple key operation.

SOME FUNCTIONS OF REFERATION DATABASE

This gives a brief outline of some functions for referation database from the point of view of software design.

(A) Input Procedure
1. Definition of fields of document file
   The name, the number of bytes, and the attribute of each field are defined.
2. Document file
   By assisting of indexes such as document, author, title, source, and publication year, document data is entried.
3. Reference file
   By assisting of author index, reference data is entried using code.

(B) Preparation of Auxilliary Files
1. Field file
   a. Merge of fields : author
   b. one data a field : source, publication year
   c. some data a field : title keyword.
2. Index and name file
   Author, source, form, publication year, title, and keyword is prepared.
3. Citation and referation file
   Citation : inverted of reference
   Referation : merge of citation, document, and reference.
4. Rank file
   Referation, citation, reference, author, source, and keyword are

Table 1 Three Types of Keyword List Searched by 47 Documents with "BIBLIOMETRIC" in Title

No	Keyword	Count	Keyword	Coefft	Keyword	Ratio
1	CITATION	4866	BIBLIOME	0.0977	BIBLIOME	90.52
2	SCIENCE	4393	EMPIRICA	0.0879	BRADFORD	82.32
3	SCIENTIF	4161	BRADFORD	0.0864	LAW	79.18
4	BIBLIOME	3196	GENERAL	0.0852	DISTRIBU	75.07
5	BRADFORD	2826	LAW	0.0846	SCIENCE	66.48
6	LAW	2510	ZIPF	0.0841	SCIENTIF	66.17
7	LITERATU	2397	LOTKA	0.0833	CITATION	65.58
8	ANALYSIS	2354	DISTRIBU	0.0829	EMPIRICA	63.76
9	DISTRIBU	2073	APPLICAT	0.0808	ZIPF	63.63
10	INFORMAT	1947	STATIONA	0.0787	GENERAL	59.51

ranked.

5. Bibliometric distribution file
   Bradford, Zipf, Lotka, growth, and life distributions are prepared.

(C) Output Procedure

1. Referation list of a given document
2. Index list
   KWIC, KWOC, author, source, title index
3. Rank, name, and frequency list, and graph
4. List of document searched.

(D) Search and Association

1. Traditional search
   Keyword, author, source, publication year
2. Referation search on a given document
   Referations and associated documents.
3. Referation search on set of documents
   Associated documents, keywords, authors, and sources.

(E) Auxillary Procedure

1. Sorting routine
   The usage to index and rank file.
2. Boolean Logic
   AND, OR, NOT operation in search.
3. Association measure
   Association count, coefficient, and ratio are obtained.

CONCLUSION

When the document related to a specific topic in references is represented by code, a new attribute of document, called for referations, can be easily made of citations, references, and document itself. And a new information retrieval system using referations can be realized on microcomputers.

Boolean search retrieves no weight and only documents. On the other hand, referation search produces the set of weighted documents, keywords, authors, and sources. These are some kinds of professional information in a specific topic.

This micro-based pilot system is very simple and feasible. The consept of referations and referation database could be applied to not only document information but also the patent and precedent information. It is desirable to develop a general software for referation database.

References

(1) Eugene Garfield, "Citation Indexes for Science," Science, 122 (1955) 108-111.

(2) Henry Small, "Co-citation in the Scientific Literature; A New Measure of the Relationship Between Two Documents," Journal of the American Society for Information Science, 27 (1973) 265-269.

(3) Eugene Garfield, Citation Indexing: Its Theory and Application in Science, Technology and Humanities. (John Wiley & Sons, 1979).

(4) G.Vladutz and J.Cook,"Bibliographic Coupling and Subject Relatedness," Proceedings of the 47th ASIS Annual Meeting, 21 (1984) 204-207.

(5) M.M.Kessler,"Bibliographic Coupling Between Scientific Papers," American Documentation, 14-1 (1963) 10-25.

(6) Isao Asai, "Development of a "Referation" Database Using Micro-computers," Proceedings of the 21th Annual Meeting on Information Science and Technology,(1984) 21-31 (in Japanese).

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

(8) Isao Asai, "The Structure and Usage of A "Referation" Database on Micro-computers For Personal Document Management,"(Lehmann,K.-D. ed.The Application of Micro-Computers in Information, Documentation, and Libraries, North Holland, Elsevier, 1986) 510-517.

INFORMATION:

The Transformation of Society

ASIS '87

Proceedings of the
50th Annual Meeting
of the American Society
for Information Science

Boston, Massachusetts
October 4-8, 1987

Volume 24

asis