

# A MATHEMATICAL EXTENSION OF THE IDEA OF BIBLIOGRAPHIC COUPLING AND ITS APPLICATIONS

SUBIR K SEN  
SHYMAL K GAN  
INSDOC, New Delhi-110067

*The idea behind bibliographic coupling was put forward independently by Fano and by Kessler. The credibility and applicability of the idea have however been questioned. In the present paper the idea has been extended and mathematically formalised. A measure of the strength of coupling vis-a-vis clustering has been proposed. Use of coupling for assessing cognitive scatter and progression of cognitive linkage has also been discussed. The limitations as well as usefulness of the principle have been critically analysed.*

## INTRODUCTION

During the last two decades citation indexing has become popular as a device for retrieval and searching scientific publications on a topic or a field of research. Citation indexes are providing information scientists with useful materials for studying literature patterns, information generation and propagation etc. All such activities are based on the hypothesis, taken as a major premise, that between a citing item and the members of the cited set of items, there is a cognitive relationship linkage of some form. In other words and in weaker terms it may be said that the citing item makes use of some piece of information contained in each of the cited items for some purpose relevant to the context and content of it.

As a consequence of this it has been suggested that citations can be considered as index of cognitive linkage between documents, just by the nature of use by the authors instead of actually assessing the content and subject expression of them. Therefore Fano[1] and Kessler[2] independently suggested that if

two different documents refer to a common item, they should have a sort of closeness in their approach, study, context or cognition. Those which have no common cited ancestors are more likely to be unrelated or conceptually afar[3]. If the number of common citations for two different papers be multiple rather than single, their strength of coupling is said to be more, implying that their cognitive contents are much close to each other. Kessler in a series of papers has further tried to establish the usefulness of the idea[4]. The notion of bibliographic coupling and these works have attracted the attention of experts in bibliometrics and information analysis as interesting but have hardly been taken seriously. The main criticism is against the hypothesis that a common parental citation may be considered as ensuring cognitive relation[5]. Even multiple common citations may not guarantee the breed of brotherhood in terms of content. At best one can say that more the common citations, more is the probability of their being cousins. We therefore felt that the idea of bibliographic coupling should be theoretically elaborated and a general mathematical framework be supplied.

## MATHEMATICAL FRAMEWORK

Instead of considering one step citation, we propose to formulate in terms of generations of citations. Further, instead of considering the common citation or citations of two papers we think in terms of citation populations taken together. In this way we can then prepare correlation matrices and quantitative measures of the bibliographic coupling strength and propose direct applicability to clustering

of items. Let  $D_{1j} = d_{ij}, j=1..n$  be the set of papers at a certain period of time  $t_1$  considered in the context of the study. Each of  $d_{ij}$  cites a set of papers which we call as the parent set or the first generation ancestor-set of citations.

Let this generator set be denoted by

$$D_{0j} = \{d_{0i}\} i=1, m.$$

Now, the totality of the papers of all these first generation ancestor-sets, we call the first generation ancestors or the cited pack,  $G_p$ .

Therefore

$$G_p = \bigcup_{j=1}^m D_{0j} = \{g_{0i}\} i=1, N.$$

Now, each of  $g_{0i}$  generates citing pack: which includes  $d_{ij}$ 's i.e.  $D_1 \subseteq C(G_p)$  where  $C(G_p)$  denotes the Citor-generated by the members of  $G_p$ . This relationship may be represented by the form of a matrix of the following sort:

	CITED PACK					
	$g_{01}$	$g_{02}$	$g_{03}$	$g_{04}$	...	... $g_{0N}$
C						
I	$d_{11}$	1	0	1	1	...
T						... 1
I	$d_{12}$	0	1	1	1	...
N						... 0
G	$d_{13}$	1	1	1	0	...
P						... 1
A	$d_{14}$	0	0	1	0	
C	...	...	...	...	...	0
K	...	...	...	...	...	
	$d_m$	1	0	1	0	...
						... 1

n x N Boolean matrix

Fig. 1. A hypothetical Boolean matrix. 1 denotes citation, 0 denotes absence of citation.

The Boolean matrix in Fig. 1 shows the citation links in general of the two populations and therefore the nature of coupling for the whole set of papers in question. This may be used for generalising the notion of coupling into bibliographic cliques and clusters.

Clusters would be formed by the populations which have at least one member having coupling with another member whereas, no

member of one cluster will have coupling with any member of another separate cluster (Fig.2).

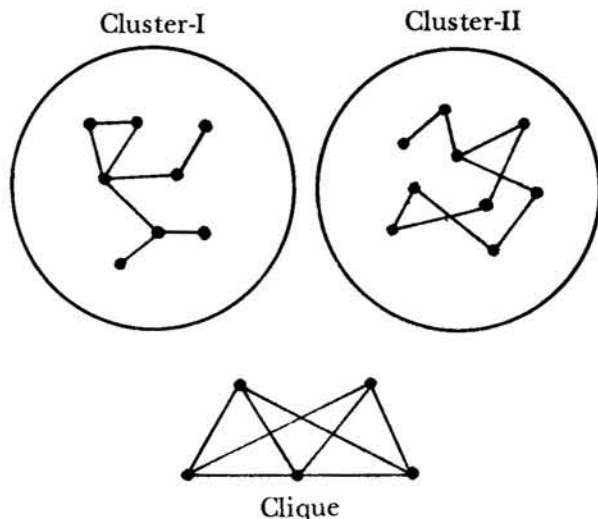


Fig. 2. Clusters are formed when an article (a member of the cluster) has coupling link with atleast another member. No member of one cluster would have link with a member of another separate cluster. Cliques are formed when every member has linkage with every other member. Nodes denote articles and lines-linkages.

Now, let us attempt at defining some measures of strength of single link (one generation or first step) citations coupling.

1.  $D_{0k}$  denotes the set of documents cited by  $d_{jk}$  and  $D_{0j}$  by  $d_{0j}$

Therefore, the set of common citations =  $D_{0j} \cap D_{0k}$  and total cited set of both =

$$D_{0j} \cup D_{0k}$$

We define the duplet coupling co-efficient (C.C) for the two papers  $d_{ij}$  and  $d_{jk}$  as

$$C.C. = \left| \frac{D_{0j} \cap D_{0k}}{D_{0j} \cup D_{0k}} \right| \text{ where } 0 \leq C.C. \leq 1; \dots (1)$$

Generally, we may write for n papers —

$$C.C.(\text{multiplet}) = \left| \frac{\bigcap_{j=1}^n D_{0j}}{\bigcup_{j=1}^n D_{0j}} \right|, \dots (2)$$

2. Considering the matrix, taking each row against a document or paper as a vector of Boolean elements we may define another

measure calling it Coupling Angle (C.A.).

$$C.A. = \frac{(D_{0j} \cdot D_{0k})}{\sqrt{(D_{0j} \cdot D_{0j})(D_{0k} \cdot D_{0k})}} \dots (3)$$

We understand that both of C.C. and C.A. give coupling strength and relative inclination of the papers but the cognitive association cannot be ensured unless a valid cut-off factor or threshold point is determined. For determining this threshold there is no theoretical basis at present. The best course is probably to project a value phenomenologically by actual experimentation. But, such an experimentation would involve assessing the information content and subject expressions which would bring in imprecision and vagueness. Hence, such a phenomenological cut-off factor has to be semi-arbitrary. Heuristically we may attempt to arrive at some value. As we know  $\cos \frac{\pi}{3} = 1/2$ , so for C.A. threshold may be taken at 0.5 which should give high correlation or strong coupling.

#### *Vertical generalisation*

In the previous sections we have attempted at a horizontal generalisation of the idea of bibliographic coupling taking only single time of citations. If we can consider instead of the first generation citation links, a number of generations, the possibilities of generating most cognitively coupled clusters would be enhanced.

Let us consider the set of papers ( $T_n$ ) which we take as n-th generation papers and among which we intend to establish coupled clustering. The cited set of  $T_n$  be  $D_{n-1}$ ; similarly, the cited set of  $D_{n-1}$  papers be  $D_{n-2}$  and so on to the set  $D_0$ . At each generation level we would then be in a position to find the coupling strength. By taking a particular cut-off factor of multiplet C.C. (or a generalised C.A.) we should then get a particular chain of papers having relationships of being cited and citing. Therefore we may find a coupled cluster of papers among  $D_0$  having common citation ancestors at each upward generation. Taking different values of the cut-off factor we are

able to generate radial clusters of diminishing coupling strength out of  $D_0$ . In this way we shall be able to ensure formation of clusters having various degrees of closeness of cognitive relationships. Because, with a single stage coupling cluster, the natural pitfalls frustrating the fundamental hypothesis adopted in the introduction can be avoided. These drawbacks and their elimination through the vertical generalisation procedure will be discussed later.

#### IMPORTANCE OF COUPLING

We must show where and why the horizontal and vertical coupling procedures, as set forth above, can be meaningfully and mechanically exploited. Information processing is a complex yet urgent requirement. Almost all facets of it either depend upon analysing the content thoroughly or on very simple enumerations of the type of document, the number of pages, the title, the number of items and such other highly artificial and elementary features. The difficulty with the first approach is that it involves penetration into semantics and syntactics of expression and language, and prior knowledge of at least the basics of the subjects with which the documents deal. Such processings are time consuming, and prone to vagueness and misrepresentation. This approach is liable to be highly qualitative and very difficult to automate. For the sake of feasibility recourse to arbitrary classifications, key word representations, terms frequency counts etc. have been superimposed but quantification of the devices employed are too elementary, too simple, incomplete and arbitrary to achieve meaningful result. So far, none has been able to show any functional relationship between the cognitive implication of the first approach and the arbitrary attributes of the second approach. Citation measures as a tool fall in the second category but stand as a third alternative between the two approaches. Citation claims a number of features of the second approach, whereas vindicating requirements of the first approach. But unprocessed simplistic citation approach without preconditioning and preprocessing is usually unworthy and too weak to meet the demands of information processing. The simple bibliographic coupling de-

vice proposed so far has all these inherent weaknesses of the simplistic citation approach [6,7]. The areas where the notion of generalised coupling as formalised in this paper may be exploited are listed below -

1. Establishing cognitive chain of information growth and dispersion;
2. Mechanisation of Information Retrieval without recourse to actual analysis of the content;
3. Clustering and classification of items of information according to their relative cognitive distances and/or linkages; and
4. Preparation of prospective and retrospective bibliographies of different spectral radii of topics with intended width.

As we have mentioned in the introduction, it is usually assumed that citing an item in the bibliographic reference implies that the author has used at least one piece of information from the cited items which is relevant to his work. So, citation emphasizes an information-based cognitive relation. As we have shown, this in itself does not warrant actual closeness of the two documents. Only with repeated citations and with formation of respective citations among a group of items we can be more or less sure of saying that the more linked items of this group have inherent conceptual and contextual affinities and our search for automatic mechanisable features for controlling and representing information processes can be rewarded. The formalised and generalised coupling technique as presented here has these qualities.

### CRITICISM AND DISCUSSIONS

Time as a parameter plays a very important role in citation. Except on rare occasions, a cited item precedes the citing item in date of publication. Almost all exceptional cases are difficult for identifying the cited item and without further corroboration of the date element and therefore are useless for our purpose. We should therefore neglect this small peripheral set although this act in itself may create some marginal effects. While referring

an item of publication at a certain point of time an author is in principle in possession of all the preceding items of information published over a long period of time. Usually, the sheer speciality of the topic and the space available in the publication medium restrict the number of items which can be actually cited. In all practical cases the time available to the author, circumstances or chances of knowing and getting hold of a particular past item randomize the selection of the set of bibliographic references. So, the items in this set have publication dates statistically scattered over a period of time through usually with a modal concentration over a span of time and publication media.

A document is specified by its date of publication along with other attributes eg. title, authorship, type of publication, pagination etc. For citation purposes the date is the chief parameter. But, for a number of items we cannot consider a specific date. So we have to account for the time parameter in periods or intervals. We can take such intervals of time in units of year, quarter or month. If we suppose that a particular item is published during the period  $T_n$ , then its cited items can span through all of the previous intervals  $T_{n-1}$  to  $T_0$ , taking  $T_0$  as the date of publication of the earliest item among the cited ones. Similarly, an item published at  $T_n$  can have citations in the ensuing periods from  $T_m$  to  $T_{m+1}$ ,  $T_{m+2}$ ,  $T_{m+3}$ ,  $T_{m+4}$ ,  $T_{m+5}$ ,  $T_{m+6}$ ,  $T_{m+7}$ ,  $T_{m+8}$ ,  $T_{m+9}$ ,  $T_{m+10}$ ,  $T_{m+11}$ ,  $T_{m+12}$ ,  $T_{m+13}$ ,  $T_{m+14}$ ,  $T_{m+15}$ ,  $T_{m+16}$ ,  $T_{m+17}$ ,  $T_{m+18}$ ,  $T_{m+19}$ ,  $T_{m+20}$ ,  $T_{m+21}$ ,  $T_{m+22}$ ,  $T_{m+23}$ ,  $T_{m+24}$ ,  $T_{m+25}$ ,  $T_{m+26}$ ,  $T_{m+27}$ ,  $T_{m+28}$ ,  $T_{m+29}$ ,  $T_{m+30}$ ,  $T_{m+31}$ ,  $T_{m+32}$ ,  $T_{m+33}$ ,  $T_{m+34}$ ,  $T_{m+35}$ ,  $T_{m+36}$ ,  $T_{m+37}$ ,  $T_{m+38}$ ,  $T_{m+39}$ ,  $T_{m+40}$ ,  $T_{m+41}$ ,  $T_{m+42}$ ,  $T_{m+43}$ ,  $T_{m+44}$ ,  $T_{m+45}$ ,  $T_{m+46}$ ,  $T_{m+47}$ ,  $T_{m+48}$ ,  $T_{m+49}$ ,  $T_{m+50}$ ,  $T_{m+51}$ ,  $T_{m+52}$ ,  $T_{m+53}$ ,  $T_{m+54}$ ,  $T_{m+55}$ ,  $T_{m+56}$ ,  $T_{m+57}$ ,  $T_{m+58}$ ,  $T_{m+59}$ ,  $T_{m+60}$ ,  $T_{m+61}$ ,  $T_{m+62}$ ,  $T_{m+63}$ ,  $T_{m+64}$ ,  $T_{m+65}$ ,  $T_{m+66}$ ,  $T_{m+67}$ ,  $T_{m+68}$ ,  $T_{m+69}$ ,  $T_{m+70}$ ,  $T_{m+71}$ ,  $T_{m+72}$ ,  $T_{m+73}$ ,  $T_{m+74}$ ,  $T_{m+75}$ ,  $T_{m+76}$ ,  $T_{m+77}$ ,  $T_{m+78}$ ,  $T_{m+79}$ ,  $T_{m+80}$ ,  $T_{m+81}$ ,  $T_{m+82}$ ,  $T_{m+83}$ ,  $T_{m+84}$ ,  $T_{m+85}$ ,  $T_{m+86}$ ,  $T_{m+87}$ ,  $T_{m+88}$ ,  $T_{m+89}$ ,  $T_{m+90}$ ,  $T_{m+91}$ ,  $T_{m+92}$ ,  $T_{m+93}$ ,  $T_{m+94}$ ,  $T_{m+95}$ ,  $T_{m+96}$ ,  $T_{m+97}$ ,  $T_{m+98}$ ,  $T_{m+99}$ ,  $T_{m+100}$ . So two different items published in the interval  $T_n$  can have cited items falling in different intervals. It may so happen that one item cites a particular item at  $T_{n-p}$  which in itself cites another item at say  $T_1$ , whereas the other item at  $T_n$  directly cites the item at  $T_1$ . In such a case both the items of  $T_n$  may have cognitive linkage with each other but would not be reflected by our coupling scheme. For better precision, the scheme has to be modified to accommodate such a situation.

Another important interfering feature of temporality is that, unless we are restricting the choice of ancestors (1st generation) within limited interval, the enumeration of coupling strengths and of generating clusters would become cumbersome for the system.

Referring to the Boolean matrix for citing and cited packs d's & g's we should consider

di's having publication dates at a certain intervals as well as all the  $g_i$ 's also having publication dates at a particular interval.

Now, while going backwards in time for further ancestors we must not only take the citation links from the  $g_i$ 's but also directly from the  $d_i$ 's and for this accommodation we must modify our formulae. Such modifications later on. The scheme not only provides us with the generalisation of coupling data and their utilisation more effectively but also the scope for generalising and providing co-citation links. Indeed we are now in a position to generate profiles of clusters bound together through both coupling and co-citation. For co-citation data we need to take the column vectors instead of row vectors in our matrix diagram. One coefficient for co-citation would then be given by

$$\frac{G_i \cap G_j}{G_i \cup G_j}, \dots \quad (4)$$

where  $G_k$  represents the items among  $d_{1j}$ 's citing  $g_{ok}$

We can derive from the matrix the set of most strongly related family of 'ancestors—progeny' papers determined by the relationship

$$G_i \cap G_j \longleftrightarrow D_k \cap D_l \dots \quad (5)$$

We shall call such a coupled offspring and co-cited parent papers as a bibliographic family.

## CONCLUSION

The scheme of generalisation provided here need to be further refined. Some manual experimentation and development of suitable computer programme are being taken up by us. Such a practical operation requires carrying out the process backward and forward a number of times starting with a single paper or a group of papers published in a journal of specialised subject or from a book or a classified abstracting service.

Indeed some works of generalisations of the concept of bibliographic coupling[8,9] have come to our notice but we have not been able to identify any paper anticipating our ideas of direct generalisation both horizontally and

vertically of the notion of bibliographic coupling.

An intensive research on clustering has been done in the Institute of Scientific Information including considerations of bibliographic coupling. However, we have not been able to know the exact theoretical models being used by others or by ISI from published reports[10,11].

We would like to express our thanks to Mr S Arunachalam, Mr R Ojha and Dr D K Bandyopadhyay for their interests shown in this work.

## REFERENCES

- [1] Fano, R.N. Documentation in action. New York, Reinhold, 1956, 238-44.
- [2] Kessler, M.M. Concerning some problems of inter-science communication. Lincoln Lab. Group Report. 1958, 35-45.
- [3] See also An elementary theory of classification and prediction. IBM Tech. Report. Nov. 17, 1958.
- [4] Kessler, M.M. An experimental study of bibliographic coupling between technical papers. IEEE Trans. PTGIT IT-9, 1963, 43. Comparison of the result of bibliographic coupling and analytic subject indexing. Amer. Doc. 1965, 16, 223-33.
- [5] Martyn, J. Bibliographic coupling. J. Doc. 1964, 20, 236.
- [6] Price, D.J. de Solla. Network of scientific papers. Science 1965, 149, 510-15.
- [7] Salton, G. Automatic indexing using bibliographic citation. J. Doc. 1971, 27, 98-110.
- [8] Kretschmer, H. Representation of a complex structure measure for social groups and its application to the structure of citations in a journal. Scientometrics, 1983, 5, 5-30.
- [9] Cawkell, A.E. Understanding science by analysing its literature, 1976, 10, 3-10. Reprinted in Garfield, E. Essays of an information scientist. Vol 2, 1977, 543-49.
- [10] Small, H. Cocitation in scientific literature: a new measure of the relationship between two documents. JASIS, 1973, 24, 265-69 and other papers.
- [11] Garfield, E. Citation indexing - its theory and applications in science, technology and humanities. New York: John Wiley, 1979. Also various articles in Essays of an information scientists. 1-5 Vols.