M. Kochen
E. Wong

Concerning the Possibility of a Cooperative Information Exchange

Personal exchanges play an important role in meeting the information needs of the scientific community. A recent study[1] suggested that about 40% of all the articles, reports, textbooks, symposia and annual reviews read during a two-week diary period by scientists and technologists were obtained through personal recommendations from colleagues. This suggests the possibility of establishing an information exchange system utilizing to a greater extent the channels of communication among scientific workers.

In this note both the existing channels and potential channels of communication are considered. Theoretical arguments are advanced to establish the conditions under which a semi-automated information exchange system, advantageous to the participants, can be established. The principal factors to be considered and the parameters to be estimated prior to the implementation of such a system are discussed.

Description of the information exchange

According to a recent study[2] a researcher obtains about half of his reading material from about ten major journals. The other half of the material that he reads comes from a large number of different journals. The materials from these secondary sources read by two individuals of even closely related interests overlap but little. The coverage by an individual of the secondary sources is a problem representing the greatest individual need, and personal exchanges can substantially contribute to its solution. Therefore, in what follows, attention will be focused on the material which originates from these secondary sources.

Consider an individual i of a population of N users of the scientific literature. Let v_i be the number of items per unit time that he reads from secondary sources of his own, i.e., sources that exist without any system of exchange. A simple way of implementing a system of exchange is to have each individual of the population submit to the system material from his own sources which he finds new and significant. The operator of the system then directs the material to a selected number of participants to whom the material is likely to be significant and new. The selection process is characterized by a matrix (a_{ij}), where a_{ij} equals 1 if an item submitted by j is to be sent to i, and 0 otherwise. Let u_i be the number of items per unit time that i receives from other participants through the system which are both new and significant. In order for the system to be successful the ratio

$$\frac{u_i}{\sum_{j\neq i} a_{ij}v_j}$$

must be high for all i. Furthermore, the total amount that i is willing even to scan per unit time is limited by some capacity c_i. It follows, therefore, that

$$\sum_{i\neq j} a_{ij}v_j + v_i \leqq c_i\,.$$

To compute u_i, it is necessary to know s_{ij}, the conditional probability that i finds an item *significant* and *new* given that j found the item significant. Since we are dealing only with secondary sources of i and j which overlap little, it will be assumed that s_{ij} can be reliably approximated by the probability that i finds an item *significant* given j found it significant. That i also finds it new will be assumed.

The central problem of designing an exchange system along the stated lines is as follows: Given the vectors **c** and **v** and the matrix S, determine the matrix $A(a_{ii} = 0)$ such that each of the components of the vector **u** are maximized. Thus $\mathbf{u} = A^*S\mathbf{v}$, where A^*S is a matrix with elements $a_{ij}s_{ij}$. The maximization is subject to the constraint $A\mathbf{v} + \mathbf{v} \leqq \mathbf{c}$.

Complete maximization of **u** may be very difficult to achieve and may not be warranted. An example will be given below to illustrate the fact that application of simple criteria can lead to a good working system without having to maximize **u**.

Let s_{ij} be given by

$$s_{ij} = j^{-\alpha} \qquad \text{if } i \neq j \qquad \text{and}$$

$$s_{ij} = 1 \qquad \text{if } i = j \qquad \text{for } i, j = 1, \cdots, N,$$

where α is a positive constant less than one. A matrix A can be easily constructed as follows: We choose some threshold ε, $(0 < \varepsilon < 1)$, and let $a_{ij} = a_{ji} = 1$ if $\min\{s_{ij}, s_{ji}\} \geqq \varepsilon$, $a_{ij} = a_{ji} = 0$ otherwise. It is easily seen that in this example there is a subset σ of size m (m being the nearest integer less than or equal to $\varepsilon^{-1/\alpha}$), which can be said to form a *cluster*. That is, $a_{ij} = a_{ji} = 1$ if and only if i and j both belong to σ,

$a_{ij} = a_{ji} = 0$ otherwise. With this matrix A, $u_i(i\varepsilon\sigma)$ can be found to satisfy the inequality

$$u_i \geqq (c - v)\left(\frac{v}{c}\right)^{\alpha},$$

where $c_j = c$ and $v_j = v$ for all j, by assumption.

For small values of α, u_i can be seen to be a large fraction of $(c - v)$.

Measurement of variables

The variables to be estimated are the matrix S and the vectors **c** and **v**. The matrix A is under the system designer's control. One particularly simple way to estimate s_{ij} is to present both i and j with a sample of references to K items in the literature. Let $\delta_{ik} = 1$ if i considers item k significant, and 0 if not. Define

$$n_{ij} = \sum_{k=1}^{K} \delta_{ik}\delta_{jk}.$$

If we take

$$n_j = \sum_{k=1}^{K} \delta_{jk},$$

then the ratio n_{ij}/n_j can be used to estimate s_{ij}.

To estimate **v** is more difficult. Possibly, observations like those of Ackoff and Halbert,[2] who found that chemists in industry spend about 9% of their work-time on professional reading, could be used. Assuming a 60-hour week, and that everything read is significant and novel to the reader, and that on the average, ½ hour is spent on such an item, then v_i could be in the vicinity of 10 to 12 items/week, though the variance is undoubtedly large.

Estimation of **c** is even more difficult and probably has to be determined by interviewing. However, there exist indirect means of estimating **c**, as the following illustrative calculation suggests: Recall that c_i is the total number of items that i is willing to scan, v_i the amount that he reads from his own sources, and u_i the amount that he reads coming from the system. Then, we find the relationship

$$\tau_1(u_i + v_i) + \tau_2 c_i = T,$$

where τ_1 and τ_2 are the average reading and scanning time respectively, and T is the total time that i spends on literature from secondary sources. Using the figures of $u_i = 3$ and $v_i = 5$ items/week, $\tau_1 = 0.5$ and $\tau_2 = 0.01$ hours, and $T = 4.5$ hours, we find $c_i = 50$ items/week.

For each of the three quantities S, **v**, **c**, there are severe problems in the mathematical statistics of estimation and sampling, particularly since the random variable (a conditional probability) s_{ij} has a skew distribution, and $s_{i1}, \cdots, s_{iN}$ cannot be independent.[3]

Practical considerations

On elementary theoretical grounds, then, there are empirically determinable conditions under which cooperative information exchanges, such as described above, should prove advantageous to participants. A first experiment to test an aspect of this assertion was conducted. In this experiment, using estimation procedures described earlier, the participants were grouped into two clusters. The average acceptance ratios (items accepted/items received) for the two clusters were as follows: For Cluster I the figures are 19% for items initiated from members of the same cluster, 4% for items initiated from outside of the cluster, 4% for items randomly chosen. For Cluster II the respective figures are 47%, 5%, and 3%. Within the accuracy that the sizes of samples allow, these figures are statistically predictable.[4]

Estimates of S must be continually revised because people's interests change. To revise the estimates, recipients of notices are required to indicate which of the notices sent them by the system led to actual reading of novel and significant articles. These responses, periodically collected over an interval of time, constitutes the bibliographic sample for which the $N \times K$ matrix (δ_{ik}) is computed. From this, in turn, S is periodically re-estimated and A recomputed. The computation of A may be a time-consuming operation when N is large; so can the duplication of incoming notices and compiling the responses. For these systems to be effective N should be large. For a given (small) capacity c_i this makes it more likely that notices from the right set of "colleagues" are being channeled to i. To determine just how large N should be, however, is an open question. In terms of an S for the entire professional community, it is theoretically possible to define an upper limit to how much i could benefit from participation in the best of all possible systems. The success of such an exchange system is strongly dependent on the active cooperation of the participants. Their cooperation, in turn, is undoubtedly affected by the success or failure of the system. The motivation of the participants, especially during the early stages of the system operation, is a most important problem, probably even more important than those of estimation of parameters and details of implementation.

References

1. I. H. Hogg and J. R. Smith, "A Survey of the Use of Literature and Information in the R & D Branch," Mimeo. Industrial Group Headquarters, Risley, Warrington, Lancashire, 1959, pp. 29-33.
2. R. L. Ackoff and M. H. Halbert, "An Operations Research Study of the Scientific Activity of Chemists," Mimeo. Cleveland: Case Institute of Technology, O. R. Group, 1958, p. 30.
3. C. T. Abraham, "Sampling Techniques for Measures of Similarity," unpublished.
4. M. Kochen and E. Wong, "An Experimental System for the Exchange of Scientific Information," submitted for publication.

Received February 7, 1962