INTEREST PROFILE—CONSTRUCTION AND USE OF EXTRACT REPORTS

The construction of interest profiles was an experiment in itself. Various methods were tried during the period covered by this report, including selection by the client of descriptors pertinent to his interests drawn from a list of about 5,000 items, i.e., using a dictionary report. This method did not prove to be very popular, although it was definitely more acceptable than selecting terms from the bulky master thesaurus. The method recommended at present has been generally used since mid-1967 for profile construction; it involves the following steps: An initial collection of terms pertinent to written descriptions of the client's interests is assembled; this selection is made by the system manager using the master thesaurus, dictionary reports, etc. The match criteria (CV) of the various groups of terms are then determined by the manager and client. Where indicated, existing root extract reports are reviewed with the client to expand coverage. For example, in an interest involving "rubbers," the appropriate section in the root extract report where descriptors involving rubber materials are collected is shown to the client, who indicates additional terms to be included in his profile. After the prototype profile is keypunched for introduction into the operating system, a personalized theasurus extract is prepared for the client. This provides a display of cross-referenced terms for each of the terms in his prototype profile, from which he chooses additional descriptors to increase profile effectiveness.

The program developed for extracting all descriptors and all descriptor fragments containing any specified character-set, provides the option of fore-truncation, aft-truncation, fore- and aft-truncation, or a complete word. These programs also allow listing the extracted material in either a report display or in the form of punched cards in the required format for direct introduction into the operating system. Experience has shown that it is not economically practicable to use the punched-card option; it is preferable to produce reports from time to time relating to various root terms as needed, and to use these reports for general reference in the manner previously described.

Little use has been made of the term-term association report, which gives the frequency that any descriptor has been used in indexing with any other descriptor. Originally, it was believed that this kind of report would be very useful as a reference guide in profile construction; however, the cost of producing it, even on a limited number of documents and a limited number of specified entry-descriptors, was too high to enable us to justify a fuller exploration of its value.

A Selective Current-Awareness System Using Engineering Index's Plastics Data Base. II. Performance*

R. H. WAGNER
Research Laboratories, Eastman Kodak Co., Rochester, N.Y. 14650

Received October 1, 1968

The operational performance over a 17-month period of the previously described selective dissemination system is presented. Of the 21,000 notifications sent to about 20 users, 91% were evaluated; of these, 14% were of "Document-Ordered Interest," 48% were "Of Interest," 27% were "Marginal," and 11% were "Of No Interest." Recall data obtained from about half the users over a period of eight months show the precision-factor/recall-factor products are generally greater than 0.5. The effect of iterative profile adjustments on precision-recall performance is discussed. A comparison made with four other SDI systems shows a relatively high level of performance for this system.

During the 17 months of operation of the system (August, 1966, through December, 1967) covered in this paper, we have sent out 21,000 notifications to an average of 17.8 individual clients. We have received "relevance ratings" for about 19,100 of these (91%), which average out as follows:

Of Interest, Document Ordered -- 2750 references (14.4%)
Of Interest, Document Not Ordered -- 9070 references (47.6%)
Marginal -- 5100 references (26.8%)
Of No Interest -- 2140 references (11.2%)

Figure 1 shows the percentile ratings on a monthly basis.

In the first six months of operation [from August, 1966 (6608), through January, 1967 (6701)], clients were offered only three rating categories; beginning with the 6702 issue,

* Presented before the Division of Chemical Literature, 155th Meeting, ACS, San Francisco, Calif., April 1968.

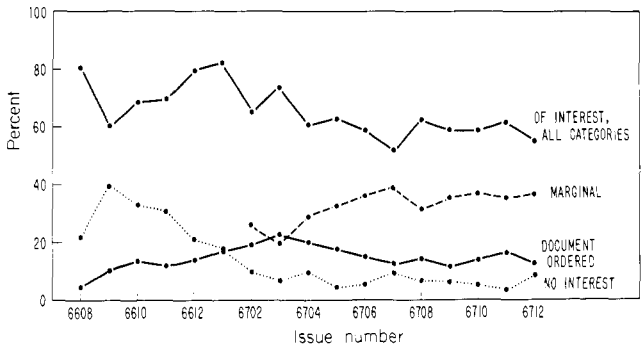

Figure 1. Monthly performance—EK/El-plastics

the Marginal category was added. Prior to its introduction, it became increasingly apparent that clients were placing many articles in the No Interest category without supplying the instructions required for profile adjustment. A survey clearly showed that some sort of rating category was needed to fill the gap between "Of Interest" and "Of No Interest."

The introduction of the Marginal category has been quite fruitful, both from the client's viewpoint and with respect to the effectiveness of the system. Many clients have reported that a considerable number of their "Marginals" have turned out to be valuable references some weeks, or months, after the initial rating. Thus, had this category not been offered, it is certain that many articles would have been classified as "Of No Interest" and subsequent potentially valuable material would have been suppressed.

Feedback is seldom received in time to be used in adjusting profiles for the next run. Most of it is received four to eight weeks after the material is sent out, but some of our clientele wait as long as 10 to 12 weeks. If the evaluation is such that little or no profile adjustment is required, tardiness is not particularly important. However, the prompt return of evaluations by clients in the first three or four months of their use of the system expedites the sharpening of profiles when such changes are most likely to be needed.

All clients are requested to scan the entire abstract bulletin for a month or two after the first and second run to determine the relevant articles missed by the operating system; such information is essential in the initial phases of profile sharpening. Clients are also encouraged to make this kind of a scan occasionally, say, every four or five months, as a means of checking the recall performance of the system, but only the more dedicated will do this.

Nevertheless, for a period of eight months (6702 through 6709), we were able to obtain information regarding missed articles from about 10 clients on a more or less regular basis. This enabled us to calculate 55 recall factors and compare them with the corresponding precision factors, and thus obtain another measure of the performance of the system. Recall factor (recall ratio) is defined as the ratio of the number of relevant articles retrieved by the system to the number of relevant articles in the bibliographic (monthly) file. The precision factor (precision ratio) is defined as the ratio of the number of relevant articles retrieved by the system to the total number retrieved. These definitions are identical to those given by Cleverdon.[1] In terms of the labels employed by Savage,[3] our recall factor is the ratio of Hits/(Hits + Misses) and our precision factor is the ratio of Hits/(Hits + Trash).

The precision-recall performance for these 55 cases is shown in Figure 2. Absolutely ideal performance would be represented by the point at P.F. = 1.0, R.F. = 1.0—denoted by the circled X. The hyperbolic curves shown represent performance limits, corresponding to P.F. × R.F. = 0.4 and 0.5, respectively. Forty-six of the 55 points lie on or above the 0.4-hyperbola and 38 lie on or above the 0.5-hyperbola.

It is our belief that, in a current-awareness system, which usually comprises a relatively small increment file, one should strive for maximum recall, at the sacrifice, if necessary, of precision. There are several reasons for this: first, even with a fairly high selective reaction (file relevance), a user will not, in general, be receiving a very large number of notifications at each search cycle, so that the impact of the number of references he deems "not relevant" will not be as damaging as the impact of ultimately learning that a substantial number of relevant references have been missed; second, many of the "irrelevant" articles may make more of a contribution to the user's thought processes and the direction of his work than is immediately apparent. For calculating precision and recall factors, we rate as irrelevant all references judged to be of marginal interest by the clients, even though few consider that all of them are truly irrelevant.

In a retrospective search, which usually involves a large cumulative file, maximum precision should be emphasized over maximum recall. The impact of missing some of the pertinent items will be less damaging than the retrieval of a substantial number of irrevelant references.

Let us now examine some of the data lying below the 0.5-line of Figure 2 by first considering the point at P.F. = 0.60 and R.F. = 0.15. In Figure 3, this point is labeled 4, indicating that it is the precision-recall result report for 6704. Subsequent reports by this client for 6705 through 6709 produced the points labeled 5, 6, 7, 8, and 9. The marked improvement in performance between 4 and 5 is largely because only four of the 13 subprofiles required were operative in 6704, whereas all 13 were operative in 6705 and later. The variations occurring between 6705 and 6709 are typical in magnitude to those observed with other recall-reporting clients; it is likely that they reflect the normal fluctuations inherent in the operating system. The introduction of the descriptor-subdescriptor search capability in 6707 may have had some bearing on the improvement in recall between 6707 and 6709.

In Figure 4, the precision-recall performances relating to three other clients are shown: these involve six of the points lying below the 0.5-line. For 6704, the system retrieved references for Client No. 020 from a file of 517 articles; of these, two were judged relevant and four marginal by the client who, in addition, indicated two articles that were not selected by the system. The client's profiles were adjusted accordingly, but these adjustments happened not to make any contribution in the selection processes in the 6705 and 6706 runs. In 6705, the system retrieved eight references out of a file of 510, all of which were deemed relevant; and the client did not report any

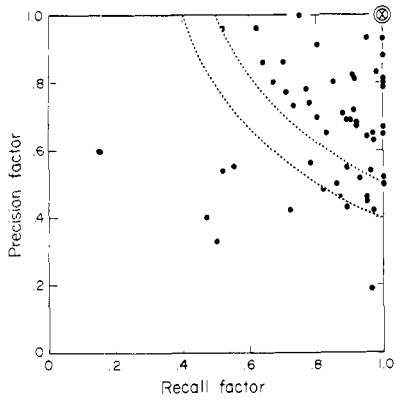

Figure 2. Precision recall performance—cumulative 6702-6709 (55 Returns)

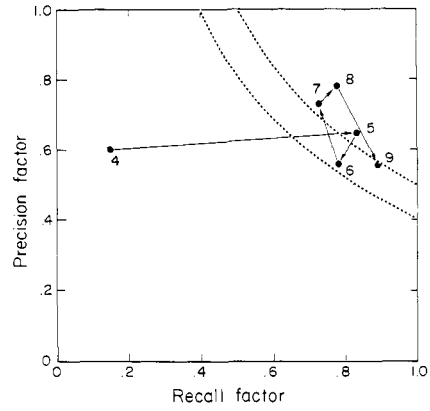

Figure 3. Performance: client No. 060

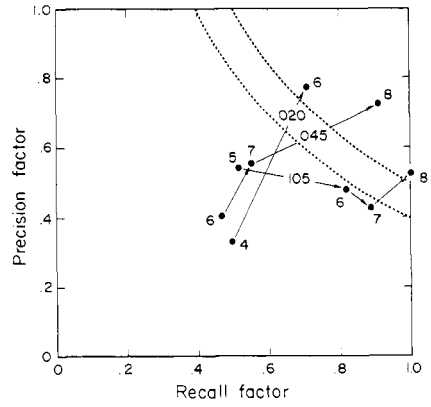

Figure 4. Performance: clients No. 020, 045, 105

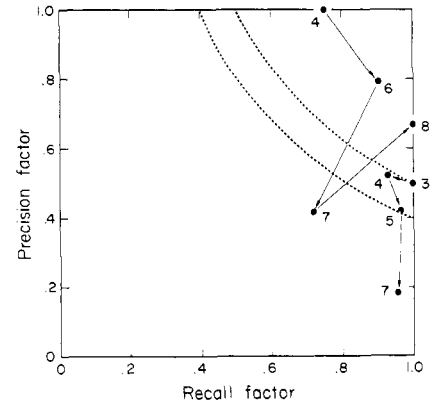

Figure 5. Performance: clients No. 010 and 050

missed articles. In 6706, the system retrieved 13 references from a file of 500, of which 10 were relevant and three were marginal; four articles were missed. Thus, the improved performance here had nothing to do with profile sharpening, but rather was related to the greater relevancy of the 6706 file compared with the 6704 file. Profile adjustments for Clients No. 045 and 105 were in large measure responsible for the improvement in performance shown in Figure 4. Between 6706 and 6708, the subprofiles of these clients were among those that had been considerably expanded by the input of descriptors selected from thesaurus-extract reports.

In Figure 5, the performances for Clients No. 010 and 050 cover the remaining points below the 0.5-line. With respect to the 4, 6, 7, 8-curve, the retrogression from 6706 to 6707 can only be explained on the basis of file relevancy variations and the low selective reaction between this client and the plastics file. In this case, the input from the thesaurus-extract selections and the introduction of descriptor-subdescriptor elements are definitely known to have contributed to the improvement between 6707 and 6708.

The performance shown by the 3, 4, 5, 7-curve is interesting. This client (No. 010) has been associated with the experiment since 6611 and, according to all accounts, apparently thought the operating system had been beneficial, even though his feedback resulted in precision factors that were consistently in the 0.50 and 0.60 range between 6611 and 6704. When the retrogressive trend, evident in the curve shown, was explored, the client stated that he had had a reassignment of job responsibilities in the spring of 1967 which should have caused him to cancel all but one of his six subprofiles, but he had failed to do so.

Another measure of performance of our system can be presented within the frame of reference provided by Savage.[3]

In the evaluation of the effectiveness of SDI systems, Savage proposes that the important parameters to be considered are the User-Document population (U·D); the percentage of U·D that expresses the number of relevant documents retrieved ("hits"), the number of irrelevant documents retrieved ("trash"), the number of relevant documents not retrieved ("miss"), and the number of irrelevant documents *not* retrieved ("pass"); the percentage of U·D that expresses the number of notices sent to users ("selective reaction"); and the precision factor ("hit ratio").

The selective reaction is a measure of the interaction of the interest profiles with the documentary material; if it is too low (perhaps less than 1 or 2%), Savage suggests that an evaluation may have little significance.

He also points out that a meaningful evaluation of systems for which no miss data, or estimates thereof, are available is very unlikely. A method is proposed whereby miss data can be estimated and this method

Table I. Comparative Performance of Various SDI Systems[3]

	SDI-1	SDI-2	SDI-3	SDI-4	This System: Full recall basis	This System: Est. recall basis
U·D,[a] No.	6670	619,550	1,860,000	532,932	23,751	144,229
Hits, %	6.2	2.3	0.9	0.6	8.3	8.2
Trash, %	9.0	1.1	0.5	1.4	5.1	5.0
Miss, %	5.8	36.6	[b]	[b]	1.2	1.2±
Pass, %	79.0	60.0	98.6	98.0	85.4	85.6±
Selective reaction, %	15.0	3.4	1.4	2.0	13.4	13.2
Hit ratio (=P.F.)	0.41	0.68	0.66	0.30	0.62	0.62

[a] U·D is the product of the number of users and the number of documents processed. [b] Data not obtained; lumped with Pass.

was used in deriving the %-miss figure given for the SDI-1 System in Table I of this paper.[2]

In Table I, the data cited in Savage's paper representing four systems labeled SDI-1, SDI-2, SDI-3, and SDI-4 are displayed together with comparable data from our system on a "full-recall basis" (U·D = 23,751) and on an "estimated-recall basis" (U·D = 144,229). The full-recall basis includes only data obtained from the 55 cases described earlier; the estimated-recall basis includes all available data between 6608 and 6712. The estimated value of 1.2± for %-miss in the last column of Table I is inferred from the preceding column and is not based on feedback from randomly selected references. Since the percentages found for the "hits" and the "trash" in these two columns are almost identical, we infer that the percentage of the missed articles will also be substantially the same.

LITERATURE CITED

(1) Cleverdon, C., *Nat. Acad. Sci.* **1**, 687 (1959).
(2) Hensley, C. B., *et al., IRE, Trans. Eng. Management* **EM-9**, 55 (1962).
(3) Savage, T. R., *Am. Doc.* **18**, 242 (1967).

Data Retrieval for a Large Organic Synthesis Project

H. J. ACKERMANN, E. H. KOBER, R. E. McARTHUR,
R. E. MAIZELL, and D. A. SHERMER
Olin Mathieson Chemical Corp., 275 Winchester Ave., New Haven, Conn. 06504

Received November 26, 1968

A data processing system for storing and retrieving experimental data from an organic synthesis project is described. The data keypunched from a specially designed notebook are processed in an IBM 1800 computer to provide multifaceted printouts for the project scientists and research management.

The laboratory project referred to in this paper involves the successful development of a new organic synthesis. Some 80 chemists and engineers have participated in various phases of the project.

The project moved along at a good pace during the first few months, and as data began to accumulate, the need for an automated system to store and access the data became clear.

By the time the data processing system became operational, the project had progressed past the initial bench stage and into the development phase. The benefits of the system were almost immediately apparent. Project scientists were relieved of the time-consuming task of tabulating data for weekly oral review meetings and for periodic progress reports. Research management was able to study advances in the progress of the project through rapid access to laboratory data. Consultants were able to supply valuable suggestions with a minimum expenditure of time. Listings of starting materials, catalysts, solvents, and products were updated by computer printout. Patent attorneys were able to utilize the file in search of information and data for patent applications.

METHOD

As our first step in developing this data processing system, we conducted group and individual in-depth interviews with the project scientists to pinpoint the kinds and relative importance of data being generated. We were then able to categorize the kinds of information most frequently used as follows: starting materials, catalysts, solvents, reactors, temperatures, time, pressures, product(s) yield, conversion, and by-product(s). Schemat-