A Computerized Current Awareness Service Using Chemical-Biological Activities (CBAC)*

LYNN BOND, CARLOS M. BOWMAN, and MARILYN T. BROWN
Computation Research Laboratory, Dow Chemical Co., Midland, Mich. 48640

Received June 9, 1969

A computerized current awareness service for selective dissemination of information using Chemical-Biological Activities (CBAC), a Chemical Abstracts Service magnetic tape service, is described. The system, now four years old, includes 40 profiles containing terms, author names, and journal coden. Terms are assigned positive or negative numerical weights, relative to an arbitrary hit level. Right truncation of terms is allowed. In searching, the titles and abstracts from each biweekly CBAC tape are inverted and matched against the alphabetized list of words from all profiles. Hit documents are printed on 4- x 6-inch card forms. Precision of weighted word profiles as determined by feedback data and user reaction to abstract searching are discussed.

Four years ago we expanded our computerized current awareness system to include Chemical-Biological Activities, or CBAC, as a supplement to *Chemical Titles*, which we had been using for two years. CBAC represented our first exposure to a service covering a specialized body of information and, therefore, possessing appeal to a limited group of users. CBAC also provided us with early experience in free-text abstract searching, using an uncontrolled vocabulary and file inversion techniques. Our four years' experience with CBAC, including user reactions, profile effectiveness, feedback results, costs, and programming problems, are discussed in this paper. Information about our other current awareness services, which include *Polymer Science and Technology* (journals and patents), *Engineering Index—Plastics Section, Biological Abstracts, Chemical Abstracts Condensates,* and internal research reports, may be found in two other papers.[1,2]

CBAC, produced by the Chemical Abstracts Service, is an index to the current literature pertaining to the biological activity of organic compounds. CBAC differs from *Chemical Titles,* or *CT*, in two important ways—the number of documents covered and the amount of information available for each document. While CBAC's journal coverage is slightly less than *CT*'s, 585 to 650, the actual number of documents per biweekly issue is only one-seventh that of *CT*, or 570 to 4000. Only articles pertinent to the defined subject area are selected for CBAC issues, whereas an issue of *CT* contains titles from all articles, without regard for their specific topics.

Secondly, CBAC contains digests, registry numbers, and molecular formulas in searchable form, whereas *CT* does not. In an effort to make abstract searching more effective, CAS has employed the concept of links. Each digest is subdivided into groups of terms tied together logically to convey a particular idea. Each group or "link"—usually a complete sentence—is coded uniquely so that the association of the words can be retained. In matching, each link is considered a separate candidate for a hit. Thus far, we have not yet made use of the registry numbers and molecular formulas; however, the experience with abstract searching using links has paved the way for the smooth addition of other tape services containing abstracts.

MECHANICS

At the start of the experiment four years ago, a new series of programs was written to handle CBAC tapes, whose physical format and content differ substantially from those of *CT* tapes. Retaining their link codes, all words in the title and digest of each document were arranged alphabetically in an inverted file for use as index terms. A rather lengthy list of stopwords was eliminated from the list. The index words were matched against an alphabetized file of weighted profile words for CBAC users only. Our search strategy required that a document, in order to be a hit, must contain enough hit words in the title, or in at least one link, to reach or exceed the arbitrary hit level. Hit documents were printed out on 8½- × 11-inch paper. The printout included title, reference information, author names, journal code, journal name, a code indicating which Dow libraries subscribe to the journal, whether it is available in translation, and all links which produced a hit. We are grateful to CAS for providing us with tapes without charge during the early period of the experiment.

Last summer, programs for a new general current awareness system which could handle each of the seven tape services we now receive were completed. Basic changes included incorporation of right truncation, or the use of prefixes, to indicate any words which contain the specified letters at the beginning of the word, consolidation of profile terms for all services into one alphabetized file, reduction of the stopword list to 63 minor words, employment of 4- × 6-inch card output forms with tear-off feedback tabs,

* Presented before the Division of Chemical Literature, 157th Meeting, ACS, Minneapolis, Minn., April 1969.

and provision for printing out the words which produced the hit beside the document citation.

The processing of a current awareness tape such as CBAC under the new system involves running three general programs written in Algol. The first, called the conversion program, breaks up the titles, author names, journal coden, and digests into an alphabetized file of index words, and at the same time sets up the written information, properly formatted, in a file ready to be printed on the output card notice. Conversion time for CBAC issues containing an average of nearly 500 documents requires 40 minutes of processor time. The second, or match program, compares the latest tape of profile terms with the index tape produced in the first program, and by referring to the profile's hit level, stored with the subscriber's name and address in a disk file, selects the documents which are hits. Matching usually consumes about 10 minutes of processor time. The third, or print program, prints the document citations. Printing normally takes 30 minutes of printer time and produces an average of 1100 hit notices. The entire job can be completed in one day, if necessary, although it has been more convenient to run only one program per night for three consecutive nights since our computer, a Burroughs B-5500, is shared with many other users in the research area of Dow.

The output cards from the print program are separated, manually packaged, mailed out within a day or so of printing, and usually reach users' desks via Dow internal mail within a week from receipt of the tape. Subscribers are asked to indicate their degree of interest in each article on the feedback tab and return it to us. If they wish to see a reprint of the article, they send the abstract card to the appropriate Dow Library and the library returns a copy.

Two profile tapes are maintained: one in profile number order for printing profiles and the other in alphabetical order for use in matching. The tapes contain over 700 profiles serving nearly 400 people. Profile changes are made as often as a reasonable number collect, and two copies of each changed profile are printed out, one for our master books and one for the subscriber. A typical profile updating run takes 15 minutes, but this figure varies in proportion to the number of corrections to be made.

The initiation of the truncation facility has cut down greatly on the number of words necessary to describe interest areas adequately. As time permits, we are backtracking to incorporate truncation where needed in old profiles and have recently finished truncating all CBAC word lists. CBAC profiles account for 6.3% of all our profiles and approximately 7.4% of the words in our collection of over 27,000 terms for all services.

PARTICIPANTS AND PROFILES

At the start of the experiment with CBAC, 18 volunteers from the Midland location were selected and their profiles developed. Some were also *CT* subscribers and benefited from their experience in profiling for that service, while others were fresh starters. The service continued with this number of users for two years while the system was perfected. Then an advertising campaign produced 36 additional subscribers. Recently, the number of users has tapered off for a variety of reasons.

Profiles are established through consultation with the profiling advisor in our group. During this interview, the prospective user is familiarized with the mechanics of the system, the output format, and the methods of obtaining reprints, as well as the techniques of selecting and weighting profile words. This procedure takes from 15 minutes to one hour, depending on the user's familiarity with computers, and how much previous thought he has given to the words he wishes to use. We believe that this time is well-spent, because a subscriber who understands the system can more easily manage his own profile and, most importantly, sell the system to other potential users.

Once on the system, subscribers can make changes at any time they wish, without charge during the first six months, and for a small fee any time after that. Most users make three or four changes before they are satisfied with their output. Thus far, the burden has rested with the subscriber to initiate his own changes as he felt they were necessary. Some users are satisfied to receive large numbers of undesired hits if they feel it is essential to be sure they are receiving every article of interest to them. If a user consistently receives an excessive number of cards, we plan to make a small extra charge per card. Usually, a few minutes spent reworking a profile can greatly improve its effectiveness, and thereby save on processing and printing costs.

The 40 current CBAC subscribers are medical doctors and other scientists who are engaged in basic research in pharmacology, toxicology, treatment of various diseases, or development of agricultural chemicals. Some of their profiles are very specific and others quite general; for example, one profile is designed to pick up all articles which mention any of a series of Dow herbicide and insecticide products.

Our profiling logic employs weighted terms; we have no provision at the present time for the use of Boolean expressions. However, most logic we wish to employ can be fairly well established using both positive and negative weights. Single words, author names, and journal coden are acceptable profile terms.

The average profile length is 48 words. Negative words appear in five profiles. Eleven profiles contain author names, generally at full weight to pick up all articles published by those authors. Nine use journal codes, all positively, to give added weight to articles within those journals. All but seven profiles contain truncated terms. Nineteen per cent of all CBAC profile words are truncated.

FEEDBACK DATA

In spite of the pre-addressed, tear-off tab, feedback data is difficult to collect. Data from about 50% of our users over four recent issues of CBAC indicated that articles of interest ran about 20%, marginal interest 15%, and no interest 65%. During this period, users averaged 27 notices per issue. Although it is too early to tell, we expect the interest ratios to improve due to the truncation of many profile terms which was recently completed. Prior to the use of truncated terms, many undesirable hits resulted from addition of weights for multiple forms

of the same concept in one link of the digest. For instance, if the words "enzyme" and "enzymatic" were present together in a profile and also in a digest link, their weights added together. Now such a profile contains the fragment "enzym*" and those words count only once.

Other possible explanations for such low interest ratios are free-text searching without vocabulary control and profiling errors. Lack of vocabulary control in searching text occasionally results in undesired hits based on unintended meanings of profile terms or combinations of terms. Free-text searching has made us aware of another cause of unwanted hits, the "dilution" factor, or the reduced significance of words when they appear in the abstract as compared with the title. Admittedly, authors usually give more thought to their choice of words for a concise title than for the abstract. It would be possible to modify our search strategy to differentiate between words appearing in the title and digest words. However, construction of a profile to take advantage of the differentiation would be extremely difficult and the resulting profile might not be consistently effective. In general, we have found that the only method of circumventing these two situations involves addition of appropriate terms with large negative weights.

Profiling errors are usually rectified easily. Subscribers quickly become aware of gross errors in assignment of weights and usually request profile changes promptly. Many unwanted hits can be avoided by the inclusion of negatively weighted terms. Other undesired hits of less obvious cause can often be eliminated by juggling weights of profile terms; the profiling consultant can assist in this endeavor.

Recall—the number of pertinent notices retrieved as compared with the number of interesting articles which should have been retrieved but were not—is extremely difficult to measure. However, most users are not aware of any important articles missed. One subscriber admits that he missed two articles over the four-year period and has no complaints about this. The few users who have missed articles attribute this to lack of coverage of that journal or to profile error which they have subsequently attempted to correct.

Nearly every subscriber indicates that CBAC has saved him time in keeping up with the literature, even if he did his literature work on his own time. However, whether it was done at home or at work, this time has now been made available for other activities. Others feel it has actually saved them research time by helping with a project directly. One subscriber indicates that through CBAC he found an article which will help define a new research area. All these intangible indications of CBAC's value are difficult to assess in terms of dollars and cents; equally frustrating is any attempt to balance the cost of the service against the return on the investment.

COSTS

In an attempt to recover part of the computer costs for running CBAC searches, we recently instituted an annual subscription rate of $150 per profile. Most subscribers felt that this charge was reasonable and elected to continue their profiles. However, the slightly reduced number of subscribers has increased our cost per profile and has led to an attempt to reevaluate the merits of free-text abstract searching using file inversion.

ABSTRACT SEARCHING

There is no doubt that for small numbers of users, abstract searching by inverting the entire file of words in titles and digests is very expensive. Analysis of the content of the first five issues of Volume 9 (1969), containing an average of 484 documents in each issue, indicates that each document includes an average of 93 different words and that for each tape there are about 10,900 unique terms, each of which appears in an average of four documents. Picking these terms off the tape and sorting them to produce the inverted file represents a significant amount of computer effort. Computer times indicate that with 40 to 45 users, the fixed costs per search (conversion to inverted file form) were greater than either match times or print times, both of which are dependent on the number of users. If we endeavored to charge subscribers enough to cover our costs completely, we would undoubtedly be left with no users at all. Cost studies on other services, *CT* in particular, indicate that for 200 to 300 users, file inversion allows us to reduce our costs enough that a price acceptable to the market will also cover our costs. The point at which file inversion for current awareness becomes worthwhile has not yet been established in our situation. However, if we plan to do retrospective searching on CBAC, files inverted during current awareness runs could easily be merged into the data bank used for retrospective searches.

User reaction to abstract searching has, of course, been favorable, especially when the service was free. The subscribers feel more secure that their profiles will pick up all interesting documents if they do not need to depend on the authors' titles alone.

Since the beginning, we have printed out on the card notices only the links in the digest which produced hits rather than the entire digest, which is sometimes quite lengthy. In some cases, links printed were reported to be so far out of context that they were worthless. Most subscribers, however, felt that the sections printed were enough to help them judge whether the article was truly of interest and whether a copy should be ordered. A few said that occasionally the portion of the digest printed provided enough information that the users had not needed to obtain a reprint at all.

Consideration of alternative methods for handling CBAC searches presents two main possibilities. First, the titles, author names, and journal coden alone could be inverted and searched, and for each hit the digest could be printed out. Employment of an arbitrary line printing cutoff could prevent long digests from running over one or two cards. Users would then continue to receive the benefits of an abstract on the cards. Searching titles only would result in a few missed documents, but the recent movement to encourage selection of accurate and complete titles for documents would undoubtedly minimize this disadvantage. A second possibility, serial searching, may be a more acceptable solution for services with small numbers of users. Serial searching involves matching profiles directly against the text of each individual document by moving through the intact sentences looking for profile

word matches. Implementation of this technique would necessitate extensive reprogramming of our system. Eventually, if we need to put CBAC on a cost-recovery basis, we will pursue one of these two possibilities.

Few subscribers rely totally on CBAC for all literature searching. Most use it primarily to gain coverage of peripheral and/or inaccessible journals. Virtually all users continue to subscribe to a few favorite journals for browsing and review *Chemical Abstracts* from time to time. One subscriber supplements CBAC coverage by perusal of *Current Contents*. It is doubtful that computerized current awareness searching would ever replace journal browsing, even if coverage of literature were complete and thorough for every journal.

FUTURE PLANS

In addition to investigation of more economical programming techniques for services which appeal to small numbers of subscribers, consideration of several other aspects of our system has high priority. These aspects include: expansion of our general search strategy to provide for Boolean logic, infixes, and suffixes; retrospective searching; revision of feedback techniques to include more extensive analysis of results at less frequent intervals, rather than simple analysis of the interest level of the output from every issue. We also hope to be able to provide more personal attention to sharpening established profiles in order to trim operating expenses.

CONCLUSIONS

Our four years' experience with CBAC has provided invaluable information in two respects. First, we learned a great deal about the literature-searching needs and habits of our research community by observing their reactions to CBAC's thorough coverage of a relatively small body of information. Secondly, we were exposed to the practical and economical problems of computerized free-text abstract searching by file inversion. These considerations have contributed to the over-all development of our current awareness system and will continue to help us define its future direction.

LITERATURE CITED

(1) Bond, Lynn, Carlos M. Bowman, and Dolores Hartman, "User Reaction to Three New Services Offered by Chemical Abstracts Service," Division of Chemical Literature, 152nd Meeting, ACS, New York, Sept. 1966.

(2) Brown, Marilyn T., "A Computerized Current Awareness System Using Chemical Abstracts Tape Services," 59th Annual Conference of the Special Libraries Association, Los Angeles, Calif., June 1968.

A Multi-Level Retrieval System

II. Medium-Sized Collections*

LEE N. STARKER, KATHERINE CRAWFORD OWEN, and BETTY COOPER BATSON
Warner-Lambert Research Institute, Morris Plains, N. J.

Received March 17, 1969

Retrieval systems based on the IBM-type variable field visual collation card can be converted to Termatrex systems as the collections grow in size and use. The conversion is accomplished by computer-generation of a corresponding term/document/IBM card deck which is used to drill a set of Termatrex cards with the J-400 Termatrex drill. The term/document cards can then be used in the preparation of a number of subsidiary search tools. Sample index work sheets and the input procedures are also discussed.

In the first paper of this series,[1] we described a visual collation or "peek-a-boo" approach to the control of relatively small collections of documents. This system was based on the common IBM card and was characterized by the fact that it could be handled readily by the ultimate user on a total do-it-yourself basis. It also could be operated with partial or total assistance from a central information group.

In our experience with the IBM card-oriented peek-a-boo technique, we found that several of the retrieval systems developed around it rapidly outgrew this particular device. The reasons varied from the large number of documents involved to the increasing number of searches required. Thus, a system for the indexing and retrieval of literature in the cosmetic field quickly passed the 6000 document mark (requiring 12 decks of peek-a-boo cards), while our library's use of the peek-a-boo cards, with a collection of 1500 papers (three decks) on Coly-Mycin, generated an increasing volume of searches. In the first case, the need to carry a relatively small number of searches through 12 decks of cards, and in the second instance, the requirement for a larger number of searches through three decks of cards, began to impose a burden on those users directly involved. The search procedures began to consume more time than was thought to be appropriate.

We did not feel that either of these situations warranted computerization at the time. Instead, we turned our considerations toward the Termatrex[2] approach. This

* Presented in part before the Third Middle Atlantic Regional Meeting, ACS, Philadelphia, Pa., February 1968, and in part before the 57th Annual Convention of the Special Libraries Association, Minneapolis, Minn., June 1966.