Searching Chemical Abstracts Condensates, On-Line and Batch†

RONALD M. KAMINECKI,* PATRICIA A. LLEWELLEN, and PETER B. SCHIPMA
Computer Search Center, IIT Research Institute, Chicago, Illinois 60616

Received December 15, 1975

IITRI's Computer Search Center provides both SDI and retrospective searches of *CA Condensates*. Questions requiring complete coverage are searched via the CSC batch system that permits both left and right truncation. Other searches are searched on-line, using either the Lockheed or SDC system. Data and observations have been recorded for the last 18 months on data base availability, ease of use, cost effectiveness, user orientation, and system reliability. The criteria for the use of one system over another are presented.

INTRODUCTION

With machine-readable data bases becoming a commercialized reality, two principles of searching techniques have arisen. Until the late 1960's, computer-aided searching of technical literature was relegated to the batch mode of data processing. This method involves the keypunching of search terms and logic parameters for entry into the software system. But now direct processing in an interactive mode is feasible, using portable cathode-ray tube (CRT) and thermal printing terminals as data entry points. What are the similarities between these two modalities? Which system is the more cost effective? How comprehensive are the results? To the user, how useful is the format of the output? Can one system be justified for use over the other on a cost basis alone? The answers to these questions lie within the context of the application. There are no clear-cut answers possible without qualifications.

Primarily, information retrieval should be user-oriented. The technical aspects of a computerized literature search are irrelevant as long as the results are satisfactory both in scope and cost. The penultimate goal is to be cost effective; the ultimate goal is to be cost beneficent. Though the process is transparent to the user, the bill certainly is not.

IITRI's Computer Search Center (CSC) has been involved with both methods of information retrieval for some time. The CSC was started in the late sixties as an outgrowth of a National Science Foundation contract. The objective of the contract was to develop software capable of searching large data bases in a short amount of computer time. When funding ceased, the CSC was continued as a self-supporting entity, existing on costs recovered from demand services. Such services include Selective Dissemination of Information (SDI), retrospective machine and manual searches of the literature. To achieve this end, batch processing and on-line retrieval methods are used. Both methods are available, either through in-house programs or via suppliers such as the System Development Corporation's "Orbit III" or Lockheed's "Dialog" service. But what is the criterion for use between batch or on-line?

SIMILARITIES AND DIFFERENCES OF BATCH AND ON-LINE RETRIEVAL

Materials. In the case of *Chemical Abstracts* (*CA*) *Condensates*, the CSC subscribes to the weekly issues of magnetic tape that correspond to the printed versions. In the same sense, the on-line vendors subscribe to the same service. Since both processors cover the same material, the same constraints apply to each. *CA Condensates* tapes contain bibliographic and indexing information. With the exception of *Chemical and Biological Activity Condensates* (CBAC) and *Polymer Science and Technology* (POST), there are no abstracts on tape owing to the great amount of material (about 300 000 citations per year). These machine-readable tapes have only been in existence since 1969. Therefore, exhaustiveness of a search is directly related to the comprehensiveness of the abstractor. But cost effectiveness is a function of the manner in which this material is searched.

Software and Hardware. The information specialist is at the mercy of the software. The interest profile must be interpreted from the natural language of the researcher's question to the machine code of the programmer. File structure, as dictated by volume and speed considerations, can be either inverted (i.e., terms are found in a linked dictionary) or linear (straight text). The inverted file is easier to search, but it is difficult to program for left truncation. On-line systems usually dictate the use of inverted files while batch systems can use either. Linear searching is much slower than inverted file searching, but improvements in software have made linear searching competitive with inverted file searching (such as Onderisin's "Least Common Bigram", LCB).[1]

On-line retrieval is inherently faster than batch, and out of necessity certain features are not used. Weighting of terms is not always feasible since there is very little time for reflective thought while interactive with computer core. Terms are weighted only after some browsing and careful ponderance concerning high-priority terms over peripheral terms. Therefore, weighting is generally not used on-line. Instead, the quick operators of Boolean logic (AND, OR, AND NOT) are the means by which material is selected. Batch systems usually include both of these features on a selective basis.

Stifled profiles are the result of insufficient term allowances. Most profiles range from 1 to over 200 terms with averages ranging from 12 to 67/profile.[2] On-line systems, through both storage and cost considerations, are usually incapable of searching over 100 terms with any degree of sophistication.

On-line means interactive (read conversational) and so, direct access devices (usually many disk drives) are employed. Tapes are not searched directly. Batch mode usually includes only one disk and a complete set of tapes. Such tape is I/O-bound, yielding a somewhat slower access but demanding a much lower hardware requirement.

There is a tradeoff, though, between slower access and hardware overhead. Batch searching is machine competive with on-line searching owing to the sharing of terms. In the former, each term is pulled from the profile and placed in a

† Presented to the American Chemical Society, Division of Chemical Information, 170th National Meeting of the American Chemical Society, Chicago, Ill., Aug 26, 1975.

* Reprint requests should be made to Mr. R. M. Kaminecki at the above address.

Table I.[a] Degree of Nesting vs. % Number of Profiles

No. of sets of parentheses	% profiles
0	24.3
1	14.6
2	18.7
3	10.8
4	9.0
5	6.7
6	3.7
7	2.6
8	1.1
9	3.3
10+	5.2

[a] Reference 3, p 285.

Table II.[a] Breakdown of Number of Logical Operators Used per Profile

No. of operators used/profile	Logical operators		
	AND	OR	NOT
0	13.1	22.4	67.5
1	25.0	22.0	27.6
2	21.3	17.2	3.4
3	14.6	5.6	1.1
4	9.7	8.6	0.4
5	5.2	7.8	
6	5.2	5.2	
7	2.6	3.4	
8	1.1	1.1	
9	0.7	0.4	
10+	1.5	6.3	

[a] Reference 3, p 284.

register. Identical terms are placed in this register only once, qualified by a profile accession number (duplicate terms are grouped and searched together). Therefore, the search step does not needlessly plough through the whole file looking for each term in each profile.

Preparation is the key to success in either technique. Computer time is expensive and human frustration in the middle of a search yields inappropriate references at high cost. In this case, batch is extremely cheaper because of its "after-the-fact" iteration. Human frustration is not allowed to enter the system during processing. Trial and error searching is not done during computer processing time. After the profile is run and the output reviewed, then hair can be torn, paper can be ripped, and cards can be spindled, folded, and mutilated.

Both systems allow nesting of Boolean operators to a degree higher than the average mind can comprehend. Some questions need nesting to the 4th or 5th degree[3] (see Tables I and II). Clearly the degree of nesting coupled with the number of operators, establishes a confusion factor. And it is an expensive factor.

The prime consideration involves the ever-present clock. Computer time is always expensive. On-line time charges range from about $36/hour to about $180/hour ($3/minute or 75 cents for the amount of time it took to read this sentence). On-line retrieval is always under this cost/time pressure. Decisions about terms and index phraseology take time. The easiest way to evaluate output is via the off-line print. Martin[4] asserts that a "fast scan by the computer of a wide range of possible decisions and their consequences allows the decision-maker to proceed to an in-depth analysis *off-line*" (emphasis added).

After a profile has been developed and run through the computer, the output can be perused at one's leisure. The batch mode always produces such output; the on-line mode must order such a luxury.

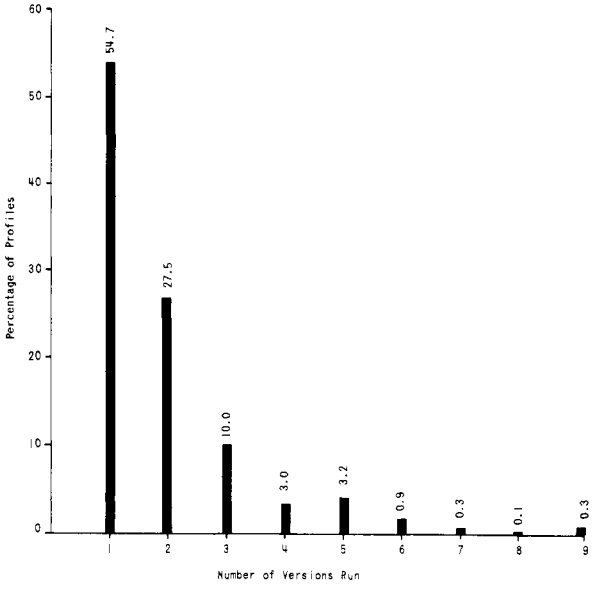

Figure 1. Iterations of SDI profiles.

Therefore, in order to be complete, an on-line search should have high recall. That is, the profile must be somewhat vague so as to allow peripheral material to be printed over and above the relevant citations. High-precision on-line searches are not, therefore, cost beneficent. If the precision is too high, a portion of the relevant material can be considered unfound. Batch mode is measurably different. High-precision searches can be developed over the several iterations used in batch technique. As shown in Figure 1, profiles are reevaluated more than once in almost half of the normal profile runs.[5]

Costs. The cost effectiveness of either system is related to the end result meeting the needs of the user. The cost beneficence is one step beyond, meaning the results should be cost justifiable with respect to timeliness, comprehensiveness, and readability of output.

Many information centers operate in a not-for-profit environment, charging only the recoverable costs. These costs are passed to the information center from the data base supplier. Research, keypunching, programming, royalties, and related charges are dictated by the data base supplier. The method of access reflects these entities according to the inflexible demands of the system. Becker[6] found that the cost of keeping inverted files is approximately one order of magnitude larger than keeping linear files of large data bases. On-line access uses inverted files thereby reflecting the high line-charges. Williams[2] found that for seven information centers using batch processing, this supplier fee accounted for an average of 7% of user cost. Computer time charges accounted for 23%. The majority of the expense involved personnel time which ranged from 29 to 84.5%.

Both methodologies involve programs that are in constant refinement. Information specialists must be constantly retrained and alerted to the dynamic aspects of data entry. Training is extremely important, not in the sense of overrunning an estimate as much as in missing a large portion of relevant material.

In Table III, the amount of training and retraining has been incorporated in the allowances for personnel time. The tasks and costing breakdown are based on our experience at the CSC. Note that within the on-line method there are two distinct ranges. This is due to the software development of different suppliers. The first estimate relies on the inclusion of a saving function whereby the searcher can save the commands of a search for future use for a small line charge. The second supplier does not have this type of function as of this date, meaning that the search has to be run in term by

Table III

Tasks		% labor/% CPU Batch	On-line
1.	Profile Development	30/0	30/0
	User questions analyzed		
	Files surveyed for fit		
	Idiosyncrasies noted		
	Search terms listed, synonyms found		
	Terms fit into logic		
	Consultation with staff		
	Final list recorded		
2.	Test Run	10/10	10/15
	Interest profile inputted		
	Corrections, editions, output changes made		
	False drop terms deleted, AND NOT terms added		
	Output reviewed		
	Output charged		
3.	Final Run	10/20	10/10-15
	Final version inputted		
	Minor changes made		
	Staff review of output		
	Interaction with requestor		
	Output charged		
4.	Updates	50/60	60/60-75
	Input final profile		
	Staff keyboarding, logon, computer expenses		
	Output charged		
	Remailing, handling		

term every update run. The major expense is in the keyboarding of the input.

SERVICES

In general, the output specification is of limited format when run on-line. The searcher usually has a set of data type options from which to choose. Certain fields can be specified or complete bibliographic information can be specified. Certain sorts can also be ordered, but output is almost always on paper. Though output can be recorded on tape, it is not yet clear how such recording of information stands on copyright grounds. The CSC does not record such information on tape through on-line vendors.

The batch mode offers output on cards for manual index files, or on regular one-part tall paper or magnetic tape (by prior arrangement with the data base suppliers). Various listings can be made using the search results as a foundation for a data base subfile. The CSC has such software established, known as Private Libraries (PRILIB). Alphanumeric sorts by any field, Keyword in Context (KWIC), Keyword out of Context (KWOC), Key Letter in Context (KLIC), and straight bibliographic listings can be generated on demand. Additions, corrections, and other manual input data can be made at any time. This capability, coupled with the advantage of having such data in machine-readable form, makes this system very adaptable and applicable to users needs.

All in all, considerations for time, need, coverage, and adaptability are used as the criteria by which a searcher uses one method over the other. Sometimes it is the case that both methods are used. On-line retrieval may be used as the initial search with a constant update via batch. A trial profile may be run on-line to get a feel for the amount of information in the field, backed up by a batch retrospective search.

Finally, note that in the case of SDI, though the time is proportioned differently, the total cost is about the same. This only lends credence to the transparency of the involved methodology to the user. Machine-readable data bases and computerized retrieval systems have opened up the library dependence of literature searching with either on-line or batch methods. Given proper profile preparation, cost effectiveness is inherent; cost beneficence is attained through proper use and selection of one system over the other by the information specialist.

REFERENCES AND NOTES

(1) E. M. Onderisin, "The Least Common Bigram: A Dictionary Arrangement Technique for Computerized Natural-Language Text Searching", IIT Research Institute, Chicago, Ill.
(2) M. E. Williams, Ed., "Cost Elements and Charge Bases in Information Centers: Proceedings of Panel Discussion", ASIDIC Meeting, Philadelphia, Pa., March 7, 1973, p 19.
(3) M. E. Williams et al., "Four Year Summary-Education and Commercial Utilization of a Chemical Information Center", Final Report, Contract No. NSF-C554, June 30, 1972, p 284.
(4) T. H. Martin, "The User Interface in Interactive Systems", C. A. Cuadra, A. W. Luke, Ed., Annual Review of Information Science and Technology, Vol. 8, 1973, p. 208.
(5) A. K. Stewart, "Iterations of Computer Search Center Information Services", presented at the 1973 Annual Meeting of the American Society for Information Science, Los Angeles, Calif., Oct 21–25, 1973, p 1.
(6) P. B. Schipma et al., "Design Specifications for Manipulation of Large Data Bases", Final Report, Contract No. NSF-C734, Oct 5, 1973, p 78.