Improvements in Cost Effectiveness in On-Line Searching. I. Predictive Model Based on Search Cost Analysis†

J. ROBERT ALMOND* and CHARLES H. NELSON

ICI United States Inc., Wilmington, Delaware 19897

Received September 22, 1977

Cost analyses of on-line searches delineate contributions from operator keystrokes, terminal speed, turnaround time, number of operators, type (on-line or off-line), and number of prints, database connect hour charges, and telecommunications charges. These components are combined to produce a general formula for prediction of search costs based on search strategy and database selection. Preliminary testing indicates the validity of qualitative cost effectiveness predictions between databases in 83% of the cases.

The value of any given piece of information has a limit to any given user. Benefits accrued to the requestor of the information must exceed the costs expended for its retrieval. It is, therefore, imperative that costs of information retrieval be predictable. The number of search requests we process continually increases, and the costs involved in information retrieval play an increasingly important role in budget preparation, staffing, internal pricing, and evaluation of operating efficiency. Some of these costs, e.g., hardware or personnel, are reasonably fixed while others such as expenditures for on-line bibliographic database usage are variable (dependent on connect time and number of citations retrieved). Many studies can be found in the literature in which these costs have been calculated for previous search years, but an analysis of search cost components and construction of a model for predicting future search costs represents a novel approach to the problem. Studies of on-line costs[1-12] are generally one of three types: (1) comparisons of actual on-line costs with costs of manual searching; (2) composite average cost estimates based on charges per unit time and estimated on-line times; (3) cost effectiveness studies in which the cost of obtaining a specific number of relevant references for a search question is compared between databases and search strategies. However, none of these analyses have concentrated on the relationship of the search strategy employed with the costs resulting from print charges and connect time and their dependence on database and software selection.

Database vendors maintain that their operations are highly cost effective although no quantitative measure is ever given. As a result many information users are unaware that the connect time per search is dependent not only on the search question but also on the contents of the database and the software used to access it. Consequently, these same users select from among relevant databases and software systems by choosing the one that is easiest to use, or has the lowest connect hour cost. However, the searcher has a number of cost factors directly under his control (type of search criteria used, degree of precision desired, time spent in selector input) and others indirectly under his control by selection of command language and database (turnaround time and connect hour charges).

This study intends to analyze the costs of on-line searching to develop a general formula for cost prediction.

Radwin[13] recently mentioned one obstacle to studies in this area—variation in the speed at which search criteria are input to the computer, i.e., a variation in the time taken by the user at the terminal in evaluation of output and entering (typing) the next input. We have circumvented this obstacle by using predetermined search criteria recorded on a magnetic tape which is activated on receipt of a user cue. This effectively converts the operation into a batch mode on-line system operating with input speed equal to machine speed (30 characters per second in this study). This limits the applicability of the predicted costs obtained from the model. To broaden the applicability of the model, each user must estimate the time spent in thinking or typing at the terminal in an interactive mode since no uniform standards exist.

A second obstacle to cost studies is the complexity of the cost components of on-line searches. Salton discusses several parameters affecting on-line costs.[14] Among these are average input rate, processor elapsed time to completion (turnaround time), number of simultaneously active users, and frequency of timeslicing. These parameters, however, are not independent. A number of them are strongly interrelated. Additionally some database vendors do not openly timeslice the user, and many other factors, e.g., character length, postings frequency, and file formats (linear or inverted), definitely affect the cost of the search. One additional objective of this study is to analyze the cost components of an on-line inverted file search to determine which components are related to known quantities, i.e., citation charges, connect hour costs, machine speed, postings levels, database size, etc. The total cost of a search (excluding planning, labor, and capital costs) is the sum of the on-line costs and the off-line costs.

On-line costs are the product of the composite rate for any given database and telecommunications network and the total time elapsed from log-in to disconnect operations. The time employed in each component operation of a search can be subdivided into components that are independent of terminal speed and components that are terminal speed dependent.

The first group represents the turnaround time for each operator (selection, limit, or combination), i.e., the time during which the operation is being performed by the computer plus the telecommunications time encompassing each respective operator. Examples of terminal speed dependent components are entry and output of selectors, limits, or combine commands; entry of print statements, on-line printing, and on-line expansion. In each case, the time used is obtained by multiplying the reciprocal of the terminal speed (seconds/character) by the number of characters processed.

Off-line costs are the product of the number of off-line prints and the dollar cost per off-line print. The interrelationships of the specific elements contributing to the total search cost are graphically repesented in Figure 1. For a given search strategy, the relationship of the relevant search components to the overall costs are summarized in Table I.

C_s, C_c, C_l, C_{pc}, and C_p represent the sum of the characters included in input, output, and user cue messages for different

† Presented at the ACS 11th Middle Atlantic Regional Meeting, Newark, Del., April 21, 1977.

0095-2338/78/1618-0013$01.00/0

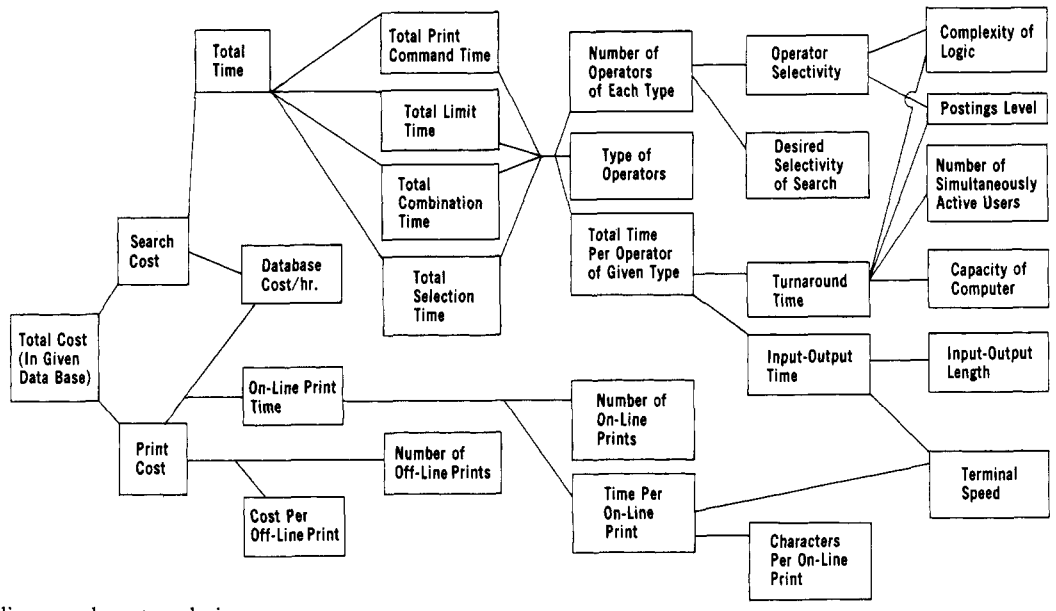

Figure 1. On-line search cost analysis.

Table I. Relationship of Relevant Search Components to Overall Costs[a]

		On-line cost								
		Select		Combine		Limit		Print		
	Component	Entry output	Process-ing	Entry output	Process-ing	Entry output	Process-ing	Entry output	Process-ing	Off-line cost
R_d	Database connect hour charge	+	+	+	+	+	+	+	+	0
R_t	Telecommunications connect hour charge	+	+	+	+	+	+	+	+	0
R_p	Cost per off-line print	0	0	0	0	0	0	0	0	+
P	Number of prints	0	0	0	0	0	0	+	+	+/0
S	Terminal speed	+	0	+	0	+	0	+	0	0
C_p	Characters/print	0	0	0	0	0	0	+/0	0	0
N_s	Number of selectors	+	+	0	0	0	0	0	0	0
N_c	Number of combinations	0	0	+	+	0	0	0	0	0
N_l	Number of limits	0	0	0	0	+	+	0	0	0
N_p	Number of print commands	0	0	0	0	0	0	+	+	0
C_s	Characters/select	+	0	0	0	0	0	0	0	0
C_c	Characters/combine	0	0	+	0	0	0	0	0	0
C_l	Characters/limit	0	0	0	0	+	0	0	0	0
C_{pc}	Characters/print command	0	0	0	0	0	0	+	0	0
T_s	Turnaround time/select	0	+	0	0	0	0	0	0	0
T_c	Turnaround time/combine	0	0	0	+	0	0	0	0	0
T_l	Turnaround time/limit	0	0	0	0	0	+	0	0	0
T_{pc}	Turnaround time/print command	0	0	0	0	0	0	0	+	0
PL	Postings/selector	0	+	0	+	0	+	0	0	0
AU	Number of active users	0	+	0	+	0	+	0	+	0

[a] +, component has an effect on cost contributions of that type; 0, component has no effect on cost contributions of that type.

operator types. For example: "# proton". The character sum for this operator in Lockheed databases = 37 characters.

User cue ?-	2
Operator input #proton	7
Carriage return (input)	1
Operator output __1__459 proton	26 (includes 16 spaces)
Carriage return (output)	1
Total	37

For the operator "&25/5/1-100", the character sum = 32

User cue ?-	2
Operator input &25/5/1-100	11
Carriage return (input)	1
Operator output printed 25/5/1-100	17
Carriage return (output)	1
Total	32

S represents the machine speed of the terminal. This figure is normally given in cps (characters per second). Terminal speeds normally used in interaction with Lockheed or System Development Corporation databases are: 10 cps for most teletypes, 30 cps for most thermal printers, and 120 cps for many CRT (cathode ray tube). If searches are conducted in batch mode, all printing is done at machine speed. In comparison, a competent typist (60 words per minute) types at 5 cps.

R_d and R_t are uniform charges per connect hour imposed by the vendor for use of the database and use of the telecommunications network, respectively. Since they represent elapsed time as opposed to processing time (CPU), they are directly proportional to the cost of all on-line operations. By adding the two values, one obtains the composite hourly rate for on-line operations (R_y) as shown in Table II.

C_p represents the characters per print in the desired format. It is important to know the approximate value of C_p to determine whether it is less expensive to print the citation off-line at a fixed cost or on-line at a cost dependent on its length and

Table II. Composite Rates for Databases

Database	Connect hour charge, $/h	Telecommunications, $/h	Total $/h
Chemical Abstracts Condensates	35	5	40
CASIA	70	5	75
Chemical Abstracts Patent Concordance	45	5	50
CLAIMS®	150[a] (90)	5	155[a] (95)
WPI (Derwent)	120[b] (90)	8	128[b] (98)

[a] Net cost = $90/h ($60/h applied to cost of Uniterm Index). Therefore, for subscriber rates $95/h is used. [b] Cost to basic subscribers of Derwent sections is $90/h. Therefore, for subscriber rates $98/h is used.

the operating terminal speed (S).

P represents the number of citations resulting from the search which are to be printed either off-line or on-line. For off-line prints, multiplying P and R_p (the cost per off-line print) gives the total off-line cost for a search. For on-line prints, P must be multiplied by C_p/S and R_y to determine the print cost.

N_s, N_c, N_l, and N_{pc} represent the number of operators of a given type (selector, combination, limit, or print command) used in a search. These are multiplied by the sum of C_x/S and the turnaround time to give the total time attributable to an operator of a given type (x).

T_s, T_c, T_l, and T_{pc} represent the turnaround time for an operator of a given type, that is, the time interval between the last character of input from the terminal to the first character of output at the terminal. This time is dependent on the size of the task (i.e., the number of items, PL, in the set created or operated upon) and the functional capacity of the computer (process time at 100% dedication diluted by the number of simultaneously active users, AU). The dependence of turnaround time on these factors will be explored in another study.

The overall cost for a given search in database y is given by the following expression:

$$\$ = R_p(P_{off}) + R_y \Sigma N_x(T_x + C_x/S)$$

where N_x = the number of operators of type x, T_x = average turnaround time for an operator of type x, C_x = average number of characters/operator of type x, S = the terminal speed (in characters per second), R_y = the composite rate for database y per connect hour, P_{off} = the number of off-line prints, and R_p = the cost per off-line print in database y.

It is obvious that for a given search strategy in a given database, P and N_x are defined by the search. C_x is defined by the selection criteria and truncation level chosen. S is defined by the terminal used. R_y and R_p are fixed for the database by the vendor's pricing policies. Therefore, if the total cost of a number of single operations (of type x) is known, the turnaround time T_x can be calculated. By utilizing these average values of T_x, approximate future total costs, may be predicted. However, since T_x is a function of the postings level of the selector, the overall number of users on the system, the system capacity, and the telecommunications time, predicted values of the total cost may not correspond exactly with the actual costs but should give qualitative indications of the relative cost effectiveness of conducting a given search in a number of databases of equal scope.

This model and its corresponding formula were tested for predictability on a qualitative basis using five appropriately constructed test questions. These questions dealt with U.S. patents available in three different databases (Lockheed's CLAIMS™/CHEM) and Chemical Abstracts Condensates databases and System Development Corporation's Derwent–WPI database). In all cases the desired patents were included in all databases. Comparisons were made of predicted and actual search costs for retrieval of these patents from each database. The order of cost-effectiveness predicted by the model was realized in 83.3% of the cases.

These results demonstrate for the first time the feasibility of employing theoretically based predictions of cost effectiveness in on-line searching with reasonable success.

REFERENCES AND NOTES

(1) D. T. Hawkins, "Impact of On-Line Systems on a Literature Searching Service", *Spec. Libr.*, **67**, 559–67 (1976).
(2) S. A. Elman, "Cost Comparison of Manual and On-Line Computerized Literature Searching", *Spec. Libr.*, **66**, 12–18 (1975).
(3) M. S. Radwin, "The New Era of On-Line Information Retrieval—Evaluation of Its Costs and Benefits—A Professional Imperative", *Proc. Am. Soc. Inf. Sci.*, **10**, 191–2 (1973).
(4) F. W. Lancaster, "The Cost-Effectiveness Analysis of Information Retrieval and Dissemination System", *J. Am. Soc. Inf. Sci.*, **22**, 12–27 (1971).
(5) B. Lawrence, B. H. Weil, and M. H. Graham, "Making an On-Line Search Available in an Industrial Research Environment", *J. Am. Soc. Inf. Sci.*, **25**, 364–9 (1974).
(6) J. S. Buckley, "Planning for Effective Use of On-Line Systems", *J. Chem. Inf. Comput. Sci.*, **15**, 161–4 (1975).
(7) F. B. Rogers, "Costs of Operating an Information Retrieval System", *Drexel Libr. Quart.*, **4**, 271–8 (1968).
(8) M. E. Williams, "Criteria for Evaluation and Selection of Databases and Database Services", *Spec. Libr.*, **66**, 561–9 (1975).
(9) D. B. Marshall, "User Criteria for Selection of Commercial On-Line Computer-Based Bibliographic Services", *Spec. Libr.*, **66**, 501–8 (1975).
(10) S. Hecht, "Comparative Costs of On-Line Retrieval via Hierarchical Coding and Natural Language", 172nd National Meeting of the American Chemical Society, San Francisco, Calif., 1976, No. CHIF-16.
(11) M. M. Bivans, "A Comparison of Manual and Machine Literature Searches", *Spec. Libr.*, **65**, 216–22 (1974).
(12) B. G. Prewitt, "Searching the Chemical Abstracts Condensates Data Base via Two On-Line Systems", *J. Chem. Inf. Comput. Sci.*, **15**, 177–83 (1975).
(13) M. S. Radwin, "Choosing A Terminal", *On-Line*, **1** (1), 11–17, 64–66 (1977).
(14) F. W. Lancaster and E. G. Fayen in "Information Retrieval on-Line", Melville, 1973, p 382