The Development of a General Model for Estimating Computer Search Time for *CA Condensates*

M. K. PARK, J. L. CARMON, and R. E. STEARNS
Computer Center, University of Georgia, Athens, Georgia 30601

Received March 25, 1970

A statistical model for the prediction of search time has been developed for computer searches of *CA Condensates*. The equation developed is $ST = a + tT + dD + fDT$ where ST is the search time, T is the number of profile terms, D is the number of documents, a is a constant, and t, d, and f are regression coefficients to be estimated. Graphs showing the effect of data base size and number of search terms on search time are presented.

The Information Science Unit of the University of Georgia Computer Center operates computer search services on data bases from several publishing organizations. Operations began May 1968 with searches initially offered only to faculty, research staff, and graduate students at the University. Since that time, several data bases in the fields of chemistry, biology, and nuclear science have been added, and services have been extended to other universities, industrial organizations, and governmental agencies.

During the first year of operation, data were collected to determine the costs associated with the operation of an Information Dissemination Center. As a part of this analysis, a study of the relative effects of various factors on computer search time was undertaken to evaluate this portion of the total costs and to develop a model for estimating search time for production jobs.

In this study, seven subfiles of the data base were constructed from Volume 71, issues 12 and 13, of *CA Condensates*. This tape service, published by Chemical Abstracts Service, includes the title of the article, the primary document citation, the secondary journal (*CA*) citation, and keyword index terms. Each test data file was analyzed to determine the number of words contained in the title and in the keyword records and the average number of characters per word in these fields. The tapes were used just as received from CAS; no processing had been done to remove "redundant" terms either between the title and the keyword phrases or within keyword phrases. Characteristics of the subfiles are shown in Table I. The number of documents in the subfiles ranged from 2000 to 11,232 with approximately 25 words per document. The average number of characters per word varied from 6.6 to 7.0. In this analysis, a word was defined as any string of characters bounded by blanks or by a blank and a period, comma, or semicolon and occurring in the title or keyword records.

Table I. Data Base Characteristics

Data Base No.	No. Documents	No. Words	Av. Words Per Citation	Av. No. Characters/ Word
1	11,232	276,083	25.6	6.7
2	7,999	200,319	25.0	6.6
3	6,999	174,794	25.0	6.7
4	5,999	150,424	25.1	6.7
5	3,870	89,928	23.2	7.0
6	2,983	76,209	25.5	6.6
7	2,000	51,794	25.9	6.6

Search profiles (questions) used in the study were selected at random from the current awareness profiles normally run in routine production operations. Nineteen groups of profiles, ranging in size from 1 to 150 profiles, were constructed. Profiles consisted of text, authors, and CODEN terms, constructed according to CAS conventions.[1] The majority of the terms, 93.7%, were text terms; Coden terms accounted for 4.6%, and author terms, 1.7%. The profile sets were also analyzed for the number of terms and the number of characters per term, as shown in Table II. The number of terms per profile ranged from 5 to 25.36 with an over-all average of 19.1 terms per profile.

The search program used was the program provided by CAS for use with the *CA Condensates* tape service.[2]

Table II. Profile Characteristics

Run No.	No. Profiles	No. Terms	Terms/ Profile	No. Characters	Characters/ Term	Characters/ Profile
1	1	5	5.00	54	10.79	54.00
2	5	26	5.19	313	12.04	62.59
3	5	27	5.40	320	11.85	64.00
4	10	167	16.69	1,514	9.06	151.39
5	10	153	15.29	1,316	8.89	136.09
6	15	194	12.93	1,689	8.71	112.59
7	15	192	12.79	1,660	8.64	110.66
8	20	250	12.50	2,191	8.76	109.54
9	25	324	12.96	2,811	8.67	112.43
10	25	316	12.64	2,719	8.60	108.75
11	30	373	12.43	3,248	8.71	108.26
12	35	592	16.91	4,752	8.03	135.77
13	40	774	19.35	6,115	7.90	152.87
14	45	877	19.49	7,138	8.14	158.62
15	50	1043	20.86	8,499	8.15	169.97
16	50	951	19.02	7,838	8.24	156.75
17	75	1504	20.05	12,679	8.43	169.05
18	100	1905	19.05	16,068	8.43	160.67
19	150	3804	25.36	32,452	8.53	216.34

It is written for the IBM 360 series computers with OS and is programmed in 360 Basic Assembler Language. There are three job steps in the system. The first step, which will be referred to as SEARCH, edits the search profiles (questions), matches the profile terms against the data base file, and selects the bibliographic data for citations which are answers to the search logic. Citations are then sorted into order by question number, question weight, and *CA* citation using the IBM OS sort program in the second step. In the third step, PRINT, the search report is printed. All computer runs were made on the Computer Center's IBM 360/65, operating in an MVT environment. The searches were run in primary core, with all intermediate files spooled to a 2314 direct access device.

A total of 67 search runs were made in the study. Not all runs were made against each data base; however, a minimum of nine and a maximum of 10 runs using varying size profile sets were made against each data base subfile. The central processing unit (CPU) time for each step and the number of answers retrieved for each run were tabulated for analysis.

RESULTS AND DISCUSSION

As can be seen from Table I, the word characteristics of each data base were fairly constant. The average number of words per citation varied from 23.2 to 25.9 words for the title and keyword records. The average number of characters per word was relatively constant at 6.6 to 7.0. The correlation between the number of words and the number of documents in the data base was very high, 0.999. Thus, the variables that characterized the different data bases were the number of documents and the number of words in the data base. For prediction of search time, either variable can adequately define the data base, due to the high correlation.

The characteristics of the profiles showed considerably more variability than did the data bases. As shown in Table II, the number of terms per profile varied considerably as did the number of characters per profile. With the exception of the first three runs, the number of characters per term was fairly constant. A smaller number of characters and terms per profile was observed in runs 1, 2, and 3.

The sort and print steps accounted for a very small portion of the total search run time, ranging from 1.3 to 6.7% of the total run time. The highest percentages occurred for runs with two minutes or less total CPU time. Both were highly correlated with the number of answers retrieved. Since it was virtually impossible to predict the number of answers retrieved, no model was developed to predict this portion of the total search time.

The SEARCH step required most of the run time in the searches, and it was to this portion of the total time that the remaining analysis was devoted. Logically, the search time should be a function of the data base size and the profile set size since the SEARCH step performs a character-by-character match between these two files.

The correlations between the variables for all search runs are shown in Table III. As mentioned previously, the correlation between the number of words and the number of documents was 0.999. Neither of these variables

Table III. Correlations Between the Variables Measured

	No. Words	No. Profiles	No. Terms	No. Characters	Search Time	No. Hits
No. Documents	0.999	0.108	0.119	0.122	0.475	0.488
No. Words	1.000	0.097	0.107	0.110	0.468	0.484
No. Profiles		1.000	0.986	0.984	0.826	0.779
No. Terms			1.000	1.000	0.835	0.761
No. Characters				1.000	0.836	0.756
Search Time					1.000	

was by itself a very good predictor of search time (correlations of 0.475 and 0.468). The correlations between the number of profiles and the number of terms or characters were very high. The correlations of characteristics of the profiles and of the data base were high and fairly constant for each file with respect to size. However, neither set of variables was sufficient to independently predict total search time. The data base characteristics accounted for 22% of the variation in search time, and the profile characteristics accounted for 69% of the variability in search time.

A plot of search time against the number of terms for each data base size is shown in Figure 1. As the graph shows, these are linear functions. For a fixed data base size, the number of terms accounted for an average of 96% of the variability in search time. Figure 2 shows a plot of search time against the number of documents for a fixed number of terms. This is also a linear function.

As can be seen from the intercepts in Figure 1, a certain amount of overhead is dependent upon the data base size. This is a reflection of the time required to pass the tape in a tape oriented batch search system. The intercepts in Figure 2 are dependent on the size of the profile set (i.e., number of terms) and represent the time necessary to load the profile terms alphabetically into core.

The average time per term was calculated, and these are plotted in Figure 3. The time per term searched reaches a constant for all data base sizes at approximately 195 terms. If we consider that the average profile for *CA Condensates* contains 19 terms, then the minimum time per term is obtained with more than 10 profiles (questions) per run. For example, in the data base containing 11,232 documents, search time was 0.26 minute

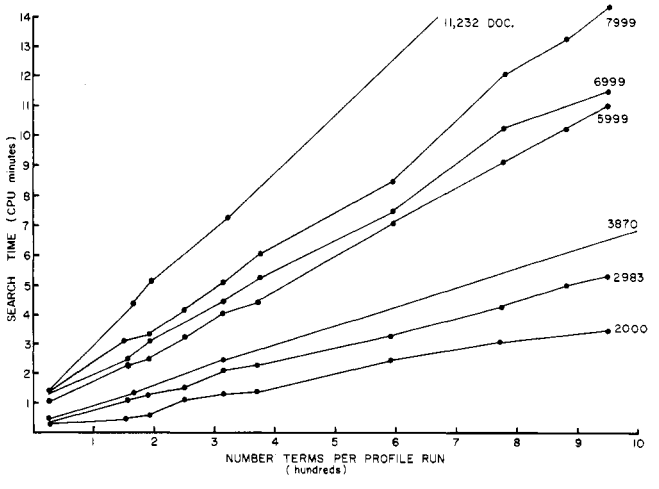

Figure 1. Search time versus number of profile terms

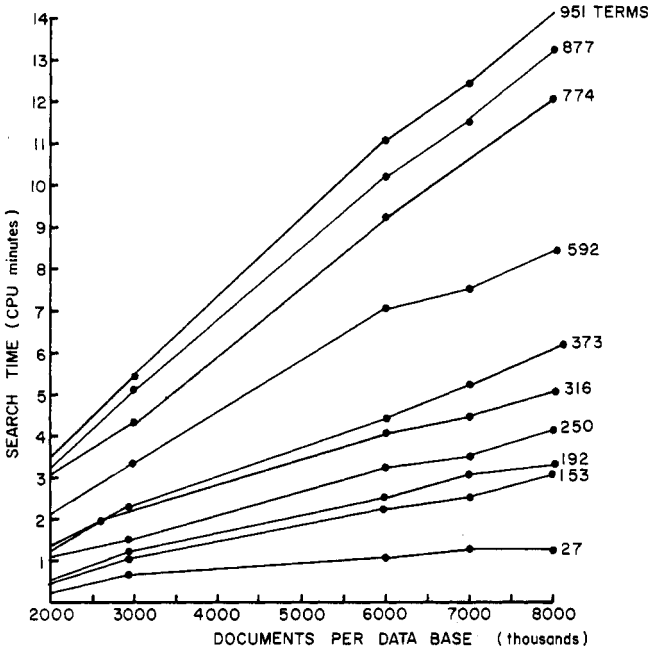

Figure 2. Search time versus number of documents in the data base

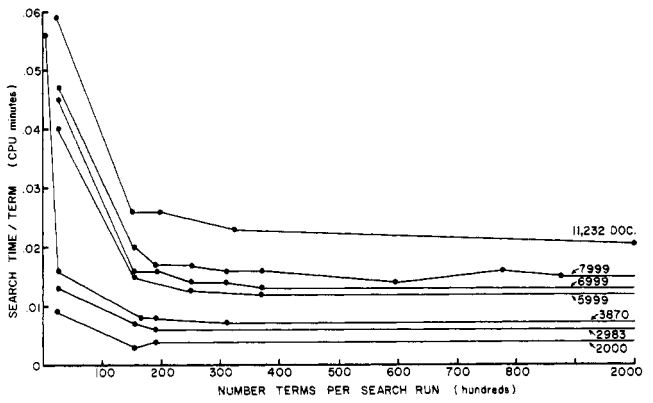

Figure 3. Search time per term for different data bases

per term with one profile containing 5 terms and 0.012 minute per term with 150 profiles and 3804 terms. In the smaller data base, 2000 documents, the average search time per term with a run of 27 terms was 0.009 minute as compared to 0.003 minute per term with a run of 951 terms. With fewer than 10 profiles (200 terms), the search time per term was comparatively very high.

As mentioned previously, search time is a function of both terms searched and data base size. Each is a linear function of search time when the other variable is held constant. Three linear regression models for predicting search time using the number of profiles, terms, and characters in a profile run were fitted for each data base using standard statistical methods. Each of these models accounted for approximately 95% of the variability in search time for a given data base size. In most cases, the number of terms as the predictor gave a smaller error of estimate of search time than did either the number of profiles or the number of characters. For this reason, a general model was developed using the number of terms.

A plot of the regression coefficient (slope of the line) for search time on the number of terms over data base size is a straight line. Thus, since the regression coefficients are also linear with respect to data base size, a general model can be developed from the following relationships.

For a given data base, search time follows the function

$$(1) \qquad ST = \overline{ST} + b(T - \overline{T})$$

where ST equals search time, b is the regression coefficient, and T is the number of terms. $\overline{ST}$ and $\overline{T}$ are the respective means.

Since b is linear with respect to data base size, it can be extended to a data base of any size by the following relationship.

$$(2) \qquad b'' = \overline{b}' + b' (D - \overline{D})$$

where b' is the regression of b on data base size and D is the number of documents. Thus, search time then becomes

$$(3) \qquad ST = \overline{ST} + b'' (T - \overline{T})$$

Substituting (2) into (3), search time can then be expressed as

$$ST = a + tT + dD + fDT$$

where a is a constant and t, d, and f are regression coefficients to be estimated.

The estimates as determined for the *CA Condensates* data base on the IBM 360/65 in this study are shown in the following equation:

$$ST \text{ (mins.)} = (2031.552 + 3.46839T + 1.20747D + .01724TD)10^{-4}$$

This model accounted for 99% of the variation in search time. A table of search times for profile sets ranging from 50 to 5000 terms and for *CA Condensates* data bases of 200 to 260,000 documents has been constructed for use in estimating CPU run times in the information center. Similar tables will also be constructed for *Chemical-Biological Activities*, *Biological Abstracts*, and *BioResearch Index* after analysis of each for the number of words per document. All these data bases are searched with the same CAS search system.

ACKNOWLEDGEMENT

This work was partially supported by the National Science Foundation under Grant GN-851.

LITERATURE CITED

(1) "Preparation of Search Profiles," Chemical Abstracts Service, Columbus, Ohio, 1967.

(2) "Text Searching," Chemical Abstracts Service, Columbus, Ohio, 1968.