DOCUMENT RESUME

ED 077 522 LI 004 349

AUTHOR	Wilmoth, James Noel
TITLE	Methods of Cost Reduction in Information Retrieval.
INSTITUTION	Auburn Univ., Ala. Foundations of Education Dept.
PUB DATE	27 Apr 73
NOTE	19p.;(2 References); Paper presented at the Mid-Southeast Chapter, Association for Computing Machinery (Gatlinburg, Tenn., 27 April 1973)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Computer Programs; *Cost Effectivenes. Data Bases; Educational Research; *Educational Resources; Electronic Data Processing; *Information Retrieval; Information Systems; Search Strategies
IDENTIFIERS	Auburn University; *Educational Resources Information Center; ERIC; QUERY

ABSTRACT

Cost effectiveness of the QUERY program for searching the Educational Resources Information Center (ERIC) data base has been an important issue at Auburn University. At least two broad categories of costs are associated with information retrieval from a data base such as ERIC: fixed costs or overhead and data base associated costs. The concern at Auburn reflected in this paper, has been particularly with three data base associated costs: preparation of records before searching, selection of a software search system, and implementation of search strategy. Computational algorithms associated with costing particular jobs should have terms applicable to each of these variables. The first two are rather straightforward to implement, but the third, because it is a human variable, is difficult to access. The most powerful approaches to cost reduction seem to be associated with partitioning of the data base into components specific to the needs of the user population. (Author/SJ)

METHODS OF COST REDUCTION

IN INFORMATION RETRIEVAL*

James Noel Wilmoth
Coordinator, Information Retrieval
Foundations of Education Department
School of Education
Auburn University

*Paper presented at the April 27, 1973 meeting of the Mid-Southeast Chapter, Association for Computing Machinery, at Galinburg, Tennessee.

I. INTRODUCTION

Traditionally computers on university campuses have found their greatest level of usage among students and faculty capitalizing on their very efficient "number-crunching" capabilites. While the mε ·hine has supported, in this manner, rich activity in the natural sciences and productive development of researchers and business managers in the social sciences; it has, nevertheless, been of little service to a larger segment of students and faculty whose activities do not have a fundamental numerical data basis.

As might be expected, the Federal Government has pioneered, largely as a by-product of its space exploration and department of defense efforts, other applications of the computer. One of these applications having significance to scholars is "information retrieval" which may be defined as recovery of document references from data sets constructed from highly sophisticated and scientific indexing procedures. Each recovered reference, or "hit", under control of a search program, may provide the scholar or researcher with several information fields all of which may serve to indicate the potential value of the original document to the topical area for which information is desired. Figure 1 lists the searchable fields of the ERIC[1] data base used in teacher education and the teaching profession.

[1] ERIC is an acronym for Educational Resources Information Center. The system is encoded under sponsorship of HEW.

FIELD NAME	FIELD ID FOR QUERY
Accession Number	A
Clearinghouse Accession Number	B,C
Program Area	D
Publication Date	E
Title	F
Personal Author	G
Institution Code	H
Sponsoring Agency Code	I
Descriptor	J
Identifier	K
EDRS Price	L
Descriptive Note	M
Issue	N
Abstract	O
Report Number	P
Contract Number	Q
Grant Number	R
Bureau Number	S
Availability	T
Journal Citation	U
Institution Name	V
Sponsoring Agency Name	W
Not Used	X,Y,Z

Figure 1.--Searchable fields in each record of the ERIC data base are combined with logical operators into "search requests" for references to documents originating in the teaching profession. Indexing in most fields is very precisely controlled.

At least two broad categories of costs are associated with information retrieval from a data base such as ERIC: (1) Fixed costs or overhead and (2) data base associated costs. Our concern will be with three dimensions of data base associated costs particularly (1) preparation of records before searching, (2) selection of a software search system, and (3) implementation of search strategy.

II. PREPARATION OF RECORDS BEFORE SEARCHING

If two years of experience at Auburn University has shown anything concerning optimization of information retrieval efforts, it has shown the need for reformatting files preparatory to high volume searching by students and faculty. Initially, we installed the QUERY search system and the ERIC tapes in their essentially unaltered forms; but, as our experience with the system increased we discovered the need for adjusting the data base to better accomodate the constraints of our budget and the priority assignment schedules of our installation. Also, since the complete ERIC data base as received from Washington resides on four tapes, adjustments permitting the mounting of fewer tapes decrease our turn around times.

One of our first moves was the partitioning of each data set of the large data base into 25 smaller ones. In each search request, thereby, we searched a fraction of the total records in the full data base. In our case, we partitioned on the clearinghouse field according to 23 valid alphabetic combinations, a set of invalid alphabetic combinations, and records for which a clearinghouse entry was missing. Figure 2 presents the latest analysis (December, 1972) on which partitioning was based. (ERIC Tapes consist of two data sets: RESUMAST AND CIJEMAST.)

The partitioned data sets were placed on nine tapes in sequence such that those data sets for clearinghouses usually searched together resided

Clearinghouse	Cumulative	
	CIJE Dec., 72	RIE Dec., 72
AA	13,255	986
AC	2,054	4,060
AL	193	1,789
CG	4,810	3,074
CS	34	192
EA	2,838	2,959
EC	2,945	2,665
EF	280	1,403
EM	2,710	2,882
FL	3,091	2,465
HE	2,598	2,320
JC	371	2,071
LI	2,372	2,141
PS	2,049	1,997
RC	978	2,854
RE	4,106	1,932
SE	6,451	3,452
SO	1,182	1,393
SP	1,427	2,808
TE	2,558	2,795
TM	640	1,646
UD	1,717	2,796
VT	4,005	5,049
bb	65	3,789
Errors	22	41
Total Records:	62,751	59,559
Avg. Record Length (Bytes):	429	1,595
Size of File (Bytes):	26,941,991	95,004,065
Size of longest Record (Bytes):	886	3,370
Size of Shortest Record (Bytes):	150	67

Figure 2.--A specification of clearinghouses and the number of records contained in each for the December, 1972 cumulative issues of Research in Education (RIE) and Current Index to Journals in Education (CIJE) ERIC Tapes.

on the same tapes. Moreover, clearinghouses from the two source data sets (RESUMAST and CIJEMAST) were respectively ordered such that concatenation provided ready access to both sources in a single jobstep. With this alteration data set patterns of Figure 3 resulted. Everyone searches the

Tape Organization of Data Sets Searched Together by Tape Number and by Label--
The Suffix "DEC72" Should be Added to Each Name

SER=	TP0889	TP0217	TP03600	TP0880	TP0092	TP0186	TP0863	TP0091
LABEL=								
1	C.XX	C.EM	C.VT	C.PS	C.CG	C.SP	C.AL	C.FL
2	R.XX	R.EM	R.VT	R.PS	R.CG	R.SP	R.AL	R.FL
3	R.AA	R.VT	R.RC	R.EC	R.EA	R.HE	A.FL	R.TE
4	C.AA	C.VT	C.RC	C.EC	C.EA	C.HE	C.FL	C.TE
5		C.LI	C.UD	C.UD	C.EC	C.JC	C.RE	C.SE
6		R.LI	R.UD	R.UD	R.EC	R.JC	R.RE	R.SE
7		R.PS		R.SP	R.SP	R.AC	R.TE	R.SO
8		C.PS		C.SP	C.SP	C.AC	C.TE	C.SO
9		C.EF						C.TM
10		R.EF						R.TM
Content Area Alphabetic Character-ization	Errors and Misc.	Materials and Facilities	Vocational Rural, and Disad-vantaged	Elementary Methods and Theories	Counselling, Guidance, and Admin-istration	Post High School Education	Language Arts	Subject Matter Methods and Theories
Appended in each DSN	A	A	B	C	D	E	F	G

Figure 3.--The Auburn University pattern for partitioning ERIC Tapes into clearinghouse data sets and of sequencing on specific tapes those clearinghouses usually searched together. Note the back to back arrangement of RIE (R.) with CIJE (C.) sources and the appearance of some clearinghouses on more than one tape.

"Errors and Miscellaneous" tape in addition to the particular clearinghouse data sets of interest to him. (The release date of the original data set is reflected in the partitioned data set names.)

The next alteration to be implemented for the Auburn University working tapes was the reversing of records on the tapes. As the original data bases are obtained from Washington, the most current records are the last ones on the tapes. With records reversed the user obtains the latest information from the data sets being searched, perhaps getting all the information in which he is interested even if his job does not complete for some reason. Before implementing the reversing technique a job having a broad search request might fail on time, on number of pages, or on exceeding the working space set aside for storage or sorting of hits; in which cases the job would yield the oldest references only, obliging the user to enter one or more follow-up jobs. With data sets now reversed, a user, having obtained hits containing the latest references to his topic, rarely feels compelled to submit follow-up jobs unless he is interested in having a more complete historical perspective on his topic. In this case he may choose to search the unpartitioned, originally ordered data base with his follow-up job.

Partitioning may also be profitable in non-education fields. The chemist may wish to create sub-sets according to natural sub-divisions within his discipline; for example, subsets reflecting inorganic, organic, physical, instrumentation, etc. Likewise the physicist, the biologist, or the medical scientist may fracture his data base into more manageable components tailored to the interests of major users at his installation.

A third adjustment that may conserve machine time for each search request and which we are studying at Auburn University will result in changing the order of information within the records. The overwhelming majority (probably over 99%) of search requests contain search elements for inspecting one or both of the same two fields in each record: the clearinghouse field and the descriptor field. As the records are fed to the system these are the second and tenth fields respectively. They are accessible by chaining from one field to another until the field of interest is located. It is probable that we will restructure the next versions of our data bases such that the clearinghouse field will be the first in each record followed in order by the descriptor field. All other fields will appear behind the descriptor field.

Another possibility for conserving time in the search step relates to deleting some of the little used fields from each record, particularly in the records of the partitioned data sets. It is rare that educators at Auburn use, if they ever have used, information from the following fields: Institution Code, Sponsoring Agency Code, Report Number, Contract Number, Grant Number, and Bureau Number. Both the institution and the sponsoring agency are named in separate, specific fields. We have simply not had occasion to use the other surplus fields mentioned.

III. CONSIDERATIONS IN THE SOFTWARE SEARCH SYSTEM

Our experience with the QUERY search system on tapes having records reversed such that the latest records are searched first and having all fields for each record in their original, unaltered order has shown that features built into the software system can significantly affect the cost of information retrieval. Three principles are apparent from this experience.

The first principle we discovered is that it is much less expensive of machine time to search on several search requests per job step rather than utilizing several job steps, each job step searching a data set to satisfy a single search request. This capability of QUERY to do variable batch searching is one of its salient features which, for our purposes, make it more attractive than "on-line" systems. The batch mode allows us to service more than one need or client at a time recognizing our trade-offs with this method to be sacrifice of

(1) Interactive capabilty,

(2) High precision in selection (editing) of printed output,

(3) Browsing capability for studying before printing, and

(4) On-line search alteration potential.

As one considers the costs associated with batching versus on-line hookups he should also be aware that the latter may tie up tape-drives and telephone lines; therefore, respective cost formulas for information retrieval may differ considerably in the formula terms associated with equipment items. At some installations there is an added $25 per CPU terminal hour and an additional $10 per telephone hour with a dial-up connection. The ideal might be to search in batch processing mode, and edit or study in time-share mode.

A second principle that should be considered for cutting search costs is to select or implement a system that cuts down on core load for retrieval. A system that preprocesses all coordinate indexed descriptors to their accession numbers for logical operations between accession numbers should be more efficient with retrieval core load than one that loads the complete content of each record and performs logical operations on the contents of one record at a time. Software called RIC and GANDALF[1] are reported to capitalize on this principle in their application to ERIC type data bases.

A third principle relates to provisions in the software for handling gross, vague, generalized, "shotgun," or otherwise "dirty" types of search requests. A provision for terminating searching on a search request, but perhaps continuing to count the number of hits, say 50, could circumvent the output of thousands of pages each of which is likely to be too wide of the specific interest of the user to be of value to him. After the 50th hit the software could issue a message, perhaps

> n,nnn,nnn records have been processed, fifty hits have been found for search request nnnn. Saving and printing are terminated, counting is continued for this search request.

[1] RIC (Research Information Center) of Grand Forks, North Dakota and GANDALF from Southwest Research Associates in Albuquerque, New Mexico.

A further provision for terminating the counting operation might be implemented after, say, 300 hits have been counted, printing a message such as the following

> Three hundred hits have been counted with the processing of n,nnn,nnn records for search request number nnnn. The count operation is terminated for this search request.

IV. IMPLEMENTATION OF SEARCH STRATEGY

The programmer is more important in holding down the cost of information retrieval than any other variable. Given a debugged search software package the programmer may write a search request having varying levels of efficiency. These levels of efficiency are dependent, in part, upon logical considerations such as the following:

1. Narrowing a search by tying it as soon as possible to fields having the most limited number of matchable retrieval constants.

2. Searching on the shortest fieids first.

3. Using the "AND" operator early in the search field in conjunction with those retrieval constants having the fewest number of occurrences in the data base.

4. Selecting major descriptors as retrieval constants rather than minor, or rather than the major form together with minor form.

If an educator is searching an unpartitioned ERIC data base the efficiency of his search request will be improved if he narrows the search base by immediately tying the request to a field having but a few possible constants as potential retrieval keys. The most practical of these fields are the clearinghouse and the date fields. By limiting the request to those records associated with one or a few clearinghouses the number of records searched for the later fields specified in the search request will be reduced to a fraction of the records in the data base being searched. Similarly, a search confined to the activity on a topic over a subset of time covered by the tape(s) will reduce the average machine time used for each record searched. Searches over subsets of time may also be effected with the accession number fields.

Searching on shortest fields first likewise conserves machine time as fewer comparisons need to be made to determine if a record ıs a potential hit. The clearinghouse, date, and accession number fields mentioned above are each short fields in comparison with the descriptor or abstract fields. Another short field of practical value to fulfillment of this principle is the author field. If the researcher has reasonable certainty that all records treating a topic will contain a particular word or phrase in their title fields then that field may also be an efficient one to search because of its relative shortness. We have not experienced much success with the title field, however.

The judicious use of the "AND" operator requires a statistical reporting of the frequency with which constants occur in specific fields of the data base. Such information is available to the educator by data base (RIE and CIJE) for two fields: (1) the descriptor field,[1] and (2) the identifier field. Having consulted such a statistical report the information retrieval specialist is prepared to structure his search request with a logical form such that the absence of the less frequent retrieval constants is detected early and the record is edited from further searching. The researcher may also discover from the statistical report that such a pivotal term appears only a few times so his most

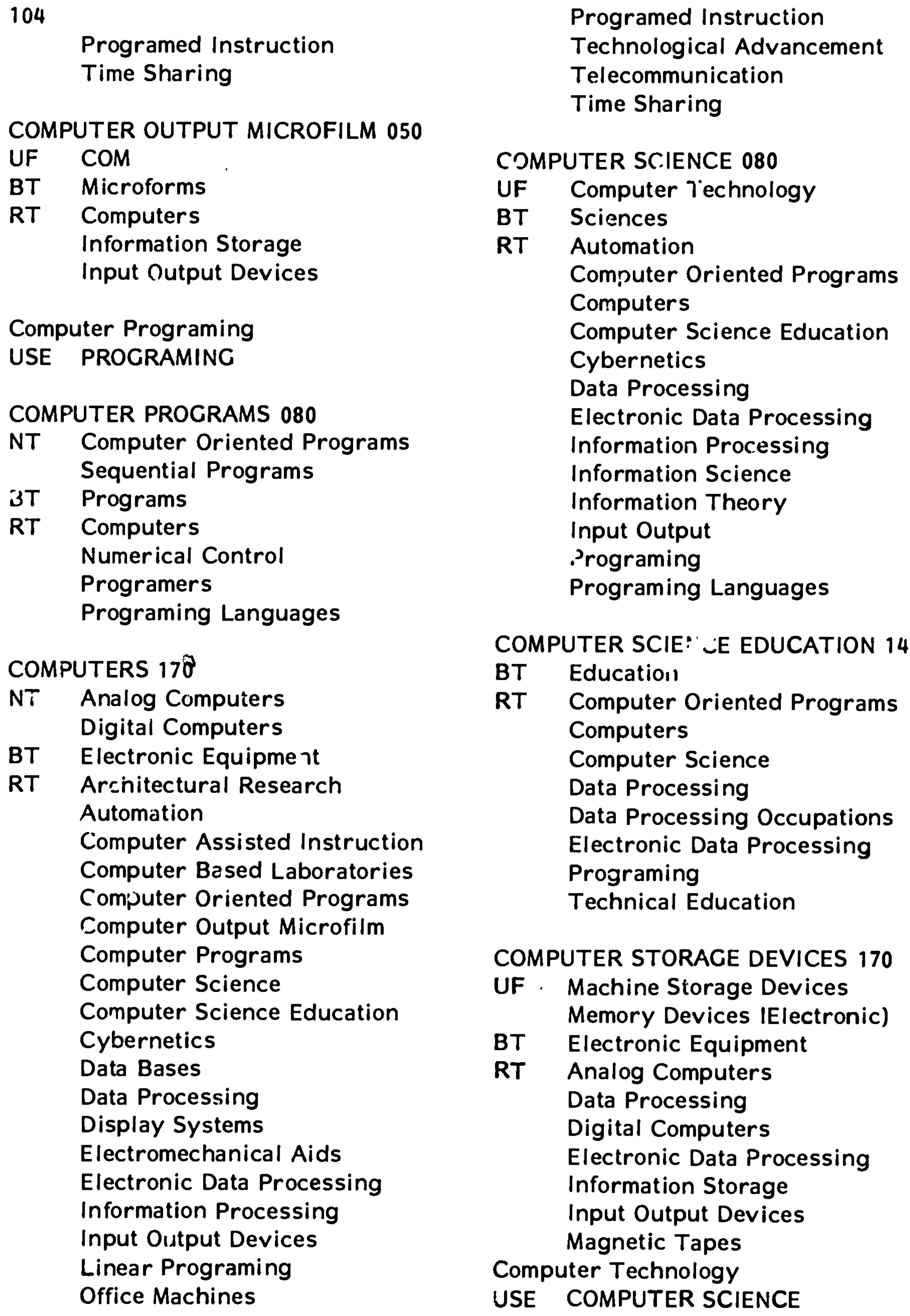
[1] ERIC Processing and Reference Facility, ERIC Descriptors (RIE Edition): Term Postings & Statistics for Research in Education (Bethesda, Md.: Leasco Systems and Research Corporation, December, 1972).

productive search request is one focused exclusively on that term; therefore the search request will contain just one search element having that term as its retrieval key.

One of the most difficult principles for student programmers to realize is that of basing the first search for a topic on major descriptors only. This seems to relate somehow to the listing of descriptors in the Thesaurus.[1] Many students operate under the misconception that a Thesaurus entry, such as Figure 4, contains the major term in bold print and that appearance of the term as a minor descriptor in a search element activates all records having the non-bold broader terms (BT), narrower terms (NT), and related terms (RT) as potential hits. Needless to say, this misconception, besides being wasteful of programmer and machine time by requiring complete reanalyses of the search request and a second or third job for the machine, creates many frustrations for the student on examination of his output.

Program writing variables, in the larger context of complete decks for each job submitted, may also include profitable selection of JCL options to produce information recovered at the time an error occurs. At Auburn we found the need to include conditional parameters whereby sorting and printing would be done even if a job abended. Our users abnormally terminate most frequently with S322 (time) and SB37 (too many hits for allocated search space). In the first case a time parameter on the EXEC card of one minute less than the time specification on the JOB card permits execution of the SORT

[1] ERIC: Thesaurus of ERIC Descriptors (New York: CCM Information Corporation, 1972).



104
Programed Instruction
Time Sharing

COMPUTER OUTPUT MICROFILM 050
UF COM
BT Microforms
RT Computers
Information Storage
Input Output Devices

Computer Programing
USE PROGRAMING

COMPUTER PROGRAMS 080
NT Computer Oriented Programs
Sequential Programs
BT Programs
RT Computers
Numerical Control
Programers
Programing Languages

COMPUTERS 170
NT Analog Computers
Digital Computers
BT Electronic Equipment
RT Architectural Research
Automation
Computer Assisted Instruction
Computer Based Laboratories
Computer Oriented Programs
Computer Output Microfilm
Computer Programs
Computer Science
Computer Science Education
Cybernetics
Data Bases
Data Processing
Display Systems
Electromechanical Aids
Electronic Data Processing
Information Processing
Input Output Devices
Linear Programing
Office Machines
Programed Instruction
Technological Advancement
Telecommunication
Time Sharing

COMPUTER SCIENCE 080
UF Computer Technology
BT Sciences
RT Automation
Computer Oriented Programs
Computers
Computer Science Education
Cybernetics
Data Processing
Electronic Data Processing
Information Processing
Information Science
Information Theory
Input Output
Programing
Programing Languages

COMPUTER SCIENCE EDUCATION 140
BT Education
RT Computer Oriented Programs
Computers
Computer Science
Data Processing
Data Processing Occupations
Electronic Data Processing
Programing
Technical Education

COMPUTER STORAGE DEVICES 170
UF Machine Storage Devices
Memory Devices (Electronic)
BT Electronic Equipment
RT Analog Computers
Data Processing
Digital Computers
Electronic Data Processing
Information Storage
Input Output Devices
Magnetic Tapes
Computer Technology
USE COMPUTER SCIENCE

Figure 4.--A part of a page from the Thesaurus of ERIC Descriptors, 1970, page 104.

and PRINT steps. The catalogue procedure provides condition codes for executing SORT and PRINT steps even if the job abends in the SEARCH step.

If the information retrieval programmer forsees considerable demand for particular search requests, it may be beneficial for him to maintain files of preprocessed requests. It is much cheaper repeatedly to dump such a file than to repeatedly search a larger file on recurrent requests having essentially the same search logic. At Auburn University we have seen the need for this approach with "Educational Accountability," "Career Education," "Rehabilitation," and with descriptors related to "performance based teacher education."

V. IMPLICATIONS FOR DATA BASE COST DETERMINATIONS

It can readily be seen that intensive research and development efforts within installations after they have begun amassing experiences with information retrieval are essential background activities before implementation of a formula for determining the non-fixed cost associated with each search request. The human factor seems to prohibit specification of a generalizable algorithm that would be highly reliable within or between installations. Some attempts at analyzing retrieval costs, however, have circumvented all variables associated with the data base by charging the user according to one or a combination of the following techniques:

1. Number of terms searched.
2. Ratio of high frequency to low frequency terms.
3. Ratio of frequency of OR's to frequency of AND's in the search request.
4. Number of hits generated.
5. Flat rate charges varying from nothing beyond actual CPU charges to CPU charges plus ten, fifteen, or twenty-five dollars.
6. Actual CPU charges plus twenty-five percent.
7. Annual charges associated with a contract; such as

   A. For a department of education, $500 per professor per year witn the requirement that all professors in the department be enrolled.

B. For a school district, twelve cents per student (in average daily attendance) per year with searching options available to professional personnel only.

VI. SUMMARY

As noted above, experience with data bases for teacher education at Auburn University has indicated the necessity of considering the organization of the data base searched, the properties of the software package used for searching, and the talents of the search strategist, in any attempt to tally the non-fixed costs of information retrieval. Computational algorithms associated with costing particular jobs should have terms associated with each of these variables, the first two being rather straightforward to implement; but, the third, because it is a human variable, is difficult to assess.

Perhaps the most powerful approaches to cost reduction are associated with partitioning of the data base into components specific to the needs of the user population. In addition to breaking down the data set into sub-files by areas (such as can be done in education with the clearinghouse field) one might also restructure his data-base into other subfiles according to level, for example, elementary, secondary, and higher education; according to the individuals or groups served--teacher, counsellors, and administrators; according to recurrent requests for "critical issue" topics or products: "vandalism in the schools," "disadvantaged," "teacher negotiations," or others; according to journals reported--Harvard Educational Review, AREA Journal, Chemical Education, or others; and according to exlempary documents such as state of the art monographs, curriculum packages, legislation, court decisions (San Antonio vs Rodriquez), or others.