Analysis of Term Distribution in the TOXLINE Inverted File

TAMAS E. DOSZKOCS*
College of Library and Information Services, University of Maryland, College Park, Maryland 20742

ROBERT J. SCHULTHEISZ and BRUNO M. VASTA
Toxicology Information Program, Specialized Information Services, National Library of Medicine, Bethesda, Maryland 20014

Received January 26, 1976

The paper describes the term generation algorithm for the TOXLINE free-text retrieval system and examines the statistical frequency distribution of search terms. The authors propose several suggestions for the utilization of analysis data.

INTRODUCTION

TOXLINE (Toxicology Information On-Line), which has been in operation on the National Library of Medicine's computer system since April 1974, is an interactive bibliographic retrieval system for toxicology information.[5] At present, this data base consists of approximately 374 000 references to published human and animal toxicity studies, effects of environmental chemicals and pollutants, adverse drug reactions, and analytical methodology. Each record in the TOXLINE data base contains a full bibliographic citation, most with abstracts and/or indexing terms and Chemical Abstracts Service (CAS) Registry Number used to uniquely identify chemical substances.[12] On-line retrieval from TOXLINE is based on free-text searching of most terms appearing in titles, keywords, and abstracts, either singly or in combination, by means of Boolean operators.[10]

The availability of search terms for the textual data in TOXLINE is controlled by two distinct factors: (1) term generation and (2) stopwords. In the process of generating the free-text terms to be used in the inverted file, we became interested in these terms with respect to their frequency, distribution, and recursiveness. We attempted to determine whether an analysis of these terms would be of use in preparing search aids for the users of TOXLINE.

TOXLINE COMPONENT SUBFILES

The TOXLINE data base is derived from five major secondary sources and one archival collection of bibliographic references. Included in the TOXLINE file are over 374 000 abstracts and citations from *Toxicity Bibliography*; *Pesticides Abstracts* (formerly *Health Aspects of Pesticides Abstract Bulletin*); *Abstracts on Health Effects of Environmental Pollutants*; *Chemical-Biological Activities*; and the Hayes File on Pesticides. These component subfiles, shown in Table I, are arranged so that they can be searched simultaneously in response to a single query. A complete description of the TOXLINE component subfiles has been reported previously.[10,13]

TOXLINE INVERTED FILE TERMS

The TOXLINE inverted file text terms were derived from words in document titles, abstracts, and keywords. Other inverted file terms, e.g., author, CAS Registry Number, CODEN, publication year, etc., are taken directly from an input data element without further processing.

This collection of literature contains terms with various significant special characters, e.g., chemical names with hyphens, commas, and parentheses. Since many chemical names denote a chemical structure and have special characteristics, an algorithm was devised to create search terms reflecting structural fragments. For example:

> A substance having a definite chemical structure and a complex systematic name also has a relatively simple synonym, MER-25. Since this naming convention—alphabetics and numerics separated by hyphens—is commonly used in the literature, a rule was devised to retain this synonym as an entity. But, additional terms derived from the systematic name
>
> alpha, (4-(2-(diethylamino)ethoxy)phenyl)-4-methoxy-alpha-phenyl-
>
> were also created as a set of primitives for search purposes:
>
> alpha diethylamino ethoxy phenyl methoxy

Table I. Major Component Subfiles of TOXLINE

Acronym	File name	Source of file	No. of records
HAPAB PESTAB	*Health Aspects of Pesticides Abstract Bulletin*	EPA	15 593
TOXBIB	*Toxicity Bibliography*	NLM	98 724
CBAC	*Chemical-Biological Activities*	CAS	193 211
HEEP	*Abstracts on Health Effects Environmental Pollutants*	BIOSIS/ NLM-TIP	34 449
IPA	*International Pharmaceutical Abstracts*	ASHP	22 272
HAYES	Hayes File on Pesticide Toxicology	EPA	10 043

RULES TO GENERATE INVERTED FILE TEXT TERMS FROM TITLES AND ABSTRACTS

The rules to create the terms are:

Step 1. Terms are created from the character string which occurs between blanks and the special delimiters:

< > · (+ & · $ *) ; / % _ ? : # @ = ''

Step 2. The term from step 1 is examined for a hyphen (-) (EBCDIC Hex 60). If no hyphen is found, then step 3 is executed. Otherwise, one of these procedures is followed.

First—If a hyphen is followed by a single alphabetic character and preceded by a numeric and a comma (EBCDIC Hex 6B) and another numeric, then the hyphen is retained. The term string becomes a candidate for an inverted file term (e.g., 2,4-D) and is processed through step 5.

Second—If a hyphen is preceded by a one or two or three character alphabetic string and followed by a numeric string, then the hyphen is retained. The term string becomes a candidate for an inverted file term (e.g., MER-25). The candidate term string is processed through step 5.

Third—If a hyphen does not meet the first or second requirement, then the hyphen becomes a blank and the candidate term strings are processed through step 3.

* Author to whom inqueries should be addressed. Also with the Toxicology Information Program of the National Library of Medicine.

TOXLINE INVERTED FILE STATISTICS

(TOTAL TERMS)

POSTINGS

LOW RANGE	HIGH RANGE	NUMBER OF TERMS
1	1	379143
2	3	185138
4	7	42405
8	15	37423
16	31	16167
32	63	9296
64	127	6133
128	255	4129
256	511	2893
512	1023	2041
1024	2047	1216
2048	4095	756
4096	8191	400
8192	16383	191
16384	32767	107
32768	65535	33
65536	131071	9
131072	262143	3
262144	524287	1

NO OF POSTINGS	NO OF TERMS
20086330	687464

Figure 1a. Binary logarithmic distribution of all search terms.

TOXLINE INVERTED FILE STATISTICS

(TEXT)

POSTINGS

LOW RANGE	HIGH RANGE	NUMBER OF TERMS
1	1	117368
2	3	59442
4	7	14642
8	15	15608
16	31	9065
32	63	6475
64	127	4876
128	255	3445
256	511	2510
512	1023	1818
1024	2047	1115
2048	4095	719
4096	8191	391
8192	16383	186
16384	32767	102
32768	65535	29
65536	131071	8
131072	262143	3
262144	524287	1

NO OF POSTINGS	NO OF TERMS
17535101	237803

Figure 2a. Binary logarithmic distribution of text terms.

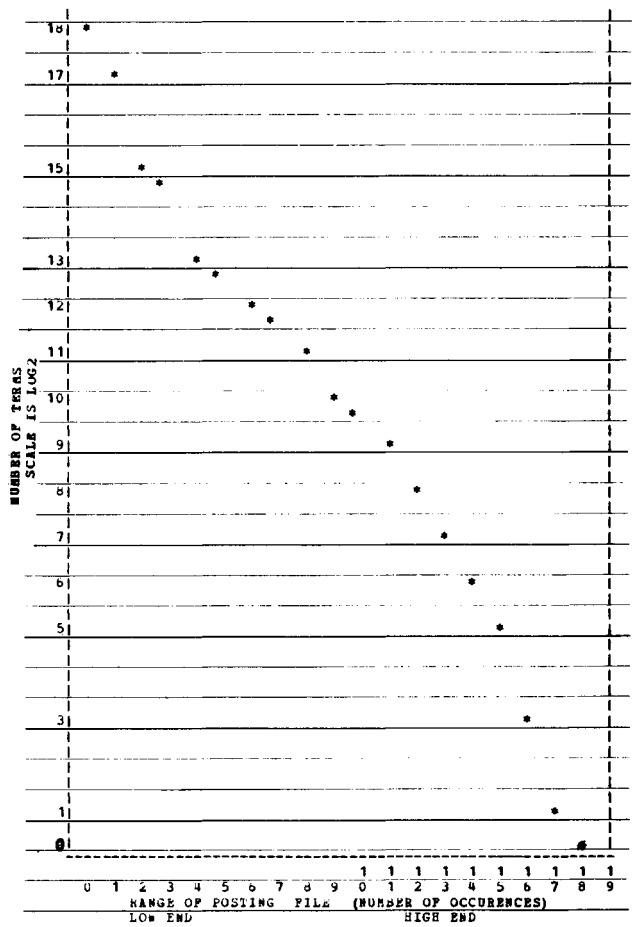

Figure 1b. Graphic representation of Figure 1a (all terms).

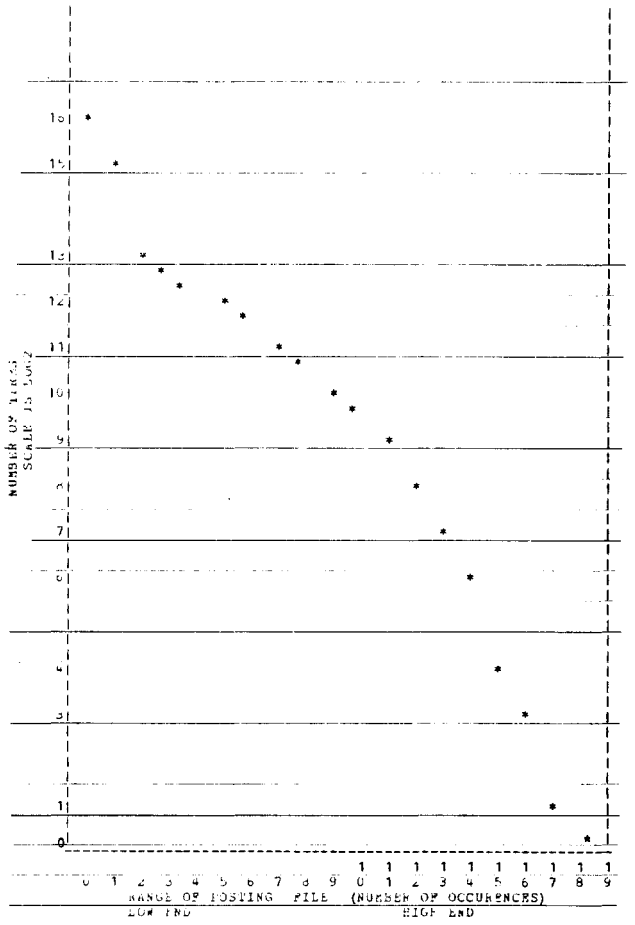

Figure 2b. Graphic representation of Figure 2a (text terms).

Step 3. The candidate term string is examined for commas. Any commas are replaced with blanks and the remaining candidate term strings are processed through step 4.

Step 4. The candidate term string is examined for alphabetics. If the term does not contain at least one alphabetic character, then the term does not become inverted file term (e.g., U235 becomes a term, 235 does not).

Step 5. If the candidate term string is greater than the maximum inverted file term length (36 characters), then the string is truncated on the right at the correct length, and the truncated term becomes the inverted file term. All truncated terms are documented for review.

**** 275051 *******	**** 27666 *******	**** 19808 *******	**** 14782 *******	**** 11299 *******	**** 9272 *******
COPYRIGHT	CAUSED	ENGLISH	CHEMICAL	GROUPS	RESIDUES
ABS	RATE	DUK	COMPARED	CHRONIC	CASES
CHEM	68	WATER	AMINO	INDICATED	RESULTED
CBAC	INCREASE	METHYL	BODY	TIMES	SENSITIVE
EFFECTS	TREATED	CONCENTRATION	COMPLICATIONS	INHIBITORS	CHILD
TOXBIB	SOME	HOWEVER	PLASMA	DEPENDENT	CONCNS
EFFECT	INHIBITION	REDUCED	INJECTED	VIVO	ROLE
INDUCED	AGENTS	ADMINISTERED	CARBON	REPS	CONTROLS
I	STUDIES	MOST	ACUTE	OBSD	DATA
HUMAN	DOSE	ALPHA	KIDNEY	65	ADOLESCENCE
NOT	CHANGES	L	OCCURRED	AGED	HORMONE
73	ABSTRACT	LESS	MECHANISM	14C	FOLD
DRUG	AGE	C	STIMULATION	AFFECTED	RNA
AFTER	ALL	66	OBTAINED	INHIBITORY	50-99-7
72	PRODUCED	REVIEW	GROUP	ACTIVITIES	FIRST
RATS	SYSTEM	RUG	FREE	HCL	PREGNANCY
71	74	MIDDLE	REACTION	O	PER
ACTIVITY	IPA	FORMATION	TESTED	APPLICATION	INTRAVENOUS
INCREASED	ASEP	ETIOLOGY	CLINICAL	ADDITION	PLANT
ACID	VITRO	ACTIVE	EXPOSURE	ETHYL	BINDING
ADVERSE	SYNTHESIS	HIGHER	STIMULATED	METHODS	DETERMINED
ADMINISTRATION	STUDY	EXPERIMENTAL	RELATED	UPTAKE	REACTIONS
CHEMICALLY	NORMAL	EFFECTIVE	WITHIN	EACH	INCUBATION
BLOOD	67	MIN	NEW	POTASSIUM	COMPLETELY
MALE	CELL	ORAL	WEIGHT	GAMMA	DIAGNOSIS
FEMALE	POISONING	ACIDS	CONDITIONS	ADDED	DNA
ANIMAL	GROWTH	BRAIN	CHLORIDE	PPM	PIGS
EXPERIMENTS	APPENDED	BETA	PRESSURE	SEVERAL	APPEARED
70	SODIUM	SAMP	DEVELOPMENT	SPECIES	CENTRAL
DECREASED	ANIMALS	LOW	PERIOD	ANY	ABSORPTION
TREATMENT	SH	OBSERVED	APPARENTLY	ABS	DESCRIBED
USE	DOSAGE	AGAINST	TESTS	CBAC	BIOSYNTHESIS
THERAPY	COMPOUNDS	SERUM	METHOD	CHEM	DIET
LIVER	ANALYSIS	PATIENTS	USING	MEMBRANE	INCREASING
BIOL	DISEASES	RABBITS	INFLUENCE	PLANTS	EXCRETION
METABOLISM	FACTORS	SIGNIFICANTLY	CONTAINING	INCORPORATION	RESISTANCE
KEEP	USED	ENGLISH	PHOSPHATE	APPLIED	FATTY
DURING	II	VARIOUS	FUNCTION	OVER	COMPARATIVE
69	DAY	TOXIC	GUINEA	PH	OXYGEN
RAT	HIGH	V	SPECIFIC	POSSIBLE	CALCIUM
BOTH	FOLLOWING	SIGNIFICANT	URINE	CONCN	PESTICIDES
ACTION	RESPONSE	DECREASE	PATHOLOGY	ORALLY	LARGE
INHIBITED	PRESENCE	CONCENTRATIONS	M	FOLLOWED	DISTRIBUTION
MORE	N	HEART	PROPERTIES	RABBIT	WEEKS
MAY	INJECTION	DIFFERENT	HAPAB	PRODUCTION	RUSSIAN
TOXICITY	TISSUE	GERMAN	PRESENT	RELEASE	E
CONTROL	PROTEIN	TOTAL	SINGLE	X	DDT
LEVELS	FOUND	GREATER	WELL	CASE	PREVENTION
MICE	LEVEL	EITHER	SMALL	DERIVATIVES	FOOD
OTHER	DOSES	CONTENT	SULFATE	ACETATE	RESP
CELLS	RESULTS	D	AFFECT	TEST	METABOLIC
SHOWED	STUDIED	RUSS	RESPECTIVELY	CHANGE	CHROMATOGRAPHY
TIME	PHARMACODYNAMICS	GLUCOSE	MEDIUM	THROUGH	ENHANCED
BETWEEN	ENZYME	DISEASE	MUSCLE	NEOPLASMS	GENERAL
P	DRUGS	INJECTIONS	VALUES	RELATION	MARKEDLY
DAYS	6	COULD	DAILY	ENZYMES	ALONE
S	LABELED	DOGS	LOWER	TISSUES	NERVOUS
ADULT	ISOLATED	UNDER	BEFORE	MOUSE	HYDROCHLORIDE
THERAPEUTIC	SIMILAR	DISCUSSED	SKIN	INDICATING	REQUIRED

Figure 3. Decreasing frequency listing of inverted file terms.

Step 6. The candidate term string is compared to the stopword list for TOXLINE. If the term is on the stopword list, then the candidate term string does not become an inverted file term.

STOPWORD LIST

The Stopword List consists of 171 words. These words commonly occur in the literature, but are not necessarily helpful in searching since the words show up in almost all documents. The Stopword List currently in use was taken from an earlier implementation of TOXLINE under a different retrieval software system. One of the objectives of our term frequency analysis was to evaluate and refine the Stopword List.

TERM DISTRIBUTION

In TOXLINE, if a term occurs more than once within the same document in the text field (title and abstract), the term will appear on the inverted file only once with the same document number reference, i.e., freq. (max) = 1 for any one term within a single document. Thus, the number of postings given for a term in the inverted file represents the number of distinct documents in which the term does occur, and not the total number of actual occurrences of the term over all the documents.

The inverted file term distribution in TOXLINE generally conforms to the Bradford–Zipf–Mandelbrot law[1,4,8,15] with over one-half (55.2%) of the 687484 unique terms having only one posting. In other words, more than half of the inverted file terms occurs in only one document.

Figure 1a represents the condensed binary logarithmic distribution of all free text search terms. These include authors, journal CODENS, language codes, CAS Registry Numbers, publication source and time delimiters, publication year and title, and abstract terms. The postings range indicates the number of postings expressed as a range value between successive powers of 2. The number of terms denotes the total number of terms that fall between the corresponding postings range.

Figure 1b is a "smoothed" graphic representation of the information in Figure 1a. The points along the axes are base 2 logarithmic values. The distribution of text terms (title words and words in the abstracts) is presented in Figures 2a and 2b.

Figure 3 is a listing of the inverted file terms in decreasing frequency (postings value) sequence. The number on top of each column represents the postings value of the first term in the column.

Figure 4 is a list of alphabetically arranged inverted file terms with element codes and actual postings values. The element code indicates the bibliographic field from which the term was derived, e.g., 165 stands for a text term.

The statistical analysis of the TOXLINE inverted file was undertaken with the objective of providing system and search improvement tools for the TOXLINE professional staff and our user community. Similar efforts in other organizations are given in references 2, 3, 6, 7, 9, 11, and 14.

DISCUSSION

In considering the potential use of a term distribution analysis for a particular file, one must consider whether this

TERMS	ELEMENT NUMBER	NUMBER OF POSTINGS	TERMS	ELEMENT NUMBER	NUMBER OF POSTINGS
ABBIA	168	1666	ADDITIONAL	165	2089
ABDOMINAL	165	1110	ADDITIVE	165	1457
ABERRATIONS	165	1210	ADDITIVES	165	1174
ABILITY	165	6376	ADDN	165	4310
ABLE	165	1680	ADENINE	165	2529
ABNORMAL	165	1990	ADENOCARCINOMA	165	1161
ABNORMALITIES	165	4592	ADENOSINE	165	5086
ABOLISHED	165	4443	ADEQUATE	165	1356
ABORTION	165	1151	ADIPOSE	165	2405
ABOVE	165	5207	ADMINISTERED	165	18883
ABS	165	9850	ADMINISTRATION	165	50291
ABS	165	6490	ADOLESCENCE	165	9010
ABS	165	228663	ADP	165	2152
ABSENCE	165	6798	ADRENAL	165	8036
ABSENT	165	1269	ADRENALECTOMIZED	165	1554
ABSORBED	165	2974	ADRENALECTOMY	165	1377
ABSORPTION	165	8616	ADRENALINE	165	3610
ABSTRACT	165	24852	ADRENALS	165	1045
ABUSE	165	2892	ADRENERGIC	165	5709
ACCELERATED	165	2465	ADSORPTION	165	1537
ACCIDENTS	165	1227	ADULT	165	28209
ACCOMPANIED	165	3979	ADULTS	165	1721
ACCORDING	165	2004	ADVERSE	165	52395
ACCOUNT	165	1770	AEHLA	168	1037
ACCOUNTED	165	1439	AEROBIC	165	1280
ACCUMULATED	165	2768	AEROSOL	165	1148
ACCUMULATION	165	6563	AEROSOLS	165	1376
ACETATE	165	9488	AERUGINOSA	165	1234
ACETIC	165	3811	AFFECT	165	11756
ACETONE	165	2122	AFFECTED	165	10679
ACETYL	165	2594	AFFECTING	165	3192
ACETYLCHOLINE	165	5587	AFFINITY	165	2147
ACETYLCHOLINESTERASE	165	1560	AFLATOXIN	165	1413
ACHIEVED	165	1803	AFTER	165	65067
ACID	165	4315	AGAINST	165	16641
ACID	165	57361	AGAR	165	1181
ACIDIC	165	1192	AGE	165	24822
ACIDOSIS	165	1375	AGED	165	10758
ACIDS	165	17062	AGENT	165	7844
ACRE	165	1357	AGENTS	165	25953
ACROSS	165	1196	AGGREGATION	165	1370
ACT	165	2894	AGRICULTURAL	165	2293
ACTED	165	1732	AGRICULTURE	165	1047
ACTH	165	2728	AIPTA	168	2828
ACTING	165	2124	AIR	165	7320
ACTINOMYCIN	165	3422	AJPHA	168	1720
ACTION	165	35216	AL	165	1122
ACTIONS	165	3683	ALANINE	165	3715
ACTIVATED	165	3836	ALBINO	165	1329
ACTIVATION	165	4878	ALBUMIN	165	3701
ACTIVE	165	17747	ALC	165	1188
ACTIVITIES	165	10634	ALCOHOL	165	7768
ACTIVITY	165	62024	ALCOHOLIC	165	2535
ACTS	165	1413	ALCOHOLISM	165	1937
ACUTE	165	14217	ALCOHOLS	165	1575
ADDED	165	9941	ALDOSTERONE	165	1214
ADDICTION	165	4485	ALDRIN	165	1698
ADDITION	165	10238	ALKALINE	165	3325

Figure 4. Alphabetic frequency listing of inverted file terms.

analysis will have value at the user/system interface level, or at the system performance level. Many useful pieces of information can be assembled and analyzed by the aforementioned term distribution procedure. For example, the user could be appraised of the search terms with variant forms that appear in the inverted file. This would serve to alert the user to consider using a truncated search term in order to take advantage of stems and term clusters. Additionally, a microfiche listing of highly posted terms could be prepared as a user guide. From a system performance point-of-view, the inverted file terms could be analyzed with regard to the possible elimination of those terms with only one posting. As shown previously, this would effectively reduce the size of the inverted file by at least 40%. However, one should only consider this action in conjunction with some mechanism which incorporates new terms and postings from updates to the file on the basis of adjusted frequency counts, thereby assuring terms with postings greater than one entry into the index file.

The term distribution analysis would also be of value when developing the Stopword List. Terms which are highly posted and have little or no retrieval value could become candidates for the Stopword List instead of carrying them in the inverted file. For example, terms like AFTER and AGAINST which have 65 067 and 16 641 postings, respectively, could be added to the Stopword List since they offer little, if any, value as a free-text retrieval term.

One could also make some intelligent decision as to what series of words could be clustered together under a single term to facilitate search retrieval. For example, the concept to DETERMINE has many variant forms including abbreviations appearing in the inverted file. Even if one were to

Term	Postings
DETERMINATION	8022
DETERMINATIONS	1082
DETERMINE	3919
DETERMINED	8884
DETERMINING	2064
DETG	*2
DETN	2059

truncate the term at DETERMIN*, one would miss the generally used abbreviations for DETERMINING & DETERMINATION as used by Chemical Abstracts Service in their CBAC component of TOXLINE.

Another aspect of the potential usefulness of term distribution analysis is the examination of the term length, i.e., how many characters are used to describe terms of high relative postings. From the systems optimizing approach, this would reflect the total number of characters required to store the terms in the inverted file.

The analysis identifies those terms which are the most highly posted. For example, it may prove useful to know the 100 most highly posted terms in the index file to serve as a guide to the users. The user would try to stay away from these terms, if possible, in order to achieve maximum retrieval efficiency, which translates in saving money in the on-line system. Once again, these user guides could be made available in the form

of microfiche. The usefulness of a microfiche of the entire inverted file as a potential user aid could be considered.

Following this same line of thinking, the analysis of terms in the inverted file can be broken down into individual components, for example, by identifying the high-frequency journal title (CODEN's) that appear in a file, or the high-frequency *author's names*, and the distribution of Chemical Abstracts Service (CAS) Registry Numbers that one finds in the inverted file. In each instance, a certain amount of implicit information is available which could influence the user's search strategy. One could easily produce the 100 most highly posted authors in the index file as a guide. In a similar fashion, the 100 most highly posted journal titles (realizing some inherent variability) could also be assembled as a user guide.

Finally, it seems possible that the analysis of term distribution could serve as a valuable tool in weighing the effects of merging one inverted file with another. In the existing software system, used by NLM, the merging of two free-text bibliographic files would essentially consolidate the inverted files, since many of the terms would be common to both files.

LITERATURE CITED

(1) A. D. Booth, "A "Law" of Occurrences for Words of Low Frequency," *Info. Control*, **10**, 4 (April 1967).
(2) Defense Documentation Center, "DDC Retrieval and Indexing Terminology", Vol. I, 1st ed, AD/A-001 200/5GA, 1975.
(3) Defense Documentation Center, "DDC Retrieval and Indexing Terminology", Vol. II, 1st ed, AD/A-001 201/3GA, 1975.
(4) R. A. Fairthorne, "Empirical Hyperbolic Distributions (Bradford-Zipf-Mandelbrot) for Bibliometric Description and Prediction", *J. Doc.*, **25**, 4 (Dec 1969).
(5) H. M. Kissman and D. J. Hummel, "TOXICON - An On-Line Toxicology Information Service", *Chem. Technol.*, 72 (Dec 1972).
(6) P. H. Klingbiel, "Multimillion Word Data Bases: A Preliminary Report", Vol. 1, Defense Documentation Center, AD-777 200/7, 1974.
(7) P. H. Klingbiel, "Multimillion Word Data Bases: A Preliminary Report", Vol. 2, Defense Documentation Center, AD-777 210/6, 1974.
(8) B. Mandelbrot, "An Information Theory of the Statistical Structure of Language", in "Proceedings of the Symposium on Applications of Communication Theory", Butterworth, London, 1953, pp 486-500.
(9) National Technical Information Service, "NTIS Master Frequency List of Subject Terms, July 69-Dec 72", NTIS/SR-74/04.
(10) "On-Line Services Reference Manual", National Library of Medicine, March 1975.
(11) P. B. Schipma, "Term Fragment Analysis for Inversion of Large Files", IIT Research Institute, Chicago, Ill., 1971, 17 pp.
(12) B. M. Vasta, "Use of TOXLINE/CHEMLINE for Retrieving Drug Information", *Drug Inf. Assoc. J.* (1975); presented Oct 28, 1974, at Drug Information Association.
(13) B. M. Vasta, "TOXLINE - NLM's On-Line Information Resource", Proceedings of the National Aeronautics and Space Administration's Annual Conference of Clinic Directors, Environmental Health Officials and Medical Program Advisors, March 18, 1975.
(14) M. E. Williams, "Handling of Varied Data Bases in an Information Center Environment", IIT Research Institute, Chicago, Ill., 1971, 24 pp.
(15) G. K. Zipf, "Human Behavior and the Principle of Least Effort", Addison-Wesley, Cambridge, Mass. 1949.

A Comparison of On-Line and Manual Modes in Searching Chemical Abstracts for Specific Compounds

JOSEPH SANTODONATO

Center for Chemical Hazard Assessment, Life and Material Sciences Division, Syracuse Research Corporation, Syracuse, New York 13210

Received March 10, 1976

Manual searching of *Chemical Abstracts* was compared with computer-searching *CA Condensates* and CBAC, especially with regard to retrieval of broad information on a specific compound. Differences are apparent owing to deficiencies in indexing and in the capability for free-text searching of abstracts in the computer-based systems. The success of manually searching the *Chemical Abstracts* Substance Indexes could not be equalled by the on-line systems, either alone or in combination.

Recent developments in the computerized searching of bibliographic sources which allow for on-line interaction and "browsing" of data bases have aided immeasurably in the efficient retrieval of chemical information. However, inherent limitations in the usefulness of these research tools have become apparent in the searching of specific chemicals as opposed to more generalized search strategies.

While developing a state-of-the-art review on the pesticide toxaphene, we felt that it would be helpful to compare the advantages of manually searching *Chemical Abstracts* with the combined results of machine-searching *Chemical Abstracts Condensates* and the CBAC portion of TOXLINE. To retrieve as many references as possible, we did not restrict our search strategy to provide only the most pertinent citations. In this sense, our literature search was intended to be a quantitative reflection of the published literature relating to toxaphene without making a value judgment regarding the relevance of individual articles.

It has been noted[1,2] that the indexing of *CA Condensates* is less systematic and exhaustive than that of the *Chemical Abstracts* Volume or Cumulative Indexes. In particular, *CA Condensates* appears to be inferior in the retrieval of citations for specific compounds.[3] Our experience has shown this to be especially true when the scope of a search on individual chemicals is extremely broad, requiring information in biological, chemical, and environmental areas. In other comparisons of results with *CA Condensates* and the weekly Keyword Indexes of *Chemical Abstracts*,[4] it was found that in certain instances either method may be superior or both may be equivalent. In quantitative comparisons of our own search results on individual chemicals, we have found that manual searching of Keyword Indexes yields the same number of citations as computerized retrieval because *CA Condensates* is constructed by using only the weekly Keyword Indexes. Superiority of manual searching becomes evident once an annual Volume Index is available. It is important to recognize that the weekly Keyword Indexes of *Chemical Abstracts* are constructed from characterizing words or phrases (keywords) selected from the title or context of an abstract. The Volume Indexes, on the other hand, are derived usually from a searching examination of the original documents, not the abstracts.[5]

The CBAC (Chemical-Biological Activities) data base, which was also used for comparison in the toxaphene search example, is offered through the National Library of Medicine's TOXLINE system. Coverage encompasses Sections 1 through