Data Retrieval by Text Searching

JOHN O'CONNOR

Center for Information Science, Lehigh University, Bethlehem, Pennsylvania 18015

Received January 31, 1977

Sixty percent of the data papers in an experiment were retrieved by human-computer text searching, in which the human contribution consisted of selection of search words for input to the computer search. Most of the successful retrieval consisted of identifying within papers those figures containing data asked for by the retrieval questions, and automatically labeling those data within the figures. The retrieval procedures are economically feasible now because they primarily require only that words from figures be in computer-readable form.

I. INTRODUCTION

A. Current Problems in Finding Data in Papers. Scanning titles and abstracts often fails to identify papers containing wanted data. For example, a random hundred data entries in a handbook[1] were examined together with the papers cited as sources for those data. In only 30% of the cases did a paper's title or abstract indicate that it contained such data.

Data-bearing papers are also frequently missed by retrieval services searching humanly assigned index terms. It is for this reason that the National Science Foundation is supporting studies of expanding such indexing to include "data flagging and/or tagging".[2]

Further, even when papers containing wanted data are identified, finding the data within the papers may take considerable time.

B. Use of Text Searching. The increasing availability of computer-readable full text of scientific journals suggests use of text searching to identify papers containing data of a desired kind and to locate the data within those papers. Computer-readable full text is becoming increasingly available as computer typesetting is more widely used for economic reasons;[3-7] for example, the American Chemical Society already publishes four journals in this way and expects to add three more during 1977.[8] Moreover, even journals which are not yet in computer-readable form can be effectively included in data retrieval by text searching. For most data in a paper are contained in its figures, and it is feasible to keyboard the words and numbers from figures into computer-readable form. For example, in the random hundred papers referred to above, the average number of figure words per paper was about 200 (counting chemical and other formulas as words); the average paper's figures also contained about 100 numerical entries (usually three digits or less).

Retrieval by text searching has several well-known advantages compared to the alternative of searching humanly assigned index terms. It saves the considerable cost of human indexing. It is also not limited by indexers' predictions of what content of papers will be of interest to users. It also has the further advantage, not yet realized by many information specialists, that it permits "passage retrieval".

In passage retrieval each output reference is accompanied by the passage(s) in the paper which caused it to be retrieved. This provides more precise retrieval than merely outputting whole papers-as-units, as do present retrieval services for scientific users. Many lawyers have had passage retrieval service since the mid-1960's.[9-11] A recent experiment suggests that passage retrieval for scientists is now feasible and will be very helpful.[12]

A passage retrieved for a data-seeking question is likely to *contain* the wanted data. In that case data retrieval or "fact retrieval" has been achieved. For example, for the question, "What are values for the surface tension of potassium?," tables such as Table I can be retrieved from within a paper[13] by text searching. The underlining of "surface tension" and "potassium" in this table directs the questioner's attention to the part of the table containing the requested data. Such underlining can be added during the text searching for a particular data retrieval question.

Table I. Surface Tension Data

Sodium		Potassium		Rubidium		Cesium	
Temp., °C	Surface tension, dynes/cm	Temp., °C	Surface tension, dynes/cm	Temp., °C	Surface tension, dynes/cm	Temp., °C	Surface tension, dynes/cm
141	208.0	77	116.9	104	85.4	71	71.6
169	206.0	95	114.0	129	84.8	70	74.4
213	197.6	120	111.9	153	79.5	123	69.9
256	191.6	151	111.0	179	79.8	163	68.5
261	195.4	204	109.0	222	76.5	162	73.1
261	192.7	242	106.9	252	74.2	218	64.1
308	190.7	259	103.8	264	72.7	217	69.5
356	184.7	297	101.1	300	72.8	268	64.9
408	176.0	349	93.6	313	71.6	326	61.5
418	168.0	401	90.5	351	67.7	376	59.5
463	176.5	464	86.5	370	69.7	422	54.5
516	163.9	511	82.8	393	65.3	462	51.3
570	171.2	564	75.6	415	65.3	486	53.3
627	152.8	616	75.7	457	62.0	516	47.6
649	152.7	673	68.8	463	61.4	567	48.0
684	134.8	729	66.0	465	60.7	566	45.9
684	140.0	783	61.9	509	61.3	624	45.2
729	132.6	846	59.4	513	57.6	674	43.6
730	135.2	897	53.6	515	60.5	742	34.8
784	129.2	947	56.4	564	58.2	793	34.2
782	124.0	993	49.6	613	57.9	851	29.6
793	122.4			613	55.9	858	29.0
842	115.1			619	52.8	867	27.9
890	116.0			649	50.7	883	30.6
893	111.0			671	51.8	962	25.7
992	99.5			724	49.9	986	29.1
				738	48.7	1011	26.6
				779	48.0		
				841	45.4		
				906	40.5		
				906	39.8		
				963	40.4		
				1006	33.5		

The example just given is relatively simple in two ways. First, there might be a language difference between a question and a table (or other part of a paper) reporting data for that question; for example, potassium might be represented by its chemical symbol K. Second, the text-searching retrieval procedure might only look for certain kinds of passages such

as figures and sentences, but a paper might report data only in a more complex kind of passage. An example is a combination of a text sentence reading "Table I gives surface tension values for (a) potassium, (b) sodium,. . .", and Table I with columns of values labeled only (a), (b),. . . . Various kinds of text-searching procedures are intended to deal with such problems.

This paper reports an experimental study of the effectiveness for data retrieval of a variety of text-searching procedures.

II. GENERAL NATURE OF THE EXPERIMENT

A random hundred data entries from the Thermophysical Properties Research Literature Retrieval Guide, Supplement I[1] (hereafter TP) were used in the experiment. Fifty were studied in the "development phase" to develop retrieval procedures, which were then tested on the other 50 in the "test phase".

A TP data entry consists of a property name, a substance name, and a reference to a paper reporting a value or values for that property of that substance. For instance, the example given above was obtained from the TP data entry

surface tension	potassium	50587 [no. of a reference in TP bibliography]

From each data entry, a data-retrieval question was derived, e.g., "What are values for the surface tension of potassium?" All questions were of the form: What are values for the

_____[property] of _____[substance]?

Given a data-retrieval question so derived and a passage retrieval procedure P being studied or tested at that time, the text of the paper cited in the TP data entry was studied to determine whether data for the question could be retrieved from that paper by procedure P. This determination was done by "manual" simulation of computer processing, because of the present unavailability of appropriate computer-readable text.

Independently of the human simulation of computer processing, it was also necessary in the experiment to determine where in a paper data were reported in a way a human subject specialist reader would recognize. This work was greatly aided by the editor of TP, Y.S. Touloukian, and by Peter Hilton (Mechanics and Mechanical Engineering, Lehigh University) and Ned Heindel (Chemistry, Lehigh University).

All text searching procedures studied in the experiment were "person-computer", where the human participation consisted of selection of "search words" for input to the computer text searching. Given the input of two lists of search words, one for the question-property and one for the question-substance, the computer then searched for passages within papers which satisfied both search word lists. The kinds of passages searched for included figures, text sentences, and some extensions of these which will be described later. A passage satisfying both search word lists was output with underlining added to those passage words matching search words.[15]

It should be noted that human selection of search words for input to computer text searching is used in operational text searching retrieval services in law.[5] It is also used in the many scientific retrieval services which now text search titles and abstracts, e.g., in TOXLINE, COMPENDEX, and other bibliographic databases containing computer-readable abstracts.

III. DEVELOPMENT PHASE

A. Data Retrieval from Figures. Consider first computer output of each "figure" which satisfies both input search word lists ("figure co-occurrence").

A "figure" here means the alphanumerics in the caption and body of a table, graph, photograph, etc. Alphanumerics include words, numbers, Greek letters, chemical formulas, etc. Graphics, such as curves, diagrams, and photographs are excluded (except for any alphanumeric material on or in them, such as a label on a curve). Thus all results for figure searching reported here require only searching of figure alphanumerics. Of course, graphic material can also be stored digitally or in coordinated microform and output as part of the output of total figures as they appear in the original papers.

Question-Word Matching. The simplest kind of property and substance search word lists consist of simply the names of the property and substance used in the question. For the question, "What are values for the surface tension of potassium?", the two search word lists input for computer searching would then be simply:

surface tension potassium

Searching using just such search words will be called "question-word matching".

Question-word matching in searches using only figure co-occurrence retrieved data (of the kind asked for) from 25% of the papers.

Synonym, Etc., Matching. Question words were looked up in a number of scientific dictionaries[16-21] to find synonyms, including symbols such as K for potassium and C_p for specific heat. The dictionaries also led to some terms specific to question words which were useful for retrieval, e.g., "plated" for the question word "coating" (in the question substance name, "copper coating"). Further, some of the search "words" led to by dictionaries were chemical formulas, e.g., "KBO_2". An unusually complex case was the question substance name, "monodeuterated methane", which required several dictionary steps before leading to a chemical formula, "CH_3D". Dictionary-found search words retrieved data from figures for another 20% of the papers.

Some TP substance names were multi-word, e.g., "iron alloy", and some papers reported data for those substances by other multiword names, e.g., "Fe(Ni) alloys", which separate the words in the TP substance name. In some other cases the words in a TP substance name, e.g., "stainless steel 440C", appeared in a different order in a data paper, e.g., "440C stainless steel". Such cases could be retrieved by requiring that the matches to the component words of a question substance name occur in the same "figure expression". Here "figure expression" means the caption (or a sentence of a multisentence caption), a column or row heading, a table entry, or a label on a graph axis or curve. (This list is subject to expansion.) This procedure retrieved data in another 15% of the cases.

Two of the successful retrievals described above also involved a rearrangement of the words in a multiword TP property name. For example, "thermal linear expansion coefficient" (TP) was represented in a figure by "coefficient of linear thermal expansion".

B. Retrieval of Data-Reporting and Data-Indicative Sentences. For the question, "What are accommodation coefficient values for boron iodide on tungsten?", data were reported by the following sentence:

> Evidence for thermal nonaccommodation of BI_3 on Pyrex is presented, however the thermal accommodation coefficient of BI_3 on tungsten was calculated to be essentially unity from critical supersaturation temperature dependence data.

This sentence was retrieved from a paper's summary with the aid of the search word "BI_3". The paper reported other data

in its figures but not accommodation coefficient data.

A sentence in a summary might not report data but be "data-indicative"; i.e., it can be inferred from the sentence that wanted data are reported in the paper elsewhere. An example, for the question, "What are viscosity values for Portland cement?", is the sentence:

> By using a device fitted to the ORGRES apparatus and an improvement in the research method we obtained figures for the viscosity in the pyroplastic state (apparent viscosity) at up to temperatures of 1500–1600 °C for a number of Portland cement raw mixtures distinguished by their lime saturation coefficients.

(This paper's figures reported viscosity data, but the figures could not be retrieved because only reading them in conjunction with the text would reveal that the data were for Portland cement.) Retrieving a data-indicative sentence from a summary enables the questioner to reliably identify the paper as one containing data, but he still must examine the paper to find the data.

A sentence in the main text of a paper might also be data reporting or data indicative. A data-indicative sentence in main text can sometimes lead quickly to the data themselves, for example, if it cites a figure reporting the data. This can be useful if the figure was not directly retrieved for some reason.

Retrieval of sentences from summaries added to the figure retrieval results described in section III.A another 5%. However, retrieving data-indicative sentences only increased retrieval of data papers, not direct retrieval of data.

Retrieval of sentences from main text added a further 5% to retrieval. These sentences were only data indicative, but one cited a figure which itself reported the data (but had not been retrieved because of an inadequate search word list).

A title may also be data indicative for a question, e.g., "Temperature Dependence of Surface Tension for Polytetrafluoroethylene (Supercooled Liquid) Estimated from Contact Angles". Ten percent of the development phase papers had data-indicative titles, but in all those cases retrieval of figures from the papers provided data directly.

C. Retrieved and Unretrieved Data in the Same Paper. In some cases data were retrieved from a figure for a question, but other data for the question reported in the same paper were not retrieved. As an example, for the question, "What are specific heat values for iron–aluminum alloys?", Figure 1 was correctly retrieved from a paper because it was captioned, "Fig. 1. Temperature dependence of the specific heat C_p of Fe-Al alloys annealed for 6 hr. at 300 °C"; but Figure 2, containing other data for the question, was not retrieved because it was captioned, "The same as Fig. 1, for alloys quenched from 800 °C." Three of the 50 development phase papers were of this kind.

Retrieval of data from such a paper constitutes successful retrieval of a *data paper* but only partially successful retrieval of data. A second of the three papers of this kind also involved a figure caption beginning, "Fig. 6. The same as Figure 5 . . .". In the third case a figure was not retrieved because of an inadequate search word list.

On the other hand, all three papers contained data-indicative summary sentences or titles retrieved by the search words used, which would reduce the chance of data in the unretrieved figures being missed by the questioner. In addition, computer searching could easily output (e.g.) any figure in a paper whose caption contained the expression "Fig. 1" once the computer search had retrieved the expression, "Fig. 1" from that paper.

D. Retrieval Failures. In 15% of the development phase papers a figure reported data in a way clear to a subject specialist reader, but the figure was not retrieved because it did not match a search word list. Some examples are given in Table II.

Table II

Question expression	Figure expression
Surface tension	σ, dynes/cm
Thermal linear expansion coefficient	[Caption] Dilatometric curves. . ., [Graph axes] t °C, Δl
Potassium nitrate + sodium nitrate mixture	$(Na\text{-}K)NO_3$ mixtures
Chlorinated paraffin wax	Cereclor 42
Interface, water–gas	Water and water–ethanol mixtures in contact with air saturated with ethanol

Table III

Question expression	Figure expression	Needed context
Thermal conductivity	λ	[In section headed "Nomenclature" at beginning of paper], λ, total thermal conductivity
Thermal linear expansion coefficient	$\alpha \cdot 10^7 deg^{-1}$	[In first sentence of main text] linear thermal coefficient α. . .
Deuterium faujasite	DY	[In a subheading] Deuterium Faujasite (DY)
Polytetrafluoroethylene, asbestos-reinforced	L	[In an earlier figure, captioned "Test Materials"] L Asbestos-reinforced Teflon

In 25% of the development phase papers a figure contained data, but some words or symbols used could not be definitely interpreted, even by a subject specialist reader, without reading another part of the paper as well. Some examples are given in Table III. About a third of these cases also involved retrieval failures of the kind described in the preceding paragraph.

Another 5% of the figure retrieval failures were caused because the question involved a complex substance which could only be identified by combining either a figure and a text sentence or two text sentences. The substances were: aqueous solution of benzyldodecyldimethylammonium chloride + titanium oxide, and paint (ZnO + potassium silicate).

A further 5% of the figure retrieval failures were caused because the papers reported data for the questions involved but did not do so in figures (see section III.B).

E. False Retrieval. False retrieval could not be tested for directly since computer-readable figures, not to mention full text, were not available for the experiment. However, it was investigated indirectly in the ways described below.

False retrieval can be caused by an ambiguous search word or by a "false relation", e.g., a property and a substance named in a figure but the figure giving no data for that property of the substance. These two kinds of false retrieval were investigated separately.

Most property search words involved in successful retrieval were either the question words or reasonably unambiguous synonyms, such as "coefficient of linear expansion" for "thermal linear expansion coefficient" and "C_p" for "specific heat". However one search word, "transmission", used as a search word for "transmittance", was seriously ambiguous; it frequently occurs in scientific literature with other meanings than transmittance and so might cause significant amounts of false retrieval. It was involved in 5% of the successful retrievals described above.

None of the substance search words used were ambiguous in the way that "transmission" was, but some of them involved risk of another kind of ambiguity. A simple and common substance name like "air" (which was the TP substance name in one question) might also occur in complexes of (e.g.) the form "air cooled X" where property data are given for sub-

stance X but not for air. Such a risk can be dealt with by pre-search use of a general concordance or post-search use of a special concordance. In the former, in a concordance to the whole collection to be searched the entries for "air" (e.g.) are scanned and the search word "air" is accompanied by instructions to the computer to ignore "air" occurrences in specified unwanted contexts such as "air cooled". In the latter, computer output of figures, sentences, etc., is accompanied by a concordance to that output, which is used in the same way. The first approach involves more human scanning but might be computationally easier in some circumstances.

One risk of false retrieval caused by "false relation" is involved in multiword question substance names, e.g., "iron-aluminum alloys". A figure caption, e.g., might contain "iron", "aluminum", and "alloy", or their synonyms, but not be about any iron–aluminum alloy. A possibly satisfactory solution to this is the familiar text-searching technique of word proximity, e.g., requiring that the search word matches to "iron", "aluminum", and "alloy" occur within a certain number of words (or characters—to allow for nonverbal material like chemical formulas) of each other; similarly for multiword property names (e.g., "thermal linear expansion coefficient"), being satisfied by slightly discontinuous rearrangements, e.g., "coefficient of linear thermal expansion".

A different kind of risk of false retrieval by "false relation" involves a property P and a substance S each being correctly named in a figure but the figure giving no P data for S. The 300 figures in the 50 development phase papers were examined for such cases. These figures collectively contained about 5000 different property–substance pairs (i.e., a property and a substance named in the same figure). In only five cases were a property and a substance named in a figure but no data for that property of the substance given in the figure. Specifically, one figure gave emissivity curves for two substances and a transmittance curve for a third substance. Another figure did not give specific heat values for two of the substances it listed. These quantitative results strongly suggest that "false relation" of property and substance words in figures will not cause significant false retrieval.

The titles and summary sentences of the development phase papers were similarly examined, and no cases found of a property and a substance named in a title or sentence and yet the latter not being data-indicative or data-reporting for that property of the substance.

A random hundred main text sentences which each contained a property name and a substance name were similarly examined, and only five were found to be not data indicative or data reporting for that property of the substance.

IV. TEST PHASE

A. Retrieval Procedures to be Tested. At the beginning of the test phase all retrieval procedures suggested by the development phase work were explicitly described, as follows:

Search Words. "Word" here means a word, a right-truncated word stem (e.g., "conductivit:"), a phrase, or a nonverbal symbol (e.g., "C_p", "KBO_2").

There will be one property search word list for a question. It will contain the property name in the question and any synonyms led to by the scientific dictionaries used.[16-21] Property search words will be modified to allow for plurals, e.g., "thermal conductivit:", "specific heat, specific heats".

There may be more than one substance search word list for a question, and usually will be for a multiword substance name in the question, e.g., "iron–aluminum alloys". Each substance search word list will consist of a question substance word and synonyms and specifics to it led to by the scientific dictionaries. Substance search words will also be modified to allow for plurals and singulars, e.g., "alloy, alloys".

Computer Search: Co-occurrence. The computer will output each figure which satisfies all the search word lists, underlining all matching words in the figures.

A property search word which is a phrase (e.g., "thermal linear expansion coefficient") will be matched if its component words occur "close" to each other in any order. Explicitly, "close" for property word matches in figures means at most two intervening words. (Thus the matching words must also be within the same caption, column heading, graph label, or other verbal expression in the figure.)

If there is more than one substance word list, a set of words satisfying all substance lists must be "close" in the figure. Explicitly, "close" for substance word matches in figures means at most 16 intervening characters. (Thus the matching words must also be within the same caption, column heading, graph label, or other expression in the figure.)

The computer will also output each title, summary sentence, and main text sentence which satisfies all the search word lists, underlining all matching words. The "closeness" requirements for property words and for substance words are the same as those for figures.

Computer Search: Abbreviation Translations. The procedures described here are suggested by the development phase results described above in section III.D (second paragraph) and related observations.

Where X is a sequence of at most three characters (none blank), at least one of which is alphabetic (English, Greek, or other alphabet), and S in a search word, then add X to the same search word list as S for the paper being processed if one of the following conditions is satisfied:

(1) In a figure, X is at the left end of a row or the top of a column, with blanks on each side of it, and S is in the corresponding row or column.

(2) In a section headed "Nomenclature" or "Appendix", X is followed by one or more blanks, then an equals sign or comma or "is" or "are", then further blanks, and then S within the same line. Alternatively, X may be to the right of the line containing S and be preceded by = or, or "is" or "are", or one or more blanks.

(3) Anywhere in the title, summary, or main text is one of the following:

SX (separated by one blank)
S,X
S(X)
S is given by X = (may be on different lines)
X(S)
X, . . . S—at most ten characters apart
X, the . . . S—at most ten characters apart
X is . . . S—at most ten characters apart

Whenever S and X are found in such a combination, and that X causes retrieval of a passage P (figure or sentence) which would not otherwise be retrieved, then P is output accompanied by the S–X passage.

Computer Search: Figure and Figure-Citing Sentence Combinations. Add any sentence citing a figure to that figure, and treat them as a unit for search word matching purposes.

B. Retrieval Results. As noted earlier, 50 data retrieval questions and a corresponding 50 papers were used in the test phase.

Figure data were retrieved from 50% of those papers. Retrieval of data-reporting main text sentences added another 5%. Retrieval of combinations of figures and sentences citing them added a further 5%.

Concerning kinds of search words involved in retrieval, figure search using only question words retrieved data from 25% of the papers. Adding search words found by scientific dictionaries added another 33% to retrieval. Using "translated

Table IV

Question expression	Figure expression
Thermal linear expansion coefficient	Lattice parameters. . .at various temperatures [including crystal lengths]
Thermal conductivity	Thermal resistance
Sodium nitrate + silver nitrate	(Ag–Na)NO_3
Sodium iodide, CO adsorbate	CO adsorbed on NaI
Tributylamine	[Dictionary led to tri-*n*-butylamine], Amines [in figure caption], Tri-*n*-butyl [at left of a row in figure]

abbreviations" (see above, section IV.A, Computer Search: Abbreviation Translation) as well added another 2%.

In some cases data were retrieved from a figure for a question, but other data for the question reported in the same paper were not retrieved. This happened for 10% of all test phase papers. These were counted as successful retrievals in the preceding two paragraphs. Thus all data were retrieved from 50% of the papers, and another 10% of the papers were retrieved as data papers but not all data were located within them. Three of the four papers of this latter sort contained data-indicative titles or summary sentences but only one was retrieved.

C. Retrieval Failures. In 25% of the test phase papers a figure reported data in a way clear to a subject specialist reader, but the figure was not retrieved because it did not match a search word list. Some examples are given in Table IV.

In 20% of the test phase papers a figure contained data but some words or symbols used could not be definitely interpreted, even by a subject specialist reader, without reading another part of the paper as well. See Table V for some examples. About a third of these cases also involved retrieval failures caused by absence of search word match.

One other retrieval failure occurred because the paper reported the answer data in two scattered main text sentences.

D. False Retrieval. As noted earlier, false retrieval could not be tested for directly because of the absence of suitable computer-readable material. It was investigated indirectly, as in the development phase, in the ways described below.[22]

One search word used in the test phase was seriously ambiguous, i.e., "K" used as a search word for "potassium", since K also has other scientific uses. It was involved in retrieval of some of the data from one paper.

Some substance words, e.g., "air", could be expected to occur in many contexts, e.g., "air cooled X", in which no property was given for air. Use of a concordance to prevent such false retrievals would be advisable (see above, section III.E, fourth paragraph).

The "closeness" requirements included in the retrieval procedures (see above, section IV.A, Computer Search: Co-occurrence) should minimize false retrieval caused by "false relation" between components of a multiword substance or property name, but this needs to be investigated.

Concerning possible "false relation" false retrieval caused by a property P and a substance S each being correctly named in a figure but the figure giving no P data for S, this was investigated as in the development phase. Only 14 P–S pairs occurred in figures without P data for that S being given in the figure (sentences citing the figures were also included). Concerning similar P–S false relations in titles and summary sentences, none were found. In a random hundred main text sentences each containing a P name and a S name, only three such false relations were found.

It should be noted that a falsely retrieved figure or sentence can be screened out far more quickly than a falsely retrieved whole paper, which is the unit of false retrieval produced by present scientific retrieval services.

E. Comment. Further experimentation will be appropriate when computer-readable full text and/or words from figures are available. This will permit direct testing for false retrieval and larger scale testing for recall.

In the test phase of the experiment reported here, figure search retrieved 50% of the data papers and main text searching added another 10%. This contrasts with the retrieval results in a related study, in which answer passages were retrieved from papers for questions about cardiovascular drug effects.[12] In that study's test phase, involving about 100 answer papers, a search of figures and summaries retrieved 80% of the answer papers, and main text searching added another 10%. The difference in results appears to be caused primarily by the far greater extent to which nonverbal symbols were used in the physical science and technology papers involved in the experiment reported here, compared to the biomedical papers involved in the other study. This is illustrated by the examples of retrieval failures given above (sections III.D and IV.C). Further study of text searching retrieval for the physical sciences should include concentration on developing retrieval procedures to handle such cases.

Concerning the costs of data retrieval by text searching, there is some uncertainty about when *on-line* retrieval by full-text searching will become operationally feasible. But there are two alternatives which are now economically feasible. One consists of adding to the present on-line text searching of titles and abstracts the text searching of words from figures as well (see section I.B, first paragraph). The other alternative consists of off-line text searching of tape by minicomputers, which is relatively slow (because of the sequential tape access) but inexpensive. For example, a $50,000 minicomputer configuration (including one tape drive, a five megabyte disk, 128 kilobyte core, and 180 character/second printer) could do the searching described in this paper at the rate of 6×10^4 characters per second (tape-reading speed). Several hundred search words, e.g., for a number of different questions, could be searched at once without decreasing this rate.[23] The papers in the experiment reported here averaged about 2000 characters each of title, abstract, and figure words (excluding stop-words[24]); thus about 10^5 papers per hour could be abstract-and-figure searched. The papers' main texts averaged 8000 characters (excluding stop-words); thus full text could be searched at about 2.5×10^4 papers per hour.

ACKNOWLEDGMENT

This work was supported by National Science Foundation Grant No. SIS75-09282.

Table V

Question expression	Figure expression	Needed context
Combustion products of methane + nitrogen + oxygen	Combustion products	[In summary] The thermodynamic properties of combustion products of methane-oxygen-nitrogen mixtures. . .
Air ← octane [diffusion coefficient wanted]	Octane [and "Diffusion Coefficients"]	[Title] Diffusion Coefficients of Some Organic and Other Vapors in Air
Reflectance	R	[In section headed Nomenclature] R and T are the reflectivity and transmittivity. . .
Interface, heptane–water	γ_{ab}	[In summary] γ_{ab}, the interfacial tension againse water, . . .

REFERENCES AND NOTES

(1) "Thermophysical Properties Research Literature Retrieval Guide", Supplement I, Y. S. Touloukian, J. K. Gerritsen, and W. H. Shafer, Ed., IFI/Plenum, New York, N.Y., 1973.
(2) National Science Foundation, "Program Solicitation—Improved Dissemination and More Productive Use of Scientific and Technical Information" (NSF 75-23), Washington, D.C., 1975, pp 2–3.
(3) R. Bemer and A. Shriver, "Integrating Computer Text Processing with Photocomposition", *IEEE Trans. Prof. Commun.*, PC-16, 92-7 (1973) (Special Issue: Record of the Conference on the Future of Scientific and Technical Journals).
(4) J. Phillips, "Integrating Computer Processing with Photocomposition", *IEEE Trans. Prof. Commun.*, PC-16, 165 (1973) (Special Issue: Record of the Conference on the Future of Scientific and Technical Journals).
(5) R. Lerner, "The Use of the Computer in Converting Primary Information", *J. Chem. Doc.*, **14**, 112–4 (1974).
(6) Aspen Systems Corporation, Editorial Processing Centers: Feasibility and Promise (Report on NSF Contract C769), Aspen, Germantown, Md., 1975.
(7) S. Rhodes and H. Banford, "Editorial Processing Centers: a Progress Report", *Am. Sociologist*, **11**, 153–9 (1976).
(8) Chemical Abstracts Service, Features of the American Chemical Society's Primary Journal Composition System, American Chemical Society, Columbus, Ohio, 1976.
(9) R. May, Ed., "Automated Law Research", American Bar Association, Chicago, Ill., 1973, pp 40, 62, 104.
(10) "United States Air Force Judge Advocate General Law Review" (Special Issue on LITE), 1966.
(11) "United States Air Force Judge Advocate General Law Review" (Special Issue on LITE), 1972.
(12) J. O'Connor, "Retrieval of Answer-Sentences and Answer-Figures from Papers by Text Searching", *Inf. Process Manage.*, **11**, 155–64 (1975).
(13) F. Roehlich et al., "Surface Tension of Four Alkali Metals to 1000 °C", *J. Chem. Eng. Data*, **13**, 518–21 (1968).
(14) For compactness, TP also uses numerical codes for substance names. Thus TP entries *appear* somewhat different from the entry given here.
(15) Since the computer procedures studied were "manually" simulated, as already noted, the last three sentences of this paragraph are rather "mechanomorphic" in style, but that style permits quick understanding of the computer procedures involved and will be used throughout.
(16) D. Lapedes, Ed., "McGraw-Hill Dictionary of Scientific and Technical Terms", McGraw-Hill, New York, N.Y., 1974.
(17) H. Bennett, Ed., "Concise Chemical and Technical Dictionary", Chemical Publishing Co., New York, N.Y., 1974.
(18) G. Hawley, Ed., "Condensed Chemical Dictionary", Van Nostrand-Reinhold, New York, N.Y., 1971.
(19) H. Gray and A. Isaacs, Ed., "A New Dictionary of Physics", Longmans, London, 1975.
(20) J. Grant, Ed., "Hackh's Chemical Dictionary", McGraw-Hill, New York, N.Y., 1969.
(21) A. Merriam, Ed., "A Concise Encyclopedia of Metallurgy", American Elsevier, New York, N.Y., 1965.
(22) If this section seems too concise, re-reading section III.E will provide sufficient context.
(23) I thank Preston Marshall, Center for Information Science, Lehigh University, for the specific computer cost and performance information given here.
(24) Stop-words include both function words such as "of" and "the" and general words such as "described" and "details".

An Algorithmic Computer Graphics Program for Generating Chemical Structure Diagrams

PAUL G. DITTMAR,* JOSEPH MOCKUS, and KATHYRN M. COUVREUR

Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43210

Received March 2, 1977

An algorithmic computer graphics program has been developed at Chemical Abstracts Service (CAS) which has the capability of transforming connection tables into chemical structure diagrams of a quality suitable to meet editorial proofing requirements at CAS. The program, which runs on an IBM 370/168 computer, begins with a connection table record and a reference file of basic chemical ring shapes. Then, by using a set of defined rules, the order of processing and the preferred placement of the structural pieces are systematically determined and the diagram is constructed. An automatic adjustment capability is available to handle situations where one structural fragment overlaps another within a given diagram.

INTRODUCTION

The CAS Chemical Registry System grew out of CAS staff research in the early 1960's in which an algorithm developed by DuPont was perfected for generating a unique and unambiguous computer-language description of the two-dimensional structure and stereochemical details of a chemical substance.[1] The Registry System uses this algorithm as the foundation for identifying chemical substances on the basis of their composition and structure and for linking their structural descriptions to the various names by which they are identified in the literature through unique computer-checkable Registry Numbers. Registry III went into operation early in 1974 as an extension to this system.[2] This extension expanded the system's chemical substance naming capability and made possible, for the first time, the re-creation algorithmically of structure diagrams from computer structure records. In Registry III the parent ring systems of a structure are identified separately by a modified form of computer-readable structure notation. These parent ring systems can thus be retrieved from a computer file to provide a starting point for recreating the structure.

The CAS Chemical Registry System utilizes a manual file of hand-drawn structural diagrams to support processing operations. [Note: Structures used throughout this paper are hand-drawn for illustrative purposes. Examples of actual chemical structures generated by the ASD program are given at the end of this paper (see Figure 19).] This file currently contains over 3.7 million entries. The ability to produce structural diagrams from the computer-readable Registry records minimizes the need to reference and maintain this extensive manual file.

When the system is fully operational, the ability to produce structure diagrams by computer will depend upon two complementary capabilities required to handle the full range of structures. Most structures will be automatically "constructed" by an Algorithmic Structure Display (ASD) program starting with information contained in the Unique Chemical Registry Records (UCRR's) of the CAS Registry File. For certain special classes of structures, this capability will be supplemented by a system which will cause structures to be "played back" from a file of display instructions recorded from input via an interactive graphics terminal.[3]

The purpose of this paper is to further describe the ASD program.

BASIC FEATURES OF ASD

The ASD program was designed with the intention of supporting the substance indexing operations within the CAS production system, primarily that of naming substances new to the CAS Chemical Registry System. A second objective