Use of Word Fragments in Computer-Based Retrieval Systems*

D. S. COLOMBO and J. E. RUSH
Department of Computer and Information Science,
The Ohio State University, Columbus, Ohio 43210

Received October 28, 1968

A number of computer-based text-search systems currently in operation permit the use of word fragments as search terms. The full potential of this feature is not realized because of the uncertainties associated with the use of word fragments. A pair of dictionaries is described which makes possible the selection of appropriate fragments both for "generic" retrieval and for retrieval of specific words. The general nature of the computer program used to produce the dictionaries is described.

A number of computer-based text-search systems are currently in operation which, among their facilities, allow the searcher the use of word fragments as search terms.[1] Word fragments, in contrast to full words, provide a means of retrieving documents on the basis of concepts which may have been coded in a variety of forms. For example, consider the simple case of singular and plural forms of a word. Rarely does the plural form of a word represent a concept sufficiently different from the singular form to make their collective retrieval undesirable. Therefore, retrieval systems using data stored in a form such that a concept may be represented by more than one code require a means of "generic" searching such as that afforded by word fragments. Three types of word fragment may be distinguished: prefixes, suffixes, and infixes. Some retrieval systems allow the use of all three types,[1a] while others are more limited.[1b]

Not only do word fragments provide a means of simulating generic searching, they should produce savings in search time 1) through a reduction in the number of search terms used and 2) by virtue of their smaller size relative to the words from which they were derived. Thus the motivation for using word fragments is a very pragmatic one. The basic difficulty encountered in the use of word fragments as search terms is that of deciding the right fragment to use. For example, if one were to choose the prefix ION* (an asterisk, *, is used throughout this paper to indicate whether a fragment is a prefix, XXX*, suffix, *XXX, or infix, *XXX*) to retrieve both ION and IONS (and perhaps IONIC), he would quickly discover that IONIZATION, IONOSPHERE, etc., are also retrieved. Similar difficulties will be encountered with many other prefixes and suffixes. But it is far easier, in most instances, to predict the result which would obtain from the use of a particular prefix or suffix than to predict the result which would obtain from the use of a particular infix. Suppose one wished to retrieve documents having to do with azo dyes, and therefore used *AZO* as a search term. This infix would certainly retrieve such dyes as BENZENEAZONAPHTHALENE, 1-[(*p*-NITROPHENYL)AZO]-2-NAPHTHOL and *p*-[(2-HYDROXY-1-NAPHTHYL)AZO]BENZENESULFONIC ACID, but it would also retrieve a whole set of unwanted material such as IMIDAZOLES, ACETAZOL, OXAZOLE, OSAZONE, BETAZOLE, BORAZOLE, CARBAZOLE, and ACINITRAZOLE. Clearly a poor choice of search term was made. This is a simple example; how many words can one recall which contain the infix *UTTI* (e.g., ABUTTING, CUTTINGS, GUTTIFERAE) or *TERO* (e.g., ASTEROID, AZASTEROIDES, BUTTER OIL, HETEROLOGOUS, PTEROPSIDA, STEROL).

Search time is, in most systems, dependent on the number and length of the search terms processed. Word fragments, which will often serve as well as several complete words as search terms, should therefore reduce the search time while maintaining or increasing the level of recall.

In order to make effective use of word fragments as search terms, either for generic searching or simply for decreasing the search time, one must be able to predict fairly accurately the consequences of choosing a particular word fragment. To provide the information necessary to make such predictions, we have developed a dictionary of word fragments which enables the user to determine the probable result of using a particular word fragment as a search term.[2] The dictionary consists of an alphabetical listing of word fragments, together with the set of words in which each occurs (parent words), such sets being listed, alphabetically, subordinate to the appropriate fragment (see Figure 1).

Word fragments were generated by means of a computer program which derives all possible fragments of four or more characters from the parent word. Once the fragments are generated and sorted, those which are contained in 10 or more parent words are discarded as being of no utility as search terms. (This is a somewhat arbitrary decision which can be varied according to circumstances.) If each of a set of fragments of varying length proves to be unambiguous for a particular parent word, as for the example of Figure 2, all but the shortest fragment are discarded, since each fragment would retrieve only the one word and since the shortest fragment should require

* Presented in part before the Division of Chemical Literature, 156th Meeting, ACS, Atlantic City, N. J., September 1968.

	ACETAL
ABIN	ACETAL
CANNABINOL	ACETALDEHYDE
CANNABINOLS	ACETALS
ABRA	ACETO
ABRASION	ACETO
GLABRATUS	ACETOBACTER
	ACETOBUTYLICUM
ACACE	ACETONE
BOMBACACEAE	ACETONEMIC
MAYACACEAE	ACETONIDE
OLACACEAE	ACETONO
PORTULACACEAE	ACETONYLIDENE
STYRACACEAE	ACETOSORBATES
ACER	ACIA
ACER	ACACIA
ACERACEAE	ARMORACIA
ACEROSA	GAYLUSSACIA
	SPINACIA
ACETAB	
ACETABULARIA	ACTERI
	BACTERIA
ACETALD	BACTERIAL
ACETALDEHYDE	BACTERIOCITE
	BACTERIOPHAGE

Figure 1. Sample dictionary of word fragments

the least search time. Finally, if there result several fragments of varying length, each of which is derived from the same set of words, as illustrated in Figure 3, only the shortest fragment, with its set of parent words, is retained, since 1) any of the several fragments would retrieve all of the parent words of the set; 2) only a word fragment longer than the longest of the several fragments under consideration would differentiate, at least partially, between the members of the set of parent words; and 3) the shortest fragment provides the greatest search efficiency for the given set of parent words. These three selection rules permit a reduction of about 75% in the size of the dictionary produced.

The choice of four characters as the minimum length fragment perhaps needs some amplification. Initial experiments with shorter fragments showed that virtually all such fragments would be deleted by the first selection rule, whereas the deletion rate was markedly reduced when four characters were used as the minimum length fragment, and generation time was reduced by about 10%. Furthermore, retrieval experience has shown that use of fragments of fewer than four characters results in a high yield of spurious documents. Nevertheless, the choice of minimum fragment length may be varied as the application dictates.

Although we have been primarily interested in the techniques used to produce the fragment dictionary, some data relating to the parent words and fragments dealt with in this study are of interest. A sample of 2845 words was used to generate the fragment dictionary which we described. These parent words ranged from four to 15 characters in length; Figure 4 shows their size distribution.

The number of fragments of length i derivable from a word of length n is

$$F_i = (n - i) + 1 \qquad (1)$$

ACETOS	ACETOSORBATES
ACETOSO	ACETOSORBATES
ACETOSOR	ACETOSORBATES
ACETOSORB	ACETOSORBATES
ACETOSORBA	ACETOSORBATES
ACETOSORBAT	ACETOSORBATES
ACETOSORBATE	ACETOSORBATES
ACETOSORBATES	ACETOSORBATES

Figure 2. For fragment-parent word combinations of the type illustrated, only the combination having the shortest fragment is retained as a dictionary entry.

ABSO	ABSORPTION
	ABSORPTIONAL
	ABSORPTIONS
	ABSORPTIVE
	ABSORPTIVITIES
ABSOR	ABSORPTION
	ABSORPTIONAL
	ABSORPTIONS
	ABSORPTIVE
	ABSORPTIVITIES
ABSORP	ABSORPTION
	ABSORPTIONAL
	ABSORPTIONS
	ABSORPTIVE
	ABSORPTIVITIES
ABSORPT	ABSORPTION
	ABSORPTIONAL
	ABSORPTIONS
	ABSORPTIVE
	ABSORPTIVITIES
ABSORPTI	ABSORPTION
	ABSORPTIONAL
	ABSORPTIONS
	ABSORPTIVE
	ABSORPTIVITIES

Figure 3. When a set of fragments is generated from the same set of parent words, only the shortest fragment is retained as a dictionary entry

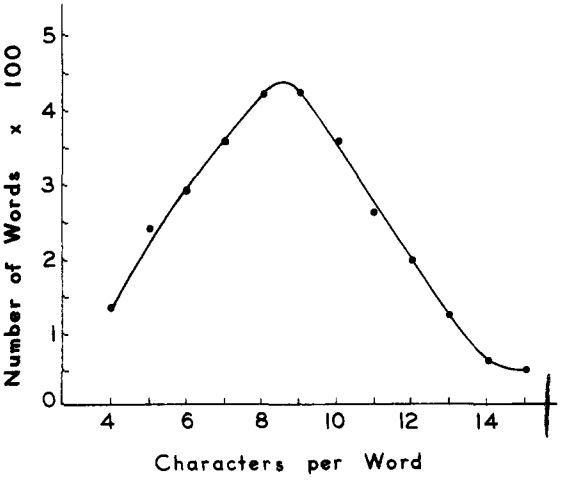

Figure 4. Distribution of word sizes in the data base used in this study (W = 2845)

and the total number of such fragments is

$$F_{i_T} = [(n - i) + 1]W_n \quad (2)$$

where W_n is the number of words in the population of length n. Figure 5 shows the values assumed by F_{i_T} for the sample of words used in this study.

Indiscriminant generation of fragments of four or more characters ($i \geq 4$) would yield 64,103 fragments, whereas application of the selection rules described earlier reduced this number to 15,590. The number of fragments, F, of any minimum length, i, derivable from a word of length n, is given by

$$F_n = \frac{n^2}{2} - \left(\frac{2i-3}{2}\right) n + C \quad (3)$$

where $C = 0$, or C is a positive integer given by $C = 2i - 5$. It follows that the total number of fragments, F_{n_T}, derivable from W words of length n is

$$F_{n_T} = F_n(W_n) \quad (4)$$

and that the total number of fragments derivable from any population of words is

$$F_T = \sum_{n=1}^{\infty} F_n(W_n) \quad (5)$$

or, from Equation 2,

$$F_T = \sum_{i=p}^{\infty} \sum_{n=q}^{\infty} F_{i_T} \quad p \leq q; p = 1,2,3\ldots \quad (6)$$

From Equation 2 we derived the theoretical distribution of fragment sizes shown in Figure 6. This distribution is contrasted with the actual distribution of fragments in the dictionary, also shown in Figure 6. The selection rules clearly bias fragment size in favor of four or five characters, although the two curves are quite similar.

During the development of the fragment dictionary, it became obvious that a considerable number of fragments occurred in single words—i.e., were unique for that word. We therefore prepared a second dictionary with the parent words listed alphabetically and with the fragments which uniquely represented them subordinate thereto. Such a dictionary is exemplified in Figure 7. Since a word may be represented unambiguously by more than one fragment, the shortest fragment is selected for the dictionary. If more than one fragment remains, that one is chosen which is highest in the alphabetical order. The distribution of fragments, by size, resulting from the application of these selection rules is shown in Figure 8.

Of the 2845 words in the input data, 2490 are represented in the unique fragment dictionary. The absence of the remaining 355 parent words is accounted for by the fact that they, as fragments ($i = n$), are each included in a larger word and are therefore represented by no unique fragments. The unique fragment dictionary makes it possible to select a fragment which will retrieve one and only one parent word and at the same time should increase search efficiency.

In order to determine whether the use of fragments as search terms would increase search efficiency, we obtained a set of search questions (from the Office of Computerized Information Services, College of Medicine, The Ohio State University) which had previously been encoded for search using *Chemical-Biological Activities (CBAC)* as the data base. We recoded the questions, using the shortest possible fragment and, where possible, the most generic fragment. We did not otherwise alter the coding. The

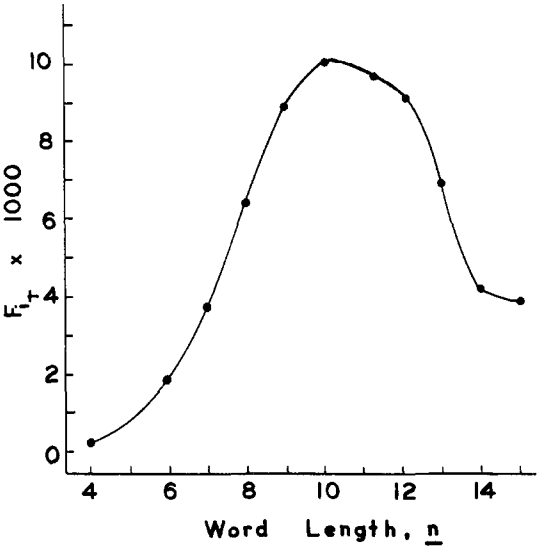

Figure 5. Plot of F_{i_T} over the range $4 \leq i \leq 15$, $4 \leq n \leq 15$ and for W_n as in Figure 4

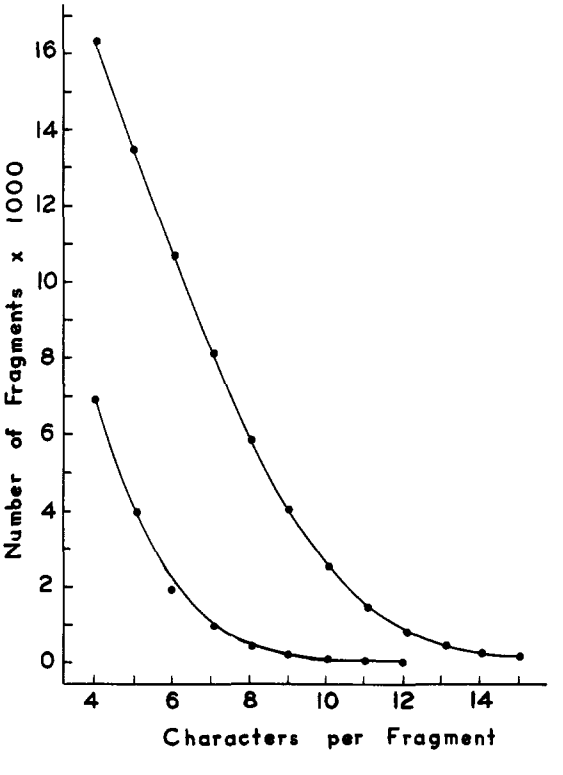

Figure 6. Theoretical (upper curve) and actual distribution of fragments by size

GLYCEROLS ROLS	GONYAULAX YAUL
GLYCEROLYSIS CEROLY	GOODENIACEAE OODE
GLYCINE LYCI	GOSSYPIUM YPIU
GLYOXIMATE XIMAT	GRACINIA GRAC
GLYOXIME XIME	GRAFTING RAFT
GOLDEN OLDE	GRAIN GRAI

Figure 7. Sample of the unique fragment dictionary

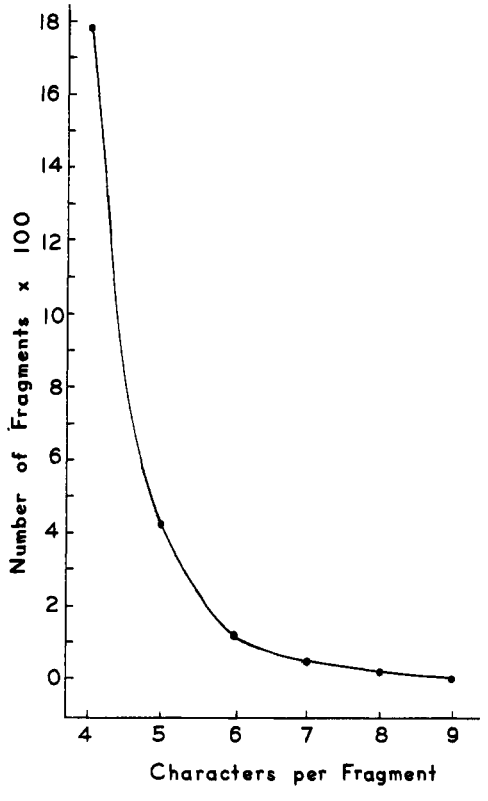

Figure 8. Distribution of fragment sizes in the unique fragment dictionary

questions, thus encoded, were searched using the same data base as was used in searching the questions as originally coded. The encoded questions are compared in Table I, while the search results, in terms of number of hits obtained, are given in Table II along with the respective search times. From these data it can be seen that use of fragments as search terms produced satisfactory results.

While the dictionaries described above can be effectively employed in manual coding of questions for searching

Table I. Comparison of Search Terms used in Conventional Coding and in Coding with Fragments

	Conventional Coding		Fragment Coding	
Q.N.	Terms	Characters	Terms	Characters
1	10	80	5	32
2	24	243	19	160
3	16	111	13	62
4	27	302	10	62
5	1	20	1	7
6	1	10	1	6
7	9	92	9	54
8	3	29	2	15
9	2	27	1	7
10	5	44	3	18
11	2	44	1	8
Totals	100	1002	65	431

Table II. Comparison of Search Results and Search Times for Conventional Coding *vs.* Fragment Coding

	Documents Retrieved	
Q.N.	Conventional coding	Fragment coding
1	4	4
2	5	1
3	4	4
4	2	1
5	0	0
6	2	2
7	0	0
8	4	5
9	2	2
10	0	0
11	0	0
Search time	9.25 min.	8.45 min.

textual data, we are at present developing procedures for using them, together with a thesaurus and word frequency data, in a completely automated process of question encoding, both for real-time and batch processing systems.

ACKNOWLEDGMENT

We wish to acknowledge support of this work by the National Science Foundation through grant GN-534. We also wish to express our appreciation to Mr. E. Nine, Systems Research Department, College of Medicine, and Miss L. K. Osborn, Office of Computerized Information Services, for their help in performing the *CBAC* searches.

LITERATURE CITED

(1) (a) For example, the Chemical Abstracts Service has available search programs which permit the use of prefixes, suffixes, and infixes; (b) DATRIX, the retrieval system of *Dissertation Abstracts*, permits the use of prefixes.

(2) During the course of this work, another novel word fragment list was described [A. K. Kent, *Svensk Kem. Tidskr.* **80** (2), (1968)].