The identification of variable-length, equifrequent character strings in a natural language data base

A. C. Clare, Elizabeth M. Cook and M. F. Lynch

Postgraduate School of Librarianship and Information Science, University of Sheffield

The words of natural language texts exhibit a Poisson (or Zipfian) rank-frequency relationship, i.e., a small number of common words accounts for a large proportion of word occurrences, while a large number of the words occur as singletons or only infrequently. Inverted-file retrieval systems using free text data bases commonly identify words as the keys or index terms about which the file is inverted, and through which access is provided. They therefore involve large and growing dictionaries and may entail inefficient utilisation of storage because of the distribution characteristics.

An alternative approach may be based on the analysis of text in terms of sets of variable-length character strings, the frequency distributions of which are much less disparate than those of words. This could lead to substantial reductions in dictionary size, and increased efficiency both in dictionary look-up times and storage ultilisation.

(Received October 1971)

1. Introduction

Computer searches of data bases of scientific and technical information are today almost commonplace. Their purpose is to provide users with bibliographic references to articles in the published literature, and they are available both as current-awareness or retrospective searches. A range of techniques is in use, including searches of the natural language texts of the titles and/or abstracts of the articles, although others employ assigned keywords or subject headings. In off-line use, the user submits a profile containing words or word-fragments, often in the form of a Boolean statement, which is designed to capture the greatest number of relevant references while excluding as much garbage as possible. In on-line situations, this profile may be re-negotiated during the search in the light of document descriptions retrieved.

The search systems used are implemented either as serial searches of magnetic tape files, or as searches of inverted files on rotating storage. The two stand in considerable contrast to one another; in the former, the profiles are stored in core and the texts of the documents are compared against the profile terms. The terms need not consist of whole words; both right-hand and left-hand truncation may be used, the first to capture morphological variants (e.g., singular and plural, or noun and verb forms), the second to allow for prefixes such as 'pseudo-' and 'poly-'. In fact, since the average length of the character string is an important factor in determining search time, the search term may be reduced to the shortest sequence of characters which uniquely identifies a known word in a particular collection (Colombo and Rush, 1969), while KLIC (Key Letter In Context) indexes are also used to determine the probable frequency of occurrence of specific character strings (Kent, 1968). In inverted file systems the word, however defined, is usually chosen as the key-forming unit or index term about which the file is inverted. While right-hand truncation is easily provided by serial examination of the dictionary, left-hand truncation, which is invaluable in disciplines such as chemistry, involves space- and time-consuming manipulation of an already large dictionary.

The disparate frequency distribution of words in written text was established by Zipf (1949), who showed that the rank of a word is related to its frequency by the expression:

$$f(r) = cr^{-a}$$

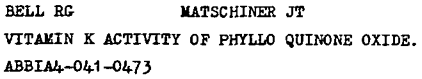
where $f(r)$ is the frequency, r the rank and a and c are constants, the values of which are approximately 1 and 0·1 respectively. The implications of this for inverted files of natural language texts are considerable. Firstly, the size of the dictionary required for key assignment is large, and grows steadily as new documents appear. Secondly, the allocation of storage space presents a substantial problem; even when the most frequent words (e.g. articles, prepositions, auxiliary verbs) are excluded, the remainder, depending on the file organisation used, may impose a large cost in terms of inefficient use of storage space. Higgins and Smith (1971) hàve suggested a technique involving exponential increase in storage allocation according as key lists overflow available fixed-length blocks. Furthermore, in some services, searches using very frequent words as profile terms, may incur special charges or may be prohibited entirely.

It seemed possible, in the light of concurrent studies on means of indexing machine-readable records of chemical structure information (Crowe, Lynch, and Town, 1970) that the Zipfian distribution of alphabetic and other characters in text might offer opportunities for solutions entailing the formation of keys for file-inversion from units other than words. In the case of chemical structures, for instance, the elements, bonds, and all other structural units exhibit extremely disparate distributions, and the solution there would seem to be to describe elements or units which occur infrequently in very general terms, but to describe the predominant units (such as the carbon atom or benzene ring, and their environments) in considerable detail.

The distribution of alphabetic characters in texts has been widely studied; early investigations have been summarised by Bourne (1963) and Schwartz (1963). Dolby and Resnikoff (1964) have examined the vowel/consonant structures of English words appearing in standard dictionaries, and cryptographic methods are, of course, especially concerned with frequency distributions of characters (Gaines, 1956). In one particularly interesting study, Newman and Gerstman (1952) analysed a 10,000 word sample of Isaiah, and computed coefficients of constraints between characters at different separations. They surmised that at distances of over five or six letters the values of the coefficients would drop, since the sequential dependencies which connect words would be weaker than those within words. This supposition was not borne out by the experiment. In the context of character recognition, Edwards and Chambers (1964) have found that letter combination frequencies could improve automatic character recognition, while more recently, Casey and Nagy (1971) have developed algorithms for a reading machine which does not require *a priori* information on character patterns, but develops its own pattern/identity associations in an adaptive mode starting from letter and digram frequencies.

2. File analysis

The file chosen for analysis was one issue of *Chemical Titles*, a periodical and magnetic tape publication of the *Chemical*



```
BELL RG              MATSCHINER JT
VITAMIN K ACTIVITY OF PHYLLO QUINONE OXIDE.
ABBIA4-041-0473
```

Fig. 1. **A typical entry from the *Chemical Titles* data base**

Table 1 Character frequencies for two 1,000-title samples

SYMBOL	SAMPLE 1	SAMPLE 2	SYMBOL	SAMPLE 1	SAMPLE 2
0	48	38	E	7440	7462
1	149	82	F	2143	2349
2	95	72	G	1042	1097
3	86	52	H	2205	2145
4	69	38	I	6902	6798
5	58	35	J	23	29
6	25	34	K	196	216
7	25	10	L	3356	3362
8	16	8	M	2330	2389
9	10	17	N	5738	5866
=	2	3	O	6659	6730
∇	12793	13143	P	2013	1885
(	199	119	Q	107	116
)	200	118	R	4586	4638
*	1	2	S	4397	4443
+	14	20	T	5882	5947
,	287	227	U	2019	2104
−	628	443	V	539	548
.	1226	1226	W	295	306
/	29	25	X	354	331
A	5769	5811	Y	1695	1519
B	965	977	Z	285	235
C	3400	3412	$	1	1
D	2724	2530	Total	89025	88958

Table 2 Extract from digram frequencies for two 1,000-title samples

DIGRAM	SAMPLE 1	SAMPLE 2	DIGRAM	SAMPLE 1	SAMPLE 2
LV	589	610	LO	305	288
L(	12	5	LP	8	9
L)	16	6	LS	72	89
L,	6	4	LT	83	73
L−	27	28	LU	215	228
L.	33	28	LV	15	12
LA	433	447	LW	1	5
LB	7	13	LY	203	207
LC	29	47	M6	1	—
LD	36	49	M∇	349	367
LE	484	514	M(	16	16
LF	70	48	M)	5	3
LG	2	6	M,	14	11
LI	467	428	M−	22	16
LK	43	29	M.	39	44
LL	190	177	M/	2	—
LM	10	12	MA	278	312

Table 3 Number of different types of *n*-grams

	NO. OF DIFFERENT *n*-GRAMS			
SAMPLE	1	2	3	5
1	47	821	5,138	25,213
2	47	803	4,902	24,872

Abstracts Service. This is issued biweekly, and includes the titles, authors' names, and bibliographic references of currently published articles of chemical interest. The issue used was No. 1, 1971, dated 11 January. A typical entry from the issue is shown in **Fig. 1**; the bibliographic reference is given as the ASTM Coden.

The titles are recorded in upper-case characters. An occasional artefact arises through the insertion of additional space symbols; the printed publication includes a KWIC (Key Word In Context) index, and the spaces ensure that certain chemical word stems such as QUINONE in Fig. 1 (the word is normally written as PHYLLOQUINONE) are indexed.

A set of simple programs (written in PLAN, the ICL 1900 series assembly language) was devised to produce counts of *n*-grams (i.e., strings of 1, 2, 3 and 5 characters), including the space character, for values of *n* between 1 and 5. The program to count single character occurrences used the binary value of the character code to address a position in a 62-word array. The digrams were counted by using a two-dimensional array (62 × 62 = 3844). Longer *n*-grams (n = 3 and 5) were created by taking a window equal to that number of characters and moving it along the title record, creating a new record at each position (a space was inserted as the initial character of each title). The records were written to tape, and subsequently sorted, counted and printed.

3. *n*-gram counts

Two samples, each of 1,000 titles, were examined; on average, each title consisted of just under 90 characters, and contained 12 words of 6·5 characters. A total of 47 different characters was encountered. **Table 1** shows the frequencies of these characters for each of the samples.

The most frequent character is obviously the space; the most frequent alphabetic character is E, closely followed by I, while the least frequent alphabetic symbol is J. The consistency in the figures is remarkable—apart from the hyphen or minus sign the figures fall within a few per cent of one another, even in the case of the less frequent characters. The overall frequencies clearly reflect the occurrence of special chemical terminology, (e.g., parentheses, and other special characters).

This same consistency is found in the case of the digrams. **Table 2** shows extracts from the data, illustrating the high degree of correspondence. Many of the possible digrams did not appear in these samples, as might be expected; the total number of different digrams generated for the first sample of 1,000 titles was 821, although the total possible number is $47^2 = 2209$.

Five digrams exceeded a total frequency of 1,500 for 1,000 titles, i.e. they occurred, on average, at least 1·5 times per title. These are the digrams IN, E∇, N∇, ON, and TI. Clearly, the frequency distribution of digrams is also highly Zipfian.

These figures, and those for trigrams and pentagrams are summarised in **Table 3**, while **Tables 4 and 5** show extracts from the listings, the first an alphabetical list of trigrams, the second a ranked list of pentagrams. Predictably, the effect of increasing the length of the character string uniformly is to increase the variety of types very substantially. A secondary effect, however, as seen from the Tables, is that the peak frequencies are steadily reduced, from one occurrence of the space symbol in seven characters in the case of single characters, to a maximum frequency of approximately 600 in 1,000 documents in the case of the most frequent pentagram, TION∇.

This is further emphasised by comparing the number of *n*-grams of each type examined which exceed particular frequency limits, as shown in **Table 6.** The data indicate unequivocally that as string-length is increased, the number of *n*-grams exceeding an arbitrarily set limit declines, and that the proportion of *n*-grams doing so is vastly reduced.

Table 4 Extract from trigram frequencies for two 1,000-title samples

TRIGRAM	SAMPLE 1	SAMPLE 2	TRIGRAM	SAMPLE 1	SAMPLE 2
LA∇	6	8	LAT	151	151
LA)	—	2	LAU	4	1
LA–	1	—	LAV	5	9
LA.	1	28	LAW	—	1
LAB	17	19	LAX	7	8
LAC	29	30	LAY	12	6
LAD	10	4	LBE	2	—
LAE	1	—	LBI	2	2
LAG	7	4	LBO	1	—
LAI	—	2	LBR	1	2
LAM	14	9	LBU	1	1
LAN	41	42	LCA	1	2
LAP	4	3	LCH	1	1
LAR	70	68	LCI	14	20
LAS	53	76	LCO	7	11

Table 5 Extract from ranked pentagram frequencies

PENTA-GRAM	SAMPLE 1	SAMPLE 2	PENTA-GRAM	SAMPLE 1	SAMPLE 2
TION∇	593	566	CTION	154	135
ATION	497	477	S∇IN∇	149	148
∇AND∇	446	463	∇OF∇A	139	137
∇THE∇	378	456	E∇OF∇	127	143
ION∇O	371	364	TIONS	123	101
N∇OF∇	369	371	∇OF∇S	122	99
ON∇OF	363	359	N∇THE	120	142
S∇OF∇	336	345	ES∇OF	119	119
∇OF∇T	205	261	HYDRO	116	72
OF∇TH	171	215	∇WITH	115	79
F∇THE	162	195	∇HYDR	113	73

Table 6 Comparison of n-gram length with numbers exceeding given frequency limits (1,000-title sample)

UPPER FREQUENCY LIMIT	NO. OF n-GRAMS EXCEEDING THIS LIMIT $n =$ 1	2	3	5
1000	20	14	2	—
500	23	49	10	1
200	27	90	44	9
100	32	187	151	28
50	36	261	405	124
20	41	361	983	519
Total no. of different n-grams	47	821	5,183	25,213

4. Variable-length equi-frequent strings

The data presented above lead to the conclusion that it is possible to select, for particular data-bases, units other than words into which natural language text can be analysed automatically so that although the length of the strings varies, their frequencies fall below predetermined limits. The following procedure was devised to examine this proposition; although it has been performed manually thus far, it can easily be automated at a later stage. A series of arbitrarily chosen frequency limits was selected, in terms of frequency of occurrence of particular n-grams in a sample of 1,000 documents. The list of digram frequencies was examined to determine which of them exceeded the particular limit. For these the list of trigrams was examined, and the trigram or trigrams which resulted in reduction of the figures below the limit were selected. If this failed, the lists of pentagrams derived from the di- and trigrams were also examined, with a similar purpose. Thus, in the case of the digram GE, with a frequency of 249 in the first sample, a frequency limit of 200 was obtained by selecting the trigram GEN, with a frequency of 155, the others, denoted by GE*, then having a residual count of 94. If the upper limit was set at 100, then the trigram GEN is over-populated. Examination of the pentagram listing shows that the most frequent tetragram was GEN∇, with a count of 100, so that the residual frequency of GEN* then becomes 55, well within the limit. **Fig. 2** illustrates this more fully for the digram LA with a frequency list of 100. The underlined n-grams are those finally selected.

At the same time, for the samples studied and with frequency limits in these ranges, certain highly frequent pentagrams still exceeded the notional limits. Examples of these are the strings ∇THE∇, ∇AND∇, ATION, ION∇O, and N∇OF∇; these were not studied further, as their value as keys in retrieval would be so low as to be useless. As expected, the distribution of n-grams selected in this manner showed a close correspondence with the data given in Table 6.

5. Possible means of implementation of file analysis and search

The implications of these findings for file-inversion of natural-language texts are considerable. Given the present method of identifying words in text, any considerable volume of text leads to a large dictionary; it is estimated that for a file of 250,000 documents representing a year's cumulation of *Chemical Abstracts Condensates* (text of titles plus added keywords), the dictionary may contain well over 100,000 terms. The system designer is thus constrained to use a large dictionary, and has little latitude in assigning storage space on rotating storage for the lists of document identifiers associated with each key chosen. A method based on the identification of variable-length character strings of approximately equal probability might provide the designer with greater flexibility in his choice of parameters for the system, trading off the size of the dictionary of text keys against their frequency distributions and the length of the lists of document identifiers.

A variety of strategies can be envisaged for implementing the technique, differentiated chiefly by their redundancy and by the level of detail stored in the document description. The possible extremes of analysis of the title shown in Fig. 1 are illustrated in **Figs. 3 and 4.**

The analysis exemplified in Fig. 3 proceeds from each character in the text, and identifies a variable-length character string from each, with a minimal length of 2. As a result, the number of key-document assignments equals the number of characters in the record. Identification of a search term in a record would

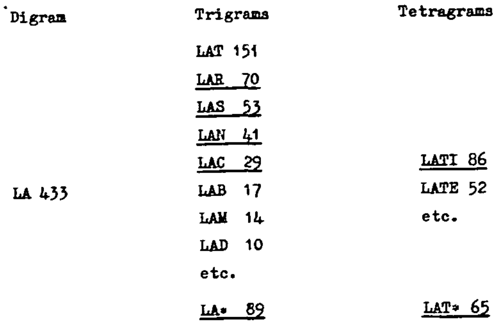


Fig. 2. Extraction of variable-length strings (The numbers represent frequency in a 1,000-title sample

```
              VITAMIN K ACTIVITY OF PHYLLO QUINONE OXIDE

∇V          ∇K          ITY          YL*          NO*
 VIT         K∇          TY           LL           ONE
  ITA         ∇ACT        Y∇O          LO*          NE∇*
   TA*         ACTI        ∇OF∇P        O∇*          E∇O
    AMI         CTI         F∇P          ∇Q           ∇OX
     MIN         TIV         ∇PH          QU           OX
      IN∇*        IVI         PHY          UI           XI
       N∇*         VIT         HY*          INO          IDE
```

Fig. 3. Character-by-character analysis of title

```
          VITAMIN K ACTIVITY OF PHYLLO QUINONE OXIDE

∇V          ∇ACT          HY*          UI          IDE
  ITA         IVI           LL           NO*
    MIN         TY            O∇*          NE∇*
      ∇K          ∇OF∇P         Q            OX
```

Fig. 4. Non-redundant analysis of title

```
          HIGHER ACTIVITY OF VITAMIN K.......

∇HI          ACTI          F∇V
  GH           VIT           etc.
    ER∇          Y∇O
```

Fig. 5. Alternative analysis of the string ∇ACTIVIT

```
                          ∇ACTIVIT

┌             ┐      ┌             ┐      ┌             ┐      ┌             ┐
  ∇ACT                 **∇                  **∇A                 *∇AC
  and  IVI       or    and ACTI       or    and   CTI      or    and  TIV
  and      T*          and     VIT          and       VIT        and      IT
└             ┘      └             ┘      └             ┘      └             ┘
```

Fig. 6. Possible key groupings for search from ∇ACTIVIT

be carried out by intersecting its constituent character strings. Thus a search for the string ∇ACTIVIT would be performed by AND-ing the keys ∇ACT, ACTI, CTI, TIV, IVI and VIT. Given such a high degree of redundancy, the probability of a record containing these keys in an order other than that intended would be very low, although this would depend in part on average document length. Clearly, retrieval of any string, including those truncated on either or both sides, would be simple. At the same time, the number of key-document assignments is much higher than in the case of word identification, and the storage requirements probably excessive.*

Moreover, the need to carry out logical operations at the search file level rather than at the document file level (for logical conditions such as FOLLOWED BY) implies that the document identifier stored with each key should carry further positional information, thus increasing storage requirements still further. If this were added (possibly in terms of character-positions) it would seem possible to reduce the number of key-document assignments quite substantially, by making the analysis non-redundant as shown in Fig. 4. In this case, the search term ∇ACTIVIT would be located in the document in question by AND-ing the terms ∇ACT, IVI, T*, and by operating on the character positions stored with the document numbers. It will be apparent, however, that the analysis of the string is not unique in this strategy, as illustrated by the possibility of a title such as that in **Fig. 5.** The string ∇ACTIVIT would be retrieved in this instance by intersection of the keys ACTI and VIT, together with any keys having a space as the last character. Accordingly, a non-redundant analysis would have to trade off reduced storage requirements and faster analysis against the need to map the search string onto the range of possible analyses, as illustrated in **Fig. 6.** As with the fully redundant analysis indicated in Fig. 3, searches for any string would be possible, but this method would require indexing of each key string to enable access to be made to any character within it.

6. Conclusions

The findings of this study suggest an alternative method of file analysis for storage and search of natural-language text. Its principal advantages are that it offers the systems designer greater flexibility in determining the size and frequency range of the dictionary of keys employed, and in this respect the objectives are strongly supported by information theoretic considerations. Thus, we quote Young (1971):

> 'Any change in a communications system which tends to equalise the probabilities of occurrence of the various symbols has the effect of increasing the information content of each message.'

The sample analysis is limited, both in provenance and time-span. Extension of the studies to other data bases and document types is clearly necessary. Vocabulary differences in other data bases may be considerable, and the method depends on the stability of character and n-gram frequencies over a period of time. Titles show a high proportion of subject words, while abstracts, for instance, will include a wider variety of syntactic constructions, with verb forms etc., which will also influence textual characteristics. Nonetheless, the study points the way to a reduction of certain of the constraints determining the design of inverted file systems for searching free-text data bases, and, as a result of research now in hand, may well lead to more effective exploitation of information resources.

Acknowledgements

We are grateful to G. W. Adamson and M. J. Snell for invaluable suggestions during this work, and Dr. A. K. Kent, Technical Director of UKCIS, for provision of the sample file of *Chemical Titles*.

*We are grateful to Mr. I. D. McCraken for pointing this out to us.

References

BOURNE, C. P. (1963). *Methods of information handling*, New York: Wiley.

CASEY, R. G., and NAGY, G. (1971). Advances in pattern recognition, *Scientific American*, Vol. 224, pp. 56-71.

COLOMBO, D. S., and RUSH, J. E. (1969). Use of word-fragments in computer-based retrieval systems, *J. Chemical Documentation*, Vol. 9, pp. 47-50.

CROWE, J. E., LYNCH, M. F., and TOWN, W. G. (1970). Analysis of structural characteristics of chemical compounds in a large computer-based file, Part I, Non-cyclic fragments. *J. Chem. Soc., C*, pp. 990-996.

DOLBY, J. L., and RESNIKOFF, H. L. (1964) On the structure of written English words, *Language*, Vol. 40, pp. 167-96.

EDWARDS, A. W., and CHAMBERS, R. L. (1964). Can *a priori* probabilities help in character recognition? *JACM*, Vol. 11, pp. 465-470.

GAINES, H. F. (1956). *Cryptanalysis*, New York: Dover.

HEAPS, H. S., and THIEL, L. H. (1970). Optimum procedures for economic information retrieval, *Inform. Stor. Retr.*, Vol. 6, pp. 137-153.

HIGGINS, L. D., and SMITH, F. J. (1971). Disc access algorithms, *The Computer Journal*, Vol. 14, pp. 249-253.

KENT, A. K. (1968). The Chemical Society Research Unit in information dissemination and retrieval, *Svensk Kem. Tidskr.*, Vol. 80, pp. 39-46.

NEWMAN, E. B., and GERSTMAN, L. J. (1952). A new method for analysing printed English, *J. Experimental Psychology*, Vol. 44, pp. 114-125.

SCHWARTZ, E. S. (1963). A dictionary for minimum redundancy encoding, *JACM*, Vol. 10, pp. 413-439.

YOUNG, J. F. (1971). *Information Theory*, London: Butterworth, p. 49.

ZIPF, G. K. (1949). *Human behaviour and the principle of least effort*, Cambridge, Mass.: Addison-Wesley.