*Inform. Stor. Retr.* Vol. 9, pp. 331–337. Pergamon Press 1973. Printed in Great Britain

# ANALYSIS OF THE MICROSTRUCTURE OF TITLES IN THE INSPEC DATA-BASE

MICHAEL F. LYNCH, J. HOWARD PETRIE and MICHAEL J. SNELL

Postgraduate School of Librarianship and Information Science,
University of Sheffield, Sheffield, United Kingdom

**Summary**—A high degree of constancy has been found to exist in the microstructure of titles of samples of the INSPEC data-base taken over a 3-year period. Character and digram frequencies are shown to be relatively stable, while variable-length character-strings characterizing samples separated by 3 years in time show close similarities.

## INTRODUCTION

THE design of inverted-file retrieval systems for computer searches of bibliographical data-bases poses considerable problems. The conventional approach is to use content words as the keys by means of which the document citations are organized and accessed. The content words may either be assigned subject headings, or text words identified in natural-language data elements (titles, abstracts, keywords), after exclusion of the most frequent, and hence least useful, by means of a stop-word list. In the latter case, the well-known hyperbolic distribution of text words [1] implies a large and constantly growing dictionary; in the case of *Chemical Abstracts Condensates*, for instance, it has been estimated that a 5-year cumulation will result in a dictionary of over a million word types. Since, in disciplines such as chemistry, the ability to carry out searches involving left-hand truncation of profile terms is essential, the items in the dictionary itself may have to be indexed too. In addition, the very disparate postings of document references to the dictionary items may lead to inefficient use of direct-access storage.

An alternative strategy to the use of text words as keys for file organization has recently been suggested [2]. The basis for this approach is to regard text searches as searches for character strings which may or may not be bounded by spaces, and to consider the text of the data-elements, including all space characters, as character strings. Indeed, this is the manner in which text is stored in machine-readable form; fragmentation of the character strings into words is essentially a human activity, which may be reflected in an appropriate program. The task of searching then reduces to indexing the text string in such a way as to optimize both storage and retrieval functions. On general information theoretic grounds, this indexing is optimized when the characteristics chosen for identification are equi-frequent. Neither with words nor with individual characters is this condition fulfilled. In the case of characters, for instance, identification of those text portions in which the character J is present is useful in cases in which this letter appears in a profile term; this eventuality is unlikely to be frequent, since J generally occurs infrequently. The essential problem with characteristics of disparate frequency is that those which have a high selectivity, or resolving power, by virtue of their infrequency, are used only occasionally, while those which are frequently called for have relatively low resolving power. It has already been demonstrated, however, that it is possible to overcome the disparate frequency of occurrence of individual characters in text by concatenating strings of characters such that, in general, the longest strings are generated from the most frequent characters, and the shortest from those that appear least often.

This can be accomplished by the use of key sets consisting of variable-length character strings which are identified by iterative analysis of text files. At each stage of the analysis, character strings of length n (or n-grams) are generated, where n is successively increased by 1. Thus, initially, the frequencies of single characters are determined. Those whose frequencies fall below some arbitrary limit are not considered for concatenation. From those characters which exceed the limit, digrams are produced; their frequencies are again compared with the limit, and trigrams are produced which begin with those digrams, etc. The result of this procedure is a set of variable length n-grams, the frequencies of which fall within a certain range. The n-gram set, or key set, then effectively characterizes the text file. The number of keys in the key set can be modified by appropriate choice of the frequency limit, so that the size of the "dictionary" is now subject to control.

The preliminary work [2] leading to this conclusion was carried out on a single issue of *Chemical Titles* (an outline of several methods by means of which the technique might be implemented is also given in the earlier paper). Subsets of the *Chemical Titles* file showed remarkable constancy in regard to the frequencies and rank orders of individual characters and n-grams. One critical factor in determining the applicability of these keys for organizing large retrospective data bases is the degree of constancy shown by a data base over a period of time. It is obviously desirable that the distribution characteristics should remain substantially static; otherwise the ability of a particular set of n-gram keys to reflect the nature of the data base might deteriorate. Reprocessing of the files to generate and apply a new key set, or partitioning of the files over successive time periods would then be required. Both procedures would be expensive and disadvantageous in other respects.

The work reported here consists of an examination of the degree of variation in a single data-base—INSPEC—over a period of 3 years. Ten individual files were analysed. For each file, the frequencies of single characters, including the space symbol, and all digrams, were determined for the title data element, and arranged both in lexicographic and rank-frequency orders. In addition, complete sets of n-gram keys were computed for the two files most widely separated in time. Finally, for a single file, four keys sets at different frequency limits were generated in order to study changes in the sizes of key sets with variations in the limit. This was also necessary in order to assess the average number of keys likely to be assigned for each title during file inversion.

This analysis thus provides multiple cross-sections of the files by means of which the relative constancy of the characteristics may be determined.

## RESULTS

Frequency data for single characters is shown in Table 1. Only the first 29 most frequent characters are shown and these are arranged in ranked order. Means and standard deviations are shown alongside the corresponding data in this figure. The conclusion reached is that inter-file differences over this three year period are minimal, and no significant trends are apparent.

Although not shown here a similar close correspondence exists between the digrams generated from the same 10 files.

TABLE 1. NORMALIZED FREQUENCIES WITH MEANS AND STANDARD DEVIATIONS FOR THE FIRST 29 CHARACTERS (ARRANGED IN RANKED ORDER). THE FILES ANALYSED RANGE FROM INSPEC 31002 (1969) TO INSPEC 31060 (1972)

| | 31002 | 31003 | 31015 | 725 | 31016 | 710 | 31017 | 31056 | 31057 | 31060 | S.D. | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ∇ | 0·1511 | 0·1505 | 0·1488 | 0·1508 | 0·1332 | 0·1499 | 0·1504 | 0·1485 | 0·1498 | 0·1502 | 0·0054 | 0·1483 |
| E | 0·0889 | 0·0900 | 0·0885 | 0·0890 | 0·0890 | 0·0903 | 0·0900 | 0·0902 | 0·0906 | 0·0883 | 0·0039 | 0·0875 |
| T | 0·0730 | 0·0727 | 0·0729 | 0·0724 | 0·0725 | 0·0724 | 0·0722 | 0·0721 | 0·0719 | 0·0728 | 0·000360 | 0·0725 |
| I | 0·0725 | 0·0736 | 0·0755 | 0·0727 | 0·0741 | 0·0739 | 0·0738 | 0·0735 | 0·0736 | 0·0731 | 0·000856 | 0·0731 |
| O | 0·0722 | 0·0701 | 0·0705 | 0·0712 | 0·0705 | 0·0699 | 0·0695 | 0·0701 | 0·0693 | 0·0715 | 0·000913 | 0·0705 |
| N | 0·0677 | 0·0671 | 0·0677 | 0·0680 | 0·0672 | 0·0672 | 0·0678 | 0·0662 | 0·0669 | 0·0674 | 0·000527 | 0·0673 |
| A | 0·0641 | 0·0647 | 0·0659 | 0·0658 | 0·0645 | 0·0651 | 0·0644 | 0·0661 | 0·0658 | 0·0654 | 0·000712 | 0·0651 |
| R | 0·0568 | 0·0569 | 0·0565 | 0·0558 | 0·0572 | 0·0576 | 0·0571 | 0·0592 | 0·0573 | 0·0563 | 0·000949 | 0·0570 |
| S | 0·0530 | 0·0529 | 0·0542 | 0·0522 | 0·0537 | 0·0533 | 0·0535 | 0·0541 | 0·0537 | 0·0530 | 0·000608 | 0·0534 |
| C | 0·0398 | 0·0397 | 0·0401 | 0·0392 | 0·0402 | 0·0397 | 0·0403 | 0·0406 | 0·0394 | 0·0388 | 0·000545 | 0·0398 |
| L | 0·0370 | 0·0370 | 0·0379 | 0·0374 | 0·0378 | 0·0375 | 0·0370 | 0·0369 | 0·0370 | 0·0375 | 0·000368 | 0·0373 |
| M | 0·0267 | 0·0259 | 0·0271 | 0·0267 | 0·0267 | 0·0267 | 0·0247 | 0·0278 | 0·0257 | 0·0271 | 0·000870 | 0·0265 |
| F | 0·0259 | 0·0260 | 0·0261 | 0·0258 | 0·0259 | 0·0251 | 0·0262 | 0·0245 | 0·0262 | 0·0257 | 0·000540 | 0·0257 |
| D | 0·0248 | 0·0256 | 0·0256 | 0·0258 | 0·0261 | 0·0252 | 0·0259 | 0·0258 | 0·0258 | 0·0254 | 0·000380 | 0·0256 |
| U | 0·0238 | 0·0238 | 0·0238 | 0·0233 | 0·0240 | 0·0231 | 0·0234 | 0·0232 | 0·0240 | 0·0237 | 0·000331 | 0·0236 |
| H | 0·0222 | 0·0226 | 0·0208 | 0·0218 | 0·0216 | 0·0220 | 0·0222 | 0·0214 | 0·0224 | 0·0217 | 0·000529 | 0·0219 |
| P | 0·0220 | 0·0210 | 0·0217 | 0·0218 | 0·0215 | 0·0223 | 0·0213 | 0·0227 | 0·0215 | 0·0224 | 0·000531 | 0·0218 |
| G | 0·0156 | 0·0156 | 0·0159 | 0·0146 | 0·0160 | 0·0162 | 0·0151 | 0·0165 | 0·0155 | 0·0153 | 0·000554 | 0·0156 |
| Y | 0·0126 | 0·0129 | 0·0125 | 0·0124 | 0·0122 | 0·0122 | 0·0123 | 0·0119 | 0·0125 | 0·0127 | 0·000310 | 0·0123 |
| B | 0·0086 | 0·0089 | 0·0087 | 0·0092 | 0·0086 | 0·0085 | 0·0089 | 0·0084 | 0·0090 | 0·0090 | 0·000257 | 0·0089 |
| V | 0·0071 | 0·0072 | 0·0071 | 0·0071 | 0·0076 | 0·0076 | 0·0073 | 0·0078 | 0·0074 | 0·0069 | 0·000285 | 0·0073 |
| — | 0·0061 | 0·0063 | 0·0063 | 0·0064 | 0·0063 | 0·0063 | 0·0062 | 0·0061 | 0·0064 | 0·0060 | 0·000137 | 0·0063 |
| W | 0·0052 | 0·0055 | 0·0056 | 0·0048 | 0·0057 | 0·0056 | 0·0053 | 0·0056 | 0·0056 | 0·0050 | 0·000303 | 0·0054 |
| X | 0·0026 | 0·0030 | 0·0026 | 0·0031 | 0·0025 | 0·0027 | 0·0027 | 0·0025 | 0·0027 | 0·0028 | 0·000199 | 0·0027 |
| K | 0·0022 | 0·0022 | 0·0021 | 0·0022 | 0·0022 | 0·0022 | 0·0023 | 0·0023 | 0·0023 | 0·0023 | 0·000067 | 0·0022 |
| . | 0·0020 | 0·0021 | 0·0021 | 0·0020 | 0·0022 | 0·0020 | 0·0023 | 0·0019 | 0·0019 | 0·0017 | 0·000169 | 0·0020 |
| * | 0·0019 | 0·0016 | 0·0015 | 0·0019 | 0·0015 | 0·0017 | 0·0017 | 0·0013 | 0·0013 | 0·0020 | 0·000246 | 0·0016 |
| Q | 0·0018 | 0·0019 | 0·0018 | 0·0017 | 0·0019 | 0·0020 | 0·0018 | 0·0019 | 0·0018 | 0·0017 | 0·000095 | 0·0018 |
| Z | 0·0016 | 0·0015 | 0·0015 | 0·0016 | 0·0015 | 0·0015 | 0·0014 | 0·0015 | 0·0016 | 0·0017 | 0·000084 | 0·0015 |

As previously mentioned, it is important that key sets generated from a particular data base should be reasonably consistent over a substantial period of time. In order to investigate this titles from equal-sized samples of the two INSPEC files most distant in time (INSPEC 31002 and 31060—1969 and 1971 respectively) were analysed. A large proportion of each of the two samples corresponded to issues of the INSPEC publication *Computer and Control Abstracts*. An arbitary limit of 100 occurrences within an 88,000 sample was chosen as the frequency limit for key generation. For both samples this necessitated generation of strings of up to 10 characters. Most keys formed at this end of the spectrum comprised common non-content-bearing constructions, as for example, ∇OF∇THE∇S.

Key sets produced from both of these samples were remarkably similar in size; 1318 keys were generated from INSPEC 31002 and 1363 from INSPEC 31060. Table 2 shows the number of keys generated from each sample at each n-gram level. The numbers of keys which were found to be common to both samples at each n-gram level are also shown and indicate a close correspondence between the two files.

TABLE 2. DISTRIBUTION OF KEYS WITHIN KEY SETS DERIVED FROM SAMPLES MOST DISTANT IN TIME

| | No. of single char. keys | No. of digram keys | No. of trigram keys | No. of tetra-gram keys | No. of penta-gram keys | No. of hexa-gram keys | No. of hepta-gram keys | No. of octo-gram keys | No. of nona-gram keys | No. of deca-gram keys |
|---|---|---|---|---|---|---|---|---|---|---|
| INSPEC 31002 (1969) | 25 | 124 | 591 | 308 | 133 | 83 | 31 | 13 | 5 | 5 |
| No. of keys common to both samples | 22 | 103 | 487 | 228 | 106 | 61 | 21 | 9 | 5 | 5 |
| INSPEC 31060 (1972) | 22 | 124 | 584 | 308 | 144 | 90 | 46 | 15 | 17 | 13 |

Total key set sizes:
INSPEC 31002 = 1318.
INSPEC 31060 = 1363.

Understandably the key sets were not completely identical, but the differences were confined to keys of lower than average frequency. The average frequency of keys common to both samples analysed was found to be 41 occurrences within the sample, while the average frequency of those keys not found in both samples was found to be 34 occurrences. Certain of the non-common keys actually occurred in the other sample but with a frequency slightly greater than the limit and consequently were in need of extension. When these were excluded from the calculations the average frequency of non-common keys was found to be 28.

The more noticeable dissimilarities between the two key sets occurred for the longest keys, which are shown in Table 3 for n = 8,9 and 10. Again, in the majority of cases the non-common keys were below average frequency but for the few instances in which this was not the case the keys in question were indicative of the content, for example, ∇ELECTRI and ∇COMPUTE. This could perhaps be explained by the fact that the samples did not correspond exactly to issues of *Computer and Control Abstracts*, ∇ELECTRI, for instance, indicates an overlap of one sample into a section corresponding to an issue of the INSPEC publication *Electrical and Electronics Abstracts.*

TABLE 3. KEYS FROM TWO INSPEC FILES FOR n = 8,9 AND 10. THE NUMBERS IN PARENTHESES ARE THE FREQUENCY OF THE n-GRAM IN THE SAMPLE

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *OCTOGRAM KEYS* | | | | | | | |
| INSPEC 31060 | | | | INSPEC 31002 | | | |
| ICATION▽ | (89) | OF▽THE▽E | (13) | ▽COMPUTE | (95) | ION▽OF▽R | (14) |
| ION▽OF▽T | (64) | OF▽THE▽S | (13) | ▽ELECTRI | (66) | ION▽OF▽C | (12) |
| ION▽OF▽A | (39) | OF▽THE▽M | (11) | ▽SYSTEM▽ | (94) | OF▽THE▽P | (17) |
| ION▽OF▽S | (23) | OF▽THE▽P | (11) | ION▽OF▽T | (57) | OF▽THE▽C | (16) |
| ION▽OF▽D | (16) | ONTROL▽O | (53) | ION▽OF▽A | (34) | ONTROL▽S | (37) |
| ION▽OF▽C | (14) | ONTROL▽S | (42) | ION▽OF▽S | (19) | ONTROL▽O | (27) |
| ION▽OF▽P | (13) | | | | | | |
| MATION▽O | (21) | | | | | | |
| OF▽THE▽C | (13) | | | | | | |
| *NONAGRAM KEYS* | | | | | | | |
| ▽OF▽THE▽C | (13) | SYSTEMS▽A | (8) | ▽OF▽THE▽P | (17) | | |
| ▽OF▽THE▽S | (13) | SYSTEMS▽I | (8) | ▽OF▽THE▽C | (16) | | |
| ▽OF▽THE▽E | (12) | SYSTEMS▽F | (7) | CONTROL▽S | (37) | | |
| ▽OF▽THE▽M | (11) | SYSTEMS▽U | (5) | CONTROL▽O | (27) | | |
| ▽OF▽THE▽P | (11) | SYSTEMS▽O | (4) | SYSTEMS▽W | (10) | | |
| ▽SYSTEM▽F | (24) | SYSTEMS▽B | (3) | | | | |
| CONTROL▽O | (53) | SYSTEMS▽R | (3) | | | | |
| CONTROL▽S | (42) | SYSTEMS▽C | (2) | | | | |
| SYSTEMS▽W | (19) | | | | | | |
| *DECAGRAM KEYS* | | | | | | | |
| ▽CONTROL▽O | (53) | ▽SYSTEMS▽O | (14) | ▽CONTROL▽S | (37) | | |
| ▽CONTROL▽S | (41) | ▽SYSTEMS▽B | (3) | ▽SYSTEMS▽W | (10) | | |
| ▽SYSTEMS▽W | (18) | ▽SYSTEMS▽R | (3) | ATION▽OF▽T | (46) | | |
| ▽SYSTEMS▽A | (8) | ATION▽OF▽T | (46) | ATION▽OF▽A | (23) | | |
| ▽SYSTEMS▽I | (8) | ATION▽OF▽A | (27) | ATION▽OF▽S | (14) | | |
| ▽SYSTEMS▽F | (7) | ATION▽OF▽S | (16) | | | | |
| ▽SYSTEMS▽U | (5) | | | | | | |

As mentioned earlier the number of keys present in a key set can be modified by the frequency limit chosen. To investigate the change in key set size with change in frequency limit, a sample from a third INSPEC file (INSPEC 725) was analysed. This sample was of the same size as the previous two and corresponded to the year 1971. Key sets were generated from 88,000 characters of this file for four different frequency limits (30, 100, 400 and 800 occurrences), and Table 4 shows the number of keys present at each n-gram level for these four frequency limits. It can be seen that the higher the frequency limit, the sooner strings needing expansion are exhausted. Consequently the key set derived at a frequency limit of 800 is completed at the hexagram level, while that derived at a limit of 30 is not completed until keys of length 14 are generated. As had been seen with the previous two INSPEC samples, most of the longer keys are of limited use for retrieval.

Curves indicating the distribution of keys with respect to their length are shown in Fig. 1 for each of the four frequency limits.

## SUMMARY

The results described in this paper form the initial investigation into a technique utilizing variable-length character strings as retrieval keys for natural language data bases. In order to show such a key set would be applicable to a data-base over a period of

TABLE 4. KEY SETS FROM INSPEC 725 FOR FOUR FREQUENCY LIMITS

| Number of n-grams where N = | Cut-off 30 | Cut-off 100 | Cut-off 400 | Cut-off 800 |
|---|---|---|---|---|
| 1 | 12 | 20 | 29 | 32 |
| 2 | 137 | 141 | 132 | 111 |
| 3 | 950 | 644 | 165 | 33 |
| 4 | 1089 | 303 | 26 | 5 |
| 5 | 606 | 120 | 17 | 2 |
| 6 | 417 | 70 | 8 | — |
| 7 | 297 | 33 | — | — |
| 8 | 142 | 9 | — | — |
| 9 | 102 | 13 | — | — |
| 10 | 57 | 2 | — | — |
| 11 | 16 | — | — | — |
| 12 | 6 | — | — | — |
| 13 | 4 | — | — | — |
| 14 | 6 | — | — | — |
| Total | 3841 | 1355 | 377 | 183 |

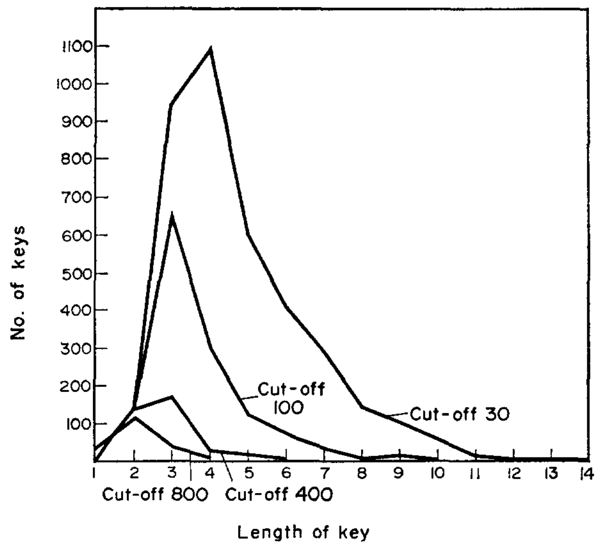


FIG. 1. Distribution characteristics of key sets generated from 88,000 characters of INSPEC 725 (1971) for four different cut-off levels.

time it was necessary to demonstrate that no radical changes in characteristics would occur within a data base during this time. The frequencies of occurrence of single characters and digrams within ten INSPEC files ranging from 1969 to 1971 were examined. Both showed little evidence of any significant trends throughout this period. Key sets derived from two INSPEC files separated in time by 3 years were also examined. These key sets showed substantial similarity in size and also in their constituent keys.

Finally, the effect a change in frequency limit had on key set size and on the distribution in length of the individual keys was studied. The results of this are shown graphically.

## PROGRAM DETAILS

The variable-length character string keys described in this paper have been generated with the use of programs written in the ICL 1900 Series assembly language, PLAN. At each n-gram level there are basically three stages in the key generation process. In the first a window of length n-characters is moved along each data element and the n-grams thus generated are matched against a list of n-grams found in the previous iteration to have occurred with a frequency greater than the stipulated limit. If a match is found the next character in sequence is picked up from the text and the $n+1$ gram thus created is written to tape. On completion the file of $(n+1)$-grams is sorted alphabetically. The second stage of the process merely counts the various $(n+1)$-gram types, and creates a new file in which these strings are sorted alphabetically on the first n characters and, within this, in descending order of the frequencies of occurrence of each $(n+1)$-gram. In the final key generation stage the frequencies of the $(n+1)$-grams are successively subtracted from the frequency with which their respective "parent" n-gram had occurred. This process is continued until the frequency of the "parent" n-gram has been reduced to a value below the stipulated limit. At each subtraction the frequency of the $n+1$-gram is itself tested—if it occurs below the frequency limit it is chosen as a key, if not it is in need of expansion and, in the next iteration, will be input into the list mentioned above.

*Acknowledgement*—We thank the Office for Scientific and Technical Information, London, for a grant in support of this work, and the Institute of Electrical Engineers, London, for the file samples and file-handling software.

## REFERENCES

[1] R. A. FAIRTHORNE: Empirical hyperbolic distribution (Bradford–Zipf–Mandelbrot) for bibliometric description and prediction. *J. Docum.* 1969, **25**, 319–343.

[2] A. C. CLARE, E. M. COOK and M. F. LYNCH: The identification of variable-length, equifrequent character strings in a natural language data base. *Computer J.* 1972, **15**, 259–262.