Quasi-Equifrequent Group Generation and Evaluation

E. J. Yannakoudakis and A. K. P. Wu

Postgraduate School of Computer Science, University of Bradford, Bradford, West Yorkshire BD7 1DP, UK

The frequency of occurrence and other statistical results derived thereupon from unique items in collections such as letters, words and records has recently formed the basis for the design of optimal information structures. A fundamental theorem of information science states that the information representing capability of a set of symbols is maximized when the probability of occurrence of any symbol in the available set becomes the same. Equifrequency however is very rarely encountered in real applications and it is in many cases desirable to have sets of items or symbols which are equifrequent within a certain deviation i.e. quasi-equifrequent. This paper presents an algorithm for generating equifrequent sets and evaluates and compares the efficiency and accuracy of (a) the entropy and (b) the variance concepts for measuring the degree of quasi-equifrequency in a set. Tests are carried out on the occurrence of the letters A–Z (out of a total of 7,908,100 letters) and on 244 unique subfields (out of a total of 1,113,447 bibliographic record subfields) and an absolutely equifrequent set of subfields is presented.

INTRODUCTION

Basic works on communication and information theory provide simple generalizations regarding efficiency in transmission and storage of information. Although the mathematical theory of communication appeared nearly thirty years ago,[1] it is only recently that an attempt was made to reinterpret the theory and to investigate its implications for information science.[2] Lynch[2] reasserts that Shannon's[1] first statement about the equifrequency of symbols and therefore about rectangular frequency distributions, stands as the ideal.

The aim of this paper is to investigate methods of equifrequent set generation and in particular to compare the efficiency and accuracy of the use of (a) the entropy and (b) the variance concepts for measuring the degree of quasi-equifrequency among a set of groups of symbols. Since absolutely equifrequent groups are rarely encountered in real applications, the term 'quasi-equifrequent' is used here to describe all intermediary arrangements prior to the one that can be characterized as optimal.

For testing purposes two different sets of 'symbols' are used: (1) The 26 letters of the English alphabet and their frequencies as calculated by Yannakoudakis out of a total of 7,908,100 letters.[3] (2) The frequency of occurrence of 244 different MARC (Machine Readable Catalogue) record subfields. A BNB (British National Bibliography)[8] file of 31,369 records was used and the frequency of occurrence of all unique fields and subfields was calculated out of a total of 1,113,447 subfields on the lines described by Ayres and Yannakoudakis.[4]

It is by no means unrealistic to consider a MARC subfield as a symbol since it is the basic element from which records are built, in more or less the same manner as words are made from letters of an alphabet. It is believed that this assumption will lead to the design of optimal record structures and hence efficient file structures.

PROBLEM FORMULATION

'Grouping' is defined here as the mapping of the alphabet

$$\mathbf{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_M\}$$

(where $M = 26$ for the letters and $M = 244$ for the MARC subfields) onto another alphabet

$$\mathbf{G} = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_N\}$$

such that the groups

$$\mathbf{g}_1 = \{\alpha_i, \ldots, \alpha_j\}$$
$$\mathbf{g}_2 = \{\alpha_k, \ldots, \alpha_l\}$$
$$\mathbf{g}_N = \{\alpha_m, \ldots, \alpha_M\}$$

are equiprobable within an acceptable deviation such that

$$c(\mathbf{g}_i) = c(\mathbf{g}_j) + \delta_{ij} \qquad (1)$$

where $c(\mathbf{g}_i)$, $c(\mathbf{g}_j)$ represent the cumulative frequencies of groups $\mathbf{g}_i$ and $\mathbf{g}_j$ respectively, and ideally, δ_{ij} is at a minimum. The problem then is how to calculate the degree of quasi-equifrequency among the members of **G** so that comparisons between alternative arrangements of all $\alpha_i \in \mathbf{G}$ can be made in order to choose the optimal. One criterion would be to minimize the variance of all $\mathbf{g}_i \in \mathbf{G}$, another to maximize the entropy of the distribution. Nugent and Vegh formulate the problem similarly but do not consider the use of entropy in their experiments.[5]

The variance method utilizes the basic distributional properties of the data set. When the items are arranged in groups, the variance of the distribution is at minimum if the groups are so arranged that their total frequencies are most evenly distributed. If we denote the mean frequency of the group set by $\bar{f}$ then the variance of the distribution is

$$\sigma^2 = 1/N \sum (c(\mathbf{g}_i) - \bar{f})^2 \qquad (2)$$

CCC–0010–4620/82/0025–0183 $02.50

Thus the variance method aims to minimize the function

$$\sum (c(\mathbf{g}_i) - \bar{f})^2 \tag{3}$$

The entropy method utilizes Shannon's expression

$$-H = \sum_{i=1}^{N} P(\mathbf{g}_i) \log_2 P(\mathbf{g}_i) \tag{4}$$

where $P(\mathbf{g}_i)$ is the probability of occurrence of group $\mathbf{g}_i$. If the groups are absolutely equifrequent then we have a maximum entropy

$$-H_{\max} = N[1/N \log_2 (1/N)] = \log_2 1/N \tag{5}$$

Therefore the relative entropy can be obtained as the fraction $-H/-H_{\max}$ or relative entropy

$$r = \frac{1}{\log_2 (1/N)} \sum_{i=1}^{N} P(\mathbf{g}_i) \log_2 P(\mathbf{g}_i) \tag{6}$$

Thus the entropy method aims to maximize $r\,(0 \leq r \leq 1)$.

Brack *et al.* used the relative entropy to measure the quasi-equifrequency of character strings (digrams, trigrams, tetragrams etc.) obtained from a number of bibliographic record files.[6] Although each measure has in the past been used in one application or another, a direct comparison of the efficiency of the two has not been carried out, and apart from Nugent and Vegh,[5] no detailed description of an algorithm to generate alternative quasi-equifrequent groups is available.

EXPERIMENTAL RESULTS

Given a set of M items in a collection, the algorithm to generate a number of quasi-equifrequent groups will require the following input: (a) identification of each item; (b) frequency of each item; (c) starting number of groups and (d) finishing number of groups. Regardless of the measure used the algorithm will terminate, optionally, upon the fulfilment of one of the following conditions, whichever appears first: (1) finishing number of groups is reached or (2) an absolutely equifrequent group set is generated.

Following a number of considerations and empirical investigations the algorithm was designed and implemented as described below. Although the method cannot guarantee an optimum solution, it will always converge to a near optimal solution. Total enumeration of all possible arrangements in order to choose the optimum would in any case be impractical due to the time constraint involved.

The algorithm

(1) Sort items by frequency of occurrence in descending order.

(2) Allocate appropriate storage areas/slots for cumulative frequencies and initialize to zero. (A slot thus becomes synonymous to a group).

(3) Perform the following steps until all frequencies in the sorted list have been exhausted: (i) Go through all storage slots and identify the slot with minimum cumulative frequency; (ii) Add next frequency in sorted list to the slot identified in step (i) above.

(4) Calculate the variance or relative entropy for the groups formed.

(5) Tentatively switch the items of each group with all items of every other group and calculate the resultant variance or entropy immediately after each switch. The best improvement, if any, subject to Eqns (3) or (6), from all switches made is then recorded and the actual switch then takes place. If an improvement is made then step (5) is repeated else step (6) is entered. With the aid of an algorithmic language step (5) becomes:

```
for all $g1 \in G$ do ⊄ $g1, g2$ are subsets within $G$ ⊄
for all ($g2 \in G$ and $g1 \neq g2$) do
for all $a \in g1$ do ⊄ $a, b$ are elements within $g1, g2$ ⊄
for all $b \in g2$ do
begin $gt1 := g1 - a + b$;
      $gt2 := g2 + a - b$;
      $Gt := G + gt1 - g1 + gt2 - g2$;
         $v := 1 - entropy\ (Gt)$; ⊄ or $v := variance(Gt)$
      if appropriate ⊄
if $v < v\ min$ then begin $v\ min := v$;
                             record $(a, b, g1, g2)$
                   end if
end od od od od;
$g1 := g1 - a + b$;      $g2 := g2 + a - b$;
```

(6) If variance becomes zero or entropy reaches one or the finishing number of groups is reached, then stop. Else increment the number of groups by one and return to step (2).

(**Note.** The switching of items is made subject to the following rules which help to improve the efficiency of the program: (a) items with equal frequencies are not switched; (b) items in single item groups are not

Table 1. (MARC fields) An arrangement into six absolutely equifrequent groups (A three digit code identifies the field and a letter the subfield)

```
690Z   015A   910A   043A   245B   260D   490A   250A   511A   945X
840A   410U   410V · 240R   710M   111A   710L   640A   910D   400H
111I   700V   400V   513A   610B   521A   110I   111F   410T   740P
710I   710H   610U   243S   900B   111C   810A   911C   910J
TOTAL  FREQUENCY =          196059

690A   001-   100A   245D   245E   500A   651X   690C   110A   021A
018A   690H   410A   245F   700C   900U   080A   410C   250D   400A
400T   110E  ·690E   600C   610B   711A   711J   911J   240Q   710K
740Q   911I   610H   710G   410E   411V   810T   611J   611X   710V
TOTAL  FREQUENCY =          196059

692A   245A   300C   100H   700A   504A   503A   010A   710A   900X
017A   610A   600H   240A   100C   022A   690F   041B   110M   745A
640S   710E   690M   700T   910B   240P   911Z   740S   243A   740R
100D   710U   690K   910B   400C   745V   640Q   410H   243P   410G
610K
TOTAL  FREQUENCY =          196059

082B   082A   008-   690D   700H   650X   650Z   245G   945A   945Z
790A   690W   245H   710C   651Y   600X   710B   111J   110L   100E
410W   002-   900C   600E   240S   910F   110K   900F   240O   610Z
600D   110G   400U   110D   740V   700D   645X   610Y   710T   490W
610I   411J
TOTAL  FREQUENCY =          196059

260A   083A   350A   083B   300B   300F   910Z   440A   900Z   910C
021B   600A   041A   110C   610C   650Y   110B   111K   651Z   600F
900F   440L   640X   690G   910H   700F   110J   740A   911K   740O
710D   110H   645A   411A   610G   410D   245C   640Z   411U   910K
910I
TOTAL  FREQUENCY =          196059

650A   260B   260C   300A   050A   910X   900A   690I   651A   900H
690Y   690V   440V   610X   690X   910U   250C   518A   505A   910E
100F   600T   911A   911X   610E   610T   700F   711K   840W   700U
711I   710J   600Z   640R   711F   243R   611A   911F   611K   610J
640P
TOTAL  FREQUENCY =          196059
```

Table 2. (MARC fields) An arrangement into nine absolutely equifrequent groups

```
690Z  043A  651X  710A  790A  690U  110C  690X  900U  505A
250P  440W  700T  690G  911Z  610B  111E  740R  740P  710H
610U  900D  111C  810A
TOTAL FREQUENCY =        130706

690A  910A  910X  440A  900X  021A  018A  100C  700C  250C
110M  651Z  410W  110E  600C  610D  740S  911J  100D  710J
690K  243S  411V  810T  911C  910J
TOTAL FREQUENCY =        130706

692A  015A  050A  245E  690I  900Z  021B  610A  410U  910U
710B  410C  600F  900F  600T  400V  513A  711J  840W  410T
110G  610H  710G  640Q  611J  611X  710V
TOTAL FREQUENCY =        130706

082B  300C  300B  504A  500A  651A  900H  945X  840A  245F
022A  111K  110L  745A  002-  640X  700V  110K  110J  243A
7400  645A  110D  410E  700P  410H  610K  490W
TOTAL FREQUENCY =        130706

260A  001-  100H  700A  245B  010A  490A  690Y  041A  410A
610C  110B  111J  640S  710F  600E  911X  240P  900E  521A
711I  110H  710I  740V  645X  243P  610I  411J
TOTAL FREQUENCY =        130706

650A  008-  100A  245D  650Z  245G  910C  511A  600H  240A
410V  111A  518A  640A  100F  690I  910H  910G  700F  610Z
240Q  600Z  910B  400C  410P  410G  610Y  411U
TOTAL FREQUENCY =        130706

260B  350A  690D  700H  503A  260D  110A  017A  245H  710C
600X  710I  080A  910E  400T  900C  240S  700F  711K  2400
700U  710U  411A  640R  245C  710T  910K  910I
TOTAL FREQUENCY =        130706

083A  260C  300A  650X  900A  945A  945Z  600A  440V  651Y
650Y  710L  100E  400A  111I  610F  610T  740A  110I  600D
710K  911I  745V  610G  640Z  911E  610J
TOTAL FREQUENCY =        130706

082A  245A  083B  300E  910Z  690C  250A  690V  690H  610X
240R  690F  041B  910D  400H  911A  690E  910F  711A  911K
710D  740Q  400U  711E  243R  611A  611K  640P
TOTAL FREQUENCY =        130706
```

switched because this can only decrease the degree of quasi-equifrequency within the group. This decrease will be due to the fact that all single item groups will involve items of higher relative frequency than any other item of a multi-item group.)

Test results have proved that in approximately 90% of the cases the terminating condition embodied in step (5) is fulfilled in one pass. The other 10% of the cases involve less than 9 loops in step (5).

A number of programs were written to implement both methods and record statistical and other information which enabled comparative evaluation under the following main factors: (i) Accuracy of results, (ii) Time, (iii) Sensitivity. For our purpose it was considered appropriate to generate between 5 and 45 groups for all MARC subfields and between 4 and 20 groups for the letters A–Z. Experiments carried out proved that both methods give similar results in terms of the actual measure used in each case. This is particularly obvious between 4 and 9 groups as shown in Figs 1 and 3 and between 5 and 12 groups as shown in Figs 2 and 4. In actual fact absolutely equifrequent groupings were obtained in 6 and 9 groups for the MARC subfields with both methods and the distributions are presented in Tables 1 and 2.

Table 3. (Letters A–Z) An arrangement into nine quasi-equifrequent groups

LETTERS	TOTAL FREQUENCY
EZX	874703
AM	874268
NUV	883046
TPWQ	874703
IBK	877821
OH	887511
RC	885749
SYG	871610
LDFJ	879345

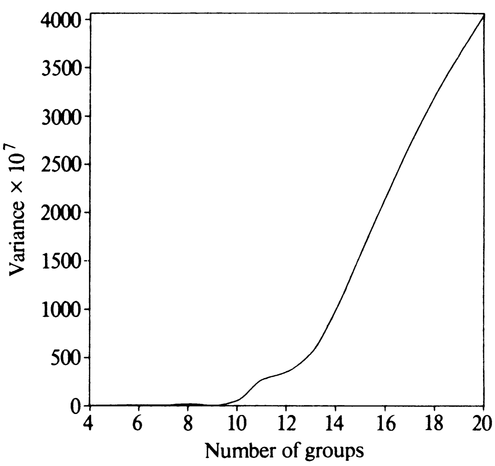

Figure 1. Variance vs number of groups (Letters A–Z).

It is interesting to note that with the frequencies of the letters A–Z no absolutely equifrequent groups could be achieved. An example of this is presented on Table 3 which contains the results for 9 groups.

Some interesting results were obtained when the time involved in each method was considered in our comparisons. Table 4 contains the results for the letters A–Z where it can be seen that as the number of groups increase the run-time of the variance method in comparison with the entropy method decreases from 30.17% in 4 groups to 21.01% in 20 groups. Table 5 contains the results from 5 up to 45 groups for the MARC subfields. Here the variance proves to be, on average, 33% more efficient (in terms of run-time) than the entropy. This can be explained by the fact that the time of fixed overheads (e.g. switching of items) becomes significant in the

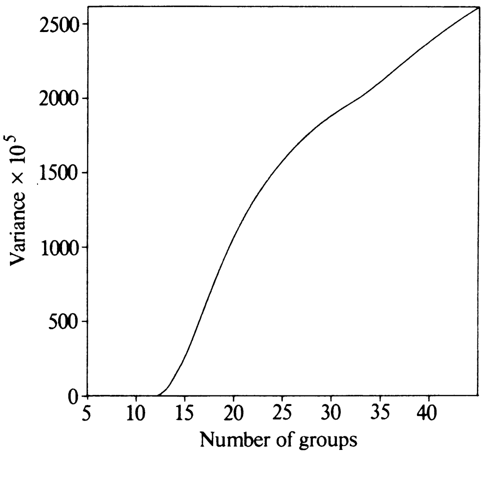

Figure 2. Variance vs number of groups (MARC fields).

© Heyden & Son Ltd, 1982

Table 4. Time used by relative entropy and variance methods (letters A–Z)

No. of groups	Entropy (s.)	Variance (s.)	Diff. (s.)
4	2.183	1.677	0.506
5	2.500	1.923	0.577
6	2.765	2.135	0.630
7	3.015	2.340	0.675
8	3.327	2.605	0.722
9	3.652	2.887	0.765
10	3.777	3.023	0.754
11	3.700	2.905	0.795
12	3.675	2.888	0.787
13	2.992	2.423	0.569
14	2.745	2.195	0.550
15	2.960	2.375	0.585
16	2.950	2.380	0.570
17	2.715	2.215	0.500
18	2.700	2.220	0.480
19	2.870	2.370	0.500
20	2.684	2.218	0.466

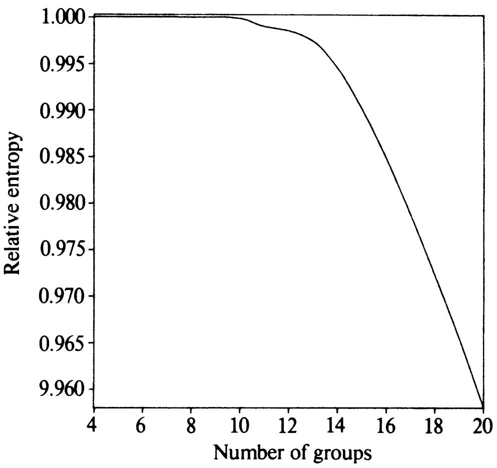

Figure 3. Entropy vs number of groups (Letters A–Z).

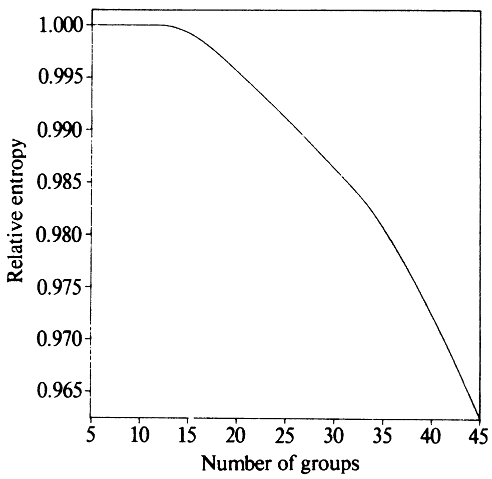

Figure 4. Entropy vs number of groups (MARC fields).

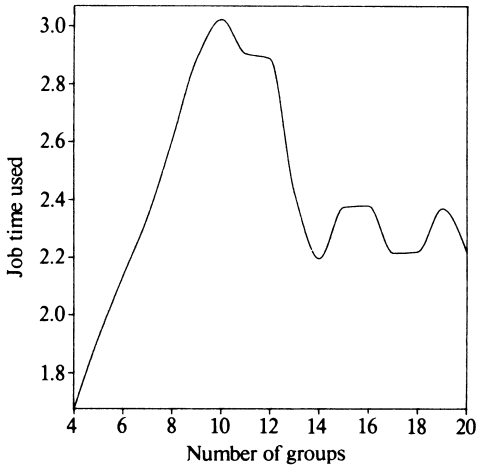

Figure 5. Time vs number of groups (Letters A–Z).

Table 5. Time used by relative entropy and variance methods (MARC fields)

No. of groups	Entropy (s.)	Variance (s.)	Diff. (s.)
5	148.53	111.95	36.58
6	155.64	116.96	38.68
7	161.59	121.02	40.57
8	165.25	124.21	41.04
9	169.49	127.02	42.47
10	172.30	129.62	42.68
11	176.56	131.79	44.77
12	179.26	133.80	45.46
13	179.12	134.07	45.05
14	181.35	136.00	45.35
15	181.19	136.10	45.09
16	184.23	138.14	46.09
17	186.87	139.95	46.92
18	189.33	141.88	47.45
19	191.16	143.67	47.49
20	193.92	145.48	48.44
21	196.35	146.60	49.75
22	198.51	148.20	50.31
23	198.43	148.42	50.01
24	199.00	150.02	48.98
25	199.90	150.03	49.87
26	202.36	151.52	50.84
27	203.37	152.87	50.50
28	206.04	154.57	51.47
29	206.39	154.41	51.98
30	208.29	155.68	52.61
31	207.15	155.30	51.85
32	203.57	152.45	51.12
33	188.68	142.35	46.33
34	191.35	143.76	47.59
35	193.21	145.71	47.50
36	195.52	146.73	48.79
37	193.30	145.36	47.94
38	195.29	146.93	48.36
39	197.24	148.24	49.00
40	195.34	146.99	48.35
41	190.78	143.88	46.90
42	190.46	143.92	46.54
43	192.75	145.67	47.08
44	195.46	146.79	48.67
45	196.89	148.24	48.65

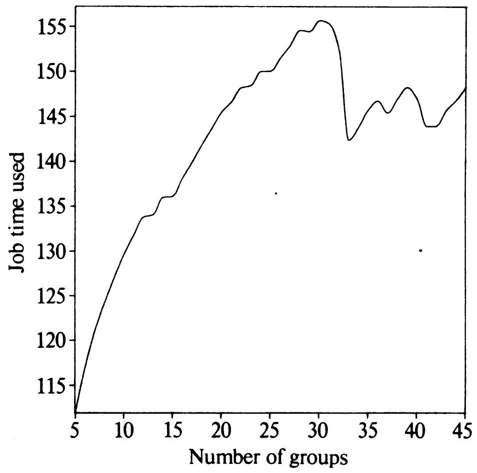

Figure 6. Time vs number of groups (MARC fields).

calculation of the overall time when the size of the collection of items is small. However, when the collection is large, the time of fixed overheads becomes negligible compared with the time taken for the other functions performed.

Figures 5 and 6 show a graphical representation of the time indicated on Tables 4 and 5, respectively, for the variance method, the pattern of which was found to be very similar to the entropy method. In both cases the time increases rapidly then decreases rapidly and finally levels off in a fluctuating pattern. We can explain this as follows: as the number of groups increases, the number of tentative switches among groups increases accordingly. However, this process reaches a turning point (see between 8 to 10 groups in Fig. 5 and between 30 to 32 groups in Fig. 6) where, as the number of groups continues to increase, the number of single item groups increases and this involves less tentative switches between individual groups (i.e. single item groups), the latter being a rule of the algorithm. Therefore the time taken for an arrangement decreases accordingly.

The sensitivity of each method was then studied in terms of the variation of each measure from one tentative switch to the next and from one arrangement to the other. To clarify the concept 'sensitivity', in its present context, let x_n be the measure used (either variance or entropy) for arrangement n and x_{n+1} be the measure of the following arrangement. Then the difference becomes much smaller in the case of the entropy as its value approaches 1 than in the case of the variance. The entropy is thus characterized as being more insensitive since it successively becomes more and more difficult to choose the best among a number of arrangements produced. The results, therefore, in view of the fact that the final groupings produced by both methods are similar, clearly indicate the superiority of the variance method in terms of speed, flexibility and reliability.

AREAS OF APPLICATION

It is hoped that the results presented herewith will be of value to communication engineers and information scientists working towards efficient transmission and communication. The variety generator seeks to reflect the microstructure of data elements in their description for storage and search, and takes advantage of the consistency of statistical characteristics of data elements in homogeneous data bases.[2] It is believed that the quasi-equifrequent algorithm can serve as a useful tool for analysing these data elements.

Research into coding for optimal record control as presented by Yannakoudakis *et al.* will be able to utilize the present results in order to generate codes for record identification.[7] This could be achieved by assigning a unique symbol to each of the letter sets of Table 3 which will then be used in the code upon the occurrence of any of the letters in a specific set. For example, given the following assignments:

EZX	0
NUV	2
TPWQ	3
IBK	4
LDFJ	8

The record title EQUIFREQUENT CODING will produce a five digit code 03248.

A fairly recent approach to the optimal file design has been to consider the statistical information of the items concerned and this has in all cases been their frequency of occurrence. If, however, this is supplemented by the frequency of access and particularly co-access it is believed that the use of the quasi-equifrequency generation algorithm will partition the items in an optimal arrangement and hence enable optimal placement on storage devices such as magnetic discs and other mass storage devices. Further research on this methodology is at present being carried out at the Computer Centre of this University.

REFERENCES

1. C. E. Shannon, A Mathematical Theory of Communication, *Bell Syst. Tech. J.* **27**, pp. 379–423, 623–656 (1948).
2. M. F. Lynch, Variety Generation—A Reinterpretation of Shannon's Mathematical Theory of Communication, and its Implications for Information Science, *J. Am. Soc. Inf. Sci.,* pp. 19–25 (1977).
3. E. J. Yannakoudakis, Towards a Universal Record Identification and Retrieval Scheme, *J. Informatics,* **3** (No. 1), pp. 7–11 (1979).
4. F. H. Ayres and E. J. Yannakoudakis, The Bibliographic Record: An Analysis of the Size of its Constituent Parts, *Program* **13** (No. 3), pp. 127–142 (1979).
5. W. R. Nugent and A. Vegh, Automatic Word Coding Techniques for Computer Language Processing. Vol. 1. Rome Air Development Centre, RADC-TDR-62-13 (1962).
6. E. V. Brack, D. Cooper and M. F. Lynch, The Stability of Symbol Sets Produced by Variety Generation from Bibliographic Data, *Program,* **2** (No. 2), pp. 64–77 (1978).
7. E. J. Yannakoudakis, F. H. Ayres and J. A. W. Huggill, Character Coding for Bibliographical Record Control, *The Comput. J.* **23** (No. 1), pp. 53–60 (1980).
8. British Library, UK MARC Manual, First Standard Edition (1975).

Received November 1980