FTD-MT-24-1456-71

FOREIGN TECHNOLOGY DIVISION

SOME RESULTS OF STATISTICAL STUDIES OF A DESCRIPTOR LANGUAGE

by

V. K. Vakhabov

Approved for public release;
Distribution unlimited.

20

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Foreign Technology Division Air Force Systems Command U. S. Air Force	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE

SOME RESULTS OF STATISTICAL STUDIES OF A DESCRIPTOR LANGUAGE

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Translation

5. AUTHOR(S) (First name, middle initial, last name)

Vakhabov, V. K.

6. REPORT DATE 1970	7a. TOTAL NO. OF PAGES 17	7b. NO. OF REFS 14
8a. CONTRACT OR GRANT NO. b. PROJECT NO. c. DIA Task Nos. T71-05-09 and d. T71-05-13	9a. ORIGINATOR'S REPORT NUMBER(S) FTD-MT-24-1456-71 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AP0154694	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Foreign Technology Division Wright-Patterson AFB, Ohio

13. ABSTRACT

The statistical regularities of a descriptor language are investigated. The distribution patterns of descriptor occurrence in document and request search patterns are examined. The distributions are shown to be close to the Zipf-Mandelbrot law. The statistical proximity of the frequency lists of descriptors from document and request files ~~has been~~ measured. The dependence ~~has been~~ evaluated of descriptor frequency on such characteristics as the number of keywords in the equivalence class of the given desciptor, the number of broader term references to the given descriptor from its underlying ones, and the average frequency of occurrence of the keywords. The validity of the findings ~~has been~~ estimated.

UNCLASSIFIED
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Information Storage and Retrieval						
Computer Language						
Natural Language						

FTD-MT-24-1456-71

EDITED MACHINE TRANSLATION

SOME RESULTS OF STATISTICAL STUDIES OF A DESCRIPTOR LANGUAGE

By: V. K. Vakhabov

English pages: 17

Source: Nauchno-Tekhnicheskaya Informatsiya. Seriya 2. Informatsionnyye Protsessy i Sistemy (Scientific and Technical Information. Series 2. Information Processes and Systems). No. 4, 1970, pp. 27-31.

This document is a SYSTRAN machine aided translation, post-edited for technical accuracy by Charles T. Ostertag.

UR/0447-70-000-004

THIS TRANSLATION IS A RENDITION OF THE ORIGINAL FOREIGN TEXT WITHOUT ANY ANALYTICAL OR EDITORIAL COMMENT. STATEMENTS OR THEORIES ADVOCATED OR IMPLIED ARE THOSE OF THE SOURCE AND DO NOT NECESSARILY REFLECT THE POSITION OR OPINION OF THE FOREIGN TECHNOLOGY DIVISION.	PREPARED BY: TRANSLATION DIVISION FOREIGN TECHNOLOGY DIVISION WP-AFB, OHIO.

FTD-MT-24-1456-71 Date 30 Oct 19 71

U. S. BOARD ON GEOGRAPHIC NAMES TRANSLITERATION SYSTEM

Block	Italic	Transliteration	Block	Italic	Transliteration
А а	*А а*	A, a	Р р	*Р р*	R, r
Б б	*Б б*	B, b	С с	*С с*	S, s
В в	*В в*	V, v	Т т	*Т т*	T, t
Г г	*Г г*	G, g	У у	*У у*	U, u
Д д	*Д д*	D, d	Ф ф	*Ф ф*	F, f
Е е	*Е е*	Ye, ye; E, e*	Х х	*Х х*	Kh, kh
Ж ж	*Ж ж*	Zh, zh	Ц ц	*Ц ц*	Ts, ts
З з	*З з*	Z, z	Ч ч	*Ч ч*	Ch, ch
И и	*И и*	I, i	Ш ш	*Ш ш*	Sh, sh
Й й	*Й й*	Y, y	Щ щ	*Щ щ*	Shch, shch
К к	*К к*	K, k	Ъ ъ	*Ъ ъ*	"
Л л	*Л л*	L, l	Ы ы	*Ы ы*	Y, y
М м	*М м*	M, m	Ь ь	*Ь ь*	'
Н н	*Н н*	N, n	Э э	*Э э*	E, e
О о	*О о*	O, o	Ю ю	*Ю ю*	Yu, yu
П п	*П п*	P, p	Я я	*Я я*	Ya, ya

* ye initially, after vowels, and after ъ, ь; e elsewhere. When written as ё in Russian, transliterate as yё or ё. The use of diacritical marks is preferred, but such marks may be omitted when expediency dictates.

SOME RESULTS OF STATISTICAL STUDIES OF A DESCRIPTOR LANGUAGE

V. K. Vakhabov

The application of statistical methods for the investigation of artificial informational languages can be promising in selecting rational methods for the organization of storage in machine retrieval. Thus in [1] some statistical adherences to the law during descriptor retrieval are examined theoretically, and an example of their utilization for the analysis of the inverse method of search is analyzed. As the basis for the cited adherences to the law two hypotheses were posed: each descriptor of the information-retrieval language has one and the same probability of appearance in the search patterns of documents and inquiries; the probability of the use of descriptors in the search patterns of documents and inquiries is subordinated to the laws of Zipf and Mandelbrot.

Below the results of an experimental check of these assumptions and adherences to the law deduced from them are described. At the same time the influence of the number of keywords in the class of equivalency, the number of hierarchical references to the descriptor, and the frequency of occurrence of keywords on the frequency of use of the descriptors have also been investigated.

The Zipf law of hyperbolic distribution, to which word frequency in the texts of natural language is subordinated, is used extensively in statistical linguistics. It proves to be suitable for the description of the most diverse phenomena, for example, it is steadily fulfilled in natural languages for the distribution of word forms and lexicon — the totality of the word forms of one changed word [2].

It has been shown in [3] and [4] that the Zipf law also extends to artificial descriptor informational languages relative to the distribution of descriptors in the search patterns of documents.

In works [5] and [6] the generality of the Zipf law with the Lotka distributive law for the productivity of scientists and with the Bradford law of the scattering of information in journals is shown, and in a number of other cases emerging beyond the frames of statistical linguistics. The universal nature of the Zipf law permits the assumption of its deducibility from the properties of any common model lying at the basis for all these phenomena.

One of such models has been constructed by Mandelbrot [7, 8] on the basis of the representations of the bond theory. Words are considered as information being transmitted by a certain communication channel. Words are formed with the help of an ergodic source of characters and are separated by spaces. With the optimum coding of words by characters, i.e., with the agreement of code with channel, when its maximum informational capacity is ensured, the a priori probability of the appearances of the i-th word P_i is determined by formula

$$p_i=\frac{K}{(B+i)^{\gamma}}, 1<i<n, \qquad (1)$$

where n — the volume of the dictionary;

K, B, γ — constants, whereupon $B \ll n$, and $1 \leq \gamma \leq 1.2$.

At B = 0 and γ = 1 the Mandelbrot formula (1) coincides with the empirical formula of Zipf.

The Mandelbrot model explains the asymptotic nature of the Zipf law. The value of γ in formula (1) is interpreted as the measure of the unbalance of the frequencies of words and characters. From the positions of the Mandelbrot model it is possible to explain the significant increase of γ (up to 1.6) in the speech of children and mental patients, which apparently differs substantially from the condition of optimum coding. The Mandelbrot law is in a better agreement with the empirical distributions of words in texts than the Zipf law, especially in the area of frequent words, where $(i \leq 15)$ the Zipf formula is generally uncertain.

The hypothesis which lies as the basis for the Mandelbrot model is disputed in [9]: here it is shown that the actual speech process is not determined by the property of optimum coding. Furthermore the Mandelbrot model is not universal.

A more common approach to the derivation of the Zipf law is given in work [5], where a model of the generation of texts by machine is examined. Here the concept of the complexity of the generation of signs and texts is introduced, during which the probability of the generation of text is a continuous decreasing function of its complexity. An inverse proportionality is postulated between the average complexity of text and its volume, and also the exponential dependence between the quantity of the signs of the alphabet and complexity. Under these conditions the appearance frequency of signs in text converges in probability to an expression analogous to the Zipf law. The advantage of the Schreider model is its greater generality than the Mandelbrot model.

Actually the concept of complexity is apparently natural for language. However, during a study, for example, of a real descriptor language the complexity which causes the frequency of descriptors cannot be interpreted by any one factor, but is defined as the generalized resulting characteristic of the random interaction of a number of factors (the number of keywords in the class of equivalency of the descriptor, the presence of the hierarchical bonds of this descriptor with others, etc.).

In examining speech or text in any language as a complex random process depending on many random and randomly interacting factors, it is possible to assume that this process is described by Bernoulli's well-known model. The derivation of the Zipf law with the utilization of a diagram of the independent trials of Bernouli was made by Andryushchenko [10], who emphasized that the probability of the appearance of the N-th word of text r times is determined from the formula of negative binomial distribution [11]:

$$p_r = C^{N-r}_{-r} p^r (-q)^{N-r}, \qquad (2)$$

where p – the probability of the appearance of the given word; q = 1 - p.

Negative binomial distribution in the case of a finite dictionary is reduced to logarithmic distribution with the absent zero class [11], to which the frequencies correspond.

$$-\frac{1}{\log(1-q)}\left\{q, \frac{q^2}{2}, \frac{q^3}{3}, \ldots, \frac{q^i}{i}\right\}. \qquad (3)$$

In a large dictionary, when p is small and q → 1, a harmonic series is obtained which corresponds to the Zipf law. In [10] the systematic divergences of empirical distributions from the laws of Zipf and Mandelbrot are explained by the fact that in reality $q \neq 1$.

In [10] it is also shown that if we evaluate parameter γ in the Mandelbrot formula (1) from the correlation

$$\frac{1}{i^\gamma} = \frac{q^i}{i}, \qquad (4)$$

then γ proves to be the increasing function from i. This agrees completely with emperical data of Frumkina [2] concerning the increase of γ with the increase in the volume of the dictionary.

Bernoulli's system at small p and q → 1 apparently is that generalized model which gives rise to the law of hyperbolic distribution. This system is apparently applicable for the derivation of the distributive law of descriptors in documents and requests.

EXPERIMENTAL CHECK OF THE LAW OF DISTRIBUTION OF DESCRIPTORS IN DOCUMENTS

The distribution of descriptors in the search patterns of documents has been checked experimentally on a group of 1600 documents on computer technology which was indexed by a dictionary which unites 1181 keywords into 578 classes of equivalency which correspond to the descriptors. Descriptors in the dictionary are partially ordered by the hierarchical relations "common - particular."

During the indexing of documents the following rule is observed: when in the search pattern of a document there is a descriptor having reference to a higher descriptor, the latter is also included in the search pattern. Therefore higher descriptors always have a frequency no less than a subordinate.

Into the search patterns of 1600 documents 502 descriptors entered 19,030 times, which corresponds to an average depth of indexing of 12 descriptors per document.

The chart of the frequency distribution of occurence of descriptors in the search patterns of documents is shown in Fig. 1, from which it is evident that the curve which corresponds to formula (1) of Mandelbrot describes somewhat better the real distribution, especially in the range of large and medium frequencies. Thus the chart testifies to the doubtless applicability of the laws of Zipf and Mandelbrot to the distribution of descriptors in the search patterns of documents.

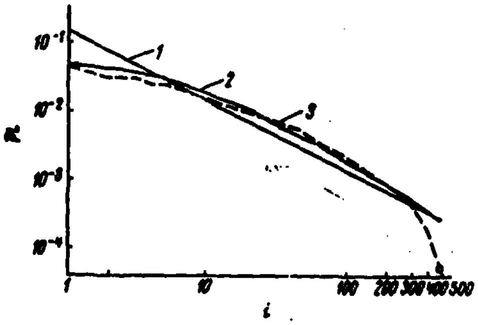

Fig. 1. The distribution of descriptors in the search patterns of documents: 1 - the Zipf law $p_i \frac{0.1472}{i}$; 2 - the Mandelbrot law $p_i = \frac{0,512}{(6+i)^{1.2}}$; 3 - the empirical distribution law of descriptors in documents.

The reliability of the results obtained can be evaluated by using the formula of the relative error of the measurement of frequency during the analysis of the sample [2]

$$\delta = \frac{Z_\rho}{\sqrt{Np}}, \qquad (5)$$

where N – the sample size; p – the frequency which should be determined according to the given sample; δ – the relative error of the measurement of frequency; ρ – the confidence coefficient, from which Z_ρ is determined [12].

At $\rho = 0.95$ it is possible to take $Z_\rho = 2$.

It is evident from formula (5) that the authenticity of the determination of frequency is lowered for low-frequency descriptors. Under the condition that relative error does not exceed the given value of δ_0 it is possible to compute the minimum probability P_{min} of the descriptor, for which

$$P_{min} = \frac{Z_\rho^2}{\delta_0^2 N}. \qquad (6)$$

Considering that the probability is determined by the Zipf law $p_i = K/i$, it is possible to find the number of the descriptor for which the condition $\delta \leq \delta_0$ is still satisfied:

$$i_{пр} = \frac{K\delta_0^2 N}{Z_\rho^2}. \qquad (7)$$

It is also possible to determine what fraction of sample size is occupied by the descriptors whose frequencies have been determined with the error $\delta \leq \delta_0$:

$$\Delta = K \sum_{i=1}^{i_{пр}} \frac{1}{i} = \frac{\sum_{i=1}^{i_{пр}} \frac{1}{i}}{\sum_{i=1}^{n} \frac{1}{i}} \approx \frac{\ln i_{пр} + C}{\ln n + C}. \qquad (8)$$

where n – the volume of dictionary; C = 0.5772.

Having taken $\delta = 0.3$ and taking into account that $N = 19{,}030$, $n = 502$, we obtain $i_{np} = 63$, $\Delta = 70\%$; i.e., 70% of samples from 19,030 descriptor-entries are made up of 63 descriptors whose frequencies have been determined with a relative error no higher than 0.3. Such an accuracy is completely sufficient for obtaining qualitative conclusions about the distribution law.

ANALYSIS OF FACTORS WHICH INFLUENCE THE DISTRIBUTION OF DESCRIPTORS IN DOCUMENTS

During research on the distribution of descriptors in documents a study was made of the influence of the number of keywords in the class of equivalency of the given descriptor on the frequency of entrance of the descriptor into the search samples of documents; the presence of references to the given descriptor from subordinate descriptors; the mean frequency of appearance of the keywords of the given descriptor in the search patterns of documents.

For evaluating the influence of each of these factors for every descriptor, besides the frequency of its entrance p into the search patterns of documents, the following were determined: t_{OCH} — the number of keywords in the class of equivalency of the descriptor and t_{Σ} — the number of keywords entering into the classes of equivalency of the given descriptor, and also all descriptors subordinate to it. Apart from these characteristics, for every descriptor $И = t_{\Sigma} - t_{\text{OCH}}$ — the number of keywords in the descriptors subordinate to that given was computed; $f_{\text{KC}} = \frac{p}{t_{\Sigma}}$ — the mean frequency of appearance of the keyword of the given descriptor in the search patterns of documents.

It should be noted that value $И$ characterizes the number of hierarchical references to the given descriptor, and t_{Σ} — simultaneously both the first factors — the number of keywords in the classes of equivalency of the descriptor and the number of references to it.

Since all these factors have a random nature, for the quantitative evaluation of each of them for the frequency of entrance of descriptors it is advantageous to use the methods of correlation analysis. The degree of influence of each of these factors v on frequency can be evaluated based on the value of the correlation factor, calculated in the method of the moments of products [13].

$$r=\frac{\sum_{i=1}^{n} v_i p_i}{\sqrt{\sum_{i=1}^{n} p_i^2 \sum_{i=1}^{n} v_i^2}}, \qquad (9)$$

where v_i, p_i – the divergence of the investigated factor v and frequency p from their mathematical expectations; n – the number of descriptors in the dictionary.

However, it is more convenient to use the coefficient of grade correlation which practically does not depend on the measuring error of the investigated factors. When using this method [13] signs being compared are ranked by a decrease (increase) of values, whereupon they operate no longer by values, but by ranks. The coefficient of grade correlation is calculated according to the formula

$$\rho=1-\frac{6\sum_{i=1}^{n} d_i^2}{n^3-n}, \qquad (10)$$

where n – the value of the greatest rank (in this case – the volume of the descriptor dictionary); d_i – the difference between the ranks of frequency and investigated factor of the i-th descriptor.

If the measured factors have identical values in a series of points, for the computation of the coefficient of rank correlation a generalized formula is used which is derived taking into account the association of ranks [13]

$$\rho=\frac{\frac{n^3-n}{6}-(T_p+T_v)-\sum_{i=1}^{n} d_i^2}{\sqrt{\left(\frac{n^3-n}{6}-2T_p\right)\left(\frac{n^3-n}{6}-2T_v\right)}}, \qquad (11)$$

where $T_v=\sum_{j=1}^{l}\frac{t_j^3-t_j}{12}$ l – the number of repeated values of the factor; t_j – the number of ranks, united by the j value of factor $(1<j<l)$.

The correlation factor in the method of the moments of products is connected with the coefficient of rank correlation by the expression

$$r=2\sin\left(\frac{\pi\rho}{6}\right). \qquad (12)$$

The results of the computations of the correlation factors of frequency and enumerated factors are given in Table 1.

Table 1. The degree of influence of various factors on the frequency of occurrence of descriptors in the search patterns of documents.

Наименование фактора (1)	Коэффициент ранговой корреляции (2)	(3) Коэффициент корреляции по способу моментов произведений
$t_{осн}$	+0,660	+0,676
t_{Σ}	+0,724	+0,738
$H=t_{\Sigma}-t_{осн}$	[+0,460	+0,475
$f_{кс}=\frac{\omega}{t_{\Sigma}}$	+0,866	+0,856

KEY: (1) Name of factor; (2) Coefficient of rank correlation; (3) Correlation factor in the method of the moments of products.

Figure 2 gives the regression curves constructed for the indicated characteristics of these factors on various intervals of frequency of the descriptors. A comparison of the correlation factors and regression curves permits the following conclusions to be made.

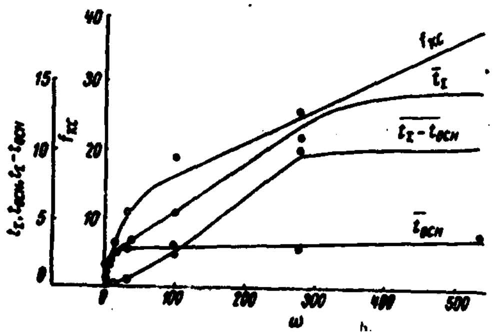

Fig. 2. The regression curves of factors which determine the frequency of descriptors in documents.

Each of three factors – the average frequency of appearance of keywords in the search patterns of documents, the number of keywords in the class of equivalency of the descriptor, the number of references to the given descriptor on the part of subordinates – substantially influences the frequency of entrance of a descriptor into the search patterns of documents.

The number of hierarchical bonds of a descriptor, measured by the number of keywords in the descriptors subordinate to that given, influences the frequency of entrance somewhat less than the remaining factors. This is explained primarily by the fact that in the descriptor dictionary in question the base lines of communications are encountered comparatively rarely. Thus from 502 descriptors only 60 have the references of subordinate descriptors. However, the analysis of regression curves $t_{\Sigma} t_{OCH}$ and $t_{\Sigma} - t_{OCH}$ (see Fig. 2) in the area of the frequent descriptors shows, the majority of which have references to subordinates, the influence of the number of references on the frequency of a descriptor proves to be more significant than the number of keywords in the class of equivalency of the descriptor. The rare descriptors whose area composes the greater part of the dictionary barely have references, and the influence of factor $H = t_{\Sigma} - t_{OCH}$ in this area is negligibly small.

CHECK OF THE LAW OF DISTRIBUTION OF DESCRIPTORS IN REQUESTS

The distribution of descriptors in requests has also been studied experimentally. Two files of requests on computer technology were examined: the artificial request, made up for the check of the quality of work of the IRS [Information Retrieval System], and real demands of specialists, assembled for input into the system for the selective distribution of information.

In the first group there were 124 requests with a total of 546 inputs of descriptors, in the second group – 113 requests having 355 descriptor-entries.

The direct check of the distribution law on such small samples is uncertain. The check of the applicability of the Zipf law for the distribution of descriptors in requests was accomplished by the indirect method, connected with the experimental check of some functionals reflecting the statistical structure of text and calculated in the execution of the Zipf law. The number of such functionals [14] include n(N) – the number of various words in a text out of N words, and n(m, N) – the number of words which are encountered m times in a text out of N words.

In Table 2 the experimental values of these functionals are compared with their mathematical expectations, calculated using formulas [14]

$$Mn(m, N)=\frac{NK}{m(m-1)}e^{-\frac{NK}{n}}\sum_{j=0}^{m-2}\frac{1}{j!}\left(\frac{NK}{n}\right)^{j}+\frac{1}{2}\frac{\left(\frac{(NK)^{m}}{n}\right)}{m!}e^{-\frac{NK}{n}}$$

at $m>2.$

$$Mn(1, N)=-NKEi\left(-\frac{NK}{n}\right)+\frac{1}{2}\frac{NK}{n}e^{-\frac{NK}{n}};$$

$$Mn(N)=n(1-e^{-\frac{NK}{n}})+Mn(1, N)-\frac{1}{2}e^{-\frac{NK}{n}}, \qquad (13)$$

which are valid in the fulfillment of the Zipf law. In formulas (13) n – the number of descriptors in the dictionary; K = (ln n + C) – constant in the formula of the Zipf law.

Table 2. The experimental and calculated values of the functionals bound with the distribution of descriptors in requests.

(1) Наименование функционала		n(N)	n(1,N)	n(2,N)	n(3,N)	n(4,N)	n(5,N)	n(6,N)	n(7,N)	n(8,N)	n(9,N)
(2) Запросы 1-й группы $N = 546$	(3) расчет	187	113	34,3	13,3	7	4	3	2	1,5	1
	(4) эксперимент	146	96	36	16	13	6	4	3	1	2
(5) Запросы 2-й группы $N = 356$	(3) расчет	143	94	24	9	4,4	3	1,8	1,3	1	0,7
	(4) эксперимент	114	53	26	12	5	4	1	5	1	0

KEY: (1) Name of functional; (2) Request of the 1st group; (3) calculation; (4) experiment; (5) Request of the 2nd group.

The satisfactory agreement of the experimental and calculated values of the functionals testifies to the applicability of the Zipf law to the distribution of descriptors in requests.

The appearance frequency of every descriptor in the requests apparently depends in the same manner on the quantity of keywords in the class of equivalency of the given descriptor and the medium frequency of appearance of the keywords of descriptors in request.

However, the limited amount of experimental material on requests does not permit a reliable quantitative evaluation of the influence of these factors on the frequency of appearance of descriptors in requests.

COMPARISON OF THE LAWS OF DISTRIBUTION OF DESCRIPTORS IN DOCUMENTS AND REQUESTS

The hypothesis concerning the agreement of the laws of distribution of descriptors in documents and requests in essence is equivalent to the assumption about the statistical proximity of frequency descriptor dictionaries for documents and requests.

The degree of statistical proximity of frequency dictionaries also can be established by the method of rank correlation.

The computation of the correlation factors for both groups of requests gives the following results:

experimental requests;

$\rho = +0{,}862$ $r = +0{,}870$

real requests

$\rho = +0{,}69$ $r = +0{,}705$

The values of the correlation factors and histograms, constructed on both series of requests (Fig. 3), testify to the significant statistical proximity of frequency dictionaries of the descriptors of documents and requests concerning the truth of the accepted hypothesis.

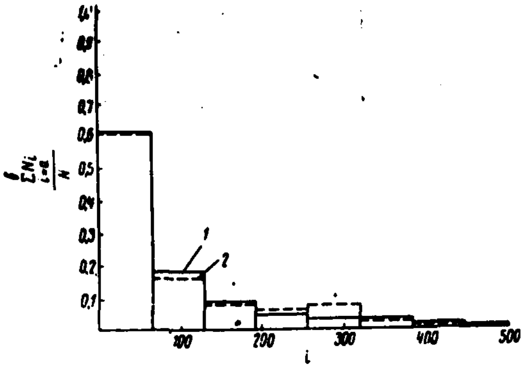

Fig. 3. Distribution of histograms for descriptors in requests: 1 — experimental requests; 2 — real requests.

The degree of divergence of the frequency dictionary of real requests from the dictionary of documents is greater than the dictionary of experimental requests. An analysis establishes the reasons for these divergences. Thus the descriptor "microelectronics" in the dictionary of real requests is one of the most frequent and has a rank of 5; in the file of documents its frequency is low-rank 101. Therefore one of the reasons for the divergence of frequencies of descriptors in requests and documents is the fact that the fund content always unavoidably lags behind the requirement of the specialists.

EXPERIMENTAL CHECK OF CONFORMITIES OF THE LAWS OF DISTRIBUTION OF DESCRIPTORS INSIDE THE REQUEST

After the analysis of t e hypotheses concerning the applicability of the Zipf law to the distribution of descriptors in the retrieval patterns of documents and requests, and also on the proximity of the given distributions between themselves on the same files of requests and documents, checks were made of the values of the functionals derived from these hypotheses which reflect the conformities of the distribution of descriptors inside the request [1].

One of such functionals is the mathematical expectation of a number (or a rank based on a frequency dictionary) Mi(j, z) of the descriptor which occupies the j place in the requests from z descriptors. It is assumed that the descriptors in the requests are arranged in increasing rank in the frequency dictionary which is constructed on the documents.

Another such functional is the mathematical expectation of a number of occurrences MN(j, z) of the j descriptor in the request from z descriptors in the retrieval patterns of the documents.

Under the indicated assumptions these values are determined from the formulas

$$Mi\,(j,\,z) \approx \sum_{i=j}^{n-z+j} zKC_{z-1}^{j-1}(K\ln Ai)^{j-1}(1-K\ln Ai)^{z-j}$$

$$MN\,(j,\,z) \approx \sum_{i=j}^{n-z+j} zNC_{z-1}^{j-1}\frac{K^{s}}{i^{s}}(K\ln A_{i})^{j-1}(1-K\ln Ai)^{z-j}, \qquad (14)$$

where n – the number of descriptors in the dictionary; K = (ln n + C)$^{-1}$ – constant in the formula of the Zipf law; C = 0.5772 – Euler constant, where upon A = e^{C}; N = fs – the total number of occurrences of descriptors in the retrieval patterns of documents, here f – the average depth of indexing of documents; S – the number of documents in the file.

For the experimental determination of the value of these functionals all the requests were divided into groups. Each group included requests containing an identical number of descriptors. The latter were ranked according to the frequency dictionary, constructed on the retrieval patterns of documents. Then in every group of requests for all descriptors standing at the j-th place $|1<j<z|$, calculations were made of the mean values of ranks and frequencies of occurrence of these descriptors in the search patterns of documents and they were compared with the calculated values of functionals obtained from formulas (14). From Figs. 4 and 5, in which the results of the comparisons are shown, it is evident that the values of mathematical expectations obtained from formulas (14) have an identical order to the mean values found experimentally. Thus formulas (14) can be considered suitable for estimated calculations.

It should be noted that on the basis of the total series of 237 requests it is not possible to obtain greater accuracy in determining the mathematical expectation of the investigated functionals based on their mean values.

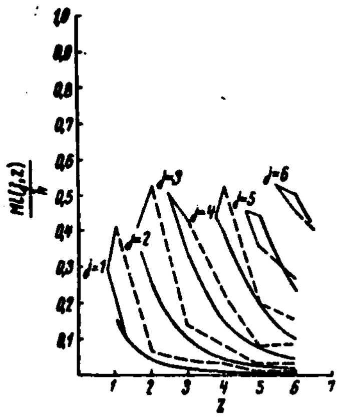

Fig. 4.

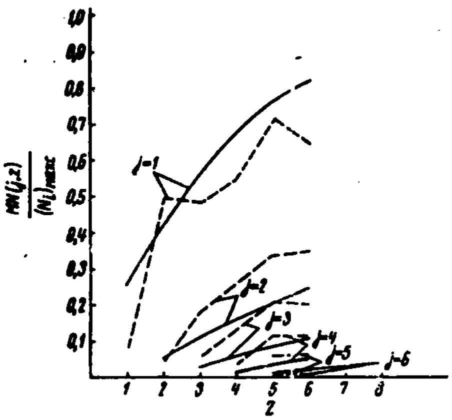

Fig. 5.

Fig. 4. Mathematical expectation of the number of the j-th descriptor of a request in the dictionary: ----- experimental value; ——— calculated value.

Fig. 5. Mathematical expectation of the number of occurrences of the j-th descriptor of the request in the retrieval patterns of documents: ----- experimental values; ——— calculated values.

Really the analysis of the variability of functionals showed that their mean square deviations in value are close to mathematical expectations. Therefore it is possible to find the relative error in the determination of mathematical expectation from its mean value, calculated in the group of t_r requests:

$$\delta=\frac{3\sigma}{M}=\frac{3\sigma}{M\sqrt{t_z}}, \qquad (15)$$

where M – mathematical expectation of the investigated functional; σ – the mean square deviation of this functional from the mathematical expectation; $\sigma_1=\frac{\sigma}{\sqrt{t_z}}$ – the mean square deviation of the mean value of the functional from the mathematical expectation.

Since $M\approx\sigma$,

$$\delta=\frac{3}{\sqrt{t_z}} \qquad (16)$$

The relative error, calculated using this formula for every group of requests depending on the number of descriptors in the request, is shown in Table 3: for the few groups of one- and seven-descriptor requests the mathematical expectations of functionals

Table 3. The error of the experimental evaluation of functionals, connected with the distribution of descriptors inside the request.

Number of descriptors in the request, z................	1	2	3	4	5	6	7
Number of requests in the group, t_z.................	4	29	85	72	37	18	2
Relative error							
$\delta = \frac{3}{\sqrt{t_z}} \cdot 100\%$	100%	55%	32%	35%	49%	78%	100%

can be obtained from their mean values with a relative error which reaches 100%. In the remaining cases this error can reach 30% and more, which fully explains the significant divergences of the experimental values of functionals from the calculated.

Nevertheless the experiment again confirms the validity of the accepted hypotheses, and also the suitability of the formulas derived in [1], for the estimated calculations of functionals.

CONCLUSIONS

On the basis of the experiments conducted it is possible to draw the following conclusions.

The appearance frequency of descriptors in the retrieval patterns of documents and requests is actually subordinated to the laws of Zipf and Mandelbrot, which are known for the distribution of words in natural languages.

The frequency of occurrence of descriptors in the retrieval patterns of documents is determined by the joint interaction of a number of random factors such as the mean frequency of use of the keywords of the descriptor; the number of keywords in the class of equivalency of the descriptor; the number of references to the given descriptor on the part of subordinates. Each of these factors influences the distribution of descriptors separately to a considerable degree which, nevertheless, is subordinated to the Zipf law. This confirms the assumption about the deductibility of the Zipf law from Bernoulli's model.

The frequency dictionaries of descriptors for documents and requests are statistically close to each other, which confirms the hypothesis about the agreement of the laws of distribution of descriptors in the retrieval patterns of documents and requests.

The formulas of mathematical expectation of rank and frequency of occurrence of the descriptors of a request in the retrieval patterns of documents, which were derived in work [1] on the basis of this hypothesis with the observance of the Zipf law, are completely suitable for the estimated calculations of these values.

The author thanks P. N. Sapozhnikov for the attentive review of the manuscript and valuable remarks.

BIBLIOGRAPHY

1. Вахабов В. К. Некоторые статистические закономерности при дескрипторном поиске информации. Доложено на международном симпозиуме стран — членов СЭВ. «Применение универсальных вычислительных машин в работе органов информации 14—17 июня 1967 г., М.
2. Фрумкина Р. М. Статистические методы изучения лексики. «Наука», 1964.
3. Bugnariu-Blaga D. Analiza aplicării metodei de indexare coordinata in domeniue documentării. «Studii si cercetări docum. si bibliol.» 1966,·8, No. 1, 3—9 (резюме).
4. Вахабов В. К., Михайлова А. А., Есилевская Л. М., Кутяева Т. С. Опыт создания информационно-поискового языка по вычислительной технике. Докл. III Всес. конференции по информационно-поисковым системам и автоматизированной обработке информации 19—22 декабря 1966 г., М.
5. Шрейдер Ю. А. О возможности теоретического вывода статистических закономерностей текста (к обоснованию закона Ципфа). Проблемы передачи информации. Т. III, вып. 1, 1967. 57—63
6. Козачков Л. С., Хурсин Л. А. Основное вероятностное распределение в системах информационных потоков. «НТИ», сер. 2, № 2, 3—12.
7. Бриллюэн Л. Наука и теория информации. М., 1960.
8. Мандельброт Б. О рекурентном кодировании, ограничивающем влиянии помех. В сб.: «Теория передачи сообщений». Под ред. В. И. Сифорова. М., 1957.
9. Miller C. A., Newman E. B. Tests of statistical explanation of the rankfrequency relation for words in written English. «Amer. J. Psychol.», 1958, 71, 1, 209--218.
10. Андрюшенко В. М. Частотный словарь-справочник немецкой обществено-политической лексики. Ч. I. Частотный словарь газеты «News Deutschland». В сб.: «Частотные словари и автоматическая переработка лингвистических текстов». Тезисы докладов II межвузовской конференции 4—6 апреля 1968 г. Минск, 1968, 25—27.
11. Кендалл М. Дж., Стьюарт А. Теория распределений. М., 1966, 185—187.
11. Вентцель Е. С. Теория вероятностей. М., 1962, 305—312.
13. Юл Дж. Э. Кендалл М. Дж. Теория статистики. Госстатиздат. М., 1960, 296—324.
14. Калинин В. М. Функционалы, связанные с распределением Пуассона и статистическим анализом текста. Тр. математического института им. В. А. Стеклова. вып. 79, 1965.

Submitted
11 Oct 1968