FTD-MT-24-1455-71

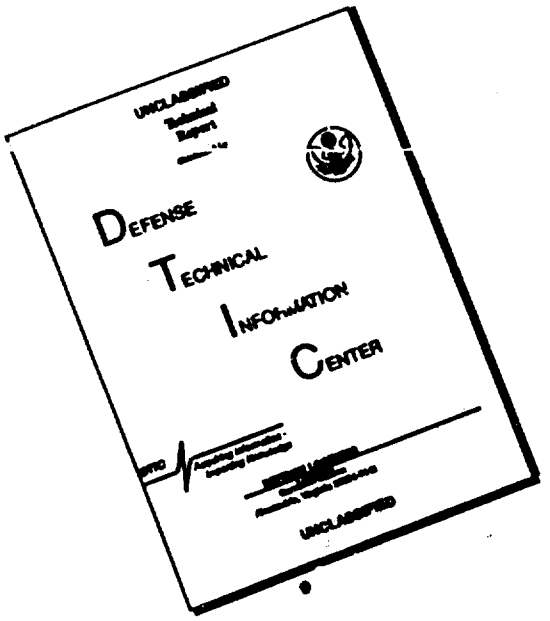
FOREIGN TECHNOLOGY DIVISION

RESULTS OF THE EXPERIMENTAL CHECK OF SOME METHODS OF AUTOMATIC CONSTRUCTION OF CLASSIFICATIONS AND INDEXING

by

S. A. Karasev

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

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
RESULTS OF THE EXPERIMENTAL CHECK OF SOME METHODS OF AUTOMATIC CONSTRUCTION OF CLASSIFICATIONS AND INDEXING

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

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

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
1970	31	7

8a. CONTRACT OR GRANT NO. b. PROJECT NO. c. DIA Task Nos. T71-05-09 and T71-05-13 d.	9a. ORIGINATOR'S REPORT NUMBER(S) FTD-MT-24-1455-71 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AP0154687

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

Two criteria are suggested which ensure complete algorithmization of automatic construction of a classification of terms. The criteria are based on statistical techniques. Two methods for automatic construction of classifications based on the factor analysis are described as in an automatic indexing method based on one of the classifications derived. The findings of a computer-based experiment are given.

DD FORM 1 NOV 65 1473

UNCLASSIFIED
Security Classification

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Algorithm Natural Language Information Storage and Retrieval						

FTD-MT- 24-1455-71

EDITED MACHINE TRANSLATION

RESULTS OF THE EXPERIMENTAL CHECK OF SOME METHODS OF AUTOMATIC CONSTRUCTION OF CLASSIFICATIONS AND INDEXING

By: S. A. Karasev

English pages: 31

Source: Nauchno-Tekhnicheskaya Informatsiya. Seriya 2. Informatsionnyye Protsessy i Sistemy (Scientific and Technical Information. Series 2. Information Processes and Systems), 1970, No. 11, pp. 5-12

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

UR/0447-70-000-011

Translator's note: On several occasions, symbols found in formulae and calculations appear to have been rendered incorrectly in the original document. They will be shown exactly as they appear in the original.

C

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.

FOLLOWING ARE THE CORRESPONDING RUSSIAN AND ENGLISH

DESIGNATIONS OF THE TRIGONOMETRIC FUNCTIONS

Russian	English
sin	sin
cos	cos
tg	tan
ctg	cot
sec	sec
cosec	csc
sh	sinh
ch	cosh
th	tanh
cth	coth
sch	sech
csch	csch
arc sin	$\sin^{-1}$
arc cos	$\cos^{-1}$
arc tg	$\tan^{-1}$
arc ctg	$\cot^{-1}$
arc sec	$\sec^{-1}$
arc cosec	$\csc^{-1}$
arc sh	$\sinh^{-1}$
arc ch	$\cosh^{-1}$
arc th	$\tanh^{-1}$
arc cth	$\coth^{-1}$
arc sch	sech^{-1}
arc csch	csch^{-1}
rot	curl
lg	log

INFORMATION ANALYSIS

RESULTS OF THE EXPERIMENTAL CHECK OF SOME METHODS OF AUTOMATIC CONSTRUCTION OF CLASSIFICATIONS AND INDEXING

S. A. Karasev

The subject of scientific-information activity is the scientific information which represents the logical information organized by means of comparison and classification of data. This circumstance requires the classification of the documents containing the scientific information [IPYa] (ИПЯ).

The main disadvantage of such IPYa is the shortage of vocabulary, which is connected with the impossibility to foresee the future changes in informational requirements. However, the noted deficiency can be eliminated if we provide the rapid reconstruction of classification in sufficiently small intervals. Such a reconstruction is no problem if the classifications are constructed by the machine method. Automatically constructed classifications can be suitable for the currently developing subject fields, for which logical classifications have not been developed, and be more rational than the traditional systems of classifications.

In the known algorithms (strategies) of the automatic construction of classifications statistical methods which digress completely from the semantic structure of the documents are used exclusively. However, statistical methods can be considerably reinforced by the application of syntactic analysis.

In spite of this, as the basis for this work the following hypothesis was set down: statistical methods of the automatic construction of classifications, based on improved criteria, will make it possible to obtain better results as compared with the known results obtained under the same assumptions.

In this case the classifications of the terms or classifications of documents can be constructed. For our purposes we use the following working definition: the system of classification is the totality, bound by the relationship of coordination or by the relationships of coordination and subordination of sets of terms, each of which satisfies a certain criterion of semantic similarity of the elements of a set with each other.

1. THE FUNDAMENTAL CHARACTERISTICS OF THE CLASSIFYING STRATEGIES

The problem of the construction of the classification of terms in the space of their signs, which in the hypothesis accepted by us are the frequencies of use of terms in the documents ($i = 1, \ldots N$), consists of the division of N-dimensional space into m areas G_1, G_2, ..., G_m, each of which is a class of terms.

If the system of classification has already been constructed, then the individual signs of each of the classified terms are replaced by the values of signs measured for the appropriate classes. With such a recoding less information is lost, and the more uniform the classes of terms in their properties.

This circumstance makes it possible to consider as the fundamental characteristics of the classifying strategies the criterion of the semantic similarity between terms, the determination of class, type, and the degree of freedom of strategy.

The view of natural language of scientific documents as a statistical phenomenon makes it possible to express quantitatively

the semantic similarity between terms on the basis of the application of a certain statistical function. In spite of the distinction in the analytical methods utilized for the representation of the criterion of semantic similarity, they all have a common logical base which consists of the following.

The bases for the terminology of any subject field are the meaning-bearing words which possess a naming (nominative) function. Meaning-bearing words separate objects of the external world, which are the subject of any scientific research. In scientific documents those objects are described which are found in a logical bond with each other. Therefore the words which designate the appropriate objects are found in the same logical bond with each other.

Therefore it is possible to claim that the words which are found in logical bond are frequently used in the same documents in various combinations with each other and much rarer (or they are in no way encountered) together with other words. Various factors of the semantic similarity of terms are also intended for the measurement of the degree of logical bond.

In examining natural language from these positions the frequency of repetition of terms is considered as the significant measure of their importance which must be considered in the statistical criterion of semantic similarity. Furthermore, the more extensive the volume of the concept of a term, the more frequently it is encountered in various combinations with other terms, the wider the frequency range of its use in documents, and therefore the greater its dispersion.

Of all known criteria of semantic similarity used in the automatic construction of classifications, a unique one, which considers both these moments, is the correlation factor. Furthermore, the correlation factor easily detects the interpretive type of dependence (positive or negative bond). This circumstance makes it possible to view the correlation factor as the most exact statistical measure of the semantic similarity of terms.

Thus, at the basis for the statistical representation of the criteria of semantic similarity lies the hypothesis that the degree of the object-logical bond of terms can be expressed by means of the statistical correlation of the frequencies of the combined occurrence of terms in documents.

The formal definition of class is not given in all the methods of the automatic construction of the classifications of terms. G. Borko [1] and G. Borko and M. Bernik [2], using the method of principal components, accomplished the formation of the classes of terms and the sampling of their necessary number intuitively, on the basis of an analysis of components obtained as a result of statistical analysis. J. Williams [3], who used discriminant analysis, intuitively constructed a system for the classification of documents, which as a result of statistical analysis, was converted into the classification of terms.

The formal definition of class, most frequently utilized in the automatic methods for the construction of classifications, is given thus: a class is that set of elements, the mean value of the factor of the semantic similarity between which is more than the mean value of the factor of the semantic similarity between the elements of the set and the elements of its complement.

Such a definition was used, for example, by R. Needham and V. Ovchinnikov. They used the fixed strategy: R. Needham [4] constructed a nonhierarchical classification of terms whereupon the number of classes was established by intuitive means; V. Ovchinnikov [5] constructed a dichotomous nonintersecting classification of terms.

There are two types of strategies: agglomerative and dividing. The first accomplish the formation of class because of the association of its individual elements; the second — separates class from the entire initial, or present at the given moment, set of elements. Since at the basis of such strategies lies the matrix of semantic resemblance, then the advantage of the dividing strategies becomes obvious, because at every given moment they use all the information placed in the matrix of semantic similarity.

Any classification can be presented in a plane. For this let us designate by T the initial set of terms

$$T=(t_1, t_2, \ldots, t_n) \tag{1}$$

and let us consider the orthogonal coordinate system Oe_1e_2, in which e_1 and e_2 are unit vectors. The position of any element (t_j = 1, 2, ..., n) in such a system is uniquely determined by coordinates (R, L), where R represents the number of the class or subject heading, and L – the level of hierarchy (R and L are positive integers). Then for any t_j, considered as a vector, the following is correct:

$$t_j=Re_1+Le_2 \tag{2}$$

Condition $R = L = 0$ satisfies initial set $T = \{t_1, t_2, \ldots, t_n\}$.

All the existing strategies have a number of degrees of freedom, not exceeding 1, i.e., previously they determine values of R or L, or R and L simultaneously. Thus, for instance, if $L = 1$, then the classification is nonhierarchical. Such classifications we refer to the methods of the fixed strategies. It is evident from (2) that methods of classification with the number of degrees of freedom equal to 2 are possible, i.e., such in which the values R and L are not previously preset. Such classifications we will refer to the methods of free strategies.

From the examination of the fundamental methods of the automatic construction of classifications it follows that they do not completely algorithmize the procedure for the construction of classifications (they do not give the formal definition of class, or they do not give the criteria which determine the necessary number of classes or the optimum number of levels of hierarchy). This deficiency substantially lowers the effectiveness of the application of computers for the creation of the required system of classification.

2. AUTOMATIC CONSTRUCTION OF CLASSIFICATIONS AND AUTOMATIC INDEXING

One of the methods for increasing the effectiveness of IRS [information retrieval system] is the application (along with the automatic construction of classifications) of automatic indexing which uses these classifications. The introduction of such methods, which ensure a sufficiently high degree of accuracy during indexing, will make it possible to develop fully automatic IRS, capable of reconstructing the systems of classifications in short intervals and thereby considering the change in the informational requirements of the users of IRS.

Automatic indexing with the application of automatically constructed systems of classifications was worked on by M. Maron [6], G. Borko and M. Bernik [2] and J. Williams [3], who obtained an accuracy of indexing of 51.8, 55.9, and 62.2% respectively.

Evidently the methods of automatic indexing with the application of automatically constructed systems of classifications possess poor accuracy even when the construction of the classification of terms procedes the intuitive organization of the system for the classification of documents.

3. COMPLETELY ALGORITHMIZED METHODS FOR THE CONSTRUCTION OF CLASSIFICATIONS

Having pointed out the deficiencies in the known methods for the automatic construction of classifications, let us consider two approaches which ensure the complete algorithmization of procedures for the construction of classifications and an increase in the accuracy of automatic indexing with the application of the resulting systems.

The question of the probability of error during indexing arises when it is necessary to refer documents to one or several subject headings. It is possible to offer various methods for the evaluation of such a probability; the shortest method consists of the following.

Let us consider event α, consisting of the assignment to the document of one of the submect headings g (g = 1, 2, ..., m). Let us assume that all outcomes of this event equally probable, that is

$$p(1)=p(2)=\ldots=p(m)=\frac{1}{m}. \tag{3}$$

Then, using entropy of event α as the measure of uncertainty of indexing, we obtain

$$H(\alpha)=-\sum_{1}^{m}p(g)\log p(g)=\log m. \tag{4}$$

It follows from (4) that the entropy of indexing has a minimum at m = 2 (if m = 1, then classification is generally absent).

The presence of only two subject classes cannot always guarantee the necessary depth of indexing. If we select such a nonhierarchical structure, in which every class on any level of hierarchy divides itself into two subclasses, then we will obtain a classification which ensures the necessary number of classes and a minimizing common probability of the error in indexing.

However, the dichotomous classification (in the definition of the subject heading as a set of terms, the mean value of the semantic similarity between which is higher than the mean value of their similarity to the terms which form the complement of the set) can be insufficient in the construction of classifications for many subject areas.

It is possible to propose two methods for the elimination of this deficiency. The first consists of the introduction of another definition of the subject heading. Such a definition can be the following. The subject heading is that set of terms of the smallest power which has a negative total resemblance to the complement of the

set. Here the complement of the set forms another subject heading. However, this approach is inapplicable in cases when the factors of the similarity of the classified objects do not take negative values.

The second method consists of the rejection of the dichotomy and the construction of classification with any necessary number of subject headings. If we reject subjective analysis for the purpose of the complete algorithmization of the procedure for the construction of classification, then it is necessary to introduce the criterion which determines the necessary number of subject headings.

It is evident that a classification obtained in this way should be reasonable from the point of view of man, i.e., possess an informative capacity, since feedback is possible between an automatic IRS and the user, when man can correct the operation of the system based on intermediate results.

The informative capacity of the system of classification is connected primarily with that information which it communicates to the user. This information is determined in turn by the semantics of subject headings and by their structure. Individual terms, if they are isolated, are indefinite. They become single-valued and informative if they have been injected into the system of classifications, i.e., if they are the elements of a semantic field along with other words, forming the context which determines the uniqueness of every element of the semantic field.

The semantic information which is carried by each individual term is connected with its unmarked nature and therefore it can be expressed statistically through the factor of activity. Then the semantic information of the subject heading can be defined as the sum of the activities of all terms entering into it.

The informative nature of the system of classification, on the

other hand, has been stipulated by the possibility of its interpretation. Experiments in the automatic construction of classifications with the application of the theory of clumps showed that the system of classification is more difficult to interpret, the greater the common mutual intersection of subject headings.

Thus the informative nature of the system of classification as the function of its structure and semantic content of the subject headings is proportional to the total activity of all terms which form a classification, and inversely proportional to the common mutual intersection of subject headings.

Let U be the initial set of the terms (j = 1, 2, ..., n), during the algorithmic division of which into groups G_1, G_2, ..., G_p the system of classification is constructed. The informative nature of classification is calculated in the following manner:

$$I = \frac{\sum_{k=1}^{p} \sum_{j=1}^{n_k} A_{jk}}{n_1 + \sum_{k=2}^{p} \sum_{i=1}^{k-1} |G_k \cap G_i|}. \qquad (5)$$

where n_k – the number of terms in k-th group; k = 1, ..., p; A_{jk} – the activity coefficient of term j entering into k-th group.

At the first stages of the construction of a classification, when nonintersecting or weakly intersecting groups are obtained, the numerator in formula (5) increases faster than the denominator, and the informative nature of the corresponding number of subject headings increases. But as soon as strongly intersecting groups are formed the denominator in (5) begins to increase faster than the numerator and the informative nature decreases. In this case the last formed group is broken down and those subject headings are examined which give the maximum of information.

For the evaluation of the activity (importance) of terms it is advantageous to use their "impressiveness" or generality within the limits of a certain subject area. The linguists call such a generality the unmarked nature of a sign and prove that it is found in direct dependence on the frequency of the sign. In the composition of dictionary-minimums for the selection of words they sometimes use the frequency of the word and the number of sources in which it was met. Designating through X_{ji} the frequency of term j in document i, let us determine the activity of term A_j by the following expression:

$$A_j = \frac{N_j \sum_{i=1}^{N} X_{ji}}{N}, \qquad (6)$$

where N_j – the number of documents into which term j enters; N – volume of sample.

Now it remains to find the criterion which determines the optimum number of levels of hierarchy in the resulting system of classification. For finding this criterion, which we will name the stop-rule, we use the syntagmatic aspect of language. It is evident that by accomplishing the series division of the set of terms it is advantageous to dwell on those groups which, from the point of view of the subject indexer, are maximally interpreting headings. For the interpretation of a group it is necessary to formulate certain information from its elements. In this case a syntagmatic bond should exist between the terms entering into the information. Psycholinguistic research showed that the syntagmatic bond is spread out on 4-5 words counting from the beginning of the information. Thus if a group contains 4-5 terms, then from them it is possible to form certain information between the elements of which a syntagmatic bond will exist. If the number of terms n_p in a certain group G_p is less than 4, then from them it is possible to form only "incomplete" information, since it can be augmented by the lacking elements in order that between all elements there would be

a syntagmatic bond (final simple information is formed). Groups with more than 5 terms represent the totality of such simple information between which logical relationships of conjunction or disjunction exist.

Thus we proposed two paths for the completely algorithmic construction of classifications: the first is based on dichotomous principle and is intended for the minimization of the probability of error during indexing, and the second is based on the criterion of information capacity and makes it possible to construct systems of classifications which possess maximum interpretability. The dichotomous principle together with the stop-rule forms the fixed strategy and therefore has a limited field of application; the principle of information capacity together with the stop-rule forms a free strategy and is therefore universal.

4. ALGORITHMS OF THE AUTOMATIC CONSTRUCTION OF CLASSIFICATIONS OF TERMS

The proposed method for the automatic construction of classifications will be realized in algorithms which include: the method of the initiation of groups, the method of the evaluation of group density and the stop-rule, which determines the level of hierarchy, on which the division of groups and factor analysis are terminated.

At the basis of the method of the initiation of groups lie the coefficients of activities of terms which are determined by formula (6). All terms are ranked in the order of their decreasing activity coefficients and for the formation of two (or m) groups the two (or m) most active terms are used.

The centers of groups, having been isolated thus, subsequently accomplish their growth because of the terms having the greatest similarities with the centers. As the measure of similarity let us take the correlation factor which makes it possible to consider not only the appearance of terms in documents, but also the number of such appearances. When a correlation matrix exists for terms R, the measure of group density is calculated from the following formula:

$$B=\frac{S/n_s}{T/n_t}, \qquad (7)$$

where S – the sum of the pairwise correlations between the terms of the group; T – the sum of the absolute values of the pairwise correlations of the terms of the group with remaining terms; n_s and n_t – the number of correlations in the sums of S and T respectively.

The introduction of the sign of absolute value in denominator (7) makes it possible to extend this method to matrices containing negative elements. At $B = 1$ the average value of the cross correlations of the selected group is equal to the average correlation of these terms with all others. Such terms cannot be considered as belonging to each other because they also belong to all the other terms.

In factor analysis the method of multiple groups (7) is known, the essence of which consists of the following.

The path has a sampling from N documents ($i = 1, 2, \ldots, N$), the subject content of which is described by n-index terms ($j = 1, 2, \ldots, n$) with a sufficient degree of completeness If we count in each of the documents the frequency of use of each term, then we will obtain the following matrix

$$X=\begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1N} \\ X_{21} & X_{22} & \ldots & X_{2N} \\ \cdot & \cdot & \cdot & \cdot \\ X_{n1} & X_{n2} & \ldots & X_{nN} \end{bmatrix}. \qquad (8)$$

If the frequencies of terms are standardized and centered, then this is reduced to

$$Z=\begin{bmatrix} z_{11} & z_{12} & \ldots & z_{1N} \\ z_{21} & z_{22} & \ldots & z_{2N} \\ \cdot & \cdot & \cdot & \cdot \\ z_{n1} & z_{n2} & \ldots & z_{nN} \end{bmatrix}. \qquad (9)$$

The problem in the construction of the classification by the statistical method is the economization of the number of characteristics documents. For this purpose it is assumed that all terms (j = 1, 2, ..., n) are found in a linear dependence on m hypothetical terms (factors) F_1, F_2, ..., F_m ($m < n$), the volume of concept of which includes the volumes of the concepts of several initial terms and a linear model of this type is selected:

$$z_j = a_{j1}F_1 + a_{j2}F_2 + \ldots + a_{jm}F_m + a_jU_j, \qquad (10)$$

where U_j – the factor of uniqueness of term j. For the sake of simplicity we will subsequently designate z_j' as z_j. There are n equations of such a type. Coefficients in the case of factors are frequently called loads.

Equation (10) for the value of the term z_j in concrete document i(= 1, 2, ..., N) can be written thus:

$$z_{ji} = a_{j1}F_{1i} + a_{j2}F_{2i} + \ldots a_{jm}F_{mi} + a_jU_{ji} \qquad (11)$$

The dispersion of the normalized frequency of the term on condition that all factors are noncorrelated can be presented in the following manner:

$$1 = S_j^2 = a_{j1}^2 + a_{j2}^2 + \ldots + a_{jm}^2 + a_j^2. \qquad (12)$$

Components to the right represent the fractions of the single dispersion of frequency of a term which are attributed to hypothetical terms and to the factor of uniqueness. Value a_j^2 indicates that fraction of the dispersion of the frequency of term f which cannot be expressed through the correlations of the frequencies of combined occurrence of terms and it is called the uniqueness of the term.

That part of dispersion of frequency of term j which is considered by hypothetical terms can be represented by the sum of the squares of the loads of term j on hypothetical terms and is called the communality of term j:

$$h_j^2 = a_{j1}^2 + a_{j2}^2 + \ldots + a_{jm}^2. \tag{13}$$

Linear model (10) can be written for all terms in an expanded form:

$$\begin{aligned} z_1 &= a_{11}F_1 + a_{12}F_2 + \ldots + a_{1m}F_m + a_1U_1 \\ z_2 &= a_{21}F_1 + a_{22}F_2 + \ldots + a_{2m}F_m + a_2U_2 \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ z_n &= a_{n1}F_1 + a_{n2}F_2 + \ldots + a_{nm}F_m + a_nU_n \end{aligned} \tag{14}$$

Such a set of terms is called a factor set. In it the hypothetical terms F_p (p = 1, 2, ..., m) can be both correlated and uncorrelated. Unique factors U_j (j = 1, 2, ..., n) are always uncorrelated.

Before beginning the analysis, it is necessary to replace the values of the "self-correlations" of terms with their communalities. As a rough estimation of "communalities" it is possible to use the values of the greatest correlations which were taken from the columns of correlation matrix R. Usually the hypothetical terms, obtained as a result of the analysis of multiple groups, are inclined to each other. Therefore the fundamental concept is the concept of the correlation matrix of hypothetical terms. Furthermore because the factors have been correlated, then the direct results of analysis should lead to two matrices: a factor set and a factor structure. The first gives the coefficients of hypothetical terms in a linear description of terms, and the second – the correlation of terms with hypothetical terms.

The essence of the method of multiple groups is the representation of hypothetical terms by axes of reference passing through the centroids of the appropriate groups of terms. Therefore it is advantageous to consider the properties of such sums of terms (centroids).

Although the frequencies of separate terms z_j are standardized, their sums are not necessarily standard. In the method of multiple groups it is assumed that the hypothetical term T_p passes through clusters n_p of terms in group G_p:

$$T_p = \sum z_k \quad (k \in G_p;\ p = 1, 2, \ldots, m). \tag{15}$$

The calculation of dispersions and correlations of hypothetical terms can be accelerated by finding the specific preliminary sums of correlations. The first of them is simply the sum of the correlations of every term z_j with all terms in each group G_p, namely:

$$w_{jp} = \sum r_{jk} \quad (k \in G_p;\ j = 1, 2, \ldots, n;\ p = 1, 2, \ldots, m), \tag{16}$$

where the addition is done on k, accepting all the values of the terms in group G_p. When $k = j$, then "self-correlation" is considered equal to communality, that is $r_{jj} = h_j^2$.

The sum of the correlations between all terms in group G_p with all the terms in group G_g (including the case p = g) has the form:

$$W_{pg} = \sum w_{jg} \quad (j \in G_p;\ p, g = 1, 2, \ldots, m). \tag{17}$$

Sum (17) can be expressed in terms of initial correlations in the following manner:

$$W_{pg} = \sum r_{jk} \quad (j \in G_p,\ k \in G_g;\ p, g = 1, 2, \ldots, m). \tag{18}$$

If $r_{jj} = h_j^2$, then the dispersion of the hypothetical term can be expressed as

$$s^2_{T_p} = \sum r_{ik}(i, \ldots) \ldots \quad (19)$$

Using (17), expression (19) can be simplified

$$s^2_{T_p} = w_{pp}. \quad (20)$$

The correlation between two hypothetical terms T_p and T_g, using the definition, can be presented in the form of

$$r_{T_p T_g} = \sum T_{pi} T_{gi} / N s_{T_p} s_{T_g}. \quad (21)$$

where addition is done for i from 1 to N. Using the previous formulas we obtain

$$\sum_{i=1}^{N} T_{pi} T_{gi} / N = w_{pg}. \quad (22)$$

Now let us find the correlation between the two inclined hypothetical terms

$$r_{T_p T_g} = \frac{w_{pg}}{\sqrt{w_{pp} \cdot w_{gg}}}. \quad (23)$$

Then it is possible to determine the correlation of terms with hypothetical terms (inclined structure) through the previous sums of simple correlations. The element of structure s_{jp} is the correlation $r_{x_j T_p}$ of the term x_j in the standard form with the hypothetical term T_p, the dispersion of which is determined by (19). It turns out that

$$s_{jp} = \frac{w_{jp}}{\sqrt{w_{pp}}}. \quad (24)$$

For obtaining inclined characteristics it is necessary to have the linear descriptions of the terms as the functions of hypothetical terms. The coefficients in these linear equations, i.e., the elements of the set, are the coordinates of the points which represent the terms relative to the inclined axes of the hypothetical terms. The matrix of set P can be obtained from known structure S and the correlations between hypothetical term Φ according to the following equation:

$$P = S\Phi^{-1} \tag{25}$$

Thus we obtained all data necessary for the characteristics of the inclined hypothetical terms. However, it is advisable to carry out the transition to orthogonal hypothetical terms because such terms, after their interpretation as subject headings, should not meet together in the same documents of the array. Therefore the description of the document content by subject headings is considerably simplified.

Of all the possible transitions from inclined to an orthogonal coordinate system during the analysis of multiple groups a special solution is used which possesses the following properties: the first axis of new system coincides with the axis of the first inclined hypothetical term, and the second lie in one plane with the common inclined hypothetical terms and is perpendicular to the first. The new third axis is perpendicular to the plane of the first two, etc. Since every time the pairwise orthogonalization of factors is realized, then for the sake of simplicity we will examine the case of two factors.

If the coordinates of term j in the system of inclined hypothetical terms T_1 and T_2 are designated by b_{j1} and b_{j2}, and in the orthogonal system F_1 and F_2 through a_{j1} and a_{j2}, then the bond between the two sets of coordinates, when F_1 and T_1 coincide, is expressed by the following equations

$$a_{j1}=b_{j1}+b_{j2}\cos\theta_{12}$$
$$a_{j2}=\quad b_{j2}\sin\theta_{12}. \tag{26}$$

where θ_{12} – the angle between inclined hypothetical terms T_1 and T_2.

Since $\cos\theta_{12} = r_{T1T2}$, then (26) can be expressed in matrix form

$$(a_{j1}a_{j2})=(b_{j1}b_{j2})\begin{pmatrix}1 & r_{T_1T_2}\\ 0 & \sqrt{1-r^2_{T_1T_2}}\end{pmatrix}. \tag{27}$$

The conversion matrix, which converts the coordinates of the inclined frame of reference into coordinates of the orthogonal system, can be obtained from the correlation matrix between inclined hypothetical terms. Using the method of the square root to a matrix Φ, a matrix is obtained which proves to be the matrix of conversion:

$$\Phi=T'T. \tag{28}$$

Now we write expression (27) for a case of n terms and m hypothetical terms:

$$A=PT', \tag{29}$$

where A and P are $n \times m$ matrices (a_{jp}) and (b_{jp}) respectively,

and T' – the matrix of dimension $m \times m$, obtained by the method of the square root from Φ. Although (29) ensures transition to the desired orthogonal coordinates, it is advantageous to express the conversion depending on the elements of the inclined structure because they are determined first in the course of analysis. In this case conversion (29) assumes the form

$$A=ST^{-1} \tag{30}$$

Matrix A gives the coordinates of terms in the system of orthogonal hypothetical terms. However, this solution possesses the deficiency that the axis of the first orthogonal hypothetical term is very close to the centroid of the first group of terms, and the axis of the second – is too removed from the centroid of the second group of terms.

Therefore the transition to such an orthogonal solution is of interest, when correlations with the factors of terms, already having high coordinates in the system of orthogonal hypothetical terms F_1, F_2, ..., F_m, approach a unit, and correlations with the factors of terms having low coordinates approach zero.

The determination of new factors B_1, B_2, ..., B_m satisfies the principle of economy (the decrease in the complexity of every term), which is expressed by the varimaks [Translator's Note: varimaksnyy – word not established, probably based on a proper name] criterion.

$$V=n\sum_{p=1}^{m}\sum_{j=1}^{n}(b_{jp}/h_j)^4-\sum_{p=1}^{m}\left(\sum_{j=1}^{n}b_{jp}^2/h_j^2\right)^2. \tag{31}$$

which it is necessary to maximize.

Let us designate the normalized correlations of term x_j with F_1 and F_2 by

$$\begin{aligned} x_j&=a_{j1}/h_j,\\ y_j&=a_{j2}/h_j, \end{aligned} \tag{32}$$

and the normalized correlations with B_1 and B_2 – through X_j and Y_j. Then the orthogonal conversion can be written in the form of

$$(X_jY_j)=(x_jy_j)\begin{pmatrix}\cos\varphi & -\sin\varphi\\ \sin\varphi & \cos\varphi\end{pmatrix}. \tag{33}$$

where ϕ is the angle of turn in plane F_1F_2. If we introduce the designations

$$u_j = x_j^2 - y_j^2,\ v_j = 2x_j y_j,$$
$$A = \sum_{j=1}^{n} u_j,\ B = \sum_{j=1}^{n} v_j \qquad (34)$$
$$C = \sum_{j=1}^{n} (x_j^2 - y_j^2),\ D = 2\sum_{j=1}^{n} x_j y_j;$$

that the degree of turn can be presented as

$$\operatorname{tg} 4\varphi = \frac{D - 2AB/n}{C - (A^2 - B^2)/n}. \qquad (35)$$

Having determined the angle of turn, let us pass from the normalized correlations X_j, Y_j to coordinates b_{j1}, b_{j2} in the system of new orthogonal hypothetical terms B_1 and B_2 according to the following equations:

$$b_{j1} = X_j h_j,$$
$$b_{j2} = Y_j h_j. \qquad (36)$$

Thus it is possible to arrange all terms j ($j = 1, 2, \ldots, n$) in the space of hypothetical terms $B_1, B_2, \ldots, B_m$, using coordinates $b_{j1}, b_{j2}, \ldots, b_{jm}$.

5. CLASSIFICATIONS OBTAINED BY THE MACHINE ROUTE

The initial data for the construction of the "term-document" matrix was a sampling of 208 abstracts from the Reference Journal "chemistry" [RZh "Khimiya"], relating to the chemistry and technology sugar-beet production (eliminating the obtaining of lime and carbon dioxide and the reprocessing of wastes), and also more than 200 keywords describing the given subject area. The average size of the selected abstracts was about 20 lines.

The composition of the dictionary of keywords was the first step in the construction of the thesaurus. Then in the dictionary of the system the synonymy of keywords was reduced by means of the association of incomplete and thematic synonyms and antonyms into classes of arbitrary equivalency. The classes were given names or descriptors, which were the most frequently encountered keywords. For instance, the descriptor which received the name "Diffu" served for the designation of the following class of keywords which are considered equivalent within the limits of the selected subject area: juice extraction, malting, leaching, recovery, extraction, crude. After the elimination of synonymy the dictionary consisted of 200 descriptors (terms). Each term represents a word form, a combination of word forms, a word, its truncated form, or the totality of truncated forms. Truncation was accomplished for the purpose of reducing the volume of the dictionary and is thus that the chain of letters which forms the truncated form and is superimposed from left to right was contained in all word forms, from which appropriate truncated form was obtained.

For all documents retrieval patterns were made up. These were the sets of terms together with the number of their appearances in documents. The compilation of retrieval patterns was carried out manually according to the following rule. If with the superposition of a certain term from left to right on any lexical unit of text it was revealed that the term coincides with the lexical unit or is contained in it, then it was considered that the corresponding term is included in the document. In the compilation of retrieval patterns the bibliographical data of the primary document were not considered.

All the retrieval patterns make up the initial term-document matrix with the dimension 200 × 208. Since the machine processing of a matrix of such a size requires considerable time, then it was reduced by means of the elimination from it of the terms, the activity coefficients of which did not exceed 1. The remaining terms, used subsequently for the construction of the classification, formed by a matrix of 91 × 208.

For the realization of the proposed procedures two programs were written in the assembly language of the BESM-6. The first program, which realizes the dichotomous principle of the construction of classification, contains 1400 one-address instructions; its block diagram is represented in Fig. 1. The flowchart of a program, containing 1430 one-address instructions and accomplishing the automatic construction of classification which maximizes information capacity, is represented in Fig. 2. Both block diagrams differ in some control instructions and unit 2A (Fig. 2), The remaining units coincide and have the following values.

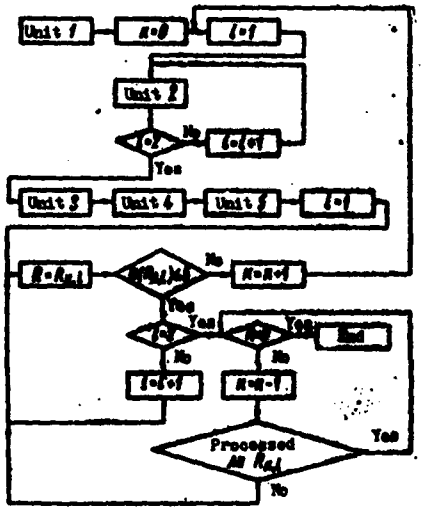


Fig. 1. The flowchart of the program for construction of a dichotomous classification.

Unit 1 uses the initial data, placed in the term-document matrix with the dimension 91 × 208, for the construction of the correlation matrix of terms of the dimension 91 × 91. Unit 2, performing as the correlation matrix and the matrix of the activity coefficients of terms, separates the semantic centers of groups and accomplishes their growth by means of the calculation of B-coefficients. The result of the operation of this unit are the groups of terms, and also the correlation matrices of terms corresponding to these

groups. Unit 3 unites the resulting groups of terms and their corresponding correlation matrices, preparing them for factor analysis, which is accomplished by unit 4.

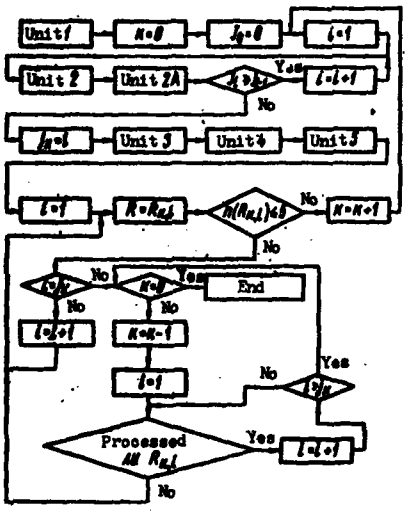


Fig. 2. The flowchart of the program for the construction of a classification which maximizes information capacity.

The result of the operation of unit 4 is the correlation matrix of terms with factors and the correlation matrix between factors. Unit 5 accomplishes the orthogonalization and analytical rotation of multiple factors.

Unit 2A (Fig. 2) is contained only in the second program and is intended for the comparison of the groups obtained as a result of the operation of unit 2, and the calculation of the criterion of information capacity.

In the construction of a classification by the dichotomous method at the first level of hierarchy two subject headings are formed. This is not sufficient for the content description of the selected subject area because they do not contain the terms relating to the chemistry and the storage of sugar beets.

The criterion of information capacity facilitated the extraction of the groups of closely connected terms: on the first level of hierarchy the information capacity of the first group of terms (subject heading 3) comprised 20.6, the information capacity of the first three groups (heading 3, 2, and 1) - 38.5. After the formation of the fourth group the total pairwise intersection of groups increased sharply and information capacity was reduced to 20.4. Therefore the fourth group was destroyed and at this stage the formation of the first level of hierarchy was terminated.

The first experiments on the further division of the groups of terms showed that the subgroups being formed intersect very strongly. This permitted the making of two conclusions. In the first place, in proportion to the specialization of the groups of terms their density or homogeneity, measured by the B-coefficient, should increase. Therefore the value of the B-coefficient, every time during the transition to the following level of hierarchy increased by 1. In the second place, as centers in the formation of subgroups one should select terms which are less active than during the formation of groups.

The appropriate changes were introduced into the program for the BESM-6 and as a result 13 factors were obtained together with the correlations of terms with factors after their orthogonalization and analytical rotation. About 4 minutes of machine time were required for obtaining the factors.

The factors represents the groups of terms organized in the sequence of reduction of their correlations with all elements of the groups. These groups were subjected to analysis depending on the semantic content of the terms and the values of their correlations with factors. The analysis was conducted for the purpose of awarding names to the groups which were equivalent in their content to the subject headings. This procedure, which is called the interpretation of factors and is subjective, is carried out depending on the semantics of those terms which have the highest correlations with the appropriate factors.

The subjectivity of interpretation does not contradict the requirement introduced by us previously for the complete algorithmization of the procedures for the construction of classifications, because it is conducted for the purpose of checking the accuracy of automatic registered indexing with the application of the classification obtained.

As an example Table 1 shows one of the 13 groups obtained containing nine terms and their correlations with the appropriate factor.

Table 1.

Carat	0.82
Storage-preservation	0.75
Beet-root	0.74
Ventilation	0.70
Waste-damage	0.49
Technolog-	0.44
Sugar	0.36
Inver- -inversion-invert sugar RV	0.35
Content	0.30

As a result of the interpretation of this group the following subject heading was adopted.

The storage of sugar beets. Physicochemical processes during storage.

The subject headings and their hierarchical structure obtained during the interpretation of all factors are given in Table 2.

6. AUTOMATIC INDEXING AND ITS RESULTS

For checking the value of the classification obtained and, therefore, checking the suitability of statistical methods for the automatic construction of classifications, the automatic indexing

Table 2. The classification of terms: TECHNOLOGY OF SUGAR-BEET PRODUCTION.

1.	2.	3.
The storage of beets. Physico-chemical processes during storage. The technological merit of sugar beets. The drying of sugar. The storage of sugar.	Obtaining of raw juice. Cleaning, clarification, and filtration of juice. Boiling of residue.	The evaporation of juice. Physicochemical processes during evaporation. Clarification of syrup. Boiling down, crystallization and centrifuzing of massecuite. The automation of boiling down of massecuite.
1.1. Storage of sugar beets Physico-chemical processes during storage.	1.2. Technological merit of sugar beets. The storage of sugar. The drying of sugar before storage.	2.1. Obtaining of raw juice. The cleaning of juice for preliminary defecation.
		2.2. Cleaning of juice for defecation and saturation. Demineralization. Filtration and the clarification of juice.
		3.1. Clarification of syrup by serbents.
		3.2. Boiling down, crystallization and centrifuzing of massecuite.
		3.3. Automation of the boiling down of massecuite.
	2.1.1. Obtaining raw juice.	
	2.1.2. Cleaning of diffusion juice for preliminary defecation.	
		2.2.1. Cleaning of juice for defecation and saturation. Filtration of juice.

of a control sampling of 100 abstracts was carried out. The criterion of accuracy of automatic indexing was manual indexing, carried out by three subject indexers (specialists in the technology of sugar-beet production) independently from one another. The subject indexers referred every abstract to the most relevant subject heading or subheading, which were taken from Table 2. Every abstract was given that subject heading (subheading) which was assigned to it by no less than two subject indexers. The results obtained were considered absolutely exact and the results of automatic indexing were compared with them.

In factor analysis the measure of the relevancy of the document to the subject heading is considered as the sum of the products of the number of appearances of terms in a document and their correlations with the subject headings. In our case such an approach is not a applicable, since the classification is hierarchical, that is, a subheading always contains a smaller number of terms than the corresponding subject heading.

In the compilation of an algorithm of automatic indexing they proceed from the following considerations. In the first place, all the subheadings should be reduced to the dimension of the appropriate headings. In the second place, the value of the relevancy of the document of the subject heading (subheading) should be greater, the greater the intersection of the retrieval pattern of the document with the subject heading (subheading).

Because of the aforesaid, the calculation of the value of relevancy RV of the document i to the subject heading (subheading) p is accomplished using the following formula:

$$RV_{ip} = \frac{a_{ip}}{S_i} \sum_{j=1}^{S_i} b^*_{jp} x_{ji}, \qquad (37)$$

where a_{ip} – the intersection of the retrieval pattern of the document with the subject heading (subheading) p; S_i – the number of terms of

the retrieval pattern of the document; b^{*}_{jp} – the correlation of term j with the subject heading (subheading) p after the reduction of the dimensions of the subheading to the dimensions of the corresponding headings.

Let X^{*} be the document-terms matrix which represents the numbers of appearances in each of the documents of the sampling of all terms; B – the correlation matrix, the term-subject heading (subheading), obtained as a result of factor analysis, orthogonalization and analytical rotation, and β_{pk} – the coefficient of reduction of the subheading to the dimensions of the appropriate heading p. Then (37) can be rewritten in matrix form:

$$(RV_{ip}) = \frac{\alpha_{ip}\beta_{pk}}{S_i}(X^{*} \cdot B). \qquad (38)$$

The appropriate program of automatic indexing was written in the ALGOL language and realized on the BESM-6. For the indexing of one document about 0.3 seconds of machine time is expended. The computer correctly indexed 64 documents out of 99, so the accuracy of automatic indexing comprised 64.6%.

One ought to note especially that the 64 correctly indexed documents did not include 17 of those which were assigned by the machine not to a subheading (as in manual indexing) but to the corresponding subject heading. Thus the introduction of a more exact criterion of relevancy can increase the accuracy of automatic indexing with the application of the classification obtained automatically by at least up to 80%.

7. THE RELIABILITY OF MANUAL INDEXING AND AUTOMATIC INDEXING.

The accuracy of automatic indexing in many respects is determined by the concordance of the results of manual categorization. This circumstance puts forth the requirement for the measurement of the

reliability of manual indexing and an evaluation of that accuracy of automatic indexing which would take place if manual indexing was absolutely reliable.

The result of any indexing is the attaching to the document of the most relevant subject heading or subheading, which emerge only by nominal definitions and do not have a quantitative expression. Therefore for the measurement of the degree of concordance of the results of indexing only those statistical criteria are used which measure the closeness of the bond between quality characteristics. As such a criterion we selected the coefficient of mutual contingency, which in a specific sense is the equivalent of the correlation factor. The criterion ensures the possibility of the comparison of our results with the results obtained by G. Borko [1]. For the calculation of the coefficient of contingency between the results of indexing by various subject indexers and by machine the appropriate program was written in the ALGOL language and realized on the BESM-6. The results of the computations are given in Table 3; manual indexing is marked by the indices 1, 2, 3, automatic – by the letter a

Table 3. Correlation of the results of indexing.

(a) Метод	1	2	3	a
1		0,9369	0,9371	0,9053
2	0,9369		0,9371	0,8946
3	0,9371	0,9371		0,8954
a	0,9053	0,8946	0,8954	
	R_{mm}=0,9370		R_{am}=0,8988	

KEY: (a) Method

The reliability of manual indexing R_{mm} is the mean of the coefficients of contingency C_{12}, C_{13}, C_{23} and comprises 0.9370. The correlation of the results of automatic indexing and manual indexing C_{a1}, C_{a2}, C_{a3} in all cases is lower than the correlation of the results of manual indexing, and its mean R_{am} = 0.8988 indicates

the degree of coordination between automatic indexing and manual indexing. It is necessary to note especially that the correlation of the results of automatic and manual indexing $R_{am} = 0.8988$ obtained by us is somewhat higher than the correlation between various subject indexers $R_{mm} = 0.870$ by G. Borko [1]. This again confirms the effectiveness of the method selected by us for the automatic construction of classification and indexing.

If manual indexing was absolutely reliable, then the correlation of automatic indexing with it comprised $R_{aam} = R_{am}/R_{mm} = 0.9592$. Now let us present the total number of documents being indexed in the form of the sum of two terms: the number of correctly and incorrectly indexed documents (with absolutely reliable manual indexing). The first value is determined by the coefficient of determination $D = R^2_{aam} = 0.9200$. This means if manual indexing was absolutely reliable, then with an automatically constructed classification and the described algorithm of indexing it was possible to correctly index 92% of the total number of documents, which is considerably higher than in G. Borko's method (67%).

CONCLUSIONS

1. The results of the test show that the criteria proposed for the complete algorithmization of the construction of classifications (the criterion which defines the number of levels of hierarchy in the resulting classification, and the criterion of information capacity which determines the optimum number of subject headings on every level of hierarchy) permit the constructing of systems of classifications which are easy to interpret and describe sufficiently fully the appropriate subject area.

2. The machine test on automatic indexing with the application of an automatically constructed system of classification confirms the effectiveness of the described method of construction of classifications, because the results were higher even in comparison with the methods of the indexing which intuitively precedes the compiled classification of documents

BIBLIOGRAPHY

1. Borko H. Measuring the reliability of subject classification by men and machines. «Amer. Docum.», 1964, 15, № 4, 268—273.
2. Borko H., Bernick M. Automatic document classification. Part 2. Additional experiments. «J. Assoc. Comput. Mach.», 1964, 11, № 2, 138—151.
3. Williams J. H. Results of classifying documents with multiple discriminant functions.—M. Stevens a. o. /eds/. «Statistical association methods for mechanized documentation. Symp. Proc. Washigton, 1964», NBS Misc. Publ. 269, 1965, 217—224.
4. Needham R. M. Application of the theory of clumps. «Mech. Translat.», 1965, 8, № 3—4, 113—127.
5. Овчинников В. Г. Автоматизированный метод построения тематических систем классификации для информационного поиска. Автореф. канд. дисс. М., 1969.
6. Maron M. E. Automatic indexing: an experiment inquiry. «J. Assoc. Comput. Mach.», 1961, 8, № 3, 404—417.
7. Harman H. H. Modern factor analysis. Chicago, Univ. Press, 1962, 471 pp.

Article was received by the editors
10 June 1970