SOVIET PHYSICS

USPEKHI

A Translation of Uspekhi Fizicheskikh Nauk

SOVIET PHYSICS USPEKHI VOL. 2(69), NO. 5, pp. 637-781 SEPTEMBER-OCTOBER, 1959

SCIENTIFIC AND TECHNICAL INFORMATION AS ONE OF THE PROBLEMS OF CYBERNETICS

G. É. VLÉDUTS, V. V. NALIMOV, and N. I. STYAZHKIN

Usp. Fiz. Nauk **69**, 13-56 (September, 1959)

Mechanized methods of information storage and processing have recently gained wide application in physics, in its technical applications, and in fields associated with physics. We therefore consider it advisable to examine both the theoretical and technical aspects of this problem in a physics journal.

1. THE EXPONENTIAL CHARACTER OF THE DEVELOPMENT OF SCIENCE

A statistical analysis of the development of science shows that the number of published papers, the number of journals, the number of workers engaged in research, and expenditures for research increase exponentially. The curve of Fig. 1[1,2] shows the increase in the total number of references in Physics Abstracts, beginning with 1900. This curve deviates from exponential only for the years of the Second World War, but again becomes parallel to the original curve after the end of the war. An analogous situation is observed for the growth in the number of publications in chemistry[3] and biology.[4] Figure 2 shows the growth in the number of scientific and abstract journals[2] (logarithmic ordinates). Here, too, an exponential growth is observed, and at the present time the number of scientific journals* approaches 100,000, while the number of abstract journals has reached nearly 300.[2] An analogous growth is observed, according to foreign data, in research budgets.[5] The lack of corresponding statistical data, makes it impossible to plot directly the number of scientific workers, but many indirect indices — the increase in number of students graduated from appropriate institutions of learning and the number of "known scientists," referred to in various dictionaries and handbooks, and also the growth in the number of publications, patent disclosures, references to scientific work, etc. allow us to state that here, too, the growth is exponential.

An analysis of the growth curves for the criteria that characterize the development of science leads to the conclusion that in all the foregoing cases the development is exponential with accuracy to 1% (reference 2). The parameters of the exponential curves of the different criteria vary within a relatively narrow range in such a manner that all the criteria double within ten or fifteen years.

An interesting mental experiment is extrapolation of the exponential curves. Extrapolation to the historical past leads to logical results. It is found that the ordinates of almost all the curves reach a value of unity at approximately 1700, i.e., Newton's time. For example, the extrapolated scientific-journal curve passes through unity at 1700. Actually, the first scientific journals appeared in 1665 and during their early years comprised a small group which does not fit the general law. Newton's era can be considered as the beginning of the historical era of great scientific discoveries, which is continuing to this day. The expon-

*Information given by various authors on the number of scientific journals vary over a wide range, since the term "scientific journal" has not yet found a universal rigorous definition. The abstract journals of the Institute of Scientific and Technical Information of the U.S.S.R. Academy of Sciences publish abstracts of articles from 11,000 or 12,000 scientific journals. This does not include journals devoted to the humanities.

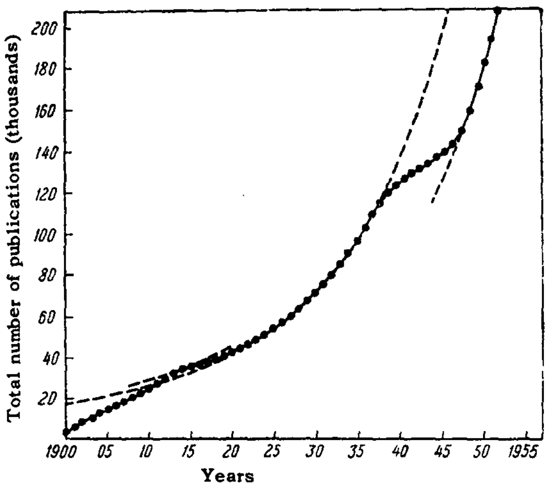


FIG. 1. Increase in the total number of references in Physics Abstracts, beginning with 1900 (the ordinates represent cumulative sums).[1,2]

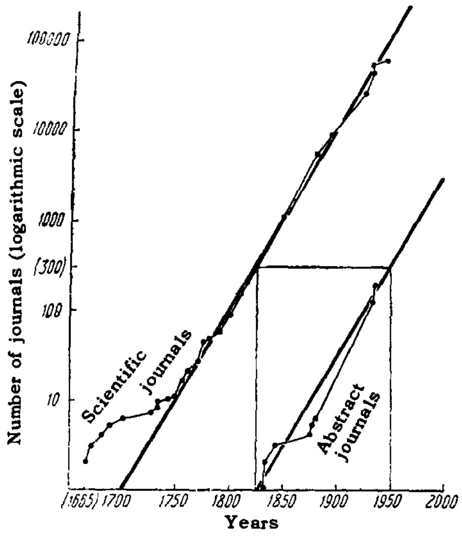


FIG. 2. Increase in the number of scientific and abstract journals.[2]

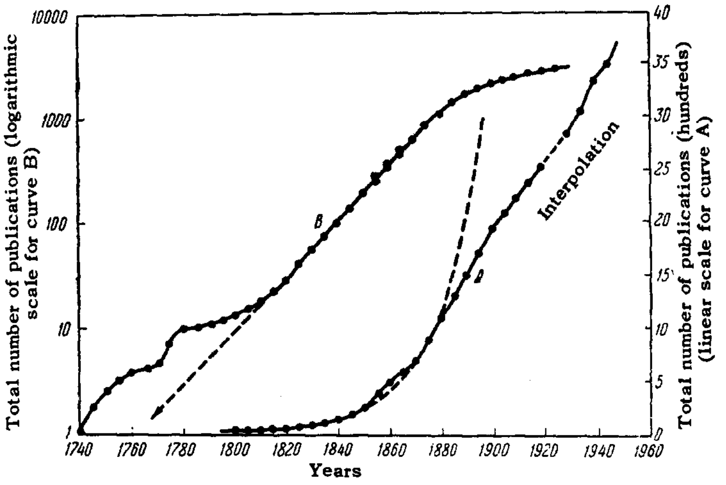


FIG. 3. Growth in the number of publications on the theory of determinants and matrices.[1]

ential character of the growth of science, based on most criteria, can be traced over 200 – 250 years.

This law of growth is observed for individual indices only in sufficiently broad fields of science, such as physics, chemistry, or biology. If we examine some narrow discipline we observe first an exponential growth, but once the development potential of such a discipline is exhausted, the increase in the number of published papers becomes linear in time. This is illustrated in Fig. 3, which shows the growth curve for the total number of publications on the theory of matrices and determinants.[1] The first publication in this field appeared approximately in 1750. Starting with 1800, when the total number of published papers was 10, the growth was strictly exponential up to 1880, after which the growth became linear. The statement concerning the exponential character of the development of science is in complete agreement with highly probable premises of sufficiently general nature. The exponential character of the development of science can be described by the relation

$$y = ae^{kt} \qquad (k > 0),$$

which is the solution of the differential equation

$$\frac{dy}{dt} = ky,$$

where the derivative dy/dt denotes the rate of growth of the index of interest to us, i.e., its increase per unit time. Thus, the exponential course of the development is the result of the fact that the relative rate of the growth is constant

$$\frac{dy}{y\,dt} = \text{const.}$$

It is quite natural to assume a priori that in the absence of limiting factors the rate of growth should be determined by the level attained: each new scientific concept should give rise to a definite number of new scientific papers — which either develop and deepen this concept or contradict it.

It is natural here for the growth "constant" to vary within a narrow range for the various criteria that characterize the development of science, so that we find a doubling within approximately the same time interval, 10 or 15 years, for all criteria. It is easy to calculate that a doubling within 10 — 15 years corresponds to a relative growth of 5 — 7% annually. The growth in the expenditures on research is somewhat faster, for according to U. S. data[5] its relative rate is 10% annually. When the growth in number of publications becomes a linear function of the time, this means that the absolute rate of growth is constant and independent of the level attained.

If we continue our mental experiment and extrapolate the exponential curves into the future, we unavoidably reach absurd results. After a tenfold doubling, i.e., after 100 — 150 years, the number of publications and the number of scientific workers should increase by a factor of a thousand, and after 200 — 300 years — by a factor of a million. The number of scientific workers would increase faster than the earth's population. According to the latest U. N. data (Izvestiya, 1959, June 2, No. 129), the relative rate of growth of the earth's population is 1.6% annually, corresponding to a doubling of the population in approximately 45 years (it was assumed earlier[6] that the rate of growth is approximately 1%, corresponding to a doubling of the population in 70 years). It is therefore natural that the growth curve of the number of scientific workers should reach saturation: the exponential curve must become an S-shaped curve. A similar situation is usually described in biology by means of the so-called logistic curve, shown in Fig. 4. The analytic expression for the logistic curve

$$y = \frac{b}{1+ae^{-kbt}} \qquad (k > 0)$$

is the solution of the differential equations

$$\frac{dy}{dt} = ky(b-y) \qquad (0 < y < b).$$

In this case the growth is limited, since b is the maximum value of y. The relative rate of growth

$$\frac{dy}{y\,dt} = k(b-y)$$

is no longer constant, but a linear function of y. The higher the level attained by the criterion of interest to us, the slower the growth. The reduction in the rate of growth becomes noticeable only at sufficiently large value of y. During the initial time, when $y \ll b$, the logistic curve almost coincides with the exponential one, as shown in Fig. 4.

From a statistical analysis of the process of development of science, several predictions can be made concerning its further development.

The number of scientific papers has now reached such a level, that the inadequacy of abstract journals as means of transmitting primary information has become evident. An important factor here is that the individual scientific disciplines no longer have clear outlines. An active research worker cannot confine himself to information in any one particular field, but must refer to allied disciplines. At the present time in many cases, the most interesting papers are now those on the border line between sciences. The intertwining of sciences also apparently increases exponentially. Even fields so foreign to each other as biology and mathematics have now become interrelated. The situation in technical research is similar. The laboratory director in a large metallurgical enterprise must keep up not only with the literature on metallurgy and metal research, but also with the development of chemistry and certain branches of physics and mathematics. Success in the operation of a laboratory is determined to a considerable degree by the extent with which new physical and mathematical (statistical) methods are applied to the investigation of technological processes and to the control of the product quality. A statistical balance sheet of the working time of a research chemist was recently compiled in the U. S.[7] Chemists were chosen for this investigation because they represent the largest group of scientific and technical workers — the American Chemical Society has 80,000 members. The results of the investigation have shown that an average of 33.4% of the total researcher's time is devoted to scientific information. The minimum value of this quantity was 15.7%, and the maximum 61.4%. The term "scientific information" covers all the processes connected with the evaluation of scientific and technical problems, search and reading of the literature, and written and oral communication. If the scope of information work is defined more broadly, by inclusion of the additional planning and thinking about the experiments on the one hand, and service and administrative information on the other, the research chemist is found to spend an average of 49.8% of his time on informational activity. The minimum of this figure becomes 20.0%, while the maximum 94.5%. Thus, even now the researcher loses an average of 50% of his time to informational activity. As the number of publications grows exponentially it is obvious that the time consumed by information work will increase, and the research chemist will no longer be able to engage in experimental work unless new more effective means are found for information service. The great significance of mechanization of scientific and administrative information activity now becomes evident.

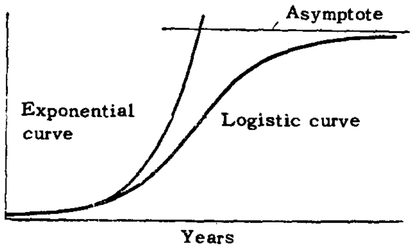


FIG. 4. Transition from the exponential curve to a logistic curve in the presence of a growth-inhibiting factor.[2]

In connection with the development of information machinery, the practice of publication of papers should change radically in the nearest future. A common statement made by scientific workers, particularly those abroad, is "that there are enough new journals." Yet the existing journals are incapable of publishing all the material received. Both in the Soviet Union and abroad the delay in publication of articles sometimes reaches two years. In some of our journals, for example "Zavodskaya laboratoriya" (Plant Laboratory), approximately 50% of the manuscripts received are rejected only because of lack of space, and the average length of articles in this journal was reduced to one half within the last ten years. According to U. S. data,[8] 48.5% of the papers delivered to conferences are printed only in the form of brief communications.

This being the situation, a considerable portion of the information received is practically unavailable to a large circle of research workers. The presently noted tendency of reducing the length of published articles will apparently continue to develop, at least for papers of experimental nature. Under such a publication system, loss of the information can be avoided only if it is placed in the long-time memories of the information machines.*

The change to new methods of information requires the development of strictly standardized methods of compact contraction of the results of the experiment and a strict quantitative estimate of that element of indeterminacy, which is connected with each experiment and which is due to unavoidable instrumental and methodological errors, to the limited nature of the experimental material, etc. Earlier, when editors did not limit the length of an article, the author could report the results of his experiments in any method convenient to him. The reader obtained a certain idea of the reliability of the results from extensive descriptions of the experimental conditions, methods for reducing the experimental data, etc. Now, when the length of the articles has decreased and when hundreds and sometimes even thousands of articles are available on each problem, no matter how narrow, a direct evaluation of the results of experimental investigations becomes impossible.

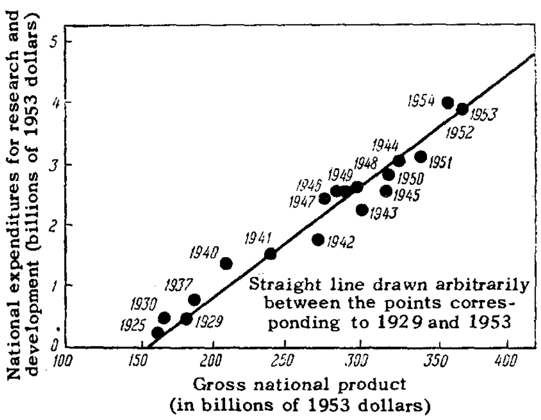


FIG. 5. Correlation between the growth of the gross national product and expenditures for research and development in the U.S.A.[5]

The reader cannot distinguish good work from bad. The presence of undisclosed poor work sometimes raises doubts concerning the results of good work. Standardization in the methods of reporting experimental results has become necessary even at that level of contraction of information, which is practiced at present in the publication of the briefest articles. This problem will become even more important in machine methods of information processing, when the information will be subject to further contraction.

If the growth curve for the number of workers engaged in research is to change in the nearest future from exponential to logistic, the question arises whether the growth of the number of scientific papers will remain exponential. To satisfy the human striving for knowledge, on the one hand, and to satisfy the increasing material needs on the other, it is apparently necessary that scientific research increase exponentially, as in the past 200 — 250 years. Heretofore there was always a correlation between the growth in gross production and the increase in expenditures on research, and consequently, also the number of research investigations, as shown, for example, in Fig. 5, which lists data that characterize the growth of gross production and expenditures on research in the U. S. A. It is obvious that the exponential growth of research work can continue only if part of the intellectual labor is performed by machines. Machines should facilitate the human intellectual activity to such an extent, that the total effort expended on research continues to increase exponentially even when the number of researchers increases along a curve close to logistic.

The exponential growth in the number of scientific investigations will undoubtedly lead to an increase in education time, unless new ways are

*Simultaneously, the role of survey literature should increase. According to statistical data,[9] even now the chemists of the U. S. make use of survey articles published in 50 different journals. Some of these journals publish survey articles only periodically or sporadically, simultaneously with original papers. So large a number of survey publications is explained by the fact that the surveys are written in response to specific requests by some group of workers. For example, the fourth (April) issue of the journal Analytical Chemistry is published in two parts, one devoted to survey articles on a great variety of problems of interest to analytical chemists.

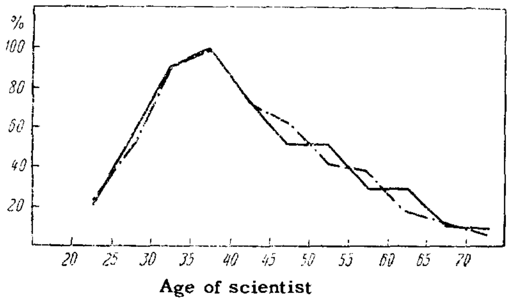


FIG. 6. Distribution of the number of published papers relative to ages of scientists. The solid curve pertains to data on 14 different countries, not including the U.S.S.R., England, France, Germany, and the U.S.A. The dash-dot curve pertains to data on the U.S.A.[10,11]

found here. In Gauss's time a man embarked on creative work at the age of 20. At the present time this age is approximately 25. Only within the last 20 — 30 years, study time has increased by 2 or 3 years. A further increase in education time can hardly be desirable, if it is considered that a scientist's maximum productivity, measured by the number of published papers,* occurs at an age of 35, as shown† in Fig. 6.[10,11]

The development of information machines should change radically the system of education. The students will be freed of the need of memorizing a tremendous amount of factual material, which at the present time burdens the consciousness of every researcher.

Until recently the problem of information in science and technology was restricted to overcoming spatial, temporal, and linguistic barriers. At the present time, in connection with the exponential character of the development of science, a new barrier has arisen, caused by the fact that the unaided human consciousness cannot perceive directly and digest the great flow of information. We proceed to consider certain attempts at overcoming this barrier.

2. MACHINE-STATISTICAL METHOD OF CONTRACTING THE TEXT OF AN ARTICLE

The machine-statistical method of abstracting articles, proposed in reference 12, can serve as an example of an attempt at mechanizing one of the most laborious and fatiguing processes of intellectual activity. The urgency of this problem becomes obvious if it is realized that the Institute of Scientific and Technical Information of the U.S.S.R. Academy of Sciences now employs the services of 12,000 free-lance abstractors. Within 100 — 150 years the number of free-lance abstractors should reach 12,000,000, if the number of publications continues to grow exponentially and if no mechanical means of processing and disseminating information are found.

The machine-statistical method of abstracting is based on the study of the frequency of appearance of words in an article and of their distribution. Among the many words used by the author of the article one must find the ones that are "significant" for the transmission of the given information. It is assumed that the significance of the words is determined by the frequency of their appearance in the abstracted paper. It is then necessary to separate the significant phrases, capable of representing in the best manner the contents of the article. The significance of the phrases is determined by how closely the significant words are associated in them.

In practice, abstracting can be carried out in the following manner: the text of the article is transferred by means of some reading device to a magnetic tape and is fed to a computer. The computer is programmed as follows: read the quoted text and separate the individual words, noting the position of each word in the article and the position of each phrase and of each paragraph in which the given words are found. Simultaneously, the machine should discard words of general character: pronouns, prepositions, articles, etc., which produce "noise." The remaining words are arranged alphabetically and the machine identifies those serving to designate the same concept. This operation is carried out by comparison of the words letter by letter. If two words contain six identical letters, they are deemed indistinguishable. Thus, for example, the words: differ, differentiate, dif-

*On the basis of statistical data, Schockley[11] believes that the number of published papers can be a measure of the scientist's creative activity. A plot of the number of papers vs. the scientist's age yields a normal-logarithmic curve. The logarithm of the number of published papers is considered as a characteristic of the scientist's intellectual ability. The creative productivities of scientists engaged in the same laboratory differ by factors of several times 10. The ability for creative work varies over a considerably greater range than the abilities that can be tested with the aid of ordinary psycho-physiological tests. This interesting phenomenon is explained by Schockley as follows: If, for example, the creation of a scientific concept requires the ability of associating four ideas, then, as follows from combinatorial rules, a scientist capable of associating six ideas will have 15 times the potential capabilities for creative work than a scientist capable of associating only four ideas. A small variation in the ability of associating ideas leads to a very large difference in creative capabilities. Not all of Schockley's conclusions are sufficiently convincing. Nevertheless, the idea of employing statistical methods for investigating processes of creative activity is undoubtedly worthy of attention.

†We have in mind here publications on creative research.

ferent, differently, difference, and differential are considered as symbols that designate the same concept. Such a method of identification of words is not beyond reproach — the identification error is estimated to be 5%; however, it cannot distort substantially the results of the statistical analysis.

Finally, after carrying out all the foregoing operations, the frequency of appearancé of the words is computed. The significant words are taken to be those whose frequency exeeds a certain value, selected on the basis of a series of preliminary experiments.

The significance of phrases is established in the following manner: The positions of the significant words are noted in each phrase and the portions of the phrase containing the significant words are enclosed in square brackets, as shown in Fig. 7. This is followed by counting the number of significant words in the square brackets, squaring this number, and dividing it by the total number of words contained in the square brackets. The quotient thus obtained is considered as a "weight" that characterizes the significance of the phrase. The abstract is prepared by joining mechanically the phrases with the maximum weights.

The foregoing method of estimating the weight of a phrase can be justified from a relation known in semantics,* according to which the multiplicity of a meaning of the word, and consequently its meaningful weight, is proportional to the square root of the frequency of its appearance.[13]

The first experiments in machine-statistical abstracting yielded favorable results.

It is interesting to note that machine-statistical abstracting can be considered as the simulation of one of the processes of abstracting work. If the editor of the abstract journal receives an article written in some unknown language, and the article has no diagrams, formulas, or other generally understandable information symbols, then, to gain some idea about the contents of the work, the editor, after scanning the text, chooses the phrases with the most frequently encountered words and then translates them literally with the aid of a dictionary.

An obvious shortcoming of an abstract written on the basis of a statistical analysis of the text is that it is somewhat dogmatic. The reader is presented with unrelated individual phrases cut out of the context. On the other hand, an undisputed advantage of such an abstract is its objectivity. The generally accepted process of abstracting at the present time is in the nature of an art. Abstracts written by different professional abstractors frequently differ greatly from each other. The abstractors unavoidably introduce a certain element of subjectivity in their work, owing to the features of their scientific viewpoint, the selectivity of their interests, etc. At the same time abstracts written by the authors of the article are frequently quite unacceptable, since authors usually lack the art of contracting the information. The inaccuracy in the representation of the contents of an article by means of abstracts is particularly obvious in the short annotation abstracts, for which the machine-statistical methods will apparently be particularly effective.

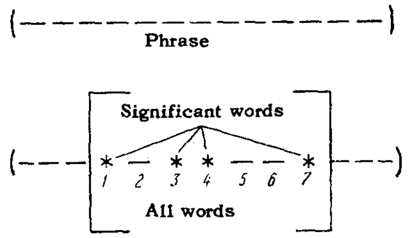


FIG. 7. An estimate of the statistical weight of a phrase. The portion of the phrase containing the significant words is enclosed in square brackets. The weight of the phrase equals the square of the number of the significant words divided by the total number of words in the square brackets: $p = \frac{4^2}{7} \approx 2.3$ (reference 12).

Statistical methods can also be used for machine indexing of a text. Several methods have been proposed for this purpose. The simplest one, the copying of the most frequently encountered words, can hardly be considered satisfactory. The contents of the article cannot be coded by means of a set of individual words, since scientific concepts are most frequently defined not by individual words, but by word combinations that produce a semantic field. For example, the noun "band" in conjunction with various defining words may serve to designate a large number of concepts: "optical absorption band, energy band," etc. The machine should therefore be programmed to write out the most frequently encountered nouns and the associated defining words. An interesting experiment was carried out in this direction by Baxendale[14] who proposed to use certain peculiarities of the English language for the programming. The indexing unit was chosen to be the "prepositional word combination," which by definition is the unit of thought expression, consisting of a preposition, a predicate or pronoun, and the associated definitions. The

*Linguistic semantics is a discipline engaged in a study of the meaningful significance of words. The problems of linguistic semantics were first formulated by M. Bréal[13] in 1883. A vigorous development of this discipline began only in recent years. The word "semantics" is derived from the Greek word σημαινειν — to have meaning.

length of the "prepositional word combination" varies from two to seven words, and equals on the average four words. The machine is programmed to identify all the prepositions (their number does not exceed 50) and to write out the first four words following the preposition, unless a new preposition or a punctuation mark are encountered first. For example, in the clause "within the scope of natural English language, an infinite number of different sentence structures is possible" the machine writes out all the underlined words. When this method is combined with an estimate of the frequency of the words in the text one can write out indexing units, as shown in Table I for the first three words most frequently encountered during the indexing of one of the articles devoted to solid state physics. The program is so set up that 0.5% of the words are picked out of the article for indexing.

Machine-statistical indexing simulates the scanning of the text, which usually precedes the reading of the article. Before he decides whether it is worth while to read the article, the scientist scans it, reading individual short prepositional word combinations (which can be scanned by a single glance) with the most frequently encountered words.

TABLE I. Example of automatic coordination of the terms as a property of "prepositional word combination"

Electron	Electron energy level Electron energy Valence electrons Electron waves Free electrons
Energy	Energy band structure Energy gap Energy spectrum Discrete energy levels Forbidden energy region Kinetic energy Energy curves
Band	Band theory Conduction band Valence band Energy band structure

Machine-statistical methods of text contraction are still in the experimental stage. New experiments indicate without a doubt that research in this direction is promising. It must be noted that all experiments were made with commercial computers such as IBM 704 and IBM 650.

This new trend in the technology of information service is based on a statistical study of the semantic and morphological laws of language. Probability methods of research have long attracted the attention of linguists, but only in the most recent time have these methods found practical application. This will apparently lead to further stimulation of theoretical work in this direction.

3. DOCUMENTATION WITH THE AID OF PUNCHED CARDS

The plethora of publications makes it necessary to try and develop a system of documentation that would permit rapid extraction of exhaustive information on a large number of logically formulated questions. Such a system of documentation is possible only with the aid of tables with multiple entries. Multi-dimensional tables can be constructed with the aid of punched cards. Each punched card can be considered as a point in a multi-dimensional space, and a stack of punched cards can be considered as a table in multi-dimensional space with a multiplicity of entries.

Figure 8 shows a standard punched card with internal perforation. This card measuring 83. × 188 mm has 80 columns and 10 horizontal rows, numbered from 0 to 10. In addition, it is also possible to punch on the card the two upper unnumbered rows. The information is entered on the punched cards with the aid of previously developed codes. The punched cards are sorted out by specified properties by special machines, at a rate of 250 — 650 cards per minute.

Nalimov and Neporent[15] have considered the ap-

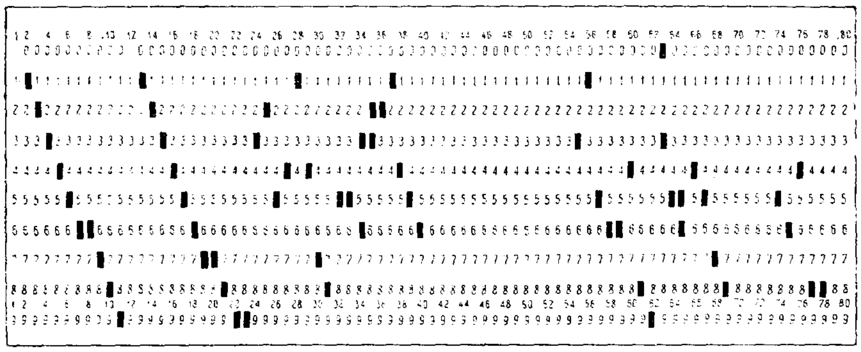


FIG. 8. Standard punched card with internal perforation.

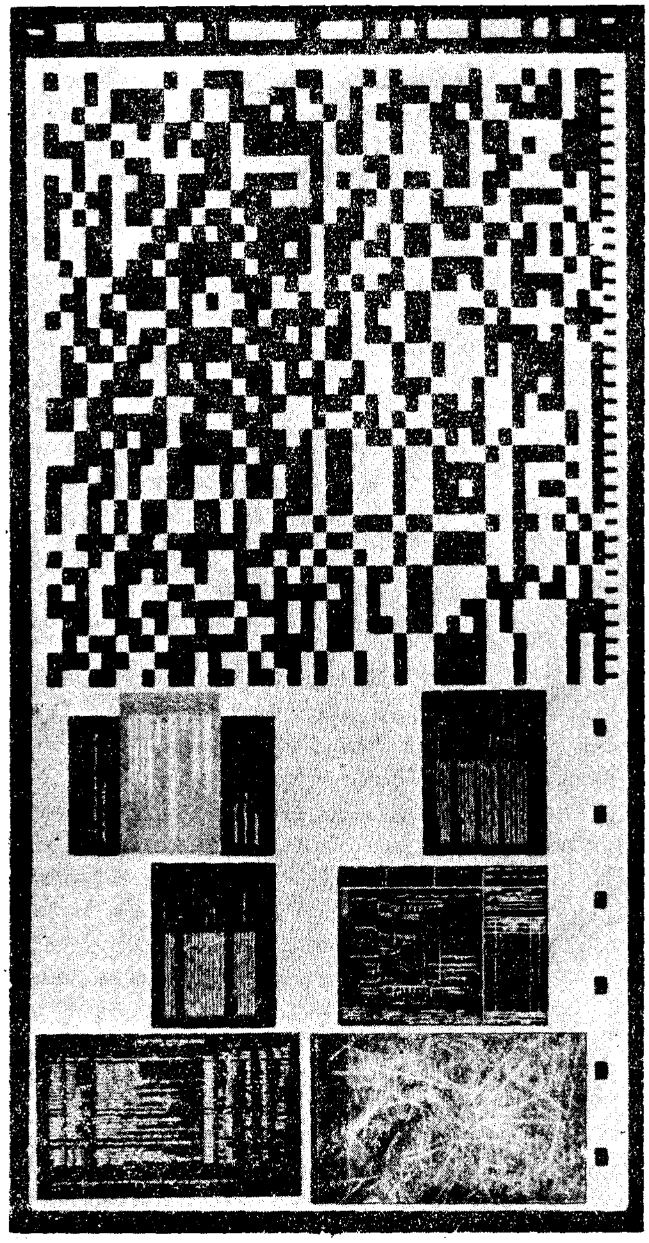


FIG. 9. The Kodak Minicard magnified 7 times. The upper portion of the film carries the coded information, and the lower one information represented in the ordinary manner (abstract, diagrams, map, etc.).

plication of punched cards for documentation in molecular spectroscopy. On the left part of the punched cards are coded the positions of the absorption bands. For example, if the number 4 is punched in the third column, this means the presence of an absorption band at 3.4μ. All the absorption bands that differ from the background by 20% or more are punched on the card. In the right half of the punched cards information is coded on the chemical composition of the molecules, on their structure, and certain physical properties of the compound. For example, column 44 contains information on the fragments of the structure that contain nitrogen, column 48 contains information on the fragments containing nitrogen and oxygen, etc. Punched cards can be used to solve many informational problems: establish a correlation between the absorption spectra and the structures of the molecules, find a group of substances that have a set of specified spectral properties or a group of spectra corresponding to specified elements of molecular structure, establish the composition of an unknown mixture from its spectra, etc. Beukelman[16] describes the use of punched cards with internal perforation for the identification of unknown substances by powder x-ray diffraction analysis. On the left half of the card are coded the interplanar spacings corresponding to the diffraction maxima, while the right half contains information on the chemical composition of the compound.

A unique modification of punched cards is the Eastman Kodak Minicard (U.S.A.) (Fig. 9) and the French Filmorex system, in which the information is recorded in a binary code of black and white dots. Together with the coded information, the card carries an abstract written in the ordinary manner, the diagrams, figures, etc. The photoelectronic selector of the Filmorex system reads and selects cards at a rate of 36,000 an hour; in the Eastman Kodak Minicard system the cards are sorted at a rate of 60,000 an hour.

In the foregoing examples, each punched card served to document information pertaining to some particular substance. In some systems the punched card is used for documentation pertaining to some other single feature or property. For example, Wachtel[17] describes a documentation system for nuclear physics, in which each card is used to document information on nuclides possessing some one property (stability, half-life, character of decay, natural radioactivity, etc.). Two sets of differently colored cards are used in this case, one for light and the other for heavy nuclides. Each nuclide is assigned a fixed position on the card and this position is punched only when the nuclide has the property designated by the card. Figure 10 shows a card with punches shown for all radioactive nuclides with 77 or more neutrons which decay with half-life greater than 100 years. Such a system of documentation is exceedingly simple and is convenient because the set of cards can be readily expanded to cover any new property that may be desired to be included in the indexing system.

The "Peek-a-Boo" system, in which each card is used to index some single concept, is extensively used. A total of 1800 small holes can be punched on a card measuring 20 × 30 cm. The holes punched in the cards correspond to the num-

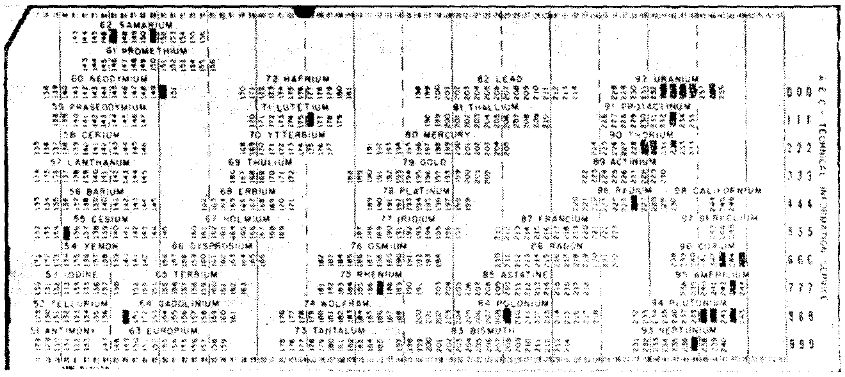


FIG. 10. Punched card used to document information on nuclear physics. The card shows the isotopes of heavy nuclei (with more than 77 neutrons) and the punched sections correspond to those isotopes, whose half-life exceeds 100 years.[17]

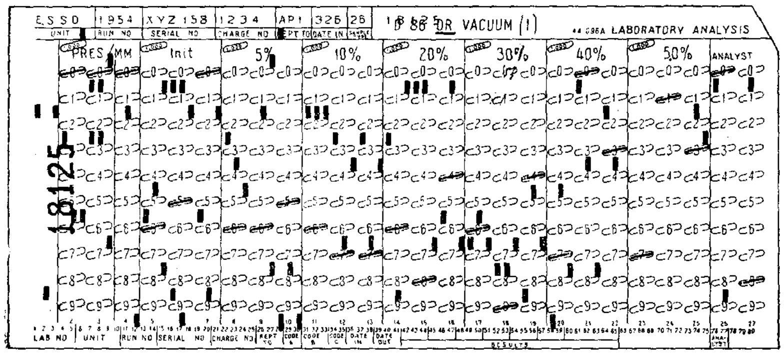


FIG. 11. Specimen of punched cards used to code and disseminate information in a large analytical laboratory. The analyst records the information on the card by darkening with a pencil the portions corresponding to the code. These portions are then automatically punched.[18]

ber of abstracts containing the term that is being indexed by the particular card. If the number of abstracts is greater than 1800, a second series of cards is introduced, etc. The abstracts are compiled in telegraphic style and the information contained in them is transferred to the punched cards with the aid of a special dictionary, frequently containing several thousand terms with cross references for synonyms. When searching for the information, the cards corresponding to the combination of concepts with which the problem is formulated are picked out. The selected cards are viewed in transmitted light or are scanned by a light beam. The holes common to all the cards indicate the numbers of the abstracts that contain the information on the particular problem. This system of documentation is used at the U. S. National Bureau of Standards to index information on instruments and measurements.

Punched cards are conveniently used to disseminate information within a laboratory. For example, the analytical laboratory of the Esso Standard Oil Company[18] uses the punched cards shown in Fig. 11. When a sample is received by the laboratory, its description, number, date of arrival, and group of analysts charged with the analysis are coded on the punched card. The analyst records his results, the conditions of the analysis, the accuracy, etc., directly in the laboratory. The punched cards are filled in by the analyst by darkening specified numbered areas, arranged in 26 columns, with a pencil that produces a current-conducting mark. Such a system of filling in the punched cards eliminates all preliminary "draft" notes, which consume much time. The filled-in card is sent to the tabulating room, where it is punched with a special machine provided with feelers and contact brushes that sense the darkened sections. The punched cards are then used in a tabulating room to print various types of summaries as required for the control of a large laboratory. These punched cards make it possible to group the analyses by various properties, such as the number of the sample, description of the sample, the conditions of the analysis, the results of the analysis, etc. Such an organization of the work facilitates subsequent statistical processing of the files, permits ready monitoring of the accuracy and correctness of the laboratory work by repeated analysis of the samples and of the standard specimens that are analyzed together with the current samples, etc.

Along with punched cards for internal performation, extensive use is also made of punched cards with external perforation. Figure 12 shows one such punched card. Certain systems use punched cards with only two rows of perforations, cards of other systems have ten rows each on one or two sides, etc. The coded information is recorded on the punched cards by cutting slots such as shown in Fig. 13. The cards are usually sorted manually, with the aid of rods: the slotted cards are lifted from the common stack. In some cases simple mechanical devices are used, permitting a sorting speed up to 20,000 per hour. In addition to the

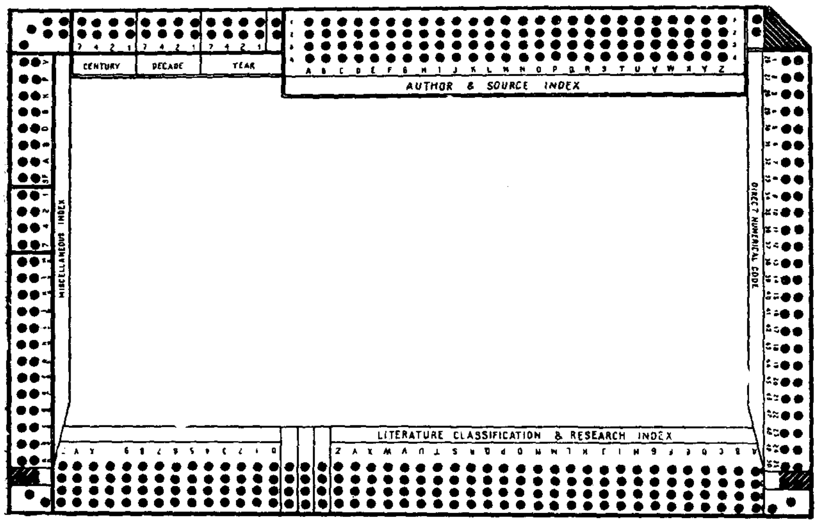


FIG. 12. Specimen of punched card with external perforation. The central (unpunched) portion can be used to record uncoded information – text of an abstract, graph, etc.[19]

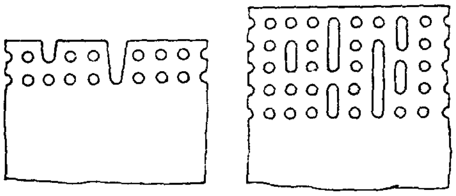


FIG. 13. Recording of coded information on punched cards with external perforation by cutting slots of various shapes.

coded information, such cards can also carry ordinary information: graphs, diagrams, text of the abstract, etc. This system of documentation is ordinarily used when not more than 10,000 cards are to be filled in.

Punched cards with external perforation are used for documentation of molecular spectra in the DMS system (Great Britain, West Germany),[15] in mass spectroscopy (where the molecular weight, boiling temperature, the elements in the compound, or the ion masses are coded), in nuclear physics, where one card can list all the properties of a given isotope.

In addition to the foregoing applications, the following punched-card applications are found in foreign references:

1) Biology, where punched cards are used more than in any other field. By way of example let us indicate that the Dow Chemical Company coordinates the work of biologists and chemists by means of a system of punched cards with internal perforation. A special code is used to record on the cards the results of 75 different tests for 11,000 substances. Approximately 200,000 punched cards have been used to store this information.

2) Astronomy. The Lick Observatory of the California University uses punched cards with internal perforation as an auxiliary device, for the compilation of a catalogue of binary stars. All the measurements of the binary stars for the northern hemisphere, either published or in manuscript, are recorded on more than 80,000 punched cards.

3) Meteorology. In 1948 the British Meteorological Service began indexing data on the upper layers of the atmosphere on punched cards with internal perforation. The observatories record their data on punched cards, which are then sent to a central organization for machine sorting, statistical treatment, tabulation, etc. It was calculated that documentation of all the information received from the British Isles requires more than 130,000 punched cards annually.

4) Geology. In this case the perforations correspond to geological objects and epochs, geographical regions, and to dates of readings or publications.

5) Inorganic chemistry. The Gmelin Institute of Inorganic Chemistry uses punched cards extensively for the preparation of its handbooks. Punched cards with external perforation are used to code articles by authors and to index materials by particular subjects (if the number of papers in this field ranges from 500 to 5000). The "Peek-a-Boo" system is used to code inorganic complexes. Punched cards with internal perforation for machine sorting are used in coding of information on the composition of compounds and minerals. By using this system it is possible, for example, to pick out all the minerals that contain germanium, all the inorganic compounds containing phosphates, etc.

6) Organic chemistry. This problem will be considered in detail in the next section.

7) Library practice. Punched cards are used as auxiliary means in the preparation of material for cataloguing, recording the flow of literature in inter-library exchange, the dissemination of infor-

mation containing the receipt of books by associated libraries, statistics, accounting, etc.

8) Bibliographical punched cards. The abstract is printed in the central portion of the punched card, and the contents of the abstract and bibliographical data (first letters of the author's last name, name of the journal, date of publication, etc.) are coded on the edges of the card. Such bibliographic systems are used in many branches of science and technology.

A detailed description of various systems of punched cards, their technology and use, the theory of coding, and on the fields of their application are given in a 1958 monograph edited by Casey, Perry, Berry, and Kent. The bibliography in this book contains 677 entries.[19] In the first edition of this book, published in 1951, the bibliography listed merely 276 titles; so large an increase in the number of publications within seven years is evidence of the great significance assumed by punched cards in information service.

Certain experiments with the application of punched cards have also ended in failure. At the end of 1947 the library of the Harwell Atomic Energy Research Center began indexing literature on a large set of problems on punched cards with internal perforation. After two years of experimentation, this system of documentation was abandoned.[20,21] The difficulties encountered here were, on the one hand, of technical nature, due to insufficient organization of the work, and, on the other hand, of fundamental nature. In information search, the scientist cannot always formulate the problem clearly, particularly if the search is in a field on the borderline of his specialty. Visual scanning of bibliographic cards, containing abstracts or annotations, is frequently a part of the creative process. Such a search may give the researcher an answer to a different problem, one which was not raised; this cannot be done in machine search with the aid of punched cards.[22] The use of punched cards for information dissemination is limited apparently to individual problems that arise wherever there is a large amount of material of a single type, and the search strategy can be formulated beforehand during the documentation.

4. DOCUMENTATION IN THE FIELD OF CHEMISTRY

Recent success in the mechanization of search for a few of the important types of chemical information is of undoubted interest to all allied sciences that deal, to some extent or another, with research on the properties of materials, and also to the general theory of documentation methods. The first applications of modern high-speed electronic computers were for the solution of certain information problems in chemistry.

The need for using complex computing devices is dictated here, on the one hand, by the tremendous quantity of factual material and, on the other hand, by the fact that in solving one of the principal information problems in chemistry (search for compounds by structural features) one is faced with a nonlinear problem.

Among all the branches of natural sciences, chemistry occupies one of the first places with respect to the quantity of accumulated information. This situation is due above all to the essentially empirical character of chemical information, i.e., to the lack of sufficiently reliable deductive rules with which to obtain the necessary information from certain relatively few principal facts. This is equivalent to a lack of sufficiently effective means of compacting the chemical information, and leads in the final analysis to a situation where in chemistry information worthy of storage is obtained by less laborious research work than, for example, in various branches of physics.

The foregoing can be readily illustrated by the following example. If one performs a physical experiment, consisting of measuring the characteristics of a certain electrical circuit, made up of a definite number of known elements such as resistances, capacitors, etc., it is obvious that the question of advisability of storing the data obtained does not arise. All the measurement results could be predicted from simple physical laws. The organic chemist, to the contrary, can obtain a new chemical compound, such as a salt, for example, by mixing any one of the numerous organic bases with any one of the no less numerous organic acids and performing a few simple operations. It is important to note here that all the physical constants of the resultant salt, beginning with its density, melting point, coefficient of refraction, etc. are of undoubted interest from the point of view of storage and dissemination, inasmuch as at the present time there is no possibility whatever of deducing this information from the values of the constants of the initial substances. Even the simple process of mixing two liquid-phase chemical compounds can be characterized by a whole series of such quantities as, for example, the heat of mixing, the shrinkage coefficient, the vapor tension over the mixture, etc., none of which can be calculated. By virtue of this circumstance, this simple process may be a source of a large amount of scientific information, intended for storage and dissemination.* We

*We do not consider here the usefulness of such information.

also note that the mixing of several components, none of which has insecticidal properties, does not preclude the possibility of obtaining a useful insecticide.

The most convincing evidence of the degree of saturation of chemistry (in particular, organic chemistry) by specific "incontractible"* information is the number of individual objects of research in this field, i.e., the number of chemical compounds described at the present time, which is not less than 600,000 (reference 23). This number increases at a rapid rate because not less than 20,000 new compounds are reported annually in the chemical literature. Even more numerous are the processes of transformation of chemical compounds from one into another, i.e., there is a particularly large number of known chemical reactions, the number of which reaches many millions.

It must be noted that these imposing figures do not mean at all, as one might think, that the corresponding fields of knowledge are well advanced. Disregarding the technical difficulties encountered in the production of micromolecular compounds in pure form, it can be stated that the set of most of all possible chemical compounds is infinite if for no other reason than that there is no limit to the number of atoms in the molecules of the compounds. However, even in the case of molecules with relatively few atoms, it is easy to verify that the number of presently known chemical compounds of a given type is only an insignificant fraction of the total number of possible similar compounds. For example, the total number of compounds with a molecule described by a common formula $C_{11}H_{10}N_2O_2$ is 125 (from Beilstein's handbook),[24] while even the lowest estimates of the expected number of compounds of this formula amounts to several times 10,000.

The great volume of information encountered in chemistry caused the problem of the development of special means and measures to facilitate the accumulation and dissemination of information to be raised in this branch of science at a relatively early stage, and led to the establishment of the first abstract journal in the history of science, "Chemisches Zentralblatt," in 1830. The same circumstance stimulated the development of a classification system for chemical abstract journals in much greater detail than is used in other abstract journals. For example, the annual index of Chemical Abstracts takes up approximately 20% of the total annual volume, and has approximately 3 or 4 times as many pages as the corresponding index of Physics Abstracts.

The highly perfected indexing system makes it relatively easy to solve the important one-dimensional problem of searching for information on a given chemical compound. This problem is of practical significance both to chemists and to a large circle of researchers engaged in the study of physico-chemical, physico-biological, and other properties of substances.

The difficulties that are nevertheless encountered in the solution of this problem are due to the absence of a universal and unambiguous system of chemical nomenclature for compounds, particularly organic compounds. Along with the names that characterize in one manner or another the molecular structure of a compound, chemistry makes use of many various names of historic origin, not connected with the structure of the compound, and different authors use in various places a great variety of names for one and the same compound. On the other hand, the individuality of the majority of organic compounds is described uniquely by the structural formula, constructed in accordance with classical representations of valence, or, where necessary, by supplements that take into account the stereochemical representations. This naturally led to the idea of using the universal physically-interpretable language of structural formulas for nomenclature of the compounds. It was necessary to construct for that purpose a method of unambiguous writing of the structural formula in the form of linear sequences of symbols, i.e., the form of a so-called linear code. Many systems of coding structural formulas were proposed during the past decade,[25] the best known of which are the systems of Dyson,[26] Wisswesser,[27] and of Gordon, Kendall, and Davison.[28]

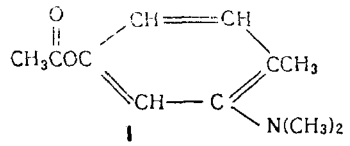
Figure 14 shows the structural formula of 2-dimethylamino-4-acetoxytoluol (I) and the linear codes corresponding to this formula in these systems.

Although these coding systems were developed initially to be used in ordinary indices, they have not yet been used for this purpose in practice.

While such codes are unwieldy from the human point of view, their use is essential for mechanization of information search, for both in the case of mechanically sorted punched cards and when electronic machines of existing types are used, the recorded and processed information must be a linear sequence of symbols.

*We do not have in mind here the non-contractibility of information in the classical sense, i.e., the impossibility of representing the observed phenomena by functional relationships. The use of statistical-theory methods and mathematical statistics for contraction of the information will be considered in the next section.

FIG. 14. Example of linear codes of a structural formula for the organic compound marked I.[25]

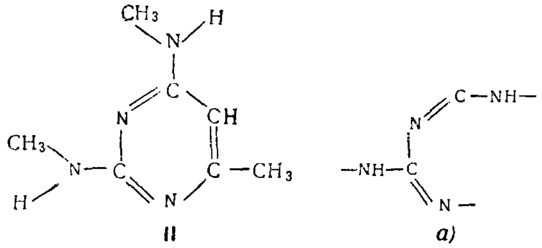


I

Dyson: B6.C.[X.C₂].4.N[C]₂,2

Wisswesser: 1N1 R1eOV

GKD: $J^1C^3OC^3MMCJ^1C^3N^3J^1J^1MX4$



II

a)

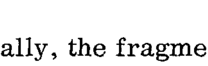
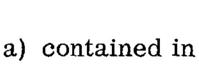
$H^1N^3J^1C^3NC^3J^1MC^3NX3N^3H^1J^1$

$H^1N^3R^1C^3R^1NC^3NR^1N^3H^1R^1$

FIG. 15. Example of notation for fragment a) contained in structure II. The linear codes are in the GKD system.

By using any coding system it is possible to place on a punched card the codes for the compounds and alongside the corresponding information on certain physical constants such as the position of the maximum in the absorption spectrum of the compound, information on the presence or absence of various types of physiological activity in the compound, etc. It is then possible to find by mechanical sorting all the compounds with a specified value of a physical constant or a specified set of physiological properties. The values of the physical constants can be indicated in the search prescription either exactly or in the form of some permissible ranges. The performance of such multi-dimensional searches, which are not feasible with ordinary single-dimensional indices will interest not only a chemist, but also the entire group of researchers mentioned above, who deal with properties of matter. The most important of the problems solved by these means is that of establishing a correlation between two or several properties of the compound. On the whole, the scope of all the problems described here corresponds in its entirety to everything mentioned in the preceding section concerning punched-card documentation methods for other branches of science.

In chemistry we deal, however, with one of the most important unique properties used to describe chemical compounds, which nevertheless cannot be used in general for search with punched cards. The property we have in mind is the presence of a definite structural element in the molecule, or more accurately, the presence of a certain specified structural fragment in the molecule. By structural fragment is meant a bonded substructure, contained in the structural formula of the compound. Examples of structural fragments are the fragment $>C=O$, contained in $(CH_3)_2C=O$ and in the foregoing structure I (Fig. 14), as well as the other fragments contained in I

 and 

or, finally, the fragment a) contained in structure (II), both shown in Fig. 15.

We can readily verify by comparing the code of fragment a) with the code of structure II, which contains this fragment, that the presence of this fragment cannot be established by a mere "scanning" of the code of the structure, since the code of the fragment is not contained as a linear part of this code. This shows that when compounds are searched for the property of a given fragment of the structure, the codes for the compounds must be subjected to a more complicated logical treatment, the realization of which is possible only with the aid of electronic computers with program control.

It must also be noted that the set of all possible structural fragments, like the set of all compounds is infinite and that the prescriptions for search may include arbitrary structural fragments.

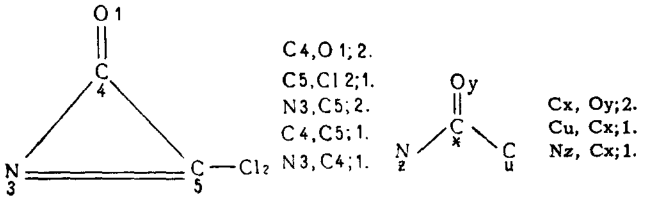
The reason why the scanning of classes of compounds for structural fragments is important is that one of the principal means of establishing chemical laws lies in ascribing to certain structural fragments the roles of carriers of definite specific types of properties, particularly physical properties (for example, spectral ones), chemical properties, i.e., the reacting ability, or biological properties (for example, physiological activity*). Therefore, in the course of establishing such laws, it becomes frequently necessary to search for classes of compounds that are specified by different structural features, in order to confirm or disprove certain hypotheses regarding the existence of a correlation between the determined properties and the structural fragments.

In view of the great practical significance of these problems, they are solved partially with the aid of punched cards, by choosing certain structural fragments of relatively small extent, considered most important from the point of view of the particular problem, and considered further as

*For example, the structural fragment a) shown in Fig. 15 is considered to be the carrier of anti-malaria activity.[29]

common features, the presence or absence of which is coded on the punched card in the ordinary manner. This is the procedure, for example, in the case of the punched cards, described in detail in reference 15, intended for the documentation of molecular spectra. For various other purposes, several versions were proposed for coding the structures of the compounds on punched cards in the form of a linear aggregate of definite structural elements.[30,31,31a] The first of these systems was extensively tested in practice at the Chemical-Biological Coordination Center (U.S.A.) and yielded favorable results in the establishment of correlations between biological activity and structure; the second one, developed under the leadership of Pietsch for inorganic compounds, is used by the Gmelin Institute (West Germany). It must be noted that the codes of the compounds, based on the foregoing principle, cannot serve as unambiguous linear ciphers for structural formulas, since one and the same code will correspond to several different structures (for example, made up of the same "basic" structural units, but differently bonded to each other).

Without denying the usefulness of such an application of punched cards for searches based on structural features, it is nevertheless easy to show, even with this example of documentation of molecular spectra, that this approach to the solution of the problem is inadequate. It is known, for example, that the vibration frequencies of the carbonyl group in different compounds vary over a sufficiently wide range (from ~1975 cm^{-1} to ~1550 cm^{-1}), and therefore only a very rough correlation can be established between the structural fragment $>C{=}O$ and the position of the corresponding maximum. The correlations can be refined substantially by considering such structural fragments as $-C({=}O)CH_2C({=}O)-$ (β-diketones), $-C({=}O)CCl<$ (α-halogenketones), or fragments of even greater extent, which characterize also classes of compounds such as "cyclic-5-member ketones," "β-lactames," etc. Much more accurate values are obtained[32] for the vibration frequencies of the C=O groups, contained in each of these groupings. As is known, the frequency of the infrared absorption spectrum bands is related to definite mutual arrangements and interactions of certain structural elements with others, and it is therefore quite natural that in order to investigate effectively the dependence of these spectra on the structures of the compounds it is necessary to be able to search for a great variety of types of structural fragments. From this point of view, the use of high-speed electronic computers to solve, in general form, the problem of search for classes of compounds by structural features, is undoubtedly of great practical interest.



III

FIG. 16. Examples of notation for a structure and a structural fragment in the form of bond tables.

The most suitable for work with these machines are the so called topological coding systems for structures of compounds. The distinguishing feature of such systems is that the structural formula is considered as a set of numbered atoms (or certain larger structural blocks) and of bonds between them, and the code indicates in principle only the pairs of the interconnected atoms.

Figure 16 shows a very simple structure, III, and one of the simplest methods (in principle) of topologically encoding it in the form of a "bond table," i.e., in the form of pairs of numbers of interrelated atoms, with indication of the type of bond between them (1 — simple bond, 2 — double bond, etc.). The atoms are numbered arbitrarily, subject only to the condition that all the atoms of the structure have different numbers. It must be noted that such a method of recording the structure, presented here in readily visualized tabular form, can be changed without difficulty into a linear sequence of symbols. It is easy to verify that the structure can be represented uniquely with such a notation and that mutations of the notation, such as interchanging two places in the numbered atomic symbols contained in the same row of the table, or interchanging any two rows of the table, do not change the structure obtained with this notation. Figure 16 shows also a structural fragment contained in structure III, and its notation in the form of the bond table, whereby the atoms of the fragment are assigned numerical symbols x, y, z, etc., instead of arbitrary specific numbers. To establish the presence of a definite fragment in a certain structure, as called for by the search prescription, it is enough to find such a mutation of the notation of the structure and such a set of specific values x, y, z, etc., for which the notation of the fragment is found in the form of a linear entry in the recorded structure. If after performing all the possible mutations in the notation for the structure and after assigning all possible values to the numbers of the atoms of the fragment, no such letter by letter correspondence between the portion of the recorded structure and the notation for the fragment is observed, we can conclude that the sought fragment is missing from the considered structure.

An atom-by-atom version of a topological code, similar to that described above, is used in the experiments of Ray and Kirsch[33] at the U. S. Patent Office, while a block by block topological variant is used by Opler and Norton[34,34a] at the Dow Chemical Company. In both variants, the atoms or the corresponding blocks of the structure can be numbered arbitrarily.

For each concrete structure notation, search programs have been developed, based on the performance of mutations of the type described, but so designed as to reduce to a minimum the number of necessary scannings of all the possibilities.

Opler and Norton have recorded on magnetic tape, used as an external machine memory, the numbers of the structural formulas of the compounds, in the form of sequences of codes together with serial numbers, brutto-formulas, and names of the compounds. The magnetic tape was connected to an IBM-704 (or 701) computer, capable of performing approximately 2.5 million elementary operations (additions, subtractions, comparisons) per minute. The search prescription, in the form of a structural fragment coded in the same way, was recorded on a punched card, which was inserted in the machine together with the standard program for the search. After pressing the starting push button, the machine started to scan all the entries on the magnetic tape in sequence and to compare the codes of the compounds with the code of the specified structural fragment. By means of a series of mutations, which included, on the average, approximately 1200 elementary logical operations for each scanned notation, the computer established the presence or absence of the sought fragment in the given compound and, in the final analysis, printed the serial numbers or names of the compounds that satisfy the search prescription. All the operation from the start of the input of the question to the production of the printed answer took approximately one minute for a search among 10,000 compounds.

In a similar investigation by Waldo and De Backer of the Monsanto Chemical Company[35,35a] the IBM-701 computer was used not only for the search operations, but also for partial automatization of the coding of structural formulas. After the chemist has recorded the structural formula on a sheet of coordinate paper, in accordance with a certain set of standards, all further operations on the translation of this record into machine language and on the verification of the resultant code are carried out only by technicians — non-chemists, with a substantial utilization of the computer itself. In these investigations, information obtained directly from the laboratory logs of the various divisions of the research laboratory was stored on magnetic tape. One division furnished information on the structure and on the physico-chemical properties of newly-synthesized compounds, while another division furnished information on tests of the biological activity of the compounds. Based on this store of information, the computer automatically prints "review reports" on problems of the following types: for what type of physiological activity is the individual compound tested, what compounds give favorable results when tested for a certain type of physiological activity but negative results with regards to another type of activity, what type of biological activity is common to all compounds containing a definite specified structural fragment, etc. In this case the machine printed not only the names or serial numbers of the compounds but also their structural formulas, albeit in a somewhat unusual "quadratized" form. Figure 17 shows by way of an example the structural formulas of phenol and potassium acetate as printed by the computer. In these formulas the hydrogen atoms have been omitted, and the numbers between the atoms indicate the type of bond.

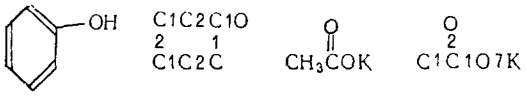


FIG. 17. Structural formulas of phenol and potassium acetate in the usual form and in the form printed by the computer in the Waldo and De Backer experiments: 1 – simple bond, 2 – double bond, 7 – ionic bond.[35a]

The first applications of high speed electronic computers for the search of chemical information have disclosed the great potential of such devices both as regards high speed in processing a considerable bulk of information, and as regards the variety of the performed search operations and operations on the transformation of the information. It must be borne in mind that the same computer machine can be used not only for a most varied type of search operation, but at the same time to solve purely computational problems. Based on the data of reference 36, approximately 100 large electronic computers are employed by the chemical industry of the U.S.A.; they are used for commercial computations and approximately every other computer is used to solve at the same time various problems of scientific nature, including problems in information search.

While general-purpose computers are quite suitable for experimental work in this field, specialized information machines are necessary to obtain serious practical results. The principal distinguishing feature of these machines is that they must have large long-time high-speed machine memories. The magnetic tapes used as memories in modern computers have a very serious shortcoming, the need of mechanical motion (rewinding the tape) when reading the information, which limits the reading speed considerably. It is therefore very important to develop machine memories capable of long-time storage of large volumes of information and of operating without mechanical motion, by using purely electric counting principles. A machine memory of this type was developed under the leadership of Prof. L. I. Gutenmakher at the Electric Simulation Laboratory of the All-Union Scientific Research Institute, Institute for Technical Information, U.S.S.R. Academy of Sciences.[37]

The availability of such a memory together with the modern technical feasibility of production of specialized devices capable of performing a great variety of logical operations, makes it possible to consider the creation of a large informational-logical machine for information on chemical compounds and chemical reactions. In addition to technical premises, an important role is played here by the fact that chemical structural formulas, readily translatable into linear codes, are a very convenient and almost universal language with which to describe the properties and behavior of chemical compounds. In this connection it becomes possible to simulate, with the aid of electronic-informational logical machines, certain aspects of "chemical thinking," in particular the solution of such problems as the use of "chemical analogies" recorded in the machine memory to obtain likely methods of synthesizing a compound not yet produced.[38]

5. REPRESENTATION OF RESULTS OF EXPERIMENTAL RESEARCH IN A FORM SUITABLE FOR CODING

In chemistry, experimental physics, and their technical applications, the foregoing documentation methods can be effective only if the measurement results are recorded in a compact form suitable for coding. The information can be contracted by using methods of modern mathematical statistics. The measurement results can usually be interpreted as a set of values of a random quantity; cumbersome tables can then be represented in a compact fashion with the aid of the distribution parameters. For many-dimensional random quantities, linear graphs can be replaced by regression and correlation coefficients. In complicated experiments included in an investigation of the action and interaction of many factors, the research results can be processed with the aid of dispersion analysis, and then the cumbersome numerical material can be represented in the form of dispersions and quantities that estimate their statistical significance.

Many experimenters still hold to the widespread incorrect opinion that mathematical statistics is applicable only to a large volume of numerical material. Modern mathematical statistics can be applied even to the processing of small samples, consisting of only three or sometimes even two measurements. Here, naturally, the element of uncertainty of the estimate of the measured quantity is increased, and this too is accounted for by methods of statistical analysis and is expressed in a form suitable for coding. The possibility of employing statistical methods for contraction of information in samples of any size allows us to represent experimental material in a strictly standard manner, regardless of its volume.

In laboratory work it becomes necessary to cope with the fact that certain regular phenomena, of no importance from the point of view of the experimenter, are superimposed on the random processes. Similar situations arise also in research in engineering, agriculture, etc. These regularities, of no importance to the experimenter, can always be eliminated with the aid of a special procedure — randomization, which makes it possible to plan the experiments in such a way as to insure the random character of variation of all the uncontrolled factors.

Experiment shows that the application of mathematical statistics to contraction of information becomes effective only in that case when the statistical methods are used not only for the final processing of the material, but also for planning the experiment (choice of experimental procedure, arrangement of material, randomization of the experimental conditions, choice of number of parallel measurements, etc.).

Mathematical statistics is beginning to be used also for compact representation of the information contained in the tremendous archives that have accumulated in research institutions and plant laboratories. As a rule, these materials have remained unused, since they are too bulky to be disseminated in unprocessed form through ordinary channels of scientific-technical information.

The idea of applying mathematical statistics to the contraction of information is due to R. Fisher.[39]

It is known that the founders of the English and American statistical school held to Machian positions and based their statistical problems by starting from these positions. Now, in connection with the ideas of cybernetics, one can approach the problem of contraction of information from principally different positions. The contraction of information must be considered not as a purpose in itself, but a certain intermediate stage in the cognition process. In many fields of research, and particularly in technical physics, one frequently deals with so complex a situation, that a single investigation still does not permit penetrating into the nature of the investigated phenomenon. If the result of the research is represented compactly and the indeterminacy element connected with the complexity of the experimental conditions is estimated, each investigation can be considered as part of a certain collective work. The information thus presented can be recorded in the long-time memory of the machine and compared with the results of other investigations. Such a comparison of a large number of investigation affords a deeper insight into the nature of the phenomena. This idea can be illustrated by the following example: in emission spectral analysis the problem of the influence of a third element is considered in hundreds of investigations. Not one of these investigations taken by itself permits penetration into the essence of this phenomenon. It is impossible to draw any general conclusions from a comparison of these investigations, since the results of their researches are represented in a non-standard manner and the element of indeterminacy, connected with the results of the investigations, has not been estimated. This being the situation, the use of information machines cannot lead to any substantial benefit. It is impossible to record in the memory of the machine the results obtained in these investigations. The only thing that the machine can do here is to select investigations that cannot be compared with each other on the basis of a given subject matter.

There are many available guides to statistical methods of planning experiments and processing experimental results.[40,41,42,43,44,45,46] There is need for strict standardization of the methods of representing the experimental results in accordance with the features of the particular field of research.

Simultaneously, in many cases it also becomes necessary to standardize the methods of performing an experiment. For example, documentation of molecular spectra can be effective only if the apparatus employed, the methods used for the measuring and representating the spectra, or the criteria for estimating the purity of the substance are sufficiently strictly standardized. The information service will be forced to impose its requirements on the experimenters.

6. DESCRIPTOR INDEXING

The examples of information search systems mentioned in Secs. 3 and 4 are only particular cases of numerous different documentation systems, developed essentially during the past ten or fifteen years, when a very great expansion has taken place in research both on the creation of new search systems and on different practical applications of these systems. A new field of theoretical and practical activity has come into being, known as documentalistics. The purpose of documentalistics is the development of methods of most effective utilization of information, stored in various collections of documents. The term document is used in documentalistics for all possible types of information records, particularly journal articles, books, reports, notes in laboratory logs, etc. During the last few years documentalistics has been acquiring more and more the character of an independent scientific discipline, striving towards a development of a general theory of information search systems with utilization of the formalism of mathematics and of mathematical logic. Many scientific journals devoted to the problem of documentalistics are now being published, the best known of these is the journal American Documentation, published since 1950.

Under the influence of cybernetic concepts, a new scientific discipline is now coming into being, with a broader purpose: the use of cybernetic machines to take over various types of intellectual labor amenable to algorithmic description, in particular, the search for the scientific information necessary to solve a definite problem, and also various processes of transformation of scientific information (translation, contraction, development of solutions, etc.). The need for such a discipline is dictated, as shown in the first section of this article, by the exponential growth of science.

Inasmuch as the accomplishments of documentalistics are of substantial significance to this new still unnamed science, and also in consideration of the independent practical interest that attaches to its results, it behooves us to consider the principal trend of modern documentalistics. Mention must be first made of the so called descriptor indexing.

This trend in documentalistics, which deals essentially in problems of indexing, has been embarked upon by the group headed by M. Taube.[48,49,50]

They propose to adopt a so called coordinate system of indexing, whereby a set of single terms (uniterms or descriptors), to some extent chosen arbitrarily, are used for the compilation of a catalogue of documents. Descriptors are concepts, each of which is necessary, and the aggregate of which is sufficient, for the identification of a given document (within the framework of a definite collection of documents). For example, for an article on the subject "Equilibrium Composition and Thermodynamic Properties of Combustible Gases" the descriptors will be the following concepts: combustible substances, gases, calculation, fuel, momentum, pressure, temperature, entropy, heat content, adiabaticity.[51] As soon as the document is received, it is assigned a definite number, which is placed on all the cards of the descriptors, actually necessary for a complete identification of the given document. The card for each descriptor (which has its own distinguishing number) is divided into 10 columns, numbered from 0 to 9 inclusive. On this card are placed the accession numbers of all documents in which the given descriptor is included, and the number of the document is placed in that column, whose number coincides with its last digit. Within each column of the descriptor card, the documents are arranged in order of increasing accession numbers. The rough scheme of search for a document, identified with the aid of n descriptors, reduces to a comparison of the numbers on n descriptor cards in order to discover the numbers common to all these n cards. The procedure of such a comparison is facilitated substantially by the fact that the accession numbers of the documents on the descriptor cards are in increasing order. An example of a descriptor card is shown in Table II.

At the present time, the firm "Information for Industry" publishes in Washington two series of indices, based on the principle of descriptor indexing and known as the Uniterm index, published separately for U.S.A. patents in the field of chemistry and electronics. These indices are in the form of descriptor cards, gathered in a loose-leaf book arranged in alphabetical order of uniterms (that is, descriptors). In order to facilitate the comparison of the numbers under two different uniterms, each binding contains two identical copies of the index arranged in such a way, that two different uniterms can be viewed simultaneously.

TABLE II. Descriptor card, containing the numbers of those papers in which information is contained on spectra

Spectra									
0	1	2	3	4	5	6	7	8	9
120	1951	1432	5713	1064	2885	4576	1277	1278	1279
190	2451	2562	6263	2894	6025	6026	2877	5728	4489
	2471	4502		4524		6106	4107		5069
	6121	5662		5714		6506	4897		
	6421	6262				6636	5707		
							6247		
							6457		
							6517		

Work has been carried out on the clarification of the logical basis of the method of coordinate system of indexing. The sum total of this investigation is the conclusion that the logic of the coordinate indexing is the calculus of categories of modern mathematical logic, which is historically based on the Boolean algebra of properties.* Each indexed term generates a class of all these objects, the description of which requires the utilization of a given term. In this case the relation between the terms in the case of coordinate indexing can be described by means of corresponding category-analysis theorems in terms of the operations of joining, meeting, and taking of complements.

An entire series of experimental investigations has been carried out on the application of the ideas and methods of coordinate indexing to library search for information. Interesting results in attempts of this kind are described and explained, for example, by Brakken and Tillit.[51] They have

*Boolean algebra, sometimes called algebraic logic, is an application of mathematical methods to formal logic. Algebraic logic is a science extensively used at present in mathematics (in probability theory), in engineering (in the theory of electric relay-contact communication networks), in logic (in so called prediction calculus, where algebraic-logic methods are used to solve, for example, the problem of derivation of all possible consequences from a given system of premises, or else the problem of searching hypotheses from which certain predictions can be obtained). An elementary introduction to algebraic logic is given in reference 47 and, in the Soviet literature, in Prof. S. A. Yanovskaya's commentaries to the Russian edition of the book by Hilbert and Ackerman.[67] Boolean algebra is named after its creator, the well known British mathematician George Boole (1815–1864). It is in part a generalization of Aristotelian syllogistics, the theory of logical deductions.

The principal operations as defined by Boole are:

1. Addition, denoted by the symbol "+"; in the calculus of categories, the Boolean formula $x + y$ corresponds to the joining of categories x and y with exclusion of their common parts; in the calculus of predictions it represents the so called strong disjunction.

2. Multiplication, denoted by the symbol "·"; in the calculus of categories this operation corresponds to a meeting, and in calculus of predictions – to conjunctions.

3. Complement to unity, denoted $1 - x$; in the calculus of categories $1 - x$ denotes the complement to category x; in the calculus of predictions it corresponds to negation of x.

considered the problem of simplifying a procedure for library search for technical information from the point of view of the method of coordinate indexing. An IBM-701 computer was used for the information search, installed in an American experimental station for naval artillery in California. The objects of search were papers pertaining to problems in the development of naval artillery and published by various agencies in the U.S.A.

A person desiring to obtain information on a definite problem, proposes a list of key words (descriptors), on which basis the search will be carried out. The purpose of programming the search for the IBM-701 computer is mechanization of the search procedure and development of a daily program of library search. Descriptor cards, prepared in the form of punched cards carrying the number of the descriptor itself together with the numbers of the corresponding papers, are copied on magnetic tape in increasing order of the numbers of the descriptors themselves. Provision is made for the insertion and erasure of new descriptors or of new numbers of papers for each descriptor. The daily program, including up to 75 independent searches, provides that each interested person propose not more than 8 descriptors, defining the paper he needs. This group of descriptors defines the independent search. The descriptors of the searches, contained in the daily program (with the exception of repetitions) are transferred from the common tape to the working tape. In the compilation of the search program the authors have taken into account the average contents of the cards with allowance for their natural increase during the next few years. Approximately 9600 descriptors and 14,000 numbers of papers were recorded on tape. Up to 120,000 bits of information were stored on $2/3$ of the tape spool. Approximately 300 new papers were added monthly. The number of new descriptors, added each month was rather small. The library search program was subdivided into three consecutive independent steps: 1) insertion, 2) search, and 3) retrieval. Phase (1) requires approximately 6 minutes; the time of phase (2) varies with the number of descriptors in each search; its average duration does not exceed four minutes. Phase (3) consumes merely one minute.

7. ATTEMPT AT THE CONSTRUCTION OF AN ARTIFICIAL LANGUAGE FOR MACHINE INFORMATION SEARCH

The natural languages are not suitable for machine information, in view of their exceeding flexibility. The large number of homonyms, synonyms, and semi-synonyms, the dependence of the meaning of a word on so called semantic fields, which are produced by virtue of the complex interrelation between words, which depends on their position in the phrase and on the construction of the phrase — all this makes mechanized information search based on natural languages impossible or inefficient. Therefore the mechanization of information search depends substantially on the construction of a certain artificial language, which we shall henceforth call information language.

Information language is a means of compact, strictly unambiguous notation for the contents of a document in a form acceptable for mechanized information search. One of the methods of translating from a natural language into information language will be considered below. We must emphasize the difference between the concepts "information language" and "formalized language" (which approximates in the scope the concept of a formal system). Although every information language includes elements of formalization, nevertheless it is not constructed fully formally, i.e., it has no axiomatics or strictly fixed rules of formal-logical deduction. The first appeals for the creation of a formalized language are found in the papers by Leibnitz and Descartes.* The first hints of information language appeared in the papers of Hollerith[54,55] in 1884, in which he also predicted the possibility of applying mathematical logic to problems connected with information search.

Much work on both the theoretical premises and the practical creation and application of information language was carried out in the U.S.A. by Berry, Perry, and Kent and their colleagues.[56,57] In reference 56 they first raise the problem of revising the terminology, a problem considered as a necessary preliminary condition for the construction of information language. The need of a special analysis of terminology is dictated particularly by the fact that searching machines, based on electronic technology, are capable of identifying several sequences of symbols, but cannot interpret the values of words. The individual stages in this analysis are: 1) it is first established what type of object or concept is characterized by a given term, 2) the field of science that it belongs to is identified, 3) the terms are tentatively classified. A possible ver-

*The most recent pioneers in the construction of formalized languages for specific fields of natural sciences have been R. Carnap[52] and Woodger.[53] Reference 52 contains attempts of axiomatic exposition of the theory of kinship (in the sense of jurisprudence and biology), and reference 53 presents a formalized language for various divisions of biology, a language that makes extensive use of the accomplishments of mathematical logic and theory of sets.

sion of such a classification as applied to chemistry could be, for example, as follows: a) machines (ozonizer, spectrograph, etc), b) processes (absorption, neutralization, etc), c) materials (acid, alcohol, etc), d) attributes (magnetic, acetate, etc), e) concepts (free energy, entropy, etc).

The principal requirement imposed on an information language is that it be capable of expressing the essential aspects of the contents of the indexed and coded document in a form suitable for search with the aid of a machine. The contents of a document must be coded in a way as to make it easier to find during the information search the more essential data contained in the document, rather than the secondary data. In reference 56 an original solution was proposed for this problem, based on the so called organized designation of the contents aspects of the document. A compact code, based on mnemonics, is introduced. Use is made of the concept of semantic units (corresponding to the features of the coded term). For example, the semantic units of the term "thermometer" are the following concepts: "intended for measurement," "included in the class of instruments," "subject to the action of temperature," etc. Various relations exist between the coded term and its semantic units. In reference 57 it is shown that a list of such relations is not only finite, but can also be made strikingly small. With the aid of a very small number of relations, and a somewhat greater but still manageable number of semantic units, reference 57 represents approximately 30,000 scientific and technical terms in information language. The relations between the terms and their semantic units are subdivided into two classes: 1) analytic and 2) synthetic. By analytic relation is meant a relation existing between terms only by virtue of their definition or field of application. For example, the statement "the navy consists of ships" expresses an analytic relation ("consists of") between the term "navy" and the concept "ship." A relation which is not analytic is considered synthetic. Examples of synthetic relations are: the relation between the final and initial material in any reaction, the relation between the coded object and a reagent that contributes to the course of a given process, etc. The formalization of synthetic relations contains, in the opinion of the authors of reference 56, one of the principal difficulties that stand in the way of creating a satisfactory system of information language. For symbolic representation of semantic units use is made of the following 13 upper-case letters

$$B,\ C,\ D,\ F,\ G,\ H,\ L,\ M,\ N,\ P,\ R,\ S,\ T, \qquad (1)$$

and for analytic relations the 10 letters

$$A,\ E,\ I,\ O,\ U,\ Q,\ M,\ X,\ Y,\ Z. \qquad (2)$$

The semantic units are of the form

$$\mu \square \lambda\theta,$$

where μ, λ, and θ are some letters from the series (1), two of which must be different, while the empty place "□" is filled by some analytic relation. The following analytic relations were proposed:

A — the inclusion of an element in a class, or of a class in a class of greater volume. Example: M—TL is an abbreviated symbol for the concept "metal"; the expression MATL indicates that a certain element is included in the class of metals.

E — relation between the coded object to the material of which it is made. For example, the expression METL indicates that a certain object is made of metal.

I — the relation between a part and a whole. For example, M—CH is the symbol for the concept "machine"; the semantic units MICH indicates that the coded object is a part of the machine.

O — the relation of a whole to its part. For example, S—HP is the symbol for the concept "ship"; the semantic unit SOPH is used to code the concept "navy" ("consists of ships").

U — intended for. For example: M—SR is the symbol for the concept "measurement"; the semantic unit MUSR indicates that a certain coded object is a measuring instrument.

W — subject to action. For example: L—CT is the symbol for the concept "electricity"; the semantic unit LWCT indicates that electricity acts on the coded object. This semantic unit will be used in the coding of the concept "ammeter."

X — absence of a certain quality or characteristic. For example: H—DR is the symbol for the concept "water"; the semantic unit HXDR is used to code the concept "anhydride" ("lacking water").

Q — is used. For example: LQCT— part of the code for "electroplating."

Y — positive characteristic that does not coincide with any of the foregoing analytic relations.

Let us give an example of coding. The code under the concept "thermometer" is as follows:

$$MACH \quad MUSR \quad RWHT \quad 4X\ 002.$$

In this expression the symbol "4X" denotes that the code preceding it satisfies not only the contents of the concept "thermometer" but also the content of several other concepts, particularly the content of the concept "pyrometer" (the code of the latter will differ from the code of the thermometer by the lack of the symbol "4X"). The numerical termination "002" serves for further differentiation of the semantic code.

Let us consider a few other examples of coding by the method discussed in references 56 and 57.

1. Absorption — BASB 001.

B—SB is a semantic unit denoting the concept of absorption. The symbol A indicates that here the semantic unit is used to designate a property that is a part of this broader concept.

2. Absorption band BWSB GARP MYPR 98X 001.

B—SB, as already indicated, denotes absorption. The symbol W indicates that the band arises as a result of absorption. G—RP serves to denote the concept "collection." The symbol A indicates that we deal here with an object that is part of a collection (absorption band — part of the absorption spectrum). M—PR serves to denote property of matter. The symbol Y indicates that the absorption band is characteristic of the properties of the matter.

3. Absorption tower BUSB MACH 005.

B—SB,. as before, denotes absorption; the symbol U indicates that the absorption is considered here as a process for which the coded object is intended. M—CH denotes a machine or device. The symbol A indicates that we deal here with a certain object which is a part of the broad concept "machine." The numerical suffix "005" allows us to segregate the concept "absorption tower" from the broad class of absorption devices.

The information is inserted in the long-time memory of the machine as follows: for each published article an abstract is written in so called telegraphic style, and its contents is then translated into information language with the aid of the semantic units and several indices that serve to denote the synthetic relations that reflect the telegraphic grammar. For example, if the paper deals with chemistry, it is important to code not only the substances that enter into reaction, but also to note the initial and final products of the reaction and the substances that contribute to the course of the reaction. For this purpose use is made of the indices: KAY — initial materials, KWY — resultant substances, KQY — reagents contributing to the course of a certain process, etc.

Searches can be carried out on the basis of a single aspect (according to the code of the term, the code for a single semantic unit, or a combination of some semantic units with an analytic relation) or else on the basis of a series of aspects of the contents of the recorded document. In the latter case the searches are usually based on the logical product of semantic units, their logical sum, their logical difference, or a combination of logical sums, products, and differences. Let us assume, for example, that we need to find substances of class A, which simultaneously have the properties* B and C (shaded area in Fig. 18). In this case the search program can be entered in the form of a logical product (meeting) $A \cdot B \cdot C$. If we need to find the substances of class A which have properties B and do not have property C (the dotted area in Fig. 18), the problem is recorded in the form of the logical difference $A \cdot B - C$. Finally, let us assume that we search all the papers of all the investigations, in which at least one of the characteristics A, B, or C has been measured for a certain class of substances X. This problem can be written in the form of a logical sum $X \cdot A + X \cdot B + X \cdot C$. More complicated stated problems are also possible. The role of the logical-mathematical relationships in multiple-aspect searches will be considered in detail later in Sec. 8.

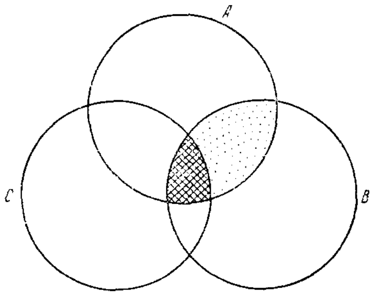


FIG. 18. Diagram of relations between properties.[56]

Much work has been done in the U.S.A. for a number of years on experimental investigation and development of information systems, based on the use of the language described above. The investigations have been performed with a slow-speed computer having electromagnetic relays as binary elements. Starting with 1959, the Center for Documentation and Communication Research of the Western Reserve University in Cleveland, headed by Perry, jointly with the American Metallurgical Society, started publishing machine information on metallurgical problems. Stored in the long-time memory of the machine are the contents of 25,000 articles on metallurgy, published since November

We use here algebraic logic (see footnote on page 654) and its topological model (circles in Fig. 18).

1955. Work began recently on the construction of a high-speed machine, type GEL-250, intended for scanning information on magnetic tape, at a rate of 100,000 abstracts per hour. In 1958 Perry and Kent published an extensive monograph,[57] which contains extensive operating instructions for the use of machine language in information service. The first examples of mechanized information search in the field of metallurgy are reported in reference 58.

Let us examine the feasibility of using information machines of this type in creative work. In the planning of experiments, the research worker, on the basis of familiarity with the literature, advances a probable (likely) hypothesis. Success depends here, on the one hand, on the worker's ability to associate ideas, and on the other hand, on his erudition and ability to digest a large amount of information.

In the U.S.A. work is frequently organized as follows: the manager of the research project gathers a group of specialists in a great variety of fields and asks for their opinion on the problem of interest to him. Each of the present may propose his explanation, even if it be of intuitive character and not rigorously founded; the only condition is not to raise any objections or doubts as regards the stated idea. All the statements are recorded in shorthand and then the project manager selects the hypotheses that seem to him the most likely.[59] The probabilistic estimates of the likeliness of the hypothesis are never converted to quantitative estimates. The mechanism of the creative process, which leads to the adoption of any particular hypothesis in the planning of the experiment, is not known to us — this process, at least at the present level of formalization of scientific theories, cannot be described formally, although it is clear that information machines can render great help to the researcher in this matter. Let us assume, for example, that we are planning experiments on the influence of a third element on the results of spectral analysis. The machine should give us information on all the investigations, in which the influence of third elements on the coefficient of diffusion in a solid body and in a gas cloud have been investigated, and also on all the investigations in which the influence of third elements on the kinetics of evaporation have been considered, etc. The machine, naturally, produces here also information that is of no interest from the point of view of the problem under consideration, for example, information on the problem of three bodies in mechanics. The experimenter will obviously be able to select without difficulty from this information all the concepts of interest to him and to use them in the formulation of a hypothesis for the planning of his experiments, aimed at studying the influence of third elements in spectral analysis. The machine will replace the pronouncements of a group of specialists in a great variety of fields.

In the construction of an information system similar to that described here, the most important and responsible element is the choice of the level of the information search, which is determined, on the one hand, by the ramifications of the information language and, on the other hand, by the completeness of the information transmitted in the coded abstracts. It is difficult to propose a clear-cut criterion by which to select the level of information search without excessively complicating and increasing the cost of this process and at the same time without obtaining superficial results. We are faced here with a situation similar to that described in the mathematical theory of games: a person designing an information system plays a game with the person interested in obtaining the information, and attempts to foresee the strategy of the search without seeing clearly the paths of development of the corresponding scientific discipline and, consequently, without knowing adequately the sphere of interests of the investigators who will turn to information search.

8. ATTEMPT AT CONSTRUCTING LOGICAL-MATHEMATICAL THEORIES OF INFORMATION RETRIEVAL SYSTEMS

The very extensive practical experience, accumulated in connection with the functioning of a variety of information retrieval systems, has made it possible to formulate several substantial theoretical generalizations, making use of the result of modern mathematical logic and algebra. In section 7 of this paper we gave an example of determining the strategy of information search with the aid of Boolean algebra. In the present section we detail a general theory of the utilization of mathematical logic and modern algebra for the formalization of various systems of information retrieval. Interesting results were obtained in this direction by Mooers,[60] who considered three different abstract retrieval systems: (1) descriptor (abbreviated DIS), (2) hierarchic type (i.e., in which the elements are arranged in order of subordination and co-subordination; abbreviated HIS), (3) logical type (i.e., in which the elements are ordered by relations of logical character: disjunctions, conjunctions, implications, etc.; abbreviated LIS).

Mooers[60] developed a general mathematical model of arbitrary information systems, which is then applied to the description of the DIS, HIS, LIS, displaying the similarity and difference between them. The following principal concepts are used in this general theory.

(1) The space P of all possible retrieval prescriptions;

(2) The space L of all possible subsets of documents, selected from a given collection of documents;

(3) Search transformation T, which establishes a connection between P and L.

The space L is described with the aid of Boolean algebra. To each point from P there correspond certain points from L but not each point in L corresponds to a point in P. The correspondence between P and L is therefore not mutually unique.

The construction of P is more complicated than the construction of L. Therefore, for the sake of convenience in investigation, it is desirable to resolve P into components. The space P is considered as a certain product of all the elements that enter in a given repository R. The repository R is the fourth principal concept of the general theory of information retrieval systems. In it, the elements of the system (descriptors, objects of hierarchic type, or else objects ordered by logical relations) are represented in the form of several partially-ordered systems. In the description of each of the R, P, L, and T successively in each of the DIS, HIS, and LIS, Mooers[60] uses essentially the concepts from a generalized theory of modern algebra (partially-ordered system, cardinal product, etc.)*

Let us now dwell on how the DIS is described in terms of the concepts introduced. We denote the descriptors by upper-case letters (with or without indices). The elements R_{DIS} will be the graphically-representable expressions shown in Fig. 19. Each of these is a partially-ordered system. If we take the entire set of documents and the entire set of descriptors pertaining to a definite more or less narrow field of science, then the relation between en the arbitrary descriptors and the arbitrary document can be described in the following manner: either a chosen descriptor is necessary for the identification of the given document (as, so to speak, a key descriptor for the document), or it is not necessary for such identification. This unavoidable alternative is indeed fixed in the graphical representation of the elements from R, where the letters A, B, C (located on the upper end of the "matchstick" of the form $\overset{\mathfrak{A}}{\underset{0}{|}}$) correspond to the case, when the descriptor is necessary, and O — to that situation, in which it is unnecessary for identification of the document. In considering a partially-ordered system of the type $\overset{\mathfrak{A}}{\underset{0}{|}}$ we see that in such a system $\mathfrak{A}$ plays the role of unity, and O plays the role of zero. Consequently, such partially-ordered systems obey the laws of algebraic logic, and, in the language of the theory of lattices — the laws of distributive lattices with complements. The logical apparatus of the DIS is therefore algebraic logic (see remark on p. 654).

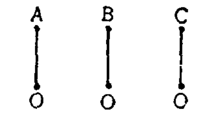


FIG. 19. Partially-ordered system for descriptors.[60]

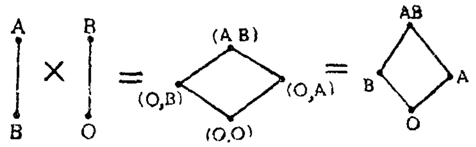


FIG. 20. Formation of a cardinal product for partially-ordered descriptor systems.[60]

P_{DIS} is formed as the cardinal product of the elements from R_{DIS}. Let, for example, the number of elements in R_{DIS} be 2. We form a cardinal product in the diagram as shown in Fig. 20 for the two descriptors A and B. The last of these figures in this diagram can be interpreted in the following manner: the documents identified by the joining of descriptors A and B (symbolically: A • B) are sought after finding all the documents identified by descriptors A and B. Each document is scanned and it is ascertained whether or not it is identified with the aid of each of the descriptors A and B. Here we have, so to speak, a hint of the program for the retrieval of docu-

*For the sake of the reader who has not made a special study of the mathematical theory of lattices, we give brief definitions of these concepts. A partially-ordered system is a set of elements together with a specific ordering relation such that for any two elements x and y from this set it is possible to decide whether x precedes y or vice-versa, or else it is impossible to determine any precedence relations between x and y. The concept of cardinal product is defined in the following manner. Let there be given two partially-ordered systems X and Y, typical elements of which are denoted by means of x and y. In this case the cardinal product X·Y is also a partially-ordered system. X·Y consists of all pairs of forms (x, y), which are ordered with the aid of the rule, that $(x, y) \leqslant (x', y')$, where $x \leqslant x'$ in X, and $y \leqslant y'$ in Y (the relation $\leqslant$ is understood in the sense of the relation of precedence). Both the partially-ordered system, and the procedure of forming a cardinal product can be illustrated with a diagram.

ments that satisfy the condition A · B for the descriptor system. Each of the expressions $\begin{matrix}\mathfrak{A}\\ \vdots\\ 0\end{matrix}$ and $\begin{matrix}\mathfrak{B}\\ \vdots\\ 0\end{matrix}$ is a Boolean structure. We now describe $T_{1_{\mathrm{DIS}}}$ and $T_{2_{\mathrm{DIS}}}$. The definition of $T_{1_{\mathrm{DIS}}}$ reads as follows: it is possible to relate each point from P_{DIS} with certain points from L_{DIS}, although not each point in L_{DIS} can be connected with some point from P_{DIS}. The definition of $T_{2_{\mathrm{DIS}}}$ can be formulated in the following manner: by specifying a given point x in P we segregate a large family X of other points of P for which the relation $x \le X$ is satisfied (in the graph all such points of family X lie above the point x; the relation "≤" corresponds to the inclusion relation). The document collection contains many documents whose assigned subset of descriptors is one of the points in the family X. Let us consider the largest class of documents such that each document in this class has a subset of descriptors represented by some point in X. Let this set of documents be represented by a point x* in L. The transformation $T_{2_{\mathrm{DIS}}}$ transforms the point x in P_{DIS} into the point x* in L_{DIS}. We denote the transforma- ι-tion of x into x* by "$\in\!\rightarrow$". Let J and O be respectively the largest and smallest elements in P and L. Then the two resultant relations are obvious:*

$$\text{(a)}\quad 0(P)\in\!\rightarrow J(L),$$
$$\text{(b)}\quad J(P)\in\!\rightarrow 0(L).$$

R_{DIS} is a lattice, and therefore the relation $x \cup y = z$ is satisfied for any x and y in P, where "∪" is the joining operation.† Where $T_{2_{\mathrm{DIS}}}$ yields $x\in\!\rightarrow x^*$, $y\in\!\rightarrow y^*$, $z\in\!\rightarrow z^*$, we have in L* (L* is the result of the retrieval from L): $x^* \cap y^* = z^*$ (where "∩" is the meeting operation). However, this relation does not hold in L. Relations containing "∪" and "∩" in P also hold in L* (provided the symbols "∪" and "∩" are interchanged).

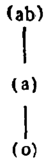


FIG. 21. Example of a partially-ordered system in hierarchic classifications.[60]

Analogous studies have been made also for HIS and LIS. Let us dwell briefly on systems of the HIS type. We represent the elements of a HIS by lower-case letters (or combinations of such letters) in parentheses. For two elements (a) and (b), the partially-ordered system (HIS) is as shown in Fig. 21 (b = ab, i.e., b ≤ a). The motion from (0) to (ab) follows the line of increasing contents and decreasing compass of the concept (i.e., along the line of concretization of the concepts used). In constructing R_{HIS}, an expression of the type (a, a) reduces to a. We distinguish between two types of hierarchies: strong and weak. In a strong hierarchy no element can be preceded immediately by more than one element. If this condition is not satisfied, we deal with a weak hierarchy. An example of a strong hierarchy is one in the decimal system, while classification of patents in the U.S. Patent Office is a weak one. In comparing the suitabilities of weak and strong hierarchies to machine search, Mooers gives preference to weak-hierarchy systems as being more flexible. In our opinion, this is explained by the fact that in systems with strong hierarchy only it is always possible to introduce the operation ∩, but not the operation ∪, for were it possible to define ∪ in them, then there would exist an element Φ, with two immediately preceding terms Φ′ and Φ″, and this would violate the condition that constitutes the very concept of strong hierarchy.

Let us consider an example of cardinal products of partially-ordered systems with hierarchies of various types. (For a strong hierarchy, each direct predecessor of an element is located exactly one place below it.) An illustration of a cardinal product is shown in Fig. 22. An analogous example for a weak hierarchy is shown in Fig. 23. The reader can readily interpret the last figures on each of these diagrams by analogy with what was done for the last figure of Fig. 20 in the case of the descriptor information system. Let us now compare the strong hierarchy with the weak hierarchy. In

*The relations (a) and (b) can be "deciphered" as follows: relation (a) expresses the thought that the absence of descriptors in the search prescription is equivalent to a requirement of selecting all the documents from the collection; relation (b) states that not a single document in the collection can be identified by the entire set of the descriptors.

†We define the operation "∪" in the following manner. Let $x \leqslant z$ and $y \leqslant z$, z being the smallest of the elements for which the two above relations are satisfied; in such case z is called the join of x and y and denoted $x \cup y$. The operation "∪" is sometimes called "cup." The operation "∩" which will be use later on is defined as follows. Let $z \leqslant x$ and $z \leqslant y$, z being the largest of the elements, for which the two foregoing relations are satisfied; in this case z is called the meet of x and y and is denoted $x \cap y$. The operation "∩" is sometimes called "cap." Taking into account the definitions in this and the preceding footnotes, it is now easy to write a general definition of a lattice. Indeed, a lattice is a partially-ordered system, in which it is possible to define the two operations "∪" and "∩" for any pair of elements. The expression $x \cap y$ is sometimes written in the literature as $x \cdot y$ and $x \cup y$ is sometimes written as $x + y$.

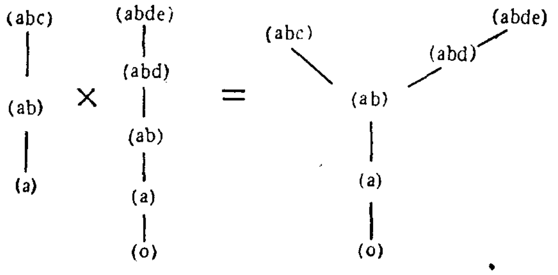


FIG. 22. Illustration of a cardinal product for a strong hierarchy.[60]

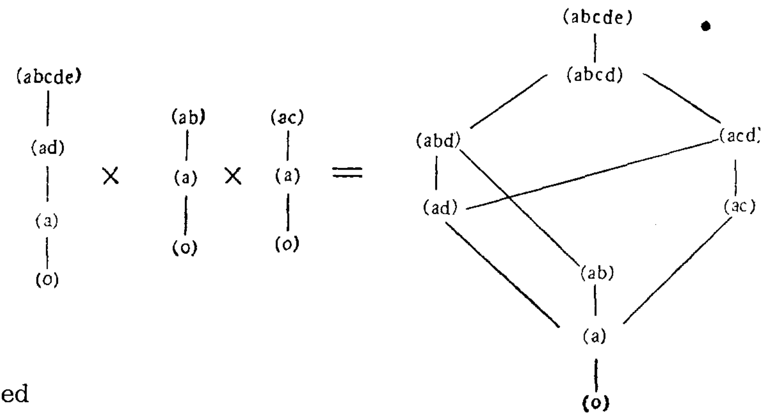


FIG. 23. Illustration of a cardinal product for a weak hierarchy.[60]

a strong-hierarchy system each element in the sequence can have only one predecessor (located one place below on the corresponding diagram). On the other hand, in a weak hierarchy each element in the sequence can have one or more copredecessors (including a null element). A strong hierarchy is represented graphically in the form of a tree, while a weak hierarchy resembles a net. Classification systems with elements of both strong and weak hierarchies are possible. We shall not dwell on these, and refer the reader interested in them and in the general problems of the application of mathematical methods to the theory of classification of systems, to reference 61. R_{HIS} and also P_{HIS}, L_{HIS}, T_{1HIS}, and T_{2HIS} are defined similar to the concepts used in DIS.

Mooers considers also information retrieval systems with elements ordered by the logical operations "and," "or" (not in the partitive sense), and "not." Each retrieval prescription is formulated in the form of a certain logical polynomial, made up of the operations "and," "or," and "not." If we bear in mind the correspondence between the abstract information systems DIS, HIS, or LIS, described in reference 60, and the actual machine (or machine-assisted) retrieval systems, then the DIS system corresponds tô the method of coordinate indexing used by Taube's group, and a combination of the DIS and LIS systems corresponds approximately to the method of Perry, Berry, and Kent (we say "approximately" in view of the fact that in the systems DIS and LIS there is no problem of formalization of certain synthetic (descriptive) relations, which, as is well known, is a component part in the methods proposed in reference 56). On the other hand, the HIS system can be considered as a formalization of the existing systems of library classification, and also of different improvements and modifications of these systems.

The problem is interestingly formulated in the paper by F. Jonker,[62] who postulates a generalized theory of indexing, in which all indexing systems are viewed as a relative continuum (spectrum). A distinguishing parameter of this continuum is the average length of the individual entries or headings used in the indexing. This parameter, which determines the position of the indexing system in the continuum, is merely the average of the number of letters in the indexing terms. At one end of the continuum or spectrum there is indexing with the aid of key words (descriptors); indexing by subject heading is somewhere in the middle, while hierarchic classifications are at the other extreme (of the spectrum). Schematically, the descriptive continuum is represented as shown in Fig. 24. The line ab symbolizes the position of the descriptor indexing ("short-term" end of the spectrum), bc represents the position of indexing by subject headings, and cd — indexing with the aid of hierarchic classifications ("long-term" end of the spectrum). Each document collection should be indexed by some optimal indexing system, i.e., a system that insures sufficient speed of retrieval and reliability in the storage of the information at reasonable cost.

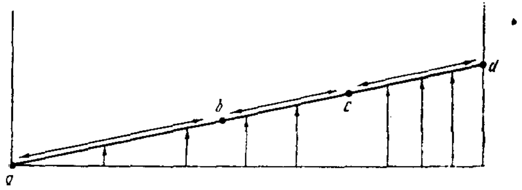


FIG. 24. Descriptive continuum.[62]

It is stated in reference 62 that all other parameters of the continuum behave as functions of the principal parameter — the average length of the heading in the indexing. Some of the derivative parameters are:

(1) Potential depth of the indexing (proportional to the number of indexing criteria used).

(2) Permutability of the indexing criteria.

(3) Degree of hierarchical definition of the indexing.

(4) Capacity for handling semantic indeterminacy.

(5) Retrieval noise (delivery of excessive information).

(6) Potential need for a coordinating mechanism.

In spite of the interesting statement of the problem, reference 62 contains several weak spots. The "principal law of descriptive continuum" formulated there (degree of hierarchical definition × depth of indexing = const) is more qualitative than quantitative in character. Secondly, there is no quantitative processing of the derivative parameters. Also noticeable is the fact that the derivative parameters segregated by Jonker are not of equal weight: thus, the permutability of the indexing criteria, is a parameter derived from another derived parameter, the potential depth of indexing. Furthermore, what is most important, the functional relation between certain derivative parameters and the principal parameter — the average length of the indexing heading — is not clear (to the extent that it is doubtful whether it exists). What is also striking is the somewhat exaggerated value given to the descriptor system over other indexing systems. Nevertheless, we must emphasize the general-methodological value of the principal idea of reference 62, which postulates the need for creating a generalized theory of indexing of documents, capable of describing both machine and "manual" methods of information processing.

It is natural to raise the following question: How can the latest theories of documentalistics be used to facilitate the laborious procedures of library search? Even in the present-day state of library science and documentalistics, we can cite several illustrations of the practical applicability of logical-mathematical representation of information retrieval. As is known, mathematics pays special attention to the properties of so-called maze structures. By a maze is meant a certain net, consisting of key points and connections between them. The library subject catalogue can be considered a maze, because it contains sets of cross references, linking the subject headings. This general idea and its consequences are discussed in reference 63.

Moore[64] posed the following elementary problem in the theory of mazes: Given an origin point and destination point in the maze, find a path from origin to destination with a minimum number of links. Moore proposed a special algorithm to solve this problem (Estrin[63] employs this algorithm to simplify the use of a subject catalog without foregoing the decimal system of document classification). The algorithm[64] for the determination of the shortest path in a maze consists of several steps and can be described in the following way:

Initial step. Select the origin and destination. Tag the origin with a zero. Enable the recognition of the destination regardless of where it may be selected.

Steps i. First (in the first of steps i) provide with tag 1 all the points adjacent to the origin. In the second step, the tag 2 designates all the points adjacent to the points already tagged 1. The process continues until the destination point is reached, which is tagged k. Thus i runs through the following values:

$$i=1,\ 2,\ 3,\ \ldots,\ k.$$

The number k is exactly equal to the number of links in the sought minimum path.

Steps j (j = k+1, k+2, k+3, . . . , 2k). Whereas steps i give information concerning the number of links in the sought shortest path, the following steps j explicitly determine a minimum path. We move backward from the destination point to certain points adjacent to it. In each of these steps we choose the adjacent point with the lowest of all tags i and assign to it an additional tag $i' = i$. Each newly tagged point is selected as a base for the next step. This process is continued until the origin point is reached, the tag of which is $i' = 0$. All the points tagged i' lie on a minimum path.

In our opinion, Moore's algorithm is obviously not the best of all the possible algorithms of finding the shortest path in a maze. More convenient, in any case, is an algorithm that begins not with the origin point alone, but simultaneously with the destination point. In this latter case the need of going through all points of the maze without exception (as we have seen done in the Moore algorithm) is dispensed with. In fact, let us imagine a maze in which two links go from each point. If we move along this maze from point 0 to the point n, using the Moore algorithm, we must go through 2^{n+1} points, while in the other possible approach we need go through not more than $2^{(n+3)/2}$ points.

However, even Moore's original algorithm is apparently quite simple and suitable for mechanization by existing computer technology. A very important feature of the maze is the possibility of its reorientation with respect to the choice of the origin point. After the reorientation is realized (calling for a preliminary ordering of the points by known levels), the shape of the maze may change radically. In information retrieval systems the applicability of Moore's algorithm is almost obvious and is connected with the solution of the following problem: let the information system be characterized graphically in the form of a certain maze; construct a Moore algorithm (or its modification) for the net in accordance with the specified retrieval prescription. In terms of library science, the situation admits of the following formulation: imagine that the points of the maze correspond to those library headings of the subject catalogues, which lead to a system of cross references. We choose the origin point of the maze corresponding to the initial subject heading, related to the retrieval prescription. In this case the shortest path between the specified origin and destination points (headings) will describe the best strategy of library search. We note that it is also necessary here to graph the catalogue maze relative to the origin point, i.e., to reorient the maze (or its fragment — a local area of the catalogue) with respect to the initial object heading.

Thus, the Moore algorithm for finding the shortest path in a maze can be used effectively to facilitate the procedure of library search. Another feature of this algorithm (important at least from the point of view of present-day conditions of non-mechanized library work) is that it does not force us to forego the decimal system of classification of documents (which is quite unavoidable, for example, when descriptor-type information search systems are used).

Among other attempts of constructing logical-mathematical models of information retrieval systems, brief notice should be taken also of two articles by Vickery.[65,66] The use of the results of modern mathematical logic is much less constructive in these articles than in the earlier papers considered. The principal concepts in reference 65 are: (1) the concept of the term as a unit of meaning and (2) the concept of the document as a unit of recorded information. The concepts of semantic connections between the terms and the documents are introduced, and various methods of information recording are considered. At the level of subject analysis, only information units are investigated, while at the level of structural analysis a transition is made to representation of information systems in the form of a certain lattice of information units. Great attention is paid to the analysis of inclusion and coordination relations for terms. In Ref. 65 there are then proposed rules for the conversion of the information lattice into a two-dimensional matrix, in which the columns contain the numbers of the documents (items) and the rows the corresponding terms. The intersection of the i-th row and the k-th column yields the inclusion (or absence) of a given term in a given item. This is followed by conversion of the two-dimensional matrix into a one-dimensional tally. In reference 66 the author investigates, in particular, the concept of modulants — categories of semantic units in the sense considered in reference 56. The general shortcoming of Vickery's approach to the solution of the problem of developing a unified theory of information systems is that he foregoes the segregation of syntactic (in some respect synthetic) connections between terms.

CONCLUSION

Searches for new ways of storage, dissemination, and processing of scientific and technical information have led to the creation of a scientific discipline (still unnamed), which can be considered a branch of cybernetics. This discipline cuts across several sciences: mathematical logic, theory of probability and mathematical statistics, documentalistics, linguistics, psychology of thinking, electronics, and computer technology. In the future (in connection with the search for new technical means of contraction of information) it will apparently make extensive use of the results of biophysical research – since the most complex contraction of information takes place in the nucleic acids of chromosomes, in which the structure of the entire organism is coded.[68,69]

Success in the development of this new discipline will depend to a considerable extent on how the principal problem of cybernetics, that of the mutual relations between the capabilities of a computing machine and thinking,[70] is solved.

Even now it is obvious that the development of this discipline will lead to new forms of organization of science and will contribute above all to its further mathematization. The development of this discipline calls for collective efforts of specialists of all the aforementioned fields of science.

[1] D. J. Price, Archives Internationales d'Histoire des Sciences **30**, No. 14, 85 (1953).

[2] D. J. Price, Discovery **17**, No. 6, 240 (1956).

[3] E. Pietsch, Arbeitsgemeinshaft für Forschung des Landes Nordrhein-Westfalen **38**, 34 (1954) (Westdeutscher Verlag, Köln und Oplanden).

[4] Biological Abstracts **31**, No. 7, X (1957).

[5] R. H. Ewell, Chemical and Engineering News **33**, No. 29, 2980 (1955).

[6] J. Thomson, The Foreseeable Future, (Russ. Transl.) IL, Moscow, 1958.

[7] M. H. Halbert and R. L. Ackoff, Preprints of papers for the International Conference on Scientific Information, Washington, D.C., November 16-21, Area 1, p. 87 (1958).

[8] F. Liebesny, ibid, Area 2,.p. 161.

[9] D. A. Brunning, ibid, Area 3, p. 7.

[10] H. C. Lehman, Scientific Monthly **78**, No. 5, 321 (1954).

[11] W. Shockley, Proceedings of the IRE **45**, No. 3, 279 (1957).

[12] H. P. Luhn, IBM Journal of Research and Development **2**, No. 2, 159 (1958).

[13] S. Ullmann, Journal de Psychologie normale et pathologique **55**, No. 3, 338 (1955).

[14] P. B. Baxendale, IBM Journal of Research and Development **2**, No. 2, 159 (1958).

[15] V. V. Nalimov and B. S. Neporent, Usp. Fiz. Nauk **65**, No. 3, 521 (1958).

[16] T. E. Beukelman, Analytical Chemistry **29**, No. 9, 1269 (1957).

[17] I. Wachtel, American Documentation **3**, No. 1, 56 (1952).

[18] Addison, Spencer, and Charlet, Analytical Chemistry **30**, No. 5, 885 (1958).

[19] Casey, Perry, Kent, and Berry (Editors), Punched Cards. Their Application to Science and Industry. 2 Ed., Reinhold, New York, p. 697 (1958).

[20] H. D. Ashthorpe, ASLIB Proceedings **4**, No. 2, 101 (1952).

[21] H. E. Stiles, American Documentation **9**, No. 1 42 (1958).

[22] J. D. Bernal, cf. reference 7, Area 1, p. 67.

[23] F. Heumann and E. Dale, Progress Report in Chemical Literature Retrieval, Interscience Publishers, p. 201 (1957).

[24] Beilsteins Handbuch der organischen Chemie, B. 29, 1 Teil General-Formelregister für das Hauptwerk und die Ergänzungswerke I und II, Berlin, Springer Verlag, 1956.

[25] Chemical and Engineering News **33**, 2838 (1955).

[26] G. M. Dyson, A New Notation and Enumeration System for Organic Compounds, 2 Ed., Longmans, Green and Co, London and New York, 1949.

[27] W. J. Wisswesser, A Line-Formula Chemical Notation, T. Y. Crowell Co., New York, 1955.

[28] Gordon, Kendall, and Davison, Chemical Ciphering: A Universal Code as an Aid to Chemical Systematics, Royal Institute of Chemistry of Great Britain and Ireland, 1948.

[29] F.H. S. Curd and G. F. L. Rose, J. Chem. Soc., 729 (1946).

[30] Chemical Codification Panel, National Research Council, A Method of Coding Chemicals for Correlation and Classification, Washington, 1950.

[31] E. Pietsch, Die IBM-Lochkarte in der Gmelin-Dokumentation, FID Manual on Document Reproduction and Selection, Part II, 1956.

[31a] W. Steidle, Pharmazeutische Industrie **19**, No. 3, 88 (1957).

[32] L. Bellamy, Infrared Spectra of Molecules (Russ. Transl.), IL, Moscow, 1957.

[33] L. C. Ray and R. A. Kirsch, Science **126**, 814 (1957).

[34] A. Opler and T. R. Norton, Chemical and Engineering News **34**, 2812 (1956).

[34a] A. Opler, Chemical and Engineering News **35**, No. 33, 92 (1957).

[35] W. H. Waldo and M. De Backer, cf. reference 7, Area 4, p. 49.

[35a] Waldo, Gordon, and Porter, American Documentation **9**, 28 (1958).

[36] J. Sherman, Industrial and Engineering, Chemistry **50**, No. 11, 2441 (1958).

[37] L. I. Gutenmakher, Вестник АН СССР Bulletin AN SSSR **88**, No. 10, (1957).

[38] G. É. Vléduts, Некоторые вопросы научной информации в области химии. I. О путях усовершенствования химических указателей (Some Problems of Scientific Information in the Field of Chemistry. I. Ways of Improving Chemical Indices.) Publ. by Inst. Sci. Tech. Inform. Acad. Sci. U.S.S.R., Moscow, 1958.

[39] R. A. Fisher, Статистические методы для исследователей (Statistical Methods for Researchers), Gosstatizdat, M., 1958.

[40] W. J. Youden, Statistical Methods for Chemists. John Wiley and Sons, Inc., New York, Chapman and Hall, Ltd., London, p. 127, 1951.

[41] W. T. Federer, Experimental Design, Theory and Application. Macmillan and Co., New York, 544 p. (1955).

[42] O. Kempthorne, The Design and Analysis of Experiments. John Wiley and Sons, Inc., New York, Chapman and Hall, Ltd., London, p. 631 (1952).

[43] C. A. Bennett and N. L. Franklin, Statistical Analysis in Chemistry and the Chemical Industry. John Wiley and Sons, Inc., New York, Chapman and Hall, Ltd., London., 724 p., 1954.

[44] O. L. Davies (Editor), Design and Analysis of Industrial Experiments. Imperial Chemical Industries Ltd., Oliver and Boyl, London, Edinburgh, 636 p., 1956.

[45] O. L. Davies (Editor), Statistical Methods in Research and Production. Imperial Chemical Industries Ltd., Oliver and Boyl, London, Edinburgh, 396 p., 1957.

[46] D. R. Cox, Planning of Experiments. John Wiley and Sons, Inc., New York, Chapman and Hall, Ltd.,

London, p. 308, 1958.

[47] Coutura, Louis, Algebraic Logic, Odessa, Publ. by Mathesis, 1909.

[48] Taube, Gull, and Wachtel, American Documentation **3**, No. 4, pp. 213-218 (1952).

[49] M. Taube and Irma S. Wachtel, American Documentation, April, pp. 67-69 (1953).

[50] M. Taube, American Documentation **6**, No. 4 (1955).

[51] R. Brakhen and G. Tillitt, Journal Assoc. Comput. Machinery **4**, No. 2, pp. 131-136 (1957).

[52] R. Carnap, Der logische Aufbau der Welt, Springer, Berlin, 1928.

[53] J. H. Woodger, The Axiomatic Method in Biology, Cambridge University Press, 1937.

[54] H. Hollerith, Art of Compiling Statistics, U.S. Patent No. 395782, January, 1889.

[55] H. Hollerith, Apparatus for Compiling Statistics, U.S. Patent No. 395787, September 23, 1884.

[56] Perry, Kent, and Berry, Machine Literature Searching. Interscience Publishers Inc., New York p. 162, 1956.

[57] J. W. Perry and A. Kent, Tools for Machine Literature Searching. New York, Interscience Publishers Inc., p. 992.

[58] J. Rees and A. Kent, American Documentation **9**, No. 4, 277 (1958).

[59] W. Platt, Information Work in Strategic Reconnaissance, (Russ. Transl.) IL, Moscow, 1958.

[60] C. N. Mooers, cf. reference 7, Area VI, pp. 57-94.

[61] R. A. Fairthore, The Mathematics of Classification, Proc. British Society of International Bibliography **9** (4) (1947).

[62] Frederick Jonker, cf. reference 7, Area VI, pp. 22-41.

[63] Gerald Estrin, ibid. pp. 113-123.

[64] E. F. Moore, International Symposium on Switching Theory, Harvard University, cf. reference 7, April 1957, in press.

[65] B. C. Vickery, cf. reference 7, Area VI, pp. 5-20.

[66] B. C. Vickery, ibid., Area V, pp. 41-52.

[67] D. Hilbert and W. Ackermann, Foundations of Theoretical Logic, (Russ. Transl.), IL, Moscow, 1947. (Chelsea, NY 1950).

[68] E. Schrödinger, Life from the Physical Point of View, (Russ. Transl.) IL., Moscow, 1948.

[69] Gamow, Rich, and Icas, article in Collection: Вопросы биофизики (Problems of Biophysics), IL, M. 1957.

[70] A. A. Lyapunov, article in collection, Проблемы кибернетики (Problems of Cybernetics), Fizmatgiz, M., 1958, No. 1, p. 5.

Translated by J. G. Adashko