A Statistical Approach To Mechanized Literature Searching

H.P. LUHN

IBM®
RESEARCH CENTER

INTERNATIONAL BUSINESS MACHINES CORPORATION
POUGHKEEPSIE, NEW YORK

A STATISTICAL APPROACH TO MECHANIZED LITERATURE SEARCHING

H. P. Luhn

International Business Machines Corporation
Research Center
Poughkeepsie, New York

Abstract

Communication of ideas is carried out on the basis of statistical probability in that an author chooses that level of specificity and that combination of words which will most probably assure him comprehension on the part of those he intends to reach. Since the execution of this process varies amongst authors and since similar ideas are therefore relayed at different levels of specificity and by means of different words, the problem of literature searching by machines still lacks a satisfactory solution. A statistical approach to this problem will be outlined and the various steps of a system based on this approach will be described. These steps include the statistical analysis of a collection of documents of a chosen field of interest, the establishment of a set of 'notions' and the vocabulary by which they are being expressed, the compilation of a thesaurus-type dictionary and index, the encoding of documents with the aid of the latter, the encoding of topological notations (such as branched structures), the recording of the coded information, the establishment of a searching pattern for finding pertinent information, and the programming of appropriate machines to carry out a search.

Research Paper
RC-3
January 30, 1957

Section I

Introduction

The problem of literature searching is to find those documents within a collection that have a bearing on a given topic. Many of the systems and devices that have been developed in the past to solve this problem are proving inadequate. New solutions are being pressed for by the rapid growth of literature and the demand for higher levels of efficiency in coping with it.

Great hopes are entertained for the application of powerful electronic devices in obtaining new and satisfactory results. The fulfillment of these hopes is in doubt if such modern devices are merely to be viewed as agents for accelerating systems heretofore fitted to human capabilities. The ultimate benefits of mechanization will be realized only if the characteristics of machines are better understood and systems are developed that exploit these characteristics to the fullest. Rather than subtilize the artful classificatory schemes now in use, such systems would replace them in large part by cold mechanical routines based on rather elementary reasoning

The sacrifice involved in substituting mechanical for intellectual means should be justified by the ends, for if relief cannot be found in automation, the demand for professional talent will become too great to fill. In view of the foreseeable strain, the most efficient use of talent will have to be made even by automatic systems. The operating requirements of these systems will, above all, have to be well adapted to the degree of education and experience of generally available personnel.

Language difficulties, too, will have to be met. The problems stemming from the mere volumes of literature to be searched are being continually aggravated by the increasing accession of foreign-language documents that rate consideration on an equal level with domestic material. To be of real value, future automatic systems will have to provide a workable means of overcoming the language barrier.

Levels of Information Systems

The general terms in which the problem of literature searching has been treated might imply the prospect of a general, or universal, solution. It would be unrealistic to assume that such is practical or desirable. It seems to be quite important to establish at this time some differentiating

criteria by which information reference arrays may be distinguished and graded as to their make-up, objectives, and uses. It will then be possible to better recognize the existence of different levels of complexity and the necessity of applying appropriately different techniques to their mechanization. The following list of six systems in the order of ascending complexity may be of use in this respect.

Levels of Information Systems

1. Ready reference look-up systems of facts such as indexes, dictionaries, and parts catalogues.

2. Systems of limited and narrowly defined specific categories especially where, as in lists of specifications, categories are repetitive.

3. Systems of the kind found commonly in chemistry that deal with inventories of uniquely definable structures and their interrelations and transformations.

4. Systems of mathematics, logic, and law that are based on disciplined concepts of human intellect.

5. Systems dealing with the exploitation of natural phenomena and things as in the applied sciences and technology.

6. Systems, of which pure fiction is the extreme, that deal with unrestricted association of human notions.

While there may be other criteria by which to graduate the spectrum of information, it is important to realize that severe differences of information make-up do exist. It therefore makes little sense to discuss a literature searching system without also identifying the portion of the spectrum to which the system is to be applied.

Distribution of Human Effort

Since the graduation of the above list goes from quite explicit factual listings to the vagueness of fancy, it seems unavoidable that the efficiency of recognition of desired information will decrease in this direction. The various systems might therefore be characterized by their recognition potential and the amount and distribution of human and machine

effort required. It seems to be an inescapable fact that the less diciplined the language, the greater the human effort that will have to be expended somewhere in the process.

There are four distinct phases of human effort involved.

1. The design, setup, and maintenance of the system proper.

2. The interpretation and introduction of information into the system.

3. The programming of wanted information for mechanical recognition.

4. The interpretation of selected records to determine whether they are relevant to the wanted information.

To arrive at an optimum process for a given information level, the question of the quality and proportion of human effort to be expended at each of these phases must be resolved.

The introduction of time as an additional variable will change proportions quite considerably. If, for instance, any kind of information must be located in a matter of minutes, the possible maximum of skilled effort will have to be spent at the input phase of the system and in an equal degree on every entry into the system. If, however, less pressing time requirements are feasible, input procedures that require medium skill and minimum effort may be chosen so that the skilled effort can be concentrated at the output phase on only a small fraction of the records of the collection. In the latter case, the fact that only a small fraction of the records of a collection will ever be selected should result in a reduction of the overall effort.

Time may affect a system in another way that makes the shift of skilled effort to the output phase more desirable. Excessive editing obviously increases the likelihood of bias due to current interests, experiences, and points of view. In consequence the usefulness of the system will be reduced as emphases and interests change in time. It would therefore appear that the less information is classified and contracted at the input, the more it will lend itself to dynamic interpretation at the output phase.

A Proposed Solution by Statistical Methods

The following paragraphs will present the basis and organization of a literature searching system that utilizes statistical methods in conjunction with a high degree of mechanization. The principles involved represent an extension and refinement of those discussed in an earlier paper.[1] Although the specific system described is primarily designed to satisfy the requirements of information level 5 of the foregoing list, it may also be found adaptable to levels 4 and 6.

Generally speaking the proposed system is based on what are variously referred to as cross-indexing, multidimensional-indexing, coordinate-indexing, multiple-aspect-indexing and encoded-abstract techniques. Actual practices vary from lifting key words of a text by a manual editing process to interpretive analysis by logical formulas of well-defined concepts. For a general description and a bibliography of methods using these techniques, the reader is referred to "Inventory of Methods and Devices for Analysis, Storage, and Retrieval of Information" by Marjorie R. Hyslop. This paper appears in the book Documentation in Action, edited by J. H. Shera, A. Kent, and J. W. Perry, and published by the Reinhold Publishing Co., New York, in 1956.

1 - Luhn, H. P., "A New Method of Recording and Searching Information", American Documentation, Jan., 1953, v. 4, pp. 14-16.

Section II

The Statistical Aspects of Communicating Ideas

Communication of ideas between humans by way of words is carried out on the basis of statistical probability. We speculate that by using certain words we will be able to produce in somebody else's mind a certain mood and disposition resembling our own state of mind which resulted from a certain actual experience or a process of thought. In order to communicate an idea, we break it down into a series of little ideas, i.e. more elementary ideas for which previous and common experience might have led to an agreement of meaning. We extend this process until we feel that we have reached a level of conventional concepts, a level at which communication can be accomplished. This level may vary depending on the degree of similarity of common experiences. The fewer experiences we have in common, the more words we have to use.

A picture of this process, if it can be drawn at all, might look something like the triangular portion of Fig. 1.

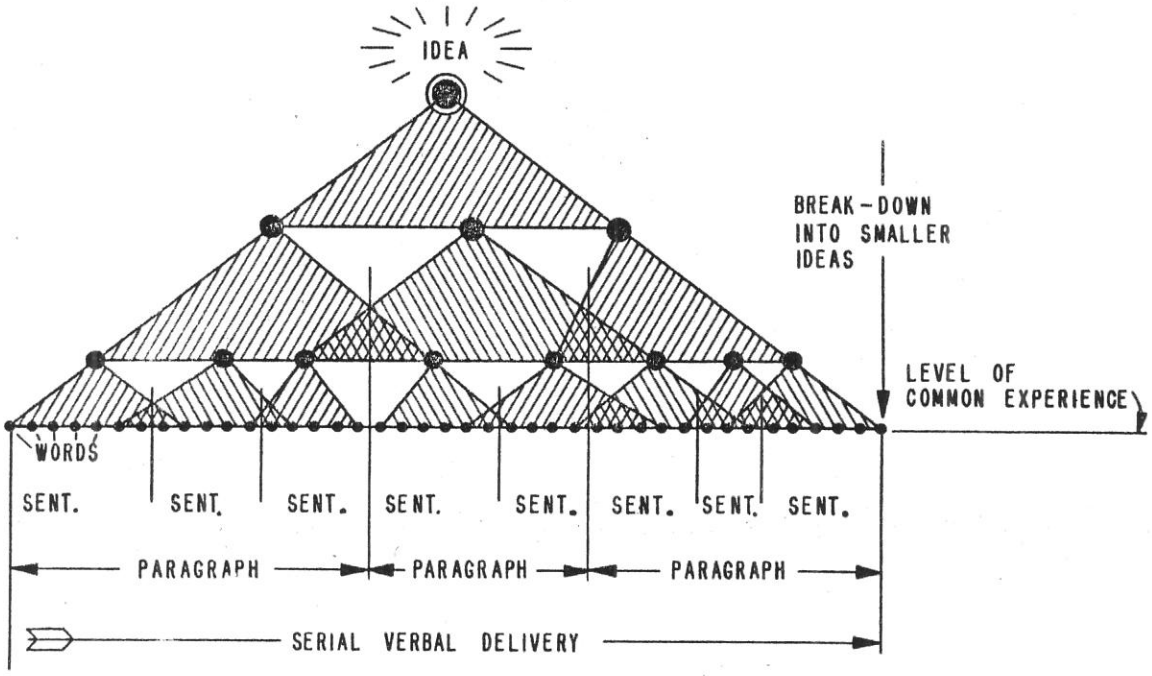

FIG. 1

The process of communicating ideas is dynamic when it can be performed by means of the spoken word between two people. In the first place, the addressor can size up the addressee and adjust the process of subdividing his idea to the level of common experience which most probably exists between the two. Secondly, guided by the feedback of the addressee's reactions and questions, the addressor may readjust to a reasonably optimum level or change his strategy of composition.

The process assumes static qualities as soon as ideas are expressed in writing. Here the addressor has to make certain assumptions as to the make-up of the potential addressee and has to make a decision as to which level of common experience he should choose. Since there is no direct feedback, the addressor has to rely on some kind of indirect feedback. The addressor might therefore take his guidance from the degree to which the written expressions of ideas of others have raised the level of common experience relative to the ones he wishes to communicate.

The most general such guidance is furnished by the dictionary. Here the verbal expressions of ideas at a given level of common experience are defined in terms of verbal expressions at other levels so that a broad domain of common experience is assured. Thus, the dictionary is a periodical report to word users on the ideas which are currently being relayed more often than not by the words they are using. The level at which the lexicographer breaks off his reporting may vary and is, of course, dictated by economical factors. This is so because in the extreme he would have to quote substantial parts of the current literature to explain the slightest differentiations of ideas. It is the task of special dictionaries to bring more remote areas of experience into the common domain by explaining ideas at higher levels.

In writing, the addressor may then take the special dictionary of his field of interest as a next approximation to a level of common experience on which to communicate with the addressee. However, since the lexicographer can never be up to date, there still remains a gap which the addressor will have to fill to permit the addressee to adjust himself to the desired level. This he may do by referring the addressee, by means of a bibliography, to that portion of the literature which the lexicographer has not as yet analyzed.

It may be assumed that the means and procedures just mentioned permit communication to be accomplished in a satisfactory manner. If it is possible to establish a level of common experience, it seems to follow that there is also a common denominator for ideas between two or more individuals. Thus, the statistical probability of combinations of similar ideas being similarly interpreted must be very high.

If it were possible to recognize such idea building blocks irrespective of the words used to evoke them, these building blocks might be considered the elements of a syntax of notions. Communication could then be carried out by relaying these notions by means of agreed-upon symbols. Since these symbols would be independent of style and language, they might serve to overcome language barriers. A symbol system of this kind would be most useful in facilitating the process of information recognition by automatic means.

Section III

Building Blocks for a Statistical System

The lack of uniformity of structure, the absence of laws, and the arbitrariness of word usage make literature an unwieldy subject for automation. Its information content must first be represented and organized in a form that can be operated on by a machine, for only then can the degree of similarity between any two records be automatically determined. The most efficient means of transforming information would be those that permit the application of a minimum of logical machine instructions to bring about the desired results.

The very nature of free-style renderings of information seems to preclude any system based on precise relationships and values, such as have been developed in the field of mathematics. A systematic approach appears to be best facilitated by treating the problem as a statistical proposition. The objectives of a system based on this proposition would be first to transform information into arrays of normalized building blocks and then to discover similarities in the respective building-block patterns of these arrays by means of a statistical analysis. It could be reasonably assumed that the more closely two arrays are matched, the greater the probability that the records they stand for contain similar information.

It is true that the principle of pattern matching has been applied in searching systems before. The emphasis here, however, is on the use of notions as a basis for pattern derivation. Where such non-precise elements can be used as building blocks, the possibility of creating a practical information retrieval system is substantially increased.

In the process of communicating ideas, an author pursues a certain plan of organizing his ideas. The outside evidence of such a plan is the grouping of his ideas into chapters, paragraphs, and sentences. Fig. 1 illustrates how this organization may come about. Notions are most closely and specifically related to each other within a sentence. A sentence immediately following might either be related in its entirety to previous notions or serve to relate these notions to new ones. The same might be

said of the succeeding sentences. However, as soon as, in the opinion of the author, a significant new argument is introduced or applied to previously expressed ideas, he will usually signify this by an appropriate division, such as a new paragraph. A still more decisive change of aspects might be denoted by the start of a new chapter.

This conscious division by the author furnishes a key to the relatedness of his notions and should be accepted as a significant and meaningful element of the information he is attempting to relay. We may therefore consider several degrees of relationship, namely, the first order relationship of notions within a sentence, a second order relationship between sentences within a paragraph and their respective notions, a third order relationship between paragraphs within a chapter, and still higher orders for larger divisions.

The final division, as far as the author is concerned, is the entirety of his message or his paper. However, since it is desirable to make the paper or document comparable with other similar documents, a still higher level of grouping is indicated and this is the level of common experience previously discussed. It was argued that a level, or field, of common experience was a requirement for communication. It follows that the more specific this field is, the closer will be the agreement amongst the notions used in the mental process of people associated with that field. It therefore seems to be important and helpful to recognize these fields and establish them as a next order or level of division.

Communications at this latter level are made as though in a foreign tongue, in that people in various specific fields each speak a "native" technical language. However, since notions are here to be considered independent of their implementation by words, we are referring to the syntax of notions of the specialist. This syntax of notions might be called "technese", for lack of a suitable existing term. We may talk, for example, about the technese of the chemist, or the lawyer, or the electrical engineer.

For each kind of technese, each notion may be expressed in the words of any desired language. The association of words and notions will of course be typical of a given field, and the more specialized the field, the more complex may be the notion expressed in a single word. It must be emphasized that language per se remains incidental. The notions, which are the essential elements in all technese, are assumed to be independent of any language.

Beyond the individual special fields, a final grouping would be required to embrace the totality of special fields. The notions to be applied at this level would necessarily be more general and the process of matching would be carried out by way of appropriately broader notions.

Besides the hierarchical organization just described, there is another kind of division which should be introduced to facilitate the adjustment of a system to the constant expansion of knowledge and the adaptations and changes of language that go with it. This may be done by starting a new division or 'age class' of documents at given intervals, as time progresses. For each new interval the system would have been updated to reflect, for the ensuing period, the changes which have occurred during the preceding period. The process of searching would then be performed first for the current period, then for the preceding period and so on, and to the extent dictated by the results obtained.

The use of age classes seems to be the only method by which a collection may be divided into mutually exclusive sections. The searching of a collection in retrogressive steps or by predetermined age groups is bound to shorten the average time of a search. It also appears useful in many instances to search the most recent literature first.

The above system of notions and their degree of relatedness is not necessarily the sole system by which comparable patterns may be derived. Certain classes of information elements such as names or symbolism of structure might demand rather specific identifications. The notations used to represent these elements would assume the same status as that accorded to notions.

Section IV

The Limitations of Serial Communication

The process of communicating notions by means of words can only be performed in serial fashion. In order to overcome this basic limitation, intricate devices have to be incorporated into a language to instruct the addressee how to relate notions in other ways than that given by the linear sequence of words. By means of so many additional words, the addressee is told how to construct a mental image of the multi-dimensional conception of the idea being communicated. Since these instructions may become rather involved and subject to misinterpretation, it is advantageous to utilize pictorial presentations. When thus supplemented, serial language lends itself much more readily to the investigation and description of multidimensional relationships.

This limitation of serial communication and its associated problems also inhere in data processing machines. Communication is carried out on a serial basis in the same sense as amongst humans. When it comes to pictorial representations, the machine is at a disadvantage, at least at the present state of the art. The best that can be done is to instruct the machine to create the equivalent of a mental image and to further instruct the machine to analyze and understand all the many relationships that are contained in this multidimensional representation. For a machine to do this, it must have an internal memory where it can store the representation and analyze it over and over again in accordance with a specific program.

The organization and recording of information capable of being analyzed in the above fashion, as well as the development of programs directing the machine to do this, is a very exacting procedure. The machine, having only logic to its credit, cannot function unless information and instructions are given it in strictly logical language. A system in which relationships between notions were to be given and explicitly recognized would therefore be dependent upon a major intellectual effort for interpreting meanings and relationships and translating them into unique notations. As with current classificatory schemes, this effort would

have to be repeated for each new document. When it came to the searching operation, inquiries would have to be similarly interpreted and encoded. The machine would then have to recognize similarity of representations through an iterative process of identifying and comparing each of the specific relationships given.

The question arises whether similarity of multidimensional representations might not be established by more direct methods without reliance on an internal memory machine. It might be argued that, while it is true that a given number of various notional or pictorial elements could theoretically be related in countless patterns, only a very limited number of these patterns represent meaningful information. Moreover, each additional pattern, in association, further limits the number of meaningful interpretations applicable in the particular case.

On these grounds it would be possible to disregard specific and explicit relationships and merely investigate whether certain elements happen to be associated and to what degree. Such a substitution of statistical for critical criteria would facilitate the establishment of similarity by matching. The more two representations agreed in given elements and their distribution, the higher would be the probability of their representing similar information. The actual matching process would be performed through a serial scanning of records. It would not be required that the machine used be capable of temporarily storing blocks of information in an internal memory.

As shall be seen, the type of scanning suggested is applicable to the statistical searching system presented in the following sections. The system concerns itself mainly with information represented by the written word. In the above discussions, however, reference was made to pictorial representations without indicating how these might be organized either for the purpose of exhaustive analysis by machines with internal memory or scanning and matching on a statistical basis. By way of example, the reader will find the first kind of system presented in a paper by Ascher Opler[1] and the second kind in a paper by H. P. Luhn[2].

---

1 - Opler, Ascher, The Dow Chemical Co., Pittsburg, Calif., "A Topological Application of Computing Machines", Proceedings of the Western Joint Computer Conference, Feb., 1956.

2.- Luhn, H. P., IBM Research Lab., Poughkeepsie, N.Y., "A Serial Notation for Describing the Topology of Multidimensional Branched Structures (Nodal Index for Branched Structures)", Dec., 1955.

Section V

The Organization of a Statistical Searching System

Objectives

The primary objective of the proposed system is to minimize the intellectual effort of professionals at the document encoding stage of the system so that this day-by-day routine may be performed by non-professional personnel and in accordance with a few simple rules. Instead, this intellectual effort of professionals is shifted to and concentrated at the creative stage of setting up the system itself. This effort will be quite substantial when a system is first installed and will call for above-average talent. Thereafter, a moderate effort will be required periodically to update the system.

It is also an objective of the system to gain simplicity of the encoding operation at the expense of the searching operation proper. In this case, however, the extra effort is mainly assigned to the mechanical means of serving the system.

Creating a Dictionary of Notions

The procedure to be described is similar to the one used by P. M. Roget for compiling his Thesaurus of English Words.[1] Roget created categories of words that had a family resemblance on a conceptual level. He arrived at approximately 1000 of these categories for the entirety of experience. Under such a category as space he lists all words and phrases that include any notion of spatiality. This procedure, as adapted here, also relies to a moderate extent upon the techniques used in the preparation of concordances of significant works in literature. Its virtue is that it provides for the greatest possible extent of mechanization. In the form presented it is most applicable to a collection of documents embracing a specialized field that would be normally pertinent to a research activity serving an industrial concern.

---

1 - The particular format of thesaurus here referred to is that of Roget's Pocket Thesaurus, Cardinal Edition C-13, Pocket Books, Inc., New York. Also, Roget's International Thesaurus, Thomas Y. Crowell Co., New York.

The first step in the procedure is the establishment of a basic sample drawn from the collection (See Fig. 2). Since the system is to be a dynamic one, such a sample should consist of the 'youngest' age group, comprising all accessions from the present back to a judiciously selected date. The choice of this date should in part be governed by the number of documents obtained thereby, a figure which in turn depends on budgetary considerations.

The next step consists of transcribing the sample documents into punched or magnetic tape, i.e., into a form which will permit subsequent mechanical operations on the information. Inasmuch as it is desirable that certain grammatical features of words be recognized in subsequent steps of the procedure, it would be advantageous to identify certain classes such as nouns, adjective qualifiers, and names by special symbols. Eventually these differentiations may be determined by machine, as they will have to be when the art of machine translation is perfected.

The third step is the preparation of a card index of all transcribed sentences. A concordance that can then be worked out with the aid of these cards will result in the grouping of words of similar or related meaning into 'notional families'. This is so similar to the work required for the creation of Roget's Thesaurus that basic organization of his book may well serve as the skeleton for this process.

The formation of notional families constitutes a major intellectual effort to be undertaken by experts thoroughly familiar with the habits of communication amongst people associated with the special field from which the subject literature stems. It would be the endeavor of these experts to differentiate the notional families in a manner which will bring about the resolution of the material in terms of an optimum number of equally weighted elements. For instance, in a field that specializes in electricity the notion 'electricity' would be common to most documents and therefore be worthless as a discriminating element. On the other hand, in this same field, the notion 'butterfly' would be entirely too specific an element to warrant the establishment of a separate notional family for it. Instead, the notion 'electricity' would have to be broken down into an appropriate number of subnotions in accordance with qualifying adjectives that might accompany it, while the notion 'butterfly' would have to be relegated to a notional family of broader aspects, such as the notions 'insects', 'animals', or 'living things', depending on the overall frequency of occurrence of such notions.

INFORMATION SEARCHING SYSTEM
CREATION OF DICTIONARY

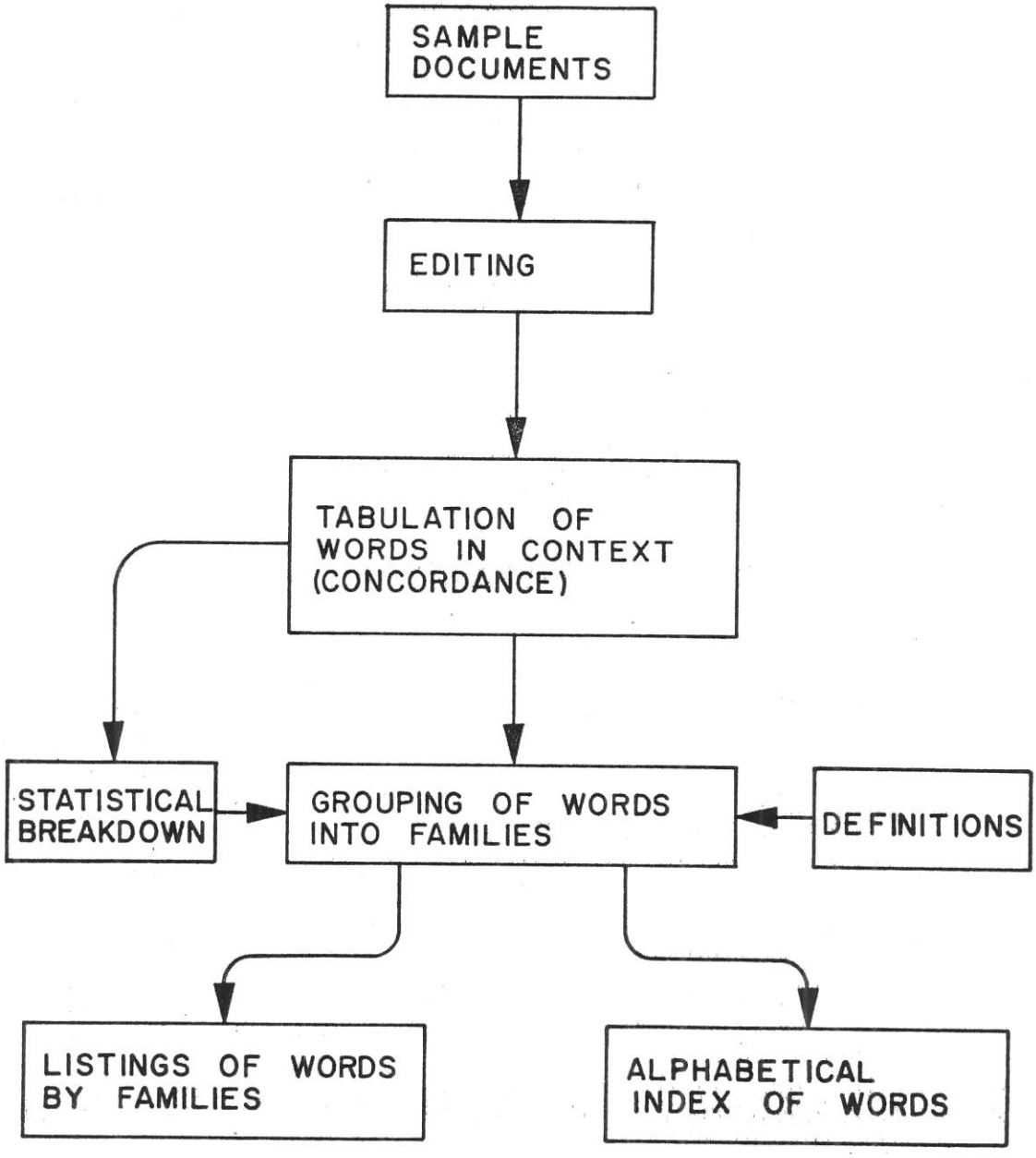

FIG. 2

If it is assumed that nouns (including gerunds) and, where necessary, qualified nouns are capable of providing an effective set of discriminating notions, such nouns would then be grouped into notional families in accordance with the principles established by Roget. The physical result would be a dictionary in two parts. The first part would be the listing in some systematic order of the notional families, each identified by an index symbol such as a number or key word. Each of these would represent a listing of the words from the sample documents which are related with respect to the notion they were made to express. If more than one language is involved, the words within a family might be segregated by language. The second part would be an alphabetic index of the words occurring in the first part, giving the key word and index number of the one or several notional families of which the given word is a member. This index may also be segregated by language.

As far as intellectual effort is concerned, the establishment of notional categories at the word level, as practiced in this system, appears to have advantages over the development of classification or subject headings. It is estimated that the number of notional categories required will be less than a thousand and that this number will grow at a very low rate. In the case of classifications or subject headings, the number and growth rate should be substantially higher. It is also believed that it is easier to agree on approximate meanings of words than on the definitions of exact classes of subject headings and their subsequent interpretations. Lastly, the reduction in the effort required to maintain and update the system should prove significant.

The Encoding of Documents

The encoding of the documents of the sample may now be carried out with the aid of the dictionary of notions. This process consists of recording each document in terms of notional elements and thereby creating patterns which later on will serve as a means of recognizing similarity of various degrees between documents.

This process might best be carried out on the basis of the prevalent patterns of literary organization. Implied here is the capability for recognizing various levels of the relatedness of notions as reflected by the author's formation of sentences, paragraphs, chapters, etc. There is also the probability that the more frequently a notion and combination of notions occur, the more importance the author attaches to them as reflecting the essence of his overall idea.

At this point the question arises as to how specifically notions and their relationship have to be encoded to arrive at a practical measure of comparison. The answer can probably not be given without a background of practical experience. It seems sensible to start with a broad system and to let experience prove whether and where refinements are needed. If, on the other hand, a system of high discriminating power operates satisfactorily, it may be difficult to determine to what degree discrimination could be relaxed without penalizing results. Inasmuch as the following procedures are meant to be performed entirely by mechanical means, a revision of a system is not as formidable an undertaking as may otherwise appear. The level on which the process of encoding will be illustrated should therefore be considered as one of the several possible levels of discrimination.

At the level chosen, concern is principally with notions related by the author within the division of a paragraph. A notion occurring at least twice in the same paragraph would be considered a major notion. A notion that also occurs in the immediately preceding or succeeding paragraph would be considered a major notion even though it appears only once in the current paragraph. Notations for major notions as just defined would then be listed in some standard order as representative of that paragraph and set apart from similar notations of other paragraphs by an appropriate division mark. It might be desirable to overcome extreme leanings of authors to very short or very long paragraphs by editorially combining short paragraphs or splitting long ones, or by some other method of division.

The notations would also include names, if present, within a paragraph. Moreover, because of the special care taken by the author in wording titles, headings, and resumes, it might be proper to rate the affected words as major notions. Formulas, tables, diagrams, and other representations not amenable to the procedure just described may call for other special methods of recording not within the scope of the present discussion.

All encoding would be carried out by a data processing machine having a direct-access information storage and look-up device. The dictionary index would be entered in one portion of the storage device and the document file previously prepared for making the concordance would now be processed in accordance with an encoding program. Each noun within a paragraph would be looked up and the corresponding family number or numbers be extracted from the index storage and stored in a separate storage portion. The machine would then determine which of the words have a notion in common by comparing each family number in turn with the other numbers stored for nouns of the current, as well as the preceding and succeeding paragraphs. Since matching family numbers are indicative of a major notion, they would be entered in a third portion of the storage device. Words which fail to attain the status of major notions would not be entered.

Upon exhaustion of this procedure and the recording of associated names, the encoded version of that paragraph would be ready for transfer to a permanent storage medium such as a reel of magnetic tape. Family numbers stored in the second portion of storage would be retained only during the analysis of the next-following paragraph. As the encoding of the new paragraph proceeds, the family numbers of its words would now also be compared with those of the immediately preceding and succeeding paragraphs to find common notions. This process would be repeated for each successive paragraph until the end of the chapter has been reached. The end result would be a mechanically prepared notional abstract.

After all the originally recorded documents were encoded in the manner just described, the input phase of the system would be complete as far as the sample documents are concerned. The encoded records would be contained entirely on magnetic tape reels and ready for subsequent searching operations.

Information Searching Procedure

When it is desired to discover amongst the encoded documents those which have a bearing on a given idea, the inquirer would be asked to prepare a document similar in format to the documents of the collection. This document would be in the form of an essay in which the inquirer describes what is prompting him to seek information, giving as many details as come to his mind concerning the problem, objectives, assumed or planned approaches to solutions, references to other authors, subjects, or, in short, anything that he feels might have a bearing on his problem.

This 'inquirer's' document would then be encoded in the exact same manner that the documents of the collection have been encoded, and the resulting notional abstract would be set up as a question pattern in a data processing machine of adequate functional ability. This procedure would overcome the inconsistency of method that would exist if a question pattern were artificially assembled by a mediator. A standard program could serve to direct the machine to compare the question pattern with the notional patterns of the documents of the collection. Since an identical match is highly improbable, this process would be carried out on a statistical basis by asking for a given degree of similarity. This gauging of the question might be accomplished in many ways; the most direct would be to ask for a match of a given fraction of the notions.

If the machine is equipped to transcribe each complying pattern onto a new tape, this searching process could be carried out most efficiently by first making an appropriately broad search to assemble a likely

selection of documents on a separate tape. With the aid of this special tape the search might be extended in several ways. One would be to change the fraction of coincidence required; another to modify conditions through requests that specific notions or combinations thereof be present.

It is a part of the system that a certain degree of intellectual effort be spent at the output phase. This would involve the sampling of the selected documents in order to determine the effectiveness of the search and to discover clues which might serve to improve the quality of response in a certain direction upon a second or subsequent runs, until an optimum return is assured.

An important adjunct to the system would be a supplementary index which would serve to 'blow up' certain notions and names in terms of definitions, component parts, or other closely related notions. This index would be used, if indicated by the result of the search, to amplify and refine the encoded version of the inquirer's essay and thereby raise its degree of specificity.

The opposite function of finding definitions or names that might apply to certain combinations of notions or names would be derived from a special collection of records of definitions. Depending on the outcome of the search, or as a first step of procedure, the question pattern might be compared with this special collection in order to pick up broader terms that could thereafter be included in the question pattern for the purpose of decreasing its level of specificity.

Maintaining, Expanding, and Updating the System

To insure a dynamic system, it is important that a means of producing quantitative information on the dictionary activities be provided. Provision should be made to register the number of times each word is looked up in the index and the number of times each family number has been used for encoding. Such a record would be an indispensable guide for making periodic adjustments.

Documents not contained in the original sample would be introduced into the system in the same manner as the sample documents except that they would not be analyzed for concordance purposes. Words not found in the dictionary would be noted by the machine during the encoding process and spelled out for subsequent manual editing. This would involve the assignment of a notional family or families to the new word and adding it to the index. As a matter of course, each new document would also be submitted to a searching operation to discover duplications and to obtain a measure of the discriminatory characteristics of the system as the collection grows.

Because the system is predicated on the division of literature into special collections, it would seem that an encoded record of collection might be unintelligible in terms of another. Not only can such mutual exclusiveness be prevented but legitimate code sharing can actually be facilitated. In a basic searching system of the type discussed, the various encoded collections would differ only in their dictionaries.

The method of transcribing the original document would be identical for all collections. If, therefore, it were desirable to incorporate a document already entered in one collection into another, no more would be necessary than to procure the original transcript. Inasmuch as the encoding of a document would be a fully automatic machine process, this transcript need only be submitted to this process with the appropriate notion index in control. Under this scheme, then, the original transcript is endowed with the universality desired, and it becomes a small matter to incorporate it, in proper 'language', into another special collection.

The same condition exists when searching. If it is desired to extend a search to another field, it is only necessary to re-encode the transcript of the inquirer's essay with the aid of the proper notion index and to proceed with searching in the other collection.

Section VI

Conclusion

The soundness and practicability of the approach and system proposed in the preceding discussion has not as yet been proved by any full-scale operations. However, some modest experiments have produced encouraging results. There are as yet many unanswered questions, such as whether nouns and adjectives or other portions of sentences furnish the most practical and, at the same time, the most effective discriminating elements for recognizing similarity of information by mechanical pattern matching. Another question is how few notional families are required, what various degrees of specificity these families should have, and how many of them need be recorded for a given kind and size of document collection.

If the proposed approach is shown to be practical, it would no longer be necessary to recognize the meaning of information for the purpose of encoding. A great advantage of this would be the elimination of professional effort from the major part of the encoding routine. However, an even greater advantage would be the assurance that the idea content of information was not being tampered with. The author of a document would reign supreme in that his ideas would not be narrowed, biased, or distorted through the intervention of an interpreter. The proposed method would merely record the various degrees of relatedness between given notions that are implied by the author's physical grouping of the corresponding words into sentences, paragraphs, and chapters.

H. P. Luhn
Jan. 30, 1957