REVIEW OF INFORMATION RETRIEVAL METHODS

H. P. Luhn

IBM

RESEARCH CENTER

INTERNATIONAL BUSINESS MACHINES CORPORATION
YORKTOWN HEIGHTS, NEW YORK

REVIEW OF INFORMATION RETRIEVAL METHODS

By

H. P. Luhn

International Business Machines Corporation
Research Center
Yorktown Heights, New York

ABSTRACT: Information retrieval objectives are outlined and the various systems that have been employed in their achievement are categorized. The encoding of source documents to prepare them for retrieval operations is also discussed, with emphasis here placed on statistical methods.

Research Report
RC-59
October 1, 1958

REVIEW OF INFORMATION RETRIEVAL METHODS

By

H. P. Luhn

1. Objectives of Retrieval Systems

In considering the subject matter of retrieval systems, two basic categories may be established, namely accountable and non-accountable information.

Accountable information deals with quantities of things. Systems serving this type of information should reflect the changes of such quantities by addition or subtraction as caused by external events. There are two elements involved; the identification of the thing and a statement of the value or quantity regarding this thing. Generally, the retrieval of this kind of information is characterized by the fact that the information wanted is the quantity and that the argument of inquiry is the identification of the thing only.

Examples typical of this category are:

Accounts of money

Inventories of material

Inventories of personnel

Retrieval systems in the area of non-accountable information serve a wide variety of purposes. They may generally be classified as follows:

a. Establishment of equivalence

Here a single argument is used to ascertain a property, usually expressed in terms of some measure. It is typical of such systems that the functions derived may be compared with each other by a common scale and then evaluated by a simple arithmetic operation. Examples of such conversions are: articles - price, material - weight, towns - population, target - location (within a coordinate system). A conversion table is probably the simplest form of such a system.

b. The identification of individual attributes

These systems comprise indexes of assignments and attributes where the relationship between attributes is not intended to be measurable by direct means. Examples of such relationships are: names - man number, properties - ownership, men - assignment. Dictionaries and directories are other applications. A typical feature of these systems is that they are frequently used in two versions so that the argument in one version becomes the function in the other version and vice versa.

In the systems thus far enumerated it is possible to order the arguments by some scale, which permits the use of look-up operations. This procedure minimizes the process of locating an argument. Inversion of the lists by changing the roles of argument and function overcomes the time-consuming process of searching within the unordered portion of the lists represented by the functions.

c. Discovery of inclusion

Retrieval systems of this type serve the function of determining whether a given item is included within a list of items. In systems of the kind just mentioned this can normally be accomplished by success or failure of looking up a given item in an ordered list. However, a more typical requirement exists where each item on the list is identified by an enumeration of its component parts. Here it becomes necessary to ascertain whether one or several given component parts are included in any of the items listed. Such assemblies of discrete elements are exemplified by bills of materials, composition of matter, bodies of people or things, functional structures, catalogs and others. The creation of inventories, listing for each discrete element the items of which it is a part, will permit retrieval by look-up. The result is an inverted list several times the length of the original index.

d. Discovery of similarity

While in the foregoing system, it was required that a given number of elements be completely included in an item, it is at times desirable to determine whether a given fraction of a number of given elements is present within items of a list. In order to retrieve items answering this requirement, systematic look-up, as previously employed, cannot be carried out. An additional problem arises here in that an intermediate inventory has to be taken of the occurrences of each single element in the various items. This process might possibly be complicated by conditional requirements of retrieval, i.e., considerations which must be expressed as logical sums, products or complements. It is apparent that the creation of individual lists becomes such a tremendous task as to become impractical, and that searching in unordered portions of the system, in one form or other, will have to be employed in the retrieval process.

It is sometimes necessary to discover similarity by way of similarity of values of discrete elements. Special designations for appropriate intervals are used to overcome some of the complexities of this kind of retrieval. Examples are: test reports, personnel records, medical records, weather reports and the like.

The discovery of similarity may concern items identified by elements which cannot be uniquely defined and which cannot be bracketed by a linear-range device. This situation exists where the identifying elements are members of two or more classes. Examples are free style statements such as messages, reports, and other communications of literary information whose format cannot be controlled.

Retrieval systems for this type of information must be able to overcome variations in the use of language and to facilitate the recognition of synonyms and word usage. Normalization of this type of information requires the extensive use of dictionaries or thesauri. The listing of material of this type for easy access is a difficult problem which cannot readily be solved by such intermediary devices as classifications and subject headings.

e. Discovery of coincidence

A desirable function of an information retrieval system is to facilitate discovery of coincidences amongst listed items. This function also assists in discovering duplications, inconsistencies, and errors. A further requirement is to discover non-obvious similarities and coincidences for the recognition of trends. New information may be synthesized with the aid of such findings.

The foregoing analysis points up the wide range of objectives of information retrieval systems. This is parallelled by a wide range of complexity of systems which are to serve specific objectives.

2. Retrieval Methods

There are two basic methods of retrieving information by means of identifying terms, irrespective of how such terms have originally been derived. These methods may be stated as follows:

1. Retrieval of information by look-up in an ordered array of stored records.

2. Retrieval of information by search in a non-ordered array of stored records.

3. A combination of 1 and 2.

Retrieval by look-up. Retrieval by look-up presupposes that a record is identifiable by a single argument. This argument may serve as an address indicating where in storage the related record is located. Retrieval may then be accomplished by going to the indicated address. Single or multiple look-up operations may be required.

For single look-up, access to given addresses is subject to the restrictions of realizable look-up storage devices. The most elementary form is an open-ended file of manipulative record carriers disposed in one, two, or three dimensions. Whether the records are ordinary index cards or machine-readable record elements, fixed or movable, a delay is encountered between the instant the address of a record has been determined and the instant the record becomes accessible for reading. This delay or "access time" varies greatly amongst various devices and is occasioned by the physical bringing together of the reading means and the designated record. In the case of fully electronic devices, where such "bringing together" is performed by switching, the access time may shrink to microseconds.

In those cases where more than one argument serves to identify a wanted record, look-up may be accomplished by storing duplicate records at each of the corresponding addresses. The larger these records are, the larger must be the storage device and the access time will be unfavorably affected.

This situation is alleviated in some systems by cross-indexing so that the full record can be stored at one address only. Since two look-up operations are usually required in this system, access time is increased.

The multiple look-up procedure is typical of retrieval schemes known as the Batten or Peek-a-boo devices and the Uniterm devices. Records are identified by descriptors and each record contains the numbers of the items which are characterized by each descriptor.

In the case of the peek-a-boo system, the storage device consists of a collection of cards whereon item numbers are represented by holes in pre-assigned locations. Retrieval of information is accomplished by drawing from the file those of the descriptor cards which characterize the inquiry. The cards thus selected are superimposed, and the locations where holes coincide on the cards may be ascertained by optical means. This process is truly a searching operation and time is required to check each location on the card for the condition of coincidence. The various matching hole locations must first be translated into the reference number of the items and then listed in an appropriate form. With the aid of this list the records of the original items are retrieved. This involves a second look-up procedure in a properly ordered file.

Since the size of cards in such a system is a function of the total number of items contained in the collection, this principle has drawbacks when it comes to mechanization by data processing equipment. In a mechanical look-up device, each descriptor will constitute an address and in each address an inventory of the serial numbers of the items characterized by the descriptor in question will be built up. This means that for each new document,. entries have to be made of the serial number at each of the storage locations of each of the descriptors.

A retrieval operation will consist of a succession of look-ups, one for each descriptor of the question. In order to determine the coincidence of serial numbers, the list of the numbers of the first descriptor will have to be compared with the list of the numbers of the second descriptor and those numbers which coincide in both of these lists will have to be stored. The thirddescriptor will then have to be looked up and the associated list compared with the list of previous coincidences. This process of look-up and search will have to be repeated as many times as there are descriptors.

There are many problems connected with this type of information retrieval. To remove obsolete items from storage, the serial number of the effective item would have to be removed from the lists of all of the affected descriptors. This is an operation of the same magnitude as that of originally entering an item. An additional operation will be necessary to reconstitute the various lists in order to keep the serial numbers in closed order.

Peek-a-boo type retrieval systems are capable of producing answers on the basis of the conjunction of all of the descriptors. If it is essential that relationships between such descriptors be specifically expressed, this will have to be accomplished by additional operations, extending the time required for completing a retrieval operation.

The addition to the system for realizing this function consists of an additional storage section where each item is stored under its serial number followed by a list of all of the descriptors which have been used to characterize the item. The relationships between the descriptors are here explicitly indicated and would become available for comparison with the explicit relationships established in the inquiry. This arrangement will make it possible to take the answers produced by the process just described and to consider them contenders for the final selection. Each document would then be looked up under its serial number, and the original question, including relationships, be compared with the notation recorded at that address.

Retrieval by Search. The organization of these systems is characterized by the fact that records are entered into the system as units and that the elements of such records are not distributed. Therefore, the store of a retrieval system of this kind consists of a chain of subsequent records, usually ordered by time of accession. Retrieval is accomplished by first setting up the identifying terms and their relationships to each other, and then comparing this data with each record of the collection in serial order. Retrieval time is here a function of the size of the collection.

Again, there are two basic systems of searching:

1. Search by scanning.

2. Search by iterative analysis.

In search by scanning, the system of information retrieval presupposes a coding structure which permits the making of decisions by a simple matching operation. Therefore a device performing this type of search does not need the capability of actually reading the stored information. Instead it is merely necessary to set up the code patterns constituting the elements of the question and to search for the presence of these code patterns while scanning the file. The events of matching are registered by some logical device and if these events occur~~s~~ in the desired configuration within a record, selection of the affected record is then brought about. In systems of this kind, the scanning of the collection may be performed at maximum speed. Where obsolescence is a function of time and where the collection is ordered in sequence of accession, the volume of stored information may be readily adjusted by dropping portions of the file by age.

Search by iterative analysis is called for where the subject matter is of such a form that the derivation of simple serial codes is not practical. Systems which answer this limitation are characterized by intermittent operation. The searching device first reads a portion of the serial record and stores it in an internal memory. Subsequently, it performs iterative searching operations on the information thus stored. These two modes follow each other as the search proceeds. Under certain conditions records required for iterative analysis can be made shorter than the records for continuous scanning. Again there are many variables which will determine which of the two methods should be employed.

Combination of Look-up and Search. From the above descriptions it is apparent that a pure look-up operation may be realized only in those cases where a single function is wanted. As soon as the record stored at a given address contains a multiplicity of identifying terms and whenever the results of retrieval are made dependent upon the co-occurrence of several of such terms, a searching operation is involved.

In the case of retrieval by search, no look-up operations are involved. While this is true for continuous scanning systems, it might nevertheless be desirable,in intermittent systems, to perform look-ups in conjunction with the iterative analysis of records while internally stored. Only a thorough analysis of a given problem can determine the relative merits of the two systems.

The access time to wanted information also vaires considerably amongst the systems. While it might be generally true that look-up systems will furnish speedier access, the application of multiplexing techniques to searching operations will bring the average access time of searching systems into comparable range. Multiplexing may consist either of performing several searches on a whole file simultaneously or by performing identical searches on an appropriate number of subsections of such a file.

3. Encoding Methods

A key problem in the development of an information retrieval system is the formulation of methods by which source information may be arranged, transformed, or reduced so that it may be effectively operated upon by devices. There are a variety of approaches that may be taken, and in the following review the criteria of a number of systems are enumerated. The over-all consideration here is to determine what the minimum amount of information can be for operating a system efficiently.

The simplest system is one where information is introduced into the system without change and where the system is equipped to use it without modification. For such a system to be efficient it is necessary to have strict control over the format in which information is first organized and recorded at the point of origin. Means for accomplishing this include well-established code words, forming part of a technical language of operations; use of style manuals; and the use of forms into which variable information is entered by checking yes-no questions or by inserting variable information in predetermined locations on the form.

In cases where such control cannot be maintained at the source of origin, recording procedures have to be provided at the input of the system proper. There are three possible systems of encoding, namely:

a. Human encoding,

b. Human encoding assisted by machine operations,

c. Fully automatic encoding.

The assignment of code terms is based on the availability of dictionaries and manuals, and the efficiency of the system is dependent upon the efficiency of such devices. The assignment of the code terms may be based on:

a. Selection of parts of the original record, such as key words,

b. Classifying certain aspects of the records,

c. Classifying the complete record by such devices as subject headings.

There are several methods of deriving classifications such as subject headings, dictionaries, etc., which may be defined as follows:

'Adopted', i.e., compiled externally for rather broad uses.

'Synthetic', i.e., created by reasoning, judgment, and experience with regard to the subject matter of the affected specific operation.

'Native', i.e., derived from the specific operation by statistical analysis and in a manner which gives each class a comparable degree of discriminating power.

4. Statistical Methods

The application of statistical methods for optimizing the assignment of code terms is an effective means for arriving at optimum dictionaries. In machine systems the periodic analysis of code usage is a means by which updating of the system becomes practical.

The ultimate goal in encoding systems is the completely automatic performance of the required steps. Here the initial requirement is that information be available in machine-readable form. Upon introduction of this type of record into the system, the character of the record is determined by machine analysis, and the record is then subjected to the appropriate automatic encoding operation.

Such automatic procedures will result in consistently uniform encodings and eliminate variations and errors attending human operations. They will also accelerate the encoding procedure considerably.

In IBM, research experiments in information retrieval have shown that coding, indexing, and abstracting can be performed wholly by machine. However, such results are forthcoming only when mechanized statistical analyses can be performed on whole documents.

Similarly, searching techniques can be automatically performed only when inquiries are themselves machinable and mechanically encoded. Thus for the design of any fully efficient retrieval system, two types of analysis are required:

a. A statistical analysis of the actual documents in file.

b. A logical and statistical analysis of the inquiries to the file.

Many factors influence the choice of information retrieval techniques. For example, documents of fixed format do not generally lend themselves to statistical analyses but should be investigated chiefly by logical techniques. Also, automatic encoding and indexing techniques for free format documents will vary with document size, ratio of total different words to total word occurrences, number of common words per document, etc.

In general, the logical and statistical properties of the documents themselves dictate the specific techniques employed. Similarly, searching techniques are dependent upon the specific logical and statistical properties of inquiries to the file. Logical inclusion and exclusion among the categories specified in typical inquiries will influence (in an obvious way) the searching procedures used. By means of appropriately constructed mechanized thesauri, time and effort will be saved by reducing the number of categories for machine search purposes.

Another important feature of the statistical approach to information retrieval is that the particular language in which the documents are phrased does not seriously affect the operation of the system. It seems clear that either documents or inquiries in any Indo-European language will be handled in a similar manner. At the present time, there is no available technique which provides for a complete mechanical translation of documents and inquiries; however, it is assumed that with the aid of specialized thesauri, translation of key terms can be accomplished. By this means, the searching and coding techniques would be able to provide documents in any language relative to a particular inquiry, even though the documents themselves would eventually have to be translated to secure their full information content.

Auto-abstracting techniques, currently being used on English texts, will, in the near future, be tested on samples of documents written in French and German. Only in languages where the entire linguistic background differs, such as Arabic, Chinese, Swahili, etc., will entirely new procedures have to be devised.

For files containing fixed format documents, it appears that statistical techniques will not provide the required information. In most cases of this sort, the background producing such files will have to be examined in detail by logical techniques.