IBM JOURNAL OCTOBER 1958

H.P. Luhn

A Business Intelligence System

Abstract: An automatic system is being developed to disseminate information to the various sections of any industrial, scientific or government organization. This intelligence system will utilize data-processing machines for auto-abstracting and auto-encoding of documents and for creating interest profiles for each of the "action points" in an organization. Both incoming and internally generated documents are automatically abstracted, characterized by a word pattern, and sent automatically to appropriate action points. This paper shows the flexibility of such a system in identifying known information, in finding who needs to know it and in disseminating it efficiently either in abstract form or as a complete document.

Introduction

Efficient communication is a key to progress in all fields of human endeavor. It has become evident in recent years that present communication methods are totally inadequate for future requirements. Information is now being generated and utilized at an ever-increasing rate because of the accelerated pace and scope of human activities and the steady rise in the average level of education. At the same time the growth of organizations and increased specialization and divisionalization have created new barriers to the flow of information. There is also a growing need for more prompt decisions at levels of responsibility far below those customary in the past. Undoubtedly the most formidable communications problem is the sheer bulk of information that has to be dealt with. In view of the present growth trends, automation appears to offer the most efficient methods for retrieval and dissemination of this information.

During the past decade significant progress has been made in applying machines to the processes of information retrieval. Automatic dissemination has so far been given little consideration; however, unless substantial portions of human effort in this area can be replaced by automatic operations, no significant overall improvement will be achieved. Even the information retrieval processes mechanized so far still require appreciable human effort to organize the information before it is entered into machines.

It is believed that techniques now being developed will greatly contribute to the solution of the problem by extending automatic processes to the preparatory phases of mechanical information-retrieval systems, to the are of dissemination and to associated functions. Ideally, an automatic system is needed which can accept information in its original form, disseminate the data promptly to the proper places and furnish information on demand.

The techniques proposed here to make these things possible are:

1. Auto-abstracting of documents;
2. Auto-encoding of documents;
3. Automatic creation and updating of *action-point* profiles.

All of these techniques are based on statistical procedures which can be performed on present-day data processing machines. Together with proper communication facilities and input-output equipment a comprehensive system may be assembled to accommodate all information problems of an organization. We call this a*Business Intelligence System.*

Objectives and principles

Before the system operation is described, the term *Business Intelligence System*should be defined and the objectives and principles stated.

In this paper, *business* is a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, et cetera. The communication facility serving the conduct of a business (in the broad sense) may be referred to as an *intelligence system.* The notion of *intelligence* is also defined here, in a more general sense, as the "ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal." (1)

The term *document* is used to designate a block of information confined physically in a medium such as a letter, report, paper or book. The term may also include the medium itself. The objective of the system is to supply suitable information to support specific activities carried out by individuals, groups, departments, divisions, or even larger units. These are the *action points* previously referred to. To this end the system concerns itself with the admission or acquisition of new information, its dissemination, storage, retrieval and transmittal to the action points it serves.

More particularly the object of the system is to perform these functions speedily and efficiently, taking advantage of novel procedures which utilize the inherent capabilities of electronic devices.

One of the most crucial problems in communication is that of channeling a given item of information to those who need to know it. Present methods of accomplishing this are inadequate and the general practice is to disseminate information rather broadly to be on the safe side. Since this method tends to swamp the recipients with paper, the probability of not communicating at all becomes great. The Business Intelligence System provides means for selective dissemination to each of its action points in accordance with their current requirements or desires. This is accomplished by the mechanical creation of *profiles*reflecting the sphere of interest of each point and by updating these profiles as dictated by changes in the attitude of the respective action points and as recorded by the system on the basis of certain transactions.

Another problem in communication is to discover the person or section within an organization whose interests or activities coincide most closely with a given situation. Presently, the difficulty of finding such relationships often results in improper decisions, wrong actions, inaction, or duplication. An objective of the Business Intelligence System is to identify related interests by use of profiles of action points. The problem of discovering information which has a bearing on a given situation has probably received the most attention in recent years, and various mechanical systems have been developed and put into operation. This phase of communication is commonly referred to as *information retrieval* or, more broadly, as the *library problem.* Information retrieval is necessarily a major function of the Business Intelligence System. Means are provided not only to integrate this function with the rest of the system but also to produce additional useful functions, as will be described later.

The achievement of these objectives is governed by principles essential to effective service and convenience of the user. Some of these are listed below:

1. Information admitted to the system includes communications, addressed to action points individually, which contain information of potential interest to other action points.

2. New information which is pertinent or useful to certain action points is selectively disseminated to such points without delay. A function of the system is to present this information to the action point in such a manner that its existence will be readily recognized.

3. Transmittal of information either as a result of dissemination or of retrieval is to be guided by progressive stages of acceptance by an action point. This procedure saves the recipient's time by reducing the amount of material to be transmitted and eliminating the non-pertinent material.

4. The system is to provide means for quickly discovering similarity of interests and activities that might exist amongst action points so that subjects and problems of common concern may be discussed and advanced through direct interchange of ideas between such points, if so desired.

5. The system is not to impose conditions on its user which require special training to obtain its services. Instead the system is to be operated by experienced library workers. Thus, in the case of an inquiry, the user will be required only to call the librarian, who will accept the query and will ask for any amplification which, in accordance with his experience, will be most helpful in securing the desired information.

6. Similarly, information lingering at an action point but of potential value to other action points is mobilized for efficient communication through inquiries of skilled reporters.

Description of the Business Intelligence System

The following description is given in rather general terms, and references to any specific type of *business* have been substantially avoided. Furthermore, the fact that certain devices are being referred to as implementation of the system, should not be interpreted as implying a specific size of the operation.

The description is given in accordance with main functional sections of the system, each illustrated by the diagram. Our assembly of these functional sections into a complete system is shown in Fig. 1.

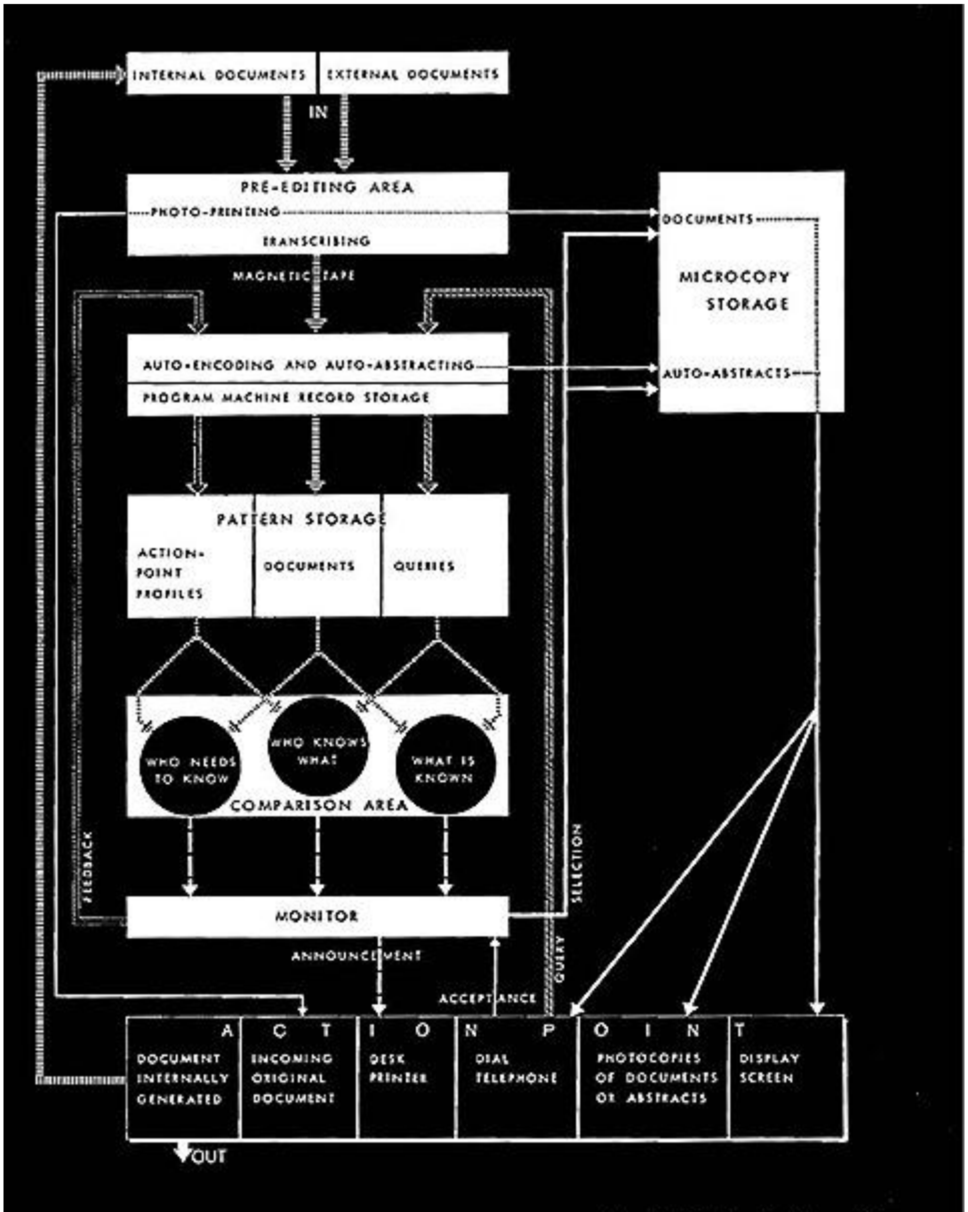

Figure 1 A Business Intelligence System

Document input

Each document entering the system shown in Fig. 1 is assigned a serial number and is photographically reproduced on some medium such as microfilm. In those cases where the document has been addressed specifically to an action point, the original is promptly transmitted to the addressee. In all other cases the original is stored in a file for a reasonably short time and thereafter destroyed, unless there are reasons for preserving it for longer periods.

The microfilm copy of the document is transcribed onto magnetic tape by a human transcriber or a print-reading device. In those cases where the original document is available in machine-readable form, the transcription is done mechanically. The document is now available both as a microfilm copy and a magnetic tape record.

The microfilm copy is then recopied onto the storage medium of a document microcopy storage device. The microfilm record is stored elsewhere to constitute a microfilm master file which may serve to regenerate records in cases of emergency.

The magnetic tape record is now introduced into the auto-abstracting and encoding device. This device submits the document to a statistical analysis based on the physical properties of the text, and data are derived on word frequency and distribution. From these data the device then selects certain sentences of the document to produce an auto-abstract. (2) This is printed out, together with the title, author, and document serial number. This printout is photographically transferred onto the storage medium of the *auto-abstract microcopy storag*e device.

The process of creating auto-abstracts consists of ascertaining the frequency of word occurrences in a document. A predetermined portion of the words of highest frequency is then given the status of significant words and an analysis is made of all the sentences in the text containing such words. A relative value of sentence significance is then established by a formula which reflects the number of significant words contained in a sentence and the proximity of these words to each other within this sentence. Several sentences which rank highest in value of significance are then extracted from the text to constitute the auto-abstract.

As soon as the auto-abstract has been created, the statistical data are further processed to derive an information pattern which characterizes the document. This process of encoding constitutes a further abstraction and involves procedures such as the categorization of words by means of a thesaurus.(3)

Useful patterns may be derived by listing a given portion of the words of highest frequency together with a selection of specific words. The interrelationship of words may also be indicated and certain frequently occurring combinations of words may be noted. Because of variation of word usage amongst authors the normalization of such words becomes an important function of encoding. Index lookup in a thesaurus-like dictionary will replace words, including those of foreign languages, by a notional family designation. The selection of specific words may also be accomplished by index lookup.

The document pattern derived by the above process is then transferred into a special pattern-storage device together with the title, author, and document serial number. This information is stored in coded form on a medium that may be subjected to serial scanning. As an alternative the resulting pattern may be rearranged and be distributed over a storage array to permit random access according to characteristics.

The tape or film transcript of the document may be stored in a library for reference if it later becomes necessary to change the method or scope of encoding.

Action-point profiles

As indicated earlier, one of the basic requirements of the system is the ability to recognize by mechanical means the sphere of interest and the type of activities that characterize each of the action points the system is to serve. This is accomplished by means of an information pattern similar to that of the documents.

Initially, the creation of these action-point profiles is best accomplished by having each action point create a document describing the various aspects of its activities and enumerating the types of information needed. Such documents are then introduced at the input of the system and are identified by action-point designation. The machine-readable transcripts of these documents are then described in connection with the document input. The resulting patterns are then stored in the Pattern Storage area in a special profile-storage device. Also stored, with each of these profile patterns, is the date of entry.

Selective dissemination of new information

Based on the document-input operation and the creation of profiles, the system is ready to perform the service function of selective dissemination of new information.

As soon as a new document has been entered into the system and its pattern developed, this pattern is set up in a comparison device which has access to all of the action-point profiles. The comparisons are carried out on the basis of degree of similarity, expressed in terms of a fraction, for each of the profile patterns. This fraction is subject to change as time goes on, depending upon conditions to be explained later.

Whenever a profile agrees to a given extent with a given document pattern, the serial number, title, and author of the affected document, together with the action-point profile designation, are transferred and stored in a monitoring device. This procedure is repeated for any subsequent similar occasion. The monitor is substantially a random-access storage device and has the functional capabilities of performing inventory operations. In this capacity it will transmit the serial number, title and author of the document in question to the desk printer at the selected action point and keep a record of this transaction. Of the various ways in which such an announcement may be transmitted to the affected action points, the most effective one is by means of a printing device at each action-point location. An objective of the system is to command attention of the recipient. The use of individual printing devices is more effective than are centrally located devices serving several action points.

Selective acceptance of disseminated information

The dissemination of information so far has consisted in furnishing the action point with the serial number, title, and author of documents selected for it. This selection, however, is considered to be a provisional one, and the system withholds any further information if the action point can determine, on the basis of information given so far, that certain of the selected subjects are not of sufficient interest. If an announcement is of interest, and more detailed information on the subject is desired, the system will produce such information on demand. This step is initiated when the action point connects itself by telephone to the monitor and dials the serial numbers of the documents affected. Upon receipt of this message the monitor will relay an instruction to the microcopy storage device to produce photoprints of the auto-abstracts of these documents and to mark them with the action-point designation. The auto-abstracts are then transmitted to the action point either in the form of a paper copy or by speedier means, such as Telefax or TV display.

The action point may now peruse the abstracts to determine which of the documents are desired in their entirety. These decisions are then entered into the system in the form of acceptances. An acceptance is made at an action point by dialing the document number, prefixed by a code symbol, whereupon the monitor will instruct the microcopy storage device to produce a photocopy of the complete document, properly marked with the action-point designation. These photocopies are then delivered to the action point.

The monitor will record the incidence of acceptance by modifying the affected records contained in its storage. At the same time the monitor will also instruct the auto-encoding device to transfer copies of the code patterns of the affected documents to the profile section of pattern storage, together with the identification of the action point involved and the date of transferal.

As a result of these operations the profile of a given action point has been updated to reflect interest in a currently communicated subject. As time goes on there is the probability that an increasing number of new documents will be announced to an action point because of possible shift of interests. In order to avoid such cumulative effects, the system is so arranged that the response to past interests is gradually relaxed. This relaxation is related to the date affixed to each new pattern that is superimposed on an action point's profile. Depending on the age of each of these patterns, an adjustment is made on the fraction of similarity that must be met in the comparison process of new documents. The older the profile pattern, the closer an agreement is needed for selection for dissemination, and consequently the fewer documents are selected. On the other hand those documents selected are more closely related to the original subject.

Information retrieval

This phase of the system concerns itself with the retrieval of those stored documents which might be relevant to a topic under consideration by an action point. The information to be discovered may vary widely and may consist of anything ranging from factual data to an extensive bibliography on a broad subject. Under the supervision of an experienced librarian the process of information retrieval is performed in the following way.

An action point telephones the librarian and states the information wanted. The librarian will then interpret the inquiry and will solicit sufficient background information from the action point in order to provide a document similar in format to that of documents normally entering the system. This query document is transmitted to the auto-encoding device in machine-readable form. An information pattern is then derived from the query document in a manner similar to that used for normal documents.

The resulting query pattern, together with a serial number and designation of the originating action point, is then sent to the queries section of the pattern-storage device. Subsequently, a copy of this query pattern is set up in the comparison device and. is compared with all of the document patterns stored in the document-pattern storage device. This operation is similar to the one described in connection with selective dissemination. In the present case, the query pattern replaces the profile pattern.

Whenever similar patterns are detected by this means, the document designation is transmitted to the monitor, where it is registered and then announced to the action point.

Although the service of a librarian is considered a convenience to the action point, in certain cases, means may be provided at the action-point location to permit direct access to the system. This would be justified where many of the inquiries concern lookup-type retrieval of data.

When an action point desires information relative to a given document, the number of the document at hand would be dialed and instructions for search given to the monitor. Thereupon the monitor would select the corresponding pattern from document-pattern storage and provide instruction for use as a query pattern in the ensuing comparison operation.

Selective acceptance of retrieved information

The considerations which prompted the step-by-step acceptance of documents in the dissemination process are also applied to information retrieval. The processes employed, therefore, are identical.

The function of information retrieval, however, differs from that of dissemination in that the choice is not that of accepting or rejecting one document, but rather a selection of one or several from a special group of potentially relevant documents. Although in some cases a first search may have produced satisfactory references, in other cases the material produced may not be satisfactory. The action point must then relay this fact to the librarian and discuss with him how the searching procedure or the query should be modified so as to improve the probability of getting relevant material.

In those cases where pertinent information has been discovered, the acceptance of the complete documents of such information will cause the updating of the action-point profile, as was the case in dissemination. The query pattern will be impressed on the profile as a matter of course, whether or not the inquiry has been satisfied, so that new documents relevant to the subject of the inquiry will be made known subsequently.

Detection of an action point having given characteristics

In the process of transacting business it is often desired to determine who concerns himself with a given subject. The usual type of question asked is: “Who does or knows a certain thing?” A function of the Business Intelligence System is to answer questions of this type.

The manner in which this function is performed by the system is similar to the information retrieval procedure. However, instead of simulating a document pattern, a profile pattern is developed which represents most closely the characteristics of an action point sought. This synthetic profile is then compared with those in the profile storage and when a given degree of similarity is discovered, the identification of the affected action point is transferred to the monitor, together with the identification of the inquirer point. Thereafter the identities are announced by the tape-printing device at the inquiring action point so that personal contacts may be made.

Document output

The functions described so far have concerned themselves with documents admitted or acquired by the system from the outside. The document-output phase deals with internally generated documents. This type of document is essentially the product of action points and may be addressed to other action points within the organization or to external points. An objective of the system is to facilitate selective dissemination and retrieval of such documents in substantially the same way as for outside documents.

When a document has been created at an action point, a copy is produced, preferably in machinable form. This copy is then dispatched for processing to the input point of the system and the original is sent to the addressee.

Since this type of document is an indication of the interest of the originating action point, the information pattern derived by the auto-encoding process is not only stored in document-pattern storage but also is impressed on the profile of its originator, thereby updating it.

In the dissemination process this internally created document is announced to other action points in the same fashion as were outside documents.

Miscellaneous functions of the system

The comprehensive system for the various functions so far described is illustrated by Fig. 1. A number of additional useful functions which may be derived from the system are briefly described here.

It might be desirable to check each new document for duplication by comparing it with all of the documents in storage. Similarly a list of related documents may be prepared to serve as references applying to a new document.

When retrieving information it might be found advantageous to compare a query first with all the queries stored, in order to discover whether similar queries have been submitted in the past. If a list of the documents retrieved is available, the process of retrieval may be greatly simplified. This method may also be used to bring together the respective inquirers to furnish an opportunity to discuss the problems which apparently brought about similar inquiries. Periodic analysis of the profiles may also furnish valuable information on trends and possible overlapping of activities or interests.

Since a history of the usage of the system is stored in the monitor, an analysis of its records will disclose the efficiency of system operation. The findings may serve to adjust the system for optimum efficiency.

There are many details which might have to be provided to adjust the general form of the system to specific applications. One such requirement might be classification, by an editor, of documents with regard to security, proprietary interests and proper utilization of information.

A plurality of systems may be organized in hierarchical fashion, in which a first system would serve a number of more specialized systems. In this case the specialized system would each assume the role of an action point in the mother system. It also appears quite feasible to share the system equipment among a number of organizations.

Prospects for establishing a Business Intelligence System

The system described here employs rather advanced design techniques and the question arises as to how far away such systems may be from realization. It may therefore be of interest to review the state of system and machine development.

The availability of documents in machine-readable form is a basic requirement of the system. Typewriters with paper-tape punching attachments are already used extensively in information processing and communication operations. Their use as standard equipment in the future would provide machine-readable records of new information. The transcription of old records would pose a problem, since in most cases it would be uneconomical to perform this job by hand. The mechanization of this operation will therefore have to wait until print-reading devices have been perfected.

The type of equipment required for processing information in accordance with the system is presently available as far as the functions are concerned. It is safe to assume that special equipment will eventually be required to optimize the operation.

The auto-abstracting and auto-encoding systems are in their early stage of development and a great deal of research has yet to be done to perfect them. Perhaps the techniques which ultimately find greatest use will bear little resemblance to those now visualized, but some form of automation will ultimately provide an effective answer to business intelligence problems.

References

(1) Webster's New Collegiate Dictionary, G. & C. Merriarn Co., Springfield, Mass.

(2) H. P. Luhn. "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, 2, No. 2, 159 (April 1958).

(3) H. P. Luhn, '*A Statistical Approach to Mechanized Encoding and Searching of Literary Information," IBM Journal of Research und Development, 1, No. 4, 309 (October 1957).

Received July I