MACHINE TECHNIQUES FOR INDEX SEARCHING AND FOR MACHINE TRANSLATION

James W. Perry
Research Associate, Massachusetts Institute of Technology, Cambridge, Mass.
Director, Scientific Literature Dept., Bjorksten Research Laboratories, Madison, Wis.

At first glance, it may seem strange to direct consideration in a single paper both to problems of indexing and to the possibilities of machine translation. Perhaps the fundamental reason for joint consideration of these two subjects is the fact that they both involve the manipulation of semantically significant symbols by machines. Furthermore, as we shall see, these manipulations have more in common than might be apparent at first sight. Another practical reason is the fact that the use of machines for scanning indexes has already been experimentally tested and is now being developed at a rapid rate for practical use on a large scale. This simple statement of fact should not be interpreted as an aspersion cast on those who are developing the strategy and tactics of machine translation. Perhaps solving their difficult problems may be facilitated by considering what has been accomplished during recent months and years in developing machines to scan indexes in order to select and correlate scientific and technical information. It also seems reasonable to hope that progress in machine translation may contribute to further development of machine techniques for index searching.

Previous speakers have made evident the problems and possibilities of machine translation. In order to provide a basis for discerning the relationship between machine scanning and indexing and machine translation, we shall now devote attention to various aspects of using machines for scanning indexes. First of all, we must consider the purpose of using machines to scan indexes. As is well known, the great expansion of research during recent decades has generated a large volume of research publication. In the field of chemistry, for example the annual indexes of Chemical Abstracts have now become as large as the first decennial index which embraced the years 1907-1916. This roughly ten-fold increase in the rate of generation of chemical literature during a single lifetime has resulted in the problem of using the chemical literature being made more difficult by at least an order of magnitude. A similar situation prevails in other branches of science and technology and, indeed, in many other realms of learning - such as law and medicine.

Making use of this new knowledge is no simple problem, and it is not intended to suggest that the use of machines for searching indexes constitutes a complete solution. In fact, it would be possible to imagine a situation in which a highly efficient machine installation might not be used to the full of its capabilities simply because there were no persons having sufficient breadth of knowledge to ask the machines those questions whose answers would contribute most effectively to the further advance of science and technology. It is axiomatic that machine searching of indexes is valuable and effective only to the extent that well-informed imaginative minds avail themselves of the possibilities offered by the machines. Without creative thinking, the ability to search recorded knowledge by machine cannot hope to provide an escape from the limitations of mediocrity.

One might well ask the question what role machine indexing can be expected to play in the realm of learning. In order to answer this question, it is necessary to recall to mind that advances in science and technology are achieved by building on past accomplishments. The starting point for any well-managed research project is the knowledge of what has been accomplished previously. Thus memory - either that of the researcher

himself or that form of memory provided by books, publications and other records - plays an important role in scientific and technical advance. The limitations of human memory being what they are, and the necessity of knowing of other persons' accomplishments being what it is, the rapid development of science and technology during the past 200 years has been accompanied by a parallel expansion in scientific and technical publication. It is true that such publication serves the purpose of providing current awareness of recent advances and, in this sense, scientific publications serve a newspaper function. Equally important, however, is the function of publications to record past advances and this memory function is, of course, the reason why scientific publications are collected and stored in libraries. The memory function, however, does not consist solely of recording. An equally important function is the ability to recall a needed record. As long as the volume of recorded material was relatively small, relatively simple methods sufficed to recall information when needed. Since the rate of publication of scientific literature has increased by an order of magnitude during the past few decades, it is to be expected that we need to improve our searching methods by a similar degree.

Machines may be used, of course, in many different ways to facilitate library operations. Automatic conveyors are used, for example, to move books from one place to another and - more recently - it has been proposed and demonstrated that television makes it possible to transfer the image of a printed page over considerable distances - thus speeding up the availability of books located some distance from the reader. Useful as such techniques may be, in providing material once it has been identified, the problem of identifying needed material is not served by such methods. The book conveyor and the television transmitter transfer physical objects or images of printed pages, but neither of these devices inform us as to which book and which page should be consulted to find what we need to know.

In order to discern how machines offer promise of being useful for locating information, we must observe first of all that our understanding of experience of any sort involves an operation of analysis. In reporting the synthetic preparation of a new compound, for example, the chemist will state what substances he caused to react, what apparatus was used; he will indicate such circumstances as temperature, time of reaction, solvents and the like, and will state what substances were produced and what their properties were. Similarly, a report of a surgical operation will be concerned with such factors as the patient's symptoms, the pathological conditions involved, the treatment and its results. In general it may be said that any record involving an experiment or observation in science or technology involves some type of interaction, the participants therein, the surrounding conditions, and the results obtained and observed. Thus a report of any experiment or observation will be a statement of some sort involving a multiplicity of substances, concepts and also – equally important - the relationships between them.

The questions posed by a research worker when developing a background of information for a research program also involve - in the most frequent general case, at least - a multiplicity of factors. Chemists will wish to be informed concerning compounds having a certain set of properties, or they may wish to know how certain compounds respond to a certain type of reaction - for example, catalytic hydrogenation at high pressures. A physician confronted by a patient having a given set of symptoms may wish to know what results a certain treatment has produced in the past. Both the information recorded by the experimentalist and the question posed by the seeker after the information can be analyzed into elements, some of which designate material objects, processes, attributes and conditions, while the relationships in which these stand to each other are indicated by still other elements. The role of the machine in mechanized index searching is to provide a way for linking the elements in the material being searched - that is, in the recorded information - to the question elements.

It would be, of course, possible to accomplish a great deal by

merely recording one-after-another in machine searchable form the various entities, processes, conditions, etc., involved in a given interaction. Such recording of the elements, appropriate and important to a given interaction, could be accomplished with available machines in some form of pattern in terms of which the machine might conduct its searching operations in much the same way that a blind man reads Braille. Machines of this type have been constructed to operate with punched cards. In order that a meaningful sequence of symbols could be detected regardless of the successive columns of the card in which they are punched, IBM was recently persuaded to develop a new type of punching worked out in such a way that a predetermined combination of five holes in any column is used to indicate some symbol, e.g. a letter or numeral. Searching is accomplished by reading the entire card photoelectrically and detecting those patterns of holes which spell out the sequence of symbols being sought. A plurality of photocells permits searching to be directed to combinations of words. This new IBM searching technique - in addition to identifying wanted criteria - also permits their relationship to each other to be specified. Thus the machine may be directed to accept all cards punched for any one of several criteria; such a search involving six criteria might be symbolized as A+B+C+D+E+F. Alternately, it might be specified that only cards bearing all six criteria would be accepted, and such a search might be represented by A·B·C·D·E·F. Combinations may be specified, such as (A+B)·(C+D+E)·F. Another possibility is to require that sane criteria or one or more of a group of criteria shall be absent. Examples of such searching might be A·B+C·D-(E+F) and (A-B)·(C-D)+(E-F).

As already noted, the basic operation with such punched-card machines is pattern matching. The wanted information is identified by specifying one or more criteria which are spelled out in terms of a pattern whose equivalent the machine detects in cards from the record file. Essentially the same thing can be accomplished if the patterns are impressed on other media. Thus, instead of patterns of holes punched in cardboard, we may have patterns of magnetic spots impressed on tape, or patterns of transparent and opaque spots on photographic film. With these and similar media, matching circuits can be set up so that individual criteria can be detected and the detection of these criteria caused to interact in the manner exemplified above by A+B+C+D+E+F or A·B·C·D·E·F or A·B+C·D-(E+F), etc.

As already noted, it is possible to capture much of the information in a document by merely listing the important elements, that is to say, the entities, processes, conditions, etc., which it is evident would be of pertinent interest to a subsequent searching operation. Thus, if one merely enumerates the elements gasoline, fire, extinguishing, foam, tank ship, there is little possibility of ambiguity as the very nature of the physical entities and processes involved will immediately suggest that foam was used to extinguish a gasoline fire occurring on a tank ship. On the other hand, if one lists the elements export, import, iron ore, steel, Germany, Sweden, there is no way of knowing or deducing what was imported or exported from which country. Similarly, a measure of ambiguity at least, would be found if one were to merely list the reactants, the products of a reaction, and the reaction conditions in a certain chemical synthesis. Such ambiguity might, it is true, be resolved by a skilled chemist but this degree of insight and interpretation would scarcely be possible with available machines designed to effect searching operations.

It is, of course, possible to make arrangements so that the patterns will permit machine searching operations to make discriminations of a syntactical nature. Natural languages suggest two general types of devices for providing needed discrimination in situations involving exports and imports, reactants and reactant products, etc. One of these devices is ordering the array patterns corresponding to words in sentences. Thus, it might be specified, for example, that the exporting country will always be placed immediately before the importing country. Another device is to use symbols of the same general type as the inflectional endings or prepositions of natural languages. As is done in natural languages, a

combination of these two devices could also be used in setting up indexes to be searched by machines. It turns out, however, that design problems - at least for punched-card machines - are made simpler if symbolism having an equivalent semantic burden to inflectional endings or prepositions is used. At first glance, it might perhaps seem relatively simple to develop appropriate symbolism for use in connection with machine searching of indexing by taking some natural language as a pattern and following its usages with regard to prepositions and inflectional endings. Our studies and experiments in machine indexing to date indicate, however, that any natural language - be it English, German or Russian - provides no more than, at best, broad general hints as to how this might be accomplished, while at the same time comparison of the prepositional and ending mechanisms of different languages reveals very clearly and strikingly a high degree of arbitrariness in well-established idioms and, more generally, in standardized patterns of words which shift depending on the meaning to be conveyed. Thus, in order to provide optimum discrimination for machine searching of indexes, it is necessary to submit index expressions to a conversion process not only with regard to their physical representation but also with respect to the semantic content of the various meaningful symbols used. It is necessary to set up a machine language in which a specially designed grammar is provided for making machine operations meaningful and efficient. This grammar must be much more consistent and logical than is the case with the grammar of human languages. The ability of the human brain to interpret signals is not even remotely approached by any corresponding ability on the part of searching machines at present in existence or likely to be constructed in the near future. We are confronted, in other words, in developing machine indexing, with the problem of designing a language for use by the machine to conduct its searching and correlating operations and also by the need to use this language as a basis for expressing the content of publications. It is perhaps obvious that a translation step is inevitably involved in using such a language. At the present time, in developing our plans for using machines to scan indexes, we have limited our expectations as to immediate application of machines to the actual searching operations. We have not anticipated that machines will be available which will be capable of transforming index entries as set up by person's analyzing documents into the machine language. We have anticipated that it is quite likely that the encoding operation will not be accomplished by the same persons who do the indexing. Yet it is not beyond the realm of possibility that this operation could be accomplished by suitably designed equipment. If this were, in fact, achieved, the human indexer who analyzed the subject matter of documents would be in somewhat the same position as the pre-editor in a machine-translating organization. The indexer would set up a series of sentence-like expressions indicating the subject matter of the document in such a way as to permit the document to be found at an appropriate time when need for it arose. Encoding by machine would then achieve something very much like a translating function. In this way, the index entries would be converted into an appropriate form - both with regard to grammar and with regard to physical recording - so as to make machine searching possible.

Up to this point, various similarities between machine searching of indexes and machine translation have been placed in the foreground of discussion. In order to forestall misunderstanding, it is perhaps advisable to point out certain features that distinguish these two important applications of automatic devices. In discussing the searching of indexes by machines we pointed out the requirement that the machine match up question elements with similar elements in the index being searched. In practice, this requirement turns out to be less simple than it might appear at first glance. If the analysis of a document reveals that it deals with a dog and we assign an appropriate pattern of holes to spell out the word d-o-g, then all is well as long as we search for that pattern, but we run into difficulties if searching is to be directed to "animal", "mammal" or

some similar term. Discussion of the problems involved here would take us far afield from machine translation. Suffice it to say that machine searches of indexes requires for full effectiveness that the index elements shall function in such a way that related generic terms, e.g. "animal", shall be available for defining the scope of a search as well as such specific terms as "dog." Conceivably a highly complex machine might be designed so as to be able to recognize a pattern spelling out "d-o-g" when search requirements require all animals to be detected and related documents selected. An alternate course - and the one we regard as more practical for the immediate future - is to design our indexing and coding system in such a way that part of the symbolism used to represent "dog" also denotes "animal" and "mammal". Although what can be accomplished by this approach is subject to certain limitations, these have been found to allow ample room for establishing the searching and correlating of information on an entirely new and much more effective basis.

6-16-52