Erich Mater
Institut für Informationswissenschaft,
Erfindungswesen und Recht der TH Ilmenau, GDR

Human Intelligence as a Precondition for the Machine Processing of Knowledge*

Mater, E.: **Human intelligence as a precondition for the machine processing of knowledge.**
Int. Classif. 15 (1988) No. 3, p. 125–132, 24 refs.

As long as basic tasks of scientific information such as content analysis have not been theoretically analyzed by man, computers cannot be expected to supply really useable practical results. Inferential processes, too, presuppose that of the three basic categories, 'source, processing and target data', two data-types are known so that the third one may be arrived at by inference. This is demonstrated in the light of such fundamental questions as document- vs. problem-oriented content analysis, selection or abstraction, recall vs. precision rates, the relationships existing between main contents, essential contents, new elements and user needs, as well as those between document vs. fact retrieval systems. The tasks of the immediate future are measured against present-day computer capabilities. (Author)

1. Aim

Ever since the start of computer use in scientific information work in the 1950's the question has been asked and today again is being asked with respect tó microcomputers:

Just what can a computer do? Can it index, abstract or even translate?

The aim of this paper is to show that the above question is based on a logically wrong approach and therefore will not get us anywhere. Instead, we claim that automatic machines as conceived by J. von Neumann can perform all those tasks in the scientific information field for which a complete, inherently noncontradictory algorithm is available whose individual actions can be formulated as program instructions.

2. The text-indexer-user relation

In non-numerical computer technology – and not only there, by the way – we have three categories of data to deal with:

1. The input data, e.g. full texts, abstracts, descriptors, etc.
2. The processing data, i.e. computer programs for indexing, abstracting, retrieval and other purposes.
3. The target data, e.g. document references, relevant titles, abstracts and text excerpts.

The usual course followed leads from the input data via the processing data to the target data, with document references and user profiles representing the input data and a comparison unit serving as processing data and leading us to the output of the references found: the target data. This usual way, however, is not the only one possible. One might also start out from the target data to arrive, by inference, at the input data via the processing data. In the nonmilitary sector this method has been used for deciphering ancient inscriptions (Maya, Linear-B) and is also being employed for finding out missing data in fact storage systems.

Now for scientific information purposes a third variant might become of interest in the future, namely the determination, by inference, of the processing data from the input and the target data.

An interesting experiment of this nature was demonstrated some years ago already and reported on in our series of publications (1); however, this report seems to have escaped general notice. Presented in somewhat simplified but understandable terms, the principle concerned is that of learning algorithms. Input data are read in as titles, abstracts or full texts, and synsemantics thereupon eliminated by means of comparing lists. The words or syntagms considered highly relevant by the user are now entered in a relevance list. The computer thereupon performs the indexing work, starting out by applying a statistical procedure depending on the transverse sum of the weights per phrase or text. This – still strictly determined – process is corrected by man through his heuristic, hence non-determined judgment by which he intuitively decides about relevance degrees. During a learning and an instructive phase the computer from then on corrects its own indexing procedure, its program, until the operations performed by both man and computer yield – even in the case of completely new texts – identical results with regard to relevance. The computer takes its bearings from the cognitive decisions taken by man. The quality of the results, however – and this is the reason why we mention this example – is completely dependent on the learning phase, i.e. on the quality of the content analysis performed by man (2, 3).

Here, now, we are at the central point of the problem: before one can try to assign any task to the computer one has to know the correct results, e.g. the target data, for some test examples. In the case of arithmetical procedures this can be accomplished in relatively simple fashion. When, on the other hand, we are dealing with tasks in the field of scientific information, be it indexing, abstracting, translation or something else, "correct" can only mean that no better result, backed up by scientific arguments, exists.

Everyone knows from practical experience how difficult it is to obtain an at least satisfactory indexing result and how greatly this result is affected by numerous, hardly assessable factors. It should therefore be all the more important for us to at last give our attention to and to analyze the causes of existing shortcomings. As

* Paper given at 15th Colloquy on Information and Documentation, 2–6 Nov. 1987 on the topic: Wissensvermittlung – Informatik – Spitzenleistungen. It was published in: Dokumentation/Information. Schriftenreihe INER d. TH Ilmenau, DDR, Heft 74, (1988) p.118–142. We gratefully acknowledge the editor's permission to publish an English translation in this journal.

long as. we have not mastered this problem theoretically, we will not have a chance – apart from occasional experiments with stochastic simulation – to bring about an improvement, however gradual, of the present unsatisfactory state of affairs. This will be illustrated by a simple diagram – greatly simplified again so as to clearly bring out the decisive processes – for Document Retrieval Systems.

Along the vertical axis we use a subdivision into object, process and result, and along the horizontal axis one into content analysis, retrieval device and user's query as follows:

	Content Analysis	Retrieval Device	User's Query
Object	text	store	problem
Process	indexing	comparison	formulation of query
Results	notations/ descriptors/ key words	relevant references/ texts/data	notations descriptors/ key words

The object of content analysis is a text. The process to which this object is subjected is called indexing. As a result of this indexing, and varying with the system employed, notations/descriptors/key words are assigned. The results are put into storage.

The object of any given user's query is a problem for which the user needs either a solution (if he has none at all) or the most up-to-date solution (if unknown to him). The process this object is subjected to is the formulation of the problem in a controlled or free retrieval language. Again, as a result of such formulation, notations/descriptors/key words are assigned. They likewise must be put into storage.

The object of the retrieval device is hence the store. The process consists in a comparison, namely of each user's query with each document reference stored. If there is no matching, the next reference will be called; if there is matching, the results will be printed out or shown on the display, as bibliographic references in the case of two-stage systems and as texts or data of the source in the case of single-stage systems like Fact Retrieval Systems.

In each case, the results are offered to the user, i.e. to the enquirer. So far, everything may be considered well-known and undisputed. But in addition, this scheme is intended to furnish support for the following claims:

1. The target function of any information activity is exclusively the user; more precisely: the reply to his retrieval query. Only this function justifies the operation of information services.

2. The results of content analysis are search characteristics, not content descriptions.

Many indexing prescriptions still demand that descriptors should simultaneously be used for content description. Thorough reflection will show this to be impermissible.

Since in the case of identical problems content analysis and query formulation should produce identical results, both should also proceed according to the same principles. While on indexing questions there is an abundance of literature, query formulation, on the other hand, although constituting our primary target function, usually is left out of consideration. Here a great deal of reflection and of making-up for past omissions is still necessary.

3. Each retrieval operation consists in a comparison of document and user profiles, each represented by notations/descriptors/key words. In the traditional vertical card file the comparison of these two profiles admittedly was a slow procedure (because the cards were manually moved and visually read) but one that took place under constant intellectual supervision. The spelling of a term (e.g. Mikroprozessor/microprocessor) or its designation (e.g. computer/Digitalrechner/EDVA/Ziffernrechner) did not play any part in retrieval, except possibly in the case of a strict alphabetical order.

This changed basically when computation was introduced. The comparison unit of the computer operates on the basis of bit patterns, which means that any variation of the spelling of a search word, even if totally irrelevant to that word's meaning, leads to the result "not identical" and thus, in turn, to the conclusion, "not relevant". From this purely technical condition, the information/documentation world, when confronted with it more than two decades ago, drew exactly the wrong conclusion, namely that, as a concession to the computer, controlled languages should be subjected to strict prescriptions. However, standardized terms in the form of descriptors not only cause considerable extra work in indexing and query formulation, they simultaneously reduce the precision rate and increase the noise rate. Evidently it is only with the concepts of "Artificial Intelligence" that informatics will be able to find its way out of this blind alley. Technically, the looking-up of words in a thesaurus could have been automated already with first-generation computers. Today it is of course not only possible but even absolutely necessary to have the computer perform a conversion of the keywords or catchwords during or after indexing as well as during or after query formulation so that conformity of bit patterns may be achieved without detracting from the precision of indexing results or of query formulation.

4. We will mention here only one further claim that may be derived from the diagram. In the case of conventional content analysis the indexer does not know the possible queries of the users. Consequently he has to deduce several unknown quantities, namely possible queries by future users, from one known quantity, namely the text. The user, in turn, when formulating his query, cannot know what documents may possibly have contents relevant to his purposes. Consequently he has to deduce several unknown quantities, namely the texts, from one known quantity, namely his problem.

This is the main reason for the fact that so far there exists neither an exclusive, "correct" indexing result, nor an exclusive, "correct" query formulation: usable processing data (indexing programs) need to be developed for known input data (the texts) without it being possible to state, be it only in approximate terms, what the "correct" target data (indexing results) would have to look like. This, of course, will hardly work! First of all, clarity will have to exist as to the causes before one can start to think about improvements from which useful concepts on automatic content analysis might then be derived.

From what has been said so far the following conclusions may be drawn:

- Human intelligence is a prerequisite for any kind of knowledge processing rather than its aim. First of all one has to know what the results, related to a concrete example, would have to look like. Thereupon, problem analysis for generalization of the individual case can be started. From this problem analysis an algorithm may in many cases be obtained which in data processing needs to employ only formal elements as basis. The algorithm, finally, constitutes the basis for the programming work, with the latter practically requiring only knowledge of computation and not, or hardly, of the given field of application. The choice of the programming language is therefore only of practical/economic importance, as is the choice of the type of computer and of its operating system.

For the foreseeable future this sequence will remain imperatively prescribed. Attempting to put computer procedures ahead, in time, of human problem analysis means to misjudge the existing causal relationships. The question "What can the computer do?" is therefore wrongly put. Correctly put, it would rcad: "What processes of human intelligence can already be analyzed and subsequently formalized to the point where they can be completely reproduced, in non-contradictory fashion, by the computer?" After this, the computer program and its implementation present only quantitative questions, no longer questions of contents.

There will be no change in this situation before the 5th computer generation has emerged.

3. Aspects of content analysis

Viewed from these premises, content analysis and query formulation are by far the most important tasks in any information activity, with abstracting being regarded in this connection as a special form of content analysis. All other activities of information centers, hence including the use of computers, are subordinated to and in fact based on these primary problems and cannot, later on, make up for what was done wrong or omitted in the beginning. Nor can this situation be corrected in any way by subsequent methods of information generation as a form of Artifical Intelligence.

It is a well-known fact that different indexers working on the same text and applying the content analysis methods used so far attain a coincidence rate of less then 50% in the document references obtained by them (4). When large systems – and only these offer representative quantities of data – search for the causes of the losses and noise yielded by their retrieval efforts, they find these causes to be, on the average, rather evenly distributed, with indexing results and query formulation each accounting for approximately half of the total (5). Causes attributable to the use of computers, on the other hand, account for less than 1% of the failure causes – and yet most of our professional colleagues currently believe that salvation lies in a larger processing capacity (16-, 32-, 64-bit processors), in a larger external memory capacity (from 50 Mbytes onward), in the clock pulse, the operating system, the programming language, the memory organization, the accesss paths, the cross-linking of computers, and last not least in Artificial Intelligence methods. But obviously, as shown above, artificial intelligence must be preceded by human intelligence, and not the other way round.

Since the literature on content analysis offers only few leads on how to conduct a systematic problem analysis, I will attempt to describe, from the point of view of nonnumerical computation, what possibilities of automatic content analysis are discernable. It will be useful in this connection to separate Document Retrieval Systems from Fact Retrieval Systems so as to let the shortcomings that have existed so far become more clearly apparent.

First of all we distinguish between document-oriented and problem-oriented content analysis. The former wishes to find out the main contents of any given document, while the latter wishes to collect all essential statements on a given problem. "Main Contents" consequently pertains to a bibliographic unit in library-science terms, while "essential" pertains to a clearly defined problem in a special field in information-science terms. If these reference quantities arc mixed up, causing e.g. "essential" to be related to a bibliographic unit or corpus – which is not unusual –, then the original approach, clear though it was in itself, will be blurred, thus giving rise to additional inaccuracies which first need to be cleared up.

Relating as it docs to only one document in any given case, the "main contents" thus is not related, in indexing, to other documents (corpora). This holds true also in the case that for 2 or more documents the same "main contents" is indicated or that several "main contents" criteria, arranged according to their quantitative rank, are indicated for one document.

Matters are different with respect to the "essential". Referring as it docs to a clearly defined problem, it bears a relation to all documents dealing with the same problem. Statements on this problem are thus distributed over several documents and require, unless they are repetitive, a high recall rate in the first retrieval step. Here the analogy to Fact Retrieval Systems immediately becomes clear: For these systems it is of no importance from what document the data were obtained, whereas it is very important that the memory contain as large as possible a number of statements about the given problem. Both the indexing process and the indexing results should be oriented accordingly.

Indexing according to the newness of information is a further aspect of indexing, not to be confused with the actuality rate of knowledge, which is measured by the length of time elapsing from the writing of a text to its becoming available in the memory. In indexing according to newness, too, clarity should exist as to the reference quantity: new to the indexer, new to the enquirer or new to the memory?

In traditional content analysis the indexer can only decide what is new to himself in which connection we leave the unreliability of human memory out of consideration. What is new to the potential user is something the indexer cannot decide, since – except in very small systems – he does not know him. In selective dissemination of information (SDI), however, an indirect checkup by computer would be possible. The only really reliable quantity that can be checked up on at the time of content analysis is the memory: if knowledge is already stored in the retrieval system, further references can only confirm, but not renew it. Here, too, a checkup would be easy in the case of Fact Retrieval Systems, but difficult in the case of Document Retrieval Systems because of their arbitrary indexing procedures. It moreover would require content analysis by dialogue, i.e. in constant communication with the memory – which, by the way, undoubtedly would constitute a major step forward towards scientifically exact indexing.

Thus, while the "newness" concept refers to individual subjects – enquirer or indexer – and may also relate to an object – the memory – we relate the "actuality" concept to the process of making knowledge available. With knowledge innovation cycles becoming shorter and shorter, the actuality rate of information, too long neglected, is now acquiring the same importance for the user as recall or precision rates: information of high actuality, even if incomplete, ranks higher than old, though highly complete, information. The interdependence existing between, on the one hand, the expenditure of time necessary for thorough content analysis and, on the other hand, the benefits obtained from a high actuality rate, will compel us to give thought to new methods of content analysis. As long as it takes an average of nine months for a document to be processed from its arrival at the information center to the point in time where it becomes available in the memory, this in the light of the less than 2 years it takes for the totality of man's knowledge to cyclically double in size – as long as this situation exists, as I keep repeating (6, 7), our professional field is not living up to its task.

Apart from indexing according to the main contents, the essential contents and the new elements, one may also index according to user's needs. Now in normal information systems the potential enquirer is not known to the indexer. But since these anonymous users form the actual target group of the entire system it seems appropriate to use at least an auxiliary construction: Round up those possible queries which reflect all main lines of research known and index at least according to them. It was exactly by such procedures that the concept termed by us the "essential" in the aforegoing was approached. However, perfection of this approach was possible only in the measure that all research subjects are known – and known well in advance at that.

This realization confronts us with the question just when indexing should preferably be performed: before or after a query has become known? We will return to this problem later on.

Now according to what points of view is content analysis being carried out in practice – the main contents, the essential contents, the new elements or user's needs? The answer to this important question is well known to everyone and is also evident from all existing indexing rules (except, of course, for Fact Retrieval Systems): what is gathered is a mixed bag of all 4 criteria.

We say this without any mocking undertone, for in intellectual retrieval this approach was entirely justified by the fact that the expert's cognitive and associative abilities enabled him to roughly reconstruct the approximate contents of the text from indexing results of the above nature. With the computer, matters are entirely different. A computer is a structure-processing automatic machine which compares bit patterns arranged in strings. A deviation of only 1 bit is already sufficient for an actually fully relevant document to be rejected. This was where the cause for the introduction of strictly controlled retrieval languages lay (even if it was a cause resulting from erroneous conclusions). It was thought necessary to adapt the contents to suit the form, rather than finding a form suitable to the contents. This error, like so many other ones, goes back all the way to the world of the library.

A computer, unless forced to do so, will not produce any syncretisms. It demands unambiguous programs whose procedures are marked by strictly one-to-one correlations. Such procedures presuppose an algorithm. Such an algorithm, in turn, is the result of a lucid problem analysis:

According to what criterion is indexing to be performed?
What does the best result thus obtainable look like?
What formal characteristics are needed for this purpose?
How does one find these characteristics in the text?
How can the logical steps necessary to this end be formulated?

These are the most important steps for arriving at the single meaningful approach in the case of a computer using procedural languages.

Under this aspect, let us consider the various indexing methods with a view to automatic content analysis.

The *main contents*, being a purely quantitative criterion, can be determined by means of statistical methods, with computer-produced results being at least equivalent here to those obtained by man. The pertinent basic idea can be traced back, like many other good ideas in this scientific field, to LUHN in the 1950's (8). Coming forth from the STEINBUCH school (9), the corresponding theoretical model was elaborated later on, with LAMPRECHT/LAMPRECHT(10) completing it by the addition of the semantic fields. Thus perfected, this procedure has been with us for some 15 years by now. It still permits of some variants, but hardly of any basic improvements.

The *essential* can be obtained by the computer through the segmentation of strings of characters or combinations thereof. The difficulty here is that results are to be obtained in the form of qualitative statements rather than of quantitative ones as in statistical procedures. More about this later.

The *new* can only be ascertained by the computer via comparison quantities stored in the reference system (machine dictionary). In the simplest case the actuality rate is chosen as reference quantity and compared with the year of publication of the document. Only the contents of the system memory or the user's present knowledge could possibly be taken into consideration as reference quantities in the proper sense, whereas the indexer's knowledge is definitely ruled out here.

User's needs present the computer basically with the same task as the essential, with the limitation, however, that "essential" is related here to only one query.

To be able to grasp this important problem we must again indulge in some theoretical reflection along the following lines: Our point of departure is a linguistically formulated text, as a rule a complete document. In content analysis, the task to be accomplished consists, quite generally speaking, in usefully transferring words (character strings) from their syntagmatic framework into a paradigmatic ordering system, hence from their linearly arranged sequence into a topographical pattern. Both the human brain and the computer will the better be able to do this the more finely the ordering pattern on the conceptual level is structured. It is only in computerized content analysis that the large measure becomes apparent in which indexing results depend on the quality of the paradigmatic order. Not only the depth, but also the precision of the analysis is almost completely determined by the quality of the conceptual classification system, since in computer procedures this system must replace man's intellectual performance.

The conventional manner of indexing, no matter whether document- or problem-oriented, permitted of only two possibilities of descriptor allocation: the assignment of descriptors obtained either by selection or by abstraction (7). We consider descriptors as having been obtained by *selection* when they appear as terms in the text. These terms may occur as lemmata, or may have been reformulated into a retrieval language, or even have been transferred onto a higher hierarchical level (hyperonym), but in any event they must be identifiable in the text as words (more accurately: as concepts).

We consider descriptors as having been obtained through *abstraction* if they have been derived or abstracted from a statement or sequence of statements. The processes by which this is done are mental, cognitive ones which the expert carries out on the basis of his wide special knowledge. In so doing he operates on the level of statements and knowledge rather than on that of individual text words.

The difference between both processes, hardly perceived in intellectual indexing, is a momentous one and brings its full weight to bear in machine procedures of every kind. While the selection process is something the computer can master, abstraction requires wholly new procedures in which the place of, say, a dictionary as reference system would have to be taken over by the total background knowledge on a limited special field. Just how ambitious such an automatic procedure would have to be cannot be examined here. Suffice it to point out that, in addition to other prerequisites, the entire range of linguistic analytic steps – graphematics, morphology, inflexion, word formation including composition and derivation, vocabulary down to the semanteme level as well as the entire field of syntax, semantics and sentence overlapping relations – would have to be run through, and only then would one have the necessary material basis for making statements on facts and processes (12). Since only the levels up to and including vocabulary have been linguistically analyzed to the point where they can be formulated as an exact system of rules, fully automatic content analysis can, for the time being, only be realized on the basis of character string selection.

These procedures may also include combinations of character strings, in linguistic terms called collocations, in terms of information science co-occurrence (13), as well as quasi-syntagms, i.e. word sequences beyond a linguistically conceived grammatical model (14, 15).

Here, attention needs to be drawn also to a momentous error which both Scientific Information and Automatic Language Processing frequently fall victim to. The error we mean is the idea that the word is the smallest linguistic unit of information. However, in actual fact it is the statement or proposition, i.e. the subject-predicate relation, which is the smallest linguistic piece of information. Therefore, descriptors can only indicate whether a certain subject is being dealt with, but not what statements are made about it.

An intermediate stage on the way toward the strict dichotomy of selection and abstraction in content analysis is formed, however, by the learning systems we already referred to in the aforegoing. Beside, below and above them there are further mathematical procedures, among which cluster analysis evidently plays a predominant part. These, too, operate on the selection level, and in their case, too, the quality of the results depends on the quality of a reference system. The best roundup of all research work under way at any given time is always to be found in INTERNATIONAL CLASSIFICATION (17), while we are indebted to PANYR (18) for systematizing this overview.

From Section 3, we can now draw the following conclusions:

- The task of Scientific Information activity consists in so organizing the processes of making information available and of processing it that, on the one hand, the recall, precision and actuality rates all reach optimal parameter values while, on the other hand, the enquirer is offered only so many data of references as he can really evaluate in the time available to him.

The solution of this problem, to the extent that electronic data processing can furnish it, requires a prior systematic inventorying and evaluation of

– the data to be processed,
– the results thereby to be obtained, and
– the pertinent methods at our disposal.

4. Possibilities of automatic content analysis

We will now consider these three points under the special aspect of computerized processing.

If information efforts are to convey the international state of the art, then data from both national and international data bases must be available. The given center's own data collection will by no means be sufficient. In the caše of Document Retrieval Systems the data available on any document consist of its title, of an abstract if possible, and in the ideal case of the full text. Processing results in a number of references, each showing, as search characteristics, the main contents, the essential contents or the new elements of the given document.

In the above we had subdivided the analysis methods into selection procedures, which segment individual character strings and on this basis obtain contents characteristics, and into abstraction procedures, which sum up statements or complexes of statements. Among the computerized selection methods the following ones have already been put to the test:

- Total, partial or floating comparison of search words (masks) with text words;
- Ejection of key words with context (also KWIC and KWOC);
- Various mathematical analysis methods;
- Use of artificial languages or of mathematical calculi.

Among the abstraction methods, experience has already been gained with the following ones:

- Evaluation of structural abstracts;
- Learning algorithms;
- Linguistic analysis and evaluation.

While the selection methods can be carried out fully automatically, the abstraction methods require co-operation by man, often on a very large scale, so that the computer furnishes primarily quantitative support.

If, now, the given objective and task are compared with the methodical apparatus that has been used so far, a discrepancy will become apparent which evidently cannot be solved by traditional methods. This is what the preceding detailed explanations concerning the relations existing between text indexing procedure, retrieval apparatus and query formulation were intended to show. Since, on the one hand, the quantity of obtainable references has increased by several orders of magnitude in recent years, partly by the copying of databases and partly by remote access to such bases, and since, on the other hand, man's receptivity remains constant, indexing should in fact be far more refined than it is now so that the number of relevant references turned up may be reduced to a measure commensurate with man's receptiveness. This, however, is not possible with traditional indexing methods, partly for economic reasons and partly because of the lacking methodical tools. It would also require that the enquirer's information needs can be formulated far more precisely. This is hardly possible, as may also be seen from the example of the two types of knowledge given byWEBER (19).

The attempt to get out of this quandary by indexing the documents only when the exact query formulations are known is not as erratic as might seem at first glance. The fact that it takes an average of nine months for a document newly arriving at an information center to become internationally retrievable from the store is not known to the user, who is happy to receive the retrieval results only a few days after his query. Should he receive them only after 2 to 3 months, due to the fact that indexing had only been performed on the basis of his query formulation, he would be highly discontented, although the actuality rate of the references provided would then be thrice as good. The real problem here lies with the mass data, however. Documents are predominantly analyzed centrally, namely by the operators of large databases. There, however, the future users will remain anonymous and their exact information needs an unknown quantity.

Under these conditions a *two-stage analysis procedure* – for which the technical prerequisites, furnished by both telecommunication and microcomputer technology, already exist – suggests itself, with the first stage concentrating on recall and actuality and the second one on precision. Accordingly, the first stage should be based on a conceptual reference system, hence a classification system with, if possible, an international range of validity, while the second stage should be reserved for detailed analysis, possibly operating also on the level of word and collocations.

In the case of such an approach, the criteria admissible in the first stage for rough analysis might even include the "main contents" criterion, except, of course, in the case of Fact Retrieval Systems. Second-stage content analysis would then have to concentrate on "essential contents", hence specific problem fields, or "new elements", both as related to special user's needs.

First-stage indexing would thus be document-oriented and second-stage indexing problem-oriented. Some computation methods are already available for both stages. For the first stage, document- oriented rough indexing, the procedures offering themselves are above all quantitative, i.e. mathematical-statistical ones. Their recall rate ranges on the average from quite satisfactory to good, with the high noise rate and low precision rate of course remaining, characteristic as they are of document-oriented indexing.

The second stage then serves exclusively for improving the precision rate on the basis of already known user profiles. Here, the dominating role should be played, on the one hand, by context procedures for obtaining statements, i.e. key words with context, and by procedures for retrieval from abstracts, while, on the other hand, recourse may be had to search control words (search masks) if a really efficient reference system (machine dictionary), capable of reproducing the vocabulary onto the paradigmatic plane, is available. Without such a well-conceived machine dictionary, however, hence when using only intuitively formulated search control words, the advantages of such floating comparison procedures will be changed into their very opposite.

In proof of this latter asertion we mention the fact that microcomputer technology presents a very real danger:- namely the temptation to draw, by analogy, erroneous conclusions. With the microcomputer enabling us, at it does, to rapidly and easily carry out experiments in small data funds, we often carelessly extrapolate the results thus obtained to assumed large data funds. Thus, using occasional search masks in miniature funds containing only a few hundred references may produce strikingly good results. Applied to real funds, however, the same method will lead to catastrophic results. We are indebted to D.C. BLAIR for having investigated this relationship on a real fund of 40.000 references, with the result that less than 20% of the relevant references were retrieved (20).

The size of the fund actually being searched through is likewise the decisive criterion for searches conducted in far removed databases. The result will be better the more intelligent use one is able to make, also in international databases, of the dialogue for the second stage, i.e. the detailed retrieval efforts. In general, however, it will be more advantageous, both from the point of view of the task at hand and for reasons of economy, to have large systems first carry out a preselection and thereupon, with the aid of these results to perform oneself the detailed retrieval operations in reduced funds.

In both cases something would be fundamentally new: the separation in space as well as in time of the user from the database would be abolished. The enquirer would, at last, sit again in front of the retrieval device, controlling the search process according to his individual needs and, in the case of remote access, also being aided by a professional searcher (21). At the microcomputer he would be guided by a menu technique to be developed, which would guide the user without requiring him to be familiar with the various command languages.

Thus, we regard the future tasks of information engineers as lying in the development of such query formulations and retrieval programs as will reliably lead the user at the screen to the retrieval results which are best for him. This includes both the syntagmatic axis in the text and the paradigmatic one in the reference system (machine dictionary), with the latter axis also comprising synonyms, hyponyms and hyperonyms – hence, in the aggregate, precisely those activities which the information engineer used to accomplish so far in oral consultations on a user's query formulation. The qualitiative change, however, consists in the fact that, in reply to his query, the user now immediately receives on his screen the number, type and contents of the results retrieved, whereupon he can then, supported by the menu, improve his query formulation. This is not a computation problem, for the programming of the process of guiding the user in a simple, if time-consuming, matter. It is a task having exlusively to do with contents, consisting as it does of mentally penetrating a technical field and analyzing the problem from the user's point of view. This is the kernel of all true information work and it is absolutely realizable by technical means.

The effort and expense required for such detailed searching depend primarily on the database, more precisely on its prior mental penetration. Searches conducted in full texts require of course the most effort and can only be performed on the selection level. Searches among document titles are the most efficient ones and the ones most readily performable by computation; since, however, they yield only some 60% of the results produced by full text retrieval, they are best suited for pre-selection (22). The most advantageous way would be, of course, to search among abstracts on the abstraction level. This, however, would presuppose structured abstracts, as the analysis of statements is something computers have not yet mastered.

For all three types mentioned here, at least solutions in principle are known, whose application to the various special fields would require relatively little effort. The biggest problem encountered here is evidently of an economic nature, since quite a few databases are accessible only for the purpose of supplying information via printer. But copying the tapes for the purpose of conducting searches in partial fields on minicomputers would cost a multiple of the basic price.

5. Differences between document and fact retrieval systems

As repeatedly pointed out in the above, most of the difficulties in content analysis, query formulation and retrieval arise only in Document Retrieval Systems, not in Fact Retrieval Systems. The advantages of Fact Retrieval Systems include the following:

- No doubt exists as to what data are to be extracted from the text, namely: the name of the object, the name of the characteristic, and the value of the characteristic. Different indexers working on the same text will arrive at the same results.
- Query formulation is just as unequivocal for the user as the extraction result is for the indexer.
- Indexing is exclusively problem-oriented; the source and its main contents are therefore wholly irrelevant.
- Content analysis takes place on the selection plane and can therefore be carried out by the computer on almost the same quality level as by man.
- Controlled languages are superfluous; at most, lists of abbreviations for names of characteristics such as physical units of measure are used.
- There is neither loss sustained nor noise produced in retrieval.
- Recall and precision rates do not behave complementarily towards each other; rather, their values are identical and close to the ideal value.
- With the retrieval results, the user simultaneously receives the total available information rather than, as in the case of Document Retrieval Systems, bibliographic data on literature he should procure and read because it contains some information relevant to his query.

The setbacks of Fact Retrieval Systems can be left out of consideration in this connection, since the only question of interest here is just where the causes of the advantages of these systems lie and just how these might be passed on to Document Retrieval Systems, too. Evidently, all the advantages of Fact Retrieval Systems can be traced back to pre-formulated fields of characteristics. Here we note a basic, though gradually differing, equality with structural abstracts or, in regard to full texts, with problem-oriented content analysis.

Before starting to index, the indexer is informed what properties and processes of an object or problem are to be deemed essential. These will be segmented and entered into fields of characteristics. While it is true that numerical 'elements are differently structured as compared to verbal ones, the statement property is common to both. The only requirement to be met in the case of either one is that the ranges of characteristics of interest should be formulated in advance and that one should keep one's mind open for newly appearing ones (23). The factual relationship to Objectified Indexing Procedures, which on their part can look back on a long history (24, 25), is obvious. In this connection the subject of facet classification, long ignored for no good reason, should also be given attention.

Thus, the detailed-indexing stage offers several possibilities of content analysis which presuppose knowledge in the field of information science and penetrate into fields as yet unexplored. This calls for theoretical and practical investigations alike which should not be postponed for too long a time.

6. Theory versus practice

We thus find ourselves confronted again with the question as to the relationship between theory and practice in information science, a question which has become unnecessarily burdened by prejudices such as embodied by the rule of thumb: practicians can do anything, but know nothing; theoreticians know everything, but cannot do a thing. Let me formulate it differently:

Theoretical reflections, no matter how valuable qualitatively, will as a rule only then be economically profitable if they have criteria of effectiveness as their object; practical efforts, no matter how productive quantitatively, will as a rule only then be economically profitable if sound reasons can be given why the method applied yields a maximum of effectiveness and why other methods would be less expedient. This being so, we should make the GDR Academy of Sciences' motto our own: *Theoria cum praxi.*

Acknowledgements:

I am indebted to the Director of our Institute (INER), Prof. Felix Weber for many invaluable suggestions and I would like to thank our colleagues for many hints in problem discussions, especially Dipl. Ing. Heidenreich, Dr.-Ing. Langbein, Dipl.-Ing. Markscheffel, Dr.-Ing. Scheller and Dr.-Ing. Stephan.

References

(1) Jarosch, H., Löw, W., Müller, H.-D., Reichert, B.: Anwendung einer Theorie des Begriffslernens bei der automatisierten Aufstellung von Nutzerprofilen. In: Dokumentation/Information. Schriftenreihe INER d. TH Ilmenau, Heft 59 (1983) p. 43–50.

(2) Klix, F.: Information und Verhalten. Berlin, DD: Dt. Verlag d. Wissenschaften 1971. 810 p.

(3) Unger, S., Wysotzki, F.: Lernfähige Klassifizierungssysteme. Berlin, DD: Akademie-Verlag 1981. 294 p.

(4) Naujocks, K.-D.: Untersuchungen zum Grad der Abweichung verschiedener Indexierungsergebnisse an gleichen Texten. Ilmenau, DD: TH Abschlußarbeit INER 1974.

(5) Lancaster, F. W.: Evaluating the performance of a large computerized information system. J. Amer. Med. Assoc. 207 (1969) 1, p. 114–120 (acc. to (1)).

(6) Mater, E.: Aufgeschobene Indexierung. ZfI-Mitteilungen, ed. by AdW d. DDR, ZfI Leipzig 64 (1983) p. 38–45.

(7) Lamprecht, H.: Eigenschaften und Kennziffern des Informationsfonds Biologie des VINITI und Schlußfolgerungen für die Recherche. Informatik 30 (1983) No. 1, p. 11–14.

(8) Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. & Developm. 2 (1958) p. 159–165.

(9) Wagner, S.W.: Automatische Stichwortanalyse nach dem Rangkriterienverfahren. Karlsruhe: TH Diss. 1966.

(10) Lamprecht, Helga, Lamprecht, Holger: Die Grundlage für ein statistisch-semantisches Verfahren zum automatischen Indexieren. Ilmenau, DD: TH Diss. 1975.

(11) Mater, E.: Zur Erhöhung der Recherchegenauigkeit. Int. Forum Inform. & Doc. 5 (1980) No. 4, p. 12–17 (engl. & russ.).

(12) Dietze, J., Mater, E., Meyer, G.F., Neubert, G., Witzmann, A.-K.: Linguistische Datenverarbeitung. Z. Phonetik, Sprachwiss. u. Kommunikat. forsch. 36 (1983) No. 6, p. 633–648.

(13) Dietze, J.: Informations-Linguistik. Leipzig: Verlag Enzyklopädie (to be published in 1988).

(14) Holzweißig, A.: Untersuchungen zum automatischen Indexieren englisch-sprachiger Erfindungsbeschreibungen für ein Patentvorrecherchesystem. Ilmenau, DD: TH Diss. 1985.

(15) Lustig, G. (Ed.): Automatische Indexierung zwischen Forschung und Anwendung. Hildesheim–Zürich–New York: G. Ohns 1986. 128 p. = Linguist. Datenverarb. Bd. 5.

(16) Mater, E., Stindlová, J. (Ed.): Les machines dans la linguistique. Prague: Academia 1968. 336 p.

(17) International Classification. Devoted to Concept Theory, Systematic Terminology and Organization of Knowledge. Ed. by I. Dahlberg, Frankfurt (Main): Indeks Verlag. 1974–; 3/ann.

(18) Panyr, J.: Automatische Klassifikation und Information Retrieval. Tübingen: Niemeyer 1986. = Sprache und Information, Bd. 12.

(19) Weber, F.: Zu Spitzenleistungen durch systematische Wissensverarbeitung und -generierung. In: Dokumentation/Information, Schriftenreihe INER d. TH Ilmenau, Heft 74 (1988) p. 34–63.

(20) Blair, D.C.: Full text retrieval: Evaluations and implications. Int. Classif. 13 (1986) No. 1, p. 18–23.

(21) Lamprecht, H.: Erfahrungen bei der Nutzung von Informationsbanken zur Informationsversorgung der Forschung an der AdW der DDR. Ber. z. Wiss. inform. d. AdW, WIZ 4 (1987) 145 p.

(22) Mater, E.: Systematisch geordnetes KWOC-Register als Expreß-Information. In: Dokumentation/Information. Schriftenreihe INER d. TH Ilmenau, Heft 66 (1985) p. 163–167.

(23) Scheller, B.: Faktographische Angaben zu elektronischen Erzeugnissen. Ilmenau: TH Diss. 1980.

(24) Manecke, H.-J.: Untersuchungen zur Fachbezogenheit des Indexierens auf der Grundlage der Untersuchung wissenschaftlich-technischer Informationsströme. In: Dokumentation/Information. Schriftenreihe INER d. TH Ilmenau, Heft 35 (1977) 112 p.

(25) Weber, F.; Mater, E.: Zur Strukturierung von Patent-Referaten in: ZfI-Mitt. (Leipzig) 129 (1987) p. 117–125.

Prof. Dr. E. Mater, Institut für Informationswissenschaft, Erfindungswesen und Recht, TH Ilmenau, DDR 6300 Ilmenau