Robert Fugmann, Jakob H. Winter
Hoechst AG, Frankfurt

Reserve Retrieval: Toward Analogy Inferences by Mechanised Classification

Treatise VII on Retrieval System Theory

Fugmann, R., Winter, J. H.: **Reverse retrieval: Toward analogy inferences by mechanized classification.** Treatise VII on Retrieval System Theory In: Intern. Classificat. 6(1979)No. 2, p. 85–91

Whether or not an object of interest is an element of a class of objects on which important statements have already been made in the literature, can be recognized mechanically by the technique of "reverse retrieval". Thus, by way of analogy inference, statements on the object of interest can be reliably suggested by this technique and submitted for validity scrutiny. Reverse retrieval is an area of information supply which exhibits, as it were, mirror symmetry to the conventional retrieval technique and thus appears as its logical complement. In particular, the former promises to open up the book stocks of libraries just as effectively as conventional retrieval does in the case of data banks. On closer inspection, creative reflection would often turn out to be a variant of analogy inference and could thus be effectively supported by the reverse retrieval technique. (Authors)

1. Introduction

It is part of the nature of the researcher that he is not satisfied merely to perceive and at most to describe disconnectedly the phenomena and objects that interest him. Rather, he seeks lawful correlations between them of the broadest possible generality, for laws make phenomena easier to understand and even predictable. If he achieves this objective, then he has not only satisfied his thirst for knowledge but also gained a great practical advantage: much time-consuming and expensive experimentation to obtain the facts which are needed for practical work can be eliminated. The greater the strength of the foundation of our theory describing lawful correlations between the phenomena and objects of our interest, the greater will these savings also be.

Cognition has frequently been achieved when our rich fund of experience is re-ordered in what is often an ingenious way. If objects or phenomena are grouped in a novel and particularly meaningful way which is based on their really essential characteristics, that flash of insight into lawful relationships will come much more readily than if we are confronted with a very large number of unordered and disconnected phenomena.

Thus, for example, Galileo in his studies of free fall considered precisely those observations to be fundamental which led to the conclusion that the velocity of a freely falling body is independent of its weight. He had the justified hope that the observed deviations from this rule could one day be attributed to inessential disturbances, disturbances that have nothing to do with the fundamental nature of free fall. He thus took a position which was diametrically opposed to that of his ancestors and contemporaries, who held just these differences in velocities of falling bodies of different weight to be fundamental. And for a long time after Galilei no one saw any analogy between the force which holds the planets in their orbits and the force which attracts the falling stone to the ground. Newton was the first one to place the two phenomena in one common, meaningful group, and he realized that the separation of terrestrial and cosmic force effects, imposed until then was inessential. From this insight easily ensued the laws of celestial mechanics, which Kepler had discovered long before using very laborious and purely empirical methods. The departures from Kepler's laws, which were already known at that time, were also better understood on the basis of Newtons theory.

If "analogy" is taken to mean conformity in the essentials, then, in other words, arriving at cognition depends very much upon discovering analogies among the objects of our interest. The analogy inference has already very frequently been used in science and is used in everyday life, though often only unconsciously. The analogy inference is one of our most powerful heuristic tools.

But a great strain may be placed on our powers of concentration and imagination and an extremely long time may be required when we seek to recognize among the great number of phenomena and objects around us precisely those which are related to one another in an essential way, and when in this process we must sort out the many other groups of phenomena the resemblance of which is only a shallow one and which only misleads us. Much knowledge which we have sought in vain may have eluded us merely because of the size of this obstacle. Progress in discovery could be greatly speeded up if we could have the genuine analogies suggested to us, analogies which we then would "only" have to test for their validity. If analogy inference receives support of this kind, it could easily happen that the achievement of many a piece of cognition is shifted into the realm of the humanly possible.

We now want to examine the essence of the analogy inference more closely and consider in detail a special variant of the ways of classifying phenomena and objects which can facilitate the analogy inference. The effectiveness of this approach is based on the fact that it opens up our store of experience and knowledge in such a way that we are no longer dependent to such an extent on our – necessarily limited – imagination, and do not require so much time and concentration and are no longer restricted to that little bit of knowledge that we carry in our memory or can compile quickly from the literature, usually in a very fragmentary fashion.

2. Analogy inference on the basis of classification

Let us consider the class I of objects of some kind, which may be defined by the generic concept A (cf. fig. 1). For example we could be dealing with a class of chemical compounds and the concept A means a particular arrangement of the atoms, an arrangement which all elements 1, 3, 5, 9; etc. of class I have in common. Many individual experiments may have already shown that all (or nearly all) substances of this class exhibit a particular property a, that is, they may be capable of lowering the blood pressure or they all (almost all) can be prepared in the same way b (e.g. by oxidation). In

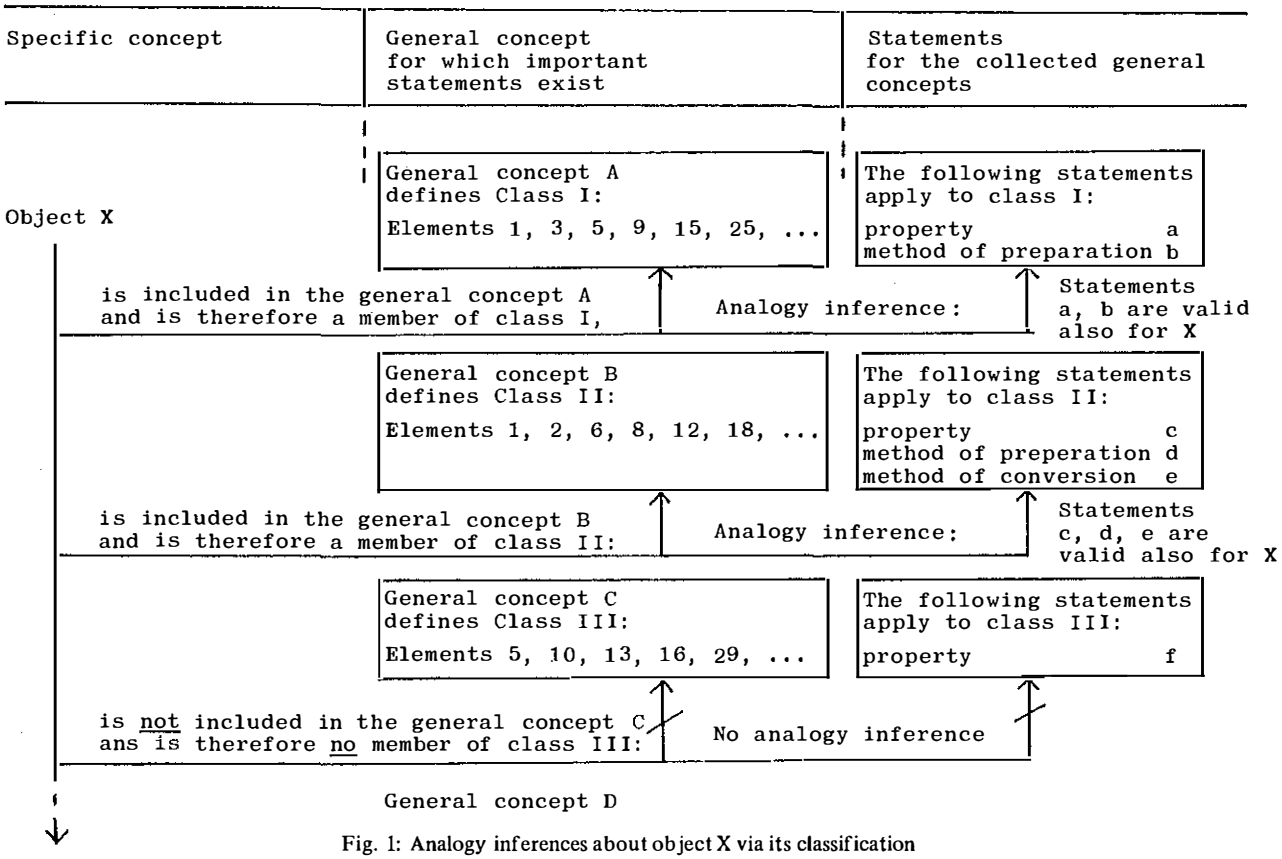

Fig. 1: Analogy inferences about object X via its classification

other words the statement is true for (nearly) all elements of class I which are known and whose properties have been appropriately tested, that they are effective in reducing blood pressure or can be obtained by oxidation. We will deal later with the question of in what way concept A and hence class I should be defined.

Let us forget for a moment these statements in the literature about class I and consider a certain object of momentary interest, say a specific chemical substance X. If this substance is available to us, then we may be interested in knowing what properties it might have, since they determine essentially to what practical uses it can be put. If we already know that substance X exhibits valuable properties, then we may be interested in finding improved manufacturing methods.

If a chemist examines compound X more closely, perhaps it will occur to him that it possesses the characteristics of the concept A previously described in the literature and is, accordingly, also an element of class I, whose members (nearly) all possess the hypotensive effect or can be manufactured by oxidation. If one has gained this insight, then it is a likely assumption that our compound X will also display this property or can be made in this way, since it has proved to be an analogue to compounds 1, 3, 5, etc. which are already known and have been appropriately tested. An analogy inference has thus been completed.

If object X is observed even more closely and still more expert knowledge is mobilized, it will perhaps be noticed that X also exhibits characteristics of the concept B. Concept B could be meaningful in that for all objects of class II, which is defined by concept B, it could be known that they possess the interesting properties c and d. Property c could be the fact that a certain harmful portion of the sunlight is blocked by substances with this property. Again it can be concluded more or less reliably that X also exhibits this property.

It is less likely, on the other hand, that the chemical substance X would also possess property f, if it does not prove to be a member of class III for which property f is characteristic.

All these conclusions can be drawn without object X ever having been manufactured, tested or even known, that is without much time and work being expended initially for preparing and testing the substance. In other words, we can quickly arrive at very many statements about object X of a more or less reliable nature, if we have previously compiled as many such statements as possible of the type a, b, c, d, etc. and assigned them to the corresponding generic concept A, B, C and their object classes I, II, III, etc.

The manufacturing methods or properties which have been "suggested" in this manner will be given preferential consideration and will be turned to good account in planning the manufacture of a chemical compound or testing its properties.

A serious obstacle is encountered in the practical application of this type of analogy inference: The human is unable to overview anything like completely all the classes of objects about which interesting statements have been made in the literature. Even those classes and statements which a person has encountered in his own work and in his literature studies can be utilized to only a very limited extent for analogy inferences. Within our biologically limited resources of time, concentration, patience and imagination we cannot possibly have all

concepts or object classes pass in review before our mind's eye in order to determine to which of these classes the object in question can be assigned and, accordingly, which statements could apply to X.

Here ist just the point at which the creative work of humans can be effectively backed up by mechanized techniques. The work of comparing the object X with very many generic concepts A, B, C, etc. (or classes I, II, III, etc.) can be left to the tireless and unerring computer, after the concepts A, B, C, etc. have been collected in a machine repertory. Once X has been identified as a member of one of these classes, then at the same time a (probably) correct statement about object X has also been found, namely the one which pertains to the members of this class.

3. Concept comparison by machine

If the comparison of two concepts (or the determination of class membership of object X)is to be accomplished by machine, then an important decision must be made and laid down in advance and in a manner that is valid for all concepts to be included in the comparison: It must be made obvious to the machine program which particular conceptual features (of the many usually present) should be shared by two concepts if they are to be considered as analogues. If, however, the human makes these comparisons himself, then he can proceed very flexibly and can allow himself to be guided by his intuition. He can also take into account the continual shifting of his interests caused by his continually encountering new knowledge and his learning new analogies. Under these circumstances he does not have a-priori to commit himself to any particular decision with respect to the set of features that should be shared by the concepts under comparison. At least this is not necessary to the same extent as when one delegates the comparison operation to a suitably programmed mechanism, for example.

4. Concept comparison in retrieval

An important variant of this machine comparison of concepts, which has long been practiced on a large scale in the information science is the retrieval, i.e. the mechanized literature search. Here the problem of the enquirer, such as the object X, is formulated as a search requirement. Object X is then compared with the numerous concepts which are contained in the documents of the store. If sufficient conformity is found between the concept X of the enquiry and a concept of a stored document, then this document will be taken as a relevant answer to the question.

Let us now consider the logic of this comparison somewhat more closely. X may be defined by the conceptual features $\alpha, \beta, \gamma, \delta, \epsilon$, and also by the absence of other conceptual features:[2]

Concept:	X	Y	Z
conceptual features of this concept:	$\alpha, \beta, \gamma, \delta, \epsilon$	$\beta, \gamma, \delta, \epsilon$	β, γ, δ

If X is a concept of an inquiry, then the comparison mechanism can obviously recognize only those concepts (and the corresponding documents) as relevant that fulfill the carefully considered and expressly formulated requirements for the conceptual features $\alpha, \beta, \gamma, \delta, \epsilon$. Any other behaviour of the comparison mechanism would be in contradiction to the explicitly formulated task. In particular, the general concept Y, and also the still more general concept Z cannot be accepted, since neither one satisfies all the requirements for the conceptual features of X.

A reverse comparison, however, in which literature on Y is sought and X is described in a publication, must have a positive outcome, for, in this case all conceptual characteristics required by Y are satisfied by X.

In an example, the concept to be retrieved may be "corrosion" and this concept may be defined by

α destruction of
β construction materials
γ by chemical processes
δ proceeding from the surface.

These conditions are not fulfilled by a stored concept that exhibits only the features

α destruction
γ by chemical processes.

This general concept would mean chemical attack on materials of every kind and include, for example, the removal of boiler scale, the fading of dyes, and paint removal with chemical paint removers.

For the same reason, an enquirer must not expect to retrieve information on the toxicity or explosivity of substance X, if he has expressly asked for the manufacturing of X, for the search parameter "manufacturing" will obviously not be satisfied by the concept of toxicity.

Or, in another example, if someone is searching for information on preparations to control ants in his garden, he must not expect any answers about the control of insects in general, of arthropods or even of animals of every kind. If he is interested in this more general type of information, he must indicate this to the comparison mechanism by generalizing the search request accordingly.

In other words, the general concept, of which it is typical that it has relatively few conceptual features, cannot be expected to be retrieved by a corresponding specific concept, for this has more conceptual features and in particular those which do not recur in the general concept.

But since, according to the above considerations the objective of our analogy inferences is always to find the more general concept in the literature, the retrieval variant of mechanized concept comparison is not suitable for our purposes.

Nor can the retrieval method be salvaged for the recognition of analogies through precautionarily generalizing the search request for object X. It is true that one would of course retrieve some general concepts pertaining to X, but there are very many different ways of generalizing the concept X, depending on which feature of the group $\alpha, \beta, \gamma, \ldots$ (or which combination of them) is omitted in the generalization. Hence it is normally impossible to take into account all conceivable generalizations in order reliably to retrieve every class to which object X belongs. Too many alternative searches would have to be formulated, each of them with a different kind of generalization, not to mention the fact that the formulation of so many generalized enquiries entails other disadvantages, which will not be discussed further here.

It is also impossible to restrict oneself to those possibly few classes and general concepts which have actually been described in the literature. This would require that we already possess complete knowledge of the literature, although it is just these concepts what is still being sought.

Nor is the situation altered if we completely eliminate all conjunctive search parameters and if we change them into alternative ones, possibly weighted in some way. In the last resort a certain minimum specificity of the enquiry will also here have to be established, though in a more flexible way; a stored document must at least match this degree of specificity in order to be recognized as relevant.

In our consideration of this question we have not restricted ourselves to the comparison of *concepts* but have always considered simultaneously also the *class* membership of object X. This may be superfluous in many fields, but occasionally we must expect that such a class is so heterogeneous that it cannot be described by a general concept but only by enumerating all the individual elements of the class. This occurs in the patent literature on chemistry by means of the so-called "Markush formula". Also it is sometimes felt that classes are easier to visualize than concepts.

5. Concept comparison in reverse retrieval

Let us take as our point of departure the assertion that complications can arise in a machine comparison between the enquirer's problem X on the one hand and a topic Y described in the literature on the other hand. In particular we have seen that a hierarchical correspondence between the two concepts X and Y *cannot be recognized* by the machine if

the *enquirer's problem* X is represented in a (specific) *query*

and

the *literature topic* Y *in the store* is general

But we have already seen that the machine comparison would reveal a hierarchical relationship between both concepts, if

the *enquirer's* (specific) *problem* X is *stored* in the machine

and

the (general) *literature topic* Y *is* formulated as a *query*.

Therefore, if in the mechanical concept comparison we exchange the functions of the enquirer's problem X on the one hand of the literature topic Y on the other, the hierarchical relationship between both will be recognized. In other words, we must employ so to speak a "reverse" retrieval (1) in which the allotment of the search problem and the recorded experience to query and store is reverse to that in conventional retrieval. Thus, through the reverse retrieval, literature topics can be found that are more general thanthe problem of the enquirer (cf. Fig. 2).

In research, we are interested in finding for a specific problem X the greatest possible number of hierarchically related general topics for the reasons explained in the foregoing. The way to achieving this is obvious: We will have to collect as many general topics of the kind suitable for analogy inferences as possible. Each of them will be phrased as query and entered into a correspondingly large repertory. The specific problem of the enquirer is, on the other hand, entered into a temporary store. The reportory of queries is then mechanically compared with the enquirer's problem (or with several of these at the same time), whereby a retrieval program of the conventional kind is used. Whenever a literature topic Y "retrieves" the given problem X, then this is equivalent

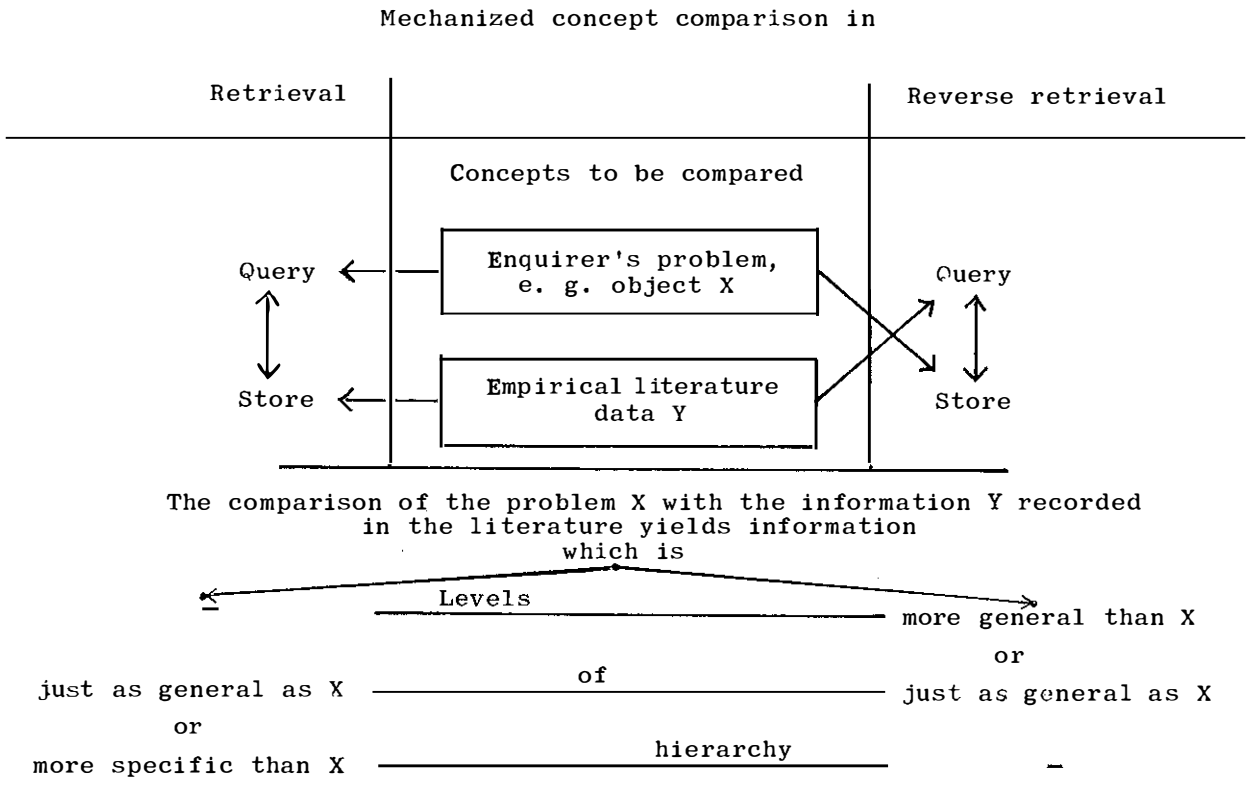

Fig. 2: Specificity of supplied information in retrieval and reverse retrieval

to the assertion that there is a hierarchical relationship between the general topic Y on the one hand and the specific problem of the enquirer on the other. In other words, X is recognized to be a member of the class defined by Y, and the statements recorded for the members of Y presumably also hold for X.

However, in reverse retrieval, particularly stringent requirements are placed on the accuracy and economy of mechanized concept comparison. For, it is to be expected that in the course of time thousands or tens of thousands of general concepts and topics accumulate as queries in the repertory. All these queries must be compared with object X.

These concept comparisons must be made with great precision. Otherwise, many general topics in the repertory would incorrectly respond to the problem X. This is equivalent to the suggestion of non-existing analogies. The greater the repertory, the greater will also be the demands made on the precision of the retrieval program to be used for the concept comparison. According to our experience, the systems best suitable for this purpose are those which are structured in accordance with the principles of the analytico-synthetic classification of Ranganathans Indian school and which are, furthermore, equipped with an efficient syntax.

Our own experience relates to a repertory of about 4,000 queries concerning classes of chemical compounds, which are often defined in several hierarchical levels of specificity. As far as the mechanical concept comparison is concerned, the retrieval system of the IDC is employed. The testing of a chemical compound X for membership in the 4,000 compound classes requires approximately 20 seconds of machine time on a Siemens 4004/151 computer.

6. Further possible applications of the reverse retrieval

It is inherent in the reverse retrieval that concepts which are more general than the subject of the enquiry can be reliably located in the literature[3].

This principle can be used for reliably finding a kind of generalized information in the literature which is particularly important to the inquirer.

The type of information that may be supplied in this way is illustrated by the following:

Object X belongs to a class of objects,

a) whose handling entails special hazards (e.g. toxicity, explosibility, radiation, etc.);
b) whose handling (e.g. storage, shipment, export) is governed by special statutory regulations;
c) for which patent protection has been granted or is pending;
d) which are starting materials for valuable products and thus are particularly important from an economic viewpoint;
e) which can typically be synthesized via a proven synthetic route (Synthesis design assistance in chemistry, cf. (1));
f) which can be purchased from certain manufacturers;
g) about whose properties an especially large amount of literature has been compiled in review articles or books, which makes these sources a mine of information for the enquirer and can save him much of his own literature search and much experimental work.

Frequently these statements are so unexpected and surprising for the enquirer that he could scarcely have made a specific search for them, or in other words, he would not have formulated them as search criteria in conventional retrieval. It is even much less likely to occur to him how the classes about which these statements are made are defined in the literature and of which of these classes his object X might be an element.

Of course such a class of objects can also be so narrowly defined and specific that in the extreme case it includes only one single object, namely precisely the one about which the statements a, b, c, etc. were made in the literature. In this case no new knowledge about object X is gained by analogy inference for the statements a, b, c, etc. are previously published facts. They could also be located through a conventional retrieval process starting with object X.

Finally, we want to take a closer look at the statements referred to under f), because of their great importance to information science. If reverse retrieval is used to obtain this type of statement, then a link is created between mechanized data banks, which are accessible to retrieval on the one hand, and libraries, on the other hand, which have hitherto proved relatively refractory to the retrieval approach. As the following discussion will show, the mechanized supply of information from the book stock of a library can function reliably only if the approach employed is that of reverse retrieval.

7. Reverse retrieval in library science

In order to demonstrate the relevance of reverse retrieval to library science let us take as an example the case that an expert is working in the field of corrosion and corrosion protection, and his special problem is the corrosion of water pipes made of copper.

The specialist will first employ the retrieval approach and formulate his – relatively specific – area of interest as search request to a suitable data bank. In this way he learns of publications in which his subject is discussed in at least the same degree of specificity in which it is phrased. They could be documents about the following subjects (cf. figure 3, left hand column):

Pitting observed at copper water pipes,
Pitting of copper water pipes caused by particulate impurities.

As the above discussion has shown, it cannot be expected that the retrieval method will locate references to documents on the subjects (cf. figure 3, right hand column):

Corrosion of copper,
Corrosion of non-ferrous metals,
Wear and tear of metal materials,

since these formulations do not satisfy all the conditions of the search request. In comparison with the concepts of the enquiry they possess fewer characteritics, i.e. they are more general.

Now, however, the general titles of these documents suggest that these documents are books or at least summarizing reviews and that our specialist could find much interesting material in them; perhaps even precisely the subject being sought will be found in a special chapter. This could be located by retrieval only if it is indexed in the machine store in almost exactly the same degree of specificity in which it is discussed in one chapter of the

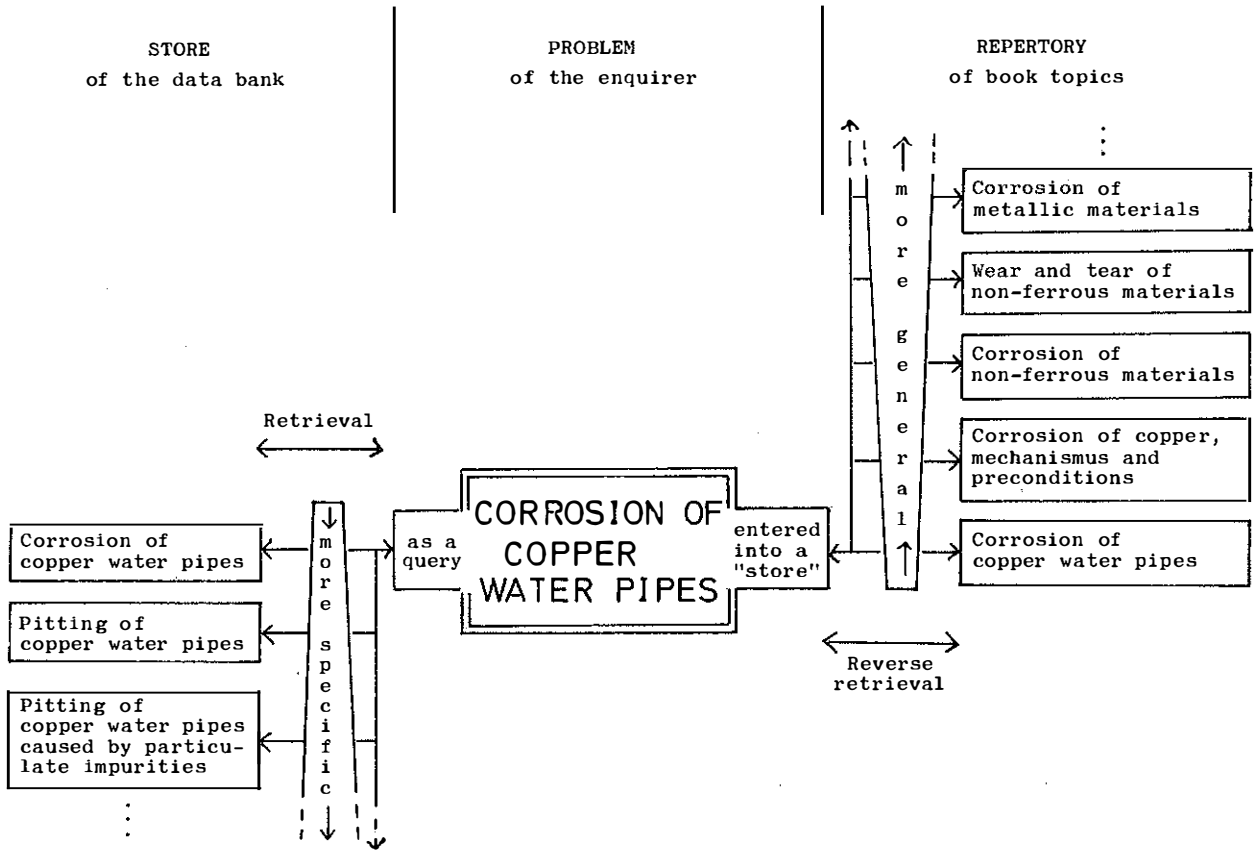

Fig. 3: The opening-up of the data bank and of the library through retrieval and reverse retrieval

book. But in view of the comprehensiveness of most of these monographs it would involve a prohibitively large amount of work to index or catalogue every detail with the same degree of specificity with which it is presented[4]. In order for the enquirer to obtain such information he would, in addition to making his retrieval request, also have to peruse the library catalogues himself or go himself to the book shelves. For again he cannot know beforehand how and to what extent his subject could be generalized in a book title on which the catalogue listing is based.

Accordingly, he will scarcely be in a position to formulate an adequately defined search request for those generalizations which lie just within his area of interest and are not all too far from satisfying his requirements. Mostly he can make his decision on the relevance of a book title, so to speak, only *a posteriori,* i.e. *after* he has seen the titles which are couched in general terms and might be of interest. In practice, however, it is observed that the enquirer only relatively seldom makes this additional effort, especially as he would have to refamiliarize himself each time with the pecularities of the catalogue and cataloguing conventions or with the special system of arranging books on the shelves.

If, however, as demonstrated in Fig. 3 the books in a library were made accessible through reverse retrieval, the very same (specific!) definition of the problem could be used for the search in the library as that already devised by the enquirer for information retrieval from a data bank. Again, his problem would be entered in a temporary "store", and all book titles in this library would be placed as "queries" in a correspondingly large repertory. The repertory of book titles would then be compared by machine with the temporarily stored problem. In this process each book title "finds" the specific subject of the enquirer only if the book title constitutes a true general heading for the enquirer's problem. Thus no additional effort would be required on the part of the enquirer to find books relevant to his topic in a library. Furthermore, this method of search, because it is mechanized and multidimensional, would be more productive and more precise than a visual search through a printed and necessarily linear catalogue upon which one must normally depend. *Just as the retrieval provides access to data banks, the reverse retrieval can do the same for libraries.*

We have already gained experience with reverse retrieval through review articles in the periodicals literature and there is no reason that this experience cannot be applied to the stock of books in a library.

In this way we can employ a single index language or classification, one and the same machine programme and a single definition of the enquiry problem to gain access to both a specialized data bank and to a library. In mechanized concept comparison we do this simply by reversing the functions of the empirical data in the literature on the one hand and the enquirer's problem on the other. At the same time the enquirer is relieved of a very time-consuming and mentally demanding task, namely that of thoroughly searching the library catalogues and mastering the corresponding cataloguing and indexing systems.

8. Concluding remarks

The only field in which the reverse retrieval has hitherto been practiced, even if in a fairly hidden manner, seems to be the field of chemical synthesis design assistance (2)–(4). A special variant of reverse retrieval could also be detected in some forms of computeraided medical diagnosis. However, the fundamental importance of this method of supplying information does not appear to have been fully recognized.

Recognition of the natural capabilities and limitations of retrieval and reverse retrieval is of great value. For, frequently in the past, demands have been made on the retrieval method which, in the light of the preceding considerations, it cannot possibly fulfill. Many types of criticism of retrieval systems evoked by these failures must therefore be reconsidered. In particular it is apparent that the capabilities of any retrieval system are overtaxed if we require it to supply *every type* of pertinent (that is to say, interesting) information. Consequently, repeated attempts to use a retrieval system for tasks exceeding the natural limitations of its capability and falling within the functional scope of *reverse* retrieval have led onto uncertain ground.

Reverse retrieval provides access, via a systematic approach, to the generalizations described in the literature. In this way it opens up a new area of information supply which forms, as it were, a mirror complement to the capabilities of the conventional retrieval method. These generalizations, once our attention has been called to them, can trigger analogy inferences and in this way exhibit heuristic value. But these generalizations may also be abstracts or summaries of writings which are in fact highly detailed. This is usually true of scientific book publications and review articles. In future a judicious combination of retrieval and reverse retrieval will be able to meet the information requirements of the scientific and technical world in a way that is substantially more complete than that possible with one of these two search strategies alone.

However, even an integrated search strategy of this type cannot always satisfy all the information requirements of scientists and engineers. An important precondition for retrieval and reverse retrieval is that the enquirer be able to *define* his problem with sufficient precision. But for good reasons this is often not possible. Then the search for information can obviously not satisfactorily be delegated, and in contrast with a "directed" information supply based on a definition of the problem, an "undirected" approach to information supply is necessary (cf. (5)). This means that the enquirer himself goes in search of technical literature in the field himself. In the present era of data banks and on-line systems it appears to us that there is a need to recall to mind the usefulness and necessity of one's own non-delegated personal literature studies and browsing and to call attention to the undiminished importance of the printed technical literature.

It is important to keep in mind the limits, of whatever type, on the delegated search for information. Otherwise one runs the risk of placing impossible demands on retrieval and also on reverse retrieval, demands which would cast an undeservedly bad light on the capabilities of every kind of mechanized, delegated search.

Notes:

1 Paper read in German at the Third Annual Meeting of the German Classification Society in Königstein/Taunus, April 6, 1979.

2 The definition of concepts through the use of, among others, negative characteristics is especially common in the definitions of chemical substances.

3 Stated more accurately, concepts can be reliably located which do not satisfy all conceptual characteristics of the enquiry.

4 This situation would not be altered by the development of advanced *automatic* indexing techniques, since the effectiveness of these techniques is very limited, at least in the field of the natural sciences.

References

(1) Fugmann, R., Kusemann, G., Winter, J. H.: The supply of information on chemical reactions in the IDC system. In: Inform. & Management 1979 (in the press)

(2) Pensak, D. A., Corey, E. J.: LHASA – Logic and Heuristics Applied to Synthetic Analysis. In: Wipke, W. T.; Howe, W. J. (Eds.): Computer assisted organic synthesis. Washington 1977, p. 1–32 (ACS Symposium Series 61)

(3) Wipke, W. T., Braun, H., Smith, G., Choplin, F., Sieber, W.: SECS – Simulation and Evaluation of Chemical Synthesis. In: Wiple, W. T., Howe, W. J. (Edss.): Computer assisted organic synthesis. Washington 1977, p. 97–127

(4) Bersohn, M.: Rapid generation of reactants in organic syntheses programs. In: Wipke, W. T., Howe, W. J. (Eds.): Computer assisted organic synthesis. Washington 1977, p. 142

(5) Fugmann, R.: Toward a theory of information supply and indexing. In: Intern. Classificat. 6 (1979) No. 1, p. 3–15