Robert Fugmann, Walter Denk, Ingeborg Nickelsen

Hoechst AG, Frankfurt**

Variations in the Order-Creating Power of Interactive Retrieval Systems

Treatise VIII on Retrieval System Theory*

Fugmann, R., Denk, W., Nickelsen, I.: **Variations in the order-creating power of interactive retrieval systems.** Treatise VIII on Retrieval System Theory. In: Intern. Classificat. 7 (1970) No.2, p. 73–78

In the interactive, mechanized literature search the phrasing of a query can be changed with great flexibility, depending on the kind of response from the search file. Thus a good deal of the flexibility of the manual literature search is preserved. The degree to which the mechanized search is superior through its high accuracy depends crucially on (1) the expressiveness of the indexing language in operation and (2) how reliably this language has been used in the indexing procedure. The syntax of an indexing language can contribute much to its expressiveness and, hence, to the accuracy of the search. A sequential file, indexed in a highly syntactical language, can be searched with a particularly high accuracy. This is in part due to the inherently high reliability of indexing in such a language. The interactive, dialogue search, however, is almost exclusively based on the vocabulary. This is due to the inverted file organization of the interactive systems, from which syntactical relations are largely omitted. Thus, the most effective, highly syntactical languages are precisely those which are least suited for the dialogue. Documents which are indexed in such a language should be translated into another, more primitive language better suited for the interactive search. Here, part of the syntax is preserved through the introduction of composite vocabulary terms which are mechanically generated. Authors

1. Introduction

A characteristic feature of the interactive, mechanized literature search is that during the search process the inquirer has a continual overview of how the file responds to his inquiry. The intermediate results of his search appear on his terminal screen or in the printout for closer inspection, and thus he is kept constantly informed of whether the search is proceeding as desired. A search which is continually monitored in this fashion can be redirected and reoriented very flexibly. In the case of simple inquiries and of low demands on the accuracy of the search results the questioner can initiate and perform the search himself.

Such information retrieval systems have been repeatedly studied and compared in recent times. However, in most of these cases ease of use, cost, and accuracy, in terms of ratios of precision and recall, have been the main subjects of interest (1)–(6). Less attention has been devoted to the disadvantages inevitably associated with an inverted, pre-sorted file organization, and to the weaknesses of the indexing language to which an interactive system must restrict itself.

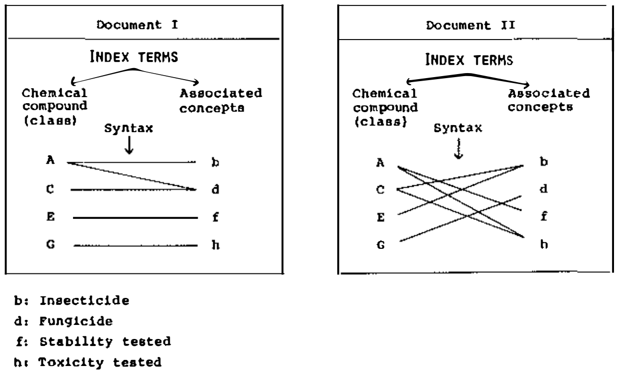
Much practical experience has been accumulated since the advent of interactive systems. Inadequacies of a less obvious nature have become more apparent, such as the failure to retrieve information despite of the fact that it is relevant and in the file. Those placing the queries have also begun to sense more fully the problematic nature of their attempts to formulate a "good" search request, with the result that collaboration between the inquirer and an information specialist is again being recommended with increasing frequency (17)–(22).

Mechanized information supply is subject to several lawful regularities, and this holds true also for the interactive version of this process. To arrive at an understanding of these regularities it is helpful to consider them in relation to the concept of "order", which is best defined for this purpose as "the meaningful proximity of the parts of a whole at a predictable place" (23), (24). It is easiest to gain an overall impression of the nature of these problems and of their interrelations when retrieval is looked upon as an order-creating process (23)–(28).

Operational retrieval systems vary widely with respect to how closely they approximate the ideal state of perfect order, of which it is typical that the relevant literature is compiled *exhaustively* (i.e. without loss of relevant information contained in the file) and *purely* (i.e. without the noise of irrelevant information) for the questioner. They also vary greatly with respect to whether and how much preference is given either to precision or to recall. Both the degree and kind of order, as attainable through the retrieval process in a collection of documents, are controlled by two important system parameters. High *predictability* (23), (35), (36) of the representation of concepts in the file is a precondition for a high recall ratio. High *fidelity* of concept representation, on the other hand, is necessary for high retrieval precision. The vocabulary and the syntax of an indexing language predetermine which degrees of predictability and fidelity and hence, search *accuracy*, are attainable through the employment of this indexing language, at least in theory.– Another important parameter is the *reliability* with which an indexing language is in fact used in the practice of indexing. The most reliable mode of working is one in which the *most appropriate index terms* are always used for representing the concepts and statements of a document (29)–(34). This goes beyond what is commonly understood by using a controlled vocabulary. This "mandatory indexing" (24), (34) requires perpetually searching through the vocabulary for the most appropriate terms and therefore also requires a very systematically organized vocabulary, easy to survey.

* Dedicated to Professor Rolf Sammet on the occasion of his 60th birthday.

** Hoechst AG, 6000 Frankfurt/M. 80, Postfach 800320, Federal Republic of Germany.



DOCUMENTS DIFFERING WITH RESPECT TO THE SYNTACTICAL RELATIONS AMONG THEIR INDEX TERMS

Figure 1

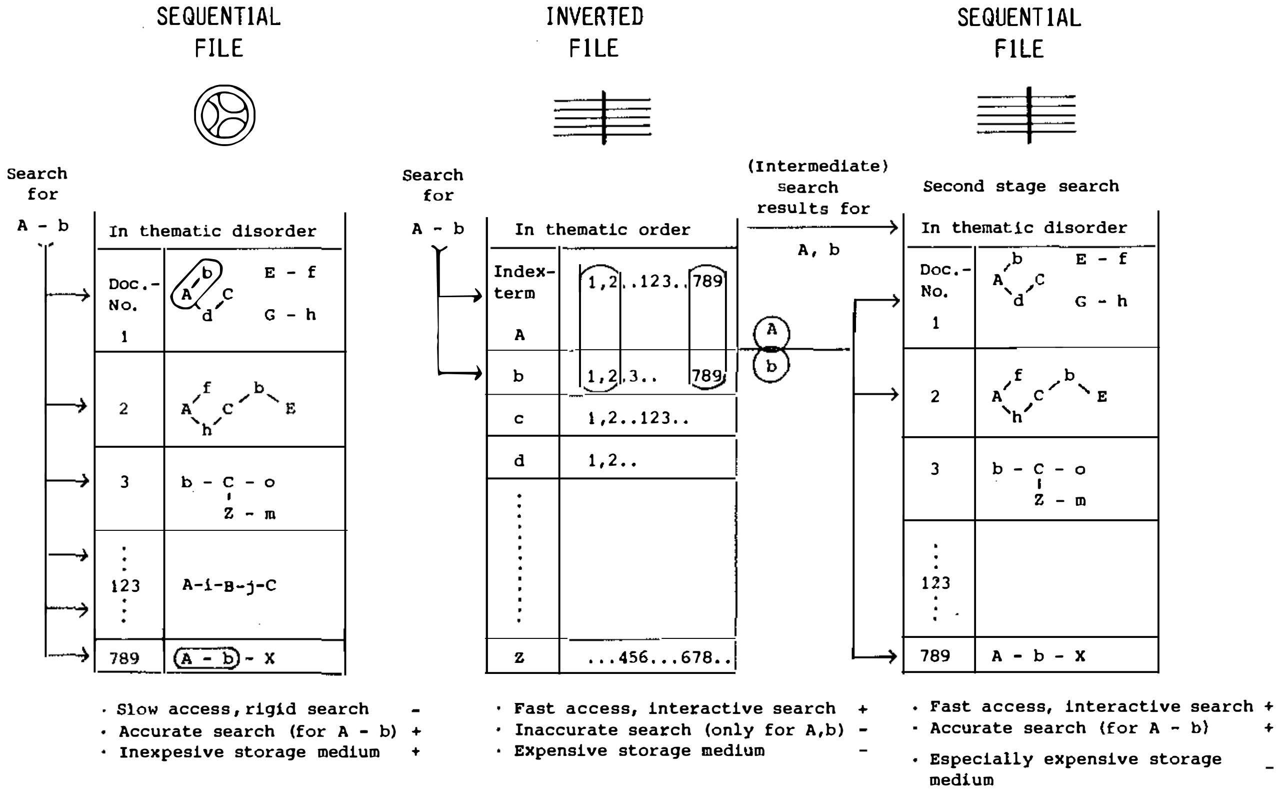
2. File organization and its implications for dialogue suitability and search accuracy

In order to achieve a sufficiently accurate information supply it is sometimes important to store the essence of documents with an extremely high degree of fidelity. Let us take as examples two publications I and II from the field of agriculture and plant protection (see Fig. 1). The same group of four chemical compounds or classes of compounds (A, C, Ė, G) and the same set of properties and test results (b, d, f, h) may be recorded in both publications. These documents may differ significantly with respect to how these properties are associated with the chemical compounds. This is symbolized in Fig. 1 by connecting lines. For example, in document I the toxicity of compound G may be reported, but in document II only its fungicidal activity is given. The accuracy with which this kind of concept coordination can be expressed in the search file will very largely depend on the type of file organization employed.

3. The sequential file

In the establishment and updating of an information system it is quite common to enter one publication after the other into the file. Each item in this file is made up of the complete set of index terms that pertain to the corresponding document. If the indexing language possesses a syntax of its own, then the syntactical relations between the index terms are also completely represented in this record, at least as far as they can be expressed in the syntax of this indexing language. This is much in the interest of the fidelity of the indexing-lingual mode of representation and, hence, in the interest of retrieval accuracy. In such a sequential file no thematic order prevails, since the publications are entered in a purely random or, at most, chronological sequence (cf. Fig. 2, left-hand column). The medium most commonly used for storing sequentially organized files is magnetic tape.

Sequential file organization requires the file to be searched from beginning to end for every inquiry, if a complete overview of the information supply in the file is to be provided. For example, compounds from class A


SEQUENTIAL FILE
INVERTED FILE
SEQUENTIAL FILE
Search for A - b
In thematic disorder
Doc.-No.
1
E - f
G - h
2
3
b - C - o
Z - m
123
A-i-B-j-C
789
A - b - X
Search for A - b
In thematic order
Index-term
A
b
c
d
z
1,2..123..789
1,2.3.. 789
1,2..123..
1,2..
...456...678..
(Intermediate) search results for
A, b
Second stage search
In thematic disorder
· Slow access, rigid search -
· Accurate search (for A - b) +
· Inexpesive storage medium +
· Fast access, interactive search +
· Inaccurate search (only for A,b) -
· Expensive storage medium -
· Fast access, interactive search +
· Accurate search (for A - b) +
· Especially expensive storage medium -

ADVANTAGEOUS (+) AND DISADVANTAGEOUS (-) PROPERTIES OF SEQUENTIAL AND INVERTED FILE ORGANIZATION

Figure 2

which exhibit fungicidal activity may be sought. This is symbolised in Fig. 2 through "A–b". Then, even the last document in the file may be a relevant one, as is shown by document no. 789 in Figure 2.

Sequential searching in large files takes much time, and it is hardly possible to intervene if the search should proceed in an unexpected and unwanted direction, since intermediate, representative results are not available. This is in sharp contrast to the requirements of the dialogue. On the other hand, due to the availability of the index terms *in their specific syntactical relations* the search is most precise. In the example of Fig. 2 only publications No. 1 and 789 are identified as relevant. They not only match the index terms of the query (A, b), but also satisfy the requirement for their correct syntactical relation.

4. The inverted file

Before examining additional properties of sequential file organization, we will consider the opposite organization principle embodied in the inverted file (see Fig. 2, centre). This file is ordered according to the terms of the indexing language (A, b, C, d, . . .). Under each term are listed the numbers of the documents to which this term pertains, a form of organization very similar to that found in a conventional index. For example, index term A was assigned to documents No. 1, 2, 123, 789, and term b refers to documents 1, 2, 3, 789. The orderly arrangement that prevails in a file which is organized on this principle also has its merits and demerits.

It is an advantage that the documents pertaining to a particular index term, for example A and b, are no longer randomly scattered throughout the file but are placed at the location of these terms. This permits rapid access to those documents which satisfy the search instructions or are at least candidates for satisfying them. Those documents which eventually satisfy all the requirements of the search instructions are identified by a simple intersection operation. It is apparent in the example that the terms A and b co-occur only in documents No. 1, 2, 789. The storage medium most commonly used for this kind of file organization is the magnetic disk.

If as a first response to a search in such a file it is reported that under term A only very few documents, or no documents at all, are registered, then the search terminates very quickly. The searcher becomes aware at a very early stage that term A was perhaps too specific a search instruction and that he would be well advised to generalize this search parameter. Thus the inverted file can be searched very rapidly and flexibly and is well suited for the interactive search.

5. The omission of syntactical relations in the inverted file

Apart from the considerable expenditure involved in updating such files, it is disadvantageous that syntactical relations between the index terms are lost in the inverted file. This is expressed in figure 2 (centre) through the suppression of all connecting lines between the index terms. Consequently, syntactical relations will also have to be omitted from the search instructions for the inverted file. Thus, one is restricted to requiring the mere co-occurrence of the search terms in the documents to be selected. However, in many cases this is a serious distortion of the topic of the inquiry, and this lack of fidelity will lead to irrelevant information. For instance, in a search for insecticidal ("b") members of compound class "A", publ. No. 2 will be provided as an irrelevant response. Here, substances C and E are reported to be insecticides. Substance A was only subjected to tests for toxicity and stability. This is obvious from the representation of document 2 in the sequential file of Fig. 2.

In other words, the dialogue search is based almost exclusively on the vocabulary of the indexing language. Even if syntactical relations have been expressed by the indexer, they cannot be utilized for the dialogue search in an inverted file.

An exception to this rule are those concept relations which the indexer has expressed through the syntactical device of "segmentation". In this case the document is subdivided into segments and those index terms which are conceptually associated in a certain, predetermined way are brought together in a common segment. Then the above-mentioned intersection procedure is not only based on document numbers, but, more specifically, also on segment numbers. This guarantees that terms of the inquiry which are syntactically related in a specific manner, will co-occur not only in a document, but also in a segment of this document, where the desired syntactical relation between these index terms prevails. For example, in an indexing system it might have been agreed to represent the particular relation between substance and property through the syntactical device of segmentation. Then, however, this device will no longer be available for representing other syntactical relations, for example the relations between the components of a mixture or alloy (part-whole relation), the sequence in time, logical relations (for example, to differentiate between "A and B" on the one hand and "A or B" on the other hand) etc. Consequently, the effectiveness of segmentation for representing and preserving syntactical relations in the inverted file is rather limited.

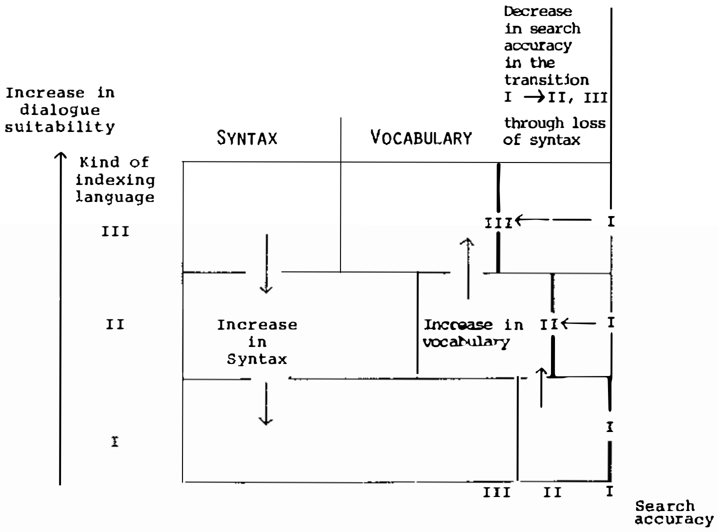
6. The partitioning of tasks between vocabulary and syntax

The designer of an indexing language has a great deal of freedom to decide how the task of representing concept relations shall be partitioned between vocabulary and syntax. For example, one might decide to introduce a specific, precoordinate index term into the vocabulary to express the concept of "insecticidal compounds of class A", and thus to dispense with the use of a syntactical device to link the terms "insecticides" and "compound class A". We shall not analyse all the properties of an indexing language vocabulary in which these precoordinate index terms dominate. It should, however, be remembered that these vocabularies are normally very large and furthermore continuously growing, and that they are interwoven with an extended and highly branched network of concept relations. This renders the search for the most appropriate terms for document representation difficult and time-consuming. In practice, mandatory indexing with such a vocabulary is impossible in most cases.

Search accuracy in a file based on such a syntaxless indexing language is correspondingly low, inspite of the highly specific, precoordinate index terms in its vocabulary. In syntactical indexing languages, on the other hand, the vocabulary can be kept small and easy to survey without sacrificing fidelity of representations at the same time. This is much in the interest of reliable indexing and, hence, accurate searching.

We refer now to our statement (cf. Fig. 2) that syntactical relations between index terms are lost in an inverted file, at least to a large extent. *Thus, the most sophisticated syntactical indexing languages, which perform excellently in searches in the sequential file, lose much of their effectiveness when exposed to interactive search conditions.* The syntax of these languages cannot contribute its share to search accuracy, and one finds oneself restricted to their relatively small and unspecific vocabularies. The opposite extreme is an indexing language the expressiveness of which is solely based on its vocabulary and the performance of which will therefore inherently be lower than that of a syntactical indexing language. However, the search accuracy in this language will not be impaired through the inversion of a file based on such a language. Thus, it may well be that in the interactive search a primitive indexing language performs better than a more sophisticated, syntactical one. This kind of inverse relation between dialogue suitability and search accuracy is schematically depicted in Fig. 3. The indexing languages I, II, III differ markedly with respect to the degree to which they employ syntactical devices.

Thus, if searches are to be made in both the sequential and inverted files and if high search accuracy is required, it is advisable to use different languages in which concept representation is differently partitioned between vocabulary and syntax.

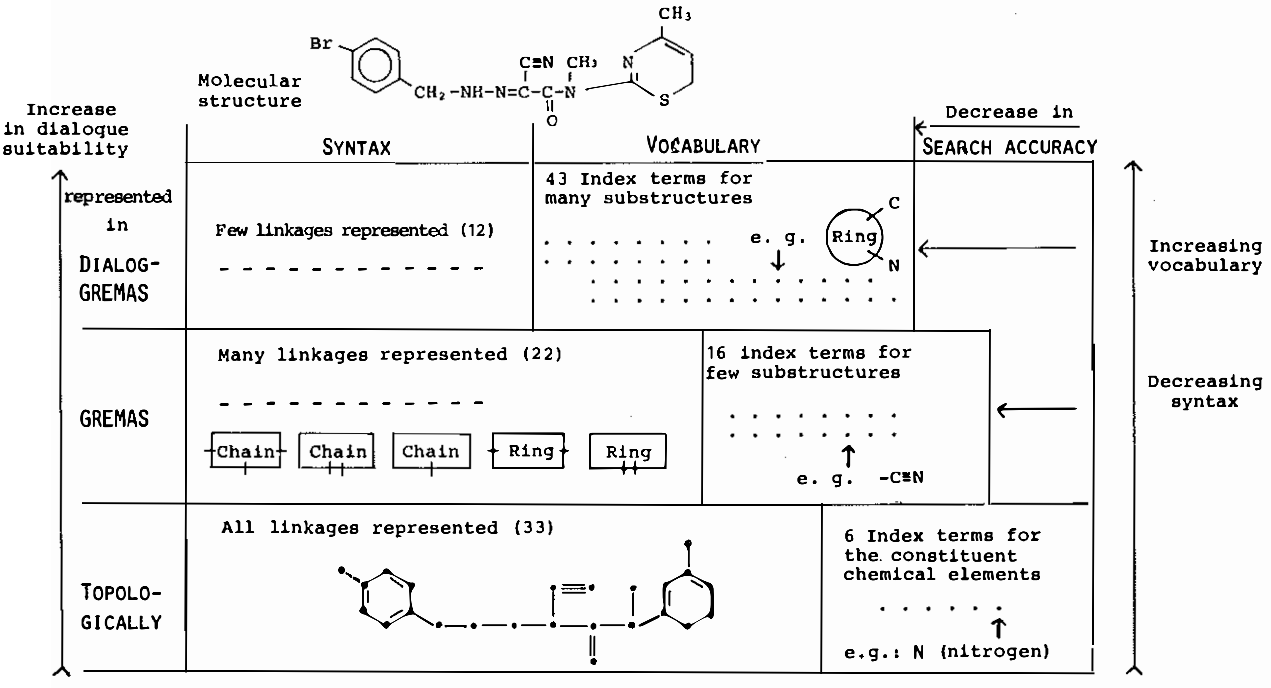

Increase in dialogue suitability
Kind of indexing language
III
II
I
SYNTAX
VOCABULARY
Decrease in search accuracy in the transition I →II, III through loss of syntax
Increase in Syntax
Increase in vocabulary
III←
II←
I
III II I
Search accuracy

RELATION BETWEEN DECREASE IN SEARCH ACCURACY AND INCREASE IN DIALOGUE SUITABILITY

Figure 3

7. Syntactical differences among chemical indexing languages

Exactly this approach is the one we employed in the IDC system for the documentation of organic chemical compounds. In Fig. 4 one and the same chemical compound is schematically represented in three languages. The smallest vocabulary is that of the topological representation. It is limited to the approximately 100 different atoms which make up all chemical compounds. To


Molecular structure
Increase in dialogue suitability
represented in
DIALOG-GREMAS
GREMAS
TOPOLO-GICALLY
SYNTAX
VOCABULARY
Decrease in SEARCH ACCURACY
Few linkages represented (12)
43 Index terms for many substructures
e. g.
Ring
C
N
Many linkages represented (22)
Chain
Chain
Chain
Ring
Ring
16 index terms for few substructures
e. g. -C≡N
All linkages represented (33)
6 Index terms for the constituent chemical elements
e.g.: N (nitrogen)
Increasing vocabulary
Decreasing syntax

• Each dot indicates an index term

- Each line indicates a syntactical relation between index terms

REPRESENTATION OF A CHEMICAL MOLECULAR STRUCTURE IN THREE INDEXING LANGUAGES DIFFERING WITH RESPECT TO THE EXPRESSIVENESS OF THEIR VOCABULARIES AND SYNTAXES

Figure 4

represent the formula of our model compound only six index terms are required from this vocabulary, namely the designations for the atoms of carbon, nitrogen, oxygen, sulfur, chlorine, and hydrogen. Extremely high expressiveness, which is typical of topological representation, is achieved mainly through the recording of all (syntactical!) relations of each atom in the compound. Thus each connecting line in Fig. 4 represents such a syntactical relation.

In the IDC system the topological representation of a chemical compound is transformed by computer programs into the indexing language GREMAS. The vocabulary of this language is considerably larger and many more index terms from this vocabulary are used to describe the same compound. The contribution made by syntax is drastically reduced, which is symbolized by the decrease in the number of connecting lines.

In the indexing language Dialogue-GREMAS the vocabulary, and also the role it plays in the representation of a chemical compound, are even larger, while the role played by syntax has become correspondingly small. This form of representation is also generated by computer, in this case from the GREMAS-code.

8. The combination of inverted and sequential files

In the actual practice of performing dialogue searches the need for highly accurate searches will steadily increase as is evident from increasing complaints about the lack of precision of on-line searches in commercially available data bases. The solution to this problem is to perform a dialogue search in the inverted file as the first stage in a two-stage process (Fig. 2, center). In the second stage the document numbers which are identified as candidate references in the first stage, e.g. Nos. 1,2,789, are looked up in another file which is sequentially organized and in which the syntactical relations between the index terms are preserved (Fig. 2, right). Hence the syntactical search instructions of the inquiry, which had to be omitted in the dialogue search of the first stage, can be fully utilized in the second file. In the second search stage it will be recognized, for example, that document No. 2 is irrelevant because it does not satisfy the syntactical search requirement. In consequence, sequential-file search accuracy is finally achieved even in a dialogue search of this kind.

Sufficiently rapid access to the documents in the sequential file of the second stage is provided by also storing this file on disks. The switch-over from the inverted file disk to the sequential file disk can be performed so fast that the inquirer has the impression he is carrying on a genuine dialogue. In fact, however, the system works, as it were, in a "pseudodialogue" mode, because it switches, though only intermittently, to the sequential search mode to weed out irrelevant responses from the genuine dialogue.

This principle of combined file organization and search is not altered if the complete set of index terms and their syntactical relations are filed after each document number in the inverted file. This is occasionally done in highly sophisticated dialogue systems and substantially speeds up access to the sequential part of the file.

9. Conclusion

Through the artifice of combining two opposite modes of file organization almost all the drawbacks inherent in each of these can be overcome. A fast and highly accurate search is made possible in the dialogue mode, even though this apparent dialogue is not always a genuine one. Furthermore, both the vocabulary and syntax of the indexinglanguage are fully employed in a subsequent sequential search. No saving of machine costs, however, is achieved by such a combined approach. On the contrary, in the combined system a particularly large amount of the expensive disk storage medium is required, an expense which is in addition to the relatively high cost of continually updating the inverted files. This is, however, at least in part counteracted through the savings in manpower which would otherwise have to be employed for weeding out the surplus of irrelevant responses in a genuine, syntaxless dialogue. The inquirer will, nevertheless, accept the increased expense if he places a high value on a fast and flexible supply of accurate information.

References:

(1) Hall, J. L.: On-line information retrieval sourcebook. London, GB.; Aslib (1977); ISBN 0-85142-106-7.

(2) Doerk, B., Fischer, R.: Erfahrungen mit externen Magnetbanddiensten. In: Nachr. Dok. 29 (1978), p. 69.

(3) Blanken, R. R., Stern, B. T.: Planning and design of online systems for the ultimate user of biomedical information. In: Inform. Proc. & Management 11 (1975), p. 207.

(4) Kaback, St. M.: A user's experience with DERWENT patent files. In: J. Chem. Inform. Comp. Sci. 17 (1977), p. 143.

(5) Kaback, St. M.: Chemical structure searching in DERWENT's World Patent Index. In: J. Chem. Inf. Comp. Sci. 20 (1980), p. 1.

(6) Kaback, St. M.: Retrieving patent information online. In: Online 2 (1978), p. 1, p. 16–25.

(7) Santodonato, J.: A comparison of online and manual modes in searching Chemical Abstracts for specific compounds. In: J. Chem. Inf. Comp. Sci. 16 (1976), p. 135.

(8) Tomea, Albert V.; Sorter, Peter F.: On-line substructure searching utilizing Wiswesser Line Notations. In: J. Chem. Inf. Comp. Sci. 16 (1976), p. 223.

(9) Weiss, Irvin: Evaluation of ORBIT and DIALOG using six data bases. In: Spec. Libr. (1976), p. 574.

(10) Almond, R. J.; Nelson, Ch. H.: Improvements in cost-effectiveness in online searching. In: J. Chem. Inf. Comp. Sci. 19 (1979), No. 4.

(11) Bowman, C. M.; Davison, L. C.; Roush, P. F.: On-line storage and retrieval of chemical infonnation. In: J. Chem. Inform. Comp. Sci. 19 (1979), No. 4, p. 228

(12) Bechtel, H.: Erfahrungen mit online-Recherchen. In: Nachr. Dok. 31 (1980), p. 41.

(13) Pichler, H.: Erfahrungen mit online-Recherchen. Nachr. Dok. 31 (1980), p. 85.

(14) Vickery, A.; Batten, A.: Large-scale evaluation study of on-line and batch computer information services. London, GB, University of London, Jan. 1978.

(14a) Collier, H. R.: The on-line tenninal. In: 1. International On-Line Information Meeting. London, GB, Dec. 13–Dec. 15, 1977. Oxford, GB: Learned Inform. (1977), p. 33–38.

(15) Hartley, D.: Investigation of some aspects of on-line searching of Chemical Abstracts Condensates. Nottingham, GB, April 1976.

(16) Hawkins Donald T.: Multiple database searching. Techniques and pitfalls. In: Online, (1978), April, p. 9–15.

(17) McCarn, D. B.: Online systems, techniques and services. In: Ann. Rev. of Infonn. Sci. and Technol. 13 (1978), p. 85–124. Washington DC, ASIS (esp. p. 104).

(18) Lawrence, B.; Weil, B. H.; Graham, M. H.: Making on-line search available in an industrial research environment. In: J. Americ. Soc. Inform. Sci. 25 (1974), p. 364.

(19) Krentz, D. M.: On-line searching-specialist required. In: J. Chem. Inf. Comp. Sci. 18 (1978), p. 4–9.

(20) Rogalski, L.: On-line searching of the American Petroleum Institute's databases. In: J. Chem. Inf. Comp. Sci. 18 (1978), p. 9.

(21) Seba, D. B.: Online in court. In: Online 1 (1977), No. 4, p. 24–27.

(22) Teitelbaum, P.: Use of multi data bases. In: Husbands, C.W. (Ed.): Information revolution. 38th ASIS Annual Meeting, Oct. 26–30, 1975, Boston, Mass. Washington DC, 1975. p. 130.

(23) Fugmann, R.: The theoretical foundation of the IDC system: Six postulates for information retrieval. In: AslibProc. 24 (1972), p. 126, 129.

(24) Fugmann, R.: Towards a theory of information supply and indexing. In: Intern. Classificat. 6 (1979), esp. p. 14–15.

(25) Landry, B. C.; Rush, J. E.: Toward a theory of indexing II. In: J. Americ. Soc. Inform. Sci. 21 (1970), p. 358.

(26) Ludwig, B. M.; Glockmann, H. P.: The formal analysis of document retrieval systems. In: J. Americ. Soc. Inform. Sci. 26 (1975), p. 51.

(27) Kraft, D. H.; Lee, T.: Stopping rules and their effect on expected search length. In: Inform. Process. & Management 15 (1979), p. 47.

(28) Robertson, S. E.: Theories and models in information retrieval. In: J. Doc. 33 (1977), p. 128.

(29) Soergel, D.: Indexing Languages and Thesauri: Construction and Maintenance. Los Angeles, CA.: Melville Publishing Co. 1974. p. 52.

(30) Smith, Linda C.: Artificial intelligence in information retrieval systems. In: Inform. Process. & Management 12 (1976), p. 189.

(31) Bates, J.: Information search tactics. In: J. Americ. Soc. Inform. Sci. 30 (1979), p. 205.

(32) Wellisch, H.: A flow chart for indexing with a thesaurus. In: J. Americ. Soc. Inform. Sci. 23 (1972), p. 185.

(33) Svenonius. E.: Good indexing: A question of evidence. In: Libr. Sci. Slant Doc. 12 (1975) Paper D.

(34) Fugmann, R.: POLIDCASYR, the polymer documentation System of the IDC. In: J. Chem. Inf. Comp. Sci. 19 (1979), p. 67.

(35) Mills, J.: Progress in Documentation. In: J. Doc. 26 (1970), p. 123.

(36) Michel, O.: Ungelöste Probleme im Vorfeld automatisch unterstützter Dokumentationssysteme. In: Nachr. Dok. 26 (1975), p. 11.