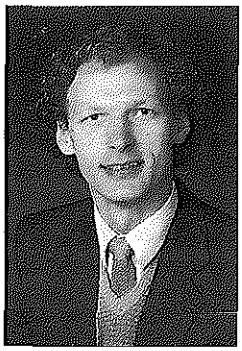
H. Peter Ohly
Informationszentrum Sozialwissenschaften, Bonn



Conceptual Information Retrieval by Knowledge-Based Programming Techniques

Ohly, P: **Conceptual information retrieval by knowledge-based programming techniques.**
Int.Classif. 18(1991)No.3, p. 148-152, 29 refs.
Information retrieval operates on the assumption that concepts can be recognized. This is only poorly supported by traditional document processing because documents are treated as separate units. Sophisticated retrieval techniques, such as bibliographic coupling or probabilistic retrieval, use statistical analyses of the context to improve retrieval results. However, extensive meta-analysis or control of the retrieval process requires knowledge-based programming techniques. The project AKCESS, which is discussed, aims to determine informants on social science topics, examining the relationships between documents, authors, institutions, and concepts, i.e. characteristics of a scientific community. (Author)

1. Information Retrieval as Concept Analysis

Like content analysis information retrieval is based on the manifest meaning of textual corpora. The unit of retrieval (i.e. analysis) is a description of a *bibliographical unit:* a book, a project and so on. The text units consist usually of different kinds of textual material: a title, an abstract, indexing terms, and bibliographical information (author, journal title, year of publication, and so on). They can be regarded as different sections, categories or fields.

The recording units in the analysis are terms. It is common to reduce to stems to broaden the meaning of the search term.

By combining terms by description section, by sentence, or by adjacent order a more precise meaning can be retrieved. Apart from the AND- and OR-connection of terms there is the negation NOT, the absense of a term.

A significant aspect of retrieval is the fact that the computerized analysis is not a final step but an intermediary means of finding the most appropriate documents that conform to a search formulation. This search statement can be considered as a *hypothesis* (or more precisely as two hypotheses):

> Concept x (as expressed in a term combination) can be found (or found to a certain quantity) in the data base
>
> The formulation of concept x in the search is identical to (or close to) representations of concept x in the retrieved documents

Only if these hypotheses are valid, can relevant literature be retrieved and utilized by the user.

As bibliographical text is stored for the purpose of *universal usage* by potential user groups, document processing must occur. The original information is transformed into a documental form: the information is attached in a condensed or standardized form to content-bound categories. To support concept searches, subject terms (descriptors) are used to express the content in a controlled vocabulary with syntactical links. This coding process can be fairly precise (e.g. by pre-combination of subjects) or more general (through facilities for post-combination). The latter is performed with uni-terms which enable more flexibility in the search - which means more recall but less precision.

To sum up, documentation and retrieval reaches a *conceptual level* (where conceptual means more or less generalized information to a user relevant topic) by:

> Retrieval technology (e.g. truncation)
>
> Informative abstracts
>
> Meaningful text-categories
>
> Controlled vocabulary with grammar

Often the controlled vocabulary has a built-in semantic structure, a so-called *thesaurus,* which determines fixed relationships like NARROWER, BROADER etc. This is not only a guide for the indexers but also for the searchers and provides by this a similar conceptual comprehension of the single descriptor terms themselves.

On the other hand, *restrictions* on the retrieval process are as follows:

> 1. Manual indexing can never be fully standardized (or if automatic: can never be complex)
>
> 2. New themes of documents and questions are increasingly outside the scope of the documentation philosophy of the design period
>
> 3. The satisfaction of the user is determined by the usability of the information and not the preciseness of the retrieved documents

The third point is the most peculiar to information retrieval and dissemination in comparison to scientific content analysis. Of ultimate interest is not an exact match, a testing of a hypothesis or a descriptive analysis, but the *usefulness* (or pertinence) of the results of the

analysis to a user-determined goal. That is, the retrieval might be perfect (relevant) but the information content unsuitable for a given problem. If one is trying to find a lecturer who can talk about "Social Inequality" it is useless to find hundreds of bibliographic records without a proper address. Similarly a user who is interested in "Social Security in Eastern Germany" for economic reasons is not interested in outdated Marxist statements on this topic.

More attention must therefore be given to a flexible and background-oriented search procedure, which enhances the information flow in a user oriented way. Structures that lie beyond the manifest content of single documentation units and determine successive retrieval sessions must be found and integrated into the search process. Until now only a few techniques have been available for a context-oriented match between users and documents within retrieval procedures.

2. Sophisticated Retrieval Techniques

Some retrieval languages like MESSENGER (used in the Scientific Technical Network) offer a procedure which computes on a given retrieval outcome a *frequency distribution* of the hits (usually of the subject descriptors). This procedure, called EXTRACT or SELECT sometimes calculates also the percentage and the relevance factor for these co-terms. Including these terms in successive search steps can be helpful in broadening or narrowing the original search concept (28).

Independent of on-line retrieval, INIST (Institut de l'Information Scientifique et Technique, Nancy) produces upon request a co-occurrence analysis for a given search output (MAPINDEXES). The result is presented in a multidimensional mapping of the *term relations* and gives an idea of the overall content structure of the retrieved documents (1).

Direct usage of document and/or term links for the retrieval process is recommended in *"cluster-retrieval"*. Documents resp. terms are clustered by similarity in the term-space resp. in the document-space. The retrieval is thus separated into a two step procedure. First, the search must fit into a document or term cluster, and secondly, the final documents must match the query. In this way a more homogeneous but also more general output is provided (21).

Another quality of tracing conceptual kinship is reached by *bibliographic coupling*. The Institute for Scientific Information, Philadelphia, offers for the SCIENCE CITATION INDEX a method of finding similar documents by using similar citation lists. Starting with one hit, the user approaches all literature which has citations in common with the first one in a succeeding step (5).

"Probabilistic retrieval" algorithms were tested in SMART and other experimental systems which classify the retrieved documents according to term distributions per document set and depth of indexing per document (20). In this way we get a weighted output, which means a rank ordering of the documents. Another approach, "relevance-feedback", incorporates user-ratings of the relevance factor to weight the terms (e.g. FAKYR (4)).

These techniques have been discussed here as sophisticated retrieval insomuch as they take into consideration the conceptual context with the aim of refining or expanding the set of retrieved documents. However, these methods are *user dependent*, that is, the researcher has to determine the next step, and they are *document oriented*, as they do not combine information which is in the data base but scattered over different documents - with the exception of statistical information.

3. Knowledge-Based Retrieval

New perspectives on retrieval and goal oriented content analysis have been opened by the introduction of expert systems (9). Documented information is not only found by a straightforward, precise algorithm but also by logical *inference* over the retrieval steps and over the sum of all relevant information provided in different documents. Therefore, this knowledge no longer has the quality of stable information content which can be hit or missed. Instead, the best combination of information details which solves a given information problem is obtained by use of a knowledge base. The query itself and the quality of the existing data sources direct the solving propositions by inferential means. A further important software tool is the *object-oriented* approach, which models relations between problem types. Specific processing algorithms belong to each object. They can be activated, inherited or transformed according to the structure between these objects. In this way, a prototype version can be easily developed and further expanded in a consistent and clear way (15, 17).

The linkage of knowledge processing with data banks is acquired by two approaches. One tries to handle *growing knowledge bases* by support of data bank retrieval and updating features. The other attempts to make more use of *existing data banks* for problem solving, e.g. in office information (KOFIS (2)), (29).

Yet another approach is the use of expert systems to support the special tasks of *documentation intermediaries*. In this case the knowledge base is independent of the data bank and assists indexing or retrieval processes.

Examples of expert systems in *indexing* are AUTOCAT (7), which automatizes the bibliographic categorization of articles, and TOPIC (11), which summarizes text material in different ways (text, graphics, tables). An extension of the latter is WIT, the condensation of different documents into one state-of-the-art report (16).

Retrieval access to data banks is supported by systems that guide selection and retrieval in different data banks (EASYNET (25), CASS (27), (14)). A transformation of user requests into adequate search formulations is usually provided for specialized subject fields (CANSEARCH (22), MOSS (18), TOME SEARCHER

(26), SAFIR (23), KONDOR (19)). Automatic analysis of retrieval sessions and incorporation in a dynamic semantic space enables user dependent optimization of thesaurus structures (e.g. TEGEN (10)). Recent developments aim at integration of and communication between different (or distributed) expert knowledge bases for different types of problems, users, and documents (INSTRAT (3), I3R (6), CODER (8)), (13).

4. AKCESS: Assistauce by Knowledge-based Context Evalnation in Social Science information retrieval

The Social Science Information Centre (Informationszentrum Sozialwissenschaften) is developing a knowledge-based system to enable a more flexible search than by conventional means (12). The existing system consists of two data bases containing bibliographic information on literature (SOLIS) and information on research projects (FORIS) both concerning the social sciences. These data bases can be accessed by conventional information retrieval tools, that is Boolean searches. Transformation of the retrieval output onto a meta-level will be performed by the analysis of *network information* which is of different types and degrees of completness manifested by the documents of the data bases. Conventional retrieval is thus only used to refine the section of the world that is analysed.

The aim is to process searches which take into account contexts going far beyond the frame which is given by the individual documentation units themselves. Instead of considering only the information within the given limits of bibliographic records, a flexible knowledge base system must be introduced which allows an *information link* between different sources of information, namely between different documentation units.

As is shown in *Figure 1* a chosen *content* (approached in the retrieval process by subject descriptors) is embedded in many relations with other knowledge objects. This contextual information can be used to evaluate the original information material or even improve or alter parts of it.

The focused content is part of a *document* representing a project or an item of literature. A bibliographical description of literature can itself be related to a project description specifying this research project. Literature or projects (that is, documents referring to them) are mostly authorized by *persons* who for their parts are often members of institutions. In other cases *institutions* are declared as authors. Both persons and corporations point to other documents (in our case bibliographic descriptions of projects or literature) of these authors, which are (via these links) connected to the starting set of documents and can be used to enrich it.

Furthermore, within the level of contents, documents, persons and institutions, links can be found which arise from co-occurrences. In a given document a descriptor term is combined with another descriptor term; a content is discussed in relation to another thesis, a

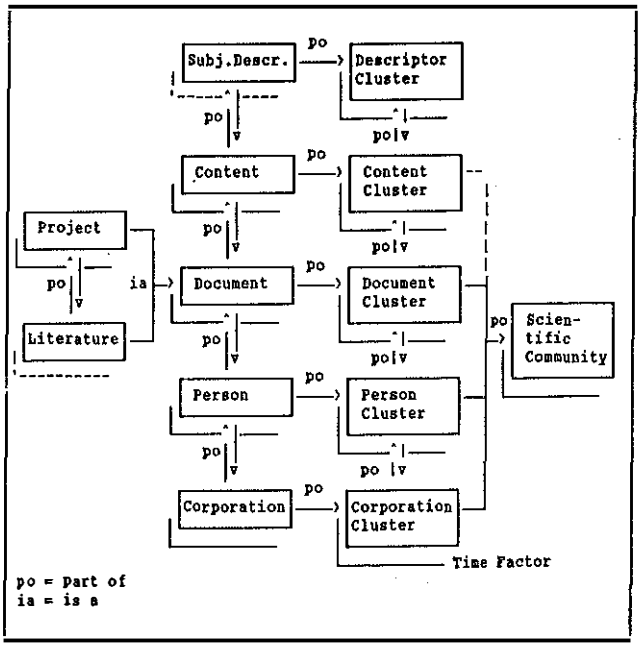


Fig.1: Network of Information Objects

document is itself located in a certain neighborhood (conglomerations of projects, preceding or following articles), persons have collaborators and institutions are in cooperation with other institutions. These links lead in each case to formations, or *clusters* (in a non-statistical sense), which constitute a context that can be used for further information.

These contexts of clusters lead to a super-context which approximates the concept of a *scientific community*. This community is formed by the set of all different context relationships. A community context is strong if all links that can be traced remain in the same set of contents, documents, persons or institutions, that is, it is reflexive or homogenous over a finite number of nodes.

This conceptual scheme of content-links over bibliographic documents is not only multidimensional in the sense that it is connected with other documents via persons, institutions or themes in multiple ways, but also because of dynamic differentiation. Research projects and literature are *time* bound which leads to successive instances of documents and gives a time dependent quality to edges. In this way the existence or non-existence of linked nodes (or objects) for a given period is determined. In the same way, location dimensions could be stated but will be omitted here neglected as the existence of relations itself reflects the importance of regional boundaries, which become more and more relative with respect to physical space.

Initially, this conceptual scheme will be elaborated only for the task of evaluating persons as **informants** on a given subject, e.g. to find an author for a paper on "Mathematical Sociology" or to bring together a discussion group on "New Perspectives of a United German Sociology". It is evident that by restrictions on limited problem fields the world with rules which establish links

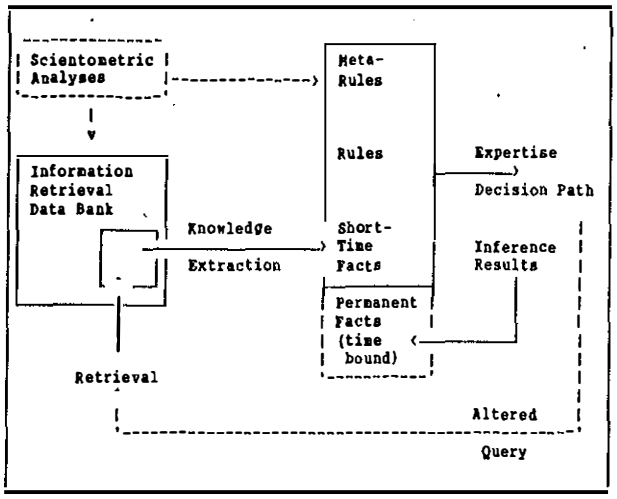


Fig.2: Model of Knowledge Processing on Retrieval Output

between information atoms becomes smaller and easier to handle. On the other hand, it must be said that additional information other than that available in the bibliographic data bank is needed to draw a practical conclusions. A human expert would easily take into account information on reputation, availability, cost factors and so on to reinforce or to suppress the findings of the information system. This leads to occasional context knowledge which is external to the data banks and will be ignored for the moment since no indicators for these evaluations can be found in the given or additional information sources.

As the underlying data base is designed as a relational data system using ADABAS and available expert shells are self-contained systems using PROLOG or other logic or object-oriented languages, the AKCESS-system will be developed in a prototype version as a stand-alone system which uses the preprocessed output of a retrieval search on a given topic of interest as a *factual* component for a knowledge base (see *Figure 2*). This procedure guarantees that only those knowledge facts which are in the context of the actual topic of interest are held in the memory and that the most recent version of the information retrieval data bank is used. This substitute for retrieval output will be stored together with *rules* which direct the information flow over the different contexts onto a starting term (or logical expression on terms) (see *Figure 3*).

```
person and co-person are part of a person cluster
     if person and co-person are in the same institute

or

the total relation between person and content is high
     if the number of individual relations is greater than 3

or

person and co-person are part of a scientific community
     if the number of clusters between them is greater than 2
```

Fig.3: Example of Rules (Conclusion with premise)

Furthermore meta-rules are to be stored in the knowledge base dependent on *feedback* from the results of the inference procedure and from empirical probabilities which are derived from external *scientometric analysis* on comparable data (see *Figure 4*).

```
if inconsistencies are found
     then elect the shortest connections

or

if the number of documents per person is lower than average
     then ignore all relationships of this person

or

if ties leave the diagnosted scientific community
     then use these ties only for the evaluation of members
```

Fig.4: Example of Meta-Rules

This knowledge base consisting of the basic facts, the rules to connect them and the rules to supervise the rules is used to give suggestions for selecting the most suitable informants (lecturers, authors and so on). The decision will be accompanied by a structure of the *decision path,* the provisional *results* of the inference process and the final *expertise* in the form of a check list. Lacking but necessary information for the inference process and user parameters (mostly on meta-rules) give the user the opportunity to interfere during a consultation session by adding new knowledge or by restarting the evaluation procedure with alterated decision weights. If desired the evaluation process can be used to retrieve all documentation units which can be found in connection with the final proposed persons or which have been referred to in the knowledge inference procedure. This provides some kind of visible material as proof and gives possibilities of external testing at the development stage.

If an acceptable outcome is found, some results of the knowledge processing will be stored as time-dependent *permanent facts* that might be referred to in later evaluations for new topics. Information on clusters are an example of facts of this kind. All other initial or derived facts should be erased as they are only of temporary interest for the specified search content. A new topic requires a new search in the retrieval base and the extraction of new knowledge facts.

Dependent on the future availability of standardized tools for knowledge retrieval in external data banks, an interactive solution which retrieves successively only those parts of the data base which are actually used in the knowledge inference procedure will be chosen. For instance. if required, all documentational information on one subject or one author under current examination is loaded and analysed only. When implemented, this design is more powerful than the stand-alone version. In this way it is possible to handle a much larger knowledge store and a variation on the one-world section in one session is allowed. What should not be forgotten is that if we once leave the small domain, additional efforts

must be undertaken to allow different views, enlarge the meta-rule system, and avoid infinite conclusion processes.

Acknowledgement: The author thanks Prof. Agnes Kukulska-Hulme, Ashton Univ., Birmingham, for a critical inspection of this paper.

Note: This is a revised version of the paper "Information retrieval and knowledge", presented at the ISKO pre-conference of the Society for Conceptual and Content Analysis by Computer, SCCAC, Darmstadt, Germany, 14 August 1990.

References

(1)INIST product information 1989 (INIST, Château du Montet, F-54514 Vandoeuvre-les-Nancy Cedex)

(2)Appelrath, Hans-Jürgen: Die Erweiterung von DB- und IR-Systemen zu Wissensbasierten Systemen. In: Strohl-Goebel, H.: Dt. Dokumentartag. München: Saur 1984, p. 408-421

(3)Belkin, N. J.; Seeger, Thomas; Wersig, Gernot: Distributed expert problem treatment as a model for information system analysis and design. In: J. of Inform. Science 5. 1983, p. 153-167

(4)Bollmann, Peter; Konrad, E.; Zuse, H.: FAKYR - A method based system for education and research in information retrieval. In: Salton, G.; Schneider, H.: Research and development in information retrieval Lecture notes in computer science 146. Berlin: North Holland 1983, p. 13-19

(5)Bonitz, Manfred: SCI auf CD-ROM oder das größte Expertensystem der Welt?. In: Informatik 37, 1. 1990, p. 37-40

(6)Croft, W. Bruce; Thompson, R. T. : I3R: A new approach to the design of document retrieval systems. In: J. of the Amer. Society for Information Science 38 (6). 1987, p. 389-404

(7)Endres-Niggemeyer, Brigitte; Knorz, Gerhard : AUTOCAT - Wissensbasierte Formalkatalogisierung von Zeitschriftenaufsätzen. In: Brauer, W.; Wahlster, W.: Wissensbasierte Systeme 2 Internationaler GI-Kongress München. Berlin: Springer 1987, p. 53-62

(8)Fox, Edward A.: Development of the CODER system: A testbed for artifical intelligence methods in information retrieval systems. In: Inform. Proc. Managt. 23 (4). 1987, p. 341-366

(9)Gebhardt, Friedrich: Querverbindungen zwischen Information-Retrieval- und Expertensystemen. In: Nachrichten für Dokumentation 36. 1985, p. 255-263

(10) Güntzer, Ulrich; Jüttner, Gerald; Seegmüller, Gerhard: TEGEN - ein lernfähiges Information Retrieval System. In: Czap, Hans; Galinski, Christian: Terminology and knowledge engeneering. International Congress on Terminology and Knowledge Engeneering, Trier, 29 Sep. - 1 Oct. 1987. Proceedings. Frankfurt: INDEKS 1987, p. 323-337

(11) Hahn, Udo; Reimer, Ulrich : Semantic parsing and summarizing of technical texts in the TOPIC system. In: Kuhlen, R.: Informationslinguistik: Theor., experimentelle, curriculare und prognostische Aspekte einer informationswissenschaftlichen Teildisziplin. Tübingen: Niemeyer 1986, p. 153-193

(12) Herfurth, Matthias; Ohly, H. Peter : Von bibliographischen Datenbanken zu Wissensbanken. In: Deutsche Gesellschaft für Dokumentation: Dt. Dokumentartag 1989. Informationsmethoden: Neue Ansätze und Techniken. 4-6. Oktober, Bremen (DOK-2). Frankfurt/M.: DGD 1990, p. 405-418

(13) Jochum, Friedbert: Einsatzmöglichkeiten und -schwierigkeiten von Expertensystemen zur Vereinfachung von Information Retrieval. In: 4. OTTI-Fachtagung "Informationen in den 90er Jahren", Regensburg, 1-2.3.90 Tagungsbd. 1990, p. 1-6

(14) Klaus, Hans G.: USA: Information Gateways. In: Strohl-Goebel, Hilde: Deutscher Dokumentartag 1987: Von der Information zum Wissen - vom Wissen zur Information. Weinheim: VHC 1988, p. 374-387

(15) Klöppel, Bert; Paul, Manfred : Objektorientierte Programmierung. In: Chip-Spezial, Programm. 7. 1989, p. 5-20

(16) Kuhlen, Rainer; Reimer, Ulrich; Sonnenberger, Gabi: Automatisierter Aufbau von Wissensbanken durch Wissensakquisition aus Texten. Anlage zum Projektantrag "wit". Bericht 4/88 (Oktober). Konstanz: Universität 1988

(17) Mittendorfer, Josef: Objekt-orientierte Programmierung mit SmallTalk und C + +. Bonn: Addison-Wesley 1989

(18) Morris, Anne; Tseng, Gwyneth M.; Walton, Kathryn P.: MOSS: A prototype expert system for modifying online search strategies. In: Online89 information. 13th International Online Information Meeting , London, 12-14 Dec. 1989. Proceedings. Oxford: Learned Information 1989, p. 415-434

(19) Cremers, Armin B.; Oechtering, Veronika; Heege, Rainer: KONDOR - Ein wissensbasiertes Unterstützungssystem zur individuellen Optimierung von Online-Suchstrategien. In: Nachrichten für Dokumentation 39 (1988), S. 257-261.

(20) Panyr, Jiri: Probabilistische Modelle in Information-Retrieval-Systemen. In: Nachr. f. Dok. 37, Nr. 2. 1986, p. 60-66

(21) Panyr, Jiri: Vektorraum-Modell und Clusteranalyse in Information-Retrieval-Systemen. In: Nachrichten für Dokumantation 38, Nr. 1. 1987, p. 13-20

(22) Pollitt, Steven.: An expert system approach to document retrieval. Inform. process. management 23 (2). 1987, p. 119-138

(23) Reiter, Monika: Improving online information retrieval with an intelligent front-end system. In: Online89 information. 13th International Online Information Meeting , London, 12-14 Dec. 1989. Proc. Oxford: Learned Inform.1989, p. 597-604

(24) Reuter, Andreas: Kopplung von Datenbank- und Expertensystemen. In: Informationstechn. 29, 3. 1987, p. 164-175

(25) Roberts, Kay L.: Evaluation of the Easynet Gateway. In: Procceedings of the 7th National Online Meeting, New York, 6-8 May 1986. 1986, p. 375-381

(26) Vickery, Alina: The experience of building expert search systems. In: Online88 information. 12th International Online Information Meeting , London, 6-8 Dec. 1988. Proceedings Vol. 1. Oxford: Learned Information 1988, p. 301-313

(27) Voigt, Kristina; Benz, J.; Eder, A.: Approach for an optimal access to data sources for environmental chemicals. In: Online89 information. 13th International Online Information Meeting , London, 12-14 Dec. 1989. Proceedings.. Oxford: Learned Information 1989, p. 391-402

(28) Keitz, Wolfgang von: Erfahrungen mit EXTRACT SHOW für scientometrische Analysen online- erfaßter Datenpools. In: Strohl-Goebel, Hilde: Deutscher Dokumentartag 1987: Von der Information zum Wissen - vom Wissen zur Information. Weinheim: VHC 1988, p. 202-220

(29) Woehl, K.: Automatic classification of office documents by coupling relational databases and PROLOG expert systems. In: 10th International Conference on Very Large Databases: Proceedings (Singapore, Aug. 1984). Saragota, Ca.: VLDB Endowment 1984, p. 529-532

Address: Dipl.Vw. H.Peter Ohly,
IZ Sozialwissenschaften, Lennéstr.30, D-5300 Bonn 1.