Reports and Communications

Semantic Retrieval
Workshop on 15-16 March 1994 at Heidelberg

The workshop was organized by the Informationszentrum Sozialwissenschaften Bonn and IBM Informationssysteme GmbH, Wissenschaftliches Zentrum, Heidelberg with Matthias HERFURTH and Gerhard RAHMSTORF respectively in charge. With a realistic review, based on experience, of the status and current problems of traditional information retrieval, the two gentlemen mentioned, together with Prof. HOEPELMANN (head of the Computer Linguistics Department at IBM) and the project reporters H. LEIN, H.J. STEFFENS, and G. GOESER, presented an overview of the current development status of the "Semantic Retrieval" project. The problems existing in this field were discussed with the experts present.

The core of the project is formed by the "semantic thesaurus", which consists of a computer-resident network whose nodes are assigned to concepts and whose edges are made up of the relations existing between these concepts, such as these relations are encountered in texts and queries. The concepts are represented by natural-language words or by phrases of various types. The relations are not merely represented in a formal topological fashion, but are also subdivided by type. This is an essential characteristic of the semantic thesaurus.

The definitions for the words of the thesaurus are expressed as phrases and serve for the algorithmic generation of the conceptual network with its extensive array of relations to super- and subordinated concepts. The recording capacity of such a network exceeds by far that of traditional thesauri, because the network also permits such concepts to be recorded for which no lexical expression has been developed yet and which therefore, at least for the time being, can be expressed only phrasally. Such a conceptual network, because of the possibilities it offers as to completeness and machine operability and because of its high systematicness, is suitable for various purposes in terminology, artificial intelligence and information science. At the workshop, attention was centered on its applicability in the retrieval field.

On the present experimental scale, the contents of texts to be made retrievable are expressed with the aid of a highly defined vocabulary from which thereupon, in the usual case, the expert will form specific phrases for a more precise mapping of the given subjects. Through the phrasal mode of expression the otherwise unmanageably large variety of uncontrolled natural-language expression is curtailed and moved into the realm of promising linguistic analysis. Each of these phrases is to automatically find its place in the network of the semantic thesaurus, with the words occurring in them forming the basis for this dovetailing process and the significance of the relation-indicating prepositions occurring between them identified by linguistic analysis.

In the same manner the queries of the users of the system are expressed in phrase form. Their (only temporary) place in the semantic thesaurus is likewise determined by the algorithm, which will then, in this fashion, retrieve the phrases stored by the memory in the near and more distant environment. These phrases are then selected, on the basis of a relevance calculation with a relevance limit, preset as desired, and printed out as output in the order of their relevance.

This procedure promises to exceed by far the precision of the traditional purely Boolean linkage of query descriptors, since the mere co-occurrence in the Boolean sense has been replaced by a well-secured linkage of concepts and since in addition the semantic nature of the concept linkage sought can likewise be made effective as a retrieval condition.

The approach distances itself in realistic fashion from the illusory hope, still widespread elsewhere in the AI field, for a satisfactory, purely algorithmic processing of an uncontrolled, natural-language mode of expression. Instead, it bases itself on the intellectual translation of the essence of texts in the form of phrases. Despite this preparatory work by experts, the reliable identification of the meaning of natural-language words, particularly of prepositions, continues to present some problems.

In the course of the two conference days, the reporters of the organizing agencies offered the 30-odd participants ample opportunity to get acquainted in theory and practice with the problems still awaiting solution and to discuss them in detail for mutual profit and advantage. The conference had been well prepared, was perfectly carried through and was generally rated a success. Robert Fugmann

Dr.R.Fugmann, Alte Poststr. 13, D-65510 Idstein

Subject Representation and Information Seeking. Summary of a Doctoral Thesis, Göteborg 1993

by Birger Hjörland

(*Editor's Note: We are grateful to the author for his permission to include his summary in this Section. His doctoral dissertation defended at the University of Göteborg, Sweden is written in Danish. Its title is supplemented by the following subtitle:* **Contributions to a Theory based on the Theory of Knowledge.** *To the summary of 8 pages put together in a binding, the 181 references of the thesis have been added. A few of these appear at the end of this communication. We would like to refer to the author for any further information on this thesis and its references. His address: The Royal School of Librarianship, Birketinget 6, DK-2300 Kopenhagen S.*)

This thesis is based on the assumption that information seeking is the key problem in Information Science (IS). Other problems, such as document representation is subordinate to the problem of information seeking. A general theory of information seeking therefore has the possibility of serving as a theoretical basis for IS.

Information seeking has mainly been studied in two large subareas of information science: "User studies" and "Information Retrieval". It is this author's opinion that both areas are, and have always been, in a crisis and that they are relatively isolated from each other. "User studies" can take a rather holistic perspective over the users' relationships to the system of information sources. "Information retrieval" research typically adopts a very atomistic perspective in studying the "math" between a representation of a query and a representation of a document (e.g. a match based on statistical or linguistic analysis of questions and document representations).

In this work it is assumed that a study of information seeking, which critically analyses the positivistic and idealistic assumptions about knowledge and science in information science, and introduces an alternative view of knowledge, can help overcome the crisis in both "user studies" and in "information retrieval research". In addition it can unite these areas. A non-idealistic view of knowledge and science inspired from a pragmatic philosophy, understands knowledge as a tool shaped in order to increase man's adjustment to his physical, biological, and cultural environment, and sees knowledge as historically and culturally developed products organized in scientific disciplines. Such a view of knowledge is the opposite of a philosophical "idealistic" point ofview. In short it can be called a "realistic" view, but it covers different traditions in philosophy: pragmatism, materials and "scientific/qualified realism".

The users' behaviour (and the subjective perceptions and assumptions behind that behaviour) must be interpreted in the light of the scientific situation in a given area. In the same way, representation of knowledge in documents and databases must be interpreted on the basis of the scientific situation. It is meaningless to investigate the "micro events", the micro behaviour of information searching and representation, if you have no indication whether this behaviour contributes to human knowledge or not. Discussions of positivism, hermeneutics, pragmatism and materialism are almost unknown in the subject literature of Information Studies (and is much underrepresented in English-language literature as compared to Scandinavian, German literature and other European languages; a book like Tolman's (1) is an important exception.)

Knowledge of such problems in the philosophy of knowledge and in the philosophy of science makes researchers much better equipped to interpret information-scientific problems like obsolescence, "overload", the cumulative nature of science, the structures of the information-landscape, the users and retrieval of information and subject representation. According to the empiricist point of view, knowledge grows in one way, fact being added to fact. From Kuhn's theory, knowledge does not accumulate in this way at all, but shifts with the "paradigm" in the field. It is almost unbelievable, that "user studies" are made without any such relation to analysis of theories of knowledge. How can you empirically examine the users' behaviour, when you do not have an adequate model of the users' role in the creation of knowledge or in the development of knowledge in a holistiic perspective? How can information science deal with the problem of "match" if it has no knowledge of how a single paper fits into the structure of science? Words can match, as can sentences, but concepts mean different things in different areas and information science needs to establish how subject-specific terminology is generated. The problem of "match" is most often seen in IS as a simple mechanical question, not as a humanistic/social scientific question of interpretation.

The study of user behaviour is made on the basis of a positivistic theory, and this is assumed to be "objective" research. But the users' behaviour reflects of course their subjective knowledge and attitudes. These subjective attitudes must be interpreted in some way. It is oflittle help to know what information sources are used, if you do not know whether the sources and strategies used are adequate and represent the optimum. How can you assist users by giving access to information sources without some knowledge of what is important and what is trivial?

It is therefore important that empirical studies give up their positivistic assumptions and begin to study user behaviour from the perspective of history, sociology, and the theory of science, etc. Kuhn's famous book (2) is well-known within information science, but it has never really influenced the methodology of the field. Of course, Kuhn's work should be further developed and questioned, but as it stands, it has very important methodological implications for research in information seeking and IS. In this book we try to explicate the methodological consequences of non-positivistic epistemologies such as Kuhn's. Our basic methodological principle is that the point of view in information science should be seen as "methodological collectivism", as studies of knowledge domains (and e.g. "paradigms"), not as "methodological individualism" (as dominating in "the cognitive point-of- view" and other approaches).

This dissertation is organised in the following way:

In the *introductory chapter* we look at the problem of "subject retrieval". A well-reputed Danish dictionary (*Informationsordbogen*, 1991) defines 'subject retrieval' as retrieval of information by use of subject-representation data - which is defined in opposition to 'descriptive data'. We do not agree with this definition, and show that an adequate theory of information retrieval must be based on an adequate definition of 'subject retrieval'.

Our proposal for a definition is: "Subject retrieval is the search for unkown documents (as opposed to 'known item search') whose contents can contribute to the solution of a concrete problem or satisfy a concrete need for information". All kinds of data which can give a cue (even a vague one) regarding the identification or evaluation of potentially relevant documents can be used in subject retrieval, including the document's own data (such as title, abstracts, list of references, author), or data different from the document itself (including classification codes, descriptors, book-reviews, evaluations and citations in other documents).

Chapter two takes a very short view of the very large field of "subject representation data" or "information retrieval languages" and introduces some important distinctions and points-of-view. Among others, we differentiate between explicit subject representation data (which are data constructed explicitly in order to facilitate information retrieval) and implicit subject representation data (which are data constructed for other purposes, but sometimes useful in retrieval). If a publisher is called "Danish Psychological Publisher", this name can sometimes be useful in searching for books about psychology. This holds also if, for example, a journal's name can contribute valuable implicit subject retrieval data.

We also consider the difference between 'content-oriented subject description data' and 'request-oriented subject data' (introduced by Soergel (3) and others). We state that our work is an attempt at consequently applying the "request-oriented" or "need-oriented" line of thought. In this we find support in the philosophical hermeneutics of Gadamer, which states that it is meaningless to claim that a text has a meaning of its own, independent of any interpretation. If it is meaningful at all to say that a text has a meaning in itself, "an objective meaning", this should be seen as the sum of all prior contemporary and future interpretations of that text.

Chapter three treats subject analysis, which is the interpretational process (made by man or eventually by machine), by which documents are analyzed and their explicit subject retrieval data are created.

It is stated that the classification system, the thesaurus, or in general: the 'Information Retrieval Language', which the subject analysis should be expressed in, works back on the subject analysis and functions as a "decision support system" for subject analysis. It is, however, very important to distinguish between the subject analysis itself and the following 'translation process' or 'expression process' in which the result from the subject analysis is expressed in some retrieval language. If these two processes (of subject analysis and subject expression) are not separated analytically, we can never form adequate theories about either subject analysis or about retrieval languages (this important principle is well pointed out in the works of Lancaster and Langridge (see e.g. (4) and (5).

The subject analysis could be more general or more specific (as pointed out in the literature by Lancaster, Soergel and others: An analysis of a document in a pharmacological database like Ringdok would and should be more specific - suited to the needs of the pharmacological industry than an analysis of the same document in e.g. Chemical Abstracts). Subject analysis can have other dimensions too (not previously discussed in the literature). The analysis could be more "abstract" or more "concrete". Concrete analysis is seen as a predominant empirical / positivistic/nominalistic influence. An "abstract" analysis is seen as an important, but underdeveloped alternative or supplementary analysis in line with "realistic" philosophies of knowledge.

Chapter four looks at the concept of 'subject' or 'subject matter'. The 'subject' of a document is seen as that object, that "something", which the subject analysis focuses on and tries to identify.

The discussions of the concepts of 'subject' and 'aboutness' in the literature of library and information science are presented, analyzed and criticized.

Existing theories are interpreted, characterized and criticized from three fundamental conceptions of knowledge and concepts:

1. "*Objective idealism*"/"Conceptual realism" (Plato and scholastic realism), which operates with "permanent, inherent characteristics of knowledge". These permanent knowledge structures exist prior to the individual, subjective perception, and are first and foremost studied by rationalistic methods.

The works on 'subject' by Ranganathan (6) and Langridge (5) are interpreted as examples of this view.

2. "*Subjective idealism*" (Berkley and empiristic epistemology), which sees knowledge and concepts as individual, subjective creations, which are best studied by empirical, psychological methods.

The works on 'subject' by Hutchins (7-8)) (and many other adherents to the concept of 'aboutness') and the 'cognitive viewpoint' are interpreted as examples of this view.

3. *"Realism", "pragmatism" and "materialism"* (John Dewey, "the cultural-historical school in Russian psychology" and others) which see knowledge as biologically, culturally and individually developed structures, suited to increase man's ability to accomodate his physical, cultural and psychological environments, and primarily organized in scientific disciplines. From this perspective, knowledge cannot be studied by either rationalistic or empirical methods alone, but must be studied by both rationalistic, empirical, and historical methods. The method must reflect the object under study.

Melvil Dewey's classification theory states: "No other feature of the DDC is more based than this: that it scatters subjects by discipline". This may be interpreted as an expression of a realistic philosophy of knowledge, because disciplines are historically developed structures which determine the way in which subjects are interpreted and organized. However, the only explicit theory of 'subject' building on this "realistic" epistemology is our own theory, which defines: "*The subject of a document is the epistemological potentials of that document*".

This "realistic" philosophy of knowledge is, in our opinion, essential not only in order to define the concept of 'subject matter', but to remove a fundamental theoretical barrier in information science as a whole.

Where 'objective idealism' will search 'subject matter' in "permanent, inherent characteristics of knowledge" or in permanent, inherent semantic relationships and tries to establish standardized, permanent, fixed ways of analyzing documents (disregarding their potential use), 'subjective idealism' will search for the subject matter of a document in either the author's or in the user's subjective perception of the documents and tries to develop a theory of subject analysis based on the author's psychological world (as done in parts of the modern "cognitive viewpoint"). None of these viewpoints are, however, developed or stated explicitly in the literature. A reason for this might be that these viewpoints - and especially 'subjective idealism' is in contradiction with reality, and therefore it is impossible to formulate the theory clearly without quickly being contradicted by concrete examples from real life. In spite of this, the existing theories of subject analysis and subject matter tend to build on such idealistic philosophies of knowledge.

From "realistic" positions we do not look on 'subject' as either 'inherent characteristics' or as something subjective in an individual way. The interpretation of a document's "epistemological or informative potentials" is theoretically a never-ending process. This interpretative process is part of the same historical-cultural development as knowledge production itself. The discussion about the possibilities of an "objective" subject analysis is therefore intimately linked to the discussions about scientific objectivity. This is the philosophical debate concerning scientific realism. The conditions of subject analysis are linked to the conditions of the scientific creation of knowledge. The state of scientific knowledge functions as the background from which the interpretation of the single document's subject matter is formed by individuals on the basis of their subjective knowledge. The better this subjective knowledge "matches" the state of scientific knowledge, the more "objective" is the analysis.

Chapter five analyzes some methodological problems in information science. IS is dominated by "methodological individualism", that is, it studies knowledge by studying the individual subjects which are carriers of this knwoledge. Collective knowledge is often seen as the sum of the knowledge of single persons. This point of view is related to the formerly described 'subjective idealism', most clearly to the 'reductionism' of positivism.

The alternative point of view is to see knowledge as a developed historical-cultural-social product, that is 'methodological collectivism'. The alternative to studying individual subjects and individual information seeking behaviour is to study knowledge domains, e.g. to study their informational structures, their terminology, knowledge representation, communication patterns and all this in connection with their theories of knowledge and theories of science. Individual subjects' behaviour in relation to use and to representation of information should be interpreted in the light of a disciplinary context.

In information science you could say that bibliometrics represents a methodological collectivism. However, bibliometrics is itself a very positivistic and criticized methodology. Buckland (9, p.22-23) says that in his opinion a reason for the crisis and pathology in the theory of IS is that an area as bibliometrics, being easy to study quantitatively, has had such a big place in the field. Therefore bibliometrics should not be seen as the main methodology for studying knowledge domains. It could be a supplement to other methods, including historical, sociological, and philosophical methods.

It is important that the link between the psychological and the social level is covered. Information seeking is mainly an individual act. In psychology some researchers are working in order to overcome methodological individualism. Information science has lent itself to a psychology (cognitivism), which is based on methodological individualism (sometimes even methodological solipsism!). Information Science should try to keep up with these collective tendencies in psychology. Important modern contributions in English are e.g. Resnick (10), Sinha (11) and Tolman (1). In the last mentioned work is an important discussion of the role of language as a methodological key to psychology and as a means of perceiving the objective world. This has never been grasped in the empirical tradition from Aristotle to modern positivism and cognitivism, but it has been present in other lines of theories from Plato to modern interpretative tendencies in the humanities.

From these studies we must conclude that human concepts and human knowledge emerge as a result of human cooperation and communication. The individual structures of knowledge can only be understood from a collective analysis of the language users. The knowledge of an individual person, his benefits from information systems, the problems and barriers he meets in the utilization of knowledge, is not primarily illuminated by psychological studies of the capacity and mechanics of the brain or by a differentiation between long-term memory and short-term memory, between semantic and episodic memory, etc., but by the knowledge of the social background of the person, his or her social roles and working commitments, educational background and cooperative relationships in addition to knowledge about the nature of the concrete domain of knowledge.

Chapter six is an analysis of information seeking from a methodological-collectivistic point of view. 'The principal uncertainty of information seeking' (a concept introduced by Swanson) is discussed and supplemented by 'degrees of freedom' determined by the scientific cooperation in a given field. Fields with a well-defined terminology and with well-established standards of publishing (and a relatively samll "scattering" of the literature) are giving the individual researcher less "degrees of freedom" than areas with very loose terminology and without established standards for publication. In these last mentioned areas, there are greater possibilities of individual self-expression, but this increases the uncertainty of information seeking by others.

It is argued that the laws orrules of information seeking are not bound to the process or technology of searching (such as the fixation on technology in information science would often like us to believe), but by scientific cooperation, the scientific organization and the nature of the scientific object. The general conditions of informaiton seeking can only be comprehended by going from a methodological individualism to a methodological collectivism.

Chapter seven is about the concept of "information needs" and contains a reinterpretation of R.S.Taylor's classical psychological study of the development of "information need" from the point-of-view of methodological collectivism.

Chapter eight shows that our proposed concept of 'subject' and our conception of subject representation data as related to"structure of relevance" are in accordance with important tendencies in philosophy and psychology.

The *concluding chapter* briefly examines the consequences of this dissertation, points out what kinds of subject searching are relatively well established and where the problems are located. It is concluded that the subject representation of data of libraries and databases have the possibility of giving research a better beginning, but that the problems are very hard and that we have to take a humble attitude towards them, contrary to the ideology of "technological fixes", which has characterized information science since the days of Vannevar Bush.

References

(1) Tolman, Ch.W. (Ed.): Positivism in psychology. Historical and contemporary problems. Berlin: Springer Verlag 1992. 221p.

(2) Kuhn, Th.S.: The structure of scientific revolutions. 1962. 2nd ed. 1970.

(3) Soergel, D.: Organizing information. Principles of data base and retrieval systems. London: Academic Press 1985. 450 p.

(4) Lancaster, F.W. et al: Subject analysis. Ann.Rev.of Inform.Sci. & Technol. 1989. p.35-84

(5) Langridge, D.W.: Subject analysis: Principles and procedure. London: Bowker-Saur 1989.

(6) Ranganathan, S.R.: Prolegomena to Library Classification. London: Asia Publ.House 1967. 640 p.

(7) Hutchins, W.J.: On the problem of "aboutness" in document analysis. J.Informatics 1(1977)p.17-35

(8) Hutchins, W.J.: The concept of 'aboutness' in subject indexing. Aslib Proc. 30(1978)p.172-181

(9) Buckland, M.K.: Informationand information systems. New York: Greenwood 1991. 225p.

(10) Resnick, L.B. et al. (Eds.): Perspectives on socially shared cognition. Washington, DC: Amer.Psychol.Assoc. 1991. 429p.

(11) Sinha, Ch.: Language and representation: A socio-naturalistic approach to human development. London: Harvester 1988. 235p.

New Developments in Subject Analysis, Classification, Indexing, and Retrieval

This was the topic of a further education workshop for librarians during the Annual Conference of the German Society for Classification on March 8, 1994 at Oldenburg University Library.

The following six papers were presented: J.KINGMA, Groningen: Entstehungsgeschichte, Zweck und Perspektiven der Basisklassifikation in den Niederlanden. - H.-J.ZERBST, Braunschweig: Zum Verhältnis von Basisklassifikation und RSWK am Beispiel des Bibliotheksverbundes Niedersachsen/Sachsen-Anhalt. - I.RECKER-KOTULLA, Osnabrück: Praxis der Sacherschließung im Verbund nach der Basisklassifikation und den RSWK. - Ursula SCHULZ, Hamburg: Was wir über OPAC-Nutzer wissen: Fehlertolerante OPAC-Gestaltung. - Friedrich GEISSELMANN, Regensburg: Online-Version einer Aufstellungssystematik. - The workshop closed with a panel discussion on Verbale und klassifikatorische Sacherschliessung im OPAC, chaired by Hermann HAVEKOST, Oldenburg. For further information turn to: Bibliotheks- und Informationssystem/ Universitätsbibliothek, Uhlhornsweg 49-55, D-26129 Oldenburg, Tel.: 0441-7984010.

In Pursuit of Excellence: Quality, Quantity, and Efficiency in the Provision of Bibliographic Records

The Library Association Cataloguing and Indexing Group (CIG) in England will hold its annual seminar under the title given abovefrom July 8-10, 1994 at Olde Bell Hotel, Retford. After the Annual General Meeting there will be a keynote speech by Derek LAW, an expert on automation in libraries and the electronic library. The next morning will be opened by Stuart EDE, the Director of the British Library's National Bibliographic Service: *Fitness for purpose - the future evolution of bibliographic records and their delivery.* A second paper will be given by Dorothy THOMSON (winner of the first Frank McAdams Memorial Award): *OPAC research project.* The afternoon is free for visits. A guest speaker has been announced for the evening of this day. There will be two more papers on Sunday morning: Ruth ALSTON, Principal Assistent Librarian for Essex Libraries: *Running bibliographic services as business units*, and Joan HOLAH, Bibliographic Services of Reed Consumer Books: *The bibliographic continuum.* For further information turn to Mr Stuart James, University Librarian, University of Paisley, PA1 2BE.