Robert Fugmann

The Navigational Index

Fugmann, R.: **The Navigational Index.**
Int.Classif.18(1991)No.2, p. 122-132, 6 refs.
Description of a set of indexes - called Navigational Index -to the volumes 1 and 2 of "Tools for Knowledge Organization and the Human Interface" (Frankfurt: Indeks Verlag 1990/91). This index is intended to provide advanced subject access. In its Basic Index part the *immediate* hierarchical environment of each descriptor is displayed, whereas the systematic part displays the *extended* hierarchical embedment of each descriptor. Together with the Alphabetical Subject Index, included in proceedings vol.2, they facilitate "navigation" in the vocabulary, which is conducive to reliable indexing and appropriate query phrasing. The Systematic Index has been added in full length to the article. The Basic Index is only available on diskette. (Author)

1. Introduction

In this article it is intended to demonstrate some of the capabilities of computer generated printed subject indexes using as an object the Proceedings of the First International ISKO Conference, held at Darmstadt, Germany, on August 14-17, 1990. This subject index constitutes a complement to the Name Index in the Proceedings Volume No. 2, compiled by I. Dahlberg.

Indexing was done and the corresponding thesaurus constructed according to the technique described in the paper presented at the Thesaurus Software Seminar, August 14, 1991 in Darmstadt (1). A second perusal of the papers as recommended there had to be dispensed with so as not to cause further delay in the appearance of the Proceedings Volumes.

It is also intended to provide a kind of subject access which exceeds the capabilities of any printed index by supplying the index on disk. The location of a subject heading or descriptor in a hierarchic arrangement is uncertain in cases of terminological vagueness. Furthermore, a subject of interest to an individual searcher may still be scattered in the vocabulary. For example, in the index described in the following, the phenomenon of conceptual transparency is encountered in descriptors such as "Concept relations, transparency of", "Indexing, transparency of", "Knowledge representation, transparency of", "Search results, transparency of", "Software transparency" etc. In these cases the possibility of undertaking a computer search for the fragment "transparency" will be welcomed.

In the following, the three individual parts of this subject index are described, and the experience gathered during their compilation as well as some of the discrepancies existing between some of the papers of the conference are reported on. The type of printed subject index as described is believed to be particularly conducive to an orientation in the vocabulary both for effective indexing and query phrasing; therefore, as well as for the reasons mentioned in the preface to Volume 2, it is called "navigational".

2. The Navigational Index

The navigational index consists of three parts, namely

. the conventional *Alphabetical* Subject Index,

. the Basic Index

. the *Systematic* Subject Index

The three parts were mechanically generated from one single data base which was formed during the perusal of the individual papers, accompanied by simultaneous subject heading assignment. All subject headings are contained as descriptors in a thesaurus, which was interactively generated during the indexing procedure, using the LIDOS^R software. Some final editing of the indexes was done using the text processing software WORD^R.

Descriptors were assigned to text passages independent of the mode of expression used by the author of a paper. Often, not only identical meanings but also closely related ones were collated under one descriptor. Minor differences in shades of meaning were ignored, often because of terminology vagueness.

2.1 The Alphabetical Subject Index

A search in the conventional alphabetical subject index is started with any promising descriptor that comes to mind. This approach is preferred by those searchers who want fast and simple access and content themselves with the limited search capabilities that are typical of this kind of display.

Whenever a capital letter occurs at the end of a descriptor, this is meant to show that there is at least one other descriptor which should be taken into consideration for searching (and also for indexing). A quotation

mark at the end of a descriptor indicates that - in the opinion of the indexer - this descriptor was used in an unusual meaning by the author of the cited paper. Examples are given in chapter 3.1 (Essence recognition). The Alphabetical Subject Index has been published in the second proceedings volume.

2.2 The Basic Index

The conventional basic index (often also called "entry vocabulary") comprises all descriptors and non-descriptors in alphabetical arrangement. It shows for each descriptor or accepted non-descriptor

. the corresponding *superordinate* descriptor (marked "+"),

. the preferred term for each non-descriptor (marked "use", indented)

. all *subordinate* descriptors of a descriptor (marked "-", indented).

The basic index displays the *immediate* hierarchical vicinity of each descriptor used in indexing, thus assisting in extending the search to closely related concepts. It constitutes the "lead-in" or "entry-" vocabulary and enables the searcher to trace those immediately related descriptors of whose existence he was made aware through the capital end letter in the alphabetical index. A specimen of the basic index is depicted in Figure 2 on page 269 of Volume 2 of the Proceedings. There, the appertaining co-ordinate descriptors for each entry descriptor are likewise presented, marked by a preceding "/". Because of space limitations the BasicIndexhas not been appended to this article. It is available on the diskette which includes all parts of the Navigational Index (see Note 4).

2.3 The Systematic Subject Index

The *systematic* subject index displays the *extended* hierarchical vicinity of the descriptor under consideration. The "-" characters in front of a descriptor indicate its level in the (generic or partitive) hierarchy. This systematic part of the navigational subject index displays at the same time the appertaining citations. Thus, it provides a lucid overview of even extended concept relations and shows how heavily the corresponding descriptors were used in indexing. It saves the user the trouble of looking up all related descriptors that should be taken into consideration for searching (or for indexing) in the alphabetical subject index. It is therefore most helpful for the phrasing of a comprehensive query in which all promising descriptors are included and all those excluded which cannot provide any citations at all.

In this respect the systematic part resembles Ranganathan's Classified Catalogue and its variations. It differs from them in that no notations were used. True, notations would have improved access to the hierarchy when one starts from the alphabetical or the basic index, but this holds true only for the professional searcher who is not deterred by systematic notations. For better orientation the systematic index is preceded by an overview of the top levels of the hierarchy of the thesaurus.

At the same time the hierarchical structure is effective in computer searches in the index diskette for mechanized descriptor grouping in upward and downward searches, if appropriate search software is available.

The hierarchy was constructed specifically for the concepts occurring in the conference papers and according to the meaning of terms and phrases assumed by the present author. It is not claimed that this hierarchy constitutes a comprehensive systematic arrangement of all information science concepts.

The arrangement of the subordinate descriptors is as follows: First, the generically related descriptors are listed for each descriptor, followed by the partitively related ones. Then, all associatively related descriptors follow, introduced by a descriptor of the format Related to < >, where the space between the arrows is filled by the corresponding superordinate descriptor. This kind of descriptors resembles the "false link" in the terminology of the Indian school (see p. 261 (Vol. 1)).

For the purposes of this index, the variations of associative relations are not distinguished from each other and specified, although this might be done according to the following model (1):

-- Indexing
-- - Related to Indexing **by device**
-- -- Index languageS
-- -- Indexing software
-- -- Etc.
-- - Related to Indexing **by results**
-- -- Indexing quality
-- -- Indexes
-- -- Etc.

Any systematic arrangement of concepts is inherently subjective from various angles of view. Subjectivity is, for example, involved in the decision which degree of specificity is to be represented in the hierarchies (see chapter 3.2.2, "Representational fidelity"). It must also be decided under which and under how many superordinate descriptors a specific descriptor should be located (polyhierarchy). For example, the Colon Classification was located both under universal and under analytico-synthetic classifications.

The meaning of natural language terms is often vague or ambiguous, though this often becomes apparent only after discussion or in looking them up. Their embedment into the hierarchies will become equally uncertain, crucially depending as it is on their (often subjectively selected) definition on which any hierarchical embedment of a descriptor is based. In cases of terminological and, hence, hierarchical vagueness of a descriptor, the basic index provides secure access to its hierarchical location.

3. Some discrepancies observed in the conference papers

During the indexing of the two Proceedings Volumes, some discrepancies between terminologies and the views of the contributors became apparent. In the following some of them are discussed. We base these considerations on a definition according to which indexing is

. the recognition of the essence of documents which is to be made retrievable, and

. the representation of this essence to a sufficient degree of predictability and fidelity[1], [2].

3.1 Essence recognition

Essence recognition includes *meaning* recognition. In natural language texts the meaning of words, phrases and paraphrases is highly context dependent. For example,

the meaning of "*management*" in "information management" (as often encountered in the papers) is quite different from the meaning of "management" in "Information needs of business management" on p. 2:194.

"*Order*" in the meaning of a demand for a certain commodity (p. 2:194) is quite different from "order" as understood as a meaningful arrangement (e.g. pp. 2:114, 2:229).

The meaning of "*category*" varies widely in the conference papers, extending from the meaning of semantic, fundamental, or elemental category to the meaning of a merely more or less well defined group of objects (cf. e.g. pp. 207, 209, 2:60)

The "matrix" on p. 234 is quite different from the matrix on p. 23.

"*Subject access*" on p. 2:239 is the mere matching of uncontrolled query words with uncontrolled text words. This is quite different from the subject access meant by the other authors (cf.pp. 11, 35, 2:9, 2:151), who base their work at least on an intellectual subject analysis as reflected in classification notations, and it is also quite different from what is commonly understood by subject access (see also chapter 3.2., Essence representation).

The "*hits*" on p. 2:245 of the same paper reveal themselves as the number of uncontrolled text words recurring in an index, some of which *may* only under certain circumstances give rise to a hit in its classical meaning of a *real* response to a query (cf. e.g. 2:157, see also chapter 3.2.1 Representational predictability).

It has sometimes been said that in a homogeneous field of knowledge polysemy does not constitute a serious problem. This is refuted by the examples in the foregoing, which are only a small selection from those encountered even in such a homogeneous corpus of knowledge as that of the topics of one and the same conference.

Hence, if text words are merely extracted and, thus, isolated from their interpretative context, as is done in the PARADOKS approach described on p. 2:239, neither will their meaning be preserved nor will the role be expressed that is played by the corresponding referent (cf. 2:239, 2:234).

Essence recognition also includes a decision as to what is essential and worthwhile for being stored for retrieval. Concepts in only marginal context have to be omitted in order not to overload the searcher with useless citations. For example, a definition merely reprinted from a common textbook and constituting basic knowledge or a commonplace statement required for the introduction into a topic should not be included in an index. The mere existence of the descriptor "thinking", as an other example, should not cause the indexer to assign this descriptor to a passage beginning with "She was thinking even at that time that..." (2:168).

The framework of the eight facets, however, which constitute the supreme level in the thesaurus hierarchy, provides some certainty for the searcher as to which sort of concepts are available as search parameters and which ones, on the other hand, should be omitted from a query to avoid a blank sort.

3.2 Essence representation

Here again discrepancies become apparent concerning the opinions on the usefulness of uncontrolled natural language as opposed to that of controlled terminology. The Axiom of Representational Predictability (4, p. 121), however, presents the key to an unravelling analysis of this problem.

3.2.1 Representational Predictability

General concepts (as opposed to *individual* ones) and topics are expressed

either

* in a *lexical unit* (i.e. a linear string of characters specifically agreed upon to denote a concept or statement, blanks included (4, p.127) such as a word or a word phrase

or

* in a non-lexical, paraphrasing manner, e.g. by definitions, anaphora (cf. e.g.(5)), etc.

An author as well as an inquirer has an entirely free hand as to his choice of the expression for his concepts or topics. This holds true especially for general concepts. Often, and for good reasons, the non-lexical expression is preferred.

The multitude of markedly different, *conceivable* paraphrases for a general concept or topic, among which an author has an entirely free choice, is unlimitedly large. Often, sentences or parts of sentences quite distant from each other in a paper combine to make up a clear description of what is meant. Therefore, this process of expressing ideas is a markedly *indeterminate* one in Mater's view (p. 47), one whose course and outcome

cannot be predicted and which is therefore not amenable to programming which requires *a priori* commitments as to the decisions which the algorithm should make.

Hence, if uncontrolled natural language expressions are permitted to enter the search file and if these very expressions constitute the material to be searched, the searcher is confronted with an unpredictable, unlimitedly large variety of expressions for (general) concepts or topics. Each of these expressions would have to be made effective as a search parameter. This would have to be done *in advance*, i.e. without knowing the wording of relevant texts beforehand. Since this is inherently impossible *in practice*, the searcher must content himself with *some* of those expressions which come to his mind, expressions which have happened to occur in some texts he encountered recently or which were compiled *a priori* for processing by the program, inevitably more or less incompletely so. Information loss to an undeterminable and unforeseeable extent is the inherent consequence. In the following, some examples for unpredictable variations of paraphrases encountered in the Proceedings Volumes for the phenomenon of prevailing or lacking *transparency* in information systems are submitted.

- "...reducing the complexity of design making it more intelligible..."
- "... the reader may get lost in hyperspace in such a network..."
- "...Browsing through nonlinear networks often leaves people with a feeling of disorientation..."
- "...Our knowledge... for better overview... should be clearly arranged..."
- "... offers good possibilities for browsing and navigating..."
- "... the possibility to visualize the network becomes more and more important.." (lest) "...a tendency to be lost..." (prevails).
- "... could also profit from transparent machine assisted access..."
- "... forms a very complex network of relations which is probably impossible to envision in its entirety..."
- "...inconsistencies become obvious...since the structure might be clearly displayed to the user..."

"Polyhierarchy", as one of many other examples, occurs as "multiple inheritance" which is also expressed as the phenomenon that "a class inherits".. ("characteristic attributes and operations") ... "of other classes in a lattice of multiple parents" (2:138).

Even in one and the same paper both non-lexical expressions and the corresponding lexical expression for one and the same concept are encountered. "Serendipity" is an example:

- "Non-linear, multidirectional, flexible..."serendipitous" browsing" (p. 2:12)
- "... the "exploratory paradigm" to describe unfocussed information seeking..." (p. 2:11)

It is evident that the compilation of all these expressions for their utilization as search parameters is by no means merely a matter of expert knowledge which can be expected of the questioner (as stated in the abovementioned paper on p. 2:244) or of merely looking up synonyms in appropriate dictionaries or glossaries. Much more it is a matter of guesswork, contingency and (fading) memory to compile all these expressions and many others, too. It is only in artificial experimental situations that the deficiencies of such memory- and contingency-based approaches will not yet come into effect or be latent, and only under these restricted and premature conditions will mere word access, where subject access is needed, appear satisfactory.

All this corroborates the statement that "it is insufficient just to extract words from the initial users statement." (p. 225) and that "there is good reason to believe that availability of verbal access points only is not sufficient for an effective retrieval" (p. 2:157) and "that retrieval only for free text elements is suitable only to a limited extent" (2:158) and "The ideal information language is a classification which can be used for searching as a thesaurus" (2:110).

This consideration does not exclude uncontrolled natural language text as a valuable *complement* to intellectual indexing, as is recommended, for example, in the "hybrid" approach on p 2:61.

3.2.2 Representational fidelity

Sufficient representational fidelity (cf. 4, p.123) counteracts noise in retrieval and helps to avoid false or insufficiently precise references in an index. In the navigational index the desirable degree of representational fidelity has not always been achieved, as is apparent from the occurrence of some overly posted descriptors. Here, subdivision into more specific descriptors would be desirable.

In this index to the conference Proceedings Volumes highly specific descriptors have also been avoided. For example, we contented ourselves with the descriptor "citation" and, here, we collected also all the "co-citation" references. Various mutually closely related concepts were also collated under a more comprehensive or even slightly different descriptor. Thus, "transparency" was not distinguished from its antonym "conceptual complexity". Homonyms, homographs, and polysemes are still collected under "polysemes". The degree of specificity chosen for concepts like these might not suffice for another community of users and for more extended files.

The compilation of an index of this kind requires continuous consultation of textbooks and of colleagues and the incessant utilization and acquisition of new background knowledge in an *unforeseeable* manner and extent. Indexing is thus likewise revealed as a markedly

indeterminate process, whose course and result cannot be predicted and, hence, not satisfactorily be programmed. Many decisions are possible only *a posteriori* and are markedly subjective in their goal of achieving the best possible adaptation to the *assumed* requirements of the majority of the later users.

There is no indication that one day a similar level of subject access can be obtained by those purely *a priori* decisions, that are always necessary in programming. Claims to the contrary which pretend to achieve genuine and equivalent or even superior *subject access* in a simple and cheap mechanical way, for example by merely selecting textwords and offering them for match with natural language query words in retrieval, are highly destructive because they jeopardize what exists and fail to provide a workable substitute. They merely provide *word* access rather than *subject* access. The ethical perspective of this problem is discussed by Wellisch[3].

The three parts of the navigational index are available from the ISKO secretariat[4] as reprint or as MS-DOS diskette in the formats of WORD[R] or LIDOS[R].

Notes

1 By "indexing" is meant the description of the essential contents of a document, by extraction and/or assignment of significant terms with or without syntactical relationships with a sufficient degree of fidelity and predictability for retrieval demands. (see (2))

2 "Indexing is the translation of the essence of a document into an indexing-lingual mode of expresseion" and "the task of an indexing language is to represent concepts and statements with a sufficient degree of predictability and fidelity" (see 3)

3 The ethical perspective of this issue is discussed in (6, p.11).

4 ISKO Secretariat, Woogstr. 36a, D-6000 Frankfurt 50, Germany. (Prize for copy of diskette DM 25.-)

References

(1) Fugmann, R.: An interactive classaurus on the PC. Int.Classif.17(1990)No.3/4, p.133-136

(2) FID/CR Terms of Reference. Int.Classif. 8(1981)No.2, p.96

(3) Fugmann, R.: Toward a Theory of Information Supply and Indexing. Treatise VI on Retrieval System Theory. Int.Classif. 6(1979)No.1, p.3-15 (here p.14)

(4) Fugmann, R.: The Five Axiom Theory of Indexing and Information Supply. JAmer.Soc.Inform.Sci. 36(1985)No.2, p.116-129

(5) Liddy, E., Bonzi, S., Katzer, J., Oddy, E.: A study of discourse anaphora in scientific abstracts. J. Amer. Soc. Inform. Sci. 38(1987)p.255-261

(6) Wellisch, H.: The literature on indexing. In: Weinberg, B.H.(Ed.): Indexing - The State of Our Knowleldge and the State of Our Ignorance. Medford,NJ: Learned Inform.1989.

Systematic Subject Index to the papers presented on the First International ISKO-Conference Darmstadt, Aug.14-17, 1990

Overview of the Subject Index Hierarchy

KNOWLEDGE ORGANIZATION

Epistemological issueS

SemioticS
- TerminologY
- Specific semiological conceptS
- LinguisticS

Knowledge representation, audio- visuaL

Information sciencE
- Semiotic trianglE
-- Objects
-- DesignationS
-- ConceptS
- Meaning recognitioN
- Information storage and retrievaL
-- Storage procedureS
-- IndexinG
--- Related to IndexinG
--- Index languageS
--- Index language vocabularY
--- --- Classification systemS
--- --- ThesaurI
--- --- Free text vocabularieS
--- Index language grammaR
-- Search procedureS
-- Data base selectioN
-- Search request maturatioN
-- Subject searcH
-- Fact searcH
-- Searching by expertS
-- End user searcH
-- Free text searcH
-- Search, interactivE
-- Query phrasinG
-- ISAR data baseS

Data bases of special fieldS

Information technologY
- Information softwarE
- Information hardwarE
- Telecommunication

Artificial intelligencE
- Machine learning
- Expert systemS
- Connectionist systemS
- Computer linguisticS
- Decision making, algorithmic
- Pattern recognitioN

IndividualitieS
- ProjectS
- Institutions for knowledge organizatioN
- Countries and languages, reported oN

Quotation marks (") are assigned to those descriptors which had been used by the author of the corresponding paper in an unusual meaning.

A capital letter at the word ending informs the user of the alphabetical subject index that at least one other related descriptor should be taken into consideration for subject search.