

# INFORMATION RETRIEVAL *I*NTERACTION

PETER INGWERSEN

## Chapter 4

*This book is available for download at [ [www.db.dk/pi/iri](http://www.db.dk/pi/iri) ]*



© Peter Ingwersen 1992, 2002

Peter Ingwersen

E-mail: pi@db.dk

Homepage: www.db.dk/pi

Royal School of Library and Information Science

Department of Information Studies

Birketinget 6

DK-2300 Copenhagen S

Denmark

*Information Retrieval Interaction* was first published in 1992 by Taylor Graham Publishing. This electronic version, published in 2002, was converted to PDF from the original manuscript with no changes apart from typographical adjustments. It has been ensured that the page numbering of the electronic version matches that of the printed version. Both versions can therefore be cited as:

Ingwersen, P. *Information Retrieval Interaction*. London: Taylor Graham, 1992. X, 246 p.

Published in 1992 by  
Taylor Graham Publishing  
500 Chesham House  
150 Regent Street  
LONDON W1R 5FA  
United Kingdom

Taylor Graham Publishing  
12021 Wilshire Boulevard  
Suite 187  
LOS ANGELES, CA 90025  
USA

ISBN 0 947568 54 9

Converted to PDF in November 2002 by Rasmus Bruun

This classic approach to IR reaches far back into the history of library and information science. It has its roots in the first keepers of registers in early human history and is heavily influenced by the current type of recording and information processing technology: from parchment and paper to electronic disks, and from manual copying over printing into computerization. Already in Roman times it was familiar to separate the storage order of manuscripts on shelves from the order of author and subject entries in catalogues. In the Middle Ages simple classification and indexing of manuscripts took place (Witty, 1973) and, as pointed out by Umberto Eco (1980), such classification systems mirror the socio-philosophical understanding of the world, its ideas and knowledge – like today. The mysticism associated with the work of ‘keepers of books’ originates mainly from such coding schemes and subject ‘labyrinths’, which only an alphabetical key makes accessible for others than librarians.

With a changed attitude in society towards the use of information and the access to scientific knowledge, mainly in North-Western Europe and in USA, the role of librarians and documentalists shifted towards the end of the last century. Libraries opened up their shelves and the core abstract journals in physics, chemistry, medicine and engineering begin to emerge. The overall goal of IR – to ease the users’ access to and the availability of information in documents – is established.

Up to the mid-seventies in this century, however, all researchers concerned with IR focused on (scientific) documents, their content, and how to relate them in a proper way – not on their users. Simple or advanced, ‘marking and parking’ of documents or their surrogates became the objective in IR. Consequently one may state that this approach, which also could be called the ‘system-driven’ or the ‘document-driven’ approach (Ingwersen, 1988), forms a dominant tradition of *paradigmatic* force in IR research.

In retrospect it is interesting to observe that this tradition produces its best results

concerning theories of representation, when still linked to the paper tools but under pressure from the emerging computer and online media after the fifties. It is in this period – the fifties into the seventies – that indexing comes of age and develops into practice based on theory. First, with the computer age, advanced retrieval techniques come into focus. However, it is also the computer technology itself, aided by the demand for information in society – not documents – that shakes the approach in the eighties. Modern information technology makes transparent the ‘garbage in – garbage out’ syndrome in IR.

Notwithstanding, one must keep in mind the outstanding results which the traditional approach has achieved – and still does on a minor scale. Under the influence of progress in linguistic text analysis and AI techniques some traditional IR researchers keep up their spirits. As A. Smeaton optimistically put it at an ESPRIT II research meeting (1990): “Super indexing makes perfect retrieval!”

In contrast, B. Croft describes the theoretical situation (1987, p.249):

..there have been many advances in the field of IR (in the last 30 years), but some fundamental issues remain unsolved. To put it simply, we do not know the best way of representing the content of text documents and the users’ information needs so that they can be compared and the relevant documents retrieved. We cannot even agree on a definition of relevance. Statistical approaches to the analysis of text and retrieval of documents have significant advantages in terms of efficiency and performance relative to other techniques (Belkin and Croft, 1987), but one need only look at the absolute performance levels measured in terms of recall and precision to see their limitations. Dissatisfaction with the current state of affairs is one of the two major factors in the recent upsurge of interest in intelligent IR. The other factor is the increasing awareness of the importance of IR as an application area, brought about by the proliferation of systems that handle text and multimedia documents (van Rijsbergen, 1986).

In the traditional approach the following characteristics can be demonstrated:

1. *Aim and foci:*

Study of text representation (classification, indexing, natural language processing) theory, retrieval techniques, and mechanical components of sources and systems in laboratory settings.

Emphasis on maximisation of retrieval performance by means of comparisons of techniques, theory and experimental design in a controlled manner in database test collections.

2. *Type of results and consequences:*

Ad hoc-based refinement of methods and models for text analysis, representation and IR technique.

IR is understood as a paradigmatic process, i.e. that systems designers, indexers and authors, as well as searchers (i.e. human intermediary and end-user) *per se* do share similar scientific views, terminology, etc.

3. *Understanding of information:*

Information understood as scientific information (and associated with meaning of text).

4. *Use of supporting disciplines:*

Linguistics, mathematics, logic, and computer science as basic supporting disciplines.

Text linguistic (syntactic) methods applied to problems of representation; mathematics and computer science, including AI in recent years, are related to (theory and) design of components and IR techniques.

Associated with the traditional R&D approach one may easily observe the influence of a variety of underlying research traditions. Theory building in relation to classification of document contents has constantly been attempted based on an *ontological* tradition (Vickery, 1975; Boyd Rayward, 1992). Theoretical and applied issues concerning indexing and IR technique development have basically been founded in *linguistic* traditions or approached from a *logico-statistical* perspective.

Very recently, Blair has put forward strong arguments, from a philosophy of language standpoint, against the traditional way of thinking on information retrieval, and representation in particular (1990). In addition, he warns against the assumptions underlying the development of the variety of IR techniques discussed below in Chapter 4.4. Essentially, he suggests avoiding assumptions guided by questions like “what does an expression *mean/signify*?”. Instead he advocates asking: “how is an expression *used*?” (1990, p. 136). One may observe certain similarities to the problems addressed concerning aboutness, in particular the problem of user-aboutness (Chapter 3.1.1). A further discussion of Blair’s context-based position in relation to IR, but essentially limited to representative issues, is carried out in Chapter 7.5.

In order to understand the basic problems faced by the traditional mainstream tradition in relation to *representation*, we may briefly look at theories and developments in classification, indexing and natural language processing. This is followed by an outline of the major *IR techniques* and a brief discussion of relevance adhering to this tradition, in Chapters 4.4 and 4.5. A summary is provided in Chapter 4.6.

#### 4.1 Classification theories

The great universal classification systems, e.g. Dewey’s designed in 1876 for shelf ordering purpose in libraries, and UDC constructed by Otlet and La Fontaine in the nineties, attempt to categorize all (scientific) knowledge stored in documents. These and other derived systems are still in use and under constant development. In relation

to aboutness it is interesting to note that several universal classification systems applied to public libraries are in fact either off-springs of the pure scientific ones, or, as in the Danish case (DK5), is a hybrid between Dewey's original system, UDC and the organisation of scientific disciplines at Copenhagen University before World War II.

Major characteristics of any classification system are that classes must be mutually exclusive and the system must be exhaustive within its domain, so that any document can be placed in one distinct category. As the world changes around the system it becomes more and more difficult to classify new topics and concept relations.

In the fifties and sixties faceted classification theories appear, initiated by the Classification Research Group, UK. These are based on the subject categories in the body of literature concerning a domain, e.g. in engineering or the natural sciences (Foskett, 1962) (Vickery, 1975), and more universal in the Broad System of Ordering (BSO) (Coates, 1983). They all adhere to the famous universal facets – PMEST (Personality, Matter, Energy, Space, Time) – by Ranganathan (1952). Faceted classification, also incorporated in UDC, implies a specific order (e.g. decreasing complexity) of facets in a string.

With the introduction of the computer, several attempts at automatic classification have been made without profound success (Sparck Jones, 1976), and classification expert systems are currently under development (Sharif, 1988).

In terms of aboutness the advantage of faceted classification is that several aspects (instead of one) of the contents of a document can be made searchable. We are here close to actually indexing the document.

Following these lines of IR research, a comparison between for example Ranganathan's or Vickery's facet schemes, and Fillmore's (1968) linguistically-based case grammar, or Lindsay and Norman's LSD scheme (1977), demonstrates intriguing similarities which in the eighties are explored in order to design retrieval systems, e.g. based on morpho-syntactic text analysis and probability (Smeaton, Voutilainen and Sheridan, 1990).

## 4.2 Indexing theory, controlled vocabulary issues

Indexing theory has in general developed around two concepts: use of a controlled vocabulary, or use of natural language inherent in the document text, or a mixture of both.

With respect to *controlled vocabulary*, A.J. Foskett's work on subject indexing provides a review of theories and methods at the edge of automation and commercial online retrieval (1971). This is followed up by F.W. Lancaster (1986b). The theories (and the technology) suggest a string of predefined terms or keyword phrases that typically represent the indexer aboutness of a document. A document with the title 'Danish exportation to India', mainly on butter export, might be indexed: *Denmark; Export; Butter export; India*. Today's commercial online bibliographic databases run

on this type of representation, invented in the sixties to be applied to printed reference tools. In the online databases we may talk of post-coordinate searching for documents, since the term coordination, i.e. the query structure, is made at search time. In printed tools, this structure is pre-coordinate, i.e. the four keywords exemplified above, possibly only in form of single terms, together with other sets of four keywords representing other documents, form a permuted index always headed by one of the four terms. The remaining three terms create a *context* in the printed index which serves to disambiguate the meaning of the entry for the searcher. As may be observed, the example demonstrates that disambiguation is difficult: Who exports to whom? Albeit, most commercial systems have gone no further – for economic reasons.

From a theoretical viewpoint, H. Spang-Hanssen argued the concepts of ‘roles and links’ (1976), based on linguistic cases and other linguistic means. Commercially, this theory has only been made available in certain chemical databases for chemical compounds.

The only commercial general-domain retrieval system actually attempting to solve this problem is the British National Bibliography (BNB). Based on intensive research, Austin developed the PRECIS system for the printed version of BNB (1974), containing a large controlled vocabulary and applying syntactic roles (prepositions) which, in the *human indexing* process, are used to form the string(s) automatically, representing monographic documents. For example: *Denmark; Butter export to India*. However, in an online situation the roles are lost, except in the record display.

The advantage of representation using controlled vocabularies should be that (the content of) new documents can be linked to old ones by consistent use of terms. It is the combination of terms at search time which separates documents or sets of documents. However, the theory presupposes that searchers as well as indexers share the same vocabulary, also with respect to any new concepts that become translated into old ones. This assumption may be dubious. For instance, the query posed to our example above should not be expressed by the terms ‘exportation’ or ‘Danish export’, but by ‘export’ or ‘butter export’ and ‘Denmark’. From a cognitive viewpoint this assumption of vocabulary concordance may only be justified for a limited period of time in confined domains among smaller groupings of collective cognitive structures.

As a proof, also stated in Chapter 3, *indexer inconsistency* occurs, i.e. that only from 10 to approx. 80 % of the index terms added to the same document by different human indexers are similar or identical, mainly attributable to the presence, completeness and stringency of decision rules for applying index terms (Cleverdon, Mills and Keen, 1966), (Jones, 1983) and depending on the source field in the document the indexers rely upon (Tell, 1969).

If *user-aboutness*, including simple weighting of major aspects, as well as roles and links, at least to some extent, should have been introduced in the example above, it might look like Figure 4.1.

As one may observe (Figure 4.1), this solution provides several additional access points for potential users, e.g. ‘import’ or ‘Danish butter’, and improved understanding of the content in the full-text document on the display. Given that user’s query contains roles and links, non-relevant documents on ‘Indian export to

Denmark' are avoided. In addition, this solution is also extremely expensive (time-consuming) to carry out by human indexing – and hence, with the exception of simple weighting, never applied commercially.

**Denmark(a)(l')(x); Danish export(ac)(x); Export(ac)(x); Butter(o); Butter export\*(ac)(x);  
Danish butter export\*(ac)(x); Import(ac)(z); India(a)(l)(z); Danish butter(o); Indian  
import(ac)(z);**  
...

**Roles:** (a) = agent; (ac) = activity  
(l) = location (to); (o) = object  
(l') = location (from)  
**Links:** (x), (z) = agent-activity  
**Weight:** \* = major terms

Fig. 4.1. Indexing example based on simple user-aboutness with controlled vocabulary.

Further, there is a problem of context exhaustivity: how many combinations of terms are relevant and required. The number of index terms in this solution might be reduced or constrained by means of an elaborated domain-dependent search thesaurus which (automatically) might lead the searcher to adequate index terms. Then, of course, the time consumption and introduction of new conceptual relations has moved to the maintenance of the thesaurus.

Surprisingly enough, *automatic indexing techniques* based on the *single words* that occur in document texts are rather effective. Reducing words to stems, excluding stopwords and the incorporation of a simple thesaurus with only synonym relations is the only vocabulary control that has been shown to have definite advantages (Croft, 1987). Indexer inconsistency is avoided, as in natural language representation. This is an example of mainly author aboutness, further demonstrated in Chapter 4.3.

However, the constant theoretical problem is, as pointed out by Rijsbergen (1990), that the indexer (mechanism) does not really know which user-aboutness or semantic value to apply in order to meet a potential user. Perhaps 'butter transport' or 'Indian butter import' might have been an adequate index term? – seen from some users' viewpoints.

*Thesaurus theory* is associated with vocabulary control. Its focus is on concept relations and it displays general relations between terms like generic relations, i.e. broader and narrow terms, part-whole relations, i.e. top and part terms, as well as synonyms and homographs. It serves as a tool for indexers and searchers in a domain, e.g. leading from non-used terms to preferred index terms. A thesaurus can be used for automatic validation of terms generated by human or automatic text analysis. In addition to the general relations, a thesaurus would traditionally operate with so-called 'related terms' which actually contain concepts with the properties of the linguistic cases or facets mentioned above. Often, the related terms mirror the *situational relations* between concepts. Although the semantic relations between such terms are unspecified the concepts themselves may be of great value to users who, in a cognitive and hermeneutic sense, are more familiar with processes than with the

more abstract hierarchies of the objects involved in such events or processes. See Chapter 2.4 for a discussion of cognition seen as a ‘concernful acting’ in a hermeneutic (Heidegger) sense, as well as Chapter 6.1.1 on situational categorisation seen from the cognitive point of view.

Therefore, thesaurus theory is an adequate means for designing knowledge bases in specific domains, e.g. in the form of semantic networks or case frames.

It is intriguing – but quite unsatisfactory – that the indexing theories and their applications do not demonstrate a high effectiveness in terms of IR performance, and do not solve the IR problems mentioned, although a rather *intelligent, knowledge-based component* guided by elaborated rules is involved in the processes: the human indexer. Obviously, the lack of consistency is one reason. Another seems to be the absence of the user’s influence in relation to the IR models, theory, and solutions. Although semantics is introduced by human indexing, a third cause is that the analysed, extracted and translated concepts fall out of context and *immediately drop* from a cognitive to a structural level in a cognitive sense, see Chapter 2.1.

### 4.3 Natural language representation

Natural language representation (NLR) demonstrates a rather different attitude towards retrieval processes by deliberately omitting the human indexer and replacing him with simple or more advanced algorithms. We may here talk only of ‘author aboutness’ at a monadic or structural level of information processing. The sources for representation are document titles, author generated abstracts or full-texts, including citations of other documents.

By avoiding the problems of indexer inconsistency and moving it to a ‘natural’ author inconsistency, the general theory behind is to place the user and his information need closer to the source in the communication process. The theory presupposes a similarity in use of terminology, concept relations, etc. between generators (i.e. authors) and potential searchers. In contrast to vocabulary control, by which users either retrieve nothing or a great deal, depending on the concordance between search and index terms, NLR normally retrieves something because of the variety of author and user generated terms. From a cognitive viewpoint the vast variety of individual knowledge structures, also within the one and same domain, makes this presupposition doubtful, and at the same time inadequate in IR.

A fundamental problem in NLR is an inherently simplistic conception of ‘meaning’ and the absence of an original conception of information. Information is seen as identical to meaning which is identified as author generated expressions.

The classic NLR is in general based on documents in machine readable form and may therefore be carried out automatically and with low costs. Four different approaches towards NLR are discussed:

1. *structured*
2. *single term*
3. *single term in context*
4. *single term with weighting*

#### 4.3.1 *Structured natural language representation*

Structured NLR implies making use of a) the term structures in document texts, e.g. in abstracts or full-text; b) the citation structure related to a (scientific) document. Method a) relies on clustering theory which is discussed in relation to IR techniques (Chapter 4.4.2). Method b) is based on the idea that a citation in document A of document B relates A and B; a document C, also citing B, must hence be related to A. Further, if one document cites A and B, then A and B must be related.

These coupling and co-citation methods, explored by E. Garfield (1979), can be used to generate 'citation clusters' for browsing and searching purposes. The theory follows the relations 1/4 and 3/5 (Figure 1.3, Chapter 1.2). In relation to bibliometric analysis of citations within scientific domains, certain limitations have been expressed as to its completeness (Cronin, 1984), whereas there is no doubt about its validity as science indicator (De Solla Price, 1976). In terms of representation the most severe problem in citation clustering is the weight or representativeness as well as the direction, i.e. the *qualitative cognitive impact*, of each citation in a document.

#### 4.3.2 *Single term natural language representation*

Single term NLR is the most used indexing method in commercial online systems. In combination with added controlled index terms, single words from document titles, and from abstracts (since the end of the seventies), form an inverted basic index as well as individual inverted fields. The Boolean exact match technique extended with proximity operators makes it possible to search for concepts in all these fields – called free-text searching. In full-text systems the vocabulary control combination rarely occurs.

By combining vocabulary control, which could be supported by a thesaurus, and NLR it is theoretically possible to obtain good retrieval performance – providing that the searcher is very good at manipulating the query language and conceptual structures interactively. One may say that this combination creates an *author+indexer aboutness*, superior to each of the two forms of representation individually, in terms of number of entry points and performance. However, their disadvantages also accompany this solution: lack of indexer consistency, and for NLR: *all* the single terms have *equal weight*. When searching on, for instance, 'Danish butter' and

‘India’ in our example, therefore, we may retrieve more documents (e.g. via abstract words and titles), but we have no control of the degree of relevance. The best way to keep high relevance in this free-text mode is to search on title terms combined with added index terms, i.e. in general avoiding the abstract field, or only searching within sentences.

In addition, it is important to note that several investigations have shown that “in general 34–86% of the index terms, assigned by human indexers, can be derived from *title words* only. Depending on the discipline, titles of articles usually describe or at least imply the contents of the document more or less sufficiently; more in the fields of science and engineering; less in the social sciences and humanities” (Borko and Bernier, 1978, p. 163–164). In our first indexing example on ‘Danish exportation to India’ at the beginning of Chapter 4.2, 75% of the index terms added are similar to those in the title.

### 4.3.3 *Single term in context NLR*

NLR based on single term in context has therefore been explored in two directions: a) automatic use of title words from journal articles in printed indexing tools; b) (semi)automatic use of title and chapter heading words from monographs in online databases.

KWOC or KWAC (KeyWord Out of Context; KeyWord And Context) are methods following direction (a). In scientific domains their performance is high, and the cost low. Article titles are scanned and significant words, i.e. the keywords, are extracted, forming a permuted, alphabetic index with the complete title ‘hanging’ as a tail in each entry, providing a context (Stevens, 1965). This solution is similar in principle to the PRECIS solution in the previous chapter. The difference is that all keywords are NLR, and that the context displays natural linguistic syntax. In an online search situation we have no means to control these linguistic cases or roles – as for PRECIS.

It is necessary to point out that KWOC and alike title-based NLR has less value concerning monographs, since their title terms usually are too broad as subject terms. The rich and nuanced content of a book may hardly be described by the few words of the title.

In direct competition with the PRECIS system, direction (b) attempts to apply the *features inherent* in monographic publications. Originally suggested by P. Atherton-Cochrane, the SAP (Subject Access Project) theory makes use of titles and chapter headings in order to improve the subject accessibility in library catalogues (1978).

Wormell extends the SAP method to incorporate captions of tables and figures, as well as back-of-the-book index terms, all referring to specific pages in the publication. This deep-indexing theory, practiced at moderate costs, is well argued (1985) and demonstrates an advantage to most other NLR approaches by pointing *directly* at the portions of text in documents containing the combinations of terms searched for in their headings. SAP does not have to be limited to monographs like

books or reports. It may also be applied to journal articles, their captions and sub-section titles, as suggested by Ingwersen and Wormell (1986). The theory presupposes that captions and chapter headings possess higher representativity than full-text single terms combined.

From a relevance viewpoint this theory is interesting, since relevance must be extended from the usual 'document relevance' concept, which in this case merely is without interest, to part-of-document relevance. Another interesting feature is the possibility of searching *graphics* (figures, graphs, etc. by their headings) which makes it applicable to office information systems and other multimedia environments. Due to the contextual and the deep-indexing characteristics, the SAP principles for NLR come closer to providing 'user-aboutness' associated with a document, than most other indexing theories in use. Note also that Wormell's theory is workable in a general domain environment. If provided with a synonym thesaurus in a more specific domain, it might perform equal to or better than the automatic term extraction supported by thesaurus, as suggested by Croft (1987) in the previous section on vocabulary control. The Esprit Project 2083 (SIMPR) automatically makes use of chapter and sub-chapter headings to form a so-called 'heading hierarchy'. The hierarchy is produced by application of the SMGL standard and can be applied as a navigation tool during searching.

It must be mentioned that neither SAP nor KWAC, nor single term NLR – with or without weighting – may avoid non-relevant retrieval of documents/captions on 'Indian export to Denmark' when users enter 'Denmark', 'India' and 'Export' as search terms.

#### 4.3.4 *Single term extraction incorporating weighting*

This type of NLR is based on document abstracts or full-text words, applying word frequency analysis. Along with the developments of manual indexing theories several methods and theories for automatic extraction have been put forward and tested. Originally based on Zipf's *rank-frequency law* which implies that the frequency of a given word in a text multiplied by the rank order of that term approximates to a constant for that text (1932), automatic NLR has developed around various applications of term frequency.

Intensive research during the fifties and sixties has provided IR with knowledge of the best ways of exploring the issue, leading to strong theoretical frameworks concerning advanced IR techniques in particular. Much of this research is summarized by Salton (1968), Sparck Jones (1974) and (Salton & McGill, 1983). By applying the term-frequency approach Sparck Jones carried out large-scale experiments on *term weighting* (1973). Hitherto, all of the single term NLR approaches outlined produce representations in the form of a very simple 'author aboutness' on a monadic level, where all terms (and documents) are equal. Term frequencies may provide estimates of their relative value.

The intentions behind weighting are twofold: to *rank* texts that contain single query terms, and/or to allow query terms to carry weights whereby documents can be ranked according to those term weights. Relative term weights for each text may be calculated at indexing time or at search time. The various methods of weighting lead to advances in relevance weighting (Robertson & Sparck Jones, 1976) and directly to theories for experimental IR techniques, as from the seventies.

The fundamental theory can be expressed by the *tf.idf* value which, for each term *T* multiplies its relative term-frequency in the text (*tf*) by the inverted document frequency (*idf*) in the collection for term *T*. For (*idf*) the relative frequency ratio  $\log N/n$  is often used, where *N* is the number of texts in the collection and *n* the number of texts that contain the term *T*. Terms with low values are terms appearing rarely in the text and documents containing it appear often in the collection, or the term appears often in both text and in collection. Such terms are poor discriminators. An appropriate lower threshold value is determined in order to select the proper single index terms. By means of other mathematical calculations term-term relation values and term-document association values can be established. These principles are used to relate terms or documents in networks. Note however, that the same term in different IR systems will carry different *tf.idf* values.

In relation to term-frequencies, Salton has suggested that since medium-range frequencies of terms in a text possess higher 'resolving power' or are better discriminators than low or high frequency words, one may handle such terms differently (1975). Low-frequency words may be grouped in classes to increase their effective frequency and high-frequency terms combined as phrases to reduce theirs. This solution implies support from either a thesaurus or a human indexer as 'validator', but omits new, and hence rare terms to be searched. Still better, according to Croft (1987), is extraction of the single terms as stems and calculating statistically their resolving power, supported by a synonym thesaurus. Automatic indexing techniques based on identification of syntactic elements of the document text have been used (Salton, 1968) (Dillon and Gray, 1983), but have not demonstrated high retrieval performance (Sparck Jones & Key, 1973).

Although syntactic NL processing has not in itself proven to improve IR performance, syntactic text analysis *combined* with elaborated rules for concept relationships and *empirically based knowledge* of categories of user goals seem applicable in specific domains to represent and use domain knowledge (Cohen & Kjeldsen, 1987). Their GRANT system makes use of domain knowledge (research funding bodies) represented as a semantic network of concepts, linked by a large number of types of relations and categories. Search is carried out by constrained spreading activation.

This semantic network approach is similar to but more elaborated and constrained, than knowledge representation based on thesaurus theory or roles and links. Because of knowledge of *user preferences* this approach is an example of moving towards a cognitive research approach, discussed in Chapter 7.

#### 4.4 IR technique developments

Mainly based on the single term NLR theories and achievements, laboratory experiments in the seventies demonstrate that *partial match techniques*, such as probability and relevance weighting, may improve the retrieval effectiveness fairly dramatically. A profound review of IR research for this period is produced by McGill and Huitfeldt (1979). A major trend in the eighties was to refine further the effectiveness of partial match and other retrieval and text representation techniques, within the confines of the source system. In connection with the various techniques, improved support of query (not the request) formulation is analysed.

In a recent review Belkin and Croft summarize the current state of R&D on IR techniques (1987). The review provides a new and appropriate classification of IR techniques by categorizing them as *exact match* and *partial match* techniques.

Also recently, in his introduction to "Document Retrieval Systems", Willett provides an excellent review of theories, models and results produced in the classic IR research environment (1988).

*Exact match* presupposes that information needs are identical to queries which again are equivalent to document representations and texts, that provide the information sought. One may say that this technique requires the model of the request, Figure 3.2, be contained, precisely as represented in the query formulation, within the text representation. Boolean or string searching is the common implementation of this IR technique in current operational IR environments. Belkin and Croft (p. 113) point to the most well-known and documented *disadvantages*:

.. a variety of search aids such as thesaurii are required to achieve reasonable performance. In the simple case exact match searching: 1) misses many relevant texts whose representations match the query only partially; 2) does *not rank* retrieved texts [except chronologically]; 3) cannot take into account the *relative importance of concepts* either within the query or within the text [except for weighting e.g. title terms higher than other terms, leading to a simple ranking formula]; 4) requires complicated query logic formulation, and 5) depends on the two representations being compared having been drawn from the same vocabulary.

In addition, the Boolean 'not' logic always results in the omission of relevant texts.

The reasons for not abandoning exact match in the large-scale commercial systems are several, of which the most important are given in a recent empirical survey (Smit and Kochen, 1988). Traditionally, one answer is that the vendors have invested too much in present software to change it for new, non-tested techniques, that is, not tested in large-scale systems. Another reason seems to be that users may apply other (partial match) techniques on their micros so why change policy? Vendors are also stating that results of alternative techniques are not sufficiently better even in experimental environments to justify any changes. A significant argument, actually demonstrating an *advantage*, is that *Boolean statements are structured*, representing important aspects of user requests and problem spaces. A recent review edited by Radecki demonstrates several approaches to improvements of exact match by partial match techniques in online environments (1988).

In *partial matching* the request is regarded as being the query, consisting of the significant terms from the request.

Partial match techniques are categorized into 'individual feature-based' and 'network-based' techniques. The first category contains *formal models*, e.g. the *vector-space model*, fuzzy-set theory and *the probabilistic model*. This category denotes that we are dealing with individual texts' author aboutness features, such as their terms. The second category implies that we are operating on a network of texts, such as in the *clustering* technique, browsing or spreading activation.

The formal models behind the first group of techniques are discussed by van Rijsbergen (1979) and Salton and MacGill (1983). More recent refinements and developments of the models are outlined by A. Bookstein, who compares probability and fuzzy-set theory for applications in IR (1985).

In contrast to exact match technique which *operates on* text representations in the form of manual indexing or simple single word extraction, all formal feature-based and network-based techniques may either similarly compare queries with documents represented as sets of features or index terms, or they may in addition be regarded as *indexing techniques*. They can work on all types of representation, whether controlled or NLR. Features can represent single words, stems, phrases, or concepts and can have weights associated with them. Query features are of similar nature.

Further, in common to *all* partial match techniques is their potential for *ranking* retrieved documents.

#### 4.4.1 *The vector space model*

The vector space model is one of the first models to appear (Salton, 1968) and has been developed and refined up to the present. Documents and queries are vectors in an  $n$ -dimensional space, where each dimension corresponds to an index term. In general, the number of query terms defines this dimensionality. Term weights are calculated by tf.idf values, discrimination ratios found and documents are ranked in decreasing order of similarity to the query using the cosine correlation  $\cos v$  to retrieve documents closest to the query in vector space:  $\cos v = \frac{\sum d_i q_i}{\sqrt{\sum d_i^2} \sqrt{\sum q_i^2}}$ , where  $d_i$  is the tf.idf weight and  $q_i$  the weight for query term  $i$ , with  $0 \leq \cos v \leq 1$ . The model can be used at indexing time, but provides higher performance at search time, because of the dynamic nature of the collection. It may be supported by a thesaurus to include important relationships among words or to expand terms to classes, since it is those relationships and classes that are relevant to an individual query, that should be identified (Croft and Thompson, 1987).

An interesting performance improvement has very recently been reported by Wendlandt and Driscoll who apply query-document similarity measures that include tf.idf weights of the linguistic thematic roles, in addition to the calculated weights of the content-bearing words in texts (1991). In this way semantic values, or conceptual relationship properties of text portions can be retrieved and measured against similar

values in requests on a syntactic linguistic level.

Related to and expanding the vector space model – in particular in association with operational Boolean systems – is the *extended Boolean IR technique* (Fox, 1983; Salton, Fox and Wu, 1983). The model allows for structured Boolean queries to be used, e.g. (*Denmark OR India*) *AND Export*. Texts containing one to all three terms are found and a similarity measure is defined that ranks the documents, e.g. in vector space. Documents that match all or parts of the Boolean query are given precedence. In an example we may have the text on ‘Danish exportation to *India*’ (1), one on ‘*Denmark – India*, an import & transport guide’ (2), and a text on ‘*Export statistics: India*’(3). Naturally, the model cannot solve the (at least), threefold ambiguity problem of what the query really is about: Danish export to India or visa versa, or Danish export as well as Indian export. Ignoring this problem, which only to a certain degree is solvable in the techniques and models, by including ‘relevance feedback’ from the user, and ignoring effects of term weights, the extended Boolean model will give precedence to text (1) and (3) since they both contain two query terms and match the Boolean query specification. Text (2) also contains two query terms (Denmark, India) but they do not match the query specification (Denmark OR India). Note that text (1) and (2) would outrank text (3) if a thesaurus was used, adding the terms ‘Danish’ (= Denmark) and ‘Import’ (Export).

#### 4.4.2 *The probabilistic model*

The probabilistic model generates the most researched and developed IR techniques. The version most often referred to was introduced by S.E. Robertson (1977a) and major contributions to its further development and success come from van Rijsbergen (1977), Sparck Jones and Webster (1980), van Rijsbergen, Robertson and Porter (1980) and, in the eighties, Robertson, Maron and Cooper (1982). Also Bookstein (1985) makes significant contributions, and yet more refined modifications to the theory, increasing its performance, have very recently been proposed and tested by N. Fuhr and C. Buckley (1990).

The techniques are similar to those developed from the vector space model. Belkin and Croft state (1987, p. 117–118):

The basic aim is to retrieve documents in order of their probability of relevance to the query. If we assume that document term weights are either 1 or 0 and that *terms are independent* of each other, this can be shown to be achieved by ranking documents according to:  $\sum d_i q_i$ , where  $q_i$  is a weight equal to:  $\log \frac{p_{ri}}{(1-p_{ri})/p_{nri}} \frac{1-p_{ri}}{p_{nri}}$ , where  $p_{ri}$  is the probability that term  $i$  occurs in the relevant set of documents, and  $p_{nri}$  is the probability that the term  $i$  occurs in the non-relevant set of documents.

The problem in applying this ranking function is the estimation of the probabilities in the query term weights, at search initiation. Laboratory experiments have solved this problem as well as other estimation and weighting issues, e.g. by use of the *tf.idf*

weight. van Rijsbergen has proposed to remove the term independency assumption (1977) and to allow structured Boolean queries to be applied with the probabilistic retrieval model. If term dependencies are used to modify document rankings they must be accurately identified, e.g. by the user or by NL processing methods (Croft, 1986).

It is interesting to note from an office automation viewpoint that 'formal' types of representations, e.g. addresses, dates, zip-codes, etc., may be included with weights in probability IR, as demonstrated for instance by Croft et al. (1990).

#### 4.4.3 Clustering techniques

Clustering is that method among the network-based IR techniques to which most research effort has been devoted during the last twenty years. "A cluster is a group of texts whose contents are similar. A particular clustering method gives a more detailed definition of a cluster and provides techniques for generating them" (Belkin and Croft, 1987, p. 121).

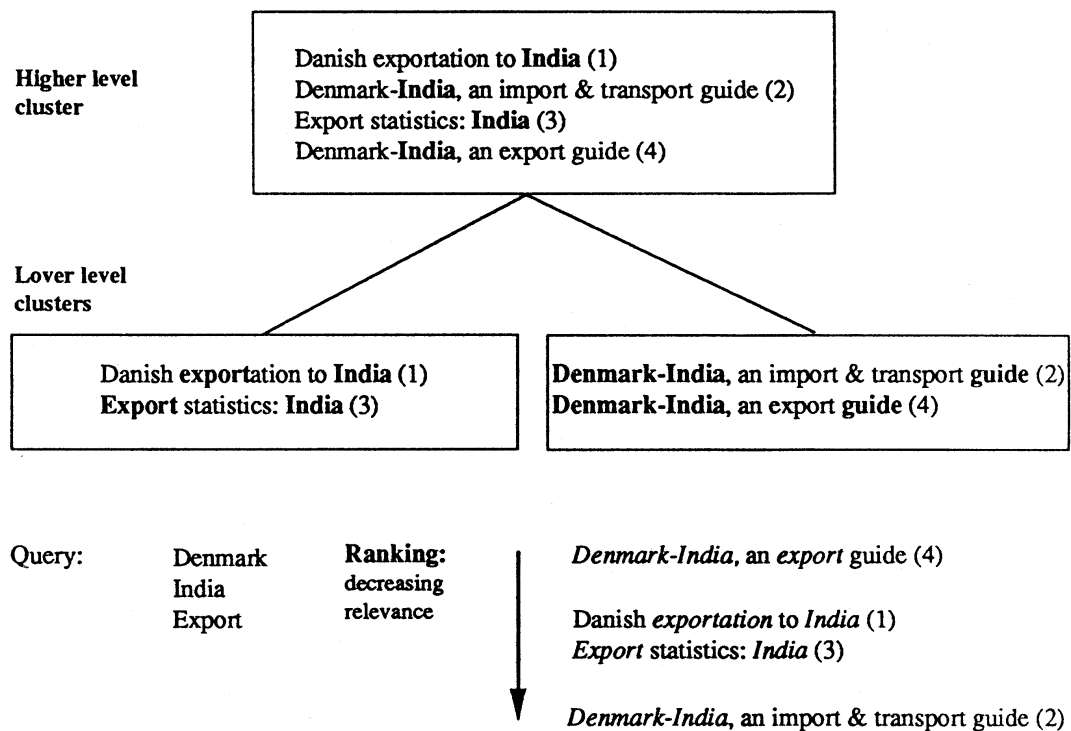


Fig. 4.2. Illustration of clustering of documents and their ranking according to a query. Words in bold imply example of 'cluster representatives'; words in italics identify query matching.

G. Salton's SMART project studied a variety of methods, mainly leading to top-down searching of cluster hierarchies (1968). Large clusters are formed automatically according to the similarity of (index) terms they contain. These are divided into smaller (and denser) clusters, etc. and a query is compared to term representatives of the large clusters by a similarity coefficient. The best cluster is chosen and comparison continues downwards in the hierarchy, resulting in a ranked list of lower-level clusters. The top-ranked clusters' documents are then ranked in relation to similarity to the query. Figure 4.2 illustrates the technique in a simple way. To the three documents on 'export, Denmark, India' mentioned above under extended Boolean logic is added a fourth: 'Denmark - India, an export guide' (4).

Also applying top-down searching Jardine and van Rijsbergen introduced a formal clustering method applying a 'single link' technique where clusters were retrieved without ranking individual documents (1971). A bottom-up method is also possible, as demonstrated by Croft (1980).

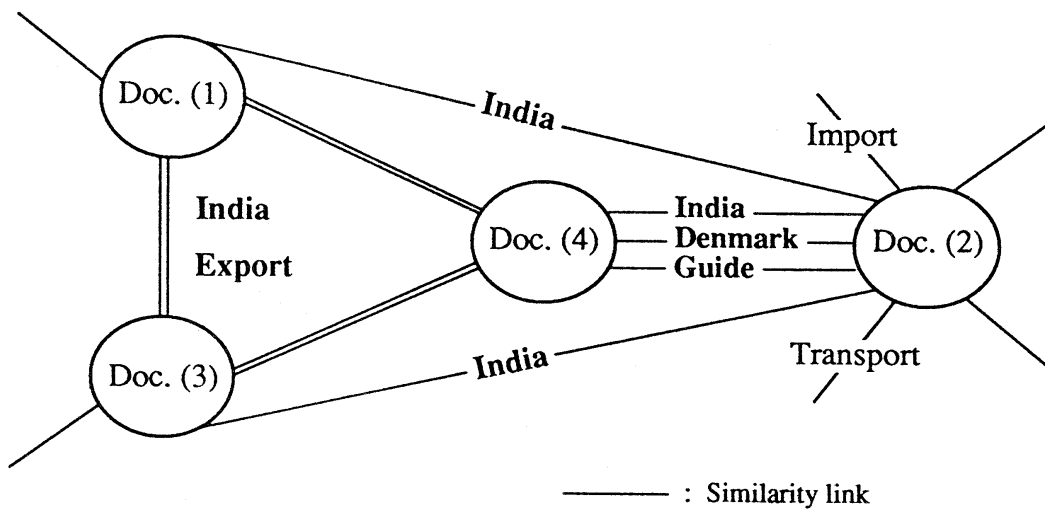


Fig. 4.3. Illustration of nearest neighbour clustering.

The query is compared to the lowest-level clusters, and documents in the highest ranked clusters are retrieved. This technique is similar to a cluster technique using the 'nearest neighbour' method (Willett, 1984). By means of term similarity measures, a document has its nearest neighbour documents linked in a network which, at search time, is used to generate clusters and their representatives. In Figure 4.2 this means for instance, that the four documents at the lower-level have their nearest neighbours linked, as shown in Figure 4.3. In the group to the right the two documents are linked very closely because of the similarity of three terms (Denmark, India, Guide). Individually they are linked to the two documents to the left, the Doc. (1) and Doc. (3).

and (3), in a slightly weaker way: Document (2) is linked with a one-word similarity (India) to both, while document (4) is linked to them with a two-word similarity (India, Export). The two documents (1) and (3) are linked to one another by the same two-words.

This latter method has been explored and refined through the eighties, because it may save storage considerably (Willett and El-Hamdouchi, 1987). Perhaps more important, it can be used to demonstrate for users the conceptual *potentiality* of the IR system – its author dependent *domain space* or state of knowledge – in a structured way, serving as a feedback feature and allowing for browsing. In a recent review P. Willett analyses the present state and discusses its effectiveness (1988).

Besides the capability to retrieve documents *similar* to a known (and relevant) document, the technique may be used to create term clusters. However, the most severe problems encountered in relation to the clustering techniques, which can be seen as a form of automatic classification, are the issues of determining the *linguistic basis* for term associations and the provision of formal definitions of association between a pair of terms, and association among a class of terms. As stated by Sparck Jones and Kay (1973, p 163–164):

Most experiments have tended to involve a great deal of grouping in the dark and theoretically unsatisfactory procedures being adopted in an ad-hoc manner on the basis of largely unjustified assumptions, and being inadequately tested.

The Figures 4.2 and 4.3 demonstrate these problems to a certain extent, as yet not solved. For instance, how do we automatically create adequate semantic connections between, say 'India' and 'Export', the left-hand side in Figure 4.3? By observing Figure 4.2 it is obvious that we require more context than simply provided by the surrounding terms in the texts. Following S.E. Robertson (1977, p. 126–127):

..the assumption underlying term-clustering experiments is that semantic connections between terms can be discovered by considering their co-occurrences. But attempts to incorporate such semantic connections into retrieval procedures have generally been disappointing. Is this because the relation between these semantic connections and some traditional retrieval operations is purely superficial, and the one cannot usefully be substituted for the other? Or is it because we lack a vital part of the overall theory, which would make this relation explicit and show us how to use it? The experiments do not tell us.

An important aspect lies in the fact that clustering techniques applied to the same collection with the same queries as the probability technique, result in like performance effectiveness, but provide slightly different output of ranked documents. In other words, they demonstrate a kind of *indexer inconsistency*, similar to two different indexers' interpretations of identical documents. The complementarity of the two IR techniques may hence be used to refine relevance in certain retrieval situations.

Feedback in the IR environment is understood slightly different from the notion of Wiener (1948, 1961). By feedback is meant either: *relevance feedback*, as developed by Salton (1968) in relation to the feature-based IR techniques, or: *system feedback*, as used by the author in relation to the Zoom feature and other means in an IR system that display author/indexer aboutness to a user (Ingwersen, 1984a, 1986; Ingwersen and Wormell, 1986). The cognitive impact of elaborated system feedback on retrieval processes is discussed in Chapter 7.3.

Salton's notion refers to the process where a user's request is modified automatically *in the system*. The modification or refinement is based on 'system feedback', e.g. displayed documents, graphic representations of concepts, ranked lists of terms, thesaurus structures, etc. Documents (or concepts) identified by the user to be relevant lead to adjustments of weights in relation to the query terms. The relevant document(s) term weights are used in a repetitive search, providing retrieval of documents similar to the document judged relevant by the user (Fuhr and Buckley, 1990). Relevance feedback may therefore be applied to all IR techniques, including clustering. An example of both types of feedback can be shown by looking at Figure 4.3; stage 1: the user enters the words 'India, Export, Denmark'; stage 2: the nearest neighbour clustering technique produces the network display as in the figure, i.e. as *system feedback*, focusing on the document node (4) as most relevant; stage 3: the user observes the network and points to Doc.(4) to see its title; stage 4: Doc. (4) 'Denmark-India, an export guide' is displayed (system feedback); stage 5: the user indicates its relevance; stage 6: this *relevance feedback* modifies the query to include the term 'guide', whereby the focus of the search shifts towards the right-hand side of Figure 4.3. The nearest neighbour documents to Doc. (2) will be displayed, e.g. linked by the terms 'Import' and 'Transport', and an eventual ranking order of documents will shift accordingly. Relevance feedback is thus an attempt to reach into the information space of the system, i.e. one of several methods which have been tried to overcome the Dark Matter problem in IR.

Both types of feedback are extremely important features in IR interaction, ensuring higher performance. Note however, that the system does not know whether to keep the original query, adding it to the modified one, or to exchange it completely for the modified one. That would require an intermediary mechanism. In association with probability techniques, relevance feedback is essential to the system in order to know about query term weights and the *pr* value.

## 4.5 Relevance measurement techniques

Since the sixties, a core issue of research in IR has been performance studies. The most common method applied within the context of the traditional approach is the

application of laboratory tests on databases designed for that purpose. Sparck Jones (ed.,1981) and Sparck Jones and van Rijsbergen (1976) outline the perspective of the methods and describe the test beds used.

The principle is scientific, i.e. that for each test collection a number of fixed queries are used, and the total number of 'objectively' relevant documents for each query is known to the researcher. Variables in the tests are either a specific IR technique, or a particular method of representation. Comparative studies of relative performance may hence be carried out. The test collections are small, from approximately 3,000 items to approximately 20,000 items, far from the size of large-scale operational systems which measure up to 9,000,000 document records. Relevance, defined as the measure or degree of a correspondence or utility existing between a text or document and a query or information requirement as determined by a person (van Rijsbergen, 1990), is normally measured in form of 'bibliographic relevance', i.e. that judgements of relevance are based on document representations, such as titles, abstracts, etc. 'Document relevance', i.e. judgements on full-text documents, may also be applied. Part-of-document relevance is not in use.

In the laboratory experiments human *users do not take part* in the experiments. This is the main draw-back. More recent user-oriented approaches to the design of IR systems naturally take into account the cognitive effects of a dynamic interaction and the feedback on the user's problem space, his request, and query structure. The fixed queries are then no longer constants, but become variables, with loss of knowledge of the total number of relevant texts as a consequence. The test collections and the comparative methods are consequently only operational in relation to the traditional approach.

The *standard criteria* for evaluation are *recall and precision*, and more recently *fallout* (Robertson, 1977). 'Recall' is defined as the number of relevant documents retrieved  $R$ , related to the total number of relevant documents in the collection  $C$ , i.e.  $R/C$ . Recall may therefore only be defined by exactly knowing  $C$ , normally impossible in operational systems and in real-life experiments, and thus resulting in a degree of uncertainty, in addition to the uncertainty inherent in IR itself.

'Precision' is the relationship between number of relevant retrieved documents  $R$  and number of retrieved documents  $L$ , i.e.  $R/L$ . 'Fallout' is often used as a replacement for precision, because it takes into account the total collection  $N$ . It involves the relationship between the number of non-relevant, retrieved documents, and all non-relevant documents:  $(L-R)/(N-C)$ . In general, an inverse relationship exists between 'recall' and 'precision'.

By standard performance measures it has been possible to compare the various techniques mentioned in this chapter. Partial match techniques all demonstrate significantly higher performance than exact match techniques, i.e. Boolean techniques. Probability is the major feature-based technique, in particular when incorporating the tf.idf weights. Following Belkin and Croft (1987, p. 127) the use of term dependencies to modify document rankings may also improve performance, but only if the dependencies are accurately identified by the user or NL processing (Croft, 1986). Thesaurus information automatically applied to expand queries is only really effective if the terms expanded and the type of thesaurus information used is

tightly controlled. Clustering techniques can achieve levels of performance similar to the feature-based IR techniques but tend to be better for high-precision results (Croft, 1980). For certain queries clustering works better.

#### 4.6 Summary statements

The traditional or classic IR approach has its limitations, as several researchers have stated since the end of the seventies, especially in relation to the issues concerning the user's problem space and its development into request and query formulations (Ingwersen, 1988, p. 153–154). In view of more recent R&D approaches and theoretical developments in IR, it is important, however, to emphasize the potential of the traditional approach, in particular serving as one of the basic instruments in knowledge-based IR environments. The theories underlying the techniques are at present often exported to other disciplines, such as classification theory to AI in relation to software reuse (Albrechtsen, 1990), and online IR techniques to office automation and work station research landscapes (McAlpine and Ingwersen, 1989; Croft et al., 1990; van Rijsbergen, 1986b).

It is evident, that for each method applied to text analysis, representation and IR technique, similar methods are used in relation to the *representation of requests*. This implies that indexing with vocabulary control or NLR requires queries, either consisting of identical or similar controlled terms structured properly, e.g. by Boolean logic, or consisting of similar single and independent terms with no or vague relationships between them. We may have control of the syntax in the query but not its meaning; not to speak of the *potential information* it carries with it, from the problem space of the user towards the system. One may easily uncover examples of request formulations which, notwithstanding their rather elaborated nature, demonstrate very different semantics, looked at from the side of the system.

It is symptomatic that all the various approaches within the classic IR research tradition take the *query or request expression for granted*. Relevance feedback in probability or clustering IR simply creates a new query, although rather complex, and moves the search spotlight from one place in the collection to another. Then, of course, even if the technique shows high performance test-results relatively speaking, some unknown but relevant texts fall into darkness. Exactly identical queries from two users may often result in totally distinct relevance judgements. The same query applied to different statistical or network based IR techniques produces overlapping, but not identical results. More poor retrieval techniques show high performance for certain type of queries. It is exactly in relation to these problems of IR theory that we are referred to by Sparck Jones' (1979) and van Rijsbergen's (1990) statements and arguments previously mentioned in this chapter. Essential issues are the deficiency of adequate conceptions of 'meaning' and 'information', the constantly inherent 'generator aboutness' versus application of 'user aboutness', and intentional usage of information in knowledge, not document, representation, as well as the

problems of increasing retrieval uncertainty.

All the methods and principles outlined in this chapter are *ad hoc theories* but linked to one another, originating from mathematics, linguistics or philosophy. However, they provide a definitive step towards a unifying theory of IR. The importance for progress in IR research is to have an idea of when to use the different principles and techniques, to have *precise understanding of parameters* for their appropriate application. Because of the intensive research under the umbrella of this mainstream approach, and given their premises, we possess a fair portion of knowledge concerning their advantages and disadvantages. But, unfortunately, we do not know which of the principles, or their combinations, that may suit the various kinds of IR situations.

From the outcome of the classic position we have only vague ideas about what may happen when users put their cognitive efforts into the game, actually applying the refined techniques in real IR interaction. This ought to be profoundly tested.

One may in addition anticipate a 'trap' easily fallen into, namely a tendency to apply linguistic theories that are better suited to research into automatic language translation. Unambiguous semantic analysis of elaborated request statements as well as of texts will supposedly be technically feasible in the near future. Such analysis techniques may demonstrate a high retrieval performance ratio – but of what? Of sentences carrying identical, or very similar semantics, to the semantics of the request, but not necessarily carrying *information*. In other words, the searcher may thus retrieve what he already knows from various texts, rather than what he does not know. Naturally, in certain verificative subject retrieval situations this mode of IR is evidently valid – but it might be achieved with less analytic and processing effort invested.

If we wish to ask the user about his desire for information we need a *platform of knowledge about users* to ask from and to relate answers to. This platform constitutes the intermediary mechanism. In this case we leave the traditional IR approach, accepting a more user-driven or cognitive one. Belkin and Croft's profound review on IR techniques (1987) is in itself a serious step towards the latter, providing an interesting introduction in the form of research questions to be answered in the near future.

Such attempts at synthesis, made out of results from the classic tradition and the more user-oriented research position, are mandatory if IR research seriously wishes to take hold at levels of information processing and knowledge transfer above the monadic and structural levels - as suggested by the cognitive point of view.

[ This page intentionally left blank ]