Libri 1995: vol. 45, pp. 160–177
Printed in Denmark. All rights reserved.

Libri
ISSN 0024-2667

An Introduction to Algorithmic and Cognitive Approaches for Information Retrieval

PETER INGWERSEN AND PETER WILLETT

This paper provides an over-view of two, complementary approaches to the design and implementation of information retrieval systems. The first approach focuses on the algorithms and data structures that are needed to maximise the effectiveness and the efficiency of the searches that can be carried out on text databases, while the second adopts a cognitive approach that focuses on the role of the user and of the knowledge sources involved in information retrieval. The paper argues for an holistic view of information retrieval that is capable of encompassing both of these approaches.

1. Introduction

The subject of information retrieval, or IR, involves the development of computer systems for the storage and retrieval of (predominantly) textual information. IR techniques were initially developed for the retrieval of references to documents from bibliographic databases, and the discussion that follows assumes this form of textual information. However, the techniques that have been developed for searching bibliographic databases are equally applicable to any sort of textual information, such as reports of meetings, legal contracts, newswire stories, film scripts, technical manuals and, increasingly over the last few years, multimedia information systems.

Interactive IR from bibliographic databases has now been available for some two decades, either *via* in-house systems or *via* dial-up to online hosts. While the number and the size of the databases have increased hugely over this period, the great majority of them have continued to employ the familiar Boolean retrieval model, in which the query terms are linked by the logical operators (AND, OR and NOT) and in which there is a range of supplementary pattern-matching facilities for truncation and proximity searching. Similar comments apply to many of the CD-ROM-based retrieval systems that have been introduced in the last few years.

The Boolean model is well understood, but has several inherent limitations that lessen its attractiveness for text searching (1–3). The first major problem is that it is difficult to formulate any but the simplest of queries using the Boolean operators without a fair degree of training; accordingly, trained intermediaries often have to carry out a search on behalf of the user who has the actual information need. Secondly, there is very little control over the size of the output produced by a particular query. Without a detailed knowledge of the contents of the file, the searcher will be unable to predict *a priori* how many records will satisfy the logical constraints of a given query. There may be several hundreds if the query has been phrased in very general terms, or there may be none at all if too detailed a query has been input; in both cases, the searcher will need to reformulate the query in some way and then to carry out a second search which, it is hoped, will retrieve a more useful

Peter Ingwersen, Department of Information Retrieval Theory, Royal School of Librarianship, Copenhagen S, Denmark.
Peter Willett, Department of Information Studies, University of Sheffield, Sheffield S10 2TN, United Kingdom.

number of records. A third problem is that Boolean retrieval results in a simple partition of the database into two discrete sub-sets, *viz* those records that match the query and those that do not. All of the retrieved records are thus presumed to be of equal usefulness to the searcher, and there is no mechanism by which they can be ranked in order of decreasing probability of relevance. Finally, there are no obvious means by which one can reflect the relative importance of different components of the query, since Boolean searching implicitly assumes that all of the terms have weights of either unity or zero, depending upon whether they happen to be present or absent in the query.

One should not overestimate the scale of these problems, since Boolean searching has provided an effective way of accessing machine-readable textual data for many years. That said, these characteristics of the Boolean model mean that non-specialists may find great difficulty in carrying out searches. There has thus been substantial interest in the development of alternative methods for text searching that are more appropriate for end-users: IR systems based on such methods are normally referred to as *best-match, nearest-neighbour, ranked-output, vector-processing* or *probabilistic* retrieval systems (3–5). The research that has been undertaken in this area focuses principally on the algorithms and data structures that are needed to maximise *retrieval effectiveness*, i.e., the ability of the system to retrieve documents from a database that are relevant to a user's query, whilst maintaining a reasonable level of *retrieval efficiency*, i.e., the ability of the system to carry out its functions with the minimal use of machine resources.

An algorithmic focus, whether Boolean or best-match, is not inappropriate if one considers the design of IR systems for trained professionals who can be expected to make themselves fully conversant with the particular systems that they need to use in their day-to-day business. Examples of such professionals are librarians, lawyers, online intermediaries and an increasing number of academic researchers. However, such a focus neglects many of the social and cognitive processes that are involved in IR, and these processes are likely to be of great significance if one is to design effective retrieval systems for inexperienced users, for whom database searching is of only minor importance. Specifically, the algorithmic approach has two principal limitations, as detailed below.

The first limitation is that no account is taken of the large body of studies that have been carried out on users' *information seeking behaviour* (i.e., on the formation, nature and properties of a user's *information need* (6–8); and the second limitation is that there is an almost-total lack of real-life investigations of the impact of the algorithmic techniques on users in socio-organisational contexts. These limitations have provided the driving force for a range of communicative and psycho-sociological studies of IR systems. The studies have been motivated by the belief that an understanding of user behaviour and user-system communication will permit the construction of *knowledge-based intermediary systems* that can support an individual's search for information in various ways, e.g., by identifying a suitable combination of retrieval techniques (9). Thus far, these studies have considered only large-scale Boolean systems but they have sufficed to show that the user's background knowledge of the information that is being sought can play a vital role in the retrieval process, as do the reasons for the information request and the subject domain. As a result, several models of *intermediary functionality* have been formulated and partially tested over the last few years (10, 11).

Research on user-centred approaches to IR led to the observation that individual information needs may be stable, but that they may also change during the course of an interaction with an IR system; moreover, these needs may be ill-defined owing to a lack of appropriate background knowledge. The research that has been carried out has also shown that it is necessary to *contextualise* the information need by means of supplementary information on intent, purpose and goals. Information seeking and the formation of the information need are hence assumed to be a *process of cognition* by the individual searcher, in which the retrieval system and the intermediary functionalities are the crucial components of the contextualisation process. An immediate consequence of this approach to information retrieval is that the wide range of representational and searching techniques now available are seen as complementary information structures of different nature and cognitive origin. This, in turn, leads to the notion of a *cognitive theory of information retrieval*, which signifies an attempt to globalise information retrieval by regarding all of its components as representing cognitive structures of varying degrees of com-

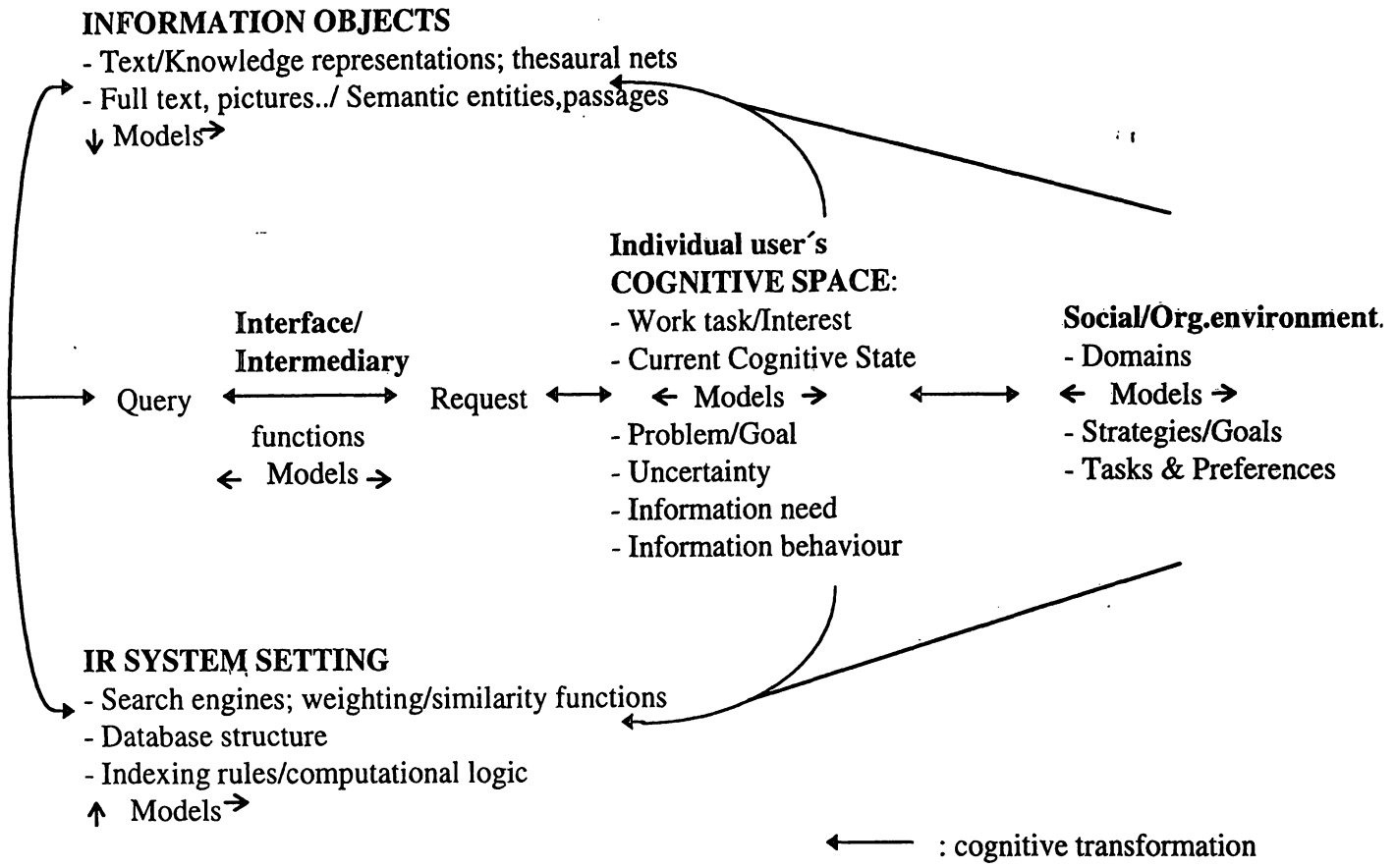

Fig. 2. Cognitive model of IR interaction. Extension of [3, p. 16].

plexity that co-operate in an interactive communication process (12).

The next two sections of this paper outline the algorithmic and cognitive approaches that we have introduced above, with the relationships between the various components of the two approaches being summarised by the diagram shown in Fig. 1. The paper concludes by noting some of the current research areas that may help further to define these two very different, but complementary, approaches to the design of IR systems. It is not possible, in a survey paper such as this, to provide detailed accounts of the algorithmic and cognitive approaches; however, the listed references should provide an entry point to the very large body of research that has been undertaken to date. More detailed accounts from the algorithmic viewpoint are provided by Belkin and Croft (1), Frakes and Baeza-Yates (4) and by Salton (5), while Ellis (13) and Ingwersen (14) provide comparable accounts from the cognitive viewpoint. Current research in both of these areas is reported in the proceedings of the *International Conference on Research and Development in Information Retrieval*, which is held annually under the auspices of the Specialist Group in Information Retrieval of the Association for Computing Machinery.

2. *Algorithmic approaches*

2.1. *Characteristics of best-match retrieval*

Best-match retrieval involves comparing the set of terms representing a query with the sets of terms corresponding to each of the documents in the database, calculating a measure of similarity between the query and each document based on the terms that they have in common, and then sorting the documents into order of decreasing similarity with the query. The output from the search is a ranked list, where the documents which the system judges to be most similar to the query are located at the top of the list and are thus displayed

first to the user. Accordingly, if an appropriate measure of similarity has been used, the first documents inspected will be those that have the greatest probability of being relevant to the query that has been submitted.

Retrieval systems based on ranking can help to alleviate many of the problems associated with Boolean searching, in that: there is no need to specify Boolean relationships between the terms in the query (since just an unstructured list of terms is required as the input); and the ranking of the database in response to the query allows complete control over the amount of output which needs to be inspected (since the user can browse down the list just as far as is needed). Both of these characteristics serve to increase the usability of systems based on the best-match model, as compared with Boolean systems. It is also normally very easy to take weighting information into account when calculating the degree of similarity between the query and the documents in the file: moreover, these weights may derive from user judgements of relevance for previously-inspected output, and there is hence an attractive mechanism available for the automatic incorporation of relevance information if a second search is required. These characteristics of the best-match model have meant that it has formed the basis for the great bulk of the research that has been carried out into increasing the effectiveness of automated text-retrieval systems. This research has considered effective algorithmic procedures not only for searching documents but also for indexing documents (and requests), as discussed later in this section.

Apart from the inherent limitations of Boolean searching that have been mentioned already, there are two further reasons for increasing the extent to which computers are used for document retrieval. The first, and most obvious of these, is on grounds of expense, since the plummeting costs of computing mean that an increasing number of IR processes are effected more cheaply by machine than by human means. Secondly, a very large number of studies over many years have shown that human processing can be very inconsistent, with the result that it is not as effective as one might assume (15). For example, indexing has traditionally been carried out by highly-trained people, who are knowledgeable about the subject matter of the database and who are familiar with the particular indexing techniques used within their organisation. However, while manual indexing can give excellent results, there is often little agreement between the sets of terms assigned when each of a number of different people index the same set of documents. Again, there have been several studies which show that online searchers differ considerably in the retrieval strategies that they use, even for a given topic on a given database, and in the ways in which they judge the relevance of the documents retrieved by such searches.

2.2. *Automatic indexing*

In view of the studies of human indexing that have been described above, it is not surprising that many people have suggested that *automatic indexing* techniques should be developed, in which the task of selecting content descriptors is carried out by automatic, rather than by manual, means (3, 15). That said, there is still much controversy as to the relative merits of manual and automatic indexing (16, 17), and it has accordingly been suggested that the best approach may involve making use of several different indexing methods simultaneously (9).

2.2.1. *Linguistic approaches*

Manual indexing is based on a syntactic and semantic analysis and interpretation of the texts of documents or requests, and there have been several attempts to apply linguistic techniques to the problem of selecting index terms. Early work, most notably that carried out by Salton and his co-workers on the SMART project, showed that stemmed keywords gave levels of retrieval performance that were comparable with, or superior to, those obtainable from manual application of controlled vocabularies or of phrase-based indexing (15). In part, the results of studies such as these arose from a concentration upon the analysis of document texts rather than request texts, and there is much evidence to support the idea that processing in IR should be request-oriented, rather than document-oriented. Of more importance is the fact that the linguistic techniques used were quite limited in scope, and it is only as a result of many years of research that the scale of the natural-language processing problem has become apparent. Indeed, most current natural-language systems function effectively only in very restricted subject domains,

e.g., as front-ends to database management systems, whereas IR systems may well contain documents on many different subjects.

The last few years have seen a resurgence of interest in the use of linguistic techniques for indexing purposes, e.g., for the automatic identification of noun phrases in continuous text (18, 19); however, none of the experimental results to date provide unequivocal support for the belief that the application of sophisticated linguistic processing to rich requests does indeed result in increases in system performance (when compared with the simpler statistical procedures discussed below) (20). Thus, the great bulk of automatic-indexing research has involved the use of statistically-based, rather than linguistically-based, techniques, and this is likely to remain the case, at least in the short term. In what follows, the techniques discussed should be understood as being equally applicable to the processing of documents and to the processing of natural-language requests.

2.2.2. Statistical approaches

The first extended studies of automatic indexing were carried out in the late Fifties by Luhn, who suggested that terms *describing the content* of a document could be obtained by selecting words from its constituent text (15). This method has the advantage that the indexing terms are derived from the author's own words, as manifested in either the full-text or the title and abstract, and Luhn's principal contribution was to suggest that terms could be identified using statistical information about the frequencies with which words occurred in a text (i.e., the "Information Objects" shown in Fig. 1). Luhn noted that a word which occurred very frequently in a database is unlikely to be able to discriminate sufficiently between relevant and non-relevant documents if it is specified in a request; a very infrequently-occurring word, conversely, is well able to discriminate but, by its very nature, is unlikely to be specified in a request. Thus, the most useful words for retrieval purposes are those with intermediate frequencies of occurrence. Terms representing the content of a document can hence be obtained by counting the *collection frequency* of each word, i.e., its frequency of occurrence in the database, and then by using as indexing terms those words in a document which have intermediate frequencies of occurrence.

The use of frequency data as a basis for keyword selection can clearly be extended to other types of statistical information. An example of this is the use of the *term frequency*, which is the frequency with which a particular word occurs within the text of an individual document. Taking the two approaches together, we may thus expect that the most important terms will be those that have a high term frequency but low-to-medium collection frequencies, i.e., terms whose occurrences are restricted to a relatively small number of documents. There are many variants of this basic idea, and many different frequency measures have been used to evaluate the worth of words as indexing terms. In fact, rather than applying sophisticated selection criteria, the tendency is increasingly to use *all* of the words from a document or request text, and then to differentiate them by means of an appropriate weighting scheme. Thus, the idea of *selecting* some of the words for best-match searching has been replaced by *extracting* all of them, with the sole exception of some of the very high frequency terms: these are eliminated by means of a *stopword list* containing some number, typically one or two hundred, of commonly-occurring words that are unlikely to be of use for retrieval purposes.

Stopwords are primarily function words, such as AND, OF, THE, FOR, etc., but they may also be words from phrases that tend to crop up in requests, e.g., ANYTHING ON or HAVE YOU GOT, and domain-specific words, e.g., INFORMATION or PROGRAM in a computing database. Thus, a request such as I WOULD LIKE DOCUMENTS ON EXPERT INTERMEDIARY SYSTEMS FOR ONLINE BIBLIOGRAPHIC SEARCHING might be processed by an automatic-indexing routine to yield the following (alphabetically-sorted) list of query terms (as represented on the left-hand side of Fig. 1): BIBLIOGRAPHIC EXPERT INTERMEDIARY ONLINE SEARCHING SYSTEMS. An analogous list of words is used to represent each of the documents in the database. Query statements are generally quite terse, and thus any given keyword is likely to occur only once. If a document abstract, or full text, is available, terms may occur several, or very many, times and the document representatives may hence contain not only each of the selected keywords but also their term frequencies.

It will be seen that the request representative above consists merely of individual words, with-

out any of the noun phrases which characterise much manual indexing, e.g., EXPERT INTERMEDIARY SYSTEM. Phrases are normally handled in manual systems by the use of a *thesaurus*, in which terms are noted as being synonyms or otherwise related to each other. The availability of a thesaurus should help to ensure a high level of *recall* (the fraction of the total relevant material in the database that is actually retrieved in a search), since it will allow term matches on related, rather than identical, terms in documents and queries. Thesaurus construction is an extremely time-consuming process, and there has hence been much interest in the use of term co-occurrence information for the automatic construction of thesauri; unfortunately, this work has not proved to be very successful in practice (21, 22). An alternative means of enhancing recall is by the use of conflation techniques as described below.

2.3. *Term conflation*

Once the set of words representing a request or document has been identified, some means must be found of overcoming the variants in word forms that are likely to be encountered in free-text databases. These variants arise from a range of causes including the requirements of grammar, valid alternative spellings, antonyms, and problems arising from mis-spelling, transliteration and abbreviation.

The problem of word variants can be alleviated, but not eliminated, by the use of a *conflation algorithm*, a computational procedure that reduces variants of a word to a single form. The rationale for such a procedure is that similar words generally have similar meanings and thus retrieval effectiveness may be increased if the query is expanded by including words that are similar in meaning to those originally contained within it. Term conflation is normally carried out in current online systems at search time using right-hand truncation as specified by the searcher, rather than by automatic means. However, considerable experience is needed if effective truncation is to be achieved since two major types of error are possible. Over-truncation occurs when too short a stem remains after truncation and may result in totally unrelated words being conflated to the same stem, as with both MEDICAL and MEDIA being retrieved by the stem MED*; under-truncation, conversely, arises if too short a suffix is removed and may result in related words being described by different stems, as with COMPUTERS being truncated to COMPUTER, rather than to COMPUT* (which would also include other related words such as COMPUTING and COMPUTATIONAL).

The most common conflation procedure is the use of a *stemming algorithm*, which reduces all words in English with the same root to a single form by stripping the root of its derivational and inflectional affixes; in most cases, only suffixes are stripped so that the algorithm performs a comparable role to that of manual, right-hand truncation (23). Many different types of stemming algorithm have been reported in the literature, but they nearly all make use of a dictionary of common word endings, such as -SES, -ING or -ATION. When a word is presented for stemming, the presence of these suffixes is searched for at the right-hand end of the word. If a suffix is found to be present, it is removed, subject to a range of constraints which forbid, e.g., the removal of -ABLE from TABLE or of -S from GAS. In addition, a range of checks may be invoked, e.g., to eliminate the doubling of terminal consonants that occurs when the present participle is used, as with FORGETTING and FORGET. Examples of stemmers are presented by Paice (24) and by Porter (25), *inter alia*. Evaluations of stemming algorithms suggest that they produce acceptable stems for about 95% of the words presented to them, although there is no unequivocal proof that searching using these stems is significantly superior to searching using the original, unstemmed words, at least in the case of English. Studies with languages having a greater degree of morphological complexity than English suggest that more significant improvements in performance can be obtained.

Word stemming is easy to implement and provides a highly effective means of conflating words with different suffixes. However, many other types of word variant are likely to occur in free-text databases, and there have been several attempts to provide conflation mechanisms for them. For example, an online dictionary of reversed words can be provided so that when truncation is carried out, words with different prefixes are conflated, thus providing facilities for *left-hand* truncation searching (rather than the ubiquitous right-hand truncation) (26). An alternative, and more general, approach involves the system calcu-

lating a measure of *string similarity* between a specified query term and each of the distinct terms in the database. Similar words can then be displayed at a terminal for inclusion in the query if the user so desires. This approach derives from work on automatic spelling correction (27) and several methods are available, including *Soundex codes*, *longest common-subsequence algorithms* and *n-gram coding* techniques. The last of these is probably the most common, and involves fragmenting a word into a sequence of *n*-grams, i.e., strings of *n* adjacent characters. The similarity between a pair of words is then estimated by the similarity between the corresponding sets of *n*-grams (28).

2.4. Matching of documents and queries

2.4.1. Similarity measures

We have noted above that best-match searching involves ranking the documents in a database in order of decreasing similarity with a query statement, this implying the calculation of some quantitative measure of similarity between the query and each of the documents (3). A similarity measure comprises two major components: the *term-weighting* scheme, which allocates numerical values to each of the index terms in a query or a document that reflect their relative importance; and the *similarity coefficient*, which uses these weights to calculate the overall degree of similarity between a query and a document (as denoted in the lower left-hand side of Fig. 1).

A commonly-used similarity coefficient is the *vector dot product*. Here, the similarity between a document and a query is calculated as the sum of the products of the weights of the terms that are common to a document and to a query. If some particular term is absent from the query or the document, then it will not make any contribution to the similarity coefficient; a non-zero contribution, conversely, is made to the overall similarity if that term is present in both the query and the document. Thus, if a query has weights of 0.5 and 1.0 assigned to two specific terms and if both of these terms are present, with weights of 0.3 and 0.7, respectively, in a document, then the similarity between the query and this particular document is given by (0.5×0.3)+(1.0×0.7), i.e., 0.85. In the case of unweighted (or *binary*) terms, i.e., where the weights are either 1.0 (for presence) or 0.0 (for absence), the dot product corresponds to just the number of terms in common between the query and the document.

Other similarity coefficients that have been used for IR are the *Dice coefficient* and the *cosine coefficient, inter alia*, both of which are normalised forms of the dot product that take account of the differing numbers of terms in different documents.

2.4.2. Initial weights

Weights can be assigned to the terms in documents or queries or both, but the extensive research that has been carried out suggests that query weighting is of greater importance in determining the effectiveness of a search; indeed, it is often the case that documents are characterised by binary, i.e., present or absent, weights. There are two main types of query-term weights, these being the *initial weights* and the *relevance weights*. The former are used when a searcher first puts a request to a best-match system (29), while the latter are used once the searcher has had a chance to inspect the output from the initial search (30).

Collection frequency, or *inverse document frequency* (IDF), weighting involves assigning weights to the terms in a query such that the weights are in inverse proportion to the frequency of occurrence of those terms in the database that is to be searched. The rationale for this approach is that people tend to express their information needs using rather broadly-defined, frequently-occurring terms and any more specific, i.e., low-frequency, terms are likely to be of particular importance in identifying relevant material. This is because the number of documents relevant to a query is likely to be fairly small, and thus any frequently occurring terms must necessarily occur in many irrelevant documents; infrequently-occurring terms, conversely, have a greater probability of occurring in relevant documents, and should thus be considered as being of greater potential importance when searching a database. These considerations lead to the use of a weight that is inversely proportional to a term's collection frequency. The IDF weight was originally suggested on purely empirical grounds, and extensive tests show that it consistently gives results that are superior for best-match searching to those resulting from the use of unweighted query terms. More recently, it has been shown that this weight is a limiting case of the probabilistic relevance

weights (described below) when no relevance information is available, thus providing a theoretical rationale for the use of IDF weighting (31, 32).

We have noted previously that document terms are often unweighted. If this is not the case and if weighting is to be used, then the most common approach involves *term-frequency weighting*, where the term frequency of a term represents its (possibly normalised) frequency of occurrence within the text of an individual document. Increasing use is being made in experimental systems of the so-called TF×IDF weight (29), which involves multiplying the term-frequency weight by the collection-frequency weight for each document-to-query match.

2.4.3. Relevance weights

When someone carries out a search, it is natural for them to modify the query in the light of previously inspected search output. *Relevance feedback* is the name given to a body of techniques that try to carry out such query modification by automatic, rather than by manual, means. The initial search is carried out as described previously, e.g., using IDF weights, and the user inspects a few top-ranking documents (perhaps 10 or 20 of them) to ascertain their relevance to the need. In the normal approach to relevance-feedback searching, the system uses these relevance judgements to calculate a new set of weights that should more accurately reflect the importance of each of the query terms. Alternatively, rather than just modifying the weights of the original query terms, the system can modify the query by the addition or deletion of terms.

It is reasonable to assume that a query term that occurs frequently in documents that are judged to be relevant to a particular query, and infrequently in documents that are judged to be non-relevant, is a "good" term, in some sense, for that query: such good query terms should thus be allocated greater weights than the other query terms. This intuitively-reasonable idea was put on a formal basis by Robertson and Sparck Jones, who were able to provide a theoretical rationale for the use of a particular term weight that is based on the occurrence of a particular term in the sets of relevant and non-relevant documents retrieved by the initial search (33). Thus, once the user has provided relevance judgements on the top-ranking documents from the initial search, the system uses the judgements to calculate new weights that should more correctly represent the relative importance of the various query terms. These new *relevance weights* are used in the second, feedback search, and the user can then provide a further set of relevance judgements on the new set of top-ranked documents. The feedback search can be iterated at will, until the user is satisfied with the output from the search.

The discussion thus far has focused on the use of relevance data to modify the weights of the terms in the original query. This data can also be used to suggest terms that should be added to, or removed from, the original query statement. Thus, it is possible to calculate relevance weights for all of the terms in the documents that have been judged relevant, and then to expand the query by the addition of some of the most highly-weighted terms that had not previously been included in the query. Again, if one of the original terms is found to occur primarily in non-relevant documents, it may be helpful to remove it from the query prior to the feedback search. There is still considerable discussion as to precisely how relevance-based query modification should be carried out to maximise search effectiveness.

2.4.4. Matching algorithms

The similarity coefficient and the weighting scheme play an important role in determining the extent to which a system is able to retrieve relevant documents. If such a system is to be used, it must also be sufficiently fast in operation to permit interactive searching of databases of non-trivial size. The obvious way to carry out a best-match search is to take the set of index terms comprising the query, and then to compare them in turn with the sets of index terms that characterise each of the documents in a database. This is clearly far too slow if a large database is to be searched using conventional computer hardware, and there is thus a need for efficient algorithms that can maximise the speed of these matching operations.

Efficient Boolean searching is possible because of the development of sophisticated inverted-file systems, with the Boolean logical operations being applied to the lists of document identifiers in the postings lists that correspond to the terms that have been specified in the query. Work over sever-

al years has demonstrated that the same posting-list information can also be used to facilitate the identification of the terms that are common to a query and a document in a best-match environment (although the actual computer processing is quite different). Once the matching terms have been identified, the resulting sets of similarity values are sorted into decreasing order to enable the presentation of the best-scoring documents to the searcher. However, the same basic data structures are used as for Boolean searching, thus enabling the implementation of best-match searching without excessive additional costs. A review of inverted-file algorithms for best-match searching is provided by Perry and Willett (34) while a specific implementation is described by Persin (35).

2.5. *Integration of Boolean and best-match retrieval*

The limitations of Boolean searching have been spelt out in the first section of this paper; however, it must be emphasised that best-match searching also has some disadvantages. Firstly, the absence of proximity operators and of the Boolean AND operator means that it is not possible to specify phrases in an explicit manner, and the absence of the Boolean OR operator means that it is similarly not possible to specify synonyms in an explicit manner. It may seem illogical to cite the absence of the Boolean operators as a weakness of best-match searching, since we have previously cited their presence as a weakness of Boolean searching. This apparent contradiction arises because their presence allows a trained searcher to specify a query in very precise detail, but might confuse a searcher who did not have sufficient experience to make full use of them. Attempts are being made to encompass proximity information in best-match searching but unequivocal solutions remain to be established (36). Secondly, a query needs to contain several terms if the matching algorithm is to be able to provide a discriminating ranking of the database. This arises since a query containing only a few terms will enable the best-match algorithm to divide the database into only a small number of groups (e.g., those documents that have none, one or two matches, respectively, with a query that contains just two terms) and thus not be able to provide a finely-honed ranking for the searcher's inspection. A simple, and obvious way around this problem is for the user to provide a known relevant document as a request; this will often be available and also avoids the user having to provide an explicit statement of need.

It will be clear that both Boolean and best-match searching have both strengths and weaknesses. It is thus hardly surprising that comparative studies suggest that there is little difference in the effectiveness of the two approaches (as measured using the traditional performance parameters of recall and precision), although there is often only a *small degree of overlap* in the identities of the relevant documents that are retrieved in the two types of search (so that the outputs of the two approaches *complement each other*). Where there is a substantial difference is in the efficiency of the search, i.e., the amount of effort that is required to obtain these relevant documents. The computational requirements are broadly comparable in the two cases, since they both involve processing of the inverted-file postings lists that correspond to the query terms (4, 34). However, there is a large difference in the amount of effort required of the searcher; specifically, the best-match model requires only that the user is able to identify and to type in the query terms, and to judge the relevance of the documents that are presented, with all of the subsequent processing being carried out by the machine.

The complementary natures of Boolean and best-match searching has led to interest in search methods that combine features of both. Methods that have been suggested include: ranking the output of an initial, recall-oriented Boolean search; use of the *p*-norm Boolean algorithms that were developed by Salton et al. at Cornell University during the early Eighties (37); and intermediate *front-end systems* - as marked in the centre of Fig. 1 – that allow best-match searching even from databases that are designed to support only Boolean retrieval (e.g., the elegant work carried out by Robertson's group at City University (38).

It seems reasonable to suppose that, for the foreseeable future at least, a trained searcher will be able to achieve better results than an inexperienced searcher: the value of the best-match approach is that it enables the latter to carry out at least some sort of search, whereas this might not be possible if only Boolean facilities are available. The current rapid growth in end-user text searching has given a substantial stimulus to the provision of best-match facilities, as is exemplified by recent product releas-

es from commercial database hosts such as West Publishing (39) and DIALOG (40). This market push is likely to lead to a significant increase in our knowledge of the comparative effectiveness of Boolean and of best-match searching.

3. User-oriented and cognitive approaches

The user-centred approach to IR is principally based on cognitive psychology and social science methods. The approach has provided substantial insights into users' mental behaviour and in their information seeking characteristics, both on an individual basis and in social and/or organisational contexts (as shown in the centre and the right-hand side of Fig. 1). It has also supplied a fair amount of information about inter-human information interactions, such as the interaction between a librarian or information specialist and a user. Finally, the role of the (human) intermediary has been defined in relation to User and Request Model building by means of search interviewing and feedback from IR systems. However, just as the traditional algorithmic approach disregards the dynamic role of the user, so the user-oriented tradition does not encompass the full range of IR system factors.

Until the mid-eighties, no investigations had taken place that involved non-Boolean retrieval and different methods of representation as well as intermediaries and users (16). This "monolithic" situation seems understandable, since without established models of searcher (users and intermediaries) behaviour, such advanced experiments could not yield results that were valid for design and test purposes or for the development of IR theories. Such models are now becoming available, as we discuss below.

3.1. The nature of the information need

Real-life investigations have provided an understanding of the formation of the information need with respect to users' "pre-information-searching" behaviour. This all-important dimension of IR was the starting point for the ASK (Anomalous State of Knowledge) hypothesis of Belkin et al., which dates from 1978–82 and which provided a fundamental step towards a detailed understanding of the nature of the "desire for information" (41). The ASK hypothesis is based on Taylor's earlier theories regarding the development of the information need (42) and other scholars' attempts to analyse this process (43). The information need is regarded as an "incompleteness" or a "gap" in the current knowledge or understanding of the actual user, which results in a reduction in the effectiveness of the user's interaction with the world around him. Taylor suggests a four-stage process in which the final stage, called the Compromised Need, essentially mirrors the *request formulation* to (the librarian and) the system. The compromise is assumed to be caused by the user's expectations and intent balanced against the intrinsic level of verbalisation of the need. Later investigations by Belkin (44), Ingwersen (8) and others have shown that this compromise actually takes place, often in the form of *labels* (the Label Effect), with the result that request formulations *do not* necessarily exactly mirror internal needs. An additional problem is, of course, that the information need mirrors "what is known about the unknown". The ASK hypothesis takes this into account by pointing to a *problem or goal-dependent situation* which underlies the need and which is assumed to be the reason for the development of this need (as shown in the centre of Fig. 1). The Monstrat Model for interface design by Belkin et al. (10) hence relies heavily on the functional modelling of this underlying problem situation. However, these investigations have also demonstrated that the underlying problems may be very poorly defined. From both a cognitive and a sociological point of view, the formation of the information need, as well as of the underlying problem or goal throughout the search and retrieval session, can thus be seen as an individual *process of cognition*, involving the current user's emotional and cognitive states in a social and communicative context.

Information needs may be categorised into three basic forms: *Verificative* needs; *Conscious topical* needs; and intrinsically *Muddled or ill-defined* needs (14). Needs in the first category are those where the user wishes to pursue (verify) data items from known properties of the items, e.g., author names, title words, publication year, the source, etc. The second type of information need implies that the topic is intrinsically well-defined, e.g., the user has searched it previously or already has some level of understanding of it. One may thus expect well-defined request formulations (or

relevance assessments) at a certain point of the IR interaction for such a need. The third category represents those cases in which information needs internal to the user are vague, muddled or ill-defined, i.e., those where users wish to obtain new knowledge and concepts in domains with which they are unfamiliar. The requests in such cases will obviously be poorly defined, until at least some degree of interaction has taken place, and the underlying problem situation may not be well-understood by the user.

The label effect that we mentioned previously is present in all of the three different types of information need, which implies that the three types may appear similar at the start of a retrieval session from the points of view of both the intermediary and the IR system. Aside from this dimension of levels of intrinsic understanding of the need, an additional dimension of *variability* can be observed throughout an IR session (9,12). In view of these facts, it is unfortunate, at the very least, that the algorithmic approach has generally assumed well-defined and static intrinsic information needs.

Users' information-seeking behaviour thus seems to depend strongly on their background knowledge, the subject or work domain in question, and the extent to which their need, or underlying problem, is developed. This behaviour carries a high degree of *uncertainty* with respect to interpretation and meaning associated with the need. In fact, there is evidence to suggest that the degree of uncertainty may actually increase during the initial part of IR (7). Also, certain social factors play important roles, in particular regarding the nature of the work domain or interest space. For instance, investigations of information-seeking behaviour clearly demonstrate that users in certain domains (such as humanities and social science disciplines or in fiction retrieval) tend to prefer to start by means of Verificative searching (using a known text) followed by backward and forward chaining, i.e., to apply similarity searching (13).

In recent years, research both inside and outside IR, e.g., in cognitive engineering, has suggested that the *actual work task* related to a domain, or the fulfilment of socio-cultural/emotional goals related to an interest, may be the principal causes for the user to be in a problematic situation and to need information (14, 45). Investigations of work-task complexity and information requirements demonstrate that the nature of the latter changes according to the degree of task complexity: the more complex the task, the more general (or vague) the need (46). Similar findings have resulted from research on Executive Information Systems. Only the underlying task may be definable. In the IR context, it hence seems appropriate to study the actual domain, the current work task or interest and the possibly related problem space, in addition to the associated information need. This way of dealing with the nature of the information need, by going below the surface towards the underlying *intentionality* (47) for such needs, signifies a *contextualisation* of the information need. This cognitive approach has two complications: an intermediary mechanism is mandatory, ideally based on domain and user models, and including *Request Model Building* (RQMB), in order to extract these underlying reasons from the user; and the *usefulness* or utility of the IR outcome is of greater importance than traditional (topical) relevance assessments. This dichotomy between a situational type of relevance and a topical one has recently led to a renewed interest in methods for measuring retrieval outcomes (48,49).

3.2. Information retrieval interaction – the cognitive turn

The complex nature of the information need makes it obvious that research on IR techniques alone cannot provide a complete understanding of the entire process of retrieving information. This process must be seen in its totality by incorporating the *system characteristics*, including the representational and retrieval techniques that characterise algorithmic approaches, with *the user's situational characteristics* and the necessary *intermediary functionalities* (see Fig. 1). In IR interaction, the intermediary (or user-interface) is the principal mechanism linking the system and the user. The Monstrat Model (10), which has been mentioned previously, was an attempt to functionalise the human intermediary behaviour, but was mainly directed towards the user. The later, Mediator Model (14) suggests that equal importance should be given to both the user and the system.

In this framework one may observe two "schools": *intelligent IR* tries to *simulate* the human behaviour of mediation by means of extensive user model building, RQMB and computational

inference techniques; and the *supportive approach to IR* tries to *stimulate* a user's mental processes during IR by means of tailored conceptual feedback from the system driven by the underlying domain model and RQMB. Both approaches provide a common platform for researchers in IR and AI. Since IR is too broad an environment for expert-system-like AI solutions, current efforts are concentrating on finding an appropriate balance between model building, inference, and user support of conceptual nature.

3.2.1. Intermediary functionalities

Human-human interaction in IR situations, i.e., "information searching" behaviour, can be divided into a *pre-search interview* stage, followed by searching activities (50). The pre-search interview serves as an exploratory dialogue in which the intermediary (such as a librarian) tries to understand the user request and the underlying background, the level of expertise, etc., so that the subsequent retrieval activity may be effective. Such detailed and systematic interviews are common in real life, e.g., when using a modern-day online system, and are often assumed in much of the research that has been carried out. However, this is primarily because of the heavy financial costs that are currently associated with the use of such systems and it may thus not be entirely appropriate to simulate this behaviour precisely. Other studies (6, 8, 51), in which search economics have no impact on the investigations, have demonstrated an *heuristic mode* of retrieval. This is characterised by *simultaneous* interviewing, searching and systems feedback. In both modes of intermediary functionality, at least a pre-defined Domain Model and/or a stereotypical User Model is necessary, from which to structure either interviewing and inference, or the supportive means provided for the user's own inference.

The heuristic mode of IR is closely associated with the supportive approach to retrieval systems design. A famous example is the *Book-house* system for the retrieval of fiction, which was developed by A. Mark Pejtersen and which is based on extensive domain and user behaviour analyses (52). A user of the Book-house carries out a search by means of suitable metaphors, content and action icons as well as transparent search strategies, all originating from the pre-established cognitive task analyses of fiction retrieval. This may involve browsing *via* similarity searching, retrieval *via* genre, author intent, or other dimensions characteristic of fiction searching, e.g., searching by front-page colour or figures. The Book-house has been extensively tested in real-life environments.

3.2.2. User and request model building

Typical pre-defined Domain and User Models are demonstrated in the recent Mediator Model, which is due to Ingwersen (14). In addition to the RQMB functionality Mediator stresses the minimal application of user model building. This latter type of model building is assumed to encompass only two dimensions of expertise by extraction from the current user: *conceptual expertise* relating to the actual topic or domain; and the user's *current retrieval competence*. The construction of more elaborate user models (based, e.g., on general knowledge, education, age, etc.) seems to be of use only in very narrow and consistent domains. Depending on the answers that are received to the questions about the user's current expertise, the level of support and mode of man-machine dialogue may be determined and adopted by the system; an example of this approach is provided by the I^3R system described by Croft and Thomson (53). Depending on what is known of seeking characteristics from the mandatory domain analysis, information on expertise may also be used to infer how the available IR techniques should be applied, e.g., it might be decided to use all of the available techniques in the system for immediate and specific retrieval in the case of a user who has extensive domain knowledge. This concurrent application of several algorithmic techniques leads to various kinds of data (dif)fusion, of retrieval overlaps and of types of relevance feedback during IR.

The RQMB stage involves elaborating the cognitive characteristics of the request, i.e., the aforementioned underlying problem situation, work task, and domain or interest. Search preferences form part of this model-building functionality. RQMB is meant to provide the system with additional structured contexts, and is not simply concerned with the request formulation itself. The assumption is made that *several simultaneous representations* of the same personal cognitive space may yield improved retrieval results and feedback

for further modifications of request, problem or task. The advantage is that the intermediary mechanism is free to perform a kind of *cognitive fusion* of the representations, or to make separate use of each individual representative structure.

3.3. Cognitive IR Theory – information in context

The user-oriented research that has been carried out thus far gives rise to two significant questions.

The first question relates to the degree to which an IR system and IR interaction ought to be designed not only to accommodate individual users in defining their need for information and resolving it, but also to define and then to solve their underlying problems. All such activities are actually found to occur during IR. Aside from the retrieval of information itself, it seems evident that IR should accommodate both the problem and the clarification of the information need, since both processes are fundamental for successful retrieval. However, IR is not the main objective in a decision or problem-solving activity. Although decisions are constantly made during IR interaction, and users may indeed often solve their underlying problem through IR, IR must be considered as a vital but *supportive* process in problem-solving and decision-making. That said, problem-solving is not the only reason for IR. The problem space or problematic situation in the mind of the user that gives rise to IR needs to be considered in a broader manner. Associated with these questions is the fact that investigations of information-seeking behaviour have consistently demonstrated that at least four different types of *uncertainties* are present at the start of, and during, IR: to express the need for information; to retrieve information entities of potential value from the information space; to interpret the conceptual outcome of retrieval, i.e., texts and other feedback, that is, to obtain information; to understand the retrieval process itself and the structure of the information space that is being searched.

The second question is concerned with how deeply we need to understand what the user really means. Logically, this is only theoretically possible if the intermediary has found that the current user's information need belongs to the Verificative or Conscious topical types that we have introduced previously. If the need is vague or intrinsically ill-defined such an understanding is impossible, regardless of the questions that are posed to the user. This gives rise to the paradoxical situation in which if the user actually possesses a great deal of knowledge about his need for information, he is more capable of assessing the usefulness of the retrieval outcome intelligently than the system is capable of estimating the meaning of the actual need (since the system's background knowledge is severely limited and insufficient). Thus, the conclusion is that IR systems may, at most, provide some support at a structural linguistic level to the user's associative and intuitive thinking processes, which are at a cognitive and pragmatic linguistic level.

The development of a cognitive theory for IR is an attempt to understand these uncertainty situations and paradoxes in an *holistic* manner, and to propose a framework for workable solutions. A conceivable way to achieve such a framework would be to make simultaneous use of the *variety* of information structures which are to be found associated with the Information Objects, the System Setting, and the Cognitive Space of users (as described previously). The basic assumption is that, by supplying structures of suitable contextual nature to all three retrieval components during interaction, uncertainty can be reduced and improved support can be provided for heuristic searching.

Cognitive IR models suggest that we should view IR interactions as the interactions of various types of cognitive structures, as demonstrated in Fig. 1. Cognitive structures are generally understood as manifestations of human cognition, reflection or ideas (12). In IR they take the form of transformations generated by a variety of human actors, i.e., belonging to a variety of different *intentionalities* and *cognitive origins*. These include systems designers and producers, IR technique developers, indexing rule constructors, indexers, authors of texts and images, intermediary mechanism designers, and users in a domain-related societal or organisational context. In the System Setting an IR system designers' cognitive structures may be represented by specific database architectures and one or several matching algorithms or logics. Human indexers' cognitive structures are represented by the index terms added to the original Information Objects. These terms are essentially the result of an intellectual interpretation of an author's text or images, and their assignment is

often guided by pre-defined rules and a thesaurus containing semantic relations and knowledge representations that have been developed by other people. Similar problems arise in automatic indexing, where any different weighting function or similarity measure can also be regarded as a form of transformed cognitive structure. Authors' texts, which include titles, captions, headings, or cited works, are representations of cognitive structures that are intended to be communicated. Later citations pointing to that particular text imply different kinds of interpretations, each carrying its own cognitive background and intentionality. Specific portions of the texts, e.g., titles, abstracts, figures, the introduction, or the full-text sections demonstrate *different functional styles*. Each type of document exhibits an analogous set of differences, as does each domain, and should thus be treated differently.

Further cognitive structures are involved in the manipulation of user requests into query formulations during RQMB and retrieval by an intermediary (whether human or computerised), as shown in the centre of Fig. 1 where the cognitive structures might, for example, be they those of the Monstrat or Mediator models. The right-hand side of the figure summarises the major different cognitive structures of individual users. It is these structures that are identified by an intermediary mechanism and (re)presented to an IR system, i.e., these are the actual work tasks or interests that lead to the current cognitive state and that may be included in the final problem or uncertainty state for the actual user. These mental activities take place in the context of epistemological, social or organisational domains that not only influence the current searcher in a "historical" socio-semantic sense but also maintain a continuous influence on the authors of texts and on attitudes to system design. The simplest type of a domain is an academic subject field, which is essentially a *social construct* represented by the collective cognitive structures of the individuals forming that field. Other types of domain include industrial sectors, individual firms and organisations, or professional groupings, such as journalists.

One consequence of the cognitive modelling of IR interaction is the demonstration of its fundamentally *polyrepresentative* nature, in particular in relation to full-text IR (9); another is the recognition of the futility of performance competition between the different algorithmic and logical approaches to retrieval. This can be overcome by replacing it with investigations of their exact characteristics when interacting with the cognitive space of users *and* the types of Information Objects. Accordingly, we must consider how best to fit together such representations and structures during IR.

One recent step forward has been the introduction of *passage retrieval* in full-text systems (54). Another step has been to allow for *manual query modification* during experimentation, e.g., in the ongoing large-scale TREC (55) and OKAPI (56) experiments. Manual query modification is necessary for two reasons: Firstly, the feedback from the system provides the basis for relevance and utility judgements of text portions, e.g., by means of *marking up the relevant portions* of an information object like passages of text. Secondly, it also provides the basis for *improved cognition* by the user of his actual need for information, and, possibly, of his underlying problem or goal, by forcing him to interpret the search outcome. This outcome does not have to be monolithic, i.e., one simple ranked list, but may also contain pointers to *several* conceivable routes into information space, e.g., by hypertext links, condensed or structured lists of concepts, and analogous means of conceptual feedback. Any modification of the request, or problem or work task statements by selections and/or deletions of concepts then mirrors the altered intrinsic formations and conceptions of the need, problem or task. In this prospective framework the well-known issue of *inter-indexer inconsistency*, for instance, then becomes an asset, rather than a problem. Similar inconsistencies have been outlined in Section 2 in relation to Boolean and algorithmic retrieval techniques. This holistic approach is in line with the berry-picking seeking behaviour, which has been analysed and described by Bates in an otherwise purely user-centred approach to IR (57).

In fact, a cognitive theory would favour *all* kinds of inconsistencies and, in particular, the *retrieval overlaps* between the variety of different cognitive structures involved in IR. The assumption is that the more remote in cognitive origin, logic, functionality, and in time, the smaller the overlap; and the better and probably more relevant the retrieval outcome (12). The conceptions of cognitive retrieval overlaps as well as of data

and request *fusion* and *diffusion* are thus essential elements of a theory framed by the cognitive perspective.

4. Conclusions

In this paper, we have tried to provide an overview of the research that has been carried out over the last two decades into the development of IR systems that will be easier to use and more effective than the long-established Boolean techniques that underlie most current systems. The algorithmic methods that were discussed in Section 2 have been available within in-house retrieval systems for some years, in systems such as Personal Librarian (58), STATUS/IQ (59) or TOPIC (60), and they have also recently been introduced into public retrieval systems such as Dow Quest (61), TARGET (40) and WIN (39).

The discussion on Section 2 has considered only those algorithmic techniques that have been demonstrated to be of general applicability and that are already available, or are becoming available, in operational best-match systems. There are many other techniques that are under active investigation, including automatic document clustering (62, 63), searching algorithms for very large full-text files such as encyclopaedias (54) and the application of best-match searching to hypertext systems (64, 65), *inter alia*. Moreover, the continuing development of information technology means that new techniques are constantly becoming available that may have applications for IR. Examples include the availability of multiprocessor operating systems such as PVM that enable networks of PCs or workstations to be used for parallel processing (thus offering the prospect of rapid, low-cost searches of even the largest text databases) and the emergence of new algorithmic paradigms such as neural networks and genetic algorithms (both of which embody machine-learning capabilities that may make them appropriate for the implementation of relevance feedback).

It is thus clear that we shall continue to see developments in the algorithmic approach to IR, but it is most unlikely that these developments will, in themselves, suffice to enable effective searching of the increasing amounts of text that are becoming available in machine-readable form. Efforts are thus being made to embed these new best-match techniques in a holistic model that takes account of all of the factors that are of importance in the retrieval process. This model, which is illustrated in Fig. 1, has been discussed in Section 3 from a relatively user-centred approach as well as from the prospective global cognitive view. Human intermediary characteristics are well-understood, principally in relation to user and request model building, and other user-oriented primary functions. The various types of knowledge which are necessary for implementation in (or teaching to) intermediaries has been successfully specified and modelled. Knowledge of human search strategies and tactics, of seeking behaviour, and of conceptual domains and types of information needs, is also fairly well established. For request model building, it seems sufficient to separate the user's underlying work task and problem from the resulting information need, thus allowing or driving the user to elaborate on these cognitive representations during the IR interaction. From a global cognitive approach the mutual advantage for the intermediary (mechanism) is that it is free to perform a cognitive fusion of these representative structures, or to make separate use of each in relation to the system. The holistic view suggests the simultaneous use of different algorithmic techniques and modes of indexing in information space: this may lead to various kinds of data (dif)fusion, retrieval overlaps, and relevance-feedback possibilities which may support interpretation and elaboration activities in the mental space of the user.

New questions and problems emerge in all rapidly developing sciences. Thus far, the requests in *all* laboratory experiments have been pre-defined sets of *simulated* well-defined and static information needs, in order to define the exact recall (or topical relevance) ratio for each request. The re-introduction of users into the non-Boolean experimental settings brings several profound methodological issues into play:

1.) The users, of course, interpret the simulated requests differently during both initial query formulations and later query modifications (55); that is, they become an uncontrollable variable in an otherwise invariable environment. Problem: how to deal experimentally with the simple conception of topical relevance assessment measures? Any fixed pre-established measure is in principle unreliable in a statistical sense if

not carried out by a panel. We may expect inter-panel inconsistencies analogous to those observed previously in studies of inter-indexer consistency.

2.) Issues of relevance and evaluation methodology across different systems become still more controversial in experiments in which users pose real-life requests, which are likely to be more variable and more ill-defined in nature than the simulated ones that have been used previously. Problem: only post assessments can be performed, and the statistical population of users, requests and panel participants has to be large.

3.) It is likely that the concept of relevance will need to encompass "relative" as well as "partial", i.e., non-binary, assessments, differentiated into situational "usefulness" or "utility", and "topicality". Problem: how can we manage the range of variables introduced by real-life experimentation without introducing a whole range of sociological methodologies?

Other substantive problems that must be faced include issues relating to the definition of appropriate combinations of retrieval logics or algorithms for handling full-text and multimedia information objects and the range of intrinsic information needs associated with such objects? Finally: what is the role of data fusion (66) in this landscape, which fusion techniques should be used, and how can these encompass cognitive retrieval overlaps?

In conclusion, we note that the two approaches discussed here have very different origins. The algorithmic approaches derive, in large part, from quantitative disciplines such as classification theory, natural language processing, pattern recognition and probability theory, whereas the cognitive approaches derive, in large part, from more qualitative disciplines such as epistemology, organisation theory, socio-linguistics and cognitive science. This has inevitably resulted in substantial differences in the experimental methodologies that are used in the two areas and, consequently, in research in the one often being conducted with little account being taken of developments in the other. This situation is starting to change, as is exemplified by much recent work on the design of user-friendly interfaces (67) and by the publication of monographs that take full account of the traditions associated with the two approaches (14, 68).

We believe that the algorithmic and the user-centred approaches are complementary in nature, in that algorithmic techniques are necessary if one wishes to search a computerised database and that cognitive techniques are necessary if one wishes to take full account of the contexts in which the searchers operate and the retrieved information is to be used. Accordingly, we hope that this review will help to bridge the divide that still exists between these two approaches to the continuing problem of retrieving information from textual databases.

References

1. Belkin NJ, Croft WB. "Retrieval Techniques". Annual Review of Information Science and Technology, 1987; 22: 109–145.
2. Cooper WS. "Getting Beyond Boole". Information Processing and Management, 1988; 24: 243–248.
3. Willett P. ed. Document Retrieval Systems. London: Taylor Graham, 1988.
4. Frakes WB, Baeza-Yates R. eds. Information Retrieval: Data Structures and Algorithms. Englewood Cliffs NJ, Prentice-Hall, 1992.
5. Salton G. Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer Reading, Mass.: Addison-Wesley, 1989.
6. Su L. "The Relevance of Recall and Precision in User Evaluation". Journal of the American Society for Information Science, 1994; 45: 207–217.
7. Kuhlthau C. Seeking Meaning. New York: Ablex Publ., 1993.
8. Ingwersen P. "Search Procedures in the Library Analysed from the Cognitive Point of View". Journal of Documentation, 1982; 38: 165–191.
9. Ingwersen P. "Polyrepresentation of Information Needs and Semantic Entities: Elements of a Cognitive Theory for Information Retrieval Interaction". In Croft WB and van Rijsbergen CJ eds. SIGIR '94. Proceedings of the Seventeenth Annual International Conference on Research and Development in Information Retrieval Berlin: Springer Verlag, 1994; 101–110.
10. Belkin NJ, Brooks H, Daniels PJ. "Knowledge Elicitation Using Discourse Analysis". International Journal of Man-Machine Studies, 1987; 27: 127–144.
11. Fox EA, Hix D, Nowell L, Bruneni D, Wake W, Heath L et al. "Users, User Interfaces, and Objects: Envision, a Digital Library". Journal of the American Society for Information Science, 1993; 44: 480–491.
12. Ingwersen P. "Cognitive Perspectives of Information Retrieval Interaction: Elements of a Cognitive

IR Theory". Journal of Documentation, 1996; 52(1); (in press).
13. Ellis D. "A Behavioural Approach to Information Retrieval Systems Design". Journal of Documentation, 1989; 45: 171–212.
14. Ingwersen P. Information Retrieval Interaction London Taylor Graham: 1992.
15. Salton G. Dynamic Information and Library Processing Englewood Cliffs: Prentice Hall, 1975.
16. Blair DC and Maron ME. "An Evaluation of Retrieval Effectiveness for a Full-Text Document Retrieval System". Communications of the ACM, 1985; 28: 280–299.
17. Salton G. "Another Look at Automatic Text-Retrieval Systems". Communications of the ACM, 1986; 29: 648–656.
18. Fagan JL. "The Effectiveness of a Nonsyntactic Approach to Automatic Phrase Indexing for Document Retrieval". Journal of the American Society for Information Science, 1989; 40: 115–132.
19. Sparck Jones K., Tait JI. "Automatic Search Term Variant Generation". Journal of Documentation, 1984; 40: 50–66.
20. Smeaton AF. "Progress in the Application of Natural Language Processing to Information Retrieval Tasks". Computer Journal, 1992; 35: 268–278.
21. Peat HJ and Willett P. "The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems". Journal of the American Society for Information Science, 1991; 42: 378–383.
22. Smeaton AF, van Rijsbergen CJ. "The Retrieval Effects of Query Expansion on a Feedback Document Retrieval System". Computer Journal, 1983; 26: 239–246.
23. Lennon M, Pierce DS, Tarry BD, Willett P. "An Evaluation of some Conflation Algorithms for Information Retrieval". Journal of Information Science, 1981; 3: 177–183.
24. Paice CD. "Another Stemmer". ACM SIGIR Forum, 1990; 24(3): 56–61.
25. Porter MF. "An Algorithm for Suffix Stripping". Program, 1980; 14: 130–137.
26. Bratley P. and Choueka Y. "Processing Truncated Terms in Document Retrieval Systems". Information Processing and Management, 1982; 18: 257–266.
27. Kukich K. "Techniques for Automatically Correcting Words in Text". ACM Computing Surveys, 1992; 24: 377–439.
28. Freund GE, Willett P. "Online Identification of Word Variants and Arbitrary Truncation Searching Using a String Similarity Measure". Information Technology: Research and Development, 1982; 1: 177–187.
29. Salton G, Buckley C. "Term Weighting Approaches in Automatic Text Retrieval". Information Processing and Management, 1988; 24: 513–523.
30. Salton G, Buckley C. "Improving Retrieval Performance by Relevance Feedback". Journal of the American Society for Information Science, 1990; 41: 288–297.
31. Croft WB, Harper DJ. "Using Probabilistic Models of Document Retrieval Without Relevance Information". Journal of Documentation, 1979; 35: 285–295.
32. Robertson SE. "On Relevance Weight Estimation and Query Expansion". Journal of Documentation, 1986; 42: 182–188.
33. Robertson SE, Sparck Jones K. "Relevance Weighting of Search Terms". Journal of the American Society for Information Science, 1976; 27: 129–146.
34. Perry SA, Willett P. "A Review of the Use of Inverted Files for Best Match Searching in Information Retrieval Systems". Journal of Information Science, 1983; 6: 59–66.
35. Persin M. "Document Filtering for Fast Ranking". In Croft WB and van Rijsbergen CJ (eds): SIGIR '94. Proceedings of the Seventeenth Annual International Conference on Research and Development in Information Retrieval Berlin; Springer Verlag, 1994; 339–348.
36. Keen EM. "The Use of Term Position Devices in Ranked Output Experiments". Journal of Documentation, 1991; 47: 1–22.
37. Salton G, Fox EA, Wu H. "Extended Boolean Information Retrieval". Communications of the ACM, 1983; 26: 1022–1036.
38. Bovey JD, Robertson SE. "An Algorithm for Weighted Searching on a Boolean System". Information Technology: Research and Development, 1984; 3: 84–87.
39. Pritchard-Schoch T. "Natural Language Comes of Age". Online, 1993; 17(3): 33–43.
40. Tenopir C, Cahn P. "TARGET and Freestyle. DIALOG and Mead Join the Relevance Ranks". Online, 1994; 18(3): 31–47.
41. Belkin NJ, Oddy R, Brooks H. "ASK for Information Retrieval". Journal of Documentation, 1982; 38: 61–71 and 145–164.
42. Taylor RS. "Question Negotiation and Information Seeking in Libraries". College and Research Libraries, 1968; 29: 178–194.
43. Mackay DM. "What Makes the Question?" The Listener, 1960; 63: May 5, 789–790.
44. Belkin NJ. "Cognitive Models and Information Transfer". Social Science Information Studies, 1984; 4: 111–129.
45. Rasmussen J, Pejtersen AM, Goodstein LP. Cognitive Engineering: Concepts and Applications New York: Wiley and Sons, 1992.
46. Byström K, Järvelin K. "Task Complexity Affects Information Seeking and Use". Information Processing and Management, 1995; 31: 191–214.
47. Searle JR. "Intentionality and its Place in Nature". Synthese, 1984; 61: 3–16.

48. Schamber L, Eisenberg M, Nilan M. "A Re-Examination of Relevance: Toward a Dynamic, Situational Definition". Information Processing and Management, 1990; 26: 755–776.
49. Journal of American Society for Information Science, 1994; 45: Special issue.
50. Belkin N, Vickery A. Interaction in Information Systems London: British Library, 1985.
51. Fidel R. "Searchers' Selection of Search Keys". Journal of the American Society for Information Science, 1991; 42: 490–500, 501–514 and 515–527.
52. Pejtersen AM. "A Library System for Information Retrieval Based on a Cognitive Task Analysis and Supported by an Icon-Based Interface". ACM Sigir Forum, June, 1989; 23: 40–47. (Special issue).
53. Croft WB, Thomson R. "I^3R: A New Approach to the Design of Document Retrieval Systems". Journal of the American Society for Information Science, 1987; 38: 389–404.
54. Salton G, Allan J, Buckley C. "Automatic Structuring and Retrieval of Large Text Files". Communications of the ACM, 1994; 37: 97–108.
55. Callan JP, Croft WB. "An evaluation of query processing strategies using the TIPSTER collection". In Korfhage R, Rasmussen EM and Willett P. (eds): Proceedings of the Sixteenth Annual International Conference on Research and Development in Information Retrieval (New York, ACM, 1993; 347–356.
56. Hancock-Beaulieu M. "Query Expansion: Advances in Research in Online Catalogues". Journal of Information Science, 1992; 18: 99–103.
57. Bates M. "The Design of Browsing and Berry-Picking Techniques for the Online Search Interface". Online Review, 1989; 13: 407–424.
58. Lundeen GW, Tenopir C. "Text Retrieval Software for Microcomputers and Beyond: an Overview and a Review of Four Packages". Database, 1992; 15: 51–63.
59. Pearsall J. "STATUS/IQ: a Semi-Intelligent Information Retrieval System". Information Services and Use, 1989; 9: 295–309.
60. Paijmans H. "Comparing the Representations of two IR systems: CLARIT and TOPIC". Journal of the American Society for Information Science, 1993; 44: 383–392.
61. Weyer SA. "Questing for the Dao: DowQuest and Intelligent Text Retrieval". Online, 1989; 13: 39–48.
62. Voorhees EM. "Implementing Agglomerative Hierarchical Clustering Algorithms for Use in Document Retrieval". Information Processing and Management, 1986; 22: 465–476.
63. Willett P. "Recent Trends in Hierarchic Document Clustering: a Critical Review". Information Processing and Management, 1988; 24: 577–597.
64. Belkin NJ, Marchetti PG, Cool C. "BRAQUE: Design of an Interface to Support User Interaction in Information Retrieval". Information Processing and Management, 1993; 29: 325–344.
65. Smeaton AF. "Information Retrieval and Hypertext: Competing Technologies or Complementary Access Methods" Journal of Information Systems, 1992; 2: 221–233.
66. Belkin NJ, Kantor P, Fox EA, Shaw JA. "Combining the Evidence of Multiple Query Representation for Information Retrieval". Information Processing and Management, 1995; 31: 431–448.
67. Vickery BC, Vickery A. "Online Search Interface Design". Journal of Documentation, 1993; 49: 103–187.
68. Ellis D. New Horizons in Information Retrieval. London: Library Association, 1990. (Second revised edition, 1995).