Division of Research
Graduate School of Business Administration
The University of Michigan

January 1984

INDETERMINACY IN THE SUBJECT ACCESS TO DOCUMENTS

Working Paper No. 358

David C. Blair

Graduate School of Business Administration
The University of Michigan

INDETERMINACY IN THE SUBJECT ACCESS TO DOCUMENTS

by

David C. Blair

Computer and Information Systems
The University of Michigan
Ann Arbor, Michigan 48109

F/1082

Abstract

Subject access to documents is influenced by two kinds of indeterminacy: the indeterminacy of the indexer's selection of indexing descriptors, and the indeterminacy of the inquirer's selection of search terms. The possibility of successful retrieval depends on how these two indeterminacies interact. Five types of interaction are discussed, and a change in the traditional method of subject searching is suggested as a way of reducing the effect of one of these two indeterminacies and of avoiding those types of interactions where the retrieval of the desired document(s) is impossible.

Document retrieval is a special province of information management for, unlike data retrieval, the retrieval system retrieves a representation of the information (documents) requested rather than the actual requested information [2]. Consequently, the retrieval effectiveness of a document retrieval system is predicated on the manner in which the documents in the database are represented. The process of creating document representations (sometimes called "references" or "surrogates") is called indexing, and the actual terms and phrases used to represent a document are called indexing descriptors. Such indexing descriptors describe the context of the document (e.g., author(s), title, data, length, document type, publisher, etc.) or its subject(s) (e.g., database systems, concurrency control, deadlock, etc.). The ultimate goal of an inquirer who uses a particular document retrieval system is to find one or more documents which, speaking loosely, satisfy his information need. To attain this goal, the inquirer must successfully select one or more indexing descriptors which: (1) have been used to represent the document(s) that will satisfy his information need, and (2) in the form of a formal query, will retrieve a set of documents small enough for the inquirer to browse through. (1) and (2) have been referred to as the Prediction Criterion and the Futility Point Criterion [1] of formal textual inquiry. In any document retrieval system there are two kinds of indeterminacy which militate against the successful satisfaction of the Prediction Criterion. There is an indeterminacy (or uncertainty) in the assignment (by indexers) of subject descriptors to documents, and an indeterminacy (or uncertainty) in the selection of search terms by an inquirer.

Inter-indexer Inconsistency

The first kind of indeterminacy is commonly referred to as inter-indexer inconsistency. What it means is that the selection of subject indexing terms

to be assigned to a particular document will vary from indexer to indexer. Such inconsistency is the case because there is no precise, specifiable procedure for deciding whether a given subject description should be assigned to a given document. Studies of inter-indexer consistency report varying rates of inconsistency. Much of this variance is, no doubt, a function of the character of the subject area with which the documents to be indexed deal, and the levels of training and experience of the indexers involved. One would expect less indexing inconsistency among term assignments to documents in a subject area like, for example, Botany, than in a subject area like Sociology. This is, of course, because the language of Botany is more normative than that of Sociology. Another source of variance in inconsistency figures may be a function of whether or not a controlled indexing vocabulary is used by the indexer. The important point here is not the exact size or cause of inter-indexer inconsistency, but the fact that it remains a noticeable kind of indeterminacy in document retrieval.

What does this inter-indexer inconsistency mean? It means that different indexers would be likely to assign different terms to the same document. That is, if copies of the same document were contained on different databases there is a good chance that each copy of the document would have different index terms assigned to it (i.e., it would be represented in different ways). Of course not all terms assigned to the document copies would be different. There would be a certain amount of overlap. An indication of such inconsistency in subject indexing can be seen in an experiment reported by Zunde and Dexter [16]. In this experiment the same document was given to eight indexers in a blind test to determine the inter-indexer inconsistency of these eight indexers. The results for one document are shown in Figure 1.

INDEXING TERMS	INDEXER NUMBER								PROBABILITY OF AN INDEXER SELECTING I_i
	1	2	3	4	5	6	7	8	
I_1	X	X	X		X		X	X	.75
I_2	X	X		X	X			X	.63
I_3				X		X	X	X	.50
I_4				X	X	X		X	.50
I_5	X	X					X	X	.50
I_6		X	X		X				.38
I_7		X		X	X				.38
I_8		X				X			.25
I_9					X	X			.25
I_{10}		X	X						.25
I_{11}		X				X			.25
I_{12}	X								.13
I_{13}		X							.13
I_{14}					X				.13

Figure 1

It is interesting to note that no index term was selected by all eight indexers, and no two indexers selected the same set of terms for the document in question. Not every inter-indexer consistency study has shown the same level of inconsistency (Zunde and Dexter, q.v., mention several other studies, and studies have been done on this phenomena more recently [3, 4, 5, 6, 13, 14]). Greater and lesser inconsistencies have been observed. But in no study was such inconsistency completely absent or insignificant. These inter-indexer inconsistency data show that for any given document to be indexed, there is a more or less definable set of candidate index terms from which terms will be actually selected for assignment to a document. This situation can be represented by Figure 2 (where: V = the total allowable terms which may be assigned to a document on the data base; C_I = the set of candidate terms for a specific document, e.g., D_i; IS = the set of terms actually assigned to D_i). It is unlikely that any individual indexer would know the entire candidate set of indexing terms for a given document. In fact, the candidate set of index terms for one document would be difficult to determine short of having the document indexed separately by several indexers as in [16]. Even here there is a reason to believe [11] that if additional indexers were to index the same document, the candidate set of terms would continue to increase (but at a decreasing rate). What is important is not how large C_I is for any document, but that it exists, and that the terms actually assigned to that document (i.e., the index set, IS) are a subset of a larger set of potentially assignable index terms. In short, the notion of a candidate set, C_I, is a way of expressing the inherent indeterminacy of subject index term assignment to documents.

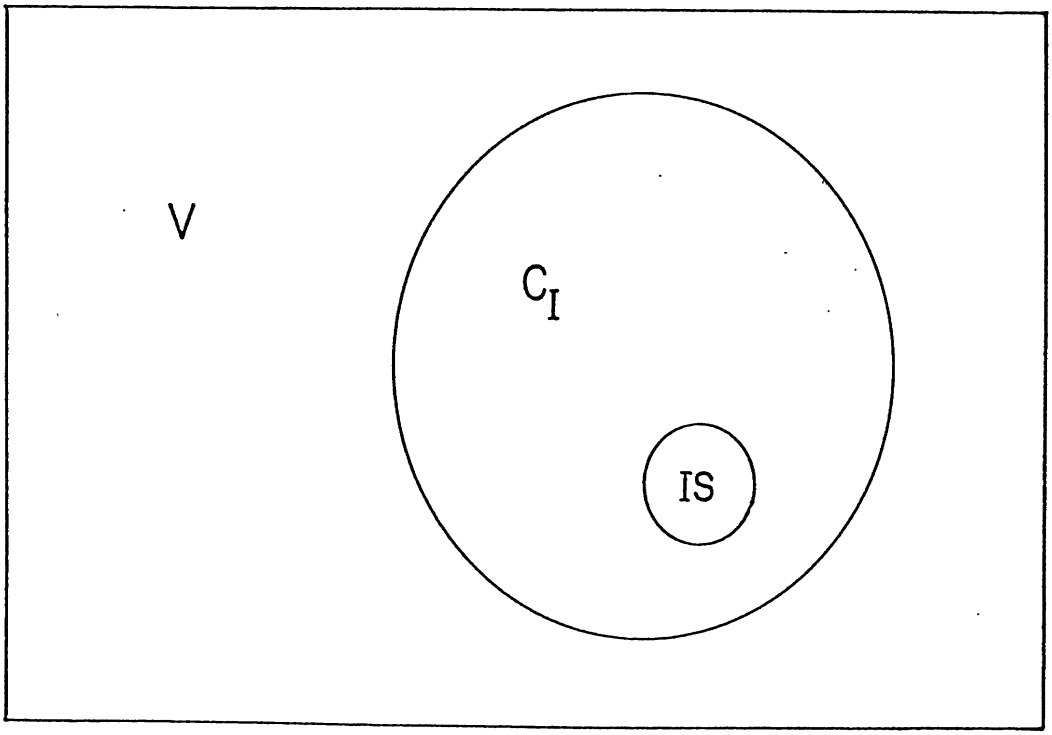


Figure 2

Uncertainty in Search Term Selection

The second kind of indeterminacy in document retrieval concerns the selection, by the inquirer, of the subject terms which he will use in some Boolean combination to construct a formal query to begin, or continue, his search for useful documents (weighted search terms will not be considered here). While we have some figures on the magnitude of inter-indexer consistency, we have little information on the variance of subject term selection by inquirers who would find the same document(s) useful. For the purposes of this discussion we shall make the reasonable assumption that for any given subject area, the inconsistency of term selection for inquirers who would find the same document useful, is at least as great as the inconsistency of indexers who might assign subject descriptors to that document. The combined forces of these two kinds of indeterminacy affect every subject search that is made on a document retrieval system.

Like the indexer, the inquirer can be said to have a candidate set of index terms from which he may select one or more terms to use in some Boolean combination to search for useful documents. This candidate set can be interpreted to mean that several inquirers, all of whom would be satisfied with the same document, e.g., D_i, would, independently of each other, use different subject terms as formal queries in their respective searches. We can represent this situation in Figure 3 (where: V = the total allowable terms which may be assigned to a document on the data base; C_Q = the set of candidate terms for the inquirer's search; QS = the set of terms the inquirer actually uses in a formal query). This is essentially the same situation as represented in Figure 2. But there is a difference. While the indexer's job is finished once he assigns a set of terms, IS, to D_i, the inquirer is

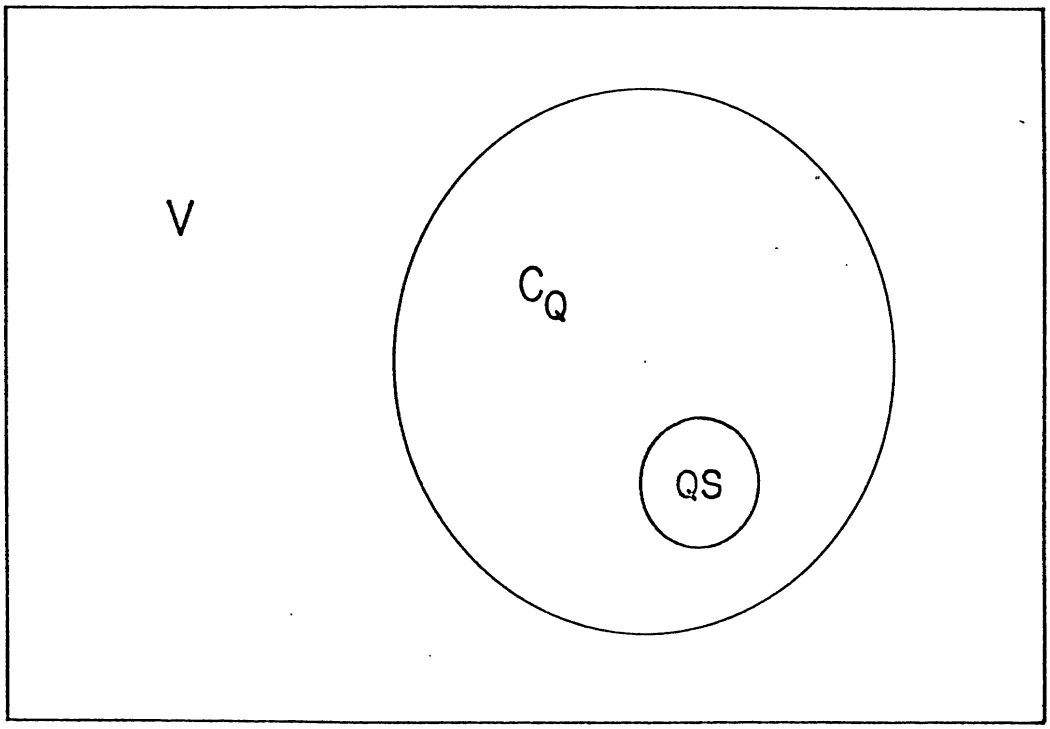


Figure 3

not limited to the selection of one set of terms for his search. Inquiry is typically a trial and error process [12], so if the inquirer does not retrieve D_i, on the first attempt, he will likely continue to formulate formal queries until he retrieves D_i or gives up in frustration (N. B. we shall assume, for purposes of simplification, that the inquirer will be satisfied only with D_i, and that a successful search is contingent upon only the satisfaction of the Prediction Criterion [1], not the Futility Point Criterion [1]).). This means that Figure 3 can be more realistically represented by Figure 4. Here the aggregate of Query Sets, QS_1, . . . , QS_n, will generally constitute a larger subset of C_Q than the single IS does of C_I (Figure 2). In set theoretic terms we can say that $QS_1 \cup QS_2 \cup \ldots \cup QS_n \subseteq CQ$ (the union of $QS_1 \ldots QS_n$ is a subset of CQ). The aggregate of Query Sets for one inquirer would not generally be equal to C_Q because C_Q represents the aggregate of Query Sets for all inquirers who would be satisfied with D_i. Presumably, each inquirer would have candidate terms that other inquirers (for varying reasons) would not have thought of, or would not consider using. C_Q represents the indeterminacy of search term selection for the inquirer searching on a document retrieval system.

Indeterminacy and Retrieval

How do these two kinds of indeterminacy influence document retrieval? We can answer this by stipulating the necessary conditions for successful retrieval of, here, D_i (which is indexed by the set of terms IS):

Condition 1. The candidate set of index terms, CI, must intersect with the candidate set of query terms, C_Q.

Condition 2. The index set, IS, and at least one of the query sets, QS_j, must be subsets of the intersection of CI and CQ

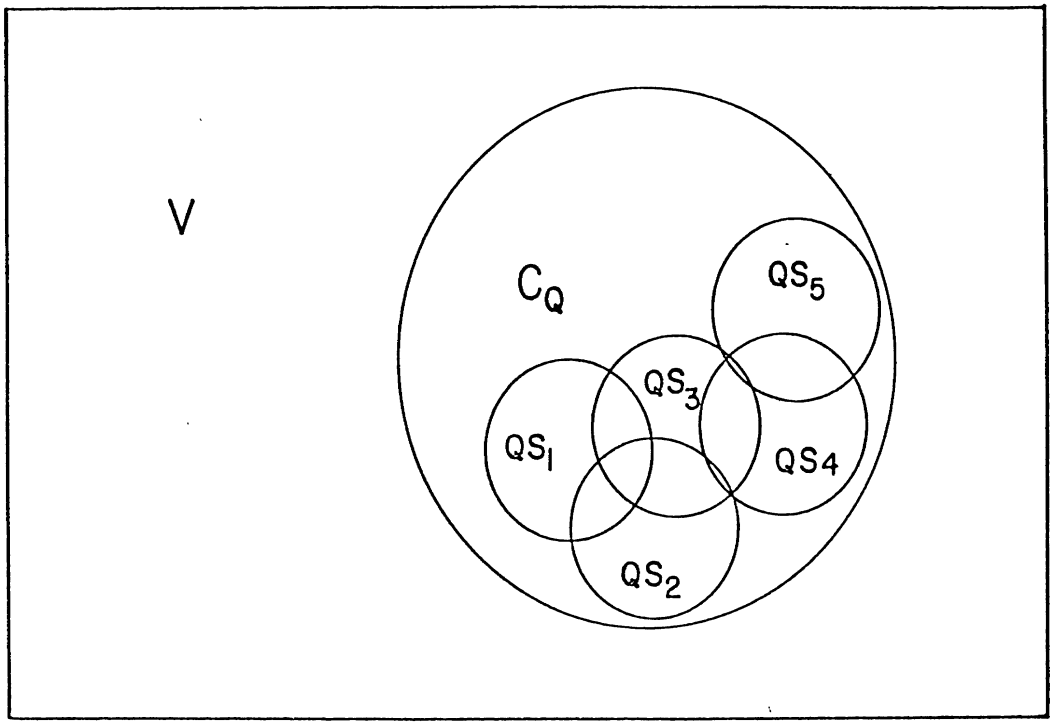


Figure 4

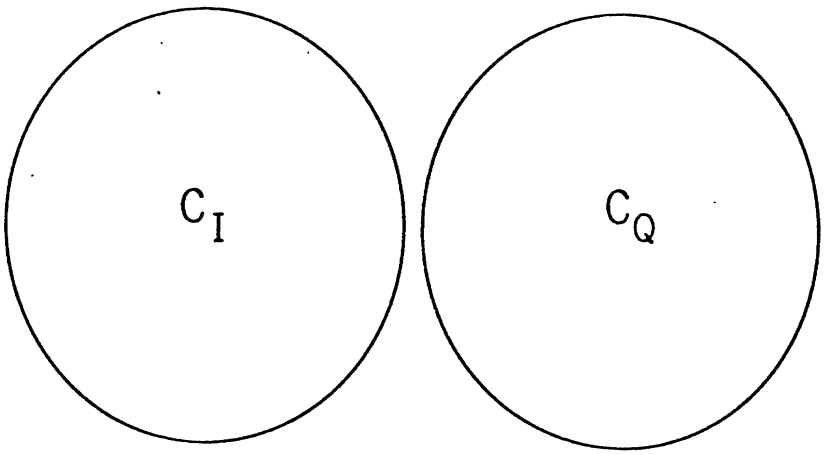
(i.e., IS U $QS_j \subseteq C_I \cap C_Q$).

Condition 3. The intersecting terms of QS_j must be a subset of IS ($\cap QS_j \subseteq IS$, where "$\cap QS_j$" = the intersecting terms of QS_j. If QS_j consists of a single term, e.g., I_k, we assume that I_k is an intersecting set, i.e., $I_k = I_k \cap I_k$.).

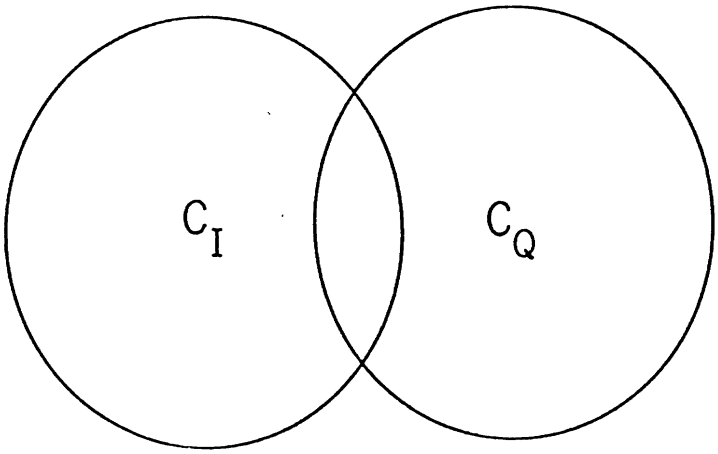
To begin with, let's just consider the first criterion of successful retrieval (above). Here we can distinguish 5 possible cases (see Figure 5). Let's look more closely at these cases.

Case 1 (the Naive Inquirer). C_I does not intersect with C_Q and the subject retrieval of D_i is impossible (sets C_I and C_Q are disjoint). This is the typical situation where the inquirer has no familiarity with the indexing philosophy of the retrieval system. Since C_Q in a subset of the indexing vocabulary, V, documents will be retrieved. The traditional solution to Case 1 situations is to educate the inquirer through instruction by professional searchers (usually indexers) who are familiar with the indexing philosophy of the system, or by expanding the inquirer's candidate set of query terms, C_Q, by the use of thesauri or associative searching algorithms. The difficulty with Case 1 is one of perception. When the inquirer queries the database and receives no satisfactory documents it may be difficult for him to determine whether he is in a Case 1 situation or whether there are no useful documents on the database. If he believes there are no useful documents on the database he may give up and not seek assistance, in spite of the fact that his belief is wrong (that is, his retrieval failure is due to Case 1, and there are, in fact, useful documents on the database).

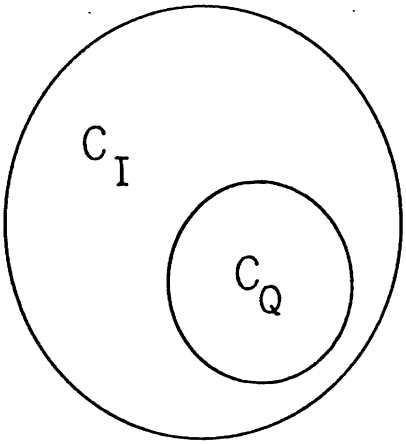
In Case 2 (the Moderately Informed Inquirer), we can identify two retrieval situations: Situation A, where IS $\subseteq C_Q$ (Figure 6); and, Situation



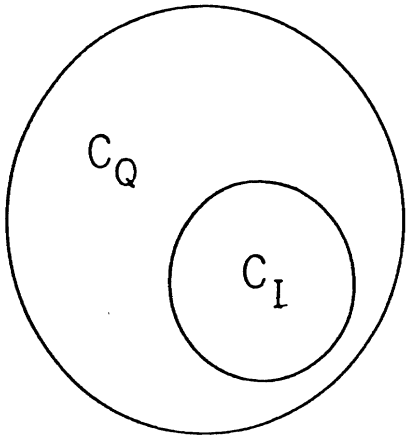
Case 1: $C_I \cap C_Q = \phi$



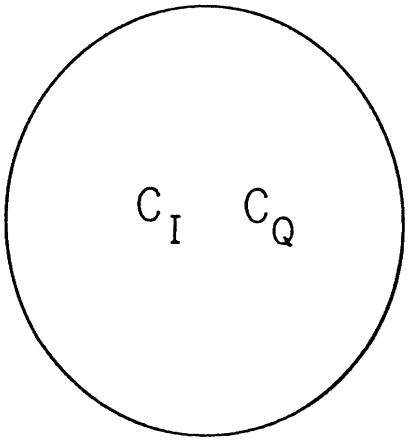
Case 2: $C_I \cap C_Q \rightarrow \phi$



Case 3: $C_Q \subset C_I$

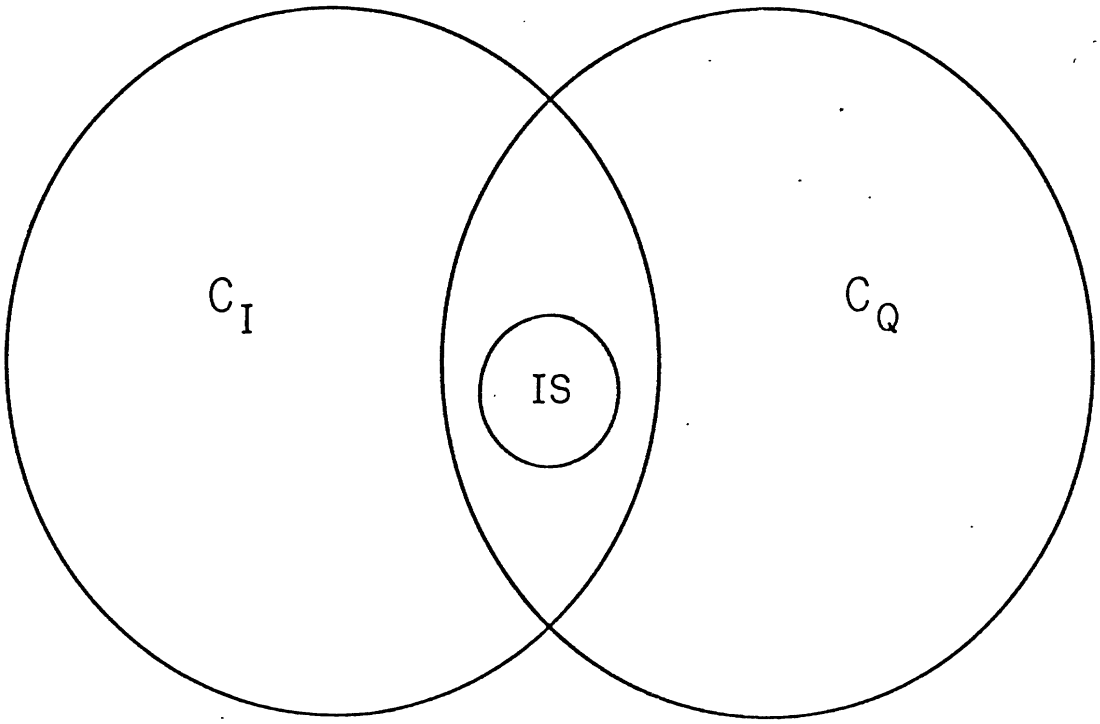


Case 4: $C_I \subset C_Q$



Case 5: $C_I = C_Q = C_I \cap C_Q$

Figure 5



Case 2: Situation A

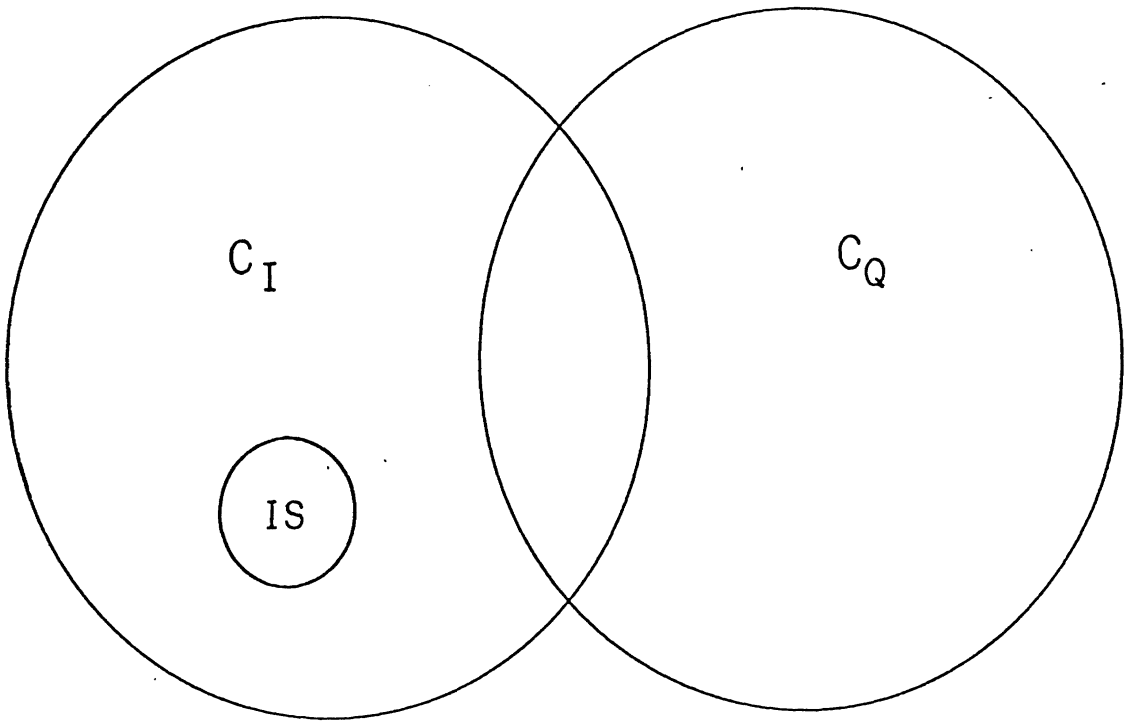
Figure 6

B, where IS $\subseteq \neg C_Q$ ("$\neg$" is read as "not") (Figure 7). In Situation A, retrieval of D_i is possible since Condition 1 (above) is met, and Conditions 2 and 3 are both possible to satisfy (i.e., since a given query set, QS_i, is, by definition, a subset of C_Q then it is possible for $QS_i \subseteq C_I \cap C_Q$ (Condition 2), and it is likewise possible for $QS_i \subseteq IS$ (Condition 3). In Situation B, on the other hand, retrieval of D_i is not possible since Condition 2 ($IS \subseteq C_I \cap C_Q$) is not met (by inference, Condition 3 cannot be met either).

Successful retrieval in a Case 2 situation is essentially a probabilistic matter based primarily on the likelihood of $IS \subseteq C_I \cap C_Q$. If we assume that C_I is an equiprobable space then the possibility of successful retrieval increases as the portion of C_I which does not intersect with C_Q decreases. This likelihood is represented in Figure 8.

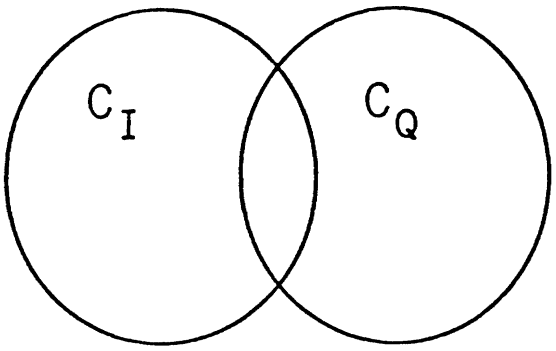
We might describe Case 2 as a situation where the inquirer has a moderate familiarity with the indexing philosophy of the retrieval system. Situation B, where successful retrieval of D_i is not possible, can be changed into Situation A by using the same solutions which were applied to Case 1 (inquirer education, thesauri or associative searching). Again, as in Case 1, there is the problem of the inquirer's perception of the cause of retrieval failure. He may believe his failure to retrieve useful documents is due to the fact that there are no useful documents on the database, and not consider the possibility that he is in a Situation B environment. If he believes the cause of his retrieval failure is due to the lack of useful documents on the database, then he may give up and not seek the assistance of professional searchers, thesauri, or associative searching techniques.

Case 3 is very similar to Case 2. Here the candidate set of terms for

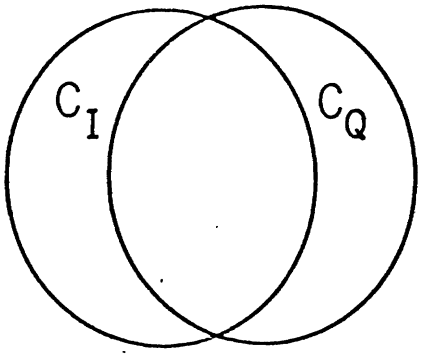


Case 2 : Situation B

Figure 7



Case 2: low likelihood of retrieval success (low likelihood of Condition 2 satisfaction)



Case 2: higher likelihood of retrieval success (higher likelihood of Condition 2 satisfaction)

Figure 8

the inquirer is a proper subset of the candidate set of terms for the indexer. We can say (as in Case 2) that the inquirer has a moderate familiarity with the indexing philosophy of the retrieval system. The same two situations occur here as they did in Case 2: Situation A (retrieval is possible) where IS $\subset C_Q$; and, Situation B (retrieval is not possible), where IS $\subset \neg C_Q$. As in Case 2, the possibility of successful retrieval increases as the portion of C_I which does not intersect with C_Q decreases.

In Case 4 (the Well-Informed inquirer) the indexer's candidate set of terms, C_I, is a proper subset of the inquirer's candidate set of terms, C_Q. We might call this the case of the well-informed inquirers, since they have a broader understanding of candidate terms for D_i than the indexers do. A likely situation for this to occur is on a research database used by experts in the database's field of coverage. In Case 4 retrieval will always be possible since Conditions 1-3 can always be satisfied. In this sense, it might be tempting to label Case 4 as an ideal retrieval situation. But we should refrain from this because there is a Case 4 situation in which retrieval is possible, but not likely. This would be the case where the indexers, as a group, are largely uninformed in the subject areas with which the database is concerned. In other words, C_I is a smaller subset of C_Q because indexing depth is very low (e.g., 1 or 2 terms per document). This is the situation for some of indexed databases on the DIALOGUE or ORBIT retrieval systems and can be a very frustrating search environment for the well-informed inquirer.

Case 5 (the Agreeable Inquirer) is the best retrieval situation for an inquirer to be in. Here the candidate sets of index and query terms are the same, from which it may be inferred that (speaking loosely) the indexer and

inquirer perceive the possibilities for representing D_i in the same way. Of course, both the indexer and the inquirer could be naive. In which case we have a situation of the blind leading the blind, and the possibilities of inquirer education are not good. But, in general, it can be argued that, ceteris paribus, Case 5 is one of the best of the 5 retrieval situations.

Since successful retrieval is, in the final analysis, contingent on the concurrence of IS and one of the inquirer's query sets, e.g., QS_j, we can summarize the likelihood of this event for each of the 5 cases:

Concurrence possibility of IS and QS_j:

Case 1: None

Case 2: Low (both IS and QS_j must be subsets of CI $\cap$ CQ. This is a Situation A.)

Case 3: Variable (IS must be a subset of C_Q. This is a Situation A.)

Case 4: Concurrence of IS and QS_j is always possible.

Case 5: Concurrence of IS and QS_j is always possible.

From this summary we can see that the satisfaction of the prediction Criterion [1] is always possible only in Cases 4 and 5, and is sometimes possible in Case 2 (Situation A) and Case 3 (Situation A). In other words, given enough search time and sufficient persistence on the part of the inquirer, retrieval of D_i is basically a matter of time for these cases (the probability of an exhaustive search by any one inquirer is not necessarily 1.0, though, because, by our definition of C_Q, any one inquirer may not know all the search terms in C_Q, and IS may include terms not considered by that inquirer). We have said that the situations where retrieval is impossible (Case 1, and Situation B of Cases 2 and 3) might be ameliorated by educating the inquirer or by making thesauri or associative searching algorithms

available to him on an elective basis. But because such aids are elective, the inquirer must be sophisticated enough to recognize his need for them, and this may not happen, due in large part to the difficulty the inquirer may have in distinguishing these 'impossible' situations from the situation where no useful documents exist on the database. This means that we could reduce the effect of the indeterminacy of index terms by designing document retrieval strategies that function primarily or solely in Case 2 (Situation A), Case 3 (Situation A), Case 4 or Case 5 environments. How could this be done?

Reducing the Effect of Subject Index Term Indeterminacy

The traditional methods of dealing with the indeterminacy of subject terms include such general approaches as "improving the indexing vocabulary/process", or the use of feedback or associative techniques to expand an inquirer's C_Q in the hope that it will eventually become large enough to make C_I a subset of it (though this is not the stated goal of such strategies). These traditional solutions have inherent flaws in them. While "improving the indexing vocabulary/process" is a noble enterprise, and no one would question its beneficial intent, it is virtually impossible to turn such lofty intents into concrete recommendations. The second solution fails for a different reason. Namely, the associative expansion of C_Q (by adding more query sets) will not in itself guarantee the eventual inclusion of C_I within C_Q. Semantic associations between index terms are easy to generate, as any casual perusal of the information science literature will indicate [9, 15]. This is due, of course, to the inherent richness and flexibility of natural language, and it means that C_Q might need to be expanded significantly before it is likely to include C_I. Why is this undesirable? It would seem that

since the eventual inclusion of C_I within C_Q is what we want, the number of terms which must be added to C_Q to effect this result would be inconsequential. Such is not the case. The expansion of the number of related terms available as search terms for the inquirer dramatically expands the number of possible search queries the inquirer might be able to formulate from them. To see this dramatic expansion let's look at a simple example.

Let's suppose that the inquirer knows all the terms in C_Q and that the number of unique such terms equals 5. This means that the total number of possible unique conjunctive queries which the inquirer may construct from these 5 terms is equal to:

$$\sum_{i=1}^{5} C_i^5, \text{ where } C_i^n = \frac{n!}{n!\ (n-i)\ !}$$

If we represent this value as C, then the total number of unique queries which the inquirer might formulate using any of 5 index terms is 31. This is a large, but not unmanageable, number of queries for the inquirer to use or select from (see [1] for a more detailed discussion of this phenomenon). But if we expand C_Q (by using a thesaurus or some other method of term association) to include 10 related terms instead of 5, then C (the number of unique conjunctive queries which can be formulated using these 10 terms) becomes 1023. Thus a doubling of associated search terms yields a <u>33 fold increase</u> in the number of possible search queries available to the inquirer. Merely expanding the number of terms available to the inquirer will be most likely to expand his available query choices beyond his ability to conveniently use them in his search. The solution is not just to add related terms to C_Q, but to add <u>all and only</u> the terms necessary to create a Case 2

(Situation A), Case 3 (Situation A), Case 4 or Case 5 environment.

Clearly, then, the traditional solutions to the indeterminacy problem do not significantly improve the subject retrieval of documents. Is there another solution? There is, and surprisingly it's a rather simple solution. Instead of having the inquirer begin his search by selecting one or more index terms with which to construct a formal query, have the inquirer begin his search by selecting a document of high utility in his present inquiry which is already contained in the database he is searching. If this can be done, the inquirer will be beginning his search with a very important piece of information: A direct representation of the indexing terms assigned to a document of high utility. This means that the conditional probability of another useful document being indexed with the specific descriptors of the selected document will approach 1.0. (It may not be exactly 1.0 because there might be other non-useful documents indexed with the same descriptors as the selected document, such a situation being due, in part, to the inherent indexing indeterminacy of term assignments by indexers as we have already discussed.) Such a strategy starts the inquirer off with, necessarily, the best representation of a useful document on this retrieval system. This method relates the two most subjective elements involved in document retrieval: The inquirer's estimation of what is useful to him (as represented by the selected document) and the indexer's estimation of how a document which is useful to the inquirer should be represented (i.e., the index terms he selects). Perhaps an analogy will prove helpful in grasping the importance of this change.

Suppose an individual needs some groceries but cannot go to the store, and let us assume that the store will deliver orders which have been placed

over the phone. Our friend knows what he wants from the store, but has recently arrived here from another country (where he grew up) and does not know what food this store stocks or what brands of similar items it carries. His grocery order looks like this:

Some breakfast meat,
Something sweet for dessert,
A variety of entrées,
Fresh vegetables,
Something special for a guest Friday night,
Several different kinds of soup,
.
.
.
.

Now when the order is delivered our friend sees that while the items he received do "fit" the order he made, they do so in a very loose way, and though many of the items were correct (they "fit" the list he made), they were not appropriate (he did not like them). Sometime later, he goes to this store and is amazed at how many desirable items there are in the store's inventory. "Why didn't you give me what I really wanted?" he asks the grocer. "Because you didn't ask for it. I can't read your mind," the grocer replies.

Now imagine a similar situation in which, like the first shopper, this individual cannot go to the store himself and is not familiar with the store's inventory. But this person takes a somewhat different approach. Instead of trying to describe what he wants in general terms, he goes through the stock of food he has so far accumulated and makes a list of all the items (including brand names) which he likes and wants more of. His list looks something like this:

1 pound Armour pork sausage,
Jello instant pudding (1 chocolate, 2 butterscotch),

2 pounds ground beef
1 pound flank steak,
1 bunch broccoli,
1/2 pound spinach,
2 Cornish game hens,
3 Campbells soup 1 @ (vegetable beef, chili bean, and Scotch broth),
.
.
.
.

Our friend still does not know whether this store stocks all these items, so he tells the grocer that these are all things which he likes and if the grocer does not have some of them in stock he is to give our friend something similar. The grocer, in filling the order, will approach the task in this way: If he does not have Armour pork sausage, he will substitute a "name brand" sausage, if he does not have flank steak he will substitute the leanest beef possible, if he does not have spinach he may substitute Bok Toy; etc.

A similar situation obtains in information retrieval. The inquirer, like our friend, cannot go to the "store" (he must use an intermediary (i.e., index terms) to access the database); and he is unsure of the inventory of the "store" (he does not know precisely what kinds of documents are stored in the database). Typically the inquirer draws up a search request somewhat like the first grocery list. He tries to describe what he wants in general terms by using the indexing language of the system, but because of the unavoidable indeterminacy in index term assignments (documents may be described differently from system to system and from indexer to indexer), his request must necessarily be vague and imprecise (no matter how precise the inquirer may believe it to be). By using the indexing terms assigned to a document of high utility for the inquirer as the first step in formal query

formulation (as described above), the inquirer can construct a formal query similar to the second grocery list. The retrieval system now has two important things: (1) It has an indication of what the inquirer has found useful (i.e., a specific document), and (2) it has a representation of that useful document in terms of the indexing descriptors and indexing philosophy of the retrieval system. The retrieval system can now concentrate its efforts towards providing information similar to what the inquirer values highly.

Search Strategy

Given that the inquirer has located on the database a document, D_i, of high utility, the indexing descriptors used to index D_i become the "seed" for the inquirer's search. Since the set of index descriptors assigned to D_i becomes the best available formal representation of the inquirer's information need, it seems reasonable that a good search procedure would be to retrieve the documents in the retrieval system whose sets of indexing descriptors most closely resemble the set of terms assigned to D_i. The documents retrieved could be weakly ordered according to the degree of fit between their individual sets of descriptors and the set of descriptors assigned to the "seed" document, D_i. A more sophisticated ranking algorithm could take into consideration the inquirer's ranking of the terms assigned to D_i in order of increasing/decreasing importance. Such possible derivative search strategies are numerous, and similar to associative search strategies already developed [9, 15].

The importance of beginning the inquirer's search with the index terms assigned to a document of high utility can be seen more clearly if we examine this strategy in light of the 5 cases discussed earlier. We can represent

this new strategy by Figure 9. Here we can see two important things: 1. IS and QS are equivalent; and, 2. there is no C_Q to speak of. (Actually, $C_Q = Q_S$.) Together these two observations indicate that an important change in searching methods has taken place. Namely, the inquirer begins his search with little or no indeterminacy in his search term selection. This means that instead of beginning his inquiry in an "impossible" situation (Case 1, Case 2: Situation B, or Case 3: Situation B) the inquirer begins his search in a Case 3: Situation A environment.

Discussion

It may be helpful to consider some objections that may be made to beginning a search with the index terms assigned to a document of high utility:

1. This form of searching is no different than methods which make use of the inquirer's feedback on the relevance of retrieved documents to modify the original formal query submitted to the retrieval system [7, 8, 10].

   Reply: Relevance feedback information can include procedures to delete, modify, or augment the index terms in the inquirer's original query based on a comparison of the similarities and differences between the subject terms in the query and those assigned to the documents judged useful or not useful by the inquirer. The problem with this search method is that the inquirer must begin his search by formulating his own search query. This means that the indeterminacy of search term selection will influence this first search query formulation. This inquirer may find himself in one of the "impossible" retrieval situations. (Case 1, Case 2: Situation B, or Case 3: Situation B), and not retrieve any highly

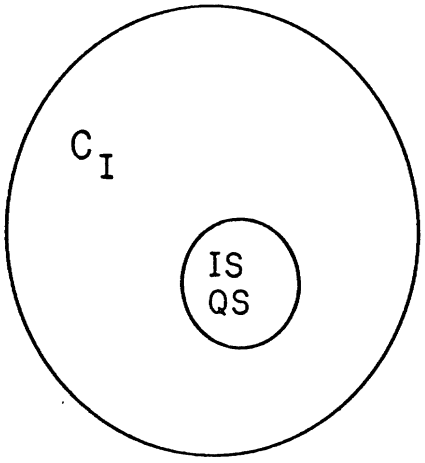


Figure 9

useful documents as a result of this first query. This could have two consequences: 1. The inquirer surmises that there are no useful documents on the database and gives up; or 2. the inquirer selects the best documents of the retrieved set (even though they may be only marginally useful) and gives this information as feedback to the system. The problem with the first consequence is readily apparent, but the difficulty with the second consequence is more subtle. The marginally relevant documents may be only the best of a bad set. Ideally, feedback information should be used to modify the search query in such a way as to improve the usefulness of the next set of documents retrieved. But because of the pervasive influence of the indeterminacy of index term assignment and search term selection it is not certain that the modification of the original inherently imprecise search query will necessarily improve the quality of the retrieved documents. In fact, there is reason to believe that the modification of the original search query based on information about marginally useful documents should create queries which will, at best, retrieve more of the same--marginally useful documents.

2. It may be difficult for an inquirer to find a highly useful document on the database with which to begin his search.

   Reply: If an inquirer cannot identify a highly useful document already on the database there are two possible conclusions which may be drawn: 1. The acquisitions policy of the database is, at least in this case, simply not doing its job. It is essential for any database to maintain, as a first priority, the most useful

documents for its regular inquirers. If these highly useful documents are not on the database the selection and retention policies of the retrieval system warrant serious review. 2. The selection and retention policies of the database are fine, but the inquirer is simply not using the right database. Retrieval systems should be designed to serve specific populations of inquirers. The ability of an individual inquirer to find a highly useful document on a database may be a reliable touchstone for determining whether he should devote a significant amount of his search time to looking for documents on this sytem.

A commitment by a retrieval system to supporting the kind of document retrieval advocated here is more than just a commitment to the development and maintenance of specific searching algorithms. It is a commitment to a certain kind of document selection and retention policy. To support this kind of retrieval the system designers should make a continual effort to find out from the inquirers who use the system regularly exactly which documents they (the inquirers) have found or do find of high utility in the normal course of their inquiry. Such seminal documents will form a "core" of access literature in the database. This direct solicitation from the inquirers of highly useful documents could be supplemented by the regular acquisition and indexing of "core" journals--those journals which contain a predictably high frequency of highly useful articles (such selection is not limited to just journals, it may include any contextually defined set of articles, reports, etc., such as the reports from a particular laboratory, or the publications authored by

a particular individual, etc.). The inquirer, knowing that, for example, all the articles from a certain journal are indexed and entered onto the database, would be on the lookout for documents in that journal which would prove valuable access points to the database. This core of highly useful documents could also be augmented with unobtrusive data. That is, whenever a document is selected as an access point by an inquirer, the retrieval system automatically adds information to that document's index record indicating that the document is now (if it has not already been) included in the core of the database.

Such a core of highly useful articles serves not only the inquirers, but also the indexers. When an indexer in the course of indexing new documents is uncertain whether to assign one or more index terms to a particular document, he could find examples of how the terms in question have been assigned by retrieving several documents from the highly useful core of documents which already have these terms assigned. Retrieving sample documents from the core of highly useful documents insures two things: 1. The indexers have readily available examples of indexing for the most important documents in the database; and 2. by comparing the document to be indexed with core documents already indexed with terms the indexer is considering, the indexer can get immediate feedback on the similarity of the document to be indexed to core documents in the database. This is especially important given that the basic searching procedure is to compare the set of terms assigned to the highly useful document selected by the inquirer and other documents

on the data base. An extension of such an indexing procedure might be to always present the indexer, when he assigns subject terms to a new document, with all (or a selection of all) the core documents that have exactly the same index terms assigned. This would help to build up the consistency of subject term assignments between new documents and core documents already on the database, and thereby help reduce the indeterminacy of subject term assignments. If statistics were kept by the retrieval system on how many times a core document was selected as an access point in an inquirer's search, the indexer, after indexing a new document, could be provided with core documents with the same index terms as the new document, but now these core documents are ranked according to the number of times they have been used as access documents. This would insure that the indexer is able to compare his work with the most frequently used (and, by inference, most valuable) core documents.

Inquirers who use such a retrieval system regularly might be encouraged to routinely submit to the database documents they believe to be significant in their work. This would insure that there would always be highly useful documents on the database for the inquirer. The inquirer could then check at regular intervals to see which documents have been indexed similarly to the ones he submitted. A highly useful document submitted by an inquirer might function like an informal SDI service. If all indexed documents included a searchable field which specified the date on which that document was entered into the database, the inquirer could use the

same document as a starting point for all his searches by restricting the retrieved set of similarly indexed documents to those acquired since his last search.

Since there is likely to be a certain amount of common interest and background among a database's regular inquirers, it might be useful for inquirers to be able to restrict their searches to the identified core documents on the database. This would ensure that an inquirer could at least begin his search by retrieving documents for which there was some consensus about their high quality. Naturally, the inquirer could expand his search beyond the core documents, but at least he would begin in a relatively rich set of useful documents undiluted by unevaluated documents.

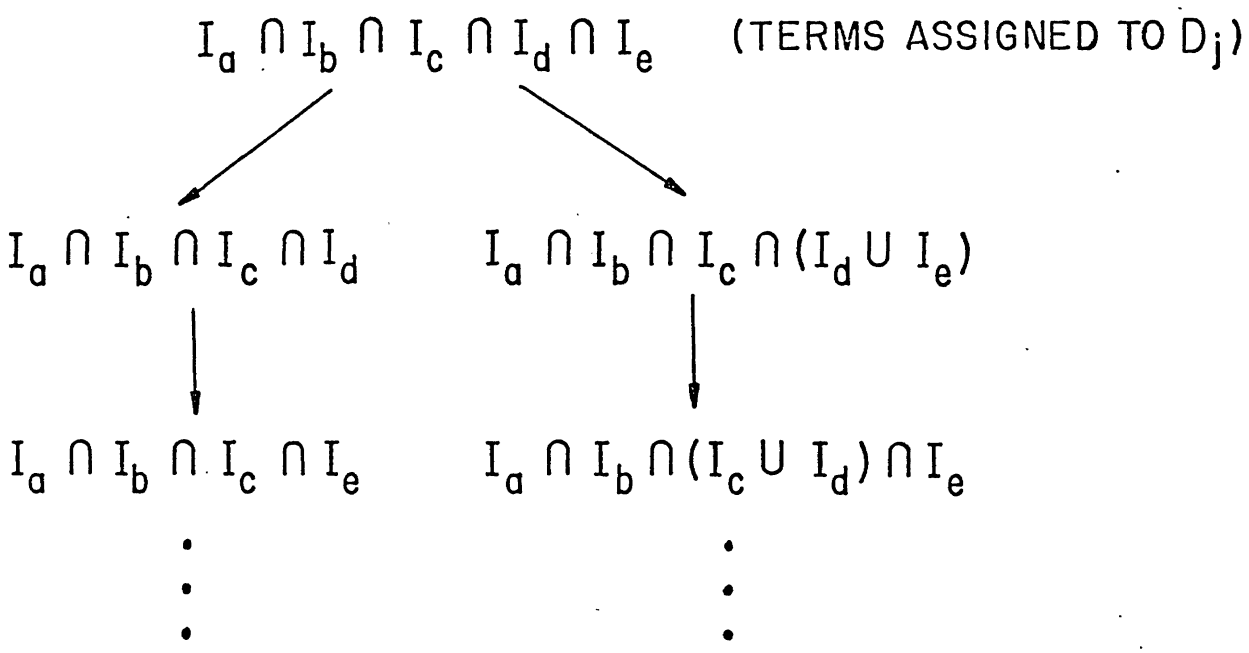
3. What if an inquirer had several highly useful documents relevant to the same inquiry, and each was indexed somewhat differently on the database?

Reply: If the inquirer has several useful documents he can use in his inquiry, so much the better. The fact that they may be germane to the same inquiry, yet indexed differently, would not be surprising. This approach to document retrieval only reduces the inquirer's search term selection indeterminacy, it does not reduce the indexer's term assignment indeterminacy (unless the indexing aids delineated in the reply to objection 2 are implemented). Thus, it would be reasonable to expect to still see varying amounts of indexing inconsistency in any document retrieval system. If the inquirer has several highly useful access documents these, in effect, give him a larger subset of the indexer's candidate set of index terms than if

the inquirer were to have only one access document.

4. The reduction of indeterminacy in search term selection affects only the retrieval of the highly useful document selected by the inquirer. Since the inquirer is already familiar with this document, the absence of search term indeterminacy in finding it is of no consequence if the search goes beyond this first document, which it must.

   Reply: Presumably, the document (e.g., D_i) selected by the inquirer will not be the only document on the database indexed in exactly this way, so the inquirer will retrieve documents in addition to the one he specifically selects. If the inquirer chooses to pursue his inquiry further than the initial retrieved set of documents, he will have a good "anchor set" of indexing terms (those assigned to D_i) on which to base his subsequent queries (see [1] for a discussion of the importance of the "anchor set" of search terms in document retrieval). The inquirer may expand his search beyond those documents with the same index terms as D_i by removing the term or terms he believes to be least important, or by forming a disjunction of some of the intersecting set of terms assigned to D_i (see Figure 10). Removing less important terms from a query that is known to retrieve at least one highly useful document (D_i), is an easier task than trying to create a query ex nihilo, which is the traditional procedure. In any event, the point of this discussion is that beginning a document search with the indexing set assigned to a highly useful document already on the database is a better place to start than traditional methods



POSSIBLE MANUAL VARIATIONS OF INITIAL SEARCH QUERY

Figure 10

provide. If an inquirer uses this method and retrieves no useful documents (other than D_i) then nothing prevents him from continuing his search by using more traditional methods.

Summary

There are two kinds of indeterminacy which affect subject searches on document retrieval systems: an indeterminacy in the assignment of subject descriptors to documents, and an indeterminacy in the selection of subject terms for use in formal search queries. These indeterminacies mean that both the indexer (who indexes D_i) and the inquirer (who would be satisfied with D_i) have their own respective candidate sets of index terms from which they may select terms to index a document or search for it. The success of a search for a document depends on whether, and in what way, these candidate sets intersect. There are five possible set relations between the indexer's candidate set of terms and the inquirer's candidate set of terms. These five cases can be divided into groups where the retrieval of the document in question is either possible or impossible. Because it is difficult for an inquirer to distinguish the retrieval situation where there are no useful documents on the database from the situation where there are useful documents on the database but it is impossible for him to formulate a query which will retrieve them, it is best to design document retrieval systems to function as much as possible in the situations where the retrieval of useful documents is at least possible. By beginning his search with the selection of a highly useful document that is already on the database, the inquirer has eliminated the indeterminacy of search term selection and has thereby improved his chances of finding useful documents like the one he selected to begin his search.

Acknowledgements

The author would like to thank Alan Merten, Tom Schriber, Dennis Severance and Michael Gordon of The University of Michigan, and M. E. Maron of the University of California, Berkeley, for their helpful comments on earlier versions on this paper.

Bibliography

1. Blair, David C. "Searching Biases in Large, Interactive Document Retrieval Systems," Journal of the American Society for Information Science, v. 31:4, July 1980, pp. 271-277.

2. _____. "The Data-Document Distinction in Information Retrieval," (Communications of the ACM, forthcoming).

3. Hooper, R. S. Indexer Consistency Tests: Origin, Measurement, Results and Utilization, IBM Corporation, Bethesda, MD, 1965.

4. Hurwitz, Frances I. "A Study of Indexer Consistency," American Documentation, v. 20, January 1969, pp. 92-94.

5. Jacoby, J. and V. Slamecka. Indexer Consistency under Minimal Conditions, Documentation Inc., Bethesda, MD, 1962.

6. Leonard, Lawrence Edwards. "Inter-Indexer Consistency and Retrieval Effectiveness; Measurement of Relationships," Ph.D. dissertation, University of Illinois, 1975.

7. Rickman, J. T. "Design Consideration for a Boolean Search System with Automatic Relevance Feedback Processing," Proceedings of the ACM 1972 Annual Conference, pp. 478-481.

8. Salton, G. The SMART Retrieval--Experiment in Automatic Document Processing, Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

9. _____. Dynamic Information and Library Processing, Prentice-Hall, Englewood Cliffs, 1975.

10. Stanfel, L. E. "Sequential Adaptation of Retrieval Systems Based on User Inputs," Information Storage and Retrieval, v. 7, 1971, pp. 69-78.

11. Swanson, Don R. "Studies of Indexing Depth and Retrieval Effectiveness," Unpublished report, National Science Foundation Grant GN 380, February 1966, p. 9.

12. _____. "Information Retrieval as a Trial-and-Error Process," Library Quarterly, v. 47:2, pp. 128-148.

13. Tarr, Daniel and Harold Borko. "Factors Influencing Inter-Indexex Consistency," in Proceedings, American Society for Information Science,

14. Tinker, J. F. "Imprecision in Meaning Measured by Inconsistency of Indexing," American Documentation, v. 17, April 1966, pp. 96-102.

15. Van Rijsbergen, C. J. Information Retrieval, 2nd Ed., Butterworths, London, 1979.

16. Zunde, Pranas and Margaret E. Dexter. "Indexing Consistency and Quality," American Documentation, July 1969, pp. 259-267.