Chapter 2: The Principal Formal Models of Information Retrieval

> "A rule stands there like a sign-post.---Does the sign-post leave no doubt open about the way I have to go? Does it show which direction I am to take when I have passed it; whether along the road or the footpath or cross-country? But where is it said which way I am to follow it; whether in the direction of its finger or (e.g.) in the opposite one?--- And if there were, not a single sign-post, but a chain of adjacent ones or of chalk marks on the ground---is there only *one* way of interpreting them?
>
> "The arrow points only in the application that a living being makes of it."
>
> (---Wittgenstein. Philosophical Investigations)

There are, of course, many proposed designs for the logical structure of document retrieval systems (and we will discuss the reason for this proliferation of designs in Chapter 3), but despite this variety, it is useful to provide a formal framework with which we can distinguish some of the major retrieval models. Figure 2.1 gives the basic structure for a document retrieval system and the following models describe the principal ways in which this system may operate. By presenting the major retrieval models in this abbreviated, formal manner we can make some useful comparisons between them, and identify some of the major issues and trade-offs in information retrieval system design.[1] (It should not be inferred that these formal models represent actual or recommended retrieval systems. Their primary purpose is to permit us to see some of the fundamental processes which combine to make up information retrieval systems)

Model 1 is the simplest document retrieval model, and it's also the most common. This model describes, of course, the retrieval characteristics of a typical library (where books are retrieved instead of documents). The advantages of this model are: (1) It is the simplest retrieval model (in a library, books are retrieved by looking up a single author, title or subject descriptor in a catalogue); and (2) It is the most

widespread type of information retrieval system, which also makes it the most familiar.

Despite the widespread acceptance of Model 1, there is one major disadvantage: Since an inquirer can use only a single descriptor as a formal request, retrieval results tend to deteriorate as the number of documents in the data base grows. That is, a request using a single descriptor (especially a subject descriptor) may "retrieve" (or, "identify") so many documents that the user cannot possibly look through them all. This problem is called "output overload"[2] and is a frequent problem in document retrieval systems. The reader who has conducted searches in a large research library (e.g., the Library of Congress, or Berkeley's Doe Library) has, no doubt, experienced this difficulty first-hand.

This single-descriptor request restriction causes another problem for the inquirer by preventing multiple-descriptor searches (such as, "retrieve all documents with the subject descriptors 'data base management systems' and 'distributed systems'"). Given the limitations

---

MODEL 1

I. FORMAL REQUESTS ARE SINGLE DESCRIPTORS.

II. DOCUMENTS ARE ASSIGNED SETS OF ONE OR MORE DESCRIPTORS.

III. IN REPLY TO A REQUEST, DOCUMENTS ARE EITHER RETRIEVED OR NOT (WEAK ORDERING).

IV. RETRIEVAL RULE: IF THE DESCRIPTOR IN THE REQUEST IS A MEMBER OF THE DESCRIPTORS ASSIGNED TO A DOCUMENT, THEN THE DOCUMENT IS RETRIEVED.

e.g.: REQUEST $= D_k$

DOCUMENT A $= < D_a, D_b, D_c >$ not retrieved

DOCUMENT B $= < D_b, D_k, D_m >$ retrieved

---

of this model, it's remarkable how often it appears as the logical structure of expensive, large-scale, computerized retrieval systems. The single-descriptor request format was a constraint imposed by the limitations of library card catalogues. But computerized systems do not have these inherent structural limitations, so the occurrence of Model 1 computerized systems can only be attributed to an inability to distinguish between data retrieval and document retrieval (*vid.* Chapter 1). That is, since most data retrieval requires only single logical access points, then why should document retrieval be any different? (Of course, even this simple model of data retrieval is now being questioned, too, with the emergence of high level Data Manipulation Languages such as SQL and the need for *ad hoc* inquiry).

Summarizing Model 1:

ADVANTAGES: simple retrieval process; widespread, familiar implementation.

DISADVANTAGES: single descriptor requests are less effective as the data base grows; no multiple-key retrieval.

Model 2 provides a slight improvement over Model 1 by offering the inquirer somewhat greater flexibility in formulating search queries. Large document/test data bases have become increasingly common in the last decade and, as a result, multiple descriptor search requests have become the predominant type of formal query. The reasons for this were discussed in Chapter 1 when we described how the typical inquirer tries to satisfy the competing retrieval goals of predicting the index terms assigned to the desired documents as well as retrieving small enough sets of documents to browse through (satisfying the FPC and the PC).

Model 2 is harder to implement than Model 1, especially if the system is manual (the "edge-notched" cards, popular in the 1950's and '60's, comprised a Model 2 citation retrieval system). Model 2 can be implemented more easily on a computerized retrieval system, although the problems of efficient multiple-key file access (necessary for Model 2 retrieval) have not yet been extensively explored.

Some readers may object that Model 2 cannot handle disjunctive queries in its present form. But this is not the case, though on a manual system (such as "edge-notched" cards) processing such queries is cumbersome. A knowledge of simple propositional logic reveals that disjunctive search queries can be non-loss transformed into queries which have only conjunctions as connectives. For example:

$$D_k \cdot (D_j \vee D_l)$$