Thoughts about intermediary systems in information retrieval

by GERARD SALTON
Cornell University
Ithaca, New York

In information retrieval, we must deal with many hundreds of different data bases, and with several dozen search and retrieval services used to provide access to these databases. In view of the common operating requirements of all retrieval services, namely large file sizes, on-line operations conducted by end-users or search intermediaries, and system responses provided to users in real time, the normal sequential search methods cannot be used. Instead, it becomes necessary to use auxiliary indexes capable of providing access to specific subsections of the files. At the present time, all operating retrieval services use large (inverted) index files to obtain fast search output with acceptable retrieval effectiveness.

Even though the internal data organization and the search strategies are effectively identical, the access protocols and command structures used by the many retrieval services are very different, and generally incompatible with each other. To bridge the gap between the common internal operating characteristics and the multiple external access protocols, user accessing aids have been designed consisting of easy-to-use front-ends, expert advice systems for query formulation and submission, and gateways capable of reaching the proper retrieval service and record files. The regrettable proliferation in noncompatible retrieval services is now matched by an equally large number of noncompatible intermediary and gateway systems.

The many different retrieval systems and intermediary services can effectively be replaced by a common query formulation and query submission system based on natural language manipulations. Such a system would provide the following facilities:

1. Initial query analysis in terms of weighted attribute vectors
2. Global collection matching facilities designed to identify the relevant document collections that must be searched in a particular instance
3. Global document matching facilities designed to provide ranked output of retrieved documents in response to submitted queries
4. Displays of expanded query vocabularies consisting of thesaurus contents and relevant phrases to be used in generating improved query formulations
5. Relevance feedback facilities designed to construct improved query formulations by automatic methods.

Such a common interface system could serve as a cornerstone for a highly effective and easily usable retrieval facility.