""""""current comments"

Full Text Searching Systems Just Around the Corner?

Number 36

September 5, 1973

In 1955, I suggested the potential value of a citation index to the Bible. Such an index would show where Biblical passages had been cited, implicitly or explicitly, in commentaries or other works. My interest in this application of citation indexing was recently revived when I learned about a project underway at the Weizmann Institute of Science in Israel. Under the direction of A.S. Fraenkel, the project is studying the use of so-called full-text searching in the retrieval of Hebrew and Aramaic legal texts.

Professor Fraenkel's project interests me both as a linguist and as a documentalist. As he points out, the rationale of FTS is to avoid the expense and inherent limitations of a priori abstracting and indexing. FTS "replaces a priori indexing by an ad hoc sifting method which can be optimized and tailored to the needs of the individual user and to his interests at any given time." In principle, FTS is simple enough. In an FTS system, a document is stored word-by-word in a computer's memory. Every stored word is addressable, and any document can be retrieved on the basis of any word it con-

FTS is easier described than done. Even in English where the form of words is relatively unaffected by the demands of grammar, the linguistic problems are enormous. Goose/geese and teach/taught are simple "exceptions" in English. Much of the interest

(and sweat) in Dr. Fraenkel's project comes from an opposing characteristic of Semitic languages. In Hebrew and Arabic the form of words may undergo rather violent change, violent to the Indo-European linguistic eye. For example, the plural of the loan-word film in Arabic is aflam. This kind of thing makes FTS in a Semitic language extraordinarily more difficult than in English. It helps explain why a straight concordance to the English Bible can be quite useful, while a similar concordance to the Hebrew Bible is of little if any use at all.

FTS is very difficult also because you can't depend upon an author's choice of words in his writing about a subject you want to research. This is the reason why one could not depend upon word-for-word "manual" scanning of documents in pre-computer days. That is why traditional systems of indexing, whether hierarchical, alphabetic, or numerical were invented. A significant purpose of these traditional schemes was to bring the facts of natural language into some sort of predictable order. In the place of natural language was substituted the artificial and highly restricted languages of hierarchies, classifications, subject heading lists, and codes. Unfortunately the use of such artificial languages has never been as successful as researchers and librarians could wish. The premise of FTS is that they have indeed been

unsuccessful enough to justify the trouble of attacking natural language. The success of numerous natural language systems such as Permuterm® & KWIC indexes justifies this hypothesis.

I suspect that we shall hear more about FTS as the cost of direct-access computer memories goes down. Surely FTS must appeal to the storage and retrieval instincts of any reprint hound. Even now, a reprint collection of 1000 papers, averaging 5000 words in length, can be stored easily on a single computer disk. And it's quite reasonable that one might want to search such a reprint collection word-by-word to retrieve an otherwise elusive document.

Disregarding the cost of computer usage, an immediate problem of FTS is presented by input, that is, getting the full text of each document into the memory. Without some sort of universal optical character recognition device-one that can read any type face and digitize it for a computer's memory-it's necessary now to key the entire text of every document. Ignoring for the moment verification or proofing of the input, you can count on a minimum of an hour per document. Keying 1000 reprints would keep a secretary busy for six months! Readying the document for keying can be even more time-consuming and expensive. In Fraenkel's project, for example, this socalled pre-editing handles only 5500 text words per day.

This input problem would be less formidable if one could obtain, along with the printed text, a machine-language version. Organizations like the AIP and ACS have done government-supported experiments to create entire journals in machine readable form. Nevertheless, we can expect FTS to be restricted primarily to full-texts of abstracts and titles for the time being.

IBM and others operate such systems with considerable success.

The Fraenkel project is particularly interesting to me, because "a citation system is imbedded in [the] full text retrieval system . . . We combine citation searching with full text searching, which indeed leads to improved performance." The same has been found true for other, modern legal information systems now in operation. Their inclusion of citations-that is, of precedents-is an obvious necessity. However, as you have heard me say ad nauseam, citation searching has been found equally useful for scientific material. Salton of Cornell has done much "full-text" searching of abstracts. Like Fraenkel he has found that inclusion of references as part of the searched text has added greatly to search precision.4

- Garfield, E. Citation index to the Old Testament. Paper presented at the annual meeting of the American Documentation Institute, Philadelphia, Penna., November 2-4, 1955.
- 2. For knowledge of this project I am indebted to Professor B. Weiss of the Jewish Theological Seminary in New York. For knowledge of Professor Weiss I am indebted to Dr. Chaim Potok, Editor of the Jewish Publication Society of Philadelphia. For knowledge of Dr. Potok I am indebted to his novels The Chosen, The Promise, and My Name is Asher Lev.
- Choueka, Y., Cohen, M., Dueck, J., Fraenkel, A.S. & Slae, M. "Full text document retrieval: Hebrew legal texts." Proc. Symposium on Information Storage and Retrieval, College Park: University of Maryland, April 1-2, 1971, p. 61-79.
- Salton, G. The evaluation of automatic retrieval procedures; selected test results using the SMART system. Amer. Documentation 16:209-22, 1965.