[*Journal of Documentation*, vol. 24, no.2]

LETTERS TO THE EDITOR

Dear Sir,

Automatic document selection

I was particularly interested in reading the paper by Hutchins (*J.Doc.*vol.23, no.4, 1967, p. 273-90) dealing with the automatic indexing and retrieval procedures incorporated into the Quantras System. Many years ago, I had ideas quite similar to those contained in his excellent article, and I believed that phrase-matching procedures using a simplified type of syntactic analysis could be used for the implementation of effective automatic retrieval systems.

While I still believe that this type of process has merit, particularly for certain types of customers requiring a high precision performance, I now know that other, much simpler automatic procedures will provide a retrieval effectiveness which, on the average, will be superior to that furnished by the syntactic phrase process outlined by Hutchins.

In particular, the extensive evaluation results published both by the Aslib-Cranfield Project[*] and by the SMART Project[†] + indicate that the requirement of syntactic connection between phrase components is too stringent for the average user, and that a syntactic process, therefore, brings with it an unwanted loss in recall. As an example, a request on 'information retrieval' would not normally retrieve a document containing a sentence such as 'Adequate retrieval performance is essential for people in need of information', if the syntactic process were used. Furthermore, most syntactic procedures, including the one described by Hutchins, are based on the use of dictionaries or thesauruses prior to the application of a syntactic process. Unfortunately, most dictionaries are not particularly suitable for use in a retrieval environment, and in the latter case, they tend to produce a loss in precision.

To summarize, the process described by Hutchins is not likely to be superior either in precision or in recall to much simpler word stem matching procedures. Additionally, it requires the storage of the full text of documents (rather than only the analyzed form as is done in the SMART system), and this in itself will probably preclude the use of the Quantras System for some time to come.

Yours faithfully,
GERARD SALTON

Department of Computer Science,
Cornell University,
Ithaca, NY 14850
USA
26 February 1968

Mr Hutchins replies:

While recognizing that there are drawbacks in the use of syntactic analysis in retrieval systems, I would suggest that in its general approach Quantras may have potential advantages over systems which analyse documents. In addition to the arguments against indexing as such (section I of my paper), the high cost in time and money of designing and operating automatic IR systems needs to be taken into account.

The cost of designing automatic systems, which might incorporate such complex components as statistical analysis, vocabulary controls, and possibly also syntactic analysis, can be very high. The systems must be precise: they cannot tolerate conflicting analyses of the same word or phrase. They must also avoid failures which result in inaccurate analyses, since such failures may mean either that the system itself must be revised and the document(s) re-analysed – at further expense – or that the store must include incorrect entries, with the consequent risk of poor retrieval performance. Since failures might easily be

[*] CLEVERDON, C., *and* KEEN, M. *Factors determining the performance of indexing systems: vol. II:* test results. Aslib-Cranfield Research Project, Cranfield, 1966.

[†] SALTON, G., *and* LESK, M.E. Computer evaluation of indexing and text processing. *Journal* of *the Association for Computing Machinery*, vol. 15, no. I, January 1968.

caused by linguistic features (e.g. neologisms) not foreseen by the designers, they are not likely to be very rare.

Because Quantras does not index it needs to deal only with the language of search requests – probably mainly noun phrases with few finite verb forms – and not with every possible sentence which might occur in texts. For this reason the theoretical and technical problems to be solved may well be less complex than those in automatic indexing systems. It would not require statistical analysis. Nor would it demand the same degree of precision: e.g. alternative syntactic analyses would be permissible (section 3); and failures to find a word of the request in the dictionary would not mean that an individual search could not proceed (section 5). Also it should be noted that, unlike failures in automatic indexing systems, no failure in Quantras could ever affect the quality and retrieval capability of the store on which searches are made.

Quantras could also prove cheaper in operation. Documents (and they need not be full texts) would merely be copied into the store instead of undergoing expensive analysis. The operating costs of analyzing and transforming; search requests by Quantras would not necessarily be higher than the cost of analyzing requests in natural language by any other fully automatic system. Lastly, the cost of searching the store would depend directly on the size of the store. There is nothing inherent in Quantras requiring that the whole of texts be searched. It could operate on any portion of text from short titles upwards (para. 2.1 and 5.4). Therefore the size of the Quantras store to be searched need not be larger than a store of documents in analyzed form. As in all IR systems, the optimum for retrieval efficiency and economic operation could only be established in practice.