Brief Communication

A Note About information Science Research

Gerard Salton
Department of Computer Science, Cornell University, Ithaca, NY 14853

This note deals with the relationship between information science research and practice. The impression that the field is moribund and that the research output is uniformly inferior is not supported by an examination of the information retrieval literature.

A provocative note by Carl Keren appeared in a recent issue of *JASIS* under the title "On Information Science" [1]. The note raises interesting questions about the relationship between information science research and practice. Keren asks four questions (slightly paraphrased):

(a) Do researchers in information science write about research that advances the state of the art?
(b) How much of the research output has really contributed to our body of knowledge?
(c) Is information science a name with a recognizable body attached to it? Is it a subject whose contents we can define?
(d) Is there a lack of feedback between researchers and technologists?

Questions such as these are of course not new, and they are reflective of legitimate concerns. In fact, I share some of Mr. Keren's misgivings about the quality and quantity of acceptable research work in our field. But we part company when Keren, by implication, answers all his questions in the negative, and specifically blames the researchers for what he considers the unsatisfactory state of information science work. In stating his arguments, Keren makes valid points about research based on questionable survey output, about lack of knowledge of the literature, about rediscovery of the obvious, and the like. Unfortunately, the note is also replete with remarks of doubtful relevance and with unsubstantiated claims about the inferior state of information science compared with that of other natural sciences. Taken as a whole, the note presents an antiresearch, if not anti-intellectual image, and requires, I believe, some amendments.

Accepted October 16, 1984

Among the less edifying parts of Keren's note is a completely gratuitous quotation of a remark, purported to have been made by Y. Bar-Hillel, to the effect that information scientists who need mathematics to describe their concepts quite probably know little of either information science or mathematics. To which one replies, first, that the remark has little to do with Keren's concern about the quality or quantity of research in information science, and second that the comment is obviously foolish and should not, in fairness to Bar-Hillel, have been repeated. (Bar-Hillel was an astute observer of our field; but he was partial to sweeping, and sometimes indefensible generalizations which might better be forgotten.) The truth of the matter is that a mathematical approach cannot transform poor information science into good research, and that few people are able to make convincing use of mathematics in the information field. But this does not, of course, imply that all mathematics is extraneous to information science, or that all information scientists who use mathematics are inferior.

What about Mr. Keren's main question? I will beg leave from answering the question about the scope and definition of information science. I just do not find that discussions relating to the definition of disciplines advance the state of the art, and personally I do not have much to contribute. I will then confine myself to the other questions concerning the quality and quantity of research in information science, and the contributions of this research to our body of knowledge and to information science practice. In discussing these questions, Mr. Keren mentions that "I did not research this, nor do I have firm data, but on the strength of my own experience, as a practitioner, I have the impression that only very rarely (does) useful and nontrivial information emanate from the published research. . . ."

Well, I do not have any firm data either, but my own assessment generally agrees with Keren's: Most of the published research in our field is probably not worth doing and ought to be forgotten. I differ with Keren in that I do not jump to the implied conclusion that therefore our researchers are all inferior and that information science research is useless. Nor do I have any evidence that rela-

tive to the number of researchers in information science, the proportion of first-rate work in our field is smaller than it is in other fields. The latter point is made explicitly at some length in Keren's note.

As I said, I cannot give conclusive evidence either way, but I can cite the literature. What follows is a list of active research topics pursued by people who generally have something to say. All of these topics belong to that part of information science which may loosely be entitled "text analysis, storage, and retrieval," and the literature citations cover the three most recent years (1981–1983). I emphasize that the list does not constitute a ranking of the best people, or the best topics, or the best papers. It is simply a sampling of recent work that offers interesting perspectives to researchers and practitioners alike. For a different view of the information science field in areas not covered here, the reader may wish to turn to other recent surveys [2,3].

(1) Under the *vector processing* retrieval strategy, both queries and stored documents are represented by sets of terms without Boolean operators. The vector processing model has been criticized because the implicit assumption is that the terms are independent of each other. Bookstein [4] has shown, however, that under various stated assumptions the preferred probabilistic retrieval approach and the vector processing systems are closely related. The vector processing model, whose effectiveness is hard to surpass in practice, has also been studied more recently by Wong and Raghavan [5].

(2) *Probabilistic retrieval* models have been preferred by researchers because they have a proper theroretical base and allow, at least in theory, the use of arbitrary combinations of dependent terms. Two different, well-known, probabilistic models have recently been compared, and a unified model including elements of both models has been proposed by Robertson, Maron, and Cooper [6,7]. An elegant, theoretically optimal probabilistic model based on the maximum entropy principle has been introduced by Cooper [8,9]. Croft has described and evaluated procedures for extending the basic probabilistic model by allowing the use of weighted terms for document representation [10,11]. Yu and co-workers [12] have introduced a generalized probabilistic model in which dependent term pairs, term triples, and higher-order term combinations can be included in decreasing order of the importance of the various term combinations. Finally, Wu and Salton [13] have evaluated some practical methods for actually obtaining the needed probabilistic parameters for the case where term independence is assumed.

(3) The probabilistic retrieval approach is theoretically attractive, but until viable methods appear for actually obtaining the probabilistic parameters for dependent-term combinations, the simpler vector processing approaches will be used in practice, or the established *inverted file* procedures based on term list manipulations. In this connection substantial attention has been paid recently to the acceleration and optimization of the similarity computations between queries and documents using inverted file strategies. Smeaton and van Rijsbergen [14] and Murtagh [15] have worked on the optimization of near-neighbor computations usable for query-document comparisons. Noreault and Chatham [16] have applied similar methods to the computation of term-term similarities, and Willett [17,18] has used optimal inverted file strategies to compute the document-document similarities needed for document clustering. All of these optimized inverted file procedures produce substantial improvements over conventional inverted file methodologies and are directly usable in current operational retrieval environments.

(4) The well-known query reformulation process based on *relevance feedback* uses relevance assessments obtained from retrieval system users in response to earlier information searches. Attar and Frankel [19] have continued working on local (and hence inexpensive) clustering and relevance feedback procedures. Term relationships obtained from a maximum spanning tree of term-term similarities are used in a relevance feedback environment by van Rijsbergen, Harper, and Porter [20]. All of these automatic query expansion and reformulation methods appear to be eminently suitable for practical implementation [21].

(5) Operational retrieval systems are currently based on *Boolean query formulations*. The conventional Boolean methodology is rudimentary: no term weighting, no output ranking of retrieved items, no automatic query formulation. A great deal of attention has been devoted by researchers to the refinement of Boolean query manipulations. Bookstein [22] has compared different term weighting systems in Boolean query environments. Buell and Kraft [23–25] and Radecki [26] have studied the *fuzzy set* retrieval models which permit the use of weighted document identifiers with Boolean queries. A different extended Boolean system was introduced by Salton, Fox, and Wu [27] which is more general than the fuzzy set system and appears to produce large improvements in retrieval output. This system can be superimposed on the currently operating retrieval methods. Regrettably, some mathematical considerations are needed to understand the operations of the extended system.

(6) Fancy *front-end procedures* have been studied by many people and beginnings have been made toward the design of *expert systems* that facilitate the access to the stored collections [28–34]. Marcus [29], has designed a flexible system that uses a common query language to access a variety of different bibliographic retrieval systems. The interactive system by Doczkocs [30] is based on flexible term weighting and automatic thesaurus display operations. Frei and Jauslin [33] use graphic terminals and multiple window displays to guide the user through the query formulation process using a hierarchical thesaurus display. Finally, Guida and Tasso [34] propose an expert-type interactive system in which user responses to appro-

priately phrased questions are utilized as a guide to the search process.

(7) Bibliographic *citation networks* have been used for many years to characterize various disciplines and literature types. Kwok [35] suggests that terms from cited titles be used for document identification purposes, and White and Griffith [36,37] continue the extensive experimentation with citation network manipulations.

(8) The area of *text passage retrieval*, as opposed to full document retrieval, has not yet come into its own, but it offers interesting prospects as a compromise between document retrieval and question-answering. O'Connor [39] has done pioneering work in this area, and Bernstein and Williamson [40] have built an interesting system for retrieval of text passages from a specially constructed handbook on hepatitis diseases.

(9) The related area of automatic indexing using *linguistic analysis* is being seriously revived by various researchers. For example, Dillon and co-worker [41,42] use phrase assignment techniques in various text processing experiments, and Sparck Jones and Tait [43] describe a sophisticated system for the automatic generation of syntactically (and semantically) acceptable term phrases.

(10) The interesting area of fast text *character matching*—for example, between query and document texts—is treated by Haskin and Hollaar [44] who propose the construction of special hardware devices for rapid character string processing. Sophisticated *simulation methods* for the construction of representative bibliographic databases are described by Tague and Nelson [45]. Simulated collections simplify the operations of controlled retrieval evaluation experiments. Finally, a large group of people including, for example, Macleod and Crawford [46], are interested in the design of *mixed search and retrieval* systems that process formatted numeric databases as well as text data in an integrated operation.

All of the foregoing research areas are directly relevant to information retrieval practice in that they point the way to the construction of flexible, interactive retrieval systems which should substantially simplify the collection search and retrieval processes for future generations of information seekers. Many other information processing topics might have been given prominence: for example, recent attempts at automatic abstracting [48], word fragment encoding experiments [49], the design of distributed retrieval systems [50], and the area of systems evaluation [51]. For present purposes, the foregoing list suffices to demonstrate that an impressive variety of research topics is being pursued by people who have something to offer.

In the face of the relative wealth of interesting work actually available, Mr. Keren's rhetorical questions about "What have you researchers done lately for us practitioners?" seems to miss the point. There are never any shortcuts in bridging the gap between research and application. That is true in our area as much as in all other intellectual areas of endeavor. It is necessary to *study* the literature; it is necessary to have sufficient know-how to discriminate, and to put matters in context. Eventually the pieces will fit together, and the observer can judge the specifics instead of being forced to rely on superficial impressions and generalizations. In our field, as in every other, it is necessary to know the field before being able to contribute.

I see no reason, however, to resign. There has been substantial progress in information system design these past 20 years. No doubt, the changes will accelerate in the years to come. The field will evolve even more rapidly if the practitioners were to stop blaming the research side and asked instead "What have we practitioners lately done for ourselves?"

References

1. Keren, C. "On Information Science." *Journal of the American Society for Information Science*. 35(2):137; March 1984.
2. Kochen, M. "Information Science Research: The Search for the Nature of Information." *Journal of the American Society for Information Science*. 35(3):194-199; May 1984.
3. Herner, S. "Brief History of Information Science." *Journal of the American Society for Information Science*. 35(3):157-163; May 1984.
4. Bookstein, A. "Explanation and Generalization of Vector Models in Information Retrieval." In: Salton, G.; Schneider, H.J., eds. *Research and Development in Information Retrieval, Lecture Notes in Computer Science*. Berlin: Springer-Verlag; 1983: 146:118-132.
5. Wong, S.K.M.; Raghaven, V.V. "Vector Space Model of Information Retrieval—A Reevaluation." In: van Rijsbergen, C.J., ed. *Research and Development in Information Retrieval*. Cambridge, England; Cambridge University Press; 1984:167-186.
6. Robertson, S.E.; Maron, M.E.; Cooper, W.S. "Probability of Relevance: A Unification of Two Competing Models for Document Retrieval." *Information Technology: Research and Development*. 1(1):1-22; January 1982.
7. Maron, M.E. "Associative Search Techniques Versus Probabilistic Retrieval Models." *Journal of the American Society for Information Science*. 33(5):308-310; September 1982.
8. Cooper W.S.; Huizinga, P. "The Maximum Entropy Principle and its Application to the Design of Probabilistic Retrieval Systems." *Information Technology: Research and Development*. 1(2):99-112; April 1982.
9. Cooper, W.S. "Exploiting the Maximum Entropy Principle to Increase Retrieval Effectiveness." *Journal of the American Society for Information Science*. 34(1):31-39; January 1983.
10. Croft, W.B. "Document Representation in Probabilistic Models of Information Retrieval." *Journal of the American Society for Information Science*. 32(6):451-457; November 1981.
11. Croft, W.B. "Experiments with Representation in a Document Retrieval System." *Information Technology: Research and Development*. 2(1):1-22; January 1983.
12. Yu, C.T.; Buckley, D., Lam, K.; Salton, G. "A Generalized Term Dependence Model in Information Retrieval." *Information Technology: Research and Development*. 2(4):129-154; October 1983.
13. Wu, H., Salton, G. "The Estimation of Term Relevance Weights Using Relevance Feedback." *Journal of Documentation*. 37(4):194-214; December 1981.
14. Smeaton, A.F.; van Rijsbergen, C.J. "The Nearest Neighbor Problem in Information Retrieval." *Proceedings of Fourth International Conference on Information Storage and Retrieval, SIGIR Forum*. 16(1):83-87; Summer 1981.

15. Murtagh, F. "A Very Fast Exact Nearest Neighbor Algorithm for Use in Information Retrieval." *Information Technology: Research and Development*. 1(4):275-283; October 1982.
16. Noreault, T.; Chatham, R. "A Procedure for the Estimation of Term Similarity Coefficients." *Information Technology: Research and Development*. 1(3): 189-196; July 1982.
17. Willett, P. "Document Clustering Using an Inverted File Approach." *Journal of Information Science*. 2:223-231; 1980.
18. Willett, P. "A Fast Procedure for Calculation of Similarity Coefficients in Automatic Classification." *Information Processing and Management*. 17(2):53-60; 1981.
19. Attar, R.; Frankel, A.S. "Experiments in Local Metrical Feedback in Full-Text Retrieval Systems." *Information Processing and Managements*. 17(3):115-126; 1981.
20. van Rijsbergen, C.J., Harper, D.J., Porter, M.F. "The Selection of Good Search Terms." *Information Processing and Management*. 17(2):77-91; 1981.
21. Pietilainen, P. "Local Feedback and Intelligent Automatic Query Expansion." *Information Processing and Management*. 19(1):51-58; 1983.
22. Bookstein, A. "A Comparison of Two Systems of Weighted Boolean Retrieval." *Journal of the American Society for Information Science*. 32(4):275-279; July 1981.
23. Buell, D.A.; Kraft, D.H. "Threshold Values and Boolean Retrieval Values." *Information Processing and Management*. 17(3):127-136; 1981.
24. Buell, D.A.; Kraft, D.H. "A Model for a Weighted Boolean Retrieval System," *Journal of the American Society for Information Science*. 32(3):211-216; May 1981.
25. Buell, D.A. "A General Model of Query Processing in Information Retrieval Systems." *Information Processing and Management*. 17(5):249-262; 1981.
26. Radecki, T. "Reducing the Perils of Merging Boolean and Weighted Retrieval Systems." *Journal of Documentation*. 38(3):207-211; September 1982.
27. Salton, G.; Fox, E.A.; Wu, H. "Extended Boolean Information Retrieval." Technical Report 82-511, Department of Computer Science, Cornell University, August 1982; also in *Communications of the ACM*. 26(11):1022-1033; November 1983.
28. Meadow, C.T.; Hewett, T.T., Aversa, E.S. "A Computer Intermediary for Interactive Data Base Searching." *Journal of the American Society for Information Science*. 33(5):325-332; September 1982; also 33(6):357-364; November 1982.
29. Marcus, R.S. "Experimental Comparison of the Effectiveness of Computers and Humans as Search Intermediaries." *Journal of the American Society for Information Science*. 34(6):381-404; 1983.
30. Doczkocs, T.E. "From Research to Application: The Cite Natural Language Information Retrieval System." In: Salton, G; Schneider, H.J., eds. *Research and Development in Information Science, Lecture Notes in Computer Science*. Berlin: Springer-Verlag; 1983:146:251-262.
31. Heine, M.H. "A Simple Intelligent Front End for Information Retrieval Systems Using Boolean Logic." *Information Technology: Research and Development*. 1(4):247-260; October 1982.
32. Pollitt, A.S. "End User Touch Searching for Cancer Therapy: A Rule Based Approach." *Proceedings of Sixth Annual Conference on Research and Development in Information Retrieval. SIGIR Forum*. 17(4):136-145; Summer 1983.
33. Frei, H.P.; Jauslin, J.F. "Graphical Presentation of Information and Services: A User-Oriented Interface." *Information Technology: Research and Development*. 2(1):23-42; 1983.
34. Guida,G.; Tasso, C. "IR-NLI: An Expert Natural Language Interface System." Conference on Applied Natural Language Processing, Assocation for Computational Linguistics, Santa Monica, CA, 1983, pp. 31-38.
35. Kwok, K.C. "A Document-Document Similarity Measure Based on Cited Titles and Probability Theory and its Application to Relevance Feedback Retrieval." In: van Rijsbergen, C.J., ed. *Research and Development in Information Retrieval*. Cambridge, England: Cambridge University Press; 1982.
36. White, H.D.; Griffith, B.C. "Authors as Markers of Intellectual Space: Cocitation in Studies of Science, Technology and Society." *Journal of Documentation*. 38(4):255-272; December 1982.
37. White, H.D.; Griffith, B.C. "Author Cocitation: A Literature Measure of Intellectual Structure." *Journal of the American Society for Information Science*. 32(3):163-171; May 1981.
38. Noma, E. "Untangling Citation Networks." *Information Processing and Management*. 18(2):43-53; 1982.
39. O'Connor, J. "Citing Statements: Computer Recognition and Use to Improve Retrieval." *Information Processing and Management*. 18(3): 125-131; 1982.
40. Bernstein, L.M.; Williamson, R.E. "Testing of a Natural Language Retrieval System for a Full Text Knowledge Base." *Journal of the American Society for Information Science*. 35(4): 235-247; July 1984.
41. Dillon, M.; Gray, A.S. "Fasit: A Fully Automatic Syntactically Based Indexing System." *Journal of the American Society for Information Science*. 34(2): 99-108; 1983.
42. Dillon, M. "Thesaurus-Based Automatic Book Indexing." *Information Processing and Management*. 18(4):167-178; 1982.
43. Sparck Jones, K.; Tait, J.I. "Automatic Search Term Variant Generation." *Journal of Documentation*. 40(1):50-66; March 1984.
44. Haskin, R.L., Hollaar, L.A. "Operational Characteristics of a Hardware-Based Pattern Matcher." *ACM Transactions on Database Systems*. 8(1):15-40; March 1983.
45. Tague, J.; Nelson, M. "Simulation of Bibliographic Databases Using Hyperterms." In: Salton, G.; Schneider, H. J., eds. *Lecture Notes in Computer Science*. Berlin: Springer-Verlag; 1983:146:194-208.
46. Macleod, I.A.; Crawford, R.G. "Document Retrieval as a Database Application." *Information Technology: Research and Development*. 2:43-60; 1983.
47. Schek, H.J. "Methods for the Administration of Textual Data in Database Systems," in: Oddy, R.N.; Robertson, R.E.; van Rijsbergen, C.J.; Williams, P.W., eds. *Information Retrieval Research*. London: Butterworths; 1981:218-235.
48. Paice, C.D. "The Automatic Generation of Literature Abstracts: An Approach Based on the Identification of Self-Indicating Phrases." In: *Information Retrieval Research*. London: Butterworths; 1981:172-191.
49. Kropp, K.; Walch, G. "A Graph Structured Text Field Index Based on Word Fragments." *Information Processing and Management*. 17 (6):363-376; 1981.
50. Moulinoux, C., Faure, J.C.; Litwin, W. "Messidor: A Distributed Information Retrieval System." In: Salton, G.; Schneider, H.J., eds. *Lecture Notes in Computer Science*. Berlin: Springer-Verlag; 1983:146:51-61.
51. Bollmann, P. "Two Axioms for Evaluation Measures in Information Retrieval." In: van Rijsbergen, C.J., ed. *Research and Development in Information Retrieval*. Cambridge, England: Cambridge University Press; 1984:233-246.