Available online at www.sciencedirect.com

ScienceDirect

Information Processing and Management 44 (2008) 971–972

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Brief Communication

An Historical Note on the Origins of Probabilistic Indexing

M.E. Maron

Professor Emeritus, University of California, Berkeley

Received 3 February 2007; accepted 12 February 2007
Available online 9 May 2007

Abstract

The motivation behind "Probabilistic Indexing" was to replace two-valued thinking about information retrieval with probabilistic notions. This involved a new view of the information retrieval problem – viewing it as problem of inference and prediction, and introducing probabilistically weighted indexes and probabilistically ranked output. These ideas were first formulated and written up in August 1958.

Keywords: Probabilistic information retrieval

In 1958 I was at the Ramo-Wooldridge Corp. (R-W) working on ways of using computers to get access to files of information: I was thinking hard about the problem of information retrieval. The traditional way of viewing the problem was, simply put, as follows: Every document has a subject matter which is what it is about. A human (indexer) must then read or scan a document and assign one or more index terms which describes what that document is about. The assignment of these subject terms was a two-valued affair – either an index term was assigned or not. There was no middle ground. Then a person in search of information would look for desired documents by searching under that term which best described his interest or concern. The "output" of a search was the set of documents to which the search term had been assigned. For every document in the collection either it was retrieved or not. Again, there was no middle ground. I felt that the two-valued assumptions incorporated in this traditional view of information retrieval were primitive. I felt that there should be degrees of aboutness; namely, that a term should be applied to a document to a degree – or with a weight. But how might these weights be interpreted in a rational way? I thought that it might be possible to interpret these weights in terms of probabilities. But how? Then I realized that the problem of information retrieval should be viewed statistically as a problem of inference and prediction–that the patron's search (query) term should be interpreted as a clue (a piece of evidence) and that via Bayes' theorem the library "system" should compute and predict, for each document in the collection, the probability that it would be wanted by the searcher in question. Thus, the IR system would compute the probability of relevance for each document and rank those documents by their computed values of probability of relevance. Such a ranking would have the great advantage of doing away with two-valued retrieval. Thus, instead of simply retrieving or not retrieving each document, given a search query, the user could be presented with a ranked list – a ranking of the output by probability of relevance. This probability ranking would constitute an optimal way for the

0306-4573/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2007.02.012

library patron to search the output; viz., look first at that document with the highest computed value of probability of relevance, then next in descending order, and so on. When I finally put it all together I had figured out how the weights could be interpreted as precisely defined probabilities and I had a probabilistic theory of information retrieval – probabilistic indexing and output ranking according to computed values of probability of relevance. I felt a great sense of excitement and quickly wrote up an internal document for senior R-W management. That document was dated August 1958.

Shortly thereafter I urged my old friend Lary Kuhns to join me at R-W. He did and for the next year we worked together to further develop, clarify and expand the new theory of information retrieval called "Probabilistic Indexing". (During that year we produced two large reports (Maron, Kuhns, & Ray, 1958, 1959), and we presented our results at a professional society meeting.) Lary Kuhns was an outstanding logician and mathematician and he made important contributions by proposing various measures of correlation between index terms by means of which a search query could be expanded. And he showed how probabilistically indexed documents could be viewed as weighted vectors and how the "distance" or similarity between documents could be measured and used to extend and expand a search. We submitted a description of our work to JACM and it was published in the July 1960 issue (Maron & Kuhns, 1960).

John Lary Kuhns, my friend and colleague, died at home in Woodland Hills, CA on February 24, 2006.

References

Maron, M. E., & Kuhns, J. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery, 7*(3), 216–244.

Maron, M. E., Kuhns, J. L., & Ray, L. (1958). *Some experiments with probabilistic indexing* (45pp.). Ramo-Wooldridge.

Maron, M. E., Kuhns, J. L., & Ray, L. C. (1959). Probabilistic indexing: A statistical technique for document identification and retrieval. Technical Memorandum No. 3. Data Systems Project Office, Ramo-Wooldridge, June 1959, 91pp.