Relevance Feedback
In An Automatic Document Retrieval System

A Thesis
Presented to the Faculty of the Graduate School
of Cornell University for the Degree of
Master of Science in Computer Science

by

Eleanor Rose Cook Ide
January, 1969

Biographical Sketch

Eleanor Ide, born in Flint, Michigan on September 7, 1940, earned a Bachelor of Arts Degree with Distinction and High Honors in Sociology from the University of Michigan in 1962. She was a research assistant for one year at the Population Studies Center, University of Michigan, and a computer programmer for three years studying adaptive logic and pattern recognition at IBM Systems Development Division, Endicott, New York. There she co-authored with Doctor Cyril J. Tunis "An Unsupervised Adaptive Algorithm," in IEEE Transactions on Electronic Computers, Volume EC - 16 #12, December, 1967, and co-invented with Dr. Tunis and Mitchell P. Marcus the 'Adaptive Logic System for Unsupervised Learning' described in Patent Application IBM Docket EN966005, now being filed.

Eleanor Ide, a member of Phi Beta Kappa, Sigma Delta Epsilon, and the Association for Computing Machinery, is presently employed as a Senior Systems Analyst by Hydra Computer Corporation in Raleigh, North Carolina.

Dedication

This thesis is gratefully dedicated to my husband, Robert Ide, whose help and patience made my graduate work possible.

Acknowledgments

The author wishes to acknowledge the programming support of Robert Williamson, and the advice and encouragement of Professor Gerard Salton and Michael Keen. The free computer time supplied by Hydra Computer Corporation is also appreciated.

Abstract

Many automatic document retrieval systems represent documents and requests for documents as numeric vectors that indicate the subjects treated by the document or query. This study investigates relevance feedback, a process that allows user interaction with such a retrieval system. The user is presented with a small set of possibly relevant documents, and is asked to judge each as relevant or non-relevant to his request. The numeric vectors representing the judged documents are usedto modify the numeric vector representing the query, and the new query vector is used to retrieve a more appropriate set of documents. The relevance feedback process can be iterated as often as desired. Several feedback algorithms are investigated in a collection of 200 documents and 42 queries. Two distinct viewpoints are taken in evaluation; one measures the movement of the query vector toward the optimum query defined by Rocchio, the other measures the retrieval experienced by the user during the feedback process. Several performance measures are reported from each evaluation viewpoint. Both evaluation methods indicate that relevance feedback is an effective process.

All algorithms tested that use only relevant document vectors for feedback provide equally good retrieval. Such algorithms should supply additional documents to any user

who judges every document presented for feedback to be non-relevant. Algorithms using non-relevant document vectors for feedback improve the retrieval obtained by these users without requiring additional relevance judgments.

The relevance feedback algorithms tested are based on the assumption that the vectors representing the documents relevant to a query are clustered in the same area of the document space. The conclusion that no tested relevance feedback algorithm is completely appropriate for the experimental environment is supported by a hypothesis that explains the observed contrasts between the behavior of strategies using only relevant documents for feedback and that of strategies using non-relevant documents. This hypothesis states that for most queries, some relevant document vectors are separated from others by one or more non-relevant document vectors. The implications of this result for future research in relevance feedback, partial search or multi-level strategies, multiple query strategies, request clustering, and document vector modification are discussed, and useful evaluation measures and new algorithms for these areas are suggested.

Table of Contents

List of Illustrations

Page

Page

		Page
Figure 37:	Example of Poor Rocchio Performance Caused by Non-Relevant Document Feedback	145
Figure 38:	An Example of Good Rocchio Performance on Separate Clusters of Relevant Documents	145a
Figure 39:	A General Two-level Cluster Feedback Algorithm	169
Figure 40:	Three Relevant Documents Forming Three Separated Clusters	176
Figure 41:	A Multiple Query Algorithm	184e