Report No.

Statistical problems in the application of probabilistic models to information retrieval

S.E. Robertson
J.D. Bovey

Final report to the British Library Research and Development Department on Project SI/G/517 (October 1981 - September 1982)

November 1982

Centre for Information Science
City University
Northampton Square,
London ECIV OHB.

Abstract

Some statistical problems arising in the use of probabilistic models in retrieval are analysed. An attempt is made to identify statistical methods from other fields which may be useful in retrieval. The logistic model, or class of models, which was developed for medical diagnostics, is identified as a promising approach. Retrieval methods based on the logistic model are developed and tested on three test collections; but the results indicate that the logistic models do not achieve the performance levels of traditional methods. Reasons for this discussed.

The logistic approach remains an attractive one for research purposes, particularly because it allows parameters to be added or removed at will. An attempt is made to derive a formal rule to determine when a new parameter (representing say an interaction between two terms) should be added. A single applicable rule is not reached, but the analysis throws some light on the problem.

Contents