5. Final Discussion

Our main feeling at the completion of this project is some disappointment that the logistic model has not so far shown that it can reach the performance level of existing methods. Since the new model is also somewhat more complex in application than the old, we have no grounds for claiming any practical use for logistic methods.

Nevertheless, there are some positive conclusions we can draw. The theoretical properties of the class of logistic models make it particularly attractive as a vehicle for experimental and theoretical work. For example, the manner in which dependence parameters can be incorporated into the model to any desired extent clearly makes the model suitable for investigation of dependence; similarly it would be suitable for query expansion work.

In both these cases. the theoretical work described in section 4 is relevant. Although we failed to achieve a simple, practical rule for when the dimension of the model should be increased, we believe we have made a substantial step in that direction. Further, in the process we have gained some insight into the problem. We can state that the possible profitable use of a new parameter depends not so much on whether that property exists or is correlated with relevance, as on whether we have enough data to make use of the property. Thus, we should no longer argue between independence and dependence models on the grounds of the existence of dependence, but rather on the basis of whether we can use the fact. Furthermore, and irrespective of the dependent/independent argument, the usefulness of a new parameter depends a great deal on how many parameters we have already.

On the problem of estimation which was one of the central problem areas originally identified, our progress has not been great. We had high hopes of the logistic approach because of the non-random

nature of the estimating sample (the same argument is used forcefully in the medical context). However, the hoped-for benefits did not emerge; we now consider, as indicated in section 2.4.2, that the use of the complement method effectively negates that advantage of the logistic model. We further failed to obtain much advantage from the ability to choose any prior distribution for the logistic model (although clearly there is scope for further experiments on these lines). It appears that the "prior" that is implied in the RSJ model, "*0.5*" formula, is as good as anything else we tried. It may be worth analysing the characteristics of this implicit prior, and trying to reproduce them in the logistic model. Further, it may be that where very small samples are concerned, the RSJ model (although not ideal) still has an advantage over the maximum posterior method used with the logistic model.

Finally, the method we have developed for realistic evaluation of feedback in searching should be useful for future experiments. Indeed, we would like to see a number of such experiments, since the few that we have done suggest that the benefits to be obtained from relevance feedback have been exagerated by past experiments.

Acknowledgements

We are grateful for the availability of the test collections used for our experiments, immediately to Dr. K. Sparck Jones and Professor C.J. van Rijsbergen, and ultimately to Mr. L. Evans of INSPEC and Dr. P.K.T. Vaswani of the National Physical Laboratory.

References

CROFT, W.B. (1981) Document representation in probabilistic models of information retrieval. Journal of the American Society for Information Science, 32, 451-457.

CROFT, W.B. and HARPER, D.J. (1979). Using probabilistic models of document retrieval without relevance information. Journal of Documentation, 35, 285-295.

DAWID, A.P. (1976) Properties of diagnostic data distributions. Biometrics, 32, 647-658.

EVANS, L. (1975a) Search strategy variations in SDI profiles. Report R75/1, INSPEC, London.

EVANS, L. (1975b) Methods of ranking SDI and IR output. Report R75/3, INSPEC, London.

HARPER, D.J. (1980). Relevance feedback in document retrieval. Ph.D. Thesis, University of Cambridge.

HARPER, D.J. and VAN RIJSBERGEN, C.J. (1978). An evaulation of feedback in document retrieval using co-occurrence data. Journal of Documentation, 34, 189-216.

PORTER, M.F. (1980) An algorithm for suffix stripping. Program, 14, 130-137. Also in : van Rijsbergen, Robertson and Porter (1980).

ROBERTSON, S.E. (1976). A theoretical model of the retrieval characteristics of information retrieval systems. Ph.D. Thesis, University of London.

ROBERTSON, S.E. (1977). The probability ranking principle in IR. Journal of Documentation, 33, 294-304

ROBERTSON, S.E. and SPARCK JONES, K. (1976). Relevance weighting of search terms. Journal of the American Society for Information Science, 27, 129-146.

ROBERTSON, S.E., VAN RIJSBERGEN, C.J. and PORTER, M.F. (1981). Probabilistic models of indexing and searching. Information Retrieval Research, Butterworths, London (pp 35-56). Also in : Van Rijsbergen, Robertson and Porter (1980).

SPARCK JONES, K. (1972) A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28, 11-21.

SPARCK JONES, K. (1979a) Experiments in relevance weighting of search terms. Information Processing and Management, 15, 133-144.

SPARCK JONES, K. (1979b). Search term relevance weighting given little relevance information. Journal of Documentation, 35, 30-48

SPARCK JONES, K. (1980). Search term weighting : some recent results. Journal of Information Science,1, 325-332.

SPARCK JONES K. and BATES, R.G. (1977) Research on automatic indexing 1974-1976. BL R&D Report No.5464

SPARCK JONES, K. and WEBSTER, C.A. (1980). Research on relevance weighting 1976-1979. BL R&D Report No.5553.

TITTERINGTON, D.M. et al (1981) Comparison of discrimination techniques applied to a complex data set of head injured patients. Journal of the Royal Statistical Society A, 144, 145-175.

VAN RIJSBERGEN, C.J. (1979) Information retrieval Butterworths, London (2nd edition)

VAN RIJSBERGEN, C.J. , HARPER, D.J. and PORTER, M.F. (1981). The selection of good search terms. Information Processing and Management, 17, 77-91. Also in Van Rijsbergen, Robertson and Porter (1980).

VAN RIJSBERGEN, C.J., ROBERTSON, S.E. and PORTER, M.F. (1980). New models in probabilistic information retrieval. BL R & D Report No. 5587.

VASWANI, P.K.T. and CAMERON, J.B. (1970) The National Physical Laboratory Experiments in statistical word associations and their use in document indexing and retrieval. National Physical Laboratory, Teddington.

YU, C.T.., LAM, K. and SALTON, G. (1982). Term weighting in information retrieval using the term precision model. Journal of the Association for Computing Machinery, 29, 152-170.