CHAPTER 4

Loose ends

The original proposal for this project contained a number of suggestions for theoretical work, some of which resulted in fruitful developments such as experimetal investigations, and some of which did not. The reasons for the latter were various, including lack of time, ideas being superseded by the work of others, and unforeseen or underestimated theoretical problems. We feel that such occurrences are inevitable in a project with a strong theoretical element such as this. But we also feel that these dead or loose ends should be recorded, for the benefit of anyone who wishes to reopen them or pursue the ideas further.

The suggestions in the original proposal which have not been followed up for lack of time are: the incorporation of document length (i.e. number of index terms) into the probabilistic models; the use of information from manually-produced thesauri, and also from previous queries, in such models; and the building of a spanning tree on the document collection (as a development of the idea of document clustering). Those to which we have given some consideration are discussed in the remainder of this chapter.

1 Harter-dependence models

We originally hoped to take the work on the Harter model a stage further than described in Chapter 3 and incorporate some ideas of dependence between terms such as those already considered for the binary case by van Rijsbergen. This idea was not pursued for the following reasons:

a) The development of a Harter-independence model proved rather more difficult, and the model itself proved rather more elaborate, than originally foreseen.

b) The number of dependencies that one might have to consider in a Harter-dependence model is alarming, as can be seen from the diagram in Chapter 3 para. 4.2.

c) Work on dependence models generally can develop in two directions: the explicit estimation and use of dependencies in a relevance feedback strategy (which was perhaps the original aim of the work), or the use of global (not relevance-related) dependencies for query expansion. Recent work has concentrated on the latter (see Chapter 2). Harper's (1980) investigations of the former produced disappointing results. But in the latter case, the dependence model can be treated separately from the weighting model.

d) One overriding problem which has beset probabilistic IR, and which is mentioned on more than one occasion in this report, is the problem of estimation. We have the strong feeling that it will now be more fruitful to concentrate on the estimation problem in the context of the probabilistic models that we now have, than to develop yet more complex probabilistic models.

2 Relevance weights and term frequency

It was suggested in the original proposal that one might seek to formalise the relationship which we know to exist between term frequency and term value, by means of a regression model which would allow the prediction of relevance weights from term frequencies in the absence of relevance feedback information. Subsequently, Croft and Harper (1979) proposed a very simple no-relevance-information model, as discussed in Chapter 3. More recently, Yu, Lam and Salton (in press) have suggested a somewhat more elaborate model.

Neither of these is, strictly speaking, a regression model, and there may still be room for such a model. But it now seems that the problem may be better treated as part of the estimation problem, with the aim of producing a Bayesian estimator which would start from some initial estimate based on term frequency, and incorporate relevance information as it became available. As with the general estimation problem, this would require more statistical expertise than is available on the present project. It is hoped to pursue this idea in the future.

3 Statistical structure of indexed collections

As part of the background to probabilistic models in general and dependence models in particular, it was hoped to investigate the overall statistical structure of indexed document collections. The aim was to develop models which would predict or generate the kinds of structure observed in real collections, and to test them by simulation. For this purpose, some theoretical background was required.

A predominant feature of the statistical structure of collections is the manner in which term-document assignments are far from uniformly or randomly distributed. This fact is reflected in the way in which we can find (by almost any clustering method) clusters of terms which tend to occur in the same documents. Conversely, one can find clusters of documents which tend to contain the same terms. The relationships between these tendencies to cluster and retrieval have been formulated as the Association Hypothesis (Chapter 2) and the Cluster Hypothesis (van Rijsbergen and Sparck Jones, 1973) respectively.

It was hypothesised that these clustering tendencies (of both documents and terms) were in some way logically connected: that is, that a collection exhibiting one type of clustering would have to exhibit the other as well. Preliminary investigation showed this hypothesis to be false, as can be seen from the following argument.

We can define a pseudo-collection which exhibits neither type of clustering as follows:

		Terms			
		a	b	c	d
Documents	A	1	1	0	0
	B	0	0	1	1
	C	1	0	1	0
	D	0	1	0	1
	E	1	0	0	1
	F	0	1	1	0

The term-term co-occurrence matrix is:

a			
1	b		
1	1	c	
1	1	1	d

so there is certainly no clustering of terms. The document-document matrix is:

A					
0	B				
1	1	C			
1	1	0	D		
1	1	1	1	E	
1	1	1	1	0	F

which also does not allow any clustering of documents.

Can we therefore similarly construct a collection which exhibits one form of clustering but not the other? Yes we can, by a very simple modification of the above. If we introduce new terms e, f, g, h which exactly duplicate a, b, c, d respectively:

	a	b	c	d	e	f	g	h
A	1	1	0	0	1	1	0	0
B	0	0	1	1	0	0	1	1
C	1	0	1	0	1	0	1	0
D	0	1	0	1	0	1	0	1
E	1	0	0	1	1	0	0	1
F	0	1	1	0	0	1	1	0

then the co-occurrence matrices become:

$$\begin{array}{cccccccc}
a \\
1 & b \\
1 & 1 & c \\
1 & 1 & 1 & d \\
3 & 1 & 1 & 1 & e \\
1 & 3 & 1 & 1 & 1 & f \\
1 & 1 & 3 & 1 & 1 & 1 & g \\
1 & 1 & 1 & 3 & 1 & 1 & 1 & h
\end{array}
\qquad
\begin{array}{cccccc}
A \\
0 & B \\
2 & 2 & C \\
2 & 2 & 0 & D \\
2 & 2 & 2 & 2 & E \\
2 & 2 & 2 & 2 & 0 & F
\end{array}$$

Thus the document-document matrix has the same structure as before, with no clustering. The term-term matrix on the other hand, shows clear clustering:

Cluster 1 : a, e
Cluster 2 : b, f
Cluster 3 : c, g
Cluster 4 : d, h

as might be expected from the way e, f, g, h were defined.

So the hypothesis is invalid. This does not imply that a particular model will not generate or explain both types of clustering simultaneously, but it does make the specific assumptions on which such a model might be based more critical. This line of work has not progressed beyond this point, but could usefully be pursued in the future. Following earlier work by Cooper (1973) and Griffiths (1978), a simulation model which may have some bearing on these problems is being developed by Tague and Nelson (in press).

REFERENCES

CROFT, W.B. and HARPER, D.J. [1979] Using probabilistic models of document retrieval without relevance information. Journal of Documentation, 35 no. 4 Dec 1979, pp. 285-295.

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

YU, C.T., LAM, K. and SALTON, G. [in press] Term weighting in information retrieval using the term precision model. Journal of the A.C.M.

VAN RIJSBERGEN, C.J. and SPARCK JONES, K. [1973] A test for the separation of relevant and non-relevant documents in experimental retrieval collections.Journal of Documentation, 29 no. 3 Sept 1973, pp. 251-257.

COOPER, M.D. [1973] A simulation model of an information retrieval system. Information Storage and Retrieval, 9 pp. 13-32.

GRIFFITHS, J.M. [1978] The computer simulation of information retrieval systems. Ph.D. Thesis, University of London.

TAGUE, J. and NELSON, M. [in press] Problems in the simulation of document retrieval systems. Paper presented at BCS-ACM Conference, Research and Development in Information Retrieval, Cambridge, June 23-27, 1980.