Chapter 6

Conclusions

6.1 Automatic query expansion

One of the objects of the present project was to investigate the effectiveness of automatic query expansion in a highly interactive environment. The experiment reported in [17] compared a system (referred to as the *QE* system) incorporating automatic query expansion and a system using both AQE and a classification browsing facility (the *FULL* system) with a "dumb" system. These systems accessed a library catalogue database, and the users were paid subjects using the systems under somewhat artificial conditions. Some extracts from the experimenters conclusions follow [17, Chapter 6].

> The results of the evaluation experiment ... show clearly that almost all the subjects who used the query expansion (QE) system found it more helpful than the dumb system. The query expansion facility was felt to be an important factor in this perceived helpfulness. Most subjects also judged the QE system to be as easy or easier [to use] than the dumb system. There was less consensus among subjects who used the full system. About four-fifths of them felt that it was more helpful than the dumb system, but more than half judged the dumb system to be easier... In short, the QE system was highly acceptable, while the full system was considered usable, but markedly less acceptable than the qe. If user acceptability is the main criterion, a system based on the qe but with the interaction improved to counter some of the criticisms ... could scarcely fail to be an improvement on most existing end user reference retrieval systems...

The writers then go on to remark that

> It is perhaps a little surprising that the automatic query expansion function was so successful. It might have been expected that automatic query expansion using system-selected terms from sparsely indexed library catalogue records would be erratic and unreliable. However, the records at the top of the ranking list of those retrieved by query expansion proved to be a somewhat richer source of relevant records than those obtained from the Dewey sequence. The screens of query expansion records were also far more consistent than the classified sequences. Several users complained that classified displays often showed records unrelated to their needs, but no user made such a complaint about query expansion. More than two-fifths of the records chosen by users of the QE system were retrieved using [query expansion]. There were very few critical remarks from users about the quality of the lists of records retrieved, although a number did comment unfavourably on the way in which records they had already seen reappeared in the query expansion lists. It seems that the present implementation of automatic query expansion is functionally satisfactory enough to form the basis of a live installation.
>
> There are two areas where further development may be needed for live use. The first is concerned with heuristics for deciding when to advise the user that it may be worth invoking an expansion search, and the second with the mechanics of selecting and rejecting records to signify their relevance or otherwise. The following two sections give no more than an indication of some of the things which need to be considered.

The authors then discuss the question of heuristics for determining when to offer query expansion. This is further discussed below in 6.1.3.

6.1.1 Query expansion in the library system

Not unexpectedly, results from live use (2) are much less striking than those reported in [17]. In that experiment participants were asked to produce printed lists of the books which they felt would be most useful to a person writing an essay on the assigned topic. The subjects tended to produce lists of between about three and 16 references. However, it is evident that in a large proportion of "real" catalogue uses people are looking for only one or a very few relevant items. In the present project 16 out of the small sample of 34 catalogue users asked about their intentions were looking for "one or two books" or "one particular item". Since automatic query expansion cannot be available until at least one item has been chosen relevant, we may estimate that an AQE facility is of no help in somewhere around 30–50% of searches.

Undoubtedly some searchers using Okapi during the present project did not know that AQE was available. In the experiment in [16] the query expansion option was (very briefly) demonstrated to each subject, although subjects were not *urged* to use it. This undoubtedly encouraged its use. It is not surprising that query expansion was only used in about a third of the searches in which it was available (Tables 2.1, 3.7 and 3.8 give the exact figures).

It was hypothesized that usage of the option might increase with system experience, but this does not seem to be the case (2.2). This suggests a further experiment where users are questioned about their search intentions (preferably by the system) before doing their search. It is likely that there would be a strong positive correlation between the exhaustiveness required and the use of query expansion.

Nevertheless, automatic query expansion was useful to a substantial proportion of users. In answer to the question as to how well did it work when it *was* used there is the evidence given in 2.5, that 47% of the times it was used query expansion led to the selection of one or more additional records. This is consistent with the figure of 60% given in 2.7 for the proportion of searches in which query expansion was not used, but in which it would have lead to the retrieval of additional probably-relevant records. Among all searches by identified catalogue users, AQE accounted for 17% of the records chosen. Restricting to searches where AQE was used, it accounted for 37% of the records chosen (Table 3.10).

6.1.2 Query expansion in searching INSPEC

In INSPEC searches AQE accounted for 15% of the records chosen (32% in searches in which query expansion was used) (Table 3.11). These figures are not very much lower than the corresponding figures for catalogue searches given in the previous paragraph. However, looking at the logs it was evident that there was often a high proportion of false drops among the records resulting from AQE. The reason appeared to be connected with the fact that too many irrelevant terms were being used in the feedback. The fields used as a source of feedback were title, feature headings and descriptors. Looking at the controlled indexing fields it is noticeable that INSPEC has a policy of fairly exhaustive indexing. This often led to a considerable number of terms describing subjects quite unrelated to the user's area of interest being included in the search. Experience with feedback in catalogue searching has shown that irrelevant terms are often in practice "swamped" by good terms, but this is partly due to the very sparse indexing of most catalogue records. Even in the INSPEC searches, if the user chose at least four or five records before invoking AQE, the feedback list would include mainly terms descriptive of the user's interest. However, it was unusual for users to choose as many records as this.

Some informal experiments were done with a view to finding out whether other combinations of feedback fields might be more appropriate. Logged searches were repeated using descriptors alone and headings alone, but neither seemed to give a consistent improvement on the original set of fields. Neither did restricting the number of terms fed back to less than the 24 which were used in the live system bring about a consistent improvement, because the most highly weighted terms would often be relatively rare ones which appeared in just one of the chosen records.

Thus, the problems with AQE in searching abstracting and indexing databases are due to two connected causes: the exhaustive indexing, and the high weight which the Robertson and Sparck Jones formula (equation 2.1) gives to rare terms, particularly when only a few records have been chosen. In [10] Robertson suggests that it may be more appropriate to use a different formula (equation 5.3) for the *selection* of feedback terms. This does not make a material difference when only one record has been chosen, but when more than one has been chosen the *selection value* of each term is mainly determined by the proportion of the relevant records in which the term has occurred. This alternative rule for the selection of feedback terms could be investigated by repetition of logged searches in which more than one record had been chosen before query expansion.

At this stage we can only make a tentative recommendation that AQE should not be encouraged until more than one record has been chosen, and that a term selection rule based on equation 5.3 should be used. More work is needed.

6.1.3 Heuristics for query expansion

In [17] the authors suggested that in a live system query expansion should not be offered except when it appears to be needed—that is, the user has not found enough relevant items—and also it appears likely that it may find some useful material. One step was taken towards the implementation of these recommendations in the present project: when at least three items match the search well then query expansion is not *offered* (although it may be *available*) unless the user has chosen at least two of the original set relevant. It is not safe to go much further here unless the user has given some explicit information about search intentions. As to the system estimating the likelihood of successful query expansion it has not been possible to arrive at any theoretically-based criteria. Experiments could be done, but they would be lengthy and the results might well not be useful enough to justify the effort.

6.1.4 Query expansion: conclusions

Automatic query expansion is a valuable feature in an online catalogue subject search facility. It can be strikingly successful, as in the search on the

City catalogue for "American electoral system" which finds nothing as an ANDed keyword search. It finds one relevant book with Okapi's best match search, and more than 20 books when this one book has been used as the basis for two iterations of query expansion. At the same time query expansion rarely produces the crazy-looking results sometimes obtained by looking at works classified in the neighbourhood of a relevant one.

In the case of abstracting and indexing databases, the results from live use of AQE seem to belie the embarrassing false drops sometimes seen by the experimenters. It was not greatly less successful in searching INSPEC than in searching the catalogue. Nevertheless, it seems likely that some form of user-aided query expansion is to be preferred. There is a note on this below.

There is no obvious way of comparing the usefulness of automatic query expansion with that of other features, but, rather subjectively comparing it with other Okapi features, it is less important than the provision of "best match" searching, which is indispensable. The results from [17] suggest strongly that it is more useful than a classification browsing feature. It is probably more useful than automatic stemming.

It is really only feasible to implement it on top of a search system which does ranked "best match" searching. Very few of the present commercial systems support best match searching. There are some indications that best match systems may become more widely available, but at the same time the move towards standardized search protocols such as NISO Z39.50 [18] may have a tendency to impose a "lowest common denominator" factor on bibliographic retrieval systems.

User-aided query expansion

In 6.1.2 above we discussed the rather unsatisfactory behaviour of automatic query expansion when used in searching exhaustively indexed databases. There is undoubtedly a place for *user-aided* query expansion. This has been used in other experimental systems, and to a limited extent in production systems. It was used for many years in the National Library of Medicine's CITE system [2] but there are no known evaluation results. In one version of the Sheffield University INSTRUCT system [13] the system was able to suggest terms from which users could choose (or suggest their own). The SilverPlatter CD-ROM search system allows users to select terms which they can see in a record and will carry out searches for each individual selected term. The resulting sets may then be combined as the user wishes. This function is not offered in such a way as to encourage its use, and it seems likely that most users would rarely take advantage of it.

The problems with user-aided query expansion are mainly connected with user interaction. If terms are to be displayed for selection they must be complete words or phrases and they must be subject-descriptive. Even if the system searches for stems, it is one or more of the words from which the stem

is derived which should be shown to the user. Further, the system should normally show nouns and noun-phrases only. Classification codes cannot be used unless the system can display the relevant portion of the classification schedules. These considerations suggest that it is terms (complete headings) from the controlled indexing fields which should be offered. Even then some care is needed: LISA, for example, uses headings like "computerised" and "test", which may not be exactly meaningless but are perhaps not suitable for offering in a list of potential search terms. If, as in the SilverPlatter case, it is up to the user to locate and choose terms from displayed records, the user has very little information about the use of each term in the database (only the knowledge that it occurs in the record currently on display). Nevertheless, with a fast and powerful system which can do very speedy "trial searches", and a direct manipulation interaction mode, it should be possible to devise a testable system where the user selects terms directly from displayed records. It would be interesting to do a three-way comparison between two types of user-aided query expansion (selection from list, selection from record) and automatic query expansion.

6.2 Towards adaptive IR systems

The preliminary studies of frequent users outlined in Chapter 4 suggest a number of ways in which the Okapi system might be improved. As regards adapting to individual users, the fact that most of the users studied tended to make identical or similar searches on different occasions (4.4.1) does make it likely that it would be fruitful to have the system store information about users' search language as well as the references which they have judged relevant.

One way of using knowledge about a user's past search terms was given in Chapter 5. A trial implementation of this is in process at the time of writing. It is hoped that it will have the effect of producing a more useful sequencing of displayed records by increasing the weights of terms which have led, in the past, to relevant references for the current user. More trivially, it would be easy and useful for the system to recognise references which have been chosen, and ones which have been rejected, in previous searches. The most obvious way of using this knowledge would be simply for the system to indicate clearly that these references had been chosen on a previous occasion. The system could ask whether the user would like to have them excluded from the display, or to have the previously chosen ones automatically included in a print or email list. The interaction would need careful design to avoid fussiness.

In connection with automatic query expansion, it is clear that the system should try to find out as soon as possible in the session something about the degree of exhaustiveness which the user requires. The most obvious source of this knowledge must be a direct question, but it may be possible

to make fairly reliable guesses from the user's behaviour (time spent looking at records[1], number of related searches in the session, number of references chosen so far, use of query expansion and classification browsing).

[1] Unfortunately, timings cannot be used when systems are accessed over slow or unpredictable networks.