1 Introduction and background

1.1 The project proposal

The research proposal "Improving access in an online public access catalogue by automatic word stemming, spelling correction and limited synonym generation" was approved by the British Library Research and Development Department (BLRDD) and funded under grant SI/G/720.

It proposed to investigate the following recall-improvement devices:

automatic word stemming

synonym and cross reference tables

Soundex-type keys for matching personal names

an *n-gram* technique for approximate matching of words

These were to be applied within a version of the Okapi online catalogue developed under an earlier project.

These devices are not, of course, new. They all have a long history, which is summarised in Chapters 3 - 5. Some of our techniques are probably new, but the main emphasis in the design of the experimental systems was on ways of *applying* and *presenting* the devices in an online catalogue for general users.

1.2 Motivation

An online catalogue is a bibliographic reference retrieval system. It differs from the "traditional" reference retrieval systems such as Lockheed's DIALOG in several ways:

searches are done by untrained end-users rather than by or with the help of intermediaries

the subject coverage is often wide

the subject description in the records is inadequate or even absent.

In traditional information retrieval (IR) intermediaries use truncation and other "fuzzy matching". Inexperienced and casual users will not generally do this, so a system designed for such users should automatically carry out some of these procedures.

There are many cases where a search intermediary would expand a searcher's query to include terms which cannot be obtained by a simple application of grammatical rules and/or recourse to lookup tables. The skills required to do this involve the use of subject knowledge, linguistic knowledge and knowledge of the file(s) being searched, and cannot be automated within the present capabilities of linguistic computing.

However, the automatic inclusion of morphologically related search terms and some tolerance of misspellings should be quite feasible with conventional hardware and software techniques.

1.2.1 Feasibility study

The likely utility of automatic query expansion had been investigated by repeating subject searches from the transaction logs of the original Okapi system (see 1.4).

More than a quarter of the searches would have benefitted from a simple stemming procedure (conflating singular and plural noun forms, "ing", "ed" and "s" verbal endings etc.). Occasionally, there would have been a decrease in precision. About a tenth of the searches contained apparently unnoticed spelling mistakes.

1.3 Staffing

The project head was Neil McLean, Head of Library Services at the Polytechnic of Central London (PCL). The proposer and director of the research was Stephen Walker. Richard Jones was appointed as Research Officer in September 1985, and Nicola Johns as Research Officer/Programmer in June 1986.

1.4 Environment

The project aimed to build on the work of Mitev, Venner and Walker who designed and developed the experimental online catalogue system, Okapi, under the BLRDD and Department of Trade and Industry funded project "Microcomputer networking in libraries". The work of this project, proposed by Neil McLean and Mel Collier, is reported in [1]. Other publications include [2,3,4,5,6,7].

The original Okapi system will be referred to here as *Okapi '84*. It is described in [1]. Okapi '84 was an implementation of an online catalogue search system on a local area

network (Nestar PLAN 4000) of Apple IIe microcomputers. It accessed a file derived from most of the monograph records in PCL's machine readable catalogue. The emphasis in Okapi '84 was on providing easy interaction - usability at sight - combined with an effectiveness at least comparable with other search systems. Okapi '84 was installed in the Riding House Street library, the largest of the polytechnic's site libraries, where up to four terminals were in daily use for nearly two years.

The same hardware and the same test site were used for the current project. The bibliographic file used for the current project is very similar in structure to that described in Chapter 4 of [1]. The only substantial difference is that Library of Congress Subject Heading (LCSH) form, content and geographical subdivisions are included.

Several generations of search and indexing software were constructed during the project. Much of the Okapi '84 search system was completely redesigned and rewritten, although all the programs rely heavily on the substantial library of utility subroutines which was written by Stephen Walker and Gill Venner for the earlier project. Of the higher level code, the record formatting and display routines remain largely unchanged.

1.5 Historical summary of the project

Because of shortage of time on the earlier project little formal evaluation of Okapi '84 had been done. Transaction log data was collected continuously from Okapi '84 until May 1986. Much of the early work on the current project involved examination and analysis of the log data in order to form a pragmatic basis for any new search system.

We wrote several versions of a Soundex-type procedure, both for personal names and for textwords, and made informal comparisons of their efficacy. The procedure eventually chosen is documented in 6.4.2 and Appendix 1. A slightly modified implementation of Martin Porter's suffix-stripping procedure (6.2.7 - 6.2.11) was tested. Both of these devices were incorporated in an 'intermediate' Okapi system which was demonstrated at the conference 'Online Public Access to Library Files' organised by the Centre for Catalogue Research at Bath University in April 1986.

The intermediate system contained a simplified and probably improved specific item search function. Author and title index displays, which were provided under some circumstances in Okapi '84 but were rarely used, were replaced by sequenced displays of brief bibliographic records. The precision of the implicit author/title acronym search function was improved by validating the resulting set by combining it with a word from the title or the author as entered by the user. There was a Soundex-

type index of personal surnames, which could lead to the display of a selection of possible matches in the case of a failed personal name look-up.

The intermediate system was never developed to the point where it was robust and finished enough for public use (although it stood up fairly well to the attention of librarians at Bath). Its specific item search will probably be used for future public versions of Okapi, and the subject search forms the basis of the systems described in this report. The intermediate system is described in [7].

The unfinished subject search facility of the intermediate system included some of the features used in the systems described in Chapters 6 and 7. The subject index consisted of words from titles, subtitles, subject headings and other subject-rich fields of the MARC record, and also Dewey numbers. All words were subjected to the Porter algorithm, and were combined using a combinatorial (AND/OR) technique with inverse term frequency weighting as described in 6.5. Records were output in decreasing weight order. There was a small cross-reference list enabling terms to be treated as synonymous: the list included CHILD and CHILDREN and some other irregular plurals and alternative spellings.

Richard Jones carried out a substantial analysis of failed searches, some of which was published in [8]. We constructed an inverted file of some 6000 subject searches submitted to Okapi '84 and used it to obtain a collection of spelling mistakes and to suggest entries for an automatic cross-reference table (Chapter 6 and Appendix 5).

During the Summer and Autumn of 1986 we designed and wrote new subject search and indexing programs. Two versions of the search program were produced. One - the "experimental" system (referred to as *EXP*) - incorporates full two-level stemming, a substantial look-up table of phrases and equivalence classes of related terms and a spelling correction procedure. The other program - the "control" or *CTL* system - has "weak" stemming but no look-up table and no spelling correction. It is unlikely that users would notice any difference between the two programs unless a search carried out on one was immediately repeated on the other.

There is also a third system, called *OSTEM*, which incorporates none of the recall-improvement devices. This was only used for the repetition of users' searches by the experimenters.

In November Okapi '86 (subject search only) was installed at two terminals in the Polytechnic's Riding House Street Library. The systems were alternated daily between the two terminals. After a trial period, during which the system was slightly improved, we began to collect data from live use. About 120 people were interviewed after they had performed subject searches, and full transaction log data was gathered from about 1100 searches (about 600 sessions).

After formal data collection had finished, the *EXP* system was left on site to collect further transaction log data. At the time of writing we have collected logs of some 7700 searches, and have done a certain amount of analysis of this data where necessary to supplement the original set.

In parallel with the work outlined above, we were also designing and developing programs for use in a second project (on the use of relevance feedback in online catalogues - SI/G/765).

While data analysis was being carried out, Nicky Johns resigned. Since it would not have been possible to obtain a suitable replacement on what would have been a very short contract, we suspended work on the relevance feedback project so as to be able to complete the present project as quickly as possible.

1.6 The report

Chapter 2 gives a brief survey of some of the problems in subject searching in online catalogues. Chapters 3, 4 and 5 survey ways of implementing the recall-improvement devices in IR systems in general as well as in catalogues.

The catalogues which we designed for this project are described in Chapters 6 and 7. Chapter 6 describes them from the inside and Chapter 7 from the outside - how the user sees them. These two chapters are interdependent. Their logical order may be 6 followed by 7, but their psychological order may be the reverse.

Chapters 8 and 9 cover data collection, evaluation, results and conclusions.

References

1 MITEV N N, VENNER G M and WALKER S. *Designing an online public access catalogue : Okapi, a catalogue on a local area network* (Library and Information Research Report 39). London : the British Library, 1985.

2 MITEV N N and WALKER S. Intelligent retrieval aids in an online public access catalogue : automatic intelligent search sequencing. In: *Informatics 8 : Advances in Intelligent Retrieval. Proceedings of an Aslib/BCS Conference, Oxford 16-17 April 1985.* London : Aslib, 1985.

3 WALKER S. The free language approach to online catalogues. In: *Keyword Catalogues and the Free Language Approach*. Edited by Philip Bryant. University of Bath, Library, 1985.

4 MITEV N N, VENNER G and WALKER S. OKAPI : an online public access catalogue on a microcomputer local area network. In: *Online Public Access to Library Files. Proceedings of a Centre for Catalogue Research Conference held at Bath University 3-5 September 1984.* Edited by Alan Seal. Oxford : Elsevier, 1985.

5 MITEV N N. The user interface in an online public access catalogue. In: *Computer Assisted Information Retrieval. RIAO 85, International Symposium organised by the Centre de Hautes Etudes d'Informatique Documentaire and the Institut d'Informatique et de Mathematiques Appliquees de Grenoble, 18-20 March 1985, Grenoble, France.* Paris : CNRS-IMAG, 1985.

6 VENNER G, WALKER S and MITEV N. Okapi : a prototype online catalogue. *VINE 59*, July 1985, 3-13.

7 WALKER S. OKAPI : evaluating and enhancing an experimental online catalogue. *Library Trends*, Spring 1987 (to be published).

8 JONES R. Improving Okapi : transaction log analysis of failed searches in an online catalogue. *VINE 62*, May 1986, 3-13.