Chapter 1

Introduction

1.1 Introduction

This report is the fifth in the series describing experiments with the British Library Research & Development Department's (BLR&DD) **Okapi** bibliographic information retrieval system [7, 16, 5, 17]. The report describes work at the City University Centre for Interactive Systems Research towards setting up the Okapi system as a generalised facility for the evaluation of interactive information retrieval systems, together with some results from two more specific experiments. Again, this work has been funded by BLR&DD.

The report has been compiled from work by a number of people, and was edited by Stephen Walker. After the introduction, describing the installation of an Okapi system at City University, the following two chapters are largely concerned with the use of automatic query expansion (AQE) in live use. Chapter 2 gives some evaluation results on the use of AQE, mainly in connection with library catalogue searching. Chapter 3 contains more general statistical results, mainly derived from transaction log analysis, on the use of the catalogue and other databases. The next two chapters are concerned with steps towards the design of information retrieval systems which adapt to their users. Chapter 4 outlines some points from examination of the transaction logs of searches by a number of frequent users of the Okapi system. Chapter 5 discusses some work towards information retrieval systems which learn about the search language of individual users and user groups. Some of the chapters are more or less self-contained, and consequently there is a certain amount of repetition of material.

1.2 Background

In July 1989 the Okapi projects moved from the Polytechnic of Central London (PCL) to the Centre for Interactive Systems Research at City University.

These projects have all been concerned with research and development in the area of bibliographic retrieval systems for direct use by end users. Most of the work has been funded by the British Library Research and Development Department (BLR&DD).

The last of the PCL projects, reported in [17, 14], consisted mainly of an investigation under controlled conditions of the use of automatic query expansion in subject searching of library catalogue data.

The present project, also funded by BLR&DD, was set up with the following general objectives:

- to build a facility which would allow extensive evaluation of many aspects of information retrieval (IR) system design, in an operational environment, with real users with real information needs
- to build associated evaluation tools;

together with two more specific objects:

- to investigate the effectiveness of automatic query expansion (AQE) in a highly interactive environment
- to conduct an initial investigation into the possibilities and problems of user modelling and profiling, and making use of this information in retrieval.

The first two objects above are discussed in this Chapter, and the two specific experiments are described in Chapters 2, 3 and 5.

1.3 Development of Okapi at City University

1.3.1 The PCL Okapi system

This was the system described in [17], which itself had been developed from earlier Okapi systems. It featured

- ranked output, "best match" keyword searching
- automatic stemming and spelling normalisation
- automatic cross referencing
- an automatic query expansion facility
- a classification browsing facility.

It had been designed for a comparative evaluation of automatic query expansion and classification browsing under controlled conditions in which participants undertook a number of assigned searches chosen from groups of topics. Automatic query expansion is described in 2.1.1 below.

The programs consisted of about 50,000 lines of C and operated on Sun workstations. Users accessed the system via dumb terminals attached to the Sun through a terminal server.

To work towards the generalised evaluation facility introduced above it was necessary to embed the Okapi system inherited from PCL in a modular and highly parameterised framework, making it more of a "generalised" retrieval system. Previous systems had been quite tightly bound to one specific database (PCL library catalogue) and contained site-specific messages. To change databases, or to change the messages, it was necessary to amend the source code and recompile. The facilities for database and index updating were primitive.

1.3.2 Parameterization

A system of parameters was evolved. These include

- Database parameters—one for each database. These contain information about the structure of a database, including the following.
  - The location of the bibliographic and index files.
  - The number of fields in each record, and the type of data in each field.
  - The number and nature of the indexes.
  - The stemming procedure to be used (Okapi permits various degrees of automatic stemming and spelling normalisation).
  - Okapi systems can make use of a database of linguistic knowledge [16, Chapter 6]. This database contains stop words and phrases, synonyms, abbreviations, prefixes, "go" phrases etc. A parameter determines which linguistic database is to be used with which index.
  - The message set to be used (obviously messages should be dependent on whether a database consists mainly of monograph records or of journal article citations).
- Record display parameters. These determine the labelling and layout of brief and full record displays for a database.
- Lists of fields to be used for various types of search and for the extraction of terms for automatic query expansion.

- Indexing parameters, describing which fields are to be indexed, the way in which terms are to be extracted from each field, and the grouping of fields for indexing and retrieval purposes (for example one might want to provide a "subject" search which accessed keywords extracted from title and subject heading fields of the bibliographic records).

1.3.3 Updating procedures

The investigation of IR systems under fully "live" conditions needs, among other things, databases which are reasonably current. A catalogue database, for example, should ideally be updated "on the fly" or at least nightly. Hence a substantial amount of work was put into designing procedures for database maintenance.[1] These included programs for selective updating of bibliographic files and their indexes.

1.3.4 Databases

An important feature of the project was to be the evaluation of Okapi accessing non-catalogue databases. Permission was sought and obtained from the suppliers of the INSPEC and LISA (Library and Information Science Abstracts) databases to make them available on the Okapi system free of charge for a limited period to staff and students of City University.

Although many catalogues are obtainable in more or less[2] standard UK-MARC format, there is no standardisation among abstracting and indexing databases. For every such database, programs have to be written to convert it to whatever format the local system requires. A MARC to Okapi conversion program already existed, but INSPEC and LISA conversions had to be written. In the case of Okapi the local format is fairly simple (still substantially as described in [7] in 1985), but the INSPEC and LISA to Okapi conversion programs consist of some 1200 lines of code each.

INSPEC supplied monthly tapes of sections C (computer science) and D (information technology) and the Okapi INSPEC database grew until it contained some 94,000 records covering October 1989–December 1990.

The Okapi LISA database was less current. It was downloaded, with some difficulty, from the CD-ROM edition of the database. It was not updated during the course of the experiments, and contained 76,000 records

[1] For a number of reasons these procedures have not yet been fully implemented. For a time it was not possible to output MARC records from City's CLSI library system. Then there is the general problem that if the experimenter needs to repeat real searches, perhaps trying the effect of variations in the search strategy, it is necessary to retain a copy of the database as it was at the time the searches to be repeated were made.

[2] *less* rather than *more*

from 1976 (when abstracts were first included) through to 1989 and 1990 (both incomplete).

The Okapi version of both the above databases contains most of the textual content of the source files, although packed into fewer fields and occupying considerably less storage.

The source for the library catalogue was obtained by downloading MARC tape files from City's CLSI integrated library system. The Okapi version of the catalogue was updated only twice during the experiments. It finally contained about 155,000 records (titles, not copies), about two-thirds of which contained subject-descriptive material in the form of Library of Congress Subject Headings (LCSH) and/or Precis verbal feature headings.

1.3.5 Logging

As with all previous Okapi systems there is an extensive automatic user logging facility which was used throughout the experiments. Okapi logs have always been something of a compromise between being understandable by humans and being suitable for accurate interpretation by computer programs. Unfortunately there were several changes to the log content and layout during the course of the experiments as shortcomings became evident. Okapi logs are described and illustrated in B.

In addition to the user logs, every time the search program is run it produces a system log giving details of program version and details, and containing system error and other diagnostic information about the internal workings of the search program.

1.4 Installation and use of Okapi at City

1.4.1 Installation

An Okapi system was installed on the Department of Information Science's Sun SPARCstation 330, which had 16 MB of memory and about 900 MB of disk storage. During the first half of 1990 a campus-wide ethernet system was installed at City, and by May of that year the Sun was connected to the ethernet.

Library system

In mid-May 1990 an Okapi terminal (a PC with ethernet card and PCTCP networking software) was set up in the university library. This terminal accessed the library catalogue only. It provided the system illustrated in A to any users who wished to try it rather than the adjacent CLSI terminals.

In the early days of the City ethernet there was a good deal of downtime due to network problems, but on most days for the rest of 1990 the library terminal was available most of the time the library was open. Later in 1990 a second Okapi terminal was installed in a seminar room in the library. This was used for the interview sessions described in 2.3.3.

Networked system

A similar system, using the same search program run with different arguments and parameters, was made available to registered users over the campus network from end May 1990. Some of the minor differences between the networked system and the library system are illustrated in Appendix A.

1.4.2 Use

Library users

Use of the library terminal was steady but not particularly heavy—there appeared to be enough CLSI terminals to eliminate queuing at all but a few peak times. From installation to the end of the 1990 there were about 1800 user sessions at this terminal, about 900 of which were by identifiable users. All sessions were logged in considerable detail (see 1.3.5 above). It is scarcely possible to require people to log in to an "open" terminal accessing a library catalogue, so it was necessary to find some way of getting round the problem of identifying user session boundaries—some experimenters have resorted to discarding apparent sessions which follow the preceding session within two or three minutes. This was of more than usual importance in the present project because the experimenters wished to study longitudinal changes in the behaviour of individual users. With this in mind, users were encouraged to identify themselves by keying in their library card numbers (see Figure A.1 for the simple but reasonably effective procedure used to entice people to enter their card numbers).

Network users

Staff and students were invited to register for use of Okapi by handouts on the library counter and by publicity in a number of faculties of the university. The registration form asked for users' library card number so that sessions by people who used the system both from the library terminal and over the network could be attributed correctly. Users were issued with advice on connecting to the Okapi machine and given a username for logging in. No passwords were allotted[3] although there was a facility within the system for peopl̇e to password their accounts on the Okapi machine.

The first network users were registered May 30 1990. By the end of the year there were about 80, most of whom had used the system at least once. During this period there was a total of 682 user sessions, of which 323 (47%) were on INSPEC, 284 (42%) on the City catalogue and 75 (11%) on LISA.

[3] This was a mistake. There is evidence that at least one user allowed an unregistered person to access his account. If this had happened on a large scale it would go some way towards invalidating the longitudinal studies.