1 INTRODUCTION

1.1 The Okapi projects

The work described in this report is the third major project in a continuing series concerned with research and development in the area of bibliographic retrieval systems for direct use by end users, and with subject access to online catalogues in particular. These projects have become known as the Okapi projects after the name of the original prototype system of 1984. Most of the work has been funded by the British Library Research and Development Department under a series of grants since 1982. Up to the present the work has been carried out at the Polytechnic of Central London (PCL), but it is expected that future projects will be done at City University, London and at the University of Bath. Published material on the Okapi projects prior to the present one includes [JONE86, JONE88, MITE85a, MITE85b, MITE85c, WALK86, WALK87a, WALK87b, WALK88, WALK89a].

The first Okapi project (1982-1985) comprised the design and development of a networked microcomputer-based online catalogue. This was originally intended to be part of an investigation of library uses of networked microcomputers. Because of the interests of the investigators, it became an attempt to develop a prototype end user retrieval system which aimed to combine instant usability with a relatively high degree of effectiveness. The resulting system, Okapi '84, was installed in one of the PCL's site libraries, giving access to a database of about 30,000 monograph records, and a limited amount of evaluation was done. Okapi '84 offered users just two types of search, which it offered as "Specific books" and "Books about something". The specific item search used some fairly elaborate techniques "behind the screen", while appearing rather simple to the user. Users were encouraged to fill in whatever they knew of the title and/or author of the sought item, and the system would take courses of action which depended on the information provided and on the results of initial searches. The subject search, on the other hand, was extremely simple. It simply looked for records which contained as many as possible of the words of the user's freely expressed search. This project was reported by Mitev, Venner and Walker in [MITE85b].

Such evaluation data as was collected suggested that Okapi '84's specific item search was not particularly successful, whereas the very simple subject search was relatively effective. Subsequent work has been mainly concerned with subject access; most of the later Okapi systems have offered subject searching only.

1.2 Notes on terminology

The following notes introduce our use of a few technical terms.

An *information retrieval* (IR) system is a computerized system for the retrieval of bibliographic references. We only consider *interactive* IR systems - ones where users obtain results while they

wait. Mostly, we only discuss *end user* interactive IR systems. These are systems which are intended for and mainly used by the people who want the information (library patrons, researchers). An *online catalogue* (online public access catalogue, OPAC) is an end user interactive IR system by means of which library users try to find references to material held in a library. The references retrieved with IR systems are often referred to as *documents* (although this is precisely what they are not).

A *query* usually means the search statement which a user types into an IR system. *Query modification* is alteration of a query by the user or the system or both. *Query expansion* is a little narrower than query modification, suggesting the addition of terms to the query. There are many possible types and varieties of query expansion. Originally, we used *automatic* query expansion to mean expansion which is applied without any user intervention, for example the addition by the system of morphologically related terms to the user's query; *semi-automatic* query expansion applied to systems where terms were selected automatically but the user chose whether to invoke an expanded search. To conform with other peoples' usage, we now refer to the latter as *automatic* query expansion, and semi-automatic expansion applies to cases where the user can intervene by selecting or rejecting, and possibly ranking, terms which the system has suggested.

Relevance is a nebulous relation between a document, a query and a user. The terms *pertinence* or *usefulness* or *utility* may be used when there is an emphasis on the satisfaction of the current needs of a specific user. *Relevance information* is information about the relevance or non-relevance of documents to a query. In principle there can be any number of degrees of relevance - a document can be "probably relevant", "fairly relevant" or "almost certainly not relevant", but in the work described here we only used a binary division into something like "likely to be relevant" and "non-relevant". *Relevance feedback* is the application of relevance information obtained in the early stages of a user's session on an IR system to modify the behaviour of the system during the remainder of the session. In particular, relevance feedback can be used in the implementation of query expansion. Automatic query expansion based on relevance feedback is the main topic of this report.

The *precision* of a search is the proportion of retrieved records which are relevant. The *recall* of a search is the proportion of the relevant records which are retrieved. Precision and recall have been widely used as performance measures for IR systems, particularly in laboratory tests of batch systems. It is not easy to apply them in experiments with interactive systems accessing databases of realistic size.

1.3 Subject access in online catalogues

We said above said that the Okapi subject search was relatively good, but this must be qualified by emphasizing that it was only good relative to other end user retrieval systems. There is a wide measure of agreement that most systems are less effective, efficient and helpful than they ought to be.

The primary subject access functions in online catalogues fall into two categories, heading or keyword. Many systems also have a search by

classification code or shelfmark, but there are numerous reports that a shelfmark search option attracts little use. Many systems offer more than one of these access methods.

In response to the user's search, systems providing access by headings usually show an alphabetically ordered display of headings around that which is "nearest" to the search. The user may then browse headings in alphabetical order or choose to see the bibliographic records which are indexed by a given heading. Fig 1.1 illustrates one of the more elaborate heading-based searches, the Library of Congress Subject Heading (LCSH) search in Ohio State University's online catalogue.

Fig 1.1 Subject heading display in the Ohio State University catalogue

```
================================================================================
COMMAND:  sub/nutrition

RESPONSE:

TBL/ ITEMS ------------------------SUBJECTS---------------------------- SAL/

1     764 Nutrition                                                        1
2       7 Nutrition--ABSTRACTS
3       3 Nutrition--Abstracts--Periodicals
          Nutrition--Aging effect
5             SEARCH UNDER:  Aging--Nutritional aspects
6       5 Nutrition and dental health                                      6
7       2 Nutrition and dental health--United States
          Nutrition and state
9             SEARCH UNDER:  Nutrition policy

MORE: PS+     BACK: PS-           FOR TITLES, ENTER: TBL/number
FOR NOTES OR RELATED SUBJECTS (ONLY WHEN NUMBER IS AT RIGHT), ENTER: SAL/number
================================================================================
```

Systems providing subject access by keyword vary greatly with respect to the source of the keywords and the way in which users' input is matched against the indexes. A few of the older systems will process only a single keyword, but most of the more recent ones work by extracting the words from the user's search and performing an implicit AND operation. Fig 1.2 shows part of a subject search for "Care of the terminally ill" in the University of California's MELVYL catalogue. Here, the words "care", "terminally" and "ill" have been looked up in an index of words automatically extracted from LCSH. The records which the system has found probably contain the LCSH "Terminally ill--Home care". This is an example of a search where keyword access has found some relevant records but heading access would have failed. Other keyword systems often use title words in addition to words from LCSH, and even words from series, corporate authors and notes fields.

Fig 1.2 Result of a keyword subject search on the MELVYL catalogue of the University of California

```
================================================================================
Search request: FIND SUBJECT CARE OF THE TERMINALLY ILL
Search result:  11 records at all libraries

Type HELP for other display options.

 1. BENSON, Hazel B.                The dying child : an annotated... 1988
 2. BROOKS, Charles H.              Cost savings of hospice home care... 1983
 3. CHILDREN'S HOSPICE ADVISORY PANEL. Conference (1984 : Washington, D.C.)
      Children's Hospice Advisory Panel Conference report : December... 1984
 4. The Dying human. 1979
 5. Home care for the dying child : professional and family perspectives. 1976
 6. LACK, Sylvia A.                 First American hospice : three... 1978
 7. LITTLE, Deborah Whiting.        Home care for the dying : a... 1985
 8. NATIONAL SYMPOSIUM ON COPING WITH CRISIS AND HANDICAP (1979 : Boston,...
      Coping with crisis and handicap. 1981
 9. SPIEGEL, Allen D.               Home health care. 1987
10. SPIEGEL, Allen D.               Home healthcare : home birthing to... 1983
11. Terminal care at home. 1986

-> display 7 full
================================================================================
```

There are many problems with both types of subject access. There is evidence that a majority of users find the more direct keyword approach preferable. There would also seem to be a greater chance of users' terms finding a match when the index language is not limited to the often stilted and out-of-date terminology of controlled subject headings. However, there seems to be little hard evidence that a keyword approach produces better results than access via headings. An important factor in the United Kingdom is that a large proportion of libraries do not have subject headings attached to their bibliographic records, so the choice is often between access by keywords from titles and access via classification. Most of the more recent systems seem to prefer keyword access, and all the Okapi systems have used keywords as the primary means of subject access.

Whichever access method - heading or keyword - is used, it is almost always found that a large proportion of searches retrieve no records at all. Wildly varying figures for the failure rate, ranging from 20% to 80%, have been quoted from various investigations under different experimental conditions, databases and search systems. Markey gives a selection of results in [MARK84]. Walker and Jones found that 34% of about 1000 searches by undergraduate users of a social sciences library would have found no records if submitted to a keyword system where terms are combined using an implicit AND operation [WALK87b, p121]. The University of California catalogue monthly statistics for March 1989 show that 30.4% of searches retrieved nothing. This figure includes all types of search.

A considerable proportion, perhaps a quarter, of these failures are due to spelling or typing mistakes. This should not be a serious problem, because the majority of such mistakes can easily be dealt with by designing the system so that it refuses to process the search until any

unknown words have been negotiated with the user, but there are few systems which do this. We are left with a substantial proportion of subject searches which would still fail, typically between 25% and 40%. The situation is really somewhat worse than this, because among "successful" subject searches there are many which are too general to be useful. It appears that many users quickly learn that searches are likely to fail unless they are broad, and this may partly account for the substantial proportion of searches like "accounting", "film", "statistics". Walker and Jones [WALK87b] found that a quarter of searches consisted of only one word. Many of these retrieve a great many records. On the MELVYL catalogue in March 1989 the mean number of records retrieved (all types of search) was 139, with an astonishingly high standard deviation of 1452. These figures include searches which retrieved no records. The University of California has an extremely large catalogue containing more than 5,000,000 titles.

1.4 Subject access in Okapi '86

The second Okapi project, completed in 1987, investigated several ways of increasing the recall of searches and reducing the proportion of search failures in end user systems. The methods used included computer-assisted spelling correction, automatic word stemming and automatic cross-referencing. The project and its subject search system Okapi '86 are described by Walker and Jones in [WALK87b]. Each of the above mentioned techniques was reasonably successful at increasing recall. Eighty-three percent of 600 searches collected from live use of the system found at least one record which the system judged to "match the search quite well" [WALK87b, p121]. At least as important as the recall-improvement devices is the fact that the Okapi systems use a method of search term combination which is weaker than the usual AND operation. All versions of Okapi, including the ones developed for the project described in the present report, combine terms on a "best match" basis. Terms are weighted in accordance with their relative frequency in the indexes, rare terms being given a higher weight than common terms. The weight given to a retrieved record is the sum of the weights of the terms common to the record and the query. The result of a search is a list of records ordered by weight, with the best matching records at the top of the list. Term weighting schemes are discussed in 2.2, and the Okapi term weighting and term combination procedure is described in Chapter 6 of [WALK87b].

1.5 Query expansion

The Okapi '86 system was reasonably effective at reducing the proportion of searches which fail completely, but there remain many searches which do not work as well as they should. It must be one of the primary aims of document retrieval system designers to produce systems which enable users to make searches which are as exhaustive or as selective as they wish. Ideally, every document on a topic should be retrievable without undue effort. In practice this is only the case for very small databases (but it is said that there are users of the British Library who go through the entire General Catalogue). All large bibliographic databases contain many items which are unlikely to be retrieved except by chance: items without subject headings, with metaphorical titles, misclassified.

As well as enabling exhaustive searches, retrieval systems should help users to refine or focus their searches. Many end user queries as

initially submitted to the system are not a good representation of the subject the user is "really" looking for. Examples are "Intelligence" for the influence of heredity and culture on intelligence, or "Child development" for the effect of the mother on the development of the child. Both of these searches are likely to find some relevant records without too much effort, unless the database is very large, but there will be many records which they do not find. Searches like "Britain as a developing country" for the economic development of Britain in the 18th century are so inappropriate that they are unlikely to work. Such queries are not uncommon. Some of them result from lack of subject knowledge, some from misapprehensions about what the system is and what it does. Nevertheless, a proportion of "dubious" queries will result in the user finding one or a few relevant records. This is more likely to be the case in a best match system like Okapi than in an implicit AND system.

Previous Okapi work has concentrated on the development of subject search systems which aim to maximize the likelihood of finding *something* relevant. Okapi '84 and '86 were relatively effective but, like most other end user retrieval systems, they were also dumb and unhelpful. Charles Hildreth remarked that they were also boring. Three ways of improving end user retrieval systems were considered. The first of these applies mainly to online catalogues, which typically suffer from records which do not contain adequate subject descriptive material. It is the enhancement of subject description in the bibliographic records. It is likely that a catalogue with records enhanced by means of data from publishers' descriptive material will soon be set up and evaluated at the University of Bath. The second area of research was concerned more with the user than with search functions or record content. This was to work towards a system which adapted its interaction according to its picture of the user's needs, aptitude and experience. A request for funding for a project in this area was not supported. Finally, there is the subject of the work reported here. This is concerned with helping users who have already found some relevant material to obtain further, related items which have not necessarily been retrieved by the original query.

There are a number of ways of tackling this. A simple and obvious one, which, surprisingly, has rarely been provided, is that of offering items classified in the same area as a relevant item. One of the few commercially available retrieval systems which offers this type of browsing in classified sequence is the BLCMP library system. Another way of extending or focusing a search is to branch to records with the same subject heading or other descriptors. At least one of the commercially available library systems (Dynix) provides this option, albeit rather clumsily. These systems and others are mentioned again in 2.4.1. We do not know of any evaluation of the effectiveness of such "pivot" techniques. Perhaps more attractive than using a single pivot is the technique of using keywords selected from relevant records as new search terms, supplementing the original query. This is usually what *query expansion* means in this report. Because of the paucity of subject description in many catalogue records, it may be important to make use of as much as possible of the available information.

1.6 Feasibility studies

There are several existing systems which offer semi-automatic query expansion. Again, three of them are briefly described in 2.4. There are

no known evaluation results, and none of the systems has quite the degree of instant usability which is regarded as essential in Okapi systems.

Some informal testing of query expansion techniques was carried out manually in 1985. Late in 1986 a modified version of Okapi '86 was made, which allowed a Dewey classification pivot search, and would accept relevance judgments and would extract terms (including Dewey numbers) from relevant records. The user could then select from these terms and instruct the system to perform a new search. Extended tests were done on this system, using a technique in which real searches were repeated by an experimenter. Query expansion using terms automatically extracted from selected records appeared extremely promising, even with no term selection by the user. Records classified near selected records were sometimes useful, but often most or all of them were far removed from the original search. The research proposal for the present project [WALK85] suggested the use of query expansion in which the original query terms were to be supplemented only by Dewey numbers extracted from relevant records. The preliminary experiments just described showed that this was unlikely to be an outstandingly useful technique, but that query expansion using subject and title words, selected by the system not the user, as well as Dewey numbers, was certainly worth trying with real users.

1.7 Development of the experimental system Okapi '88

Because query expansion using Dewey classification codes alone did not look very promising most of the development work was put into producing a practical implementation of a system providing automatic query expansion based on terms extracted from relevant records.

The retrieval systems - Okapi '88 - which were developed are described and illustrated in some detail in Chapter 3. The general appearance of the screens and also the structure of the programs and data are rather similar to previous Okapi systems, although the hardware (Sun) and operating system (unix) environment are completely different. In particular, Okapi '88 incorporates the recall improvement devices developed for Okapi '86 [WALK87b], with the exception of spelling correction.

It was originally intended that the systems should be evaluated in live use in a library, but this was not practical. Some experiments were done using volunteer subjects under controlled conditions. These are described in Chapter 4.