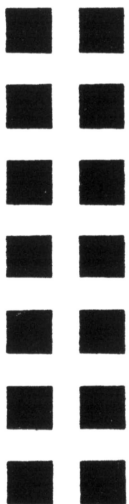
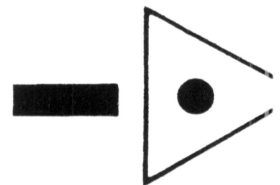
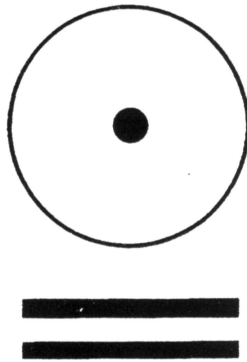


IMPROVING SUBJECT RETRIEVAL IN ONLINE CATALOGUES



Stephen Walker and
Richard M Jones



Improving subject retrieval in online catalogues

1. Stemming, automatic spelling correction
and cross-reference tables

Stephen Walker and Richard M Jones

with contributions by **Nicky Johns**

The Polytechnic of Central London

1987

British Library Research Paper 24



695
F.W.34
1987
V.1

British Library Cataloguing in Publication Data

Walker, Stephen

Improving subject retrieval in online catalogues.—(British Library research paper, ISSN 0269-9257; 24).

1 : Stemming, automatic spelling correction and cross-reference tables

1. Catalogs, On-line

I. Title II. Jones, Richard M. III. Series
025.3'13 Z699

ISBN 0-7123-3129-8

Stephen Walker worked in management services in industry, and as a mathematician, before moving into information science. He has taught information science and computing in several institutions, and worked on the design of information systems. Since 1982 he has worked exclusively in the field of online catalogues, both at the Polytechnic of Central London and as Technical Consultant to Swalcap Library Services Ltd, Bristol.

Richard Jones is a librarian who worked for three years in the Library of the University of London Institute of Education before joining the Okapi projects. He surveyed the operation of the Geac online catalogue at the Polytechnic of the South bank for his Masters' dissertation.

© The British Library Board 1987

The opinions expressed in this report are those of the authors and not necessarily those of the British Library.

SI/G/720

British Library Research Papers are published by the British Library and distributed by the British Library Publications Sales Unit, Boston Spa, Wetherby, West Yorkshire LS23 7BQ, UK. In the USA and Canada they are distributed by Longwood Publishing Group Inc, 27 South Main Street, Wolfeboro, New Hampshire 03894-2069, USA. In Japan they are distributed by Kinokuniya Co Ltd, PO Box 55, Chitose, Tokyo 156.

Printed in Great Britain by the British Library.

Preface

This report describes research into practical ways of improving subject access in online catalogues. This is part of a continuing programme which began in 1982 with the original Okapi project [1]. This is not a "final" report. The next phase, on the use of relevance feedback, will be published in 1988.

The work described here is concerned both with librarianship and with computing, particularly linguistic computing. Some knowledge of online catalogues and of computing and library jargon is assumed. We refer, for example, to "generations" of online catalogues, to "phrase" and "keyword-type" catalogues, subject searching and specific (or known-) item searching, to headings, entries, fields and records. The most important source for background material to this report is [1]. For online catalogues in general, Hildreth's 1983 monograph [2] and his ARIST survey [3] are recommended. The most useful work on subject searching problems is Markey's book [4].

Acknowledgments

This work would not have been possible without the help and support of the British Library Research and Development Department. Particular thanks are due to Maureen Grieves and to Derek Greenwood, who is now our project officer.

We are also most grateful to Neil McLean (Project Head and Head of Library Services, Polytechnic of Central London), to Maura Coghlan (site Librarian of our test library) and to Dave Roberts and his staff at the Polytechnic's Computing Services.

Others at PCL who have helped us include Winifred Abbott (Technical Services Librarian), Jacky Conroy (Systems Librarian), Pat Manson (Information Officer, Library Technology Centre) and Penny Pope (Chief Cataloguer). Many other people have helped us with advice, criticism and encouragement. They include Philip Bryant (Director of the Centre for Catalogue Research), Charles Hildreth (READ Ltd), Karen Markey (University of Michigan), Martin Porter (Scott Polar Research Institute) and Peter Willett (Sheffield University). Gill Venner (Plymouth Polytechnic, late of the Okapi team) helped with the rewriting of some of the Okapi programs.

Credits

It would not have been possible to design and program the experimental catalogues used in this project without the foundation provided by the original Okapi system. This was designed and written by Nathalie Mitev, Gill Venner and Stephen Walker. The additional programs were written by Nicky Johns and Stephen Walker. Nicky Johns did much of the detailed design and wrote some complex programs with great speed and accuracy. Richard Jones made substantial contributions to the design and did most of the evaluation work - interviewing users and repeating searches. He also wrote a survey of methods of improving retrieval in subject searching. Chapters 3, 4 and 5 of this report are condensed from this survey, which we hope will be published separately (it is too long to be incorporated in full here). Parts of Chapter 6 were written by Nicky Johns. Most of the rest of the report was written by Stephen Walker, who is responsible for the inevitable inaccuracies. It has not been easy to reconcile the demands of readability and conciseness with the need to include enough detail for other designers to be able to benefit from this work. It is hoped that complete Okapi specifications will be published later this year, together with program source code. Meanwhile, anyone wanting additional detailed information should contact Stephen Walker at PCL.

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 HILDRETH C R. *Online Public Access Catalogs : the user interface*. OCLC Online Computer Library Center, 1982.
- 3 WILLIAMS M E (editor). *Annual Review of Information Science and Technology*. Vol 20. Knowledge Industry Publications, 1985.
- 4 MARKEY K. *Subject searching in library catalogs : before and after the introduction of online catalogs*. OCLC Online Computer Library Center, 1984.

Contents

List of figures	x
List of tables	xi
1 Introduction	1
1.1 The project proposal	1
1.2 Motivation	1
1.2.1 Feasibility study	2
1.3 Staffing	2
1.4 Environment	2
1.5 Historical summary of the project	3
1.6 The report	6
2 Subject searching problems	7
2.1 Introduction	7
2.2 What users bring to the catalogue	8
2.3 Subject searching facilities in current online catalogues	9
2.3.1 Phrase searching	9
2.3.2 Problems with subject headings	9
2.3.3 Keyword searching	10
2.3.4 Access points	11
2.3.5 Access method	11
2.4 How might subject access be improved?	12
2.4.1 Truncation and stemming	12
2.4.2 Cross-reference and other lookup facilities	14
2.4.3 Spelling correction	15
2.5 CITE	16
3 Stemming and truncation	21
3.1 Introduction	21
3.2 Methods and techniques in algorithm construction	21
3.2.1 Iterative or longest match?	22
3.2.2 Conditional rules	22
3.2.3 Stem modification	22
3.2.4 Compilation of the suffix list	22
3.2.5 Users' needs	23

3.3	Conflation algorithms : a review	23
3.3.1	INTREX	23
3.3.2	RADCOL	23
3.3.3	Generation of suffix lists	24
3.3.4	INSPEC	24
3.3.5	Stemming in SMART and FIRST	25
3.3.6	MORPHS	25
3.3.7	Cercone and linguistic analysis	26
3.3.8	MARS	26
3.3.9	Porter	27
3.3.10	Dawson	28
3.4	Evaluating conflation algorithms	29
3.5	Stemming in online catalogues	29
3.5.1	Choice of stemming procedure for online catalogues	30
4	Tables and dictionaries	33
4.1	Introduction	33
4.2	Methods and techniques	33
4.2.1	Dictionaries in spelling correction	33
4.2.2	Proof-reading methods	34
4.2.3	Spelling correction using a dictionary together with a word representation technique	35
4.2.4	Using tables to match related words	35
4.2.5	Linking natural language terms with controlled language terms	36
4.2.6	Compound words and homographs	37
4.2.7	Stop lists	37
4.3	The use of tables in online catalogues	37
5	Fuzzy matching and spelling correction	41
5.1	Introduction	41
5.2	N-grams	41
5.2.1	Definition and applications	41
5.2.2	Use of n-gram techniques to improve recall	42
5.2.3	Use of n-gram techniques to detect spelling errors	43
5.2.4	Different values of n	45
5.2.5	Effectiveness tests	45
5.3	Soundex, soundex-type and other abbreviation codes	46
5.3.1	Definition and applications	46
5.3.2	Use of soundex-type codes: a survey	47
5.4	Fuzzy matching in online catalogues	49
5.4.1	Spelling correction	49
5.4.2	Spelling correction using n-grams	50
5.4.3	Soundex-type codes in online catalogues	51

6 Design and implementation	57
6.1 Introduction	57
6.2 Stemming and spelling standardisation	57
6.2.1 Background	57
6.2.2 Functional design considerations	60
6.2.3 Strong and weak stemming	60
6.2.4 Spelling standardisation	61
6.2.5 Two levels of stemming	61
6.2.6 Interaction design	62
6.2.7 Choice of stemming procedure	64
6.2.8 Stage one - weak stemming and spelling standardisation	64
6.2.9 Stage two - strong stemming	66
6.2.10 Discussions and examples	66
6.2.11 Some oddities	67
6.3 Phrases and the go/see list	68
6.3.1 Categories of equivalence class	68
6.3.2 Phrases	69
6.3.3 Problems in searching for phrases	71
6.3.4 Other go/see list categories	71
6.3.5 A note on stemming and indexing	72
6.4 Spelling correction	73
6.4.1 Dictionary	73
6.4.2 Selection of candidate replacements	74
6.4.3 Finding the nearest match	75
6.4.4 Discussion of the spelling correction technique	76
6.5 Search processing and term combination	77
6.5.1 Preprocessing and index lookup	77
6.5.2 Assignment of term weights	79
6.5.3 Calculation of 'good' and 'acceptable' weights for record retrieval	80
6.5.4 Merging the posting lists	82
6.6 The bibliographic file	83
6.7 The subject index	84
6.7.1 Indexing and the go/see list	84
6.7.2 Source fields	84
6.7.3 Index contents and size	84
6.8 Storage requirements	85
6.9 The Okapi '86 programs	85
 7 System description	 89
7.1 Introduction	89
7.2 Keyboard and display	91
7.3 User input and preprocessing	92
7.4 The search	93
7.4.1 Control system	94
7.4.2 Experimental system	97
7.5 Term combination - the merge	100

7.6	Record display	102
7.6.1	Highlighting of search terms in records	103
7.6.2	Sequencing of displayed records	104
7.6.3	Options following record display	104
7.7	Second and subsequent input screens	105
8	Evaluation	107
8.1	Objects of the evaluation	107
8.2	Methodology	108
8.2.1	Evaluation considerations	108
8.2.2	Controlled or uncontrolled experiments?	109
8.3	Data collection and collation	109
8.3.1	Acceptance tests	110
8.3.2	Observation and interviewing	111
8.3.3	The interview	111
8.3.4	Transaction log data	112
8.3.5	Searches and sessions	113
8.3.6	The <i>SRCHES</i> file	113
8.3.7	Description of the <i>SRCHES</i> file	115
8.4	Analysis of observation and interview data	116
8.4.1	Success rate reported by users	116
8.4.2	Brief analysis of the 17 "failure" sessions	117
8.4.3	Comments made by interviewed users	119
8.5	Statistical analysis of <i>SRCHES</i> file	121
8.5.1	Distribution of number of records retrieved by system	121
8.6	Repetition of searches by experimenter	124
8.6.1	Notes on method	124
8.6.2	Measures of success	124
8.6.3	Experimenters' relevance judgments	125
8.6.4	Searches which retrieved no "good" records: EXP vs. CTL	125
8.6.5	Comparison of recall on first search of session	126
8.7	Treatment of users' words which are not in the index	129
8.7.1	Misspellings and miskeyings	129
8.7.2	Legitimate words which are not in the file	131
8.7.3	The effect of stemming on spelling correction	132
8.7.4	User response to "CAN'T FIND" messages	132
8.7.5	Is spelling correction worth while?	133
8.8	Use of the go/see list	134
9	Conclusions and recommendations	137
9.1	Introduction	137
9.2	Stemming	138
9.2.1	Weak stemming	138
9.2.2	Spelling standardisation	139

9.2.3	Strong stemming	140
9.2.4	Answers to the questions on stemming	141
9.2.5	Recommendations on stemming	142
9.3	Spelling correction	142
9.3.1	Recommendations and discussion	143
9.4	Cross-reference tables - the <i>go/see</i> list	144
9.4.1	Answers to the questions on cross-reference tables	144
9.4.2	Recommendations on cross-reference tables	145
9.5	Users' perception of and behaviour with the system	146
9.6	Applicability of our findings	148
9.7	Concluding remarks	149

Appendixes

1	The soundex algorithm	151
2	The word-matching algorithm	152
3	Extracts from a log file	153
4	Notes on some failed searches	160
5	List of equivalence classes of terms used in the experimental catalogue (<i>go/see</i> list)	167

References	172
------------	-----

Index	180
-------	-----

List of figures

7.1	Introductory screen	90
7.2	Empty input screen	90
7.3	Input screen after user has started to type	91
7.4	Screen during lookup (EXP system)	94
7.5	Display while looking up - word not found	95
7.6	Retyping a misspelt word	96
7.7	Display during and after merging	97
7.8	EXP system suggests a replacement	99
7.9	User accepts suggested replacement	99
7.10	A search for two common terms which do not co-occur	101
7.11	Two "rare" terms which do not co-occur	101
7.12	Full record display	103
7.13	Subsequent input screen showing the results of previous searches	105
7.14	Improved subsequent input screen	105

List of tables

2.1 A sample of subject searches	9
6.1 Types of word used in subject searches (Okapi '84)	59
8.1 Success rate for observed searches by system	117
8.2 Proportion of 'zero hits' searches by system	121
8.3 Distribution of number of terms in searches	122
8.4 Proportion of 'good weight' searches by system by number of terms in search	123
8.5 Searches which found no records of 'good' weight on CTL repeated on EXP	126
8.6 Repetition of initial searches	127
8.7 Initial searches which were the same in CTL as in OSTEM but retrieved more records in EXP	128
8.8 Treatment of misspellings and miskeyings	130
8.9 Legitimate words which were not in the file	131
8.10 Response to 'CAN'T FIND' by system	134
8.11 List of go/see entries used in the searches	135