



UNIVERSITI PUTRA MALAYSIA

**MONOLINGUAL AND CROSS-LANGUAGE INFORMATION
RETRIEVAL APPROACHES FOR MALAY AND ENGLISH
LANGUAGE DOCUMENTS**

MUHAMAD TAUFIK ABDULLAH.

FSKTM 2006 1



**MONOLINGUAL AND CROSS-LANGUAGE INFORMATION RETRIEVAL
APPROACHES FOR MALAY AND ENGLISH LANGUAGE DOCUMENTS**

By

MUHAMAD TAUFIK ABDULLAH

**Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia,
in Fulfilment of the Requirement for the Degree of Doctor of Philosophy**

February 2006



Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

MONOLINGUAL AND CROSS-LANGUAGE INFORMATION RETRIEVAL APPROACHES FOR MALAY AND ENGLISH LANGUAGE DOCUMENTS

By

MUHAMAD TAUFIK ABDULLAH

February 2006

Chairman: Associate Professor Hajah Fatimah Dato' Ahmad, PhD

Faculty: Computer Science and Information Technology

This thesis concerns a Malay-English monolingual and cross-language information retrieval system. It presents a pioneer work in the aspects that are important for the development of Malay-English information retrieval system. An improved Malay stemming algorithm has been developed to stem the various word forms into their common root for the purpose of indexing and retrieving of Malay documents. The new stemming approaches have been introduced for Malay language, namely Rules-Frequency-Order (RFO), Minimum-Rules-Frequency-Order (MRFO), Rules-Frequency-Application-Order (RFAO), and Rules-Application-Frequency-Order (RAFO).

The performance of the new Malay stemming algorithm and approaches are tested using the first two chapters of the Malay translation of the Quranic documents. The results show that the new stemming algorithm and approaches are superior to the

previous stemming algorithm and approach. The retrieval effectiveness of the stemming algorithm and approaches are then tested on the actual Quranic collection using vector space model and latent semantic indexing. The results show that there is an improvement in performance from non-stemmed Malay to stemmed Malay, and also from previous stemming algorithm to the new stemming algorithm.

Since the employment of the new stemming algorithm and approaches achieved good performance results in Malay monolingual information retrieval, a Malay-English cross-language information retrieval experiment has been performed. The results again show that there is an improvement in performance from non-stemmed Malay to stemmed Malay, and from previous stemming algorithm to the new stemming algorithm. In addition, the results reveal that the new stemming in Malay has performed better than the English stemming in retrieving relevant document. The results can be a reference to forthcoming similar experiments and research for cross-language testing of documents retrieval.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

**PENDEKATAN DAPATAN SEMULA MAKLUMAT
MONOBAHASA DAN SILANG-BAHASA UNTUK DOKUMEN
BAHASA MELAYU DAN INGGERIS**

Oleh

MUHAMAD TAUFIK ABDULLAH

Februari 2006

Pengerusi: Profesor Madya Hajah Fatimah Dato' Ahmad, PhD

Fakulti: Sains Komputer dan Teknologi Maklumat

Tesis in adalah berkenaan dengan satu sistem dapatan semula maklumat monobahasa dan silang-bahasa Melayu-Inggeris. Ia mengemukakan kerja perintis dalam aspek-aspek yang penting untuk pembangunan sistem dapatan semula maklumat Melayu-Inggeris. Satu penambahbaikan kepada algoritma pengakar bahasa Melayu telah dibangunkan untuk mencantas pelbagai bentuk perkataan kepada kata akar yang sama untuk tujuan pengindeksan dan dapatan semula dokumen-dokumen Melayu. Pendekatan pengakar yang baharu telah diperkenalkan untuk bahasa Melayu, iaitu *Rules-Frequency-Order*, *Minimum-Rules-Frequency-Order*, *Rules-Frequency-Application-Order*, dan *Rules-Application-Frequency-Order*.

Prestasi algoritma dan pendekatan pengakar baharu bahasa Melayu ini telah diuji dengan menggunakan dua surah pertama daripada dokumen terjemahan Al-Quran

bahasa Melayu. Hasil menunjukkan algoritma dan pendekatan pengakar baharu bahasa Melayu ini adalah lebih baik berbanding dengan algoritma dan pendekatan sebelumnya. Keberkesanan dapatan semula bagi algoritma dan pendekatan ini kemudian telah diuji ke atas koleksi Al-Quran yang sebenar dengan menggunakan kaedah ruang vektor dan pengindeksan semantik terpendam. Hasil menunjukkan bahawa terdapat peningkatan prestasi daripada perkataan Melayu yang tidak dicantas kepada perkataan yang dicantas, dan juga daripada algoritma pengakar sebelumnya kepada algoritma pengakar baharu.

Memandangkan penggunaan algoritma dan pendekatan pengakar baharu ini telah menghasilkan keputusan prestasi yang baik, satu eksperimen dapatan semula maklumat silang-bahasa Melayu-Inggeris telah dilaksanakan. Hasilnya juga menunjukkan bahawa terdapat peningkatan prestasi daripada perkataan Melayu yang tidak dicantas kepada perkataan yang dicantas, dan juga daripada algoritma pengakar sebelumnya kepada algoritma pengakar baharu. Di samping itu, hasil menunjukkan bahawa pengakar baharu bahasa Melayu mempunyai prestasi yang lebih baik daripada pengakar bahasa Inggeris dalam dapatan semula dokumen yang berkaitan. Hasil-hasil boleh menjadi rujukan kepada eksperimen-eksperimen akan datang yang seumpamanya dan penyelidikan untuk pengujian silang-bahasa bagi dapatan semula dokumen-dokumen.

ACKNOWLEDGEMENTS

In the Name of Allah

The Most Beneficent, The Most Merciful

This thesis would not have been possible without the help and support of many people. First and foremost, I would like to express my sincere and deepest gratitude to the chairman of the supervisory committee Associate Professor Dr Hajah Fatimah Dato' Ahmad for her invaluable advice, guidance, discussion, co-operation and most of all for being very understanding of my situation as a student and a staff at her department.

I am also very grateful to the member of the supervisory committee, Associate Professor Dr Ramlan Mahmud and Professor Dr Tengku Mohd Tengku Sembok for their advice, motivation, comments and being very helpful during the completion of this thesis.

I am also indebted to the Universiti Putra Malaysia for the sponsorship and study leave which enables me to pursue this research. My gratitude is also extended to my parents, friends and family for being so supportive and helpful.

Finally, my special thanks and appreciation goes to my wife Ismawani and my four kids Hafiz, Syafiq, Afiqah and Syafiah for their understanding, caring and patience.

TABLE OF CONTENTS

| | Page |
|--|---------------|
| ABSTRACT | ii |
| ABSTRAK | iv |
| ACKNOWLEDGEMENTS | vi |
| APPROVAL | vii |
| DECLARATION | ix |
| LIST OF TABLES | xiii |
| LIST OF FIGURES | xxiii |
| LIST OF ABBREVIATIONS | xxv |
| CHAPTER | |
| 1. INTRODUCTION | 1 |
| 1.1 Background | 1 |
| 1.2 Problem Statement | 4 |
| 1.3 Objectives of the Research | 5 |
| 1.4 Scope of the Research | 5 |
| 1.5 Research Methodology | 6 |
| 1.6 Contributions of the Research | 9 |
| 1.7 Organization of the Thesis | 9 |
| 2. LITERATURE REVIEW | 12 |
| 2.1 Introduction | 12 |
| 2.2 Information Retrieval Models | 14 |
| 2.2.1 Boolean Model | 15 |
| 2.2.2 Probabilistic Model | 16 |
| 2.2.3 Vector Model | 17 |
| 2.2.4 Latent Semantic Indexing Model | 19 |
| 2.2.5 Comparison of the Models | 22 |
| 2.3 Cross-Language Information Retrieval | 22 |
| 2.3.1 The Problems in CLIR | 24 |
| 2.3.2 Finding Translations | 25 |
| 2.3.3 Pruning the Translation Alternatives | 26 |
| 2.3.4 Weighting the Translation Alternatives | 27 |
| 2.4 Indexing | 28 |
| 2.5 Term Weighting Schemes | 29 |
| 2.5.1 Local Weight | 30 |
| 2.5.2 Global Weight | 32 |
| 2.5.3 Normalization | 33 |
| 2.6 Automatic Word Conflation | 33 |
| 2.6.1 Table Lookup Stemmers | 35 |
| 2.6.2 N-Gram Stemmers | 35 |
| 2.6.3 Successor Variety Stemmers | 36 |
| 2.6.4 Affix Removal Stemmers | 37 |
| 2.7 English Language Stemmers | 38 |
| 2.7.1 Lovins Stemmer | 38 |

| | | |
|-----------|--|------------|
| 2.7.2 | Porter Stemmer | 39 |
| 2.8 | Malay Language Stemmers | 41 |
| 2.8.1 | Abdullah's Stemmer | 42 |
| 2.8.2 | Othman's Stemmer | 43 |
| 2.8.3 | Ahmad's Stemmer | 45 |
| 2.8.4 | Abu Bakar's Conflation Method | 47 |
| 2.8.5 | Sock's Stemmer | 48 |
| 2.8.6 | Idris's Stemmer | 48 |
| 2.9 | Summary | 51 |
| 3. | THE MAIN CHARACTERISTICS OF MALAY WORDS | 52 |
| 3.1 | Introduction | 52 |
| 3.2 | Malay Word Formation | 53 |
| 3.2.1 | Single Word Formation | 54 |
| 3.2.2 | Derivative Word Formation | 54 |
| 3.2.3 | Compound Word Formation | 59 |
| 3.2.4 | Reduplication Word Formation | 60 |
| 3.3 | Conclusions | 61 |
| 4. | RESULTS AND DISCUSSIONS FOR THE DEVELOPMENT OF A NEW MALAY STEMMING ALGORITHM | 63 |
| 4.1 | Introduction | 63 |
| 4.2 | Combination of Malay Affixes | 64 |
| 4.3 | Development of a Stop Word List | 66 |
| 4.4 | Evaluation of the New Stop Word List | 69 |
| 4.5 | Design of a Stemming Algorithm | 69 |
| 4.7 | Evaluation of the New Malay Stemming Algorithm | 95 |
| 4.7.1 | The Level of Compression | 95 |
| 4.7.2 | Percentage of Errors | 97 |
| 4.8 | Conclusions | 97 |
| 5. | THE FEATURES OF THE EXPERIMENTAL RETRIEVAL SYSTEM | 99 |
| 5.1 | Introduction | 99 |
| 5.2 | System Architecture | 100 |
| 5.2.1 | Document Collection | 100 |
| 5.2.2 | Dictionary | 102 |
| 5.2.3 | Query Processor | 102 |
| 5.2.4 | Retrieval Processor | 103 |
| 5.2.5 | Processing Module | 104 |
| 5.3 | Conclusions | 108 |
| 6. | RESULTS AND DISCUSSIONS FOR MONOLINGUAL INFORMATION RETRIEVAL EXPERIMENT | 109 |
| 6.1 | Introduction | 109 |
| 6.2 | Document Collection | 110 |
| 6.3 | Set of Queries | 111 |
| 6.4 | Relevance Judgements | 113 |
| 6.5 | Performance of the Stemming Algorithm | 116 |

| | | |
|-----------|--|------------|
| 6.6 | Data Analysis Methods | 117 |
| 6.7 | Data Collection | 119 |
| 6.8 | Analysis of Results from VSM Experiments | 121 |
| 6.8.1 | Experiment on the Usage of Hyphen | 122 |
| 6.8.2 | Experiment with the New Stop Word List | 123 |
| 6.8.3 | Experiment on the Term Weighting Schemes | 126 |
| 6.8.4 | Retrieval Effectiveness of the Stemmers | 133 |
| 6.8.5 | Significance Test | 166 |
| 6.9 | Analysis of Results from LSI Experiments | 176 |
| 6.10 | Conclusions | 216 |
| 7. | RESULTS AND DISCUSSIONS FOR CROSS-LANGUAGE INFORMATION RETRIEVAL EXPERIMENT | 219 |
| 7.1 | Introduction | 219 |
| 7.2 | Purpose of the Experiment | 220 |
| 7.3 | Methodology | 221 |
| 7.3.1 | The Test Environment | 221 |
| 7.3.2 | The Test Procedures | 223 |
| 7.4 | Analysis of Results | 224 |
| 7.4.1 | Retrieval Effectiveness of the Stemmers | 224 |
| 7.3.2 | Significance Test | 282 |
| 7.5 | Conclusions | 298 |
| 8. | CONCLUSIONS | 299 |
| 8.1 | Introduction | 299 |
| 8.2 | Research Conclusions | 300 |
| 8.3 | Suggestions for Future Work | 305 |
| | REFERENCE | 308 |
| | APPENDIXES | 313 |
| | BIODATA OF THE AUTHOR | 337 |

LIST OF TABLES

| Table | | Page |
|--------------|--|-------------|
| 2.1 | Local weighting formula | 31 |
| 2.2 | Global weighting formula | 32 |
| 2.3 | Normalization factors formula | 33 |
| 2.4 | Special affixes | 42 |
| 2.5 | List if suffix and prefix (Rule 1) | 49 |
| 4.1 | A list of the 50 most frequently occurring words in the Quranic collection | 67 |
| 4.2 | Results of the experiment using RAO stemmer | 71 |
| 4.3 | The eight new affixes added to the new set of rules | 72 |
| 4.4 | The modified spelling variations rule | 72 |
| 4.5 | The added and deleted root word entries | 73 |
| 4.6 | Results of the experiment using RAO2 stemmer | 73 |
| 4.7 | Spelling exceptions for prefixes | 74 |
| 4.8 | Spelling exception for suffix | 74 |
| 4.9 | Results of the experiment using NRAO stemmer | 77 |
| 4.10 | Results of the experiment using RFO stemmer | 78 |
| 4.11 | Results of the experiment using MRFO stemmer | 79 |
| 4.12 | Results of the experiment using MRFO stemmer | 79 |
| 4.13 | Results of the experiment using RFAO stemmer | 80 |
| 4.14 | Results of the experiment using RAFO stemmer | 81 |
| 4.15 | Comparison between the seven stemmers | 82 |
| 4.16 | Errors for all seven tests done using RAO stemmer | 85 |
| 4.17 | Distribution of unique errors using RAO stemmer | 88 |

LIST OF TABLES

| Table | Page |
|--|-------------|
| 2.1 Local weighting formula | 31 |
| 2.2 Global weighting formula | 32 |
| 2.3 Normalization factors formula | 33 |
| 2.4 Special affixes | 42 |
| 2.5 List if suffix and prefix (Rule 1) | 49 |
| 4.1 A list of the 50 most frequently occurring words in the Quranic collection | 67 |
| 4.2 Results of the experiment using RAO stemmer | 71 |
| 4.3 The eight new affixes added to the new set of rules | 72 |
| 4.4 The modified spelling variations rule | 72 |
| 4.5 The added and deleted root word entries | 73 |
| 4.6 Results of the experiment using RAO2 stemmer | 73 |
| 4.7 Spelling exceptions for prefixes | 74 |
| 4.8 Spelling exception for suffix | 74 |
| 4.9 Results of the experiment using NRAO stemmer | 77 |
| 4.10 Results of the experiment using RFO stemmer | 78 |
| 4.11 Results of the experiment using MRFO stemmer | 79 |
| 4.12 Results of the experiment using MRFO stemmer | 79 |
| 4.13 Results of the experiment using RFAO stemmer | 80 |
| 4.14 Results of the experiment using RAFO stemmer | 81 |
| 4.15 Comparison between the seven stemmers | 82 |
| 4.16 Errors for all seven tests done using RAO stemmer | 85 |
| 4.17 Distribution of unique errors using RAO stemmer | 88 |

| | | |
|------|--|-----|
| 4.18 | Errors for all seven tests using RFO and MRFO stemmers | 89 |
| 4.19 | Distribution of unique errors using RFO and MRFO stemmers | 90 |
| 4.20 | Errors for all seven tests using RFAO stemmer | 91 |
| 4.21 | Distribution of unique errors using RFAO stemmer | 92 |
| 4.22 | Errors for all seven tests using RAFO stemmer | 93 |
| 4.23 | Distribution of unique errors using RAFO stemmer | 94 |
| 4.24 | Compression achieved by the algorithms | 96 |
| 4.25 | Percentage of errors obtained by the stemmers | 97 |
| 5.1 | The Quran's chapters with their corresponding total number of verses | 101 |
| 6.1 | The main quantitative characteristics of the Malay and English sets of queries before deletion of stop words | 112 |
| 6.2 | The main quantitative characteristics of the Malay and English sets of queries after deletion of stop words | 113 |
| 6.3 | Result of the experiment on the usage of hyphen in Malay information retrieval system | 122 |
| 6.4 | Results of the experiment on the usage of hyphen in English information retrieval system | 123 |
| 6.5 | Results of the experiment using the stop word lists on Malay information retrieval system | 124 |
| 6.6 | Results of the experiment using the stop word lists on English information retrieval system | 125 |
| 6.7 | Results of the experiment on Malay VSM retrieval using various local weights on document | 126 |
| 6.8 | Results of the experiment on Malay VSM retrieval using various global weights on document | 127 |
| 6.9 | Results of the experiment on Malay VSM retrieval using various normalization weights on document | 128 |
| 6.10 | Results of the experiment on Malay VSM retrieval using various local weights on query | 129 |
| 6.11 | Results of the experiment on Malay VSM retrieval using various global weights on query | 130 |

| | | |
|------|---|-----|
| 6.12 | Results of the experiment on English VSM retrieval using various local weights on document | 130 |
| 6.13 | Results of the experiment on English VSM retrieval using various global weights on document | 131 |
| 6.14 | Results of the experiment on English VSM retrieval using various normalization weights on document | 132 |
| 6.15 | Results of the experiment on English VSM retrieval using various local weights on query | 132 |
| 6.16 | Results of the experiment on English VSM retrieval using various global weights on query | 133 |
| 6.17 | Number of relevant documents retrieved by using Malay Non-Conflation VSM search at different cutoff points | 135 |
| 6.18 | Number of relevant documents retrieved by using RAO VSM search at different cutoff points | 136 |
| 6.19 | Number of relevant documents retrieved by using RFO VSM search at different cutoff points | 137 |
| 6.20 | Number of relevant documents retrieved by using MRFO VSM search at different cutoff points | 138 |
| 6.21 | Number of relevant documents retrieved by using RFAO VSM search at different cutoff points | 139 |
| 6.22 | Number of relevant documents retrieved by using RAFO VSM search at different cutoff points | 140 |
| 6.23 | Number of relevant documents retrieved by using English Non-Conflation VSM search at different cutoff points | 141 |
| 6.24 | Number of relevant documents retrieved by using Porter VSM search at different cutoff points | 142 |
| 6.25 | The differences between the number of relevant documents retrieved by RAO and Malay non-conflation for all the 36 queries at different cutoff points | 144 |
| 6.26 | The differences between the number of relevant documents retrieved by RFO and Malay non-conflation for all the 36 queries at different cutoff points | 145 |
| 6.27 | The differences between the number of relevant documents retrieved by MRFO and Malay Non-Conflation for all the 36 queries at different cutoff points | 146 |

| | | |
|------|---|-----|
| 6.28 | The differences between the number of relevant documents retrieved by RFAO and Malay Non-Conflation for all the 36 queries at different cutoff points | 147 |
| 6.29 | The differences between the number of relevant documents retrieved by RAFO and Malay Non-Conflation for all the 36 queries at different cutoff points | 148 |
| 6.30 | The differences between the number of relevant documents retrieved by RFO and RAO for all the 36 queries at different cutoff points | 149 |
| 6.31 | The differences between the number of relevant documents retrieved by MRFO and RAO for all the 36 queries at different cutoff points | 150 |
| 6.32 | The differences between the number of relevant documents retrieved by RFAO and RAO for all the 36 queries at different cutoff points | 151 |
| 6.33 | The differences between the number of relevant documents retrieved by RAFO and RAO for all the 36 queries at different cutoff points | 152 |
| 6.34 | The differences between the number of relevant documents retrieved by MRFO and RFO for all the 36 queries at different cutoff points | 153 |
| 6.35 | The differences between the number of relevant documents retrieved by RFAO and RFO for all the 36 queries at different cutoff points | 154 |
| 6.36 | The differences between the number of relevant documents retrieved by RAFO and RFO for all the 36 queries at different cutoff points | 155 |
| 6.37 | The differences between the number of relevant documents retrieved by RFAO and MRFO for all the 36 queries at different cutoff points | 156 |
| 6.38 | The differences between the number of relevant documents retrieved by RAFO and MRFO for all the 36 queries at different cutoff points | 157 |
| 6.39 | The differences between the number of relevant documents retrieved by RAFO and RFAO for all the 36 queries at different cutoff points | 158 |
| 6.40 | The differences between the number of relevant documents retrieved by Porter and English Non-Conflation for all the 36 queries at different cutoff points | 159 |
| 6.41 | Average Recall and Precision for Malay Non-Conflation and Automatic Word Conflation using VSM | 162 |
| 6.42 | Average Recall and Precision for English Non-Conflation and Automatic Word Conflation using VSM | 165 |
| 6.43 | Frequency distribution of the direction of differences between 15 pairs of text representation (with p value) | 170 |

| | | |
|------|--|-----|
| 6.44 | Results of the experiment on Malay LSI retrieval using various local weights on document | 177 |
| 6.45 | Results of the experiment on Malay LSI retrieval using various global weights on document | 178 |
| 6.46 | Results of the experiment on Malay LSI retrieval using various normalization weights on document | 178 |
| 6.47 | Results of the experiment on Malay LSI retrieval using various local weights on query | 179 |
| 6.48 | Results of the experiment on Malay LSI retrieval using various global weights on query | 180 |
| 6.49 | Results of the experiment on English LSI retrieval using various local weights on document | 180 |
| 6.50 | Results of the experiment on English LSI retrieval using various global weights on document | 181 |
| 6.51 | Results of the experiment on English LSI retrieval using various normalization weights on document | 182 |
| 6.52 | Results of the experiment on English LSI retrieval using various local weights on query | 182 |
| 6.53 | Results of the experiment on English LSI retrieval using various global weights on query | 183 |
| 6.54 | Number of relevant documents retrieved by using Malay Non-Conflation LSI search at different cutoff points | 185 |
| 6.55 | Number of relevant documents retrieved by using RAO LSI search at different cutoff points | 186 |
| 6.56 | Number of relevant documents retrieved by using RFO LSI search at different cutoff points | 187 |
| 6.57 | Number of relevant documents retrieved by using MRFO LSI search at different cutoff points | 188 |
| 6.58 | Number of relevant documents retrieved by using RFAO LSI search at different cutoff points | 189 |
| 6.59 | Number of relevant documents retrieved by using RAFO LSI search at different cutoff points | 190 |
| 6.60 | Number of relevant documents retrieved by using English Non-Conflation LSI search at different cutoff points | 191 |

| | | |
|------|---|-----|
| 6.61 | Number of relevant documents retrieved by using Porter LSI search at different cutoff points | 192 |
| 6.62 | The differences between the number of relevant documents retrieved by RAO and Malay non-conflation for all the 36 queries at different cutoff points | 194 |
| 6.63 | The differences between the number of relevant documents retrieved by RFO and Malay non-conflation for all the 36 queries at different cutoff points | 195 |
| 6.64 | The differences between the number of relevant documents retrieved by MRFO and Malay Non-Conflation for all the 36 queries at different cutoff points | 196 |
| 6.65 | The differences between the number of relevant documents retrieved by RFAO and Malay Non-Conflation for all the 36 queries at different cutoff points | 197 |
| 6.66 | The differences between the number of relevant documents retrieved by RAFO and Malay Non-Conflation for all the 36 queries at different cutoff points | 198 |
| 6.67 | The differences between the number of relevant documents retrieved by RFO and RAO for all the 36 queries at different cutoff points | 199 |
| 6.68 | The differences between the number of relevant documents retrieved by MRFO and RAO for all the 36 queries at different cutoff points | 200 |
| 6.69 | The differences between the number of relevant documents retrieved by RFAO and RAO for all the 36 queries at different cutoff points | 201 |
| 6.70 | The differences between the number of relevant documents retrieved by RAFO and RAO for all the 36 queries at different cutoff points | 202 |
| 6.71 | The differences between the number of relevant documents retrieved by MRFO and RFO for all the 36 queries at different cutoff points | 203 |
| 6.72 | The differences between the number of relevant documents retrieved by RFAO and RFO for all the 36 queries at different cutoff points | 204 |
| 6.73 | The differences between the number of relevant documents retrieved by RAFO and RFO for all the 36 queries at different cutoff points | 205 |
| 6.74 | The differences between the number of relevant documents retrieved by RFAO and MRFO for all the 36 queries at different cutoff points | 206 |
| 6.75 | The differences between the number of relevant documents retrieved by RAFO and MRFO for all the 36 queries at different cutoff points | 207 |

| | | |
|------|---|-----|
| 6.76 | The differences between the number of relevant documents retrieved by RAFO and RFAO for all the 36 queries at different cutoff points | 208 |
| 6.77 | The differences between the number of relevant documents retrieved by Porter and English Non-Conflation for all the 36 queries at different cutoff points | 209 |
| 6.78 | Average Recall and Precision for Malay Non-Conflation and Automatic Word Conflation using LSI | 212 |
| 6.79 | Average Recall and Precision for English Non-Conflation and Automatic Word Conflation using LSI | 215 |
| 7.1 | Number of relevant documents retrieved by using English Non-Conflation search at different cutoff points | 226 |
| 7.2 | Number of relevant documents retrieved by using Porter search at different cutoff points | 227 |
| 7.3 | Number of relevant documents retrieved by using Malay Non-Conflation search at different cutoff points | 229 |
| 7.4 | Number of relevant documents retrieved by using RAO search at different cutoff points | 230 |
| 7.5 | Number of relevant documents retrieved by using RFO search at different cutoff points | 231 |
| 7.6 | Number of relevant documents retrieved by using MRFO search at different cutoff points | 232 |
| 7.7 | Number of relevant documents retrieved by using RFAO search at different cutoff points | 233 |
| 7.8 | Number of relevant documents retrieved by using RAFO search at different cutoff points | 234 |
| 7.9 | The differences between the number of relevant documents retrieved by Porter and English Non-Conflation for all the 36 queries at different cutoff points | 236 |
| 7.10 | The differences between the number of relevant documents retrieved by RAO and Malay Non-Conflation for all the 36 queries at different cutoff points | 238 |
| 7.11 | The differences between the number of relevant documents retrieved by RFO and Malay Non-Conflation for all the 36 queries at different cutoff points | 239 |
| 7.12 | The differences between the number of relevant documents retrieved | 240 |

| | | |
|------|--|-----|
| | by MRFO and Malay Non-Conflation for all the 36 queries at different cutoff points | |
| 7.13 | The differences between the number of relevant documents retrieved by RFAO and Malay Non-Conflation for all the 36 queries at different cutoff points | 241 |
| 7.14 | The differences between the number of relevant documents retrieved by RAFO and Malay Non-Conflation for all the 36 queries at different cutoff points | 242 |
| 7.15 | The differences between the number of relevant documents retrieved by RFO and RAO for all the 36 queries at different cutoff points | 243 |
| 7.16 | The differences between the number of relevant documents retrieved by MRFO and RAO for all the 36 queries at different cutoff points | 244 |
| 7.17 | The differences between the number of relevant documents retrieved by RFAO and RAO for all the 36 queries at different cutoff points | 245 |
| 7.18 | The differences between the number of relevant documents retrieved by RAFO and RAO for all the 36 queries at different cutoff points | 246 |
| 7.19 | The differences between the number of relevant documents retrieved by MRFO and RFO for all the 36 queries at different cutoff points | 247 |
| 7.20 | The differences between the number of relevant documents retrieved by RFAO and RFO for all the 36 queries at different cutoff points | 248 |
| 7.21 | The differences between the number of relevant documents retrieved by RAFO and RFO for all the 36 queries at different cutoff points | 249 |
| 7.22 | The differences between the number of relevant documents retrieved by RFAO and MRFO for all the 36 queries at different cutoff points | 250 |
| 7.23 | The differences between the number of relevant documents retrieved by RAFO and MRFO for all the 36 queries at different cutoff points | 251 |
| 7.24 | The differences between the number of relevant documents retrieved by RAFO and RFAO for all the 36 queries at different cutoff points | 252 |
| 7.25 | The differences between the number of relevant documents retrieved by RAO Stemming and Porter Stemming for all the 36 queries at different cutoff points | 255 |
| 7.26 | The differences between the number of relevant documents retrieved by RFO Stemming and Porter Stemming for all the 36 queries at different cutoff points | 256 |

| | | |
|------|---|-----|
| 7.27 | The differences between the number of relevant documents retrieved by MRFO Stemming and Porter Stemming for all the 36 queries at different cutoff points | 257 |
| 7.28 | The differences between the number of relevant documents retrieved by RFAO Stemming and Porter Stemming for all the 36 queries at different cutoff points | 258 |
| 7.29 | The differences between the number of relevant documents retrieved by RAFO Stemming and Porter Stemming for all the 36 queries at different cutoff points | 259 |
| 7.30 | The differences between the number of relevant documents retrieved by CLIR and Monolingual IR with English Non-Conflation for all the 36 queries at different cutoff points | 261 |
| 7.31 | The differences between the number of relevant documents retrieved by CLIR and Monolingual IR with Porter stemming for all the 36 queries at different cutoff points | 262 |
| 7.32 | The differences between the number of relevant documents retrieved by CLIR and Monolingual IR with Malay Non-Conflation for all the 36 queries at different cutoff points | 264 |
| 7.33 | The differences between the number of relevant documents retrieved by CLIR and Monolingual IR with RAO stemming for all the 36 queries at different cutoff points | 265 |
| 7.34 | The differences between the number of relevant documents retrieved by CLIR and Monolingual IR with RFO stemming for all the 36 queries at different cutoff points | 266 |
| 7.35 | The differences between the number of relevant documents retrieved by CLIR and Monolingual IR with MRFO stemming for all the 36 queries at different cutoff points | 267 |
| 7.36 | The differences between the number of relevant documents retrieved by CLIR and Monolingual IR with RFAO stemming for all the 36 queries at different cutoff points | 268 |
| 7.37 | The differences between the number of relevant documents retrieved by CLIR and Monolingual IR with RAFO stemming for all the 36 queries at different cutoff points | 269 |
| 7.38 | Average Recall and Precision values for English-Malay CLIR with English Query | 271 |
| 7.39 | Average Recall and Precision values for English-Malay CLIR with Malay Query | 272 |

- 7.40 Frequency distribution of the direction of differences between 1 pair of text representation (with p value) for English 283
- 7.41 Frequency distribution of the direction of differences between 15 pairs of text representation (with p value) for Malay 284
- 7.42 Frequency distribution of the direction of differences between 5 pairs of text representation (with p value) for Malay and English 290
- 7.43 Frequency distribution of the direction of differences between 2 pairs of retrieval (with p value) for English monolingual and cross-language retrieval 293
- 7.44 Frequency distribution of the direction of differences between 6 pairs of retrieval (with p value) for Malay monolingual and cross-language retrieval 294



LIST OF FIGURES

| Figure | Page |
|---|-------------|
| 2.1 Information retrieval processes | 13 |
| 4.1 Flowchart of the new Malay stemming algorithm | 76 |
| 5.1 Architecture of the retrieval system | 100 |
| 5.2 Precision and recall for a given query | 104 |
| 5.3 Processes involved in the processing module | 105 |
| 6.1 Average Recall-Precision Graph for each type of Malay text representation using VSM | 163 |
| 6.2 Average Recall-Precision Chart for each type of Malay text representation using VSM | 164 |
| 6.3 Average Recall-Precision Graph for each type of English text representation using VSM | 166 |
| 6.4 Average Recall-Precision Graph for each type of Malay text representation using LSI | 213 |
| 6.5 Average Recall-Precision Chart for each type of Malay text representation using LSI | 214 |
| 6.6 Average Recall-Precision Graph for each type of English text representation using LSI | 216 |
| 7.1 Average Recall-Precision Graph for English-Malay CLIR with English Query | 271 |
| 7.2 Average Recall-Precision Graph for English-Malay CLIR with Malay Query | 273 |
| 7.3 Average Recall-Precision Chart for English-Malay CLIR with Malay Query | 274 |
| 7.4 Average Recall-Precision Graph for the six types of text representationfor CLIR | 276 |
| 7.5 Average Recall-Precision Chart for the six types of text representationfor CLIR | 277 |
| 7.6 Average Recall-Precision Graph from the English Non-Conflation text | 278 |