

**Compression, Indexing and Searching of a Large Structured-Text
Database in A Library Monitoring and Control System (LiMaCS)**

by

Tam Wai I

A Thesis submitted in partial fulfillment
of the requirements of the Degree of

MASTER OF SCIENCE
in
SOFTWARE ENGINEERING

Faculty of Science and Technology
UNIVERSITY OF MACAU

June 1998

Contents

- 1. Introduction 1
 - 1.1 Background 1
 - 1.2 Approach of the thesis..... 1
 - 1.3 Outline of this thesis 2
- 2. Domain Analysis..... 4
 - 2.1 Basic Domain Components..... 4
 - 2.1.1 Library 4
 - 2.1.2 Client 6
 - 2.1.3 Document Provider..... 6
 - 2.2 Functions, Operations and Behaviors 7
 - 2.2.1 Acquisition..... 7
 - 2.2.2 Cataloguing..... 12
 - 2.2.3 Access and Catalogue Provision..... 14
 - 2.2.4 Circulation Control 16
 - 2.2.5 Serials Control 19
 - 2.2.6 Inter-library Loan..... 21
- 3. System Overview 25
 - 3.1 Architecture of LiMaCS..... 25
 - 3.1.1 Regional Libraries 25
 - 3.1.2 System Requirements 28
 - 3.2 Database Design (Logical)..... 28
 - 3.2.1 Entity and Relationship 29
 - 3.2.2 Compression and index 32
 - 3.3 Implementation 32
 - 3.3.1 Extendible word-based Huffman coding 32
 - 3.3.2 Implementation environment and tools 33

4. Overview of text compression, indexing and searching 34

4.1 Text compression 34

4.1.1 Huffman coding 34

4.1.2 Arithmetic coding 37

4.1.3 Lempel-Ziv approach.....41

4.1.4 Conclusion 42

4.2 Indexing 43

4.2.1 Inverted file indexing.....43

4.2.2 Signature files 45

4.2.3 Bitmaps 47

4.2.4 Conclusion 49

4.3 Searching.....49

4.3.1 Structures of the lexicon.....49

4.3.2 Boolean query 52

4.3.3 Ranking and information retrieval.....53

5. Compression, indexing and searching methods used in LiMaCS.....54

5.1 Compression in LiMaCS.....54

5.1.1 Database design (Physical).....54

5.1.2 Requirement.....55

5.1.3 Choice of compression method 56

5.1.4 Extendible Huffman coding.....56

5.1.5 Implementation of extendible word-based Huffman coding 61

5.1.6 Evaluation 65

5.2 Index construction and Search processing 68

5.2.1 Choice of index method.....68

5.2.2 Inverted file structure.....69

5.2.3 Index file structure for searching.....69

5.2.4 Index construction 70

6. Conclusions 72

6.1 Summary of achievements 72

6.2 Limitations and future research areas 73

7. Appendix A 75

8. References 92