TABLE OF CONTENTS

Acknowledgement                                           i
Certificate                                               ii
Declaration                                               iii
Dedication                                                iv
Abstract                                                  v
Table of Content                                          vi
List of Figures                                           xi
List of Tables                                           xiv

CHAPTER 1    INTRODUCTION

1.0 Outline of this chapter                              001
  1.1 Origin of Data Compression                           001
  1.2 Introduction of Data Compression                    002
  1.3 Data Compression                                    004
  1.4 Lossless Compression Algorithms                     006
    1.4.1 Statistical Methods                              006
    1.4.2 Dictionary Methods                               011
    1.4.3 Transform Based Methods                          014
  1.5 Searching in Compressed Files                       015
  1.6 Our Contribution                                    016
  1.7 Organization of Thesis                              016

CHAPTER 2    COMPRESSION TECHNIQUES

2.0 Outline of this chapter                              018
  2.1 Origins of Arithmetic Coding                         018
  2.2 Arithmetic Coding                                    020
    2.2.1 Example of Arithmetic Coding                     020
    2.2.2 Practical Matters                                024
    2.2.3 Complication in Arithmetic Coding                026
    2.2.4 Decoding                                         028
    2.2.5 Comparison with Huffman Coding                   029
2.3 Burrows Wheeler Transformation 030
  2.3.1 The Reversible Transformation 031
  2.3.2 Why the transformed string compressed well 034
  2.3.3 Move-To-Front Coding 035
2.4 Prediction By Partial Match 036
2.5 Lempel-Ziv Markov-chain Algorithm 046
  2.5.1 LZMA Algorithm 047
  2.5.2 Range Encoding 048
  2.5.3 Relationship with Arithmetic Coding 049

CHAPTER 3 STRING-MATCHING ALGORITHMS

3.0 Outline of this chapter 050
  3.1 Introduction 050
  3.2 String-matching Algorithms 051
    3.2.1 From Left to Right 052
    3.2.2 From Right to Left 053
    3.2.3 In a specific order 053
    3.2.4 In any order 054
  3.3 Karp-Rabin Algorithm 054
  3.4 Knuth-Morris-Pratt Algorithm 056
  3.5 Boyer-Moore Algorithm 057
  3.6 Quick Search Algorithm 062

CHAPTER 4 PROPOSED COMPRESSION METHODS

4.0 Outline of this chapter 064
  4.1 Introduction 064
    4.1.1 Dictionary Models 066
    4.1.2 Related work for Preprocessing Texts 068
  4.2 Idea of our Method 072
  4.3 Proposed Compression Methods 074
    4.3.1 Character based Text Compression method using Static Dictionary (CBTC-A) 074
4.3.2 Character based Text Compression method using Semi-Dynamic Dictionary (CBTC-B) 080
4.3.3 Word based Text Compression method using Semi-Dynamic Dictionary (WBTC-A) 083
4.3.4 Word based Text Compression method using Semi-Dynamic Dictionary (WBTC-B) 086
4.3.5 Word based Text Compression method using Two Dimensional Semi-Dynamic Dictionary (WBTC-C) 087
4.3.6 Word based Text Compression method using Dynamic Dictionary (WBTC-D) 090
4.3.7 Word based Text Compression method using Static Dictionary (WBTC-E) 090

4.4 Comparison among Proposed Methods 093

CHAPTER 5 IMPLEMENTATION OF PROPOSED METHODS
5.0 Outline of this chapter 097
   5.1 Implementation of CBTC-A 097
   5.2 Implementation of CBTC-B 101
   5.3 Implementation of WBTC-A 107
   5.4 Implementation of WBTC-B 117
   5.5 Implementation of WBTC-C 122
   5.6 Implementation of WBTC-D 126
   5.7 Implementation of WBTC-E 129
   5.8 Implementing Searching in compressed form. 134

CHAPTER 6 EXPERIMENTAL RESULTS
6.0 Outline of this chapter 135
   6.1 List of Files used for Testing 135
      6.1.1 E-Text Corpus 135
      6.1.2 European Parliament Corpus 136
      6.1.3 Enronsent Corpus 137
      6.1.4 Project Gutenberg Corpus 138
      6.1.5 Mixed Corpus 138
6.1.6 Summary of Corpus

6.2 Comparison of Word based Methods with Bzip2
   6.2.1 Compression statistics of E-Text Corpus
   6.2.2 Compression statistics of European Parliament Corpus
   6.2.3 Compression statistics of Enronsent Corpus
   6.2.4 Compression statistics of Project Gutenberg Corpus
   6.2.5 Compression statistics of Mixed Corpus
   6.2.6 Compression statistics of All Corpus

6.3 Comparison of Word based Methods with PPMd
   6.3.1 Compression statistics of E-Text Corpus
   6.3.2 Compression statistics of European Parliament Corpus
   6.3.3 Compression statistics of Enronsent Corpus
   6.3.4 Compression statistics of Project Gutenberg Corpus
   6.3.5 Compression statistics of Mixed Corpus
   6.3.6 Compression statistics of All Corpus

6.4 Comparison of Word based Methods with PPMII
   6.4.1 Compression statistics of E-Text Corpus
   6.4.2 Compression statistics of European Parliament Corpus
   6.4.3 Compression statistics of Enronsent Corpus
   6.4.4 Compression statistics of Project Gutenberg Corpus
   6.4.5 Compression statistics of Mixed Corpus
   6.4.6 Compression statistics of All Corpus

6.5 Comparison of Word based Methods with LZMA
   6.5.1 Compression statistics of E-Text Corpus
   6.5.2 Compression statistics of European Parliament Corpus
   6.5.3 Compression statistics of Enronsent Corpus
   6.5.4 Compression statistics of Project Gutenberg Corpus
   6.5.5 Compression statistics of Mixed Corpus
   6.5.6 Compression statistics of All Corpus

6.6 Comparison of CBTC-B with Arithmetic Coding
   6.6.1 Compression statistics of E-Text Corpus
6.6.2 Compression statistics of European Parliament Corpus 173
6.6.3 Compression statistics of Enronsent Corpus 174
6.6.4 Compression statistics of Project Gutenberg Corpus 175
6.6.5 Compression statistics of Mixed Corpus 175
6.6.6 Compression statistics of All Corpus 176

6.7 Searching Phrase in Compressed File 180
6.7.1 Searching Phrase using Karp-Rabin Algorithm 181
6.7.2 Searching Phrase using Knuth-Morris-Pratt Algorithm 181
6.7.3 Searching Phrase using Brute-Force Algorithm 182
6.7.4 Searching Phrase using Boyer-Moore Algorithm 182
6.7.5 Searching Phrase using Quick Search Algorithm 183
6.7.6 Overall Comparison of Searching Algorithm 183

6.8 Decompression Time 186
6.8.1 Decompression Time for E-Text Corpus 186
6.8.2 Decompression Time for European Parliament Corpus 187
6.8.3 Decompression Time for Enronsent Corpus 189
6.8.4 Decompression Time for Project Gutenberg Corpus 191
6.8.5 Decompression Time for Mixed Corpus 192
6.8.6 Decompression Time for All Corpus 192

CHAPTER 7 CONCLUSION AND FUTURE WORK 197
BIBLIOGRAPHY 201
APPENDIX A Coding of CBTC-B 208
APPENDIX B Coding of WBTC-A 221
APPENDIX C Coding of WBTC-B 233
APPENDIX D Coding of WBTC-C 239
APPENDIX E Coding of WBTC-D 248
APPENDIX F Coding of WBTC-E 254
APPENDIX G Coding of Searching Phrase 262