Contents

List of Algorithms page xiii
Foreword xvii
Acknowledgments xix

1 Introduction 1
1.1 Why Compact Data Structures? 1
1.2 Why This Book? 3
1.3 Organization 4
1.4 Software Resources 6
1.5 Mathematics and Notation 7
1.6 Bibliographic Notes 10

2 Entropy and Coding 14
2.1 Worst-Case Entropy 14
2.2 Shannon Entropy 16
2.3 Empirical Entropy 17
  2.3.1 Bit Sequences 18
  2.3.2 Sequences of Symbols 20
2.4 High-Order Entropy 21
2.5 Coding 22
2.6 Huffman Codes 25
  2.6.1 Construction 25
  2.6.2 Encoding and Decoding 26
  2.6.3 Canonical Huffman Codes 27
  2.6.4 Better than Huffman 30
2.7 Variable-Length Codes for Integers 30
2.8 Jensen’s Inequality 33
2.9 Application: Positional Inverted Indexes 35
2.10 Summary 36
2.11 Bibliographic Notes 36
## 3 Arrays

### 3.1 Elements of Fixed Size

- Page 40

### 3.2 Elements of Variable Size

#### 3.2.1 Sampled Pointers

- Page 46

#### 3.2.2 Dense Pointers

- Page 47

### 3.3 Partial Sums

- Page 48

### 3.4 Applications

#### 3.4.1 Constant-Time Array Initialization

- Page 49

#### 3.4.2 Direct Access Codes

- Page 53

#### 3.4.3 Elias-Fano Codes

- Page 57

#### 3.4.4 Differential Encodings and Inverted Indexes

- Page 59

#### 3.4.5 Compressed Text Collections

- Page 59

### 3.5 Summary

- Page 61

### 3.6 Bibliographic Notes

- Page 61

## 4 Bitvectors

### 4.1 Access

#### 4.1.1 Zero-Order Compression

- Page 65

#### 4.1.2 High-Order Compression

- Page 71

### 4.2 Rank

- Page 73

#### 4.2.1 Sparse Sampling

- Page 73

#### 4.2.2 Constant Time

- Page 74

#### 4.2.3 Rank on Compressed Bitvectors

- Page 76

### 4.3 Select

- Page 78

#### 4.3.1 A Simple Heuristic

- Page 78

#### 4.3.2 An $O(\log \log n)$ Time Solution

- Page 80

#### 4.3.3 Constant Time

- Page 81

### 4.4 Very Sparse Bitvectors

- Page 82

#### 4.4.1 Constant-Time Select

- Page 83

#### 4.4.2 Solving Rank

- Page 83

#### 4.4.3 Bitvectors with Runs

- Page 86

### 4.5 Applications

- Page 87

#### 4.5.1 Partial Sums Revisited

- Page 87

#### 4.5.2 Predecessors and Successors

- Page 89

#### 4.5.3 Dictionaries, Sets, and Hashing

- Page 91

### 4.6 Summary

- Page 98

### 4.7 Bibliographic Notes

- Page 98

## 5 Permutations

### 5.1 Inverse Permutations

- Page 103

### 5.2 Powers of Permutations

- Page 106

### 5.3 Compressible Permutations

- Page 108

### 5.4 Applications

#### 5.4.1 Two-Dimensional Points

- Page 115

#### 5.4.2 Inverted Indexes Revisited

- Page 116

### 5.5 Summary

- Page 117

### 5.6 Bibliographic Notes

- Page 117
# Contents

## 6 Sequences

6.1 Using Permutations
   - 6.1.1 Chunk-Level Granularity
   - 6.1.2 Operations within a Chunk
   - 6.1.3 Construction
   - 6.1.4 Space and Time

6.2 Wavelet Trees
   - 6.2.1 Structure
   - 6.2.2 Solving Rank and Select
   - 6.2.3 Construction
   - 6.2.4 Compressed Wavelet Trees
   - 6.2.5 Wavelet Matrices

6.3 Alphabet Partitioning

6.4 Applications
   - 6.4.1 Compressible Permutations Again
   - 6.4.2 Compressed Text Collections Revisited
   - 6.4.3 Non-positional Inverted Indexes
   - 6.4.4 Range Quantile Queries
   - 6.4.5 Revisiting Arrays of Variable-Length Cells

6.5 Summary

6.6 Bibliographic Notes

## 7 Parentheses

7.1 A Simple Implementation
   - 7.1.1 Range Min-Max Trees
   - 7.1.2 Forward and Backward Searching
   - 7.1.3 Range Minima and Maxima
   - 7.1.4 Rank and Select Operations

7.2 Improving the Complexity
   - 7.2.1 Queries inside Buckets
   - 7.2.2 Forward and Backward Searching
   - 7.2.3 Range Minima and Maxima
   - 7.2.4 Rank and Select Operations

7.3 Multi-Parenthesis Sequences
   - 7.3.1 Nearest Marked Ancestors

7.4 Applications
   - 7.4.1 Succinct Range Minimum Queries
   - 7.4.2 XML Documents

7.5 Summary

7.6 Bibliographic Notes

## 8 Trees

8.1 LOUDS: A Simple Representation
   - 8.1.1 Binary and Cardinal Trees

8.2 Balanced Parentheses
   - 8.2.1 Binary Trees Revisited
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.3</td>
<td>DFUDS Representation</td>
<td>233</td>
</tr>
<tr>
<td>8.3.1</td>
<td>Cardinal Trees Revisited</td>
<td>240</td>
</tr>
<tr>
<td>8.4</td>
<td>Labeled Trees</td>
<td>241</td>
</tr>
<tr>
<td>8.5</td>
<td>Applications</td>
<td>245</td>
</tr>
<tr>
<td>8.5.1</td>
<td>Routing in Minimum Spanning Trees</td>
<td>246</td>
</tr>
<tr>
<td>8.5.2</td>
<td>Grammar Compression</td>
<td>248</td>
</tr>
<tr>
<td>8.5.3</td>
<td>Tries</td>
<td>252</td>
</tr>
<tr>
<td>8.5.4</td>
<td>LZ78 Compression</td>
<td>259</td>
</tr>
<tr>
<td>8.5.5</td>
<td>XML and XPath</td>
<td>262</td>
</tr>
<tr>
<td>8.5.6</td>
<td>Treaps</td>
<td>264</td>
</tr>
<tr>
<td>8.5.7</td>
<td>Integer Functions</td>
<td>266</td>
</tr>
<tr>
<td>8.6</td>
<td>Summary</td>
<td>272</td>
</tr>
<tr>
<td>8.7</td>
<td>Bibliographic Notes</td>
<td>272</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Graphs</td>
<td>279</td>
</tr>
<tr>
<td>9.1</td>
<td>General Graphs</td>
<td>281</td>
</tr>
<tr>
<td>9.1.1</td>
<td>Using Bitvectors</td>
<td>281</td>
</tr>
<tr>
<td>9.1.2</td>
<td>Using Sequences</td>
<td>281</td>
</tr>
<tr>
<td>9.1.3</td>
<td>Undirected Graphs</td>
<td>284</td>
</tr>
<tr>
<td>9.1.4</td>
<td>Labeled Graphs</td>
<td>285</td>
</tr>
<tr>
<td>9.1.5</td>
<td>Construction</td>
<td>289</td>
</tr>
<tr>
<td>9.2</td>
<td>Clustered Graphs</td>
<td>291</td>
</tr>
<tr>
<td>9.2.1</td>
<td>(K^2)-Tree Structure</td>
<td>291</td>
</tr>
<tr>
<td>9.2.2</td>
<td>Queries</td>
<td>292</td>
</tr>
<tr>
<td>9.2.3</td>
<td>Reducing Space</td>
<td>294</td>
</tr>
<tr>
<td>9.2.4</td>
<td>Construction</td>
<td>296</td>
</tr>
<tr>
<td>9.3</td>
<td>(K)-Page Graphs</td>
<td>296</td>
</tr>
<tr>
<td>9.3.1</td>
<td>One-Page Graphs</td>
<td>297</td>
</tr>
<tr>
<td>9.3.2</td>
<td>(K)-Page Graphs</td>
<td>299</td>
</tr>
<tr>
<td>9.3.3</td>
<td>Construction</td>
<td>307</td>
</tr>
<tr>
<td>9.4</td>
<td>Planar Graphs</td>
<td>307</td>
</tr>
<tr>
<td>9.4.1</td>
<td>Orderly Spanning Trees</td>
<td>308</td>
</tr>
<tr>
<td>9.4.2</td>
<td>Triangulations</td>
<td>315</td>
</tr>
<tr>
<td>9.4.3</td>
<td>Construction</td>
<td>317</td>
</tr>
<tr>
<td>9.5</td>
<td>Applications</td>
<td>327</td>
</tr>
<tr>
<td>9.5.1</td>
<td>Binary Relations</td>
<td>327</td>
</tr>
<tr>
<td>9.5.2</td>
<td>RDF Datasets</td>
<td>328</td>
</tr>
<tr>
<td>9.5.3</td>
<td>Planar Routing</td>
<td>330</td>
</tr>
<tr>
<td>9.5.4</td>
<td>Planar Drawings</td>
<td>336</td>
</tr>
<tr>
<td>9.6</td>
<td>Summary</td>
<td>338</td>
</tr>
<tr>
<td>9.7</td>
<td>Bibliographic Notes</td>
<td>338</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Grids</td>
<td>347</td>
</tr>
<tr>
<td>10.1</td>
<td>Wavelet Trees</td>
<td>348</td>
</tr>
<tr>
<td>10.1.1</td>
<td>Counting</td>
<td>350</td>
</tr>
<tr>
<td>10.1.2</td>
<td>Reporting</td>
<td>353</td>
</tr>
<tr>
<td>10.1.3</td>
<td>Sorted Reporting</td>
<td>355</td>
</tr>
</tbody>
</table>
## Table of Contents

### 10.2 $K^2$-Trees
- 10.2.1 Reporting

### 10.3 Weighted Points
- 10.3.1 Wavelet Trees
- 10.3.2 $K^2$-Trees

### 10.4 Higher Dimensions

### 10.5 Applications
- 10.5.1 Dominating Points
- 10.5.2 Geographic Information Systems
- 10.5.3 Object Visibility
- 10.5.4 Position-Restricted Searches on Suffix Arrays
- 10.5.5 Searching for Fuzzy Patterns
- 10.5.6 Indexed Searching in Grammar-Compressed Text

### 10.6 Summary

### 10.7 Bibliographic Notes

### 11 Texts

#### 11.1 Compressed Suffix Arrays
- 11.1.1 Replacing A with $\Psi$
- 11.1.2 Compressing $\Psi$
- 11.1.3 Backward Search
- 11.1.4 Locating and Displaying

#### 11.2 The FM-Index

#### 11.3 High-Order Compression
- 11.3.1 The Burrows-Wheeler Transform
- 11.3.2 High-Order Entropy
- 11.3.3 Partitioning $L$ into Uniform Chunks
- 11.3.4 High-Order Compression of $\Psi$

#### 11.4 Construction
- 11.4.1 Suffix Array Construction
- 11.4.2 Building the BWT
- 11.4.3 Building $\Psi$

#### 11.5 Suffix Trees
- 11.5.1 Longest Common Prefixes
- 11.5.2 Suffix Tree Operations
- 11.5.3 A Compact Representation
- 11.5.4 Construction

#### 11.6 Applications
- 11.6.1 Finding Maximal Substrings of a Pattern
- 11.6.2 Labeled Trees Revisited
- 11.6.3 Document Retrieval
- 11.6.4 XML Retrieval Revisited

#### 11.7 Summary

#### 11.8 Bibliographic Notes
12 Dynamic Structures 450
  12.1 Bitvectors 450
     12.1.1 Solving Queries 452
     12.1.2 Handling Updates 452
     12.1.3 Compressed Bitvectors 461
  12.2 Arrays and Partial Sums 463
  12.3 Sequences 465
  12.4 Trees 467
     12.4.1 LOUDS Representation 469
     12.4.2 BP Representation 472
     12.4.3 DFUDS Representation 474
     12.4.4 Dynamic Range Min-Max Trees 476
     12.4.5 Labeled Trees 479
  12.5 Graphs and Grids 480
     12.5.1 Dynamic Wavelet Matrices 480
     12.5.2 Dynamic $k^2$-Trees 482
  12.6 Texts 485
     12.6.1 Insertions 485
     12.6.2 Document Identifiers 486
     12.6.3 Samplings 486
     12.6.4 Deletions 490
  12.7 Memory Allocation 492
  12.8 Summary 494
  12.9 Bibliographic Notes 494

13 Recent Trends 501
  13.1 Encoding Data Structures 502
     13.1.1 Effective Entropy 502
     13.1.2 The Entropy of RMQs 503
     13.1.3 Expected Effective Entropy 504
     13.1.4 Other Encoding Problems 504
  13.2 Repetitive Text Collections 508
     13.2.1 Lempel-Ziv Compression 509
     13.2.2 Lempel-Ziv Indexing 513
     13.2.3 Faster and Larger Indexes 516
     13.2.4 Compressed Suffix Arrays and Trees 519
  13.3 Secondary Memory 523
     13.3.1 Bitvectors 524
     13.3.2 Sequences 527
     13.3.3 Trees 528
     13.3.4 Grids and Graphs 530
     13.3.5 Texts 534

Index 549