Jewels ^{of} Stringology

This page is intentionally left blank

Maxime Crochemore

Université Marne-la-Vallée, France

Wojciech Rytter

Warsaw University, Poland & University of Liverpool, UK

Jewels *of* Stringology

Published by

World Scientific Publishing Co. Pte. Ltd.

P O Box 128, Farrer Road, Singapore 912805

USA office: Suite 1B, 1060 Main Street, River Edge, NJ 07661

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

JEWELS OF STRINGOLOGY

Copyright © 2002 by World Scientific Publishing Co. Pte. Ltd.

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

ISBN 981-02-4782-6

This book is printed on acid-free paper.

Printed in Singapore by Mainland Press

Preface

The term *stringology* is a popular nickname for *string algorithms* as well as for *text algorithms*. Usually *text* and *string* have the same meaning. More formally, a text is a sequence of symbols. Text is one of the basic data types to carry information. This book is a collection of the most beautiful and at the same time very classical algorithms on strings. The selection has been done by the authors, and is rather personal, among so many famous algorithms that were natural candidates to be included and that belong to a field that has become now fairly popular.

One can partition algorithmic problems discussed in this book into practical and theoretical problems. Certainly string matching and data compression are in the first class, while most problems related to symmetries and repetitions are in the second. However, we believe that all the problems are interesting from an algorithmic point of view and enable the reader to appreciate the importance of combinatorics on words.

In most textbooks on algorithms and data structures the presentation of efficient algorithms on words is quite short as compared to issues in graph theory, sorting, searching, and some other areas. At the same time, there are many presentations of interesting algorithms on words accessible only in journals and in a form directed mainly at specialists. There are still not many books on text algorithms, especially the books which are oriented toward undergraduate and graduate students. In the book the difficult parts are indicated by a star, so the basic text becomes painless for undergraduate students. We hope that this book will cover a gap on algorithms on words in book literature for the broader audience, and bring together the many results presently dispersed in the masses of journal articles.

March 2002

M. Crochemore, W. Rytter

This page is intentionally left blank

Contents

Preface	v
1 Stringology	1
1.1 Text file facilities	2
1.2 Dictionaries	5
1.3 Data compression	6
1.4 Applications of text algorithms in genetics	7
1.5 Efficiency of algorithms	8
1.6 Some notation and formal definitions	10
1.7 Some simple combinatorics of strings	11
1.8 Some other interesting strings	14
1.9 Cyclic shifts and primitive words	16
Bibliographic notes	17
2 Basic string searching algorithms	19
2.1 Knuth-Morris-Pratt algorithm	20
2.2 Boyer-Moore algorithm and its variations	26
Bibliographic notes	31
3 Preprocessing for basic searchings	33
3.1 Preprocessing patterns for <i>MP</i> and <i>KMP</i> algorithms	33
3.2 Table of prefixes	36
3.3 Preprocessing for Boyer-Moore algorithm	39
3.4 * Analysis of Boyer-Moore algorithm	41
Bibliographic notes	44
4 On-line construction of suffix trees	45

4.1 Tries and their compact versions 45

4.2 Prelude to Ukkonen algorithm 49

4.3 Ukkonen algorithm 51

Bibliographic notes 53

5 More on suffix trees 59

5.1 Several applications of suffix trees 59

5.2 McCreight algorithm 63

Bibliographic notes 68

6 Subword graphs 69

6.1 Directed acyclic graph 69

6.2 On-line construction of subword graphs 73

6.3 The reverse perspective 79

6.4 Compact subword graphs 82

Bibliographic notes 84

7 Text algorithms related to sorting 85

7.1 The naming technique: *KMR* algorithm 85

7.2 Two-dimensional *KMR* algorithm 90

7.3 Suffix arrays 91

7.4 Constructing suffix trees by sorting 95

7.5 The Lowest-Common-Anccestor dictionary 101

7.6 Suffix-Merge-Sort 103

Bibliographic notes 107

8 Symmetries and repetitions in texts 111

8.1 Searching for symmetric words 111

8.2 Compositions of symmetric words 114

8.3 Searching for square factors 119

Bibliographic notes 123

9 Constant-space searchings 125

9.1 Constant-space matching for easy patterns 125

9.2 MaxSuffix-Matching 127

9.3 Computation of maximal suffixes 129

9.4 Matching patterns with short maximal suffixes 131

9.5 Two-way matching and magic decomposition 133

9.6 Sequential sampling for unordered alphabets 136

9.7 Galil-Seiferas algorithm 138

9.8 Cyclic equality of words 139

Bibliographic notes 140

10 Text compression techniques 141

10.1 Substitutions 142

10.2 Static Huffman coding 145

10.3 Dynamic Huffman coding 151

10.4 Factor encoding 154

Bibliographic notes 161

11 Automata-theoretic approach 163

11.1 Aho-Corasick automaton 164

11.2 Determinizing automata 173

11.3 Two-way pushdown automata 176

Bibliographic notes 181

12 Approximate pattern matching 183

12.1 Edit distance 183

12.2 Longest common subsequence problem 186

12.3 String matching with errors 191

12.4 String matching with don't care symbols 194

Bibliographic notes 196

13 Matching by dueling and sampling 199

13.1 String matching by duels 199

13.2 String matching by sampling 204

Bibliographic notes 207

14 Two-dimensional pattern matching 209

14.1 Multi-pattern approach 211

14.2 Don't cares and non-rectangular patterns 212

14.3 2D-Pattern matching with mismatches 214

14.4 Multi-pattern matching	215
14.5 Matching by sampling	218
14.6 An algorithm fast on the average	221
Bibliographic notes	222
15 Two-dimensional periodicities	225
15.1 Amir-Benson-Farach algorithm	225
15.2 Geometry of two-dimensional periodicities	235
15.3 * Patterns with large monochromatic centers	242
15.4 * A version of the Galil-Park algorithm	244
Bibliographic notes	247
16 Parallel text algorithms	249
16.1 The abstract model of parallel computing	249
16.2 Parallel string-matching algorithms	252
16.3 * Splitting technique	255
16.4 Parallel <i>KMR</i> algorithm and application	258
16.5 Parallel Huffman coding	263
16.6 Edit distance — efficient parallel computation	268
Bibliographic notes	269
17 Miscellaneous	271
17.1 Karp-Rabin string matching by hashing	271
17.2 Shortest common superstrings	274
17.3 Unique-decipherability problem	276
17.4 Parameterized pattern matching	278
17.5 Breaking paragraphs into lines	281
Bibliographic notes	284
Bibliography	285
Index	305