Algorithms on Strings

This book is intended for lectures on string processing and pattern matching in master's courses of computer science and software engineering curricula. The details of algorithms are given with correctness proofs and complexity analysis, which make them ready to implement. Algorithms are described in a C-like language.

This book is also a reference for students in computational linguistics or computational biology. It presents examples of questions related to the automatic processing of natural language, to the analysis of molecular sequences, and to the management of textual databases.

Professor MAXIME CROCHEMORE received his PhD in 1978 and his Doctorat d'état in 1983 from the University of Rouen. He was involved in the creation of the University of Marne-la-Vallée, where he is currently a professor. He also created the Computer Science Research Laboratory of this university in 1991. Professor Crochemore has been a senior research fellow at King's College London since 2002.

CHRISTOPHE HANCART received his PhD in Computer Science from the University Paris 7 in 1993. He is now an assistant professor in the Department of Computer Science at the University of Rouen.

THIERRY LECROQ received his PhD in Computer Science from the University of Orléans in 1992. He is now a professor in the Department of Computer Science at the University of Rouen.

Algorithms on Strings

MAXIME CROCHEMORE
Université de Marne-la-Vallée

CHRISTOPHE HANCART
Université de Rouen

THIERRY LECROQ
Université de Rouen

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521848992

Originally published in French as
Algorithmique du texte by Maxime Crochemore, Christophe Hancart, Thierry Lecroq

English edition (a translation from the French-language edition) first published by
Cambridge University Press

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data
Crochemore, Maxime, 1947–
Algorithms on strings / Maxime Crochemore, Christophe Hancart, Thierry Lecroq.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-521-84899-2 (hardback)
ISBN-10: 0-521-84899-7 (pbk.)
1. Computer algorithms. 2. Matching theory. 3. Computational biology.
I. Hancart, Christophe, 1964– II. Lecroq, Thierry. III. Title.
QA76.9.A43C757 2007
005.1–dc22 2006039263

ISBN 978-0-521-84899-2 hardback

Contents

Preface

This book presents a broad panorama of the algorithmic methods used for processing texts. For this reason it is a book on algorithms, but whose object is focused on the handling of texts by computers. The idea of this publication results from the observation that the rare books entirely devoted to the subject are primarily monographs of research. This is surprising because the problems of the field have been known since the development of advanced operating systems, and the need for effective solutions becomes essential because the massive use of data processing in office automation is crucial in many sectors of the society. In 1985, Galil pointed out several unsolved questions in the field, called after him, Stringology (see [12]). Most of them are still open.

In a written or vocal form, text is the only reliable vehicle of abstract concepts. Therefore, it remains the privileged support of information systems, despite of significant efforts toward the use of other media (graphic interfaces, systems of virtual reality, synthesis movies, etc.). This aspect is still reinforced by the use of knowledge databases, legal, commercial, or others, which develop on the Internet. Thanks, in particular, to the Web services.

The contents of the book carry over into formal elements and technical bases required in the fields of information retrieval, of automatic indexing for search engines, and more generally of software systems, which includes the edition, the treatment, and the compression of texts. The methods that are described apply to the automatic processing of natural languages, to the treatment and analysis of genomic sequences, to the analysis of musical sequences, to problems of safety and security related to data flows, and to the management of the textual databases, to quote only some immediate applications.

The selected subjects address pattern matching, the indexing of textual data, the comparison of texts by alignment, and the search for local regularities. In addition to their practical interest, these subjects have theoretical and combinatorial aspects that provide astonishing examples of algorithmic solutions.

The goal of this work is principally educational. It is initially aimed at graduate and undergraduate students, but it can also be used by software designers.

We warmly thank the researchers who took time to read and comment on the preliminary outlines of this book. They are Saïd Abdeddaïm, Marie-Pierre Béal, Christian Charras, Raphaël Clifford, Christiane Frougny, Gregory Kucherov, Sabine Mercier, Laurent Mouchard, Johann Pelfrêne, Bruno Petazzoni, Mathieu Raffinot, Giuseppina Rindone, and Marie-France Sagot. Remaining flaws are ours.

Finally, extra elements to the contents of the book are accessible on the site `http://chl.univ-mlv.fr` or from the Web pages of the authors.

MAXIME CROCHEMORE
CHRISTOPHE HANCART
THIERRY LECROQ
Marne-la-Vallée, London, Rouen
June 2006