Dissertation Title

Dissertation zur Erlangung des Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.) vorgelegt von

Enrico Siragusa



am Fachbereich Mathematik und Informatik der Freien Universität Berlin

Berlin 201X

Datum des Disputation: $\boldsymbol{XX.XX.201X}$

Gutachter:

Prof. Dr. Knut Reinert, Freie Universität Berlin, Deutschland

 $\boldsymbol{\textit{Prof. Dr. XXX~XXX}, XXX}, XXX$

Abstract

Bla bla bla.

CONTENTS

Pa	art I	Approximate String Matching	1
1.	Alg	orithms for Edit Distance Verification	3
	1.1	Banded Myers' bit-vector algorithm	3
	1.2	Increased bit-parallelism	3
Pa	art II	I Full-Text Indices	5
2.	Clas	ssic Full-Text Indices	7
	2.1	Trie	7
	2.2	Suffix trie and suffix tree	7
	2.3	Suffix array	7
		2.3.1 Forward search	7
	2.4	q-Gram index	7
3.	Con	npressed Full-Text Indices	9
	3.1	Rank dictionaries	9
		3.1.1 Sequential dictionaries	9
		3.1.2 Wavelet tree	9
	3.2	Burrows-Wheeler transform	9
		3.2.1 Construction	9
		3.2.2 Inversion	9
	3.3	FM-index	9
		3.3.1 Backward search	9
		3.3.2 Locating occurrences	9
1	Onl	ine Approximate String Matching	11
т.	4.1		11
	4.2		11
		morewood one personal domes on the decisions and the control of th	
Da	.n≠ T1	II Applications	13
Γċ	art 11	II Applications	10
5.	Rea	11 0	15
	5.1	Related work	15
	5.2	The Masai mapper	15
		5.2.1 Single-end mapping	15
		5.2.2 Paired-end mapping	15

	5.3	The M	asai 2 mapper	15
		5.3.1	Single-end mapping	15
		5.3.2	Paired-end mapping	15
		5.3.3	Parallelization	15
		5.3.4	Hardware acceleration	15
	5.4	Experi	mental results	15
		5.4.1	Comparison of filtration strategies	15
		5.4.2	Rabema benchmark results	15
		5.4.3	Variant detection results	15
		5.4.4	Runtime results	15
	5.5	Discuss	sion	15
6.	Stri	ng Sim	nilarity Search / Join	17
	6.1	The co	ompetition	17
	6.2	Our m	ethod	17
		6.2.1	Implementation	17
		6.2.2	Parallelization	17
	6.3	Related	d methods	17
	6.4	Experi	mental results	17
		6.4.1	Search	17
		6.4.2	Join	17
	6.5	Discuss	sion	17
Α.		_	filters	19
			ped seeds	19
	A.2		d seeds	19
		A.2.1	<i>y</i> 01 1	19
		A.2.2	Threshold computation	19
		A.2.3	Sensitivity computation	19
D	Dool	laratio	n	21
ப.	Dec.	iai at 10.	ш	41
Bi	bliog	raphy		22

Part I APPROXIMATE STRING MATCHING

1

Algorithms for Edit Distance Verification

- 1.1 Banded Myers' bit-vector algorithm
- 1.2 Increased bit-parallelism

Part II FULL-TEXT INDICES

Classic Full-Text Indices

2

- 2.1 Trie
- 2.2 Suffix trie and suffix tree
- 2.3 Suffix array
- 2.3.1 Forward search
- 2.4 q-Gram index

3 Compressed Full-Text Indices

- 3.1 Rank dictionaries
- 3.1.1 Sequential dictionaries
- 3.1.2 Wavelet tree
- 3.2 Burrows-Wheeler transform
- 3.2.1 Construction
- 3.2.2 Inversion
- 3.3 FM-index
- 3.3.1 Backward search
- 3.3.2 Locating occurrences

4

Online Approximate String Matching

- 4.1 Banded Myers' bit-vector algorithm
- 4.2 Increased bit-parallelism using SIMD instructions

Part III APPLICATIONS

Read Mapping

– 1	D - 1 - 41	1
5.1	Related	work
$\mathbf{O} \cdot \mathbf{L}$	Itolatoa	

- 5.2 The Masai mapper
- 5.2.1 Single-end mapping
- 5.2.2 Paired-end mapping
- 5.3 The Masai 2 mapper
- 5.3.1 Single-end mapping
- 5.3.2 Paired-end mapping
- 5.3.3 Parallelization
- 5.3.4 Hardware acceleration
- 5.4 Experimental results
- 5.4.1 Comparison of filtration strategies
- 5.4.2 Rabema benchmark results
- 5.4.3 Variant detection results
- 5.4.4 Runtime results
- 5.5 Discussion

6 String Similarity Search / Join

- 6.1 The competition
- 6.2 Our method
- 6.2.1 Implementation
- 6.2.2 Parallelization
- 6.3 Related methods
- 6.4 Experimental results
- 6.4.1 Search
- 6.4.2 Join
- 6.5 Discussion

Appendix

Counting filters

A

- A.1 Ungapped seeds
- A.2 Gapped seeds
- A.2.1 Underlying principle
- A.2.2 Threshold computation
- A.2.3 Sensitivity computation

APPENDIX

В

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Enrico Siragusa October 5, 2013

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

LIST OF NOTATIONS

bp	base pair, character of the alphabet $\{A,C,G,T\}$??
$\overline{\mathrm{Mb}}$	megabase, 1 million base pairs or characters	??
MB	megabyte, $1 \text{ MB} = 1024 \text{ kB} = 1,048,576 \text{ byte}$??
GB	gigabyte, $1 \text{GB} = 1024 \text{MB}$??
Σ, Ψ, Φ	finite alphabets	??
\sum^*	set of all possible strings over the alphabet Σ	??
\sum_{n}^{n}	set of all possible strings over the alphabet Σ with length n	??
ϵ	empty string	??
s	length of string s	??
s[i]	character of s at position i (counting from 0)	??
$s_i, \operatorname{suf}(s, i)$	suffix of s beginning at position i	??
[ij]	set of integers $i, i+1, \ldots, j$??
[ij)	set of integers $i, i+1, \ldots, j-1$??
\mathbb{N}_0	set of non-negative integers	??
\prec, \preceq	strict and non-strict substring relation	??
$\langle \dots \dots \rangle$	definition of a string analogous to the set notation	??
lcp S	longest common prefix of a set $\mathcal S$ of strings $\ldots \ldots$??
$<_{ m lex}$	lexicographical order	??
$<_q$	lexicographical prefix order, compares only prefixes of length q	??
$\$, \j	(virtual) sentinel characters to well-define the suffix tree	??
concat(v)	edge label concatenation on the path from root to suffix tree node	??
· /		??
<u>s</u>	suffix tree node whose edge label concatenation is $s cdots$??
$\operatorname{rank}(a)$	rank of character a in the underlying alphabet	??
R, D, I	edit operations that replace, delete, or insert a character	??
$ T _{\mathrm{E}}$	number of edit operations in transcript T	??
$ T _{\mathbf{R}}$	number of matches, replacements, and deletions in transcript T	??
\mathcal{R}	sequenced reads, set of strings	
G	reference sequence, string	??
\mathcal{D}	database, set of strings	??
$\operatorname{freq}(\phi, \mathcal{D})$	absolute number of strings in \mathcal{D} that contain ϕ at least once.	??
$\operatorname{supp}(\phi,\mathcal{D})$	relative number of strings in \mathcal{D} that contain ϕ at least once	??