Text Indexing

Lecture 12: Longest Common Extensions

Florian Kurpicz

The slides are licensed under a Creative Commons Attribution-ShareAlike 4.0 International License: www.creativecommons.org/licenses/by-sa/4.0 | commit 47a39dc compiled at 2023-01-30-13:05

Recap: Document Listing and Top-k Retrieval

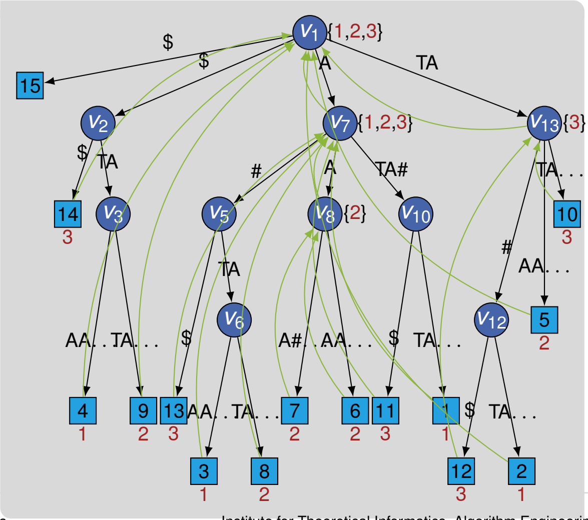
Definition: Document Listing

Given a collection of D documents $\mathcal{D} = \{d_1, d_2, \ldots, d_D\}$ containing symbols from an alphabet $\Sigma = [1, \sigma]$ and a pattern $P \in \Sigma^*$, return all $j \in [1, D]$, such that d_j contains P.

- $d_1 = \text{ATA}$
- $d_2 = \text{TAAA}$
- $d_3 = \text{TATA}$

And for queries:

- $P = \text{TA}$ is contained in $d_1, d_2,$ and d_3
- $P = \text{ATA}$ is contained in d_1 and d_3



Recap: Inverted Index and List Encodings

Karlsruhe Institute of Technology

Definition: Inverted Index

Given a set of documents and terms that are contained in the documents, an inverted index stores the terms and associated with each term t

- the number of documents f_t that contain t and
- an ordered list $L(t)$ of documents containing t

List Encodings

- Δ-encoding
- unary- and ternary-encoding
- Elia-γ and -δ-encoding
- Golomb- and Fibonacci-encoding

```
1 The old night keeper keeps the keep in the town
2 In the big old house in the big old gown
3 The house in the town had the big old keep
4 Where the old night keeper never did sleep
5 The night keeper keeps the keep in the night
6 And keeps in the dark and sleeps in the light
```

term t	f_t	$L(t)$
and	1	[6]
big	2	[2, 3]
dark	1	[6]
...	...	...
had	1	[3]
house	2	[2, 3]
in	5	[1, 2, 3, 5, 6]
...	...	...

Recap: Pattern Matching with the LCP-Array (1/3)

- remember how many characters of the pattern and suffix match
- identify how long the prefix of the old and next suffix is
- do so using the LCP-array and
- range minimum queries ℹ detailed introduction in **Advanced Data Structures**

- $lcp(i,j) = \max\{k \colon T[i..i+k)$
- $lcp(i,j) = T[j..j+k)\} = LCP[RMQ_{LCP}(i+1,j)]$
- RMQs can be answered in $O(1)$ time and
- require $O(n)$ space

Definition: Range Minimum Queries

Given an array $A[1..m)$, a **range minimum query** for a range $\ell \leq r \in [1,n)$ returns

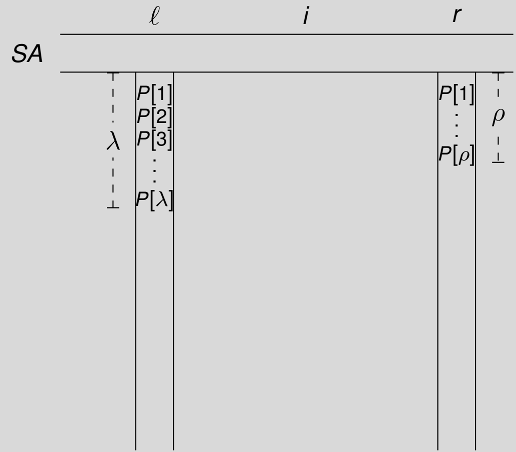
$$RMQ_A(\ell, r) = \arg\min\{A[k] \colon k \in [\ell, r]\}$$

Recap: Pattern Matching with the LCP-Array (2/3)

- during binary search matched
- λ characters with left border ℓ and
- ρ characters with right border r
- w.l.o.g. let $\lambda \geq \rho$

- middle position i
- decide if continue in $[\ell, i]$ or $[i, r]$

- let $\xi = lcp(SA[\ell], SA[i])$ ℹ $O(1)$ time with RMQs



Recap: Pattern Matching with the LCP-Array (3/3)

- let $\xi = lcp(SA[\ell], SA[i])$

$\xi > \lambda$

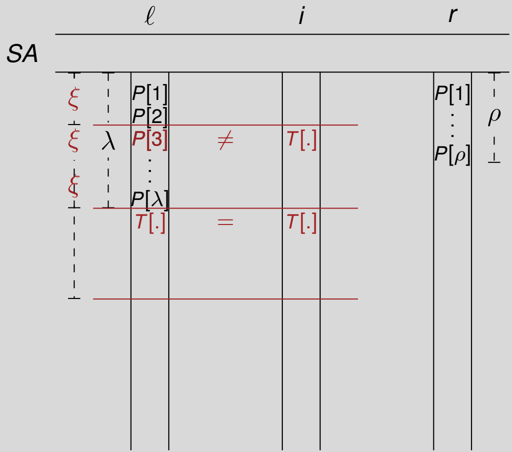
- $P[\lambda + 1] > T[SA[\ell] + \lambda] = T[SA[i] + \lambda]$
- $\ell = i$ without character comparison

$\xi = \lambda$

- compare as before

$\xi < \lambda$

- $\xi \geq \rho$ and $P[\xi + 1] < T[SA[i] + \xi]$
- $r = i$ and $\rho = \xi$ without character comparison



Old Problem, New Name

Definition: Longest Common Extensions

Given a text T of size n over an alphabet of size σ, construct data structure that answers for $i, j \in [1, n]$

$$\mathrm{lce}_T(i,j) = \max\{\ell \geq 0 \colon T[i, i+\ell) = T[j, j+\ell)\}$$

- also denoted as $lcp(i,j)$ ℹ in this lecture

Applications

- (sparse) suffix sorting
- approximate pattern matching
- ...

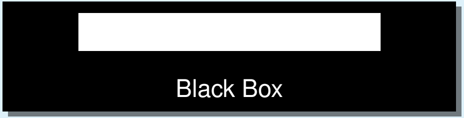
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
T	A	B	C	D	A	B	C	C	D	B	C	C	B	A	B	C	D	A	D	A

$$\mathrm{lce}_T(1, 14) = 0\ 1\ 2\ 3\ 4\ 5$$

Practical Algorithms for Longest Common Extensions [IT09]

Sophisticated Black Box (BB)

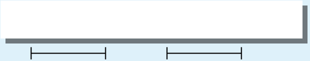
- based on ISA, LCP, and RMQ



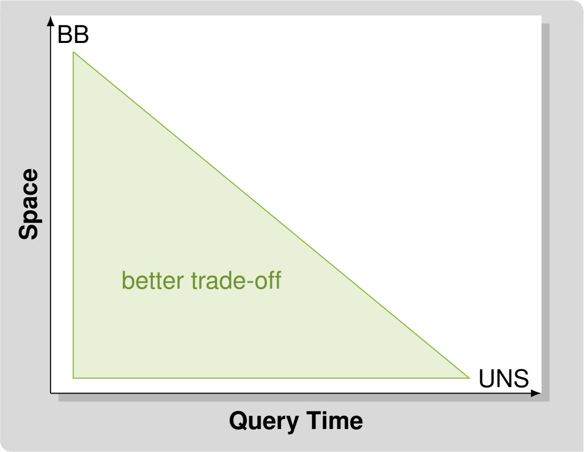
- $O(1)$ query time, $\approx 9n$ bytes additional space

Ultra Naive Scan (UNS)

- compare character by character



- $O(n)$ query time, no additional space



Monte Carlo and Las Vegas Algorithms

- setting: randomized algorithms

Monte Carlo Algorithm

- returns wrong result with small probability
- deterministic running time

Las Vegas Algorithm

- returns correct result
- only expected running time

- some Monte Carlo algorithms can be turned into Las Vegas algorithms
- depends on correctness check
- all Monte Carlo algorithms presented today can be turned into Las Vegas algorithms

Randomized String Matching

- compute 🫆s of strings
- application not limited to LCEs

Definition: Karp-Rabin Fingerprint [KR87]

Given a text T of length n over an alphabet of size σ and a random prime number $q \in \Theta(n^c)$, the Karp-Rabin fingerprint of $T[i..j]$ is

$$🫆(i, j) = \left(\sum_{k=i}^{j} T[k] \cdot \sigma^{j-k}\right) \bmod q$$

ℹ $(x + y) \bmod z = z \bmod z + y \bmod z \pmod z$

- if $T[i..i+\ell] = T[j..j+\ell]$, then

$$🫆(i, i+\ell) = 🫆(j, j+\ell)$$

- if $T[i..i+\ell] \neq T[j..j+\ell]$, then

$$\mathrm{Prob}(🫆(i, i+\ell) = 🫆(j, j+\ell)) \in O\left(\frac{\ell \lg \sigma}{n^c}\right)$$

- prime has to be chosen uniformly at random
- how to turn it into Las Vegas algorithm?

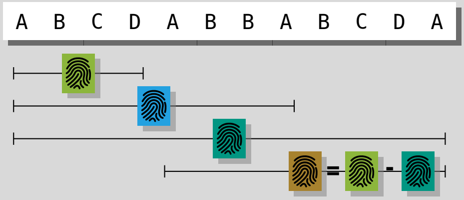
- example on the board

Overwriting the Text with Fingerprints (1/2) [Pre18]

- given a text T over an alphabet of size σ
- let w be size of a computer word ℹ e.g., 64 bit
- choose $\tau \in \Theta(w/\lg \sigma)$ ℹ 8 for byte alphabet
- choose random prime $q \in [\frac{1}{2}\sigma^\tau, \sigma^\tau)$
- group the text into size-τ blocks: $B[1..n/\tau]$ with

$$B[i] = T[(i-1)\tau + 1..i\tau]$$

- compute $P'[i] =$ 🫆$(i, \tau i)$ for $i \in [1, n/\tau]$
- $P'[i]$ can fits in $B[i]$



- overwrite text with fingerprints (in-place)

- all parts of text are restorable
- how?

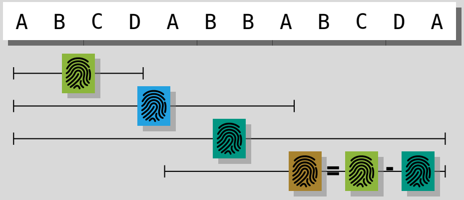
Overwriting the Text with Fingerprints (2/2)

- choose random prime $q \in [\frac{1}{2}\sigma^\tau, \sigma^\tau)$
- $B[i] = T[(i-1)\tau + 1..i\tau]$
- $\lfloor B[i]/q \rfloor \in \{0, 1\}$
- $D[i] = \lfloor B[i]/q \rfloor$ ℹ bit vector of size n/τ
- $P'[i] =$ 🫆$(i, \tau i)$ and together with D:

$$B[i] = (P'[i] - \sigma^\tau \cdot P'[i-1] \bmod q) + D[i] \cdot q$$

- this gives us access to the text(!)

- q can be chosen such that MSB of $P'[i]$ is zero w.h.p., then
- D can be stored in the MSBs



- overwrite text with fingerprints (in-place)

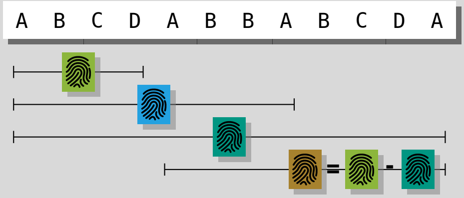
- enough to answer LCE queries
- how?

Answering LCE Queries with Fingerprints

LCEs with Fingerprints

- compute LCE of i and j
- exponential search until 🫆$(i, i + 2^k) \neq$ 🫆$(j, j + 2^k)$
- binary search to find correct block m
- recompute $B[m]$ and find mismatching character

- requires $O(\lg \ell)$ time for LCEs of size ℓ



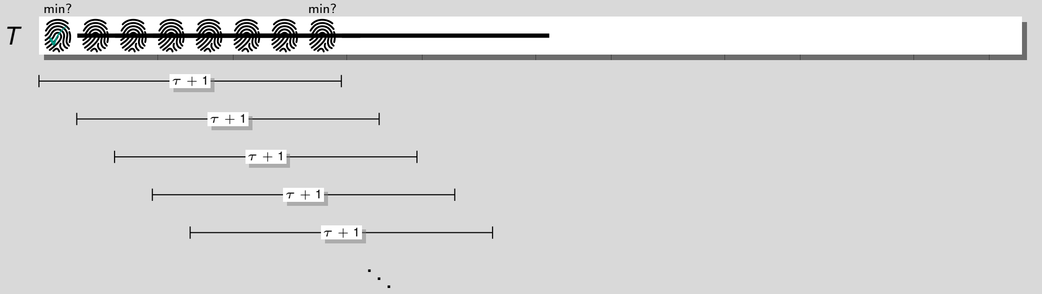
- overwrite text with fingerprints (in-place)

String Synchronizing Sets (Simplified, 1/2)

Definition: Simplified τ-Synchronizing Sets [KK19]

Given a text T of length n and $0 < \tau \leq n/2$, a simplified τ-synchronizing set S of T is

$$S = \{i \in [1, n - 2\tau + 1] \colon \mathit{min}\{\mathrm{fp}(j, j + \tau - 1) \colon j \in [i, i + \tau]\} \in \{\mathrm{fp}(i, i + \tau - 1), \mathrm{fp}(i + \tau, i + 2\tau - 1)\}\}$$



String Synchronizing Sets (Simplified, 2/2)

- $|S| = \Theta(n/\tau)$ in practice (on most data sets)
- more complex definition required to obtain this size

Consistency & (Simplified) Density Property

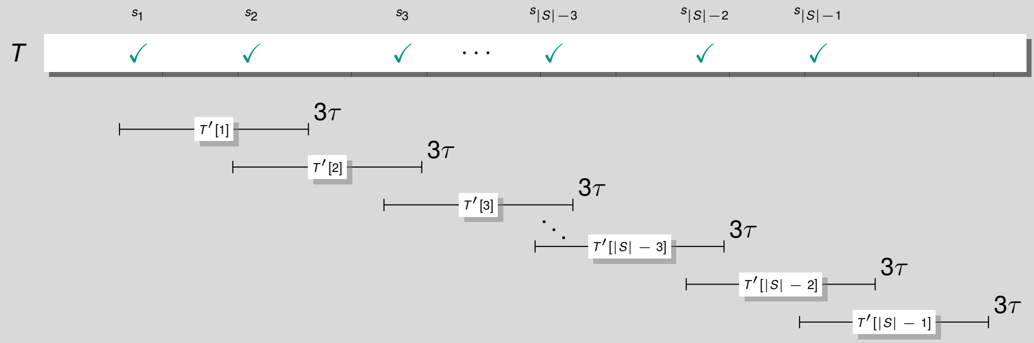
- for all $i, j \in [1, n-2\tau+1]$ we have

$$T[i, i{+}2\tau{-}1] = T[j, j{+}2\tau{-}1] \Rightarrow i \in S \Leftrightarrow j \in S$$

- for any τ consecutive positions there is at least one position in S

Answering LCE Queries with String Synchronizing Sets (1/2)

Text T' for Positions in S

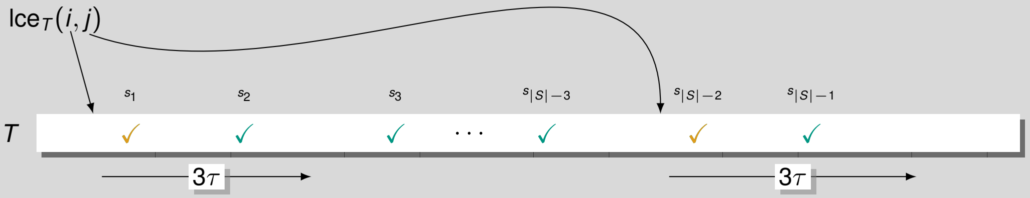


Answering LCE Queries with String Synchronizing Sets (2/2)

- in practice, we sort the substrings
- characters of T' are the ranks of substrings
- build BB LCE for T' w.r.t. length in T

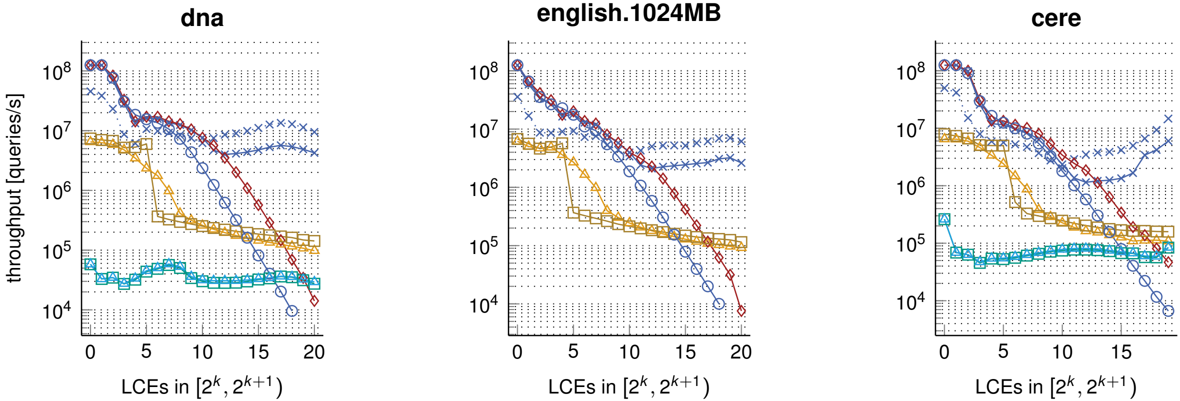
Answering Queries

- compare naively for 3τ characters
- if equal find successors of i and j in S
- compute LCE of successors in T'

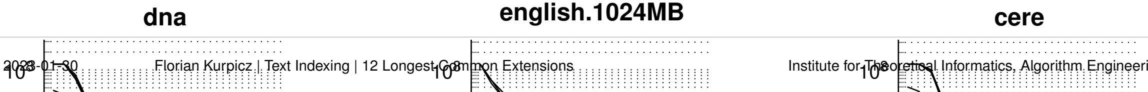


- in this example: $\mathrm{lce}_T(i,j) = s_1 - i + \mathrm{lce}_{T'}(1, |S| - 2)$
- in practice: we have a very fast static successor data structure

Practical Evaluation [Din+20]



??



Conclusion and Outlook

Karlsruhe Institute of Technology

This Lecture

- longest common extension queries
- Karp-Rabin fingerprints
- string synchronizing sets

Next Lecture

- big recap and Q&A

Thats all! We are (mostly) done.