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

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

- $\Delta$-encoding
- unary- and ternary-encoding
- Elia-$\gamma$ and -$\delta$-encoding
- Golomb- and Fibonacci-encoding

<table>
<thead>
<tr>
<th>term</th>
<th>$f_t$</th>
<th>$L(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>and</td>
<td>1</td>
<td>[6]</td>
</tr>
<tr>
<td>big</td>
<td>2</td>
<td>[2, 3]</td>
</tr>
<tr>
<td>dark</td>
<td>1</td>
<td>[6]</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>had</td>
<td>1</td>
<td>[3]</td>
</tr>
<tr>
<td>house</td>
<td>2</td>
<td>[2, 3]</td>
</tr>
<tr>
<td>in</td>
<td>5</td>
<td>[1, 2, 3, 5, 6]</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

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
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 \( \text{detailed introduction in Advanced Data Structures} \)

**Definition: Range Minimum Queries**

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

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

- \( \text{lcp}(i, j) = \max \{ k : T[i..i+k) \} \)
- \( \text{lcp}(i, j) = T[j..j+k) = \text{LCP}[\text{RMQ}_{\text{LCP}}(i+1, j)] \)
- RMQs can be answered in \( O(1) \) time and
- require \( O(n) \) space
Recap: Pattern Matching with the LCP-Array (2/3)

- during binary search matched
- $\lambda$ characters with left border $\ell$ and
- $\rho$ 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 $\xi = lcp(SA[\ell], SA[i])$ $\in O(1)$ time with
  RMQs
Recap: Pattern Matching with the LCP-Array (3/3)

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

$\xi > \lambda$
- $P[\lambda + 1] > T[SA[\ell] + \lambda] = T[SA[i] + \lambda]$
- $\ell = i$ 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
Definition: Longest Common Extensions

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

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

Applications

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

Also denoted as $lcp(i, j)$ in this lecture.

Example:

<table>
<thead>
<tr>
<th>$i$</th>
<th>$j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14</td>
</tr>
</tbody>
</table>

$\text{lce}_T(1,14) = 0 1 2 3 4 5$
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 fingerprints of strings
- application not limited to LCEs

**Definition: Karp-Rabin Fingerprint [KR87]**

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

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

$(x + y) \mod z = z \mod z + y \mod z \mod z$ (mod z)

- if $T[i..i + \ell] = T[j..j + \ell]$, then
  $$\mathcal{F}(i, i + \ell) = \mathcal{F}(j, j + \ell)$$

- if $T[i..i + \ell] \neq T[j..j + \ell]$, then
  $$\text{Prob}(\mathcal{F}(i, i + \ell) = \mathcal{F}(j, j + \ell)) \in O\left(\frac{\ell \log \sigma}{n^c}\right)$$

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

- example on the board
given a text $T$ over an alphabet of size $\sigma$
let $w$ be size of a computer word e.g., 64 bit
choose $\tau \in \Theta(w/\log \sigma)$ 8 for byte alphabet
choose random prime $q \in \left[\frac{1}{2}\sigma^\tau, \sigma^\tau\right)$
group the text into size-$\tau$ blocks: $B[1..n/\tau]$ with
\[ B[i] = T[(i - 1)\tau + 1..i\tau] \]
compute $P'[i] = \mathcal{F}(i, \tau i)$ 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?
choose random prime \( q \in \left[ \frac{1}{2}\sigma^{\tau}, \sigma^{\tau} \right) \)

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

\[
\lfloor B[i]/q \rfloor \in \{0, 1\}
\]

\[
D[i] = \lfloor B[i]/q \rfloor \text{ bit vector of size } n/\tau
\]

\[
P'[i] = (i, \tau i) \text{ and together with } D:
\]

\[
B[i] = (P'[i] - \sigma^{\tau} \cdot P'[i - 1] \mod q) + D[i] \cdot q
\]

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?
LCEs with Fingerprints

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

- requires $O(\lg \ell)$ time for LCEs of size $\ell$
Definition: Simplified $\tau$-Synchronizing Sets [KK19]

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

$$S = \{ i \in [1, n - 2\tau + 1] : \min \{ (j, j + \tau - 1) : j \in [i, i + \tau] \} \in \{ (i, i + \tau - 1), (i + \tau, i + 2\tau - 1) \} \}$$
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 $\tau$ consecutive positions there is at least one position in $S$
## Answering LCE Queries with String Synchronizing Sets (1/2)

### Text $T'$ for Positions in $S$

| $s_1$ | $s_2$ | $s_3$ | $s_{|S| - 3}$ | $s_{|S| - 2}$ | $s_{|S| - 1}$ |
|-------|-------|-------|----------------|----------------|----------------|
| ✓     | ✓     | ✓     | ✓              | ✓              | ✓              |

- $T'[1]$ to $T'[|S| - 3]$: Each position $i$ in $T$ is marked ✓ if there is a longest common extension of $T$ with $S$ starting at position $i$.
- $3\tau$: The length of the extension is $3\tau$.

---

2023-01-30 Florian Kurpicz | Text Indexing | 12 Longest Common Extensions

Institute for Theoretical Informatics, Algorithm Engineering
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 \( 3\tau \) characters
- if equal find successors of \( i \) and \( j \) in \( S \)
- compute LCE of successors in \( T' \)

in this example: \( \text{lce}_T(i, j) = s_1 - i + \text{lce}_{T'}(1, |S| - 2) \)

in practice: we have a very fast static successor data structure
Practical Evaluation [Din+20]

LCEs in $[2^k, 2^{k+1})$

throughput [queries/s]

DNA

english.1024MB

cereal
Conclusion and Outlook

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

Next Lecture
- big recap and Q&A

That's all! We are (mostly) done.