Text Indexing

Lecture 13: Recap and Q&A

Florian Kurpicz
Warning

This is just a very succinct overview. Please refer to the lecture slides for more details.
Tries & Suffix Trees

Trie Representations

- different trie representations
- space-time trade-off

Suffix Tree (Compact Trie)
### Suffix Array

Given a text $T$ of length $n$, the **suffix array** (SA) is a permutation of $[1..n]$, such that for $i \leq j \in [1..n]$

$$T[SA[i]..n] \leq T[SA[j]..n]$$

#### SA Construction in EM

- Prefix Doubling
- DC3

#### 1 2 3 4 5 6 7 8 9 10 11 12 13

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>a</td>
<td>b</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>a</td>
<td>b</td>
<td>b</td>
<td>a</td>
<td>$</td>
</tr>
<tr>
<td>$SA$</td>
<td>13</td>
<td>12</td>
<td>1</td>
<td>9</td>
<td>6</td>
<td>3</td>
<td>11</td>
<td>2</td>
<td>10</td>
<td>7</td>
<td>4</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>$LCP$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

### SAIS

- linear time suffix array construction
- induced copying and recursion
  - classification
  - sorting special suffixes
  - inducing other suffixes
### LCP-Array & LCE-Queries

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>a</td>
<td>b</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>a</td>
<td>b</td>
<td>a</td>
<td>b</td>
<td>a</td>
</tr>
<tr>
<td>$SA$</td>
<td>13</td>
<td>12</td>
<td>1</td>
<td>9</td>
<td>6</td>
<td>3</td>
<td>11</td>
<td>2</td>
<td>10</td>
<td>7</td>
<td>4</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>$LCP$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

- speed up pattern matching in suffix array
- suffix tree construction
- compression

**Longest Common Extensions**

- lcp-value between any suffix
- scan or RMQ
- Rabin-Karp fingerprints
- string synchronizing sets
Compression

Entropy

Given a text $T$ of length $n$ over an alphabet $\Sigma = [1, \sigma]$ and its histogram $Hist$, then

$$H_k = (1/n) \sum_{S \in \Sigma^k} |T_S| \cdot H_0(T_S)$$

Huffman Codes

- variable length codes
- more frequent characters get shorter codes
- canonical Huffman-codes
- Shannon-Fano codes can be worse, but
- are still optimal

LZ77

Given the text $T = abababbbaba\$:

- $f_1 = a$
- $f_2 = b$
- $f_3 = abab$
- $f_4 = bbb$
- $f_5 = aba$
- $f_6 = \$

LZ78

Given the text $T = abababbbaba\$:

- $f_1 = a$
- $f_2 = b$
- $f_3 = ab$
- $f_4 = abb$
- $f_5 = bb$
- $f_6 = aba$
- $f_7 = \$
Burrows-Wheeler Transform

Given a text $T$ of length $n$ and its suffix array $SA$, for $i \in [1, n]$ the **Burrows-Wheeler transform** is

$$BWT[i] = \begin{cases} T[SA[i] - 1] & SA[i] > 1 \\ \$ & SA[i] = 1 \end{cases}$$

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
</table>
| $T$ | a | b | a | b | c | a | b | c | a | b | b | a | $\$
| $SA$ | 13 | 12 | 1 | 9 | 6 | 3 | 11 | 2 | 10 | 7 | 4 | 8 | 5 |
| $BWT$ | a | b | $\$ | c | c | b | b | a | a | a | a | a | b | b |

LF-Mapping

Given a $BWT$, its $C$-array, and its rank-Function, then

$$LF(i) = C[BWT[i]] + rank_{BWT[i]}(i)$$

- transform back to text
- used in backwards search

Compression using BWT

- move-to-front
- run-length compression
Wavelet Tree

Wavelet Tree

- generalize rank and select to alphabets of size > 2

Wavelet Matrix

Compression

- build over text compressed with canonical Huffman codes

Bit Vectors

- rank and select queries on bit vectors in $O(1)$ time and $o(n)$ space
Function $BackwardsSearch(P[1..n], C, rank)$:

1. $s = 1$, $e = n$
2. For $i = m, \ldots, 1$ do
3.   $s = C[P[i]] + rank_{P[i]}(s - 1) + 1$
4.   $e = C[P[i]] + rank_{P[i]}(e)$
5.   If $s > e$ then
6.      return $\emptyset$
7. return $[s, e]$

**FM-Index**
- Use (compressed wavelet tree for rank)
- Compress bit vectors further

**r-Index**
- Store lots of arrays
- Containing information for each run
- Size proportional to number of runs
- Queries become harder
Compressed Indices

**Block Tree**
- answer rank and select queries
- size proportional to number of LZ-factors

**Number of Runs and LZ-Factors**

Let $T$ be a text of length $n$, then

$$r(T) \in O(z(T) \log^2 n)$$
The old night keeper keeps the keep in the town
In the big old house in the big old gown
The house in the town had the big old keep
Where the old night keeper never did sleep
The night keeper keeps the keep in the night
And keeps in the dark and sleeps in the light

<table>
<thead>
<tr>
<th>term</th>
<th>(t_i)</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>

Encodings
- unary/ternary encoding
- Fibonacci encoding
- Elias-\(\delta/\gamma\) encoding
- Golomb encoding

List Intersection
- binary/exponential search
- two levels
Document Retrieval

### Document Listing
- optimal with document array and chain array

<table>
<thead>
<tr>
<th>T</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>15</td>
<td>14</td>
<td>4</td>
<td>9</td>
<td>13</td>
<td>3</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>11</td>
<td>1</td>
<td>12</td>
<td>2</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>DA</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>CA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td>10</td>
<td>11</td>
<td>9</td>
<td>12</td>
</tr>
</tbody>
</table>

- $P = TA$

Diagagram showing a graph with vertices labeled $V_1, V_2, V_3, V_4, V_5, V_6, V_7, V_8, V_9, V_{10}, V_{11}, V_{12}, V_{13}$. The graph includes arrows indicating the relationships between the vertices. The vertices are labeled with sets such as $\{1,2,3\}$ and characters like TA. The document array $SA$, chain array $DA$, and chain array $CA$ are also included in the table.