Improving FM Index

Juha Kärkkäinen\textsuperscript{1}

\textsuperscript{1}University of Helsinki, Finland

Algodan Seminar

October 28, 2011
**FM Index** (Ferragina, Manzini)

Practical succinct self-index of a text

- Pattern counting

<table>
<thead>
<tr>
<th>space (bits)</th>
<th>query time</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n \log \sigma + o(n \log \sigma)$</td>
<td>$\mathcal{O}(m \log \sigma)$</td>
</tr>
</tbody>
</table>

$n =$ text length

$\sigma =$ alphabet size

$m =$ pattern length

Main components

- Burrows–Wheeler transform (BWT)
- Wavelet tree
- Binary rank index
Text Compression

Text $T$: $n = |T|$  $\sigma =$ alphabet size

Three levels of compression

1. No compression: $n \log \sigma$ bits
2. Zeroth order compression: $nH_0(T)$ bits
3. Higher order compression: $nH_k(T)$ bits

$H_k(T)$ is the $k$th order empirical entropy of $T$

- $\log \sigma \geq H_0(T) \geq H_k(T)$
Zeroth Order Compression of FM Index

- Huffman-shaped wavelet tree (HWT)
- RRR binary rank index

<table>
<thead>
<tr>
<th>Compression technique</th>
<th>space</th>
<th>query time</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>$n \log \sigma + o(n \log \sigma)$</td>
<td>$\mathcal{O}(m \log \sigma)$</td>
</tr>
<tr>
<td>HWT</td>
<td>$nH_0(T) + n + o(nH_0(T) + n)$</td>
<td>$\mathcal{O}(mH_0(T))$ on average</td>
</tr>
<tr>
<td>RRR</td>
<td>$nH_0(T) + o(n \log \sigma)$</td>
<td>slower by a constant factor</td>
</tr>
<tr>
<td>HWT + RRR</td>
<td>$nH_0(T) + o(nH_0(T) + n)$</td>
<td></td>
</tr>
</tbody>
</table>
Higher Order Compression by Compression Boosting

- Partition $\text{BWT}(T)$ into blocks suitably
- Compress each block using zeroth order compressor
- The result is higher order compression of $T$

**Theorem (Manzini)**

*For any text $T$ and any $k$, there exists a partitioning $L_1 L_2 \ldots L_\ell$ of $\text{BWT}(T)$ into $\ell \leq \sigma^k$ blocks so that*

$$\sum_{i} |L_i| H_0(L_i) = |T| H_k(T)$$
Compression Boosting of FM Index

- Context block boosting (Ferragina et al.)
  - Optimal partitioning of BWT into varying size context blocks
- Implicit boosting (Mäkinen & Navarro)
  - Using RRR achieves the boosting effect
- FM index with either boosting method needs
  \[ nH_k(T) + o(n \log \sigma) \]
  bits of space
Fixed Block Boosting  (Kärkkäinen & Puglisi, SPIRE 2011)

- Partition BWT into blocks of fixed size
- Never much worse than any other partitioning

**Theorem**

Let $X_1 X_2 \ldots X_\ell$ be any partitioning of $S$.
Let $Y_1 Y_2 \ldots Y_{n/b}$ be a partitioning of $S$ into blocks of size $b$.
Then

$$\sum |Y_i| H_0(Y_i) \leq \sum |X_i| H_0(X_i) + (\ell - 1)b$$

- Achieves the same space of
  $$nH_k(T) + o(n \log \sigma)$$

- Simpler, faster and smaller in practice
## Experimental Results

<table>
<thead>
<tr>
<th>Previous Implementations (from Pizza &amp; Chili)</th>
<th>New implementations (using the same components)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HWT</td>
<td>FB+HWT</td>
</tr>
<tr>
<td>- No boosting</td>
<td>- Fixed block boosting</td>
</tr>
<tr>
<td>CB+HWT</td>
<td></td>
</tr>
<tr>
<td>- Context block boosting</td>
<td></td>
</tr>
<tr>
<td>HWT+RRR</td>
<td></td>
</tr>
<tr>
<td>- Implicit boosting</td>
<td></td>
</tr>
<tr>
<td>FB+HWT+RRR</td>
<td></td>
</tr>
<tr>
<td>- Fixed block boosting</td>
<td></td>
</tr>
</tbody>
</table>
Juha Kärkkäinen (Univ. Helsinki)
Improving FM Index
Algodan 2011 11 / 14
Juha Kärkkäinen (Univ. Helsinki)  Improving FM Index  Algodan 2011  12 / 14

DNA 100MB

Time (msec/pattern)

Memory (bits/symbol)

FB+HWT+RRR  HWT+RRR  CB+HWT  FB+HWT  HWT
Summary

- Compression boosting improves FM index compression from zeroth order to higher order
- Better compression boosting using blocks of fixed size
  - Same theoretical properties
  - Simple to implement
  - Smaller and faster in practice
Further Improvements

- Algorithm engineering
  - Better boosting
  - Engineering components
  - Supporting other queries

- Getting below $nH_k(T)$
  - Highly repetitive sequences (DNA data, version databases, ...) can be compressed to less than $nH_k(T)$ bits
  - Run-length encoding: RLFM, RLWT, RLFM+
  - Measures of compressibility?