On Performance of Compressed Pattern Matching on VF Codes

Satoshi Yoshida and Takuya Kida
Graduate School of Information Science and Technology,
Hokkaido University
Background

- **Compressed Pattern Matching**

Compressed Data

Program Searching on Compressed Data

- **Variable-to-Fixed Length (VF) Code** has been attracted some interests.

<table>
<thead>
<tr>
<th>Compressed Text</th>
<th>Input Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed</td>
<td>FF Code</td>
</tr>
<tr>
<td>Variable</td>
<td>FV Code</td>
</tr>
<tr>
<td>Fixed</td>
<td>VF Code</td>
</tr>
<tr>
<td>Variable</td>
<td>VV Code</td>
</tr>
</tbody>
</table>
Research Goal

Compression method for

- high compression ratio
- fast pattern matching
# Codeword Length in VF Codes

<table>
<thead>
<tr>
<th>short</th>
<th>codeword length</th>
<th>long</th>
</tr>
</thead>
<tbody>
<tr>
<td>small</td>
<td>size of dictionary</td>
<td>large</td>
</tr>
<tr>
<td>low</td>
<td>compression ratio</td>
<td>high</td>
</tr>
<tr>
<td>low</td>
<td>cost of construct/hold</td>
<td>high</td>
</tr>
<tr>
<td>low</td>
<td>preparation cost of pattern matching</td>
<td>high</td>
</tr>
</tbody>
</table>
Motivation

Obtaining a high compression ratio with small dictionary
Two Approaches

1. **VF Coding** (short codewords)
   - Input Text
   - Intermediate Output
   - Other Coding
   - Output
   - Range Coder [Martin1979]

2. **STVF Coding** [Kida2009]
   - Input Text
   - VF Coding (long codewords)
   - Output
Pattern Matching on STVF + Range Coded Text

After decoding compressed text with range coder, we get STVF coded text.

This method works well if this part is sufficiently fast.
Experiments

- Compression methods
  - STVF Coding + Range Coder
  - STVF Coding

- Data
  - English Text (brown corpus, 6.8MB, |Σ|=96)

- Environments
  - CPU: Intel® Xeon® processor 3.00GHz dual core
  - Memory: 12GB
  - OS: Red Hat Enterprise Linux ES Release 4

- Codeword Length
  - \( l = 8\text{-}16 \text{ bits} \)

We compared compression ratios and pattern matching times between the two methods.
Pattern Matching Time

![Graph showing pattern matching time vs pattern length for different methods: STVF(16), STVF(12), and STVF(12) + Range Coder. The graph indicates that as the pattern length increases, the pattern matching time also increases, with STVF(16) generally having the highest time, followed by STVF(12), and then STVF(12) + Range Coder.]
Conclusion

- STVF + Range Coder improves compression ratios.
- STVF + Range Coder depreciates pattern matching speed.

Future Work

- Combine with other methods whose decompression speeds are fast such as gzip.
- Implement Boyer Moore type algorithm to improve pattern matching speed when pattern length is long.
Appendix: STVF Code

Input: AAABBACB$

Output: 000 001 011 110 101