Engineering Rank and Select Queries on Wavelet Trees

Roland Larsen Pedersen

Datalogi, Aarhus Universitet

roland.l.pedersen@gmail.com

June 25, 2015

Overview

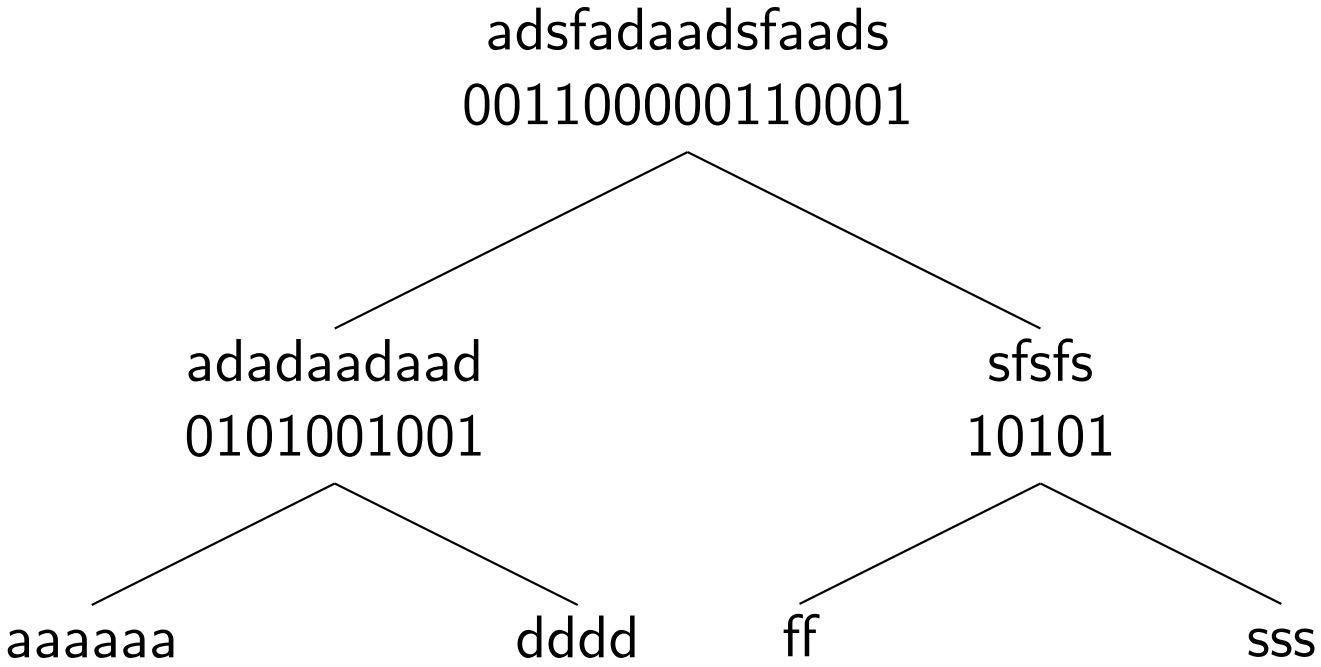
Wavelet Tree: Definitions

- In its basic form, the wavelet tree is a balanced binary tree.
- It stores a *sequence* $S[1, n] = c_1 c_2 c_3 \ldots c_n$ of *symbols* $c_i \in \Sigma$, where $\Sigma = [1 \ldots \sigma]$ is the *alphabet* of S.
- The tree has height $h = \lceil \log \sigma \rceil$, and $2\sigma - 1$ nodes, with σ of those as leaf nodes and $\sigma - 1$ as internal nodes.

Constructing the Wavelet Tree

- The wavelet tree is constructed recursively, starting at the root node and moving down the tree, with each node in the tree receiving a string constructed by its parent, except the root node that receives the full input string.
- Each node calculates the middle character of Σ and uses it to set the bits in the bitmap and split S in two substrings S_{left} and S_{right}.

Wavelet Tree Example



$S = \text{adsfadaadsfaads}, \Sigma = \text{adfs}$

Construction time and memory usage

- Construction time: $O(n \cdot h) = O(n \log \sigma)$
  - The Wavelet Tree can theoretically be constructed in $O(n \cdot h) = O(n \log \sigma)$ time as the sum of the lengths of the strings being processed at any single layer of the tree is the length of the input string to the tree.
- Memory usage: $O(n \log \sigma + \sigma \cdot ws)$ bits
  - At each level in the tree at most n bits are stored in the bitmaps in total, making $n \cdot h = n \cdot \log \sigma$ an upper bound to the total number of bits that a wavelet tree stores in its bitmaps.
  - In addition to this, each node takes some constant amount of machine words of space, and there are $2\sigma - 1$ nodes in the tree. ws is the size of our machine words. This makes the total memory consumption $O(n \log \sigma + \sigma \cdot ws)$ bits.

Queries

- The wavelet tree supports three queries:
  - **Access(p)**: Return the character c at position p in sequence S.
  - **Rank(c, p)**: Return the number of occurrences of character c in S up to position p.
  - **Select(c, o)**: Return the position of the oth occurrence of character c in S.

Blocks of Highlighted Text

Block 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

Block 2

Pellentesque sed tellus purus. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Vestibulum quis magna at risus dictum tempor eu vitae velit.

Block 3

Suspendisse tincidunt sagittis gravida. Curabitur condimentum, enim sed venenatis rutrum, ipsum neque consectetur orci, sed blandit justo nisi ac lacus.

Multiple Columns

Heading

1. Statement
2. Explanation
3. Example

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

Table

Treatments	Response 1	Response 2
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296

Table : Table caption

Theorem

Theorem (Mass–energy equivalence)

$E = mc^2$

Verbatim

Example (Theorem Slide Code)

```
\begin{frame}
\frametitle{Theorem}
\begin{theorem}[Mass--energy equivalence]
$E = mc^2$
\end{theorem}
\end{frame}
```

Figure

Uncomment the code on this slide to include your own image from the same directory as the template .TeX file.

Citation

An example of the `\cite` command to cite within the presentation:

This statement requires citation [Smith, 2012].

References

John Smith (2012)

Title of the publication

Journal Name 12(3), 45 – 678.

The End