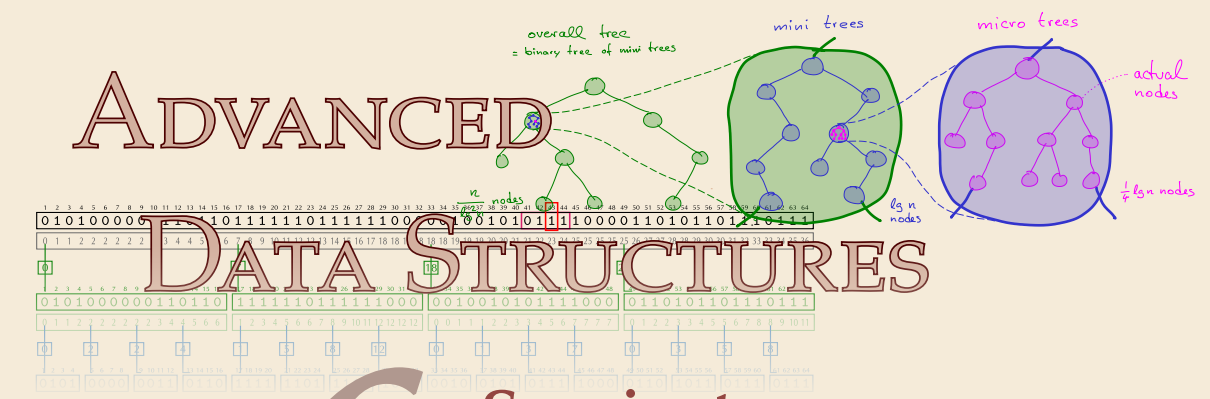


Advanced Data Structures

6 Succinct Data Structures

Advanced Data Structures · Summer 2026

Prof. Dr. Sebastian Wild

CS650 (Summer 2026)
Philipps-Universität Marburg
version 2026-06-22 16:03 H

Outline

6 Succinct Data Structures

6.1 A Motivating Example

Computing over compressed data

- traditionally:
  - compression for **archiving** data, minimize size of representation
  - computation/analysis: minimize time; use extra data structures
  - ⇝ always decompress data first!
  - reaches limits of fast memory for large datasets

- Approach in space-efficient data structures:
  - Represent data in compressed form
  - Augment with **small** index data structures to enable fast queries **directly on compressed representation**
  - *succinct* = $(1 + o(1)) \cdot$ information-theoretic lower bound

Computing over compressed data

$T[0..n) =$

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a	b	a	n	a	n	a	a	n	d	a	n	a	p	p	l	e

$C[0..m) =$

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
0	1	1	1	0	0	0	1	0	0	1	0	0	0	0	1	0	1	1	1	0	1	0	1	0	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	0

Huffman Code

- a ↦ 0
- b ↦ 11100
- d ↦ 11101
- e ↦ 11110
- l ↦ 11111
- n ↦ 10
- p ↦ 110

- Can decode from beginning ⇝ fine for storage / transmission

But how to read $T[9]$ without full decoding? How to know where its codeword starts?

- We don't. But we can store it!
  - Naive way: Store starting index for each char
    ⇝ n numbers in $[n]$ ⇝ $n \lg n$ bits.

    Much more than the (compressed) text!

entropy-compressed bitvector

Clever **succinct index** supports random access in constant time with $o(n)$ extra bits!

Dynamic Bitvectors

Recall: Dynamic rank/select lower bounds (Fredman-Saks)

Note: Can use bitvector to represent n numbers from universe $[0..u)$

- $\rightsquigarrow$ Cannot have dynamic bitvectors **and** fast rank/select queries, $o\left(\frac{\lg n}{\lg \lg n}\right)$.
- $\rightsquigarrow$ *Focus here on static bitvectors.*

6.2 Bitvectors

Bitvector Representation in C++

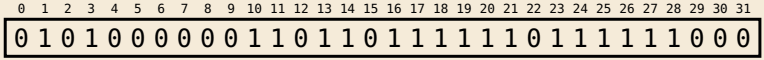
- `std::vector<char/int>`, `bool[]`
  - easy to access
  - fastest read and write access
  - ≥ 1 **byte** per bit
- `std::vector<byte>`
  - special overload
  - *may* use only 1 bit each
  - (with my GNU g++ it used same space as `vector<char>`)
- `std::vector<uint64_t>`
  - store 64 bits in one 8 byte integer
  - use bit tricks to access and write individual bits

$\rightsquigarrow$ in practice: `sdsl::bit_vector` (based on packing bits to words)

- from external library SDSL Lite
- `https://github.com/simongog/sdsl-lite`

Rank & Select on Bitvectors

- $B[0..n)$ static array of n bits (Boolean array).
  - easy to store using n bits of space
- $B.\texttt{rank}(i) = \#\,1\text{s in } B[0..i)$ (first $i \in [0..n]$ positions) (=prefix sum)
- $B.\texttt{select}_1(r) = \min\{j : B.\texttt{rank}(j+1) \geq r\} \cup \{n\}$
  $= \text{index of } r\text{th } 1 \text{ in } B, (r = 1, 2, \ldots)$



- $\texttt{rank}(12) = \texttt{rank}(13) = 4$
- $\texttt{select}_1(3) = 10$
- $\texttt{select}_1(4) = 11$

$\rightsquigarrow$ $B[i] = B.\texttt{rank}(i+1) - B.\texttt{rank}(i)$ (no need to discuss access separately)

Goal: Support rank and select on bitvector using $n + o(n)$ bits of space.

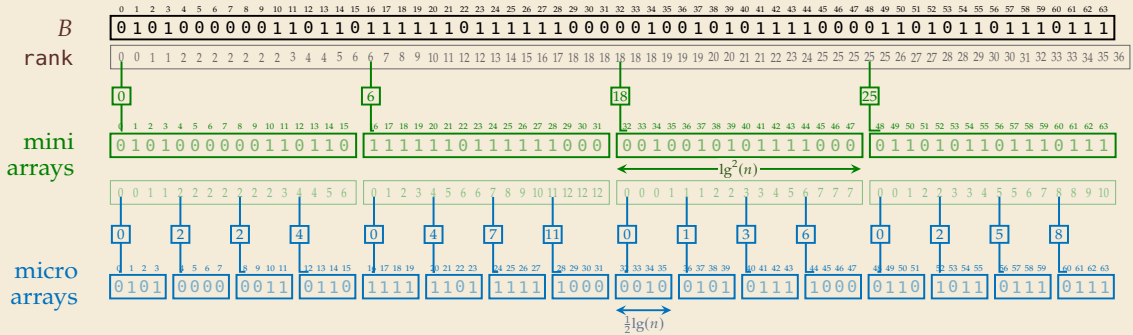
Versatile Bitvectors

- **Set of integers**
  - $S \subset [0..u) \quad \rightsquigarrow \quad B[0..u)$ with $B[x] = 1 \iff x \in S$
  - $\rightsquigarrow$ $B.\texttt{select}(B.\texttt{rank}(x+1)) \;=\; S.\texttt{predecessor}(x)$
  - $\rightsquigarrow$ efficient intersection, union, etc. via bitwise logical operations
  - used, e. g., for **"Roaring Bitmaps"**: one bitvector / integer set per attribute
- **Unary encoded integers**
  - integers $x_1, \ldots, x_k \in \mathbb{N}_{\geq 0} \quad \rightsquigarrow \quad B = 1^{x_1}0\,1^{x_2}0\,\ldots 1^{x_k}0$
  - $x_1 + \cdots + x_j \;=\; B.\texttt{rank}\big(B.\texttt{select}_{\texttt{0}}(j)\big)$
  - x_j as difference of prefix sums
- **Sparse array index**
  - suppose $A[0..n)$ is mostly 0, but with some k indices having non-zero values
  - $\rightsquigarrow$ Store in $V[0..k)$ all non-zero values compactly, in $B[0..n)$ store $B[i] = 1 \iff A[i] \neq 0$
  - $\rightsquigarrow$ Simulate $A[x] \;=\; \begin{cases} 0 & \text{if } B[x] = 0 \\ V\big[B.\texttt{rank}[x]\big] & \text{else} \end{cases}$
- **Note:** In all these cases, rather natural to expect skewed frequencies of `1`s and `0`s.

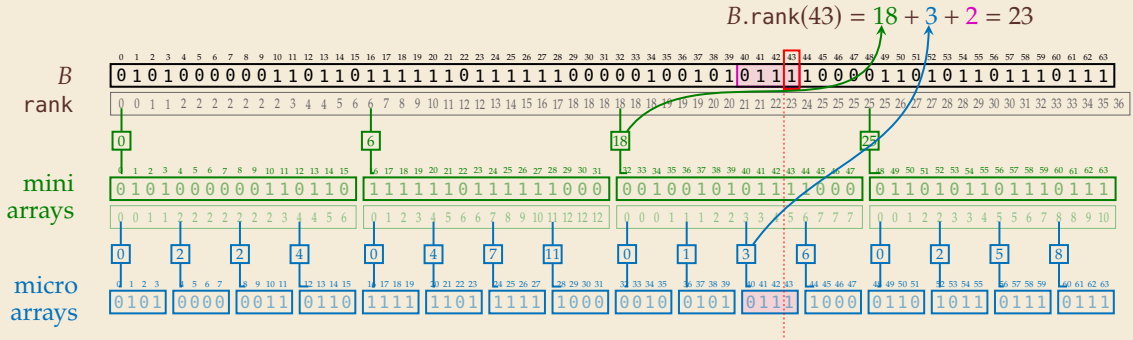
6.3 Supporting Rank

Rank Index for Bitvector

- Apart from B, we store:
  1. Rank of first pos. of each **mini array** of $L = \lg^2(n)$ bits $\rightsquigarrow$ $\frac{n}{L} \cdot \lg(n) = \frac{n}{\lg n} = o(n)$ bits
  2. Rank of first pos. in **micro array** of $\ell = \frac{1}{2}\lg(n)$ bits *relative* to its mini array
     - $\rightsquigarrow$ ranks are numbers in $[0..L]$ $\rightsquigarrow$ $\lg L = 2 \lg\lg n$ bits suffice for each
     - $\rightsquigarrow$ $\frac{n}{\ell} \lg L = \frac{n}{\frac{1}{2}\lg n} \cdot 2\lg\lg n = \frac{4n \lg\lg n}{\lg n} = o(n)$ bits



How to find rank



- How to compute $B.\texttt{rank}(i)$?
  - find rank up to element's mini array
  - add rank up to element's micro array (mini-array local)
  - add micro-array-local rank of position
    - can either do this naively by scanning $\frac{1}{2}\lg n$ bits $\rightsquigarrow$ $O(\log n)$ time
    - or use bit marks and pop-count instructions on CPUs $\rightsquigarrow$ $O(1)$ time
    - or use exhaustive *lookup table!* $\frac{1}{2}\lg n$ bits $\rightsquigarrow$ only $2^{1/2 \lg n} = \sqrt{n}$ different micro arrays

Exhaustive Tabulation

Classic word-RAM technique: bootstrap on small cases with *exhaustive tabulation*

- **Idea 1:** on word-RAM, can use integers (and hence small bit sequences) as array index
- **Idea 2:** for mirco subproblems on tiny bit seqeuences, cannot have many *different* micro subproblems

$\rightsquigarrow$ Precompute all micro subproblems in a lookup table, indexed by the subproblem's bitpattern

$\rightsquigarrow$ $O(1)$ time for micro subproblem (after preprocessing)

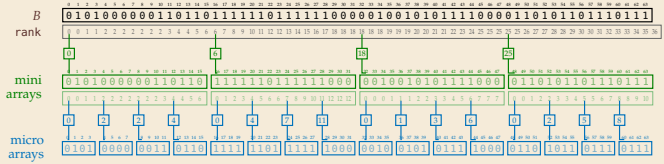
Back to Rank!

Recall: Our task is to compute the local rank in a micro array of $\ell = \frac{1}{2}\lg n$ bits.

Preprecessing: Build table for all possible queries for all *possible* micro arrays

$$\mu R[0..2^\ell)[0..\ell) \;=$$

idx	micro array	rank(0)	rank(1)	rank(2)	rank(3)	rank(4)
0	0000	0	0	0	0	0
1	0001	0	0	0	0	1
2	0010	0	0	0	1	1
3	0011	0	0	0	1	2
4	0100	0	0	1	1	1
5	0101	0	0	1	1	2
6	0110	0	0	1	2	2
7	0111	0	0	1	2	3
8	1000	0	1	1	1	1
9	1001	0	1	1	1	2
10	1010	0	1	1	2	2
11	1011	0	1	1	2	3
12	1100	0	1	2	2	2
13	1101	0	1	2	2	3
14	1110	0	1	3	3	3
15	1111	0	1	2	3	4



- Note: $\#\text{micro arrays} = \frac{n}{\frac{1}{2}\lg n} \gg$ $\#\textit{different}$ micro arrays $= 2^{1/2\lg n} = \sqrt{n}$
- To access micro-array rank(i)
  1. Read micro array *type* (bits), here from B
  2. Treat *type* as binary number $\rightsquigarrow$ idx t
  3. Return $\mu R[t][i]$

Space: $2^\ell \cdot \ell \cdot \lg \ell \;=\; \sqrt{n} \cdot \text{polylog}(n) \;=\; o(n)$

Four Russians?

The *exhaustive-tabulation technique* is often called "Four Russians trick" in the literature . . .

- The algorithmic technique was published 1970 by
  V. L. Arlazarov, E. A. Dinitz, M. A. Kronrod, and I. A. Faradžev

  Arlazarov, Dinitz, Kronrod, Faradžev: *On economical construction of the transitive closure of a directed graph*, Dokl. Akad. Nauk SSSR **1970**

  inofficial translation: `https://minorfree.github.io/FourRussian/`

- all worked in Moscow at that time . . . but Soviet ≠ Russian; Dinitz emigrated to Isreal

  (Arlazarov and Kronrod are Russian)

- American authors coined the othering term "Method of Four Russians"
  . . . name in widespread use

The Original Application: Boolean Matrix Multiplication

Suppose we want to multiply two $n \times n$ Boolean matrices $C = A \cdot B$.

We divide A, B, and C into $\ell \times \ell$ *micro matrices*.

$\rightsquigarrow$ C consists of $\left(\frac{n}{\ell}\right)^2$ micro matrices, each of which is the sum of $\frac{n}{\ell}$ micro-matrix products.

The number of *different* possible micro matrix products is $L = 2^{\ell^2} \cdot 2^{\ell^2}$.
If we pick $\ell = \frac{1}{4}\sqrt{\lg n}$, we have only $L = 2^{2\ell^2} = \sqrt{n}$ different products.

$\rightsquigarrow$ ***Exhaustive Tabulation:*** Can *precompute* all $\sqrt{n}$ *possible* micro-matrix sums/products!

For two micro matrices a and b, we store $a \cdot b$ at the offset $a_{1,1} \ldots a_{\ell,\ell} b_{1,1} \ldots b_{\ell,\ell}$, where we interpret this bitstring as a binary number.
On a word RAM, we can use this as indirect memory access in $O(1)$ time.

$\rightsquigarrow$ Any micro matrix sum/product takes $O(1)$ time
after a total of $O(\sqrt{n} \cdot \log^{3/2} n)$ preprocessing.

The total time to compute one micro matrix in C is thus $O(\frac{n}{\ell})$.
So the total time to compute C is $O(n^3/\ell^3) = O(n^3/\log^{3/2} n)$.

Note: By taking $n \times \ell$ resp. $\ell \times n$ *"micro strips"* instead of squares, we can choose $\ell = \Theta(\log n)$ and obtain final time $O(n^3/\log^2 n)$.

6.4 Supporting Select

6.5 Compressed Bitvectors

Applications

- variable-cell arrays
- monotone perfect hashing
- prefix sums

6.6 Sequences

Rank & Select on Sequences

Recall: Decoding only needs access to

1. ith char c of $\text{sort}(T) = \text{sort}(B)$
2. *position* of (that copy of) c in B

Both can be supported using rank/select on sequences.

#occurrences of c

- $\text{rank}_c(T[0..n), i) = \left|T[0..i)\right|_c$
  $= \#c$ in first i characters of T
- $\text{select}_c(T[0..n), r)$
  $= \min\{j : |T[0..j]|_c \geq r\} \cup \{n\}$
  $= $ index of rth c in T, $(r = 1, 2, \ldots)$

	0	1	2	3	4	5	6	7	8
$T[0..9)$	b	a	n	a	n	a	b	a	n

	0	1	2	3	4	5	6	7	8	9
$\text{rank}_a(T, i)$	0	0	1	1	2	2	3	3	4	4
$\text{rank}_b(T, i)$	0	1	1	1	1	1	1	2	2	2
$\text{rank}_n(T, i)$	0	0	0	1	1	2	2	2	2	3

$\text{select}_a(T, r)$	/	1	3	5	7	9	9	9	9	9
$\text{select}_b(T, r)$	/	0	6	9	9	9	9	9	9	9
$\text{select}_n(T, r)$	/	2	4	8	9	9	9	9	9	9

	0	1	2	3	4	5	6	7	8
$\text{sort}(T)$	a	a	a	a	b	b	n	n	n

$O[0..\sigma] = [0, 4, 6, 9]$

Random Access in BWT

- store offsets $O[c] = \sum_{c'=0}^{c-1} |B|_{c'}$ for $c \in \Sigma$
- ith char of $\text{sort}(B)$ = unique c for which $O[c] \leq i < O[c+1]$
- position of rth c in B = $\text{select}_c(B, r)$

Wavelet Trees

The **Wavelet Trees** for a $T \in [0..n)$ over $\Sigma = [0..\sigma)$

- supports access to $T[i]$ in $O(\log \sigma)$ time,
- $\text{rank}_c(T, i)$ and $\text{select}_c(T, r)$ in $O(\log \sigma)$ time, and
- occupies $\sim n \lg \sigma$ *bits* of space. *(Further compression possible!)* $\rightsquigarrow$ *Advanced Data Structures*
- The generalized $\sigma\text{-rank}_c(T, i) = \text{rank}_c(T, i) + \sum_{c' < c} |T|_{c'}$ is also supported in $O(\log \sigma)$ time

Storing $B[0..n]$ as a wavelet tree $\rightsquigarrow$ reconstruct ℓ chars from T in $O(\ell \log \sigma)$ time
if starting position known

e. g., $t = \lg n$

Storing **every tth entry** of $R[0..n]$ $\rightsquigarrow$ may need to go back t characters for access
$\rightsquigarrow$ $O((\ell + t) \log \sigma)$ time for decode
using $\sim n \lg n/t$ extra bits of space

Locally decodable BWT

- no longer need to store $T[0..n)$!
- compressible (e. g., Wavelet trees with compressed bitvectors)

6.7 Trees

6.8 Graphs