x3: Lossless Data Compressor

David Barina Ondrej Klima

Brno University of Technology

2022

Compression algorithm overview

1: $p \leftarrow 0$ ▷ The p is a position in input stream.
2: **while** $p \neq \text{EOF}$ **do**
3: $l_d \leftarrow \text{QueryDictionary}(p)$ ▷ Length of the fragment.
4: $l_w \leftarrow \text{SearchInWindow}(p)$ ▷ Length of the fragment.
5: **if** $l_d > l_w$ **then**
6: $\text{EncodeDictionaryIndex}(p, l_d)$
7: $p \leftarrow p + l_d$
8: **else**
9: $\text{EncodeRawFragment}(p, l_w)$
10: $\text{AddFragmentToDictionary}(p, l_w)$
11: $p \leftarrow p + l_w$
12: **end if**
13: **end while**

SearchInWindow function

Require: M : maximum number of matches; L : maximum match length
Ensure: returns length of the best match

```
1:  function SearchInWindow(p)
2:      for l ← 1 ... L do
3:          c_l ← CountOccurrences(p, l)
4:      end for
5:      for m ← M ... 1 do
6:          for l ← L ... 1 do
7:              if c_l > m then
8:                  return l
9:              end if
10:         end for
11:     end for
12: end function
```

Compression ratio on Silesia corpus. Best results in bold

File	gzip	xz	zstd	Brotli	x3
dickens	2.6461	3.6000	3.5765	3.6044	**3.7168**
mozilla	2.6966	**3.8292**	3.3769	3.6922	2.7432
mr	2.7138	3.6231	3.2132	3.5317	**4.0364**
nci	11.2311	**23.1519**	20.7925	22.0780	19.1103
ooffice	1.9907	**2.5346**	2.3587	2.4818	2.0668
osdb	2.7138	3.5456	3.2855	3.5812	**3.6151**
reymont	3.6396	5.0374	4.9060	4.9747	**5.1010**
samba	3.9950	**5.7778**	5.5267	5.7367	4.1871
sao	1.3613	**1.6386**	1.4479	1.5812	1.5042
webster	3.4372	4.9540	4.8970	4.9188	**4.9685**
xml	8.0709	12.2910	11.8004	**12.4145**	9.2249
x-ray	1.4035	1.8868	1.6457	1.8096	**1.9649**

Memory consumption on Silesia corpus

Size and MaxRSS are given in megabytes. Factor is the ratio MaxRSS / Size.

File	Size	MaxRSS	Factor
dickens	9.8	42.3	4.4
mozilla	49.0	697.2	14.3
mr	9.6	53.1	5.6
nci	32.0	53.3	1.7
ooffice	5.9	105.8	18.0
osdb	9.7	51.2	5.3
reymont	6.4	27.9	4.4
samba	21.0	163.6	7.9
sao	7.0	100.7	14.6
webster	40.0	177.1	4.5
xml	5.1	22.1	4.3
x-ray	8.1	59.0	7.3

Impact of window size

Window size is given in kilobytes. Matches indicate the optimal number of matches for the given window size.

Window size	Matches	Compression ratio
1	7	3.5359
2	9	3.5548
4	15	3.5684
8	28	3.5799
16	38	3.5963
32	73	3.6117
64	136	3.6358