DCC IEEE

Presented at the 2014 IEEE
Data Compression Conference
Snowbird, Utah, USA — March 26–28

# Improving Compression via Substring Enumeration by Explicit Phase Awareness

Mathieu Béliveau
mathieu.beliveau.2@ulaval.ca

Danny Dubé
Danny.Dube@ift.ulaval.ca

Université Laval
Canada

This work was funded by the Natural Science and Engineering Research Council of Canada.

NSERC CRSNG

## Substring enumeration [6]

| Length | Substrings | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0 | 8×ϵ | | | | | | | |
| 1 | 6×0 | | | | | | 2×1 | |
| 2 | 4×00 | | | | 2×01 | | 2×10 | |
| 3 | 3×000 | | | 1×001 | 2×010 | | 1×100 | 1×101 |
| 4 | 2×0000 | | 1×0001 | 1×0010 | 1×0100 | 1×0101 | 1×1000 | 1×1010 |
| 5 | 1×00000 | 1×00001 | 1×00010 | 1×00101 | 1×01000 | 1×01010 | 1×10000 | 1×10100 |
| 6 | 1×000001 | 1×000010 | 1×000101 | 1×001010 | 1×010000 | 1×010100 | 1×100000 | 1×101000 |
| 7 | 1×0000010 | 1×0000101 | 1×0001010 | 1×0010100 | 1×0100000 | 1×0101000 | 1×1000001 | 1×1010000 |
| 8 | 1×00000101 | 1×00001010 | 1×00010100 | 1×00101000 | 1×01000001 | 1×01010000 | 1×10000010 | 1×10100000 |

Naïve substring enumeration for '01000001'.

## Experimental results (in bpc)

| File [12] | Gzip | BWT | PPM | CSE | +SC | +EPA |
|---|---|---|---|---|---|---|
| bib | 2.51 | 2.07 | 1.91 | 1.98 | 1.88 | **1.87** |
| book1 | 3.25 | 2.49 | 2.40 | 2.27 | 2.33 | **2.24** |
| book2 | 2.70 | 2.13 | 2.02 | 1.98 | **1.93** | **1.93** |
| geo | 5.34 | **4.45** | 4.83 | 5.35 | 4.57 | 4.56 |
| news | 3.06 | 2.59 | **2.42** | 2.52 | **2.42** | **2.42** |
| obj1 | **3.84** | 3.98 | 4.00 | 4.46 | 3.99 | 3.95 |
| obj2 | 2.63 | 2.64 | **2.43** | 2.71 | 2.44 | 2.44 |
| paper1 | 2.79 | 2.55 | **2.37** | 2.54 | 2.41 | 2.39 |
| paper2 | 2.89 | 2.51 | 2.36 | 2.41 | 2.34 | **2.33** |

| File | Gzip | BWT | PPM | CSE | +SC | +EPA |
|---|---|---|---|---|---|---|
| paper3 | 3.11 | — | — | 2.73 | 2.63 | **2.61** |
| paper4 | 3.33 | — | — | 3.20 | 3.01 | **2.96** |
| paper5 | 3.34 | — | — | 3.33 | 3.10 | **3.05** |
| paper6 | 2.77 | — | — | 2.65 | 2.49 | **2.47** |
| pic | 0.82 | 0.83 | 0.85 | **0.77** | 0.81 | 0.81 |
| progc | 2.68 | 2.58 | **2.40** | 2.60 | 2.44 | 2.42 |
| progl | 1.80 | 1.80 | 1.67 | 1.71 | 1.64 | **1.63** |
| progp | 1.81 | 1.79 | **1.62** | 1.78 | 1.66 | 1.64 |
| trans | 1.61 | 1.57 | **1.45** | 1.60 | 1.47 | **1.45** |

## Implicit phase awareness in CSE+SC [4, 5, 11]

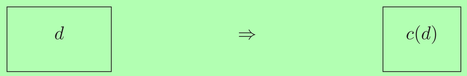
...!W*... ⟼

...00100001011101010101101110010100100111...

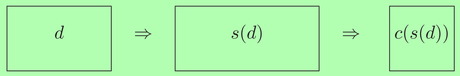
CSE+SC is still unaware of the phase of the bits
but $D$ contains a synchronization code.
CSE+SC's predictions on sufficiently long strings (e.g. 13 bits in the illustration)
do not mix substrings of different phases together.

## Use of synchronization codes

Instead of:




a pre-processing step inserts a synchronization code:




## Synchronization schemes

Per-byte mappings only; padding bytes with 9 bit strings:

$$M(b_1\, b_2 \ldots b_8) = w_1\, b_1\, w_2\, b_2 \ldots w_8\, b_8\, w_9.$$

A *c-bit scheme* inserts $c$ bits per byte, where $c = |w_1 \ldots w_9|$.

A $c$-bit scheme is *r-reliable* if it is possible to determine the phase (a number among $0, \ldots, c+7$) of any substring of the synchronized data that is at least $r$ bits long.

| | $r$ | $c$ | Synchronization scheme |
|---|---|---|---|
| ISITA'10 [4] | — | 0 | _ _ _ _ _ _ _ _ |
| | — | 1 | _ _ _ _ _ _ _ _ 0 |
| | — | 2 | _ _ _ _ _ _ _ _ 0 1 |
| | — | 3 | _ _ _ _ _ _ _ _ 0 1 1 |
| | — | 4 | _ _ _ _ _ _ _ _ 0 1 1 1 |
| | 13 | 5 | _ _ _ _ _ _ 0 _ _ 0 1 1 1 |
| DCC'11 [5] | 12 | 8 | _ _ _ 0 _ _ 1 0 0 _ _ 1 _ 1 1 0 |
| | 11 | 8 | _ _ _ 0 _ 0 _ _ 1 1 0 _ _ 1 1 0 |
| | 10 | 10 | _ _ _ _ 0 _ 0 1 1 _ _ _ 0 1 0 0 1 1 |
| | 9 | 10 | _ _ _ _ 0 0 0 1 1 _ _ _ _ 0 1 0 1 1 |
| | 8 | 15 | _ _ _ 0 0 0 1 0 _ _ _ 1 1 1 0 1 _ _ 1 1 0 0 1 |
| | 7 | 20 | 1 1 0 1 _ 1 _ 1 1 0 0 _ 0 _ 0 1 0 0 _ 0 _ 1 1 0 0 _ 1 _ |

## Phase unawareness in CSE

...!W*... ⟼

...00100001010101110010101 0...

Benchmark files are made of bytes; each byte is mapped to 8 bits.
CSE is unaware of the original bytes and the phase of the bits.
Bits on different phases are likely to have different statistics.
CSE's predictions on mixed-phase substrings are likely to be suboptimal.

## Occurrences and numbers of occurrences in CSE

The data to compress, denoted by $d$, is drawn from $\{0, 1\}$ and has length $N$. CSE works on a *circular* version of $d$, which is denoted by $D$.

A substring $w$ *occurs at position* $p$ in $D$, denoted by $w \in_p D$, if:

$$\exists u, v \in \{0, 1\}^* \,.\, \exists i \in \mathbb{N} \,.\, u\, w\, v = d^i \text{ and } 0 \leq |u| = p < N.$$

A substring $w$ *occurs* in $D$, denoted by $w \in D$, if:

$$\exists p \in \mathbb{N} \,.\, w \in_p D.$$

The *number of occurrences* of a substring $w$ in $D$, denoted by $C_w$, is:

$$|\{p \in \mathbb{N} \mid w \in_p D\}|.$$

The following equations hold:

$$C_{0w} + C_{1w} \;=\; C_w \;=\; C_{w0} + C_{w1}, \qquad \text{for any } w \in \{0, 1\}^*.$$

The predictions on the numbers of occurrences are guided by the bounds:

$$\max(0, C_{0w} - C_{w1}) \;\leq\; C_{0w0} \;\leq\; \min(C_{w0}, C_{0w}), \qquad \text{for any } w \in \{0, 1\}^*.$$

## Pseudo-code for CSE

**Send** $N$
**Send** $C_0$
**For** $l$ **:=** 2 **to** $N$ **do**
  **For every** $w \in D$ **such that** $|w| = l - 2$ **do**
    **Predict and send** $C_{0w0}$

## Explicit phase awareness in CSE+EPA [contribution]

...!W*... ⟼

...00100001010101110010101 0...

CSE+EPA is explicitly aware of the phase of the bits
relative to the boundaries of the original bytes.
CSE+EPA's predictions on substrings
do not mix substrings of different phases together.

## Occurrences and numbers of occurrences in CSE+EPA

Again, $d$ is drawn from $\{0, 1\}$ and has length $N$. $D$ is the circular version of $d$. We suppose $d$ is made of $k$-bit blocks (e.g. $k = 8$ for bytes).

A substring $w$ *occurs at phase* $q$ *and position* $p$ in $D$, denoted by $w \in_p^q D$, if:

$$\exists u, v \in \{0, 1\}^* \,.\, \exists i \in \mathbb{N} \,.\, u\, w\, v = d^i, 0 \leq |u| = p < N, \text{ and } p \bmod k = q.$$

A substring $w$ *occurs at phase* $q$ in $D$, denoted by $w \in^q D$, if:

$$\exists p \in \mathbb{N} \,.\, w \in_p^q D.$$

The *number of occurrences* of a substring $w$ in $D$ at phase $q$, denoted by $C_w^q$, is:

$$|\{p \in \mathbb{N} \mid w \in_p^q D\}|.$$

The following equations hold:

$$C_{0w}^q + C_{1w}^q \;=\; C_w^q \;=\; C_{w0}^{q \oplus 1} + C_{w1}^{q \oplus 1}, \qquad \text{for any } w \in \{0, 1\}^* \text{ and } 0 \leq q < k.$$

The predictions on the numbers of occurrences are guided by the bounds:

$$\max(0, C_{0w}^q - C_{w1}^{q \oplus 1}) \;\leq\; C_{0w0}^q \;\leq\; \min(C_{w0}^{q \oplus 1}, C_{0w}^q), \qquad \text{for any } w \in \{0, 1\}^* \text{ and } 0 \leq q < k.$$

## Pseudo-code for CSE+EPA

**Send** $N$
**For** $q$ **:=** 0 **to** $k - 1$ **do send** $C_0^q$
**For** $l$ **:=** 2 **to** $N$ **and** $q$ **:=** 0 **to** $k - 1$ **do**
  **For every** $w \in^q D$ **such that** $|w| = l - 2$ **do**
    **Predict and send** $C_{0w0}^q$

## Conclusions

- Our contribution is CSE+EPA.
- In principle, CSE+EPA is ideal on byte-oriented data.
- Still, we get a negative result, in that CSE+SC does almost as well as CSE+EPA.
- Related work: Gzip [8] as a variant of LZ77 [14], the Burrows-Wheeler transform [1], prediction by partial matching [2], antidictionaries [3], LZ78 [15]. and other results on CSE [7, 9, 10, 13].

## References

[1] M. Burrows and D. Wheeler. A block sorting lossless data compression algorithm. Technical Report 124, Digital Equipment Corporation, 1994.

[2] John G. Cleary and William J. Teahan. Unbounded length contexts for PPM. *The Computer Journal*, 40(2/3):67–75, 1997.

[3] M. Crochemore and G. Navarro. Improved antidictionary based compression. In *Proceedings of the International Conference of the Chilean Computer Science Society*, pages 7–13, 2002.

[4] Danny Dubé. Using synchronization bits to boost compression by substring enumeration. In *Proceedings of the International Symposium on Information Theory and its Applications*, pages 82–87, Taichung, Taiwan, October 2010.

[5] Danny Dubé. On the use of stronger synchronization to boost compression by substring enumeration. In *Proceedings of the Data Compression Conference*, page 454, Snowbird, Utah, USA, March 2011.

[6] Danny Dubé and Vincent Beaudoin. Lossless data compression via substring enumeration. In *Proceedings of the Data Compression Conference*, pages 229–238, Snowbird, Utah, USA, March 2010.

[7] Danny Dubé and Hidetoshi Yokoo. The universality and linearity of compression by substring enumeration. In *Proceedings of the International Symposium on Information Theory*, pages 1519–1523, Saint-Petersburg, Russia, July 2011.

[8] Jean-Loup Gailly and Mark Adler. The GZIP compressor. http://www.gzip.org.

[9] Ken-ichi Iwata, Mitsuharu Arimura, and Yuki Shima. An improvement in lossless data compression via substring enumeration. In *Proceedings of the IEEE/ACIS International Conference on Computer and Information Science*, pages 219–223, Sanya, Hainan Island, China, May 2011.

[10] Ken-ichi Iwata, Mitsuharu Arimura, and Yuki Shima. On the maximum redundancy of CSE for i.i.d. sources. In *Proceedings of the International Symposium on Information Theory and Applications*, pages 489–492, Honolulu, Hawaii, USA, October 2012.

[11] Dany Vohl, Claude-Guy Quimper, and Danny Dubé. Finding synchronization codes to boost compression by substring enumeration. In *Proceedings of the International Workshop on Constraint Modelling and Reformulation*, Quebec City, Quebec, Canada, October 2012.

[12] Ian Witten, Timothy Bell, and John Cleary. The Calgary corpus, 1987. ftp://ftp.cpsc.ucalgary.ca/pub/projects/text.compression.corpus.

[13] Hidetoshi Yokoo. Asymptotic optimal lossless compression via the CSE technique. In *Proceedings of the International Conference on Data Compression, Communications and Processing*, pages 11–18, Palinuro, Italy, June 2011.

[14] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–342, 1977.

[15] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, September 1978.