Improving Compression via Substring Enumeration by Explicit Phase Awareness

Mathieu Béliveau
Université Laval, Canada
mathieu.beliveau.2@ulaval.ca

Danny Dubé
Université Laval, Canada
danny.dube@ift.ulaval.ca

Compression by Substring Enumeration (CSE) is a recent lossless compression scheme that favorably compares to other lossless compression techniques. Experiments showed that it tends to incur a performance loss on non-textual, byte-oriented sources and it was conjectured that CSE's *phase unawareness* was responsible for this loss of performance. Subsequent work [1] confirmed the conjecture by obtaining improved compression ratios when synchronization codes get inserted in the data source, indirectly giving to CSE some kind of phase awareness, i.e. some perception of the position of the bits in the bytes. This solution does not provide a direct measure of the loss really incurred by phase unawareness. In this work, we present how CSE has been modified to be made explicitly phase aware. The original version of CSE describes its data source D by constructing a tree (called the CST) level by level. In the CST, the successive arc labels on a path from the root to any node n_w form a substring w of D. The node label associated to n_w is C_w, the number of occurrences of w in D. Thanks to the inequality $\max(0, C_{0w} - C_{w1}) \leq C_{0w0} \leq \min(C_{w0}, C_{0w})$, CSE is able to establish an upper and lower bounds on C_{0w0} based on counts for shorter substrings, which helps CSE to efficiently predict the exact value of C_{0w0}. This process continues until the CST is completely built, causing D to get fully described. On the other hand, the explicitly phase-aware version of CSE builds an 8-rooted CST, where each of the 8 subtrees describes the substrings that start at a specific phase in a byte. The count C_w^q is the number of occurrences of w that start at phase q. A more precise inequality is used: $\max(0, C_{0w}^q - C_{w1}^{q\oplus 1}) \leq C_{0w0}^q \leq \min(C_{w0}^{q\oplus 1}, C_{0w}^q)$. The tables show the compression ratios, in bits per character, for CSE, CSE with Synchronization Codes (+SC), and CSE with Explicit Phase Awareness (+EPA). To our surprise, the results show the near optimality of +SC as a measure against phase unawareness.

File	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**

[1] D. Dubé, "On the use of stronger synchronization to boost compression by substring enumeration," in *Proceedings of the Data Compression Conference*, Snowbird, Utah, USA, March 2011, p. 454.