Empirical Analysis of Burrows-Wheeler Algorithm

Yuanjie Jin

Burrows-Wheeler transform (BWT) is used to transform the original file so that it can be compressed more using a regular Huffman encoder. The basic idea is that same suffixes often follow same prefixes, which leads to grouping of same prefixes after sorting and gives rise to higher frequency of repeated letters in transformed chunks of data. To test the performance of compression by BWT followed by Huffman encoding compared to Huffman encoding only, I ran HuffMark using both text files and non-text files under the directories *calgary* and *waterloo,* respectively. The results are shown in Table 1 and Table 2, including the time to compress and the amount of compression. It is obvious that for most of files compressed, BWT generates a significant improvement on the compression ratio compared to Huffman only. This difference is especially dramatic for binary files in *waterloo,* which are not much compressed using solely Huffman. On the other hand, BWT seems to take more time than Huffman only, especially for compressing large binary files, which could be due to more operations in the algorithm. This suggests that BWT actually sacrifices time for memory.

Table 1

binary files in Waterloo	from (bytes)	Huffman only			BWT + Huffman		
		to (bytes)	time (sec)	compression%	to (bytes)	time (sec)	compression%
clegg.tif	2149096	2034591	14.997	5.33	967115	37.299	55.00
frymire.tif	3706306	2188589	15.04	40.95	720455	55.676	80.56
lena.tif	786568	766142	5.162	2.60	665511	9.323	15.39
monarch.tif	1179784	1109969	7.681	5.92	862550	12.226	26.89
peppers.tif	786568	756964	5.221	3.76	652441	9.102	17.05
sail.tif	1179784	1085497	7.619	7.99	931162	13.055	21.07
serrano.tif	1498414	1127641	7.847	24.74	286581	9.309	80.87
tulips.tif	1179784	1135857	8.699	3.72	998964	14.852	15.33
total bytes read		12466304			12466304		
total compressed bytes		10205250			6084779		
total percent compression		18.137			51.19		
compression time		72.266			160.842		

Table 2

text files in Calgary	from (bytes)	Huffman only			BWT + Huffman		
		to (bytes)	time (sec)	compression%	to (bytes)	time (sec)	compression%
bib	111261	73791	1.016	33.68	42467	0.762	61.83
book1	768771	439405	5.817	42.84	343299	5.174	55.34
book2	610856	369331	4.745	39.54	240476	3.692	60.63
geo	102400	73588	0.958	28.14	74752	1.016	27.00
news	377109	247424	3.278	34.39	169595	2.541	55.03
obj1	21504	17081	0.244	20.57	12475	0.237	41.99
obj2	246814	195127	2.558	20.94	99893	1.651	59.53
paper1	53161	34367	0.441	35.35	21228	0.318	60.07
paper2	82199	48645	0.623	40.82	33492	0.513	59.25
paper3	46526	28305	0.401	39.16	20034	0.306	56.94
paper4	13286	8890	0.119	33.09	6364	0.1	52.10
paper5	11954	8461	0.105	29.22	5931	0.092	50.38
paper6	38105	25053	0.311	34.25	15564	0.24	59.15
pic	513216	107582	1.484	79.04	116451	18.508	77.31
progc	39611	26944	0.345	31.98	16025	0.235	59.54
progl	71646	44013	0.575	38.57	22430	0.411	68.69
progp	49379	31244	0.415	36.73	15300	0.298	69.02
trans	93695	66248	0.956	29.29	28960	0.504	69.09
total bytes read		3251493			3251493		
total compressed bytes		1845499			1284736		
total percent compression		43.241			60.488		
compression time		25.027			38.758		