

Modelling Chinese For Text Compression

Peiliang Wu and W.J.Teahan

Computer Science Division, School of Informatics,
University of Wales Bangor, Dean Street,
Bangor, LL57 1UT, U.K.

email: {*perry, wjt*}@informatics.bangor.ac.uk

We have adapted the PPM model especially for Chinese text and achieve good compression results. We highlighted the importance of pre-processing work for Chinese, as unlike naturally segmented language such as English, it is not clear what are the most appropriate symbols to use for encoding. We have developed a text compression corpus for Chinese text, and this can be found at <http://www.informatics.bangor.ac.uk/~wjt/AIIA/BMCC.tar.gz>. Our experiments with the corpus show that the pre-processing work can improve the compression rate significantly. Compression results are shown in Table 1. The full paper is available at <http://www.informatics.bangor.ac.uk/~wjt/AIIA/preprint/2005/index.html>.

We made several changes in the PPM model to adapt specifically to the Chinese language. Changing the symbol encoding unit to 16 bits captures the structure of the language precisely. Sorting all the characters in the context by frequency order improves the program speed significantly and using no exclusions also leads to faster execution speed. This new PPM-Ch model should also achieve similar improvements in other large alphabet size languages such as Japanese, Korean and Thai.

file	size (bytes)	gzip (bpb)	bzip2 (bpb)	WinRAR (bpb)	ABC (bpb)	PPMC (bpb)	PPMD (bpb)	PPMII (bpb)	PPMD-Ch (bpb)
Chinese news1	15,371	4.572	4.496	4.549	4.428	4.213	4.217	4.119	3.955
Chinese news2	26,195	4.801	4.692	4.773	4.563	4.468	4.478	4.324	4.168
Chinese article1	19,245	4.883	4.787	4.863	4.657	4.529	4.550	4.340	4.248
Chinese article2	38,724	4.705	4.401	4.660	4.258	4.229	4.228	4.083	4.021
Chinese article3	80,760	4.236	3.886	4.158	3.772	3.716	3.706	3.571	3.514
Chinese book1	53,706	4.466	4.203	4.395	4.083	4.032	4.036	3.880	3.828
Chinese book2	436,656	4.670	4.013	4.385	3.830	3.872	3.857	3.689	3.737
LCMC-A	271,053	4.763	4.353	4.518	4.223	4.197	4.193	4.015	3.963
LCMC-B	170,469	4.413	4.085	4.221	3.965	3.907	3.909	3.756	3.710
LCMC-C	112,681	4.102	3.765	3.927	3.690	3.590	3.583	3.455	3.404
LCMC-D	103,031	4.478	4.238	4.336	4.113	4.056	4.055	3.900	3.812
LCMC-E	218,473	4.501	4.282	4.323	4.143	4.123	4.124	3.946	3.889
LCMC-F	266,266	4.655	4.316	4.461	4.186	4.166	4.163	3.983	3.939
LCMC-G	452,511	4.813	4.326	4.524	4.158	4.161	4.156	3.981	3.948
LCMC-H	208,035	3.508	3.119	3.229	3.045	2.998	2.988	2.875	2.878
LCMC-J	509,095	4.110	3.676	3.847	3.552	3.529	3.517	3.377	3.401
LCMC-K	159,118	4.828	4.502	4.666	4.316	4.346	4.344	4.148	4.047
LCMC-L	137,743	4.765	4.458	4.618	4.313	4.310	4.307	4.119	4.003
LCMC-M	35,159	4.736	4.575	4.663	4.473	4.378	4.393	4.232	4.043
LCMC-N	155,897	4.741	4.435	4.569	4.265	4.280	4.282	4.088	4.008
LCMC-P	273,109	4.730	3.128	2.715	3.019	3.197	3.212	2.946	2.928
LCMC-R	49,289	4.627	4.411	4.563	4.273	4.242	4.245	4.081	3.924
Average		4.550	4.189	4.317	4.060	4.025	4.025	3.859	3.789

Table 1: Chinese text compression for Small Bangor Mandarin Chinese Corpus using different compressors