

Experimental Study of Some Cleary-Witten Algorithm Modifications

Yu.M.Shtarkov[†], D.N.Zotkin^{††}

[†]Institute for Problems of Information Transmission
of Russian Academy of Sciences, 101447, Moscow, GSP-4, Russia
^{††}Moscow Institute of Physics and Technology
Institutskii per, 9, 141700 Dolgoprudnyi, Russia

1 Introduction

J.G.Cleary and I.H.Witten proposed the well known now, promising approach to data compression problem [1, 2]. Some features of their (CW) algorithm (especially the choice of conditional probabilities) seem to be chosen rather heuristically and thus we hope to improve the results (as it was made in [2] comparing with [1]). The simple (and heuristical too) way is introducing some free parameters in expressions for conditional probabilities and optimizing its values (see, for example, [3]). But the achieved improvements were not essential.

Introducing statistical model of data allows us to order such trials. CW-algorithm was proposed for arbitrary finite-state Markov (FSMX) sources. Nevertheless it does not correspond to general FSMX-model and has no Markov property sometimes. Then the model of context-tree finite-state sources with proportional conditional probabilities (PCP-model) was developed [4]. It helps us to consider some modifications of original CW-algorithm. We paid attention to efficiency of algorithms and did not consider its complexity. Some experimental results are described here.

2 Description of experiments

A compression algorithm substitutes every message (sequence $x^n = x_1, \dots, x_n$ of n letters of M -ary alphabet) with codeword, containing $l(x^n)$ symbols of M^* -ary alphabet (we shall suppose that $M = M^*$). The efficiency of compression of given message x^n is defined with coding rate

$$R(x^n) = l(x^n)/n. \quad (1)$$

And the sequential arithmetic coding (see, for example, [5]) allows us to reduce a problem of constructing of noiseless (distortionless) uniquely decodable code to a choice of conditional probabilities $\{\vartheta(a|x^k), a \in A\}$, $x^k \in A^k$, $k = 0, 1, \dots$

The values of $\vartheta(a|x^k)$ in CW-algorithm are equal to

$$\vartheta(a|x^k) = \vartheta_i(a|x^k) \prod_{j=i+1}^D P_j(x_D^k), \quad -1 \leq i \leq D, \quad (2)$$

iff letter a satisfies inequalities

$$t(x_i^k a|x^k) \geq Q, \quad t(x_{i+1}^k a|x^k) < Q \quad (3)$$

(if $t(x_{D+1}^k a|x^k) \geq Q$ then $i = D$ and if $t(a|x^k) < Q$ then $i = -1$), where D is a maximum depth of memory, x_D^k are the last D letters of x^k , $t(x_i^k a|x^k)$ is a number of occurrences of sequence of $i + 1$ letters $x_i^k a$ in x^k and $Q > 0$ is an integer threshold.

The choice of $\vartheta_i(a|x^k)$ and "escape" probabilities $P_j(x_D^k)$ are restricted by equality $P_{-1}(x_D^k) \equiv 0$ and by natural normalizing conditions only. Following [4] we considered

$$\vartheta_i(a|x_D^k) = \frac{\tau(x_i^k a) + \alpha_i}{N(x_i^k)}, \quad P_j(x_D^k) = \frac{\tau[E_j(x_D^k)] + \beta_j}{N(x_D^k)}, \quad (4)$$

where $N(x_i^k)$ provides for normalization, α_i and β_j are free parameters, $\tau(x_i^k a) = t(x_i^k a|x^k)$,

$$\tau(x_i^k a) = \sum_{v \in A} f[\tau(v x_i^k a)], \quad 0 \leq i < D, \quad (5)$$

$$\tau[E_j(x_D^k)] = \sum_{a \in A} g[\tau(x_j^k a)] - \lambda_j(x_j^k), \quad 0 \leq j \leq D, \quad (6)$$

$\lambda_j(x_j^k) = 0$ in the most of experiments, $f(z)$ and $g(z)$ are arbitrary non-negative functions.

We considered in experiment

- $Q = 1$;
- $D = 1 - 5$;
- sliding window (buffer) of size $B = 4096, 8192$ and 16384 bytes;
- three different functions: linear function (L), which equals z for all values of z , saturation function (S), which equals z if $z \leq Q$ and equals Q otherwise, and "zeroing" function (Z), which equals z if $z \leq Q$ and equals zero otherwise ;
- 10 files of approximately the same length 50000 bytes: ASM, C and Pas are an Intel 8086 assembler language program, a commented C program and Pascal program respectively; ED, ET, RD and RT are texts (E - english, R - russian, D - fiction, T - technical); EXE and OBJ are MS-DOS executive and object files respectively; SFP is an HP LaserJet soft font.

3 Results of experiments

We shall define modifications(modes) of CW-algorithm with two capital letters: SL-mode, ZS-mode, etc, which describe functions $g(z)$ and $f(z)$ respectively. For example, original CW-algorithm with the best Moffat's choice of conditional probabilities (see [2]) corresponds to SL-mode with $\alpha_i = \beta_j = 0$. Let us note that the L-function can be used as $f(z)$ only.

First we considered the influence of the choice of B and D with $\alpha_i = \beta_i = 0$ on the efficiency of three modes. $B = 16384$ appeared to be the best, but the optimal values of D depend on file, $3 \leq D \leq 5$ (see Table 1). The minimum rates R for SL and SZ-modes are almost the same

MODE	D	ASM	C	PAS	ED	ET	RD	RT	EXE	OBJ	SFP
SL	2	.219	.296	.250	.338	.371	.398	.432	.525	.540	.316
	3	.200	.248	.229	.286	.330	.360	.408	.504	.512	.301
	4	.200	.239	.228	.278	.324	.355	.408	.500	.509	.278
	5	.200	.238	.230	.278	.328	.358	.411	.501	.508	.255
ZL	2	.222	.298	.256	.341	.370	.401	.433	.541	.558	.318
	3	.204	.255	.239	.295	.337	.373	.424	.526	.534	.306
	4	.203	.248	.238	.290	.338	.374	.433	.523	.531	.286
	5	.202	.246	.239	.291	.344	.378	.440	.525	.531	.267
SZ	2	.215	.295	.246	.339	.373	.397	.438	.518	.520	.315
	3	.196	.246	.224	.286	.335	.361	.417	.498	.490	.303
	4	.197	.239	.225	.280	.333	.360	.420	.495	.488	.280
	5	.198	.241	.228	.283	.339	.366	.424	.498	.489	.255

Table 1: Dependence of rate R on D , $B = 16384$, $\alpha_i = \beta_i = 0$.

(ZL-mode is slightly worse). And 4 is the best choice of D (the rates are minimum or close to minimum for all files).

Then the rate R was minimized for every considered mode and for every D separately (but at average over all files) by the choice of α_i and β_i (see Table 2).

MODE	D	ASM	C	PAS	ED	ET	RD	RT	EXE	OBJ	SFP
SL	3	.195	.243	.224	.281	.327	.357	.409	.500	.506	.299
	4	.193	.230	.221	.270	.321	.352	.411	.493	.500	.274
	5	.191	.227	.222	.268	.325	.353	.411	.493	.499	.249
SZ	3	.191	.241	.220	.282	.333	.359	.420	.493	.482	.301
	4	.191	.231	.218	.273	.331	.358	.423	.487	.478	.277
	5	.190	.231	.221	.274	.337	.362	.426	.490	.478	.251
ZZ	3	.190	.239	.219	.280	.330	.356	.416	.488	.478	.300
	4	.189	.229	.217	.272	.328	.354	.419	.483	.473	.276
	5	.189	.229	.220	.273	.334	.359	.421	.485	.472	.250
SS	3	.183	.233	.213	.272	.323	.348	.408	.481	.474	.294
	4	.181	.220	.209	.260	.317	.343	.408	.473	.467	.268
	5	.179	.216	.210	.258	.322	.344	.410	.473	.466	.240
ZS	3	.182	.231	.212	.271	.321	.346	.405	.477	.471	.292
	4	.180	.218	.209	.259	.315	.340	.405	.470	.464	.267
	5	.178	.215	.209	.258	.319	.342	.407	.470	.462	.239
ALG	ASM	C	PAS	ED	ET	RD	RT	EXE	OBJ	SFP	
ARJ	.198	.251	.233	.297	.378	.397	.463	.503	.463	.221	
PKZ	.149	.249	.231	.294	.376	.395	.459	.502	.448	.221	
LIM	.150	.247	.229	.291	.371	.389	.451	.499	.443	.219	
HA	.153	.211	.207	.254	.307	.325	.388	.470	.452	.256	

Table 2: Table 2. The rate R : a) for different modes and D with optimal values of α_i and β_i , b) for known compression algorithms.

We used such notations of known algorithms as: ARJ = ARJ 2.41a, PKZ = PKZIP 2.04d, LIM = LIMIT 1.0, HA = HA 0.98.

The optimal values of parameters are $\lambda_D(x_D^k) = 1, \alpha_D = \beta_D = .5$ and $\lambda_i(x_D^k) = \alpha_i = \beta_i = 0, i < D$, for all modes. They are well matched with PCP-model of a source.

For ZS-mode the rate is 10% - 15% less then for original CW-algorithm and for such files as ASM, C and OBJ. In the most cases these results are close to the best known result for HA-algorithm of Mr. Harry Hirvola. It is interesting and important that he uses the zeroing function as $f(z)$. We are grateful to Mr.Hirvola for some comments of HA-algorithm which he sent us by E-mail.

The received experimental results permit us to plan more or less logical ways of the new improvements of CW-algorithm.

References

- [1] Cleary J.G., Witten I.H., "Data Compression Using Adaptive Coding and Partial String Matching". IEEE Trans. on Commun., 1984, Vol.32, No.4, pp. 396-402.
- [2] Bell T., Witten I.H., Cleary J.H., "Modeling for Data Compression". ACM Computing Surveys, 1989, Vol.21, No.4, pp. 557-591.
- [3] Shtarkov Yu.M., Volkov S.V., "Applying Universal Coding of Markov Sources to Compression of Finite-Length Strings". Proc. of 5th Joint Soviet-Swedish Intern. Workshop on Inform. Theory, Jan. 13-19, 1991. Moscow, 1990, pp. 138-142.
- [4] Shtarkov Yu.M., "About Cleary-Witten Algorithm". This Proceedings.
- [5] Rissanen J., Langdon G.G., "Universal Modeling and Coding". IEEE Trans. Inform. Theory, 1981, Vol.27, No.1, pp. 12-23.