The Zero-Frequency Problem in Predictive Self-Organizing Lists Data Compression

1 R − 5

Damras WONGSAWANG *and* Masayuki OKAMOTO
Shinshu University

1 Introduction

An adaptive or real-time data compression technique always faces with the problem of the occurrence of unexpected events called "The Zero-Frequency Problem". It is the problem of estimating the likelihood of a novel event occurring which is important in statistical adaptive data compression because the occurrence of novel events impair the coding efficiency. This problem has been described in [1] and [4], and also the estimating of the probabilities of novel events has been proposed for arithmetic coding in PPM algorithm [3]. In predictive self-organizing lists data compression (PLDC), described in [5], it is necessary to reserve a space in the code table for a novel character/word. Because all characters/words in the code table are referred by their positions, a suitable allocation of space for novel event will help improve the compression efficiency. In this presentation, the zero-frequency problem in PLDC will be studied and investigated. We then present the allocation of space for novel event by varying with times and occurrence rates of novel events. These methods are evaluated and compared with the conventional method in terms of the compression rates efficiency.

2 Novel Events in PLDC

In PLDC algorithm, both encoder and decoder maintain an identical code table which contains a list of characters previously occurred. The zero-frequency problem is encountered when a character appears for the first time. For single code table model, there is only one code table sharing among all prediction contexts. The number of novel events in this case will be limited by 2^8 (for 8-bit symbol) and a model of any orders gives the same distribution of novel events. We have studied and investigated the occurrence of novel events for many types of text files. Figure 1 (a) shows the distributions of novel events for some selected files, e.g., source programs, UNIX manual pages, electronic news, etc., plotting against time (amount of data processed).

In multiple code table model, the number of novel events will increase as many as $n * 2^8$, where n is the number of code tables, and different order models give different distributions. The same set of text files has been used to study the novel event distributions for multiple code table. Figure 1 (b) shows the distributions of novel events for multiple code table of *3rd* order model.

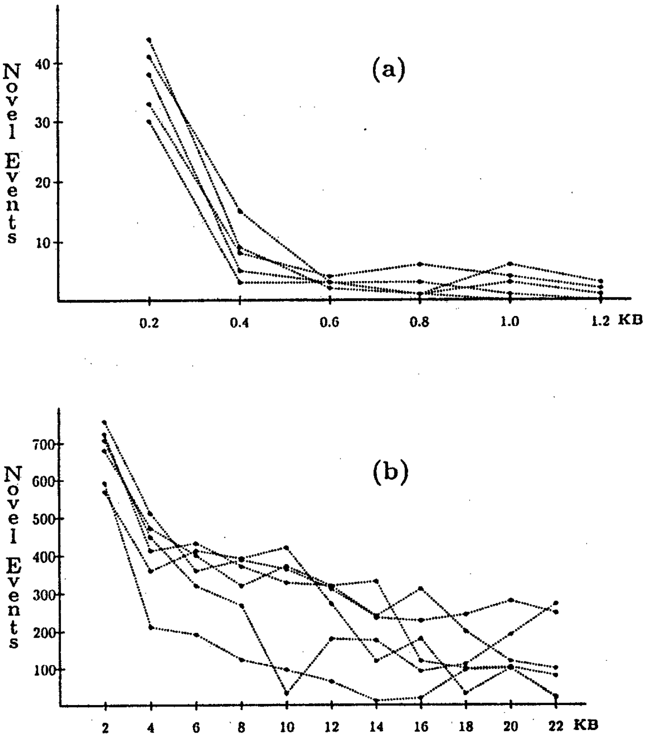

Figure 1 : The novel event distributions (a) for single code table model and (b) for multiple code table model (*3rd* order).

3 Allocation of Space for Novel Event

3.1 Conventional Allocation

This allocation is based on the assumption that novel event has the least likely to occur, i.e., has a zero probability, or has the smallest probability of occurrence. So novel symbols are always placed at the last position. The locally adaptive data compression [2] and PLDC [5] adopt this method by sending an integer $n + 1$ to signal the decoder of novel symbol, where n is the current number of symbols in the code table.

The Zero-Frequency Problem in Predictive
Self-Organizing Lists Data Compression
Damras WONGSAWANG *and* Masayuki OKAMOTO
Shinshu University

3.2 Improved Allocation

We propose the alternative allocations of space for novel events based on the assumption that novel events occur more frequently at the beginning than the later period. The space allocated for novel event is not necessary to be the last position. We have developed two allocation techniques for novel events, by varying with times, called **Method A**, as

$$f_1(t) = \begin{cases} k_i & \text{if } t_i < t \leq t_{i+1}, i = 1, 2, ..., j; \\ n+1 & \text{otherwise,} \end{cases}$$

and varying with occurrence rates of novel events, called **Method B**, as

$$f_2(t) = \begin{cases} k_i & \text{if } r_i < r(t) \leq r_{i+1}, i = 1, 2, ..., j; \\ n+1 & \text{otherwise,} \end{cases}$$

where k_i's, t_i's and r_i's are predetermined positions, amount of data processed and occurrence rates of novel events respectively,
$r(t)$ is the occurrence rate of novel events at t.

4 Implementation and Experiment

We have developed two simple ad hoc allocation techniques for novel events based on Method A and Method B. These two allocation techniques have been applied to the PLDC program. A fixed context, multiple code table model of order 3 has been tested with text files, e.g., C source programs (files 1-3), UNIX manual pages (files 4-6), text excerpted from electronic news (file 7), UNIX OS and Utilities document (file 8). We define the compression rate as the percentage of reduction in size after compression, i.e., *(original-compressed)/original *100.* We monitor the coding efficiency for the first 3000 bytes of compression process. The summary of the experimental results, comparing between conventional and two proposed allocations, are shown in Table 1. The three numbers in each block indicate compression rates at first 1000, 2000 and 3000 bytes of processing respectively.

5 Conclusions

In this presentation we have studied and investigated the conventional approach to the zero-frequency problem used in the self-organizing lists data compression, and also proposed the alternative approach to approximate the occurrence of novel events. Two simple approximations of novel event occurrences have been developed and applied to the PLDC. We compared the results with the conventional approach and found that a small improvement in terms of compression rates efficiency can be obtained by these two allocation techniques. The results show that the proposed allocations, with the same resource requirement as conventional approach, can help improve the compression rates efficiency by 4-10% during the beginning of compression process. There is no significant difference of improvement between the two proposed allocations. However, the improvement decline as process continue because of few occurrences of novel events in the later period and at that time no allocation technique has effect on the compression rate efficiency.

No	File	Conventional	Method A	Method B
1	yyy1.c	49.74 49.71 55.58	51.56 52.31 57.14	51.24 50.39 56.03
2	yyy2.c	37.85 45.02 55.16	42.20 49.45 58.69	42.20 49.29 58.55
3	yyy3.c	21.09 29.23 36.62	29.03 33.30 39.48	29.03 33.03 39.79
4	less.1	17.78 33.34 41.11	26.58 37.76 43.75	26.58 38.36 44.12
5	xterm.n	18.73 27.60 34.71	27.74 33.18 37.70	27.74 33.18 37.83
6	rn.n	26.96 29.04 35.24	35.40 34.79 38.76	35.40 34.79 38.84
7	E-News	28.09 34.58 38.95	31.10 36.74 39.74	31.10 36.47 39.68
8	UNIX Doc	37.00 30.83 36.21	41.16 36.55 40.28	41.16 36.55 39.70

Table 1 : The percentages of reduction at the first 1000, 2000 and 3000 bytes of processing comparing among 3 allocations of novel events.

References

[1] Bell T.C.,Cleary J.G. and Witten I.H.; *Text Compression*, Prentice Hall, Englewood Cliffs, NJ, 1990.

[2] Bentley J.L., Sleator D.D., Tarjan R.E. and Wei V.K.; *A Locally Adaptive Data Compression Scheme*, Comm. ACM, Vol. 29, No. 4, 320-330.

[3] Cleary J.G. and Witten I.H.; *Data Compression Using Adaptive Coding and Partial String Matching*, IEEE Trans. Comm., Vol. COM-32, No. 4, 396-402.

[4] Witten I.H. and Bell T.C.; *The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression*, IEEE Trans. Inf. Theory, Vol. IT-37, No. 4, 1085-1094.

[5] Wongsawang D., Okamoto M.; *Predictive Text Compression with Self-Organizing List*, Proc. 45th Annual Convention of IPS Japan, (Oct 1992), Vol. 1, 69-70.