Combination between static Arithmetic Coding and probability (Dynamic) Arithmetic Coding to compress data

Arshed Adham Ahmad **Mohammed S. Mohammed**

Department of computer/ Diyala Iraq

Diyala University/ college of Education for pure science

ABSTRACT

The key idea to arithmetic coding was done and implemented completely by replacing the input symbol with a specific code. A series of symbols can be coded by the interval zero to one, closed interval [0, 1]. Arithmetic coding using many methods and need many bits especially if the message is long and complex, so the compression must be found to reduce the number of bits by using probability methods. Also by combination between methods we can reduce the interval [0, 1] to less than using one method for arithmetic coding.

KEYWORDS: coding; probability; Huffman; compression; arithmetic.

BIOGRAPHICAL NOTES

Mohammed S. Mohammed received his MSc. in Computer Science from Technology University in 2008. He is currently Assistant Lecturer at the Department of Computer Science, Diyala University, Diyala, IRAQ. His current research interest of Arithmetic Coding combining with probability to get minimum range to compress to use it in a probably way with minimum values to use the remaining level of rang [0 to 1] for another compressing to achieve the better range of it .

Arshad A. Ahmed received his MSc. in mathematical Science from AL-Nahrin University He is currently a Lecturer at the Department of Computer Science, Diyala University, Diyala, IRAQ. His current research interest of using probability with a compatible technique to

combined with arithmetic coding, by using it in a multiple level with arithmetic coding as a static and dynamic according to the step that we get it.

الدمج بين التشفير الرقمي القياسي والاحتمالية (العشوائية) لضغط وتشفير البيانات

محمد سامي محمد
العراق / جامعة ديالى
كلية التربية للعلوم الصرفة
مدرس مساعد

ارشد ادهم احمد
العراق / جامعة ديالى
كلية التربية للعلوم الصرفة
مدرس

الخلاصة

ان فكرة التشفير الرقمي هي باستبدال الرمز الداخل بكود معين. وان مجموعة من الرموز من الممكن ان تضغط وتشفر خلال الفترة من الصفر – الواحد . هناك طرق عديدة للتشفير الرقمي وتحتاج الى مراتب كثيرة وخاصة اذا كانت النصوص طويلة ومعقدة ، لذلك سنستخدم هنا الاحتمالية وذلك من اجل ضغط هذه الفترة . بحيث تستخدم مع الطريقة التقليدية لتقليل هذه الفترة.

الكلمات المفتاحية: التشفير، الاحتمالية، هوفمان، الضغط، الرقمي.

INTRODUCTION

The main drawback of Huffman scheme is that has problems when there is a symbol with very high probability. Where static Huffman redundancy bound is redundancy

$$\leq P_1 + 0.086$$

where P_1 is the probability of the most likely symbol .

At first we take the text and code it with binary code, then we deal with the text and compress it in first step by normal solution, the solve the second step by wing frequency probability of the text. And then find the result in these methods only [1] .

PROBABILITY

Probability is ordinarily used to describe an attitude of mind towards some proposition of whose truth we are not certain. The proposition of interest is usually of the form "Will a specific event occur" The attitude of mind is of the form "How certain are we that the event will occur?" The certainty we adopt can be described in terms of a numerical measure and this number, between 0 and 1, we call probability. The higher the probability of an event, the more certain we are that the event will occur. Thus, probability in an applied sense is a measure of the likeliness that a (random) event will occur.

The concept has been given an axiomatic mathematical derivation in probability theory, which is used widely in such areas of study as mathematics, statistics, finance, gambling, science, artificial intelligence/machine learning and philosophy to, for example, draw inferences about the likeliness of events. Probability is used to describe the underlying mechanics and regularities of complex system(2)

SIMILAR DATA (TEXT CODING)

To code similar symbol for text in a file like "mmmm" with probability 0.3 called probability of beginning P_a and probability 0.6 called probability of end P_b, also must defined the end of file which can denoted by $P_{EoF} = 0.1$ for example. so:

$$P_a + P_b + P_{EoF} = 1 \qquad (1)$$

$$0.3 + 0.6 + 0.1 = 1.0$$

The interval which is [0, 1] is the interval working on it to compress and code, we can built a subinterval by using this fixed arithmetic coding methods:-

$$P_a = 0.3 \rightarrow low_{old} = 0 \rightarrow low_{new} = 0 + 0.3 = 0.3$$

So the interval will be [0 - 0.3], we have

$$P_b = 0.6$$

so the high will be

$$high_{new} = high_{new} + 0.6 = 0.9 \quad (2)$$

high(new) $\rightarrow$ high of the new interval. So the interval will be [0.3-0.9] and the rest of the bit is for the ending files $P_{EoF} = [0.9 - 1.0]$

For each step we take it in fixed methods we divide the probability by "2" for example. 2nd step $P_a = 0.3/2 = 0.15$

So the interval will be [0.3-0.3+0.15] = [0.3-0.45].Also we have $P_b = 0.6$ in 2nd step will be $0.6/2 = 0.3$ So the interval will be [0.45- 0.45+0.3] = [0.45- 0.75]

The probability of end of file will be $P_{EoF} = 0.1/2 = 0.05$

$P_{EoF\ new} = [0.75 - 0.8]$

With two steps was duce the interval from [0.1] to [0.3-0.8].Until were each the limit of fixed model by this equation

$$P = (P_b)^3 (P_{EoF}) = (0.6)^3 (0.1) = (0.0216). \quad (3)$$

To get these ending we divided it by "2"

$P_{Eof\,(1st\ step)} = 0.1$

$P_{Eof\,(2nd\ step)} = 0.1/2 = 0.05 > 0.0216$

Continue …

$P_{Eof\,(3rd\ step)} = 0.05/2 = 0.025$.

Is now the 0.0216 step

So with three steps only in this example we can code and compress this text "mmmm".

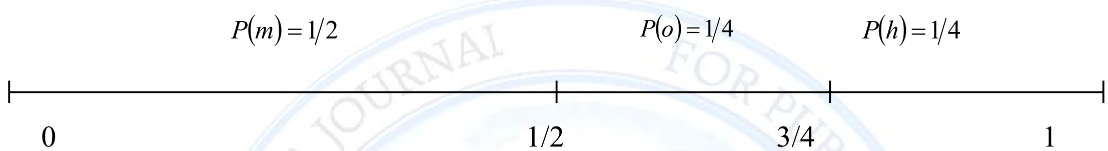
As we see the length of the symbol is don’t matter became it was similar symbol for text "mmmm" each symbol can coded by 16 bit so the total will be 16×4=64 bit while by using this methods we can find the length of the code by using[4]

$$-\log P = -\log 0.6 \quad (4)$$

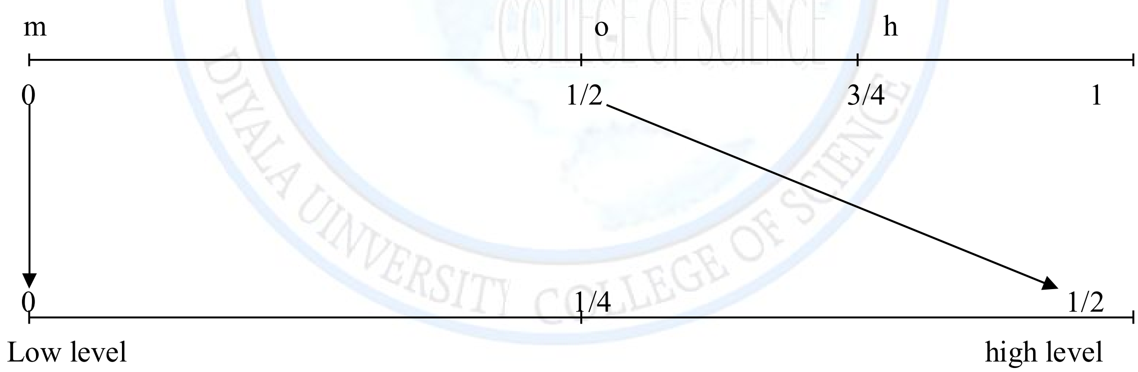
TEXT CODING AND COMPRESSION WITH DIFFERENT PROBABILITY

Depending on equation (3) , If we have the text "mohm" and with probability of $P(m)=1/2, P(o)=1/4, P(h)=1/4$

So the compression will be as follow:-



P(m) has a low level with zero and a high level with 1/2 and the probability of "o" has a low level of 1/2 and a high level of 3/4 . Also the p (h) has a low level of 3/4 and a high level of 1. The text that we have to compress and code with different probability start with "m" then



Probability of m in the second step will be half the first step

Rang = high value – low value (4)

$= 1/2 - 0 = 1/2$

Then

Pnew(m)) =P (m)old*range new (5)

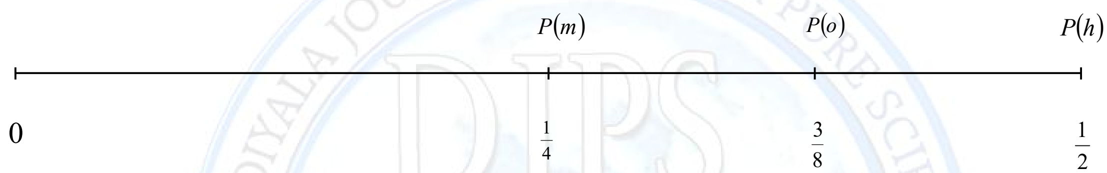
$$=1/2*1/2=1/4$$

Then the low value will be 0 and high will be 1/4 .The probability of 0old is 1/4 then the new level will start from 1/4 and end by

$$P(o)_{old}\ 1/4*range_{new}\ 1/2=1/8 \qquad (6)$$

the new probability.

So the high level will be 1/4 + 1/8 which it is the high level of mplus the probability of o which it is equal 3/8 then



And so on, until we get all the symbol of the text. If we want to get the value of each symbol (low and high) we can get it by these laws:-

Range = high – low (7)

low(new b) = low(old b) +range new *low(old b) (8)

high(new b) = low(old b) +range new *high(old b) (9)

ARITHMETIC CODING

Arithmetic coding is a form of variable-length entropy encoding used in lossless data compression. Normally, a string of characters such as the words "hello there" is represented using a fixed number of bits per character, as in the ASCII code [4].

When a string is converted to arithmetic encoding, frequently used characters will be stored with fewer bits and not-so-frequently occurring characters will be stored with more bits, resulting in fewer bits used in total. Arithmetic coding differs from other forms of entropy encoding such as Huffman coding in that rather than separating the input into component symbols and replacing each with a code, arithmetic coding encodes the entire message into a single number, a fraction n where (0.0 ≤ n < 1.0) (5).

COMBINATION BETWEEN METHODS TO COMPRESS AND CODING IN ARITHMETIC

Let's take the text "AMMR" and using combination between the frequency of symbol and the probability as follow:-

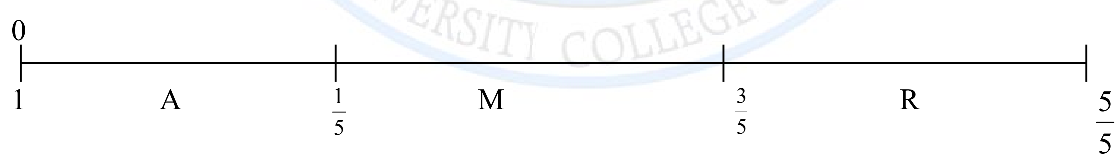
- At first we take these probabilities:

$P(A)=\frac{1}{5},\ P(M)=\frac{2}{5},\ P(R)=\frac{2}{5},$

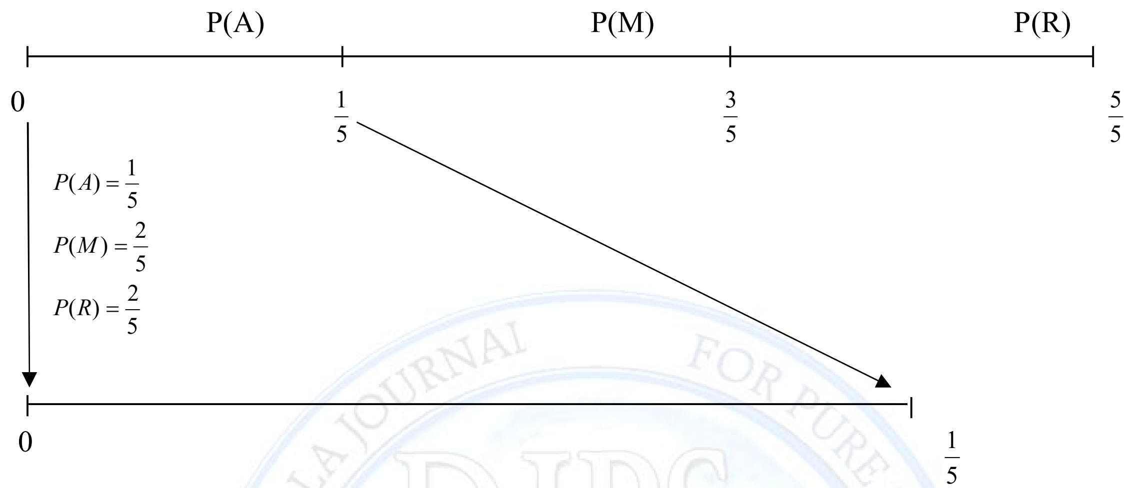
So the total probabilities equal to "1"

- Then let's draw the line below from (0 - 1):-

Write the low and the high level for each symbol

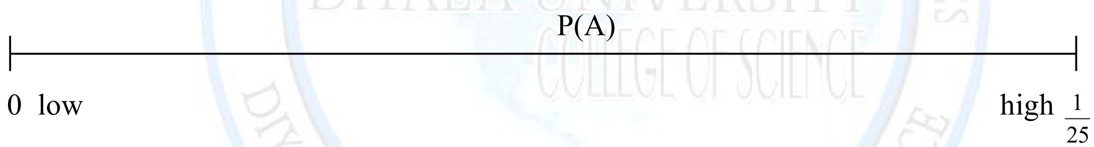


"AMMR"



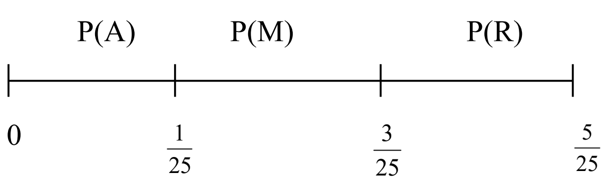
$$Rang = \frac{1}{5} - 0 = \frac{1}{5}$$

$$P(A)_{new} = P(A)_{old}\,(Rang)_{new} = \frac{1}{5}\cdot\frac{1}{5} = \frac{1}{25}$$



$$P(M) = P(M)_{old}\cdot Rang_{new} = \frac{2}{5}\cdot\frac{1}{5} = \frac{2}{25}$$

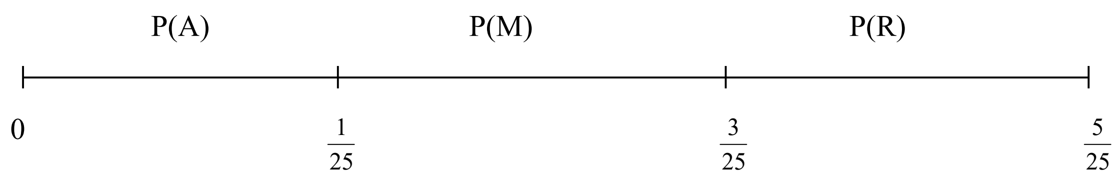
$low = \frac{1}{25} \qquad high = \frac{3}{25}$



$$P(R) = P(R)_{old}\cdot Rang_{new} = \frac{2}{5}\cdot\frac{1}{5} = \frac{2}{25}$$

4 $ow = \frac{3}{25} \qquad high = \frac{5}{25}$

Then we take other methods 1st :-

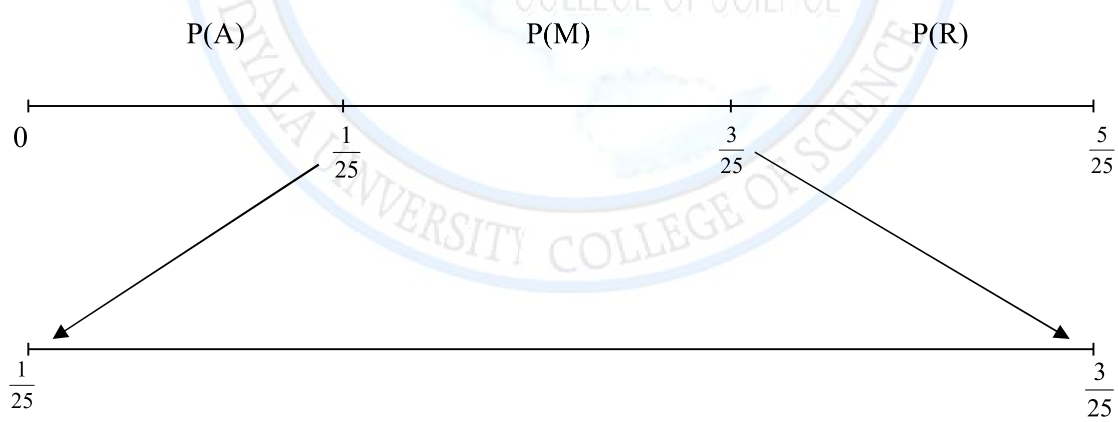


2nd we find the second probability by using other method to combing between two find the result:-

So we take A at first by using the old probability and the (now) we used M by using the new probability of (equal probability) like this:-

We take at first step a one's time and M one's time and R one's time now we take the second character M so the probability of A is one and the probability of M is 2 and probability of R is one.The total probability is A+M+M+R=4.

$P(A)=\frac{1}{4},\quad P(M)=\frac{2}{4},\quad P(R)=\frac{1}{4}.$



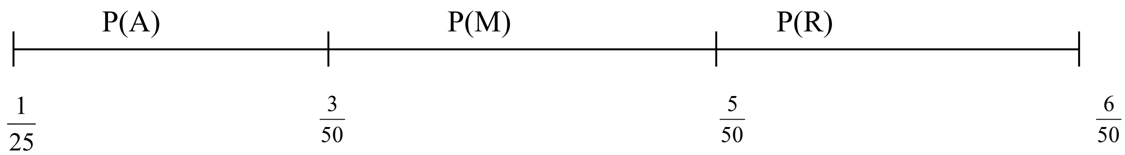
$$Range=\frac{3}{25}-\frac{1}{25}=\frac{2}{25}$$

$$P(A)=\frac{1}{4}\rightarrow P(A)_{new}=\frac{1}{4}\cdot\frac{2}{25}=\frac{1}{50}$$

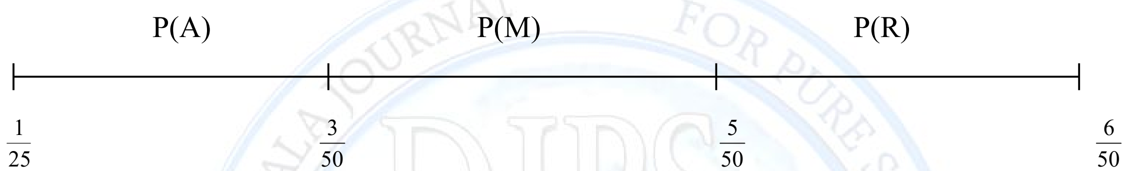
$$P(M)=\frac{2}{4}\rightarrow P(M)_{new}=\frac{2}{4}\cdot\frac{2}{25}=\frac{1}{25}$$

$$P(R)=\frac{1}{4}\rightarrow P(R)_{new}=\frac{1}{4}\cdot\frac{2}{25}=\frac{1}{50}$$

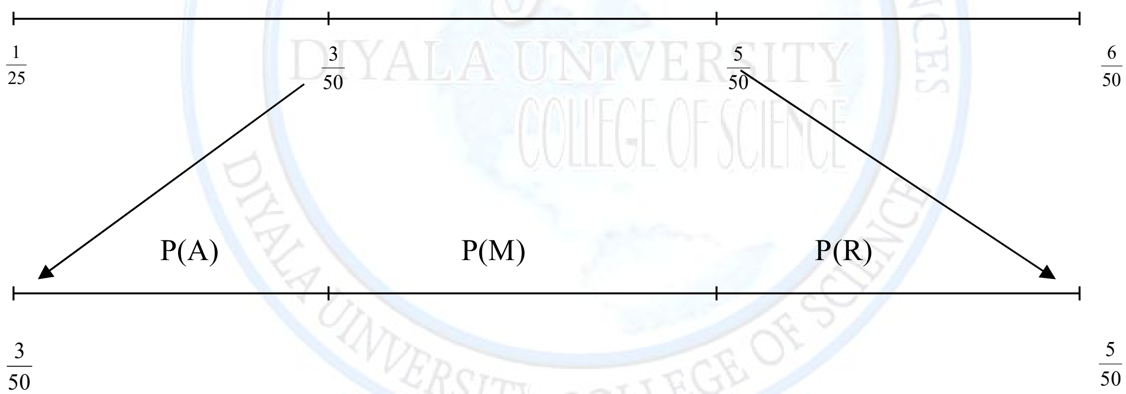
Now we find the low and high level



Now we take M again as follow:- A+M+M+M+R=5



$$P(A)=\frac{1}{5},\quad P(M)=\frac{3}{5},\quad P(R)=\frac{1}{5}.$$

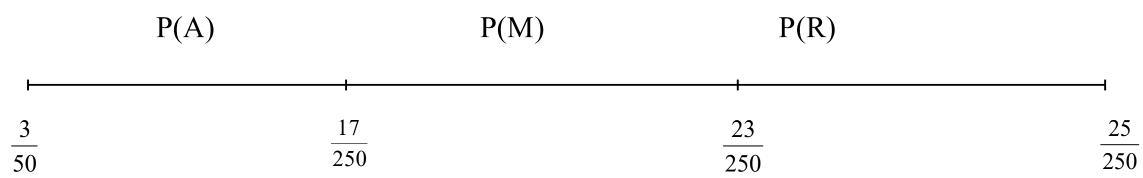


$$Range=\frac{5}{50}-\frac{3}{50}=\frac{2}{50}$$

$$P(A)=\frac{2}{50}\cdot\frac{1}{5}=\frac{2}{250}=\frac{1}{125}$$

$$P(M)=\frac{3}{5}\cdot\frac{1}{50}=\frac{3}{125}$$

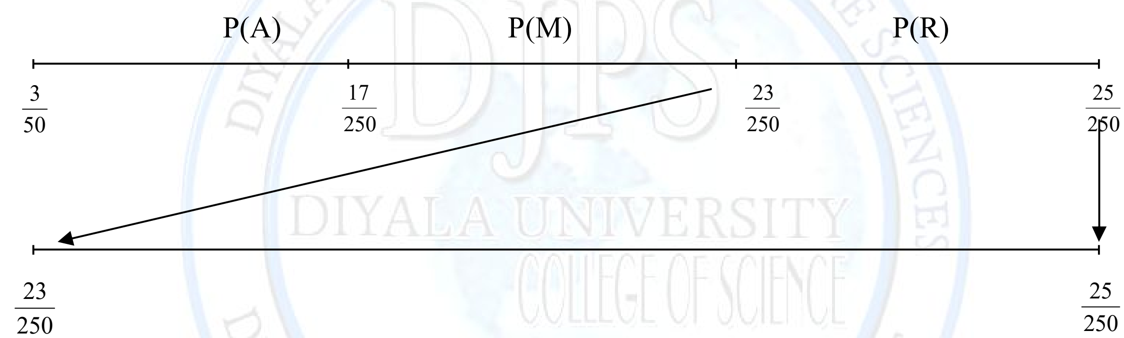
$$P(R)=\frac{1}{5}\cdot\frac{1}{25}=\frac{1}{125}$$



For the last time we take R so the probability is now:- A+M+M+M+R+R=6

$$P(A)=\frac{1}{6},\quad P(M)=\frac{3}{6}=\frac{1}{2},\quad P(R)=\frac{2}{6}=\frac{1}{3}.$$

Then



$$Range=\frac{25}{250}-\frac{23}{250}=\frac{2}{250}=\frac{1}{125}$$

$$P(A)=\frac{1}{125}\cdot\frac{1}{6}=\frac{1}{750}$$

$$P(M)=\frac{1}{125}\cdot\frac{1}{2}=\frac{1}{250}$$

$$P(R)=\frac{1}{125}\cdot\frac{1}{3}=\frac{1}{375}$$

Then the Range is 1/125 instead of 1 which mean that R=0.008 from 0.092 to 0.1 probability of coding A is [low=0.092 and high=0.093] and probability of coding M is [low=0.093 and high=0.0973] and probability of R is [low=0.0973 and high=0.1].

CONCLUSION AND FUTURE WORK

By using two methods range will be reduce too much.

Prbability of each variable will be approximately from 0.092 to 0.093 which it will be so near to each other, so the range was very useful to use the other remaining range for another compressing.

With using of many variables and symbols in this range (0-1) by compressing data by using these two combination methods.

The first method can be used at first step or at the last step will get the same result approximately.

Combine between these methods can be 0,1 which mean using one methods then another and vise versa.

REFERENCES

1. Alan Stuart and Keith Ord, (2009) ‘Kendall's Advanced Theory of Statistics’, Volume 1: Distribution Theory, 6th Ed.
2. J. Jian, (1995) ‘novel design of arithmetic coding for data compression’ IEEE proc. Compute. Dig Tech, vol. 142 (nov) 419-424.
3. P.G. Howard and J.S. Vitter, (1994) ‘Arithmetic coding for data compression proc’. IEEE vol. 82, no. 6, 857-865.
4. T. H. witten, R.M. neal, and J.G. (1987) ‘Cleary arithmetic coding for data compression’ C. ACM, vol. 30, no. 6, 520-540, jun 1987.
5. William Feller. (1986), ‘An Introduction to Probability Theory and Its Applications, 3rd Ed.

WEBSITES

6. Recommendation T.81 (1992) Corrigendum 1 (01/04). Recommendation T.81 (1992). International Telecommunication Union. 2004-11-09. http://www.itu.int Retrieved 3 February 2011.