アルファベット符号表を用いた情報量に関する解説

澤見英男

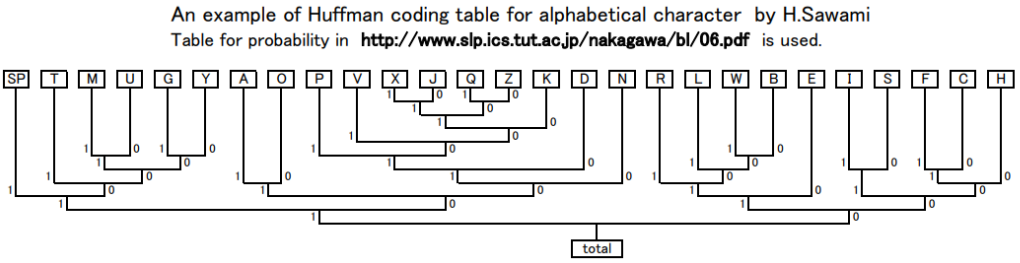
情報量はシャノン（Claude Elwood Shannon, 1916-2001）により導入された概念であり，熱力学における状態量エントロピ（entropy）に対応している。離散系のエントロピ，すなわち平均情報量 H は事象 k の生起確率 $p_k \neq 0$ を用いて以下のように定義されている。

$$H = -\sum_k p_k log p_k$$

対数の底を２にした場合の単位はビット（bit; binary digit）になる。ここでは，英文テキスト中のアルファベット文字列の生起確率を取り上げ，平均情報量の計算をする。例として，文字の大きさを区別しない場合，空白（SP）を含む組“**文字(生起確率)**”の集合には以下に示すようなものがある（http://www.slp.ics.tut.ac.jp/nakagawa/bl/06.pdf）。

{ SP(0.1817), A(0.0668), B(0.01179), C(0.0226), D(0.031), E(0.1073), F(0.02395), G(0.01633), H(0.04305), I(0.0519), J(0.00108), K(0.00344), L(0.02775), M(0.02075), N(0.0581), O(0.0654), P(0.01623), Q(0.00099), R(0.0559), S(0.0499), T(0.0856), U(0.0201), V(0.00752), W(0.0126), X(0.00136), Y(0.01623), Z(0.00063) }

これをもとにエントロピ符号化をすると，次の様なハフマン符号化テーブルが得られる。



chr	code	bit	p	-log2(p)
SP	111	3	0.1817	2.4604
A	1011	4	0.0668	3.904
B	011000	6	0.01179	6.4063
C	00010	5	0.0226	5.4675
D	10010	5	0.031	5.0116
E	010	3	0.1073	3.2203
F	00011	5	0.02395	5.3838
G	110001	6	0.01633	5.9363
H	0000	4	0.04305	4.5378
I	0011	4	0.0519	4.2681
J	1001100110	10	0.00108	9.8548
K	10011000	8	0.00344	8.1834
L	01101	5	0.02775	5.1714
M	110011	6	0.02075	5.5907
N	1000	4	0.0581	4.1053
O	1010	4	0.0654	3.9346
P	100111	6	0.01623	5.9452
Q	1001100101	10	0.00099	9.9803
R	0111	4	0.0559	4.161
SP	0010	4	0.0499	4.3248
T	1101	4	0.0856	3.5462
U	110010	6	0.0201	5.6367
V	1001101	7	0.00752	7.0551
W	011001	6	0.0126	6.3104
X	1001100111	10	0.00136	9.5222
Y	110000	6	0.01623	5.9452
Z	1001100100	10	0.00063	10.632

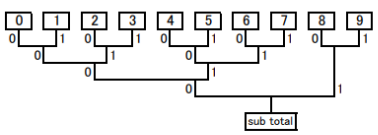
メモ

出現確率の最小となる文字はZ(p=0.0006)そして2番目に小さい文字はQ(p=0.001)であることがわかる。

数字0から9までの確率の和がこの範囲にあるとき，文字Zの確率を数字も含んだものとすると，
文字Zの元の符号1001100100（10ビット）の下に10011001001（11ビット）で表し，符号10011001000（11ビット）の下には，
数字0から9までを4ビット分拡張した符号に分岐させ組み入れることで数字を含む符号化テーブルへと拡張をすることもできる。

例えば，数字0を符号1001100100000000（15ビット），．．．．，数字7を符号1001100100000111（15ビット），
数字8を符号1001100100010（13ビット），数字9を符号1001100100011（13ビット）へと符号化することもできる（下図参照）。

このようにして，多くのビット数（情報量）をより出現確率の低い記号に割り振ることにより符号化をするのがハフマン符号化である。



chr	code	bit	p	length
0	0000	4	1/16	0.25
1	0001	4	1/16	0.25
2	0010	4	1/16	0.25
3	0011	4	1/16	0.25
4	0100	4	1/16	0.25
5	0101	4	1/16	0.25
6	0110	4	1/16	0.25
7	0111	4	1/16	0.25
8	10	2	1/4	0.5
9	11	2	1/4	0.5

<－－－－－－－
平均ビット長　4.10504
エントロピ：　4.06507
（平均ビット長はエントロピよりも長くなる）

－－－－－－－>
平均ビット長　3
エントロピ：　3
（確率が2のマイナス冪乗なので平均ビット長とエントロピは一致する）

図１　ハフマン符号化テーブルの作成例

二進木としても表現できるハフマン符号化表を組“**文字（文字毎の生起確率, 文字毎のエントロピー$-\log_2 p_k$, 文字を表す符号のビット長)**”の集合として表すと以下のようになる。

{ SP(0.1817, 2.460 , 3), A(0.0668, 3.904 , 4), B(0.01179, 6.406 , 6), C(0.0226, 5.468 , 5),
D(0.031, 5.012 , 5), E(0.1073, 3.220 , 3), F(0.02395, 5.384 , 5), G(0.01633, 5.936 , 6),
H(0.04305, 4.538 , 4), I(0.0519, 4.268 , 4), J(0.00108, 9.855 , 10), K(0.00344, 8.183 , 8),
L(0.02775, 5.171 , 5), M(0.02075, 5.591 , 6), N(0.0581, 4.105 , 4), O(0.0654, 3.935 , 4),
P(0.01623, 5.945 , 6), Q(0.00099, 9.980 , 10), R(0.0559, 4.161 , 4), S(0.0499, 4.325 , 4),
T(0.0856, 3.546 , 4), U(0.0201, 5.637 , 6), V(0.00752, 7.055 , 7), W(0.0126, 6.310 , 6),
X(0.00136, 9.522 , 10), Y(0.01623, 5.945 , 6), Z(0.00063, 10.632 , 10) }

これらの結果より，確率が低くなるほど事象の情報量の増すことすなわち符号が長くなること，確率が２のべき乗ではないことから多少の誤差を生じていることが分かる。すなわち，エントロピ H は **4.065** ビットであるのに対し，ハフマン符号の平均符号長は **4.105** ビットと少しだけ多目の値になっている（図１）。一方，ここで示しているアラビア数字に関する組“**数詞（確率)**”の集合 **{ 0(1/16), 1(1/16), 2(1/16), 3(1/16), 4(1/16), 5(1/16), 6(1/16), 7(1/16), 8(1/8), 9(1/8) }** に関しては，確率が２のべき乗であることから，エントロピ H とハフマン符号の平均符号長は，共に３ビットと，一致することになる（図１）。

ハフマン符号化は，事象 k のエントロピ値 $-\log_2 p_k$ と符号長がほぼ同じになるか，または確率が 2 のべき乗であれば一致することから，エントロピ符号化とも呼ばれている。

ここで示した具体例では，文字または数詞を区別するために用いる符号の長さが，確率の２を底とした対数の負値 $-\log_2 p_k$ に対応していることを示している。すなわち確率の低い事象に長い符号を割り当てていることになる。すなわち，ハフマン符号化すなわちエントロピ符号化では，確率の低い事象から合算していくように段階（対数に対応）を追って配置しながら二進木を構成していることになる。そして，この枝分かれを１ビットにより区別・記録することは，情報そのものを区別・記録することに対応する。すなわち、事象の有する情報量とは確率の対数の負値により評価できる量であり，確率の低い事象の方が確率の高い事象よりも情報量は多くなるということを意味している。

ところで，環境の記述を除いて確率だけで評価すると，百万分の一の生存率が意味するところは 20 ビットの情報処理により生き残りが可能ということになるので，下等生物でも条件さえ整えば繁殖できるということになる。このようなことから妄想すると，人類だと数百億ビット・レベルの情報処理が出来ることから，遠い未来には，未知の惑星に移植し生存できる様になっていても不思議ではないかもと思えてしまう。