

第2章 コンピュータの仕組み

1. 情報の表現

1. “表現”の様々な側面

情報の表現には様々な種類がある。以下にその例を示す。

一. 「手続き型表現」と「宣言型表現」

手続き型表現： 「2つ目の角を右に曲がって、少し歩いた右側です。」と、手順に沿った表現。

宣言型表現： 「高島屋の隣です。」と、対象間の関係や対象の属性などに基づく表現。

二. 「記号表現」と「パターン表現」

記号表現： 数学記号・方程式・論理式など、表現の記号とそれを解釈する規則体系に基づいた表現。

下の例〔図1〕は、地図記号や国道の記号など、その表現と意味との対応関係に基づいた地図の表現である。これは、描き手と読み手の間でこの対応関係を共有していることを前提とした表現である。

パターン表現： 表現の個々の要素にはそれ自体に意味はなく、構成要素間の時空間的パターンによって情報を表す表現。

下の例〔図2〕は、道路のみを線で表した大雑把な地図である。これは、構成要素の線などには特に意味がなく、描き手と読み手の間で前提知識を共有していなくとも通ずる表現である。

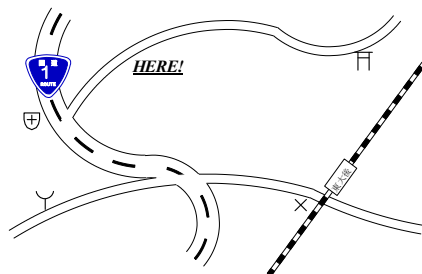


図1

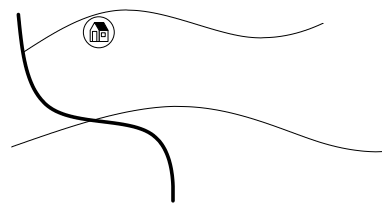


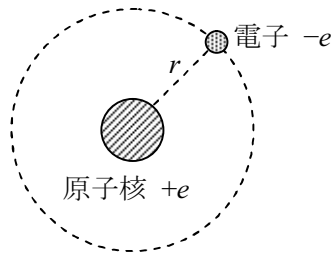
図2

※ 記号表現とパターン表現の差は明確なものではなく、例えば上の〔図2〕であっても「家の形」が「家」を表すなどという規則体系が前提として必要である。

情報の表現には、これらの他にもデジタル/アナログや情報量など多くの側面がある。これらは必ずしも独立はでない。情報の表現には、こうした側面を多角的に考慮する必要がある。

2. 情報の表現とモデル

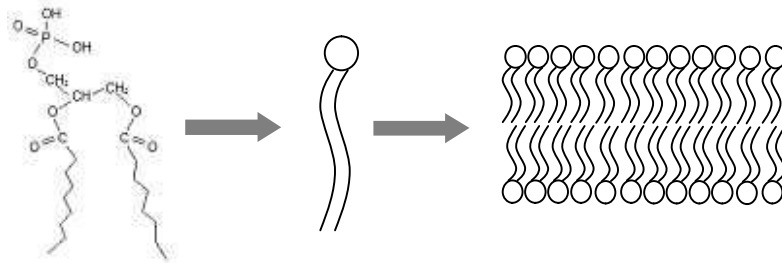
物事概念を単純化・抽象化して表したものを一般に**モデル**と呼ぶ。例えば、原子核の周りを回る電子の様子を二つの質点の関係として単純化することで(長岡モデルなど)、細かい様子を考慮せずとも電子のだいたいの動きが把握できる〔図3〕。また、リン脂質の構造を親水基と疎水基のみで単純化して表すことで、二重膜構造を解り易く表すことができる〔図4〕。



〔図3〕水素原子モデル

原子核と電子を単純化して質点と見做すモデルを考えると、細かいことを考えずとも次の式を導ける。

$$m \frac{v^2}{r} = k_0 \frac{e \cdot e}{r^2}$$



〔図4〕

リン脂質の親水基と疎水基だけ表せば、脂質二重層を表すには十分である。

モデル化の手法として用いられる表現形式としては、以下のようなものがある。

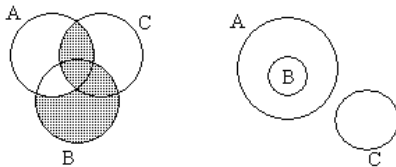
一. 表

日付	長さ
7/18	89cm
7/20	110cm
7/21	123cm
7/22	130cm
7/23	140cm
7/24	150cm
⋮	⋮
8/18	414cm
8/19	442cm
8/20	470cm
8/21	470cm
8/22	495cm
8/23	520cm

表は、込み入った事柄を整理して見易くするために一般的に用いられている表現形式である。最近では計算機上の表計算ソフトの利用も一般的であり、これにより、ヒストグラムや円グラフなどの目的に応じた見易い表現形式に変換することも容易になっている。

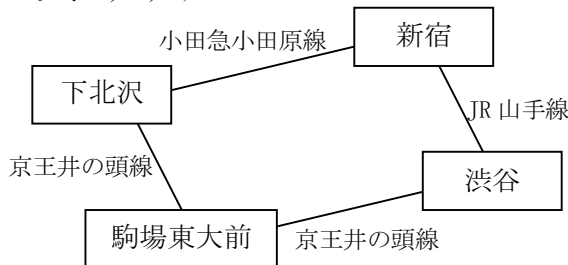
(因みに左の表は 1999/7/18~8/22 のヘチマの観察記録である ©関根)

二. 図



図は、人間が何らかの目的をもって描いた 2 次元図形を指す。上の地図 1・2 等も図である。このような図は、人間の思考・推論を支援・拡張する役割を持つ。例えば、左の Venn 図や Euler 図は、複雑な集合関係を理解するのに役立つ。

三. グラフ



左のように**ノード**(頂点)と**エッジ**(辺)から構成されるものを**グラフ**と呼ぶ。左の場合、ノードが駅、エッジが鉄道路線となっている。方向を持つエッジを**有向エッジ**或いは**弧**と呼ぶ。グラフは、組織図・PERT 図・意味ネットワーク等で幅広く用いられる。

3. 情報の表現とは

情報の表現においては、少なくとも以下の 3 つの事項について考慮する必要がある。

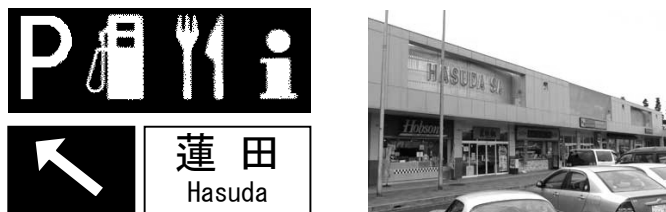
- 表現の対象**：何が表現されているのか？ / 何を表現するのか？
- 表現の目的**：何のために表現されているのか？ / 何のために表現するのか？
- 表現の方法**：どう表現されているのか？ / どう表現するのか？

II. 記号と表現

1. 図記号(ピクトグラム) —— 記号と意味

一. 抽象化された図形

情報表現に関係した具体例として、**図記号(ピクトグラム)**を考える。下図はサービスエリア(SA)を表すピクトグラム、写真(NEXCO 東日本のサイトより)は蓮田サービスエリアである。



このピクトグラムの表現の対象は SA、目的はドライバーに SA の所在を知らせることである。もしサービスエリアの写真がそのまま標識だったら、ドライバーは何の標識だか理解し辛いだろう。この例のようにピクトグラムは抽象化された図形を用いてデザインされ、視覚的に認知されやすいように工夫されている。情報の表現には、その記号のデザインや配置等を考慮する必要がある。

二. シニフィアンとシニフィエ

記号論によれば、記号は次の2つの側面を持つ。

シニフィアン (signifiant) : 何かを意味している記号表現 (上の例ではピクトグラム)

シニフィエ (signifié) : 意味されている記号内容(上の例では SA)

※ 教科書 p. 17 11 の末尾、たぶん誤植です。読めばわかると思いますが、文章が不自然です。





三. 自然言語の比喻との関係

サービスエリアを表すナイフとフォークの絵は、「ナイフ」と「フォーク」を意味しているのではなく、合わせて「サービスエリア」を意味している。これは自然言語における**提喻**(関連する別の概念であるものを表す)に相当する表現技法である。コンピュータの GUI (*Graphical User Interface*, 教科書第9章参照)における**アイコン**にも似た例がある。例えば、ファイル削除手続きの為に用いられている「ごみ箱」アイコンは、「ごみを捨てる」という行為の**隠喩**として表現されている。



四. 記号の恣意性

上の例では、「『ナイフとフォーク』が『サービスエリア』を表す」という“取りきめ”のもとに図記号が成立している。このような**記号の恣意性**は、時に混乱を招く。

日本では、記号  で禁止が表されることが多い。また、 は肯定、 は否定を表す、というように考えられている。しかし、外国では必ずしもそうではない。下の2つの図は欧州における道路標識の例であるが、(a)は「全車両通行禁止」、(b)は「二輪車以外の車両通行禁止」である。日本の感覚ではこの は肯定(通行可)と見做してしまいがちであり、注意が必要である。



記号表現の問題は他にもある。例えば、日本語の文字コードの問題である。日本語を表す文字コードには、**JIS**、**シフト JIS**、**日本語 EUC (EUC-JP)** などがある。例えば「関」の字では、シフト JIS では 35542、EUC では 46296、Unicode では 38306 となる(以上全て十進数)。逆に、同じコードであっても、それが何の文字コードでエンコード(符号化)されているのかがわからなければ元の文字が何であったかわからない。

2. 数の表現 —— 記号と解釈の規則体系

一. 表現の方法とトレードオフ

数の表現には、アラビア数字・ローマ数字・漢数字など、色々ある。アラビア数字を用いた位取り記数法は、次のように計算に用いるのに向いている。

$$\begin{array}{r} 1989 \\ + 19 \\ \hline 2008 \end{array}$$

ローマ数字を用いて MCMLXXXIX + XIX としたのではうまくいかない。しかしアラビア数字にも欠点がある。例えば、1989 と書かれたものに対し、前後に数字を足して 1231989456 などと改竄できてしまう。これを防ぐためにローマ数字(MCMLXXXIX) や漢数字の大字(壱千九百八拾九)などが用いられることもある。また、金額の場合 ¥1,989- などという表記も用いられる。

この場合の「計算にはアラビア数字が向いているが、改竄の虞を考えるとローマ数字の方が良い」などという二律背反の関係は、**情報表現間のトレードオフ**と呼ばれる。

二. コンピュータにおける数の表現

コンピュータではすべての数値が**ビット列**で表現されている。つまり、「0」と「1」の2種類の記号のみで表現されるということである。コンピュータで表現できる数値はシステムが採用しているビット数(ビット列の桁数)に依存する。例えば16ビットなら0から65535までの65536 (= 2¹⁶)種類の数値が表せる。(実際には負の数も表現する必要がある)

III. アナログとデジタル

1. アナログ表現とデジタル表現

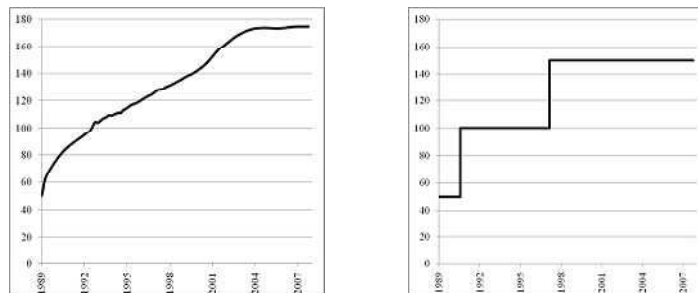
アナログ量 : 物体の色や音の大きさなどの連続的な量。この表現を**アナログ表現**という。無限の精度が必要なため、そのデータの複製は近似でしかありえない。

デジタル量 : ある目盛りの値に近似された離散的な量。この表現を**デジタル表現**という。精度が有限のため、劣化させずに複製が可能。

2. 量子化

アナログ量をデジタル量に変換するには、その値をある一定間隔の離散的な量に近似する必要がある。これを**量子化**と呼ぶ。

例えば、下の2つの曲線はいずれも同じ人の身長の変化を表したものである。身長は本来アナログ量であるが、身長計で測るときに量子化されてデジタル量となっている。下の2つは、量子化の精度が異なるのである。左は1mm単位、右は50cm単位で量子化されている。1mm単位なら実用上問題ないが、50cm単位ではもはや身長の推移がわからない。



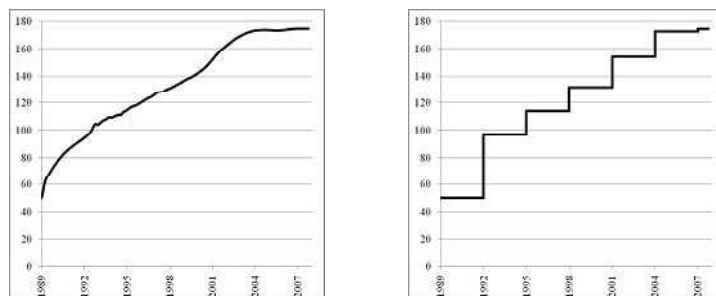
この例のように、量子化の上で重要なのは実用上問題のない程度の十分な精度である。コンピュータのディスプレイでは赤緑青の3成分をそれぞれ256段階に分けて表している。また、音楽CDでは音の振幅を65536段階で表している。

3. 標本化と標本化定理

一. 標本化

アナログ量をデジタル量に変換するには、その範囲における抽出(記録)の間隔(頻度)が重要である。この操作を**標本化**と呼び、標本化の対象となる範囲を**標本空間**と呼ぶ。

下の2つの曲線は、先程と同じく身長の変化を表したものである。今度は両者の標本化の頻度が異なる。左は1年に1度以上の頻度で標本化されているが、右は3年毎である。3年毎の標本化では身長の変化が大雑把にしかわからない。



二. 標本化定理

では、どの程度の頻度で標本化すべきなのだろうか。この疑問に答えるのが、**シャノンの標本化定理**である。これによると、標本間隔 t_0 で標本化した場合、正しく復元できる精度は標本間隔 $2t_0$ (周波数 $1/2t_0$) までである。つまり、必要な精度の $1/2$ 倍の標本間隔で標本化する必要があるということである。この周波数 $1/2t_0$ を、標本間隔 t_0 の**ナイキスト周波数**と呼ぶ。ナイキスト周波数は復元できる周波数の上限であり、これより高い周波数をもつ周期関数が成分として含まれていると誤ったデータが復元されてしまう(**エイリアシング**)。

音楽 CD の場合、周波数 44.1 kHz で記録しているので、復元できる周波数の上限は 22.05 kHz である。人間の可聴音域は最大 20 kHz 程なので、22.05 kHz 再現できれば十分である。

IV. デジタル符号化

1. 2進符号によるデジタル符号化

2進符号とは、普段我々が用いている十進数を二進数に変換したものと捉えられる。十進数の $(1)_{10}$, $(2)_{10}$, $(3)_{10}$, ..., $(10)_{10}$ は、4桁の二進数 $(0001)_2$, $(0010)_2$, $(0011)_2$, ..., $(1010)_2$ で表される。

ここで、2つの2進符号において、対応する桁の0と1が異なる個数を**ハミング距離**と呼ぶ。例えば、 $(0000)_2$ と $(0001)_2$ とは4桁目のみが異なるのでハミング距離は1、 $(0000)_2$ と $(1010)_2$ とは1桁目と3桁目の2つが異なるのでハミング距離は2、といった具合である。

2進符号において、一般に数値の差とハミング距離とは一致しない。これを一致させるようにした符号として**グレイ符号**というものがある。

V. 第2章 あとがき

纏まりのない文章で大変申し訳ない。具体例が微妙なのは関根の PC 内のデータから適当に作ったからである。

このシケプリを第2章から作り始めたのは、第1章はただの全体の概観であり、大して内容が無いからである(洒落ではない)。気になる人は(気にならない人も)、教科書を読んでみて欲しい。

尚、このシケプリに収録した範囲は、理系諸君にとって必要であろう「必須学習項目(赤色)」と「要望学習項目 B(青色)」の範囲のみである。

このシケプリは、読者の成績を保障するものではなく、関根は成績に関して一切責任を負わないものとする。かく言う関根自身の成績も全く保障されていない。