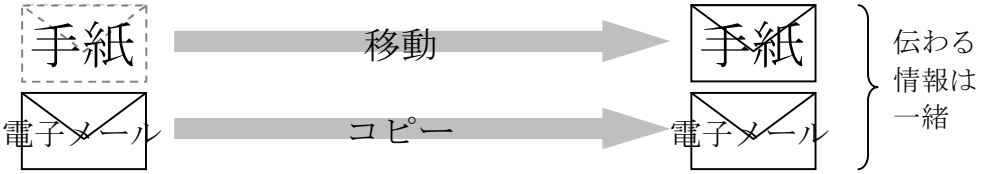
第 3 章 情報の伝達と通信

Ⅰ. 情報の伝達と情報量

1. 情報の伝達

情報の伝達には、その**メディア**(媒体)は関係ない。メディアには電話・電報・手紙・電子メール・FAX・狼煙など色々あるが、同じ**メッセージ**が伝わるならば何れであっても関係ない。

手紙の場合は送り手の手元からはメディアが消えてしまうが、電子メールの場合は残る。ここで重要なのは、送り手ではなく**受け取り側の状態の変化**であるということである。



また、興味のないメッセージが伝えられても、情報としての価値は少ない。例えば、

「今回の情報の試験は共通問題だけ」

というのなら情報としての価値があると考えられるが、

「今日の朝青龍の夕飯はカレー」

などというのであればどうでもよい。そこで、情報としての価値は

・自分に影響のある、これまで知らなかった事実を知った

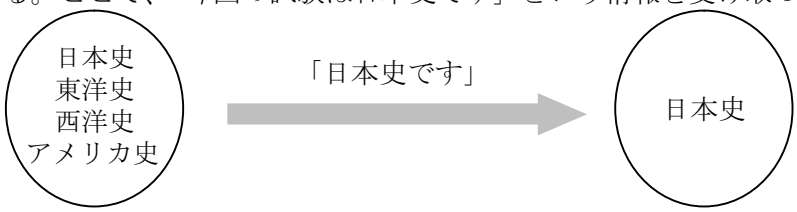
・何らかの判断の材料にできる事実を知った

という場合にあると考えられる。

このような観点から、情報の大きさを表す量を定義してみる。

2. 情報の大きさ —— 情報量

「歴史」の試験において「日本史」「東洋史」「西洋史」「アメリカ史」のうち何れか１つが出題されるとする。ここで、「今回の試験は日本史です」という情報を受け取った場合を考える。



情報を受け取ったことにより、学生は 4 分野のうち 1 分野だけを勉強すればよいことになる。では、「情報を 3 だけ受け取った」と考えて良いのだろうか。

答えは NO である。「4 分野 → 1 分野」というのは「100 分野 → 97 分野」というのと比べて明らかに有益な情報である。しかし、この定義では両方とも「情報を 3 だけ受け取った」となってしまい、情報量が情報の有益性を正しく表していないことになってしまうのである。

情報量を考える上で大切なのは、場合数(この場合は勉強すべき分野の数)の「**差**」ではなく「**比**」である。「4 分野 → 1 分野」というのは、「100 分野 → 97 分野」というのよりも寧ろ「100 分野 → 25 分野」というのと同等の情報と考えられる。

では、「比」をそのまま情報量としてよいのだろうか。情報量を物理量と同様に考えるのならば、情報をx受け取ってからy受け取ったときの全体の情報量は$x+y$であって欲しい(**情報量の加法性**)。

以上の要請は、場合数の**比の対数**(= 対数の差)を考えることで解決する。場合数が「$A \rightarrow B$」と変化したときの情報量を $\log\frac{A}{B}$ と定義するのである。こうすれば $\log\frac{A}{B} + \log\frac{B}{C} = \log\frac{A}{C}$ となり、確かに「$A \rightarrow B \rightarrow C$」と変化した場合の情報量が「$A \rightarrow C$」の場合と同じになっている。普通、二者択一の場合(2 → 1)の情報量が 1 となるように、対数の底は 2 とする。

定義

情報量 $\stackrel{\text{def}}{=} \log_2\frac{\text{事前の場合数}}{\text{事後の場合数}}$　　(単位：**ビット**)

もっと一般化すると、次のようになる。

$$\text{情報量} = \log\frac{\text{事前の場合数}}{\text{事後の場合数}} = -\log\frac{\text{事後の場合数}}{\text{事前の場合数}} = -\log(\text{確率})$$

このように考えると、確率の低い事柄の方が大きな情報であると言える。例えば、

「犬が人間を噛んだ」

というありふれた事柄はニュースにならないが、

「人間が犬を噛んだ」

という稀有な出来事はニュースとなり得る。

定義

情報量 $\stackrel{\text{def}}{=} -\log_2(\text{確率})$　　(単位：**ビット**)

3. 平均情報量

情報量の平均(期待値)を**平均情報量**という。上の例で、メッセージが「日本史」「東洋史」「西洋史」「アメリカ史」のいずれかである場合の平均情報量は、

$$\text{平均情報量} = \frac{1}{4}\times\left(-\log_2\frac{1}{4}\right)+\frac{1}{4}\times\left(-\log_2\frac{1}{4}\right)+\frac{1}{4}\times\left(-\log_2\frac{1}{4}\right)+\frac{1}{4}\times\left(-\log_2\frac{1}{4}\right)=2$$

となる。つまり、等確率の状況では、個々のメッセージの持つ情報量と平均情報量とが一致する。

等確率でなければこうはいかない。例えば、メッセージが「日本史」か「世界史」の場合、

$$\text{平均情報量} = \frac{1}{4}\times\left(-\log_2\frac{1}{4}\right)+\frac{3}{4}\times\left(-\log_2\frac{3}{4}\right)\fallingdotseq 0.811$$

となる。珍しい情報(「日本史」)と珍しくない情報(「世界史」)が相殺するのである。

平均情報量は、全てのメッセージが等確率で来る場合(予想が全くつかない場合)に最大となる。

4. 符号化と情報量

一. 符号化

メッセージの伝達では、「日本史」「東洋史」「西洋史」「アメリカ史」という日本語を用いずとも、0 と 1 という符号を用いて 00, 01, 10, 11 のように表せばそれで十分である。このようにメッセージを符号で表すことを**符号化**という。代表的な符号化の手法である 2 進符号化については第 2 章のⅣ節を参照。

メッセージの伝送を速くするには、データを小さくすればよい。そのための効率の良い符号化を考える。

二．データの圧縮

n 個のメッセージ m_1，m_2, …，m_n があるとする(例えば「日本史」「東洋史」……)。m_i を長さ l_i の符号で表すものとする。m_i が確率 p_i で現れる場合、n 個の記号を符号化した長さの期待値は $\sum_i np_il_i$ となる。従って、1 個の記号を符号化した長さの期待値は $\sum_i p_il_i$ となる。この $\sum_i p_il_i$ のことを**平均符号長**と呼ぶ。

毎年の「日本史」「世界史」の出題の様子を符号化することを考える。つまり、「07 年は世界史、06 年は世界史、05 年は日本史、……」などというデータの符号化である。ここで、先程と同じく「日本史」「世界史」の出題確率はそれぞれ 1/4，3/4 とする。

まず、1 年毎に符号化してみる。「日本史」→ 0，「世界史」→ 1 として符号化すると、平均符号長は 1 となる。

次に、2 年分を纏めて符号化してみる。右図のように､確率の出現の高いものほど符号が短くなるようにハフマン符号化(課題でやりましたね)すると、1 年分あたりの平均符号長は

$$\frac{1}{2}\sum_i p_il_i = \frac{1}{2}\left(\frac{1}{16}\cdot 3+\frac{3}{16}\cdot 3+\frac{3}{16}\cdot 2+\frac{9}{16}\cdot 1\right) \fallingdotseq 0.844$$

となり、1 年毎に符号化した場合よりも短くなっている。

出題 m_i	確率 p_i	符号	符号長 l_i
日・日	1/16	111	3
日・世	3/16	110	3
世・日	3/16	10	2
世・世	9/16	0	1

さらに 3 年毎、4 年毎、…… としていくと平均符号長はだんだん短くなっていく。平均符号長の下限は平均情報量(0.811)となることが知られている(**情報源符号化定理**)。

Ⅱ．情報通信

1．プロトコル

通信する上で、互いに意図の誤解のないように予め定めておく決めごとを**プロトコル *(protocol)*** という。例えば「電話の冒頭では『もしもし』と言う」「無線通信で発話の終わりに『どうぞ』と言う」なども一種のプロトコルである。

現在のインターネットでは **TCP/IP** と呼ばれるプロトコル群が用いられる。これについてはⅢ節で説明する。

2．通信の秘密と相手の認証

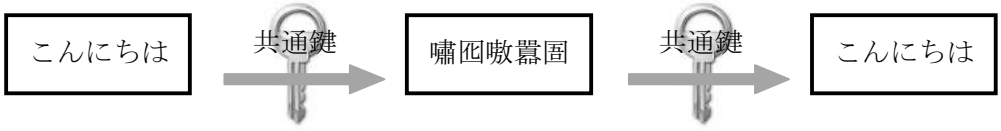
実際の通信では、秘密に送りたい情報を傍受されたり、偽の要求に騙されたりする虞がある。そこで、情報を秘密にする「暗号化」と、送信者を特定する「ディジタル署名」とが重要となる。

ここでは、暗号化の手法として「**共通鍵暗号**」と「**公開鍵暗号**」について説明する。

一．共通鍵暗号

暗号の議論では、元のデータを**平文**(ひらぶん)、暗号化したデータを**暗号文**という。また、暗号文から元の平文に戻すことを**復号**という。暗号化や復号は、専用の**鍵**を持っている人のみが行える。

共通鍵暗号においては、暗号化と復号に同じ鍵(**共通鍵**)が用いられる。つまり、送信者と受信者が同じ鍵を用いて通信を行うということである。



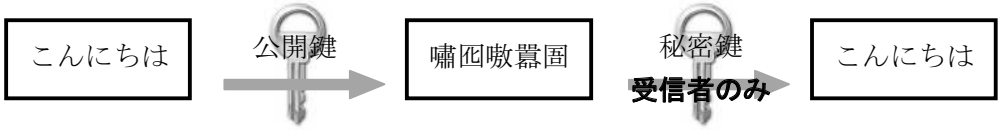
ここで、問題は如何にして送信者と受信者のみが同じ鍵を共有するかである。もし悪意の第三者に鍵がバレてしまったら、暗号文の内容は筒抜けである。

そこで、より安全な暗号化の手法として、公開鍵暗号がある。

二．公開鍵暗号

公開鍵暗号では、暗号化と復号に用いる鍵が別である。そして、**公開鍵**は誰でも知ることができる状態にし、**秘密鍵**は作成者本人以外の誰にも明かさないものとする。だから鍵がバレる可能性は共通鍵暗号と比べて格段に低い。

具体的には、公開鍵を用いて暗号化し、秘密鍵を用いて復号する。そうすれば、誰でも暗号化してメッセージを送ることができるが、そのメッセージを読めるのは秘密鍵を持つ受信者だけである。



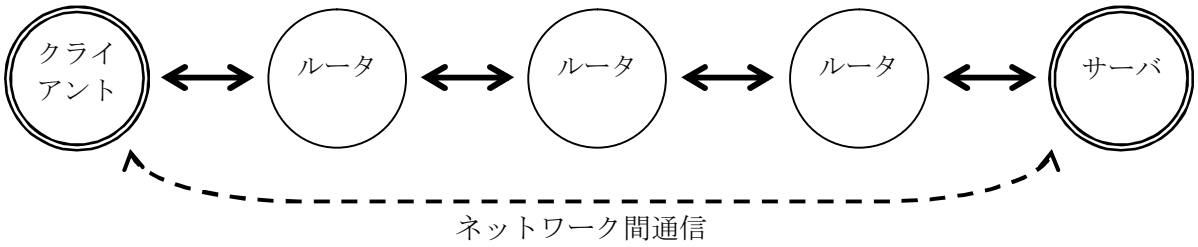
因みに、逆に秘密鍵を用いるとディジタル署名ができるが、それについては試験範囲外なので割愛する。

Ⅲ．インターネット

1．ネットワークの集合体と通信

インターネットは小規模な**ネットワーク**が互いに接続した集合体である。ネットワーク同士を接続する中継機器を**ルータ**と呼ぶ。

通信において、情報を要求する側をクライアント、その要求に応じて適切な情報を提供する側をサーバと呼ぶ。ここでは、異なるネットワーク内のクライアントとサーバの通信を見てみる。



ネットワーク間通信

1. クライアント(ブラウザなど)が相手サーバの IP アドレスを調べる。詳しい調べ方については後述。
2. クライアントはメッセージを**パケット**と呼ばれる細かい単位に分割し、パケット毎に相手の IP アドレスに向けて送信する。
3. 各パケットは、まず同一ネットワーク内のルータに届けられる。ルータはそれと同じネットワーク内の別のルータにパケットを届ける。
4. 順次ルータを回ってサーバに辿り着いたパケットは正しい順序に並べられ、元のメッセージが復元される。
5. サーバからクライアントへの応答も以上と同様に行われる。

以上のように、ネットワーク内通信(上図の↔)の連続によって、クライアントとサーバとの間のネットワーク間通信が為されるのである。

2．階層プロトコル

第Ⅱ節でも述べたように、インターネットでは TCP/IP と呼ばれるプロトコル群が用いられている。細かく見ていくと、このプロトコル群は階層的な構造を成している。

TCP/IP モデル	主なプロトコル	主な役割
アプリケーション層	**HTTP**, SMTP, DNS, DHCP	アプリケーション間の通信
トランスポート層	**TCP**, UDP	パケットの分割や合成
インターネット層	**IP**	ネットワーク間通信
ネットワークインターフェース層	（**イーサネット**）	ネットワーク内通信

上の表は、TCP/IP の主なプロトコルの階層と役割を示したものである。

先程の例で言うと、ルータ間のネットワーク内通信においては**イーサネット**と呼ばれる規格が用いられている。この規格によって初めて、クライアントとサーバとが通信できるのである。

(本当はイーサネットはプロトコルではないのだが、そんなことは気にしない)

クライアントとサーバとの間のネットワーク間通信では、より上位の **IP** と呼ばれるプロトコルが用いられている。これは、「送信者から来たパケットを受信者に届ける」という操作に関するプロトコルである。

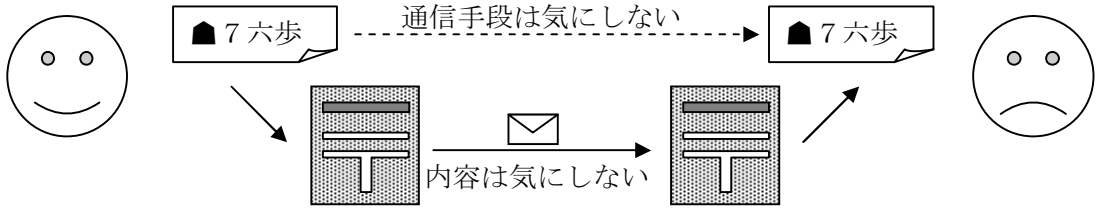
その上位の通信としては、**TCP** による通信がある。これは、「送信者が分割して送り出したパケットを受信者のところで組み立てる」ということなどを規定するプロトコルである。

最上位の通信として、**HTTP** による通信がある。これは、「ブラウザがデータを要求し、サーバがそれに応える」ということなどを規定するプロトコルである。

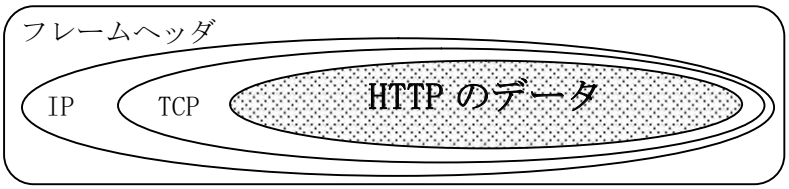
イーサネットレベルの通信では、何が通信されているのかは気にしない。ただ来たデータを次のルータ乃至サーバに送り届けるだけである。

IP レベルでの通信では、どのような経路で通信されているかなどは気にしない。ただ送信者から来たパケットを受信者に届けるだけである。

このように、他の階層の通信については気にしない(分離されている)のが**階層プロトコル**である。丁度、郵便将棋において、郵便屋は棋戦については気にせず、棋士は郵便については気にしないのと同様である。



階層的なプロトコルを実現するために、カプセル化という手法が用いられている。これは、データをプロトコル毎の制御データ(先頭の**ヘッダ**と末尾の**トレーラ**)で挟み込むというものである。データは上位のプロトコルから順にカプセル化され、下位の通信が行われる時にはより上位の様子は関係無いという状況になっている。



3. IP アドレスとポート番号

TCP/IP の通信では、相手を特定するためのアドレスとして **IP アドレス**が使われる。IP アドレスは 32 ビットの数値で、4 つに区切って十進数の組として表す(0.0.0.0 ～ 255.255.255.255)。

コンピュータ内のアプリケーションについては、16 ビットの**ポート番号**で識別する。

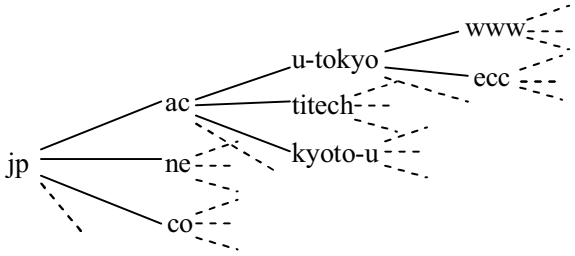
4. IP アドレスとホスト名の対応づけ —— DNS

一. ホスト名

前に述べた通りコンピュータの通信では IP アドレスが用いられるが、IP アドレスはただの数字であり人間には扱い難い。そこで、人間にわかりやすいコンピュータの識別名として、**ホスト名**が付けられる。ホスト名は www.u-tokyo.ac.jp のように英数字をピリオドで繋げたものである。

二. 分散管理

ホスト名 www.u-tokyo.ac.jp は、下図のような木構造のノードに対応している。www.u-tokyo.ac.jp など u-tokyo.ac.jp 以下のドメインは東京大学が管理しており、東工大 titech.ac.jp や京大 kyoto-u.ac.jp などから独立している。このように、ホスト名はドメイン毎に管理を分散させることができるのである。これが **DNS (*Domain Name System*)** である。



ホスト名に対応する IP アドレスを調べる場合、**反復問い合わせ**と呼ばれる問い合わせを行う。www.u-tokyo.ac.jp の場合、まずルートサーバに jp を管理するサーバを問い合わせ、次にそのサーバに ac.jp を管理するサーバを問い合わせ、今度はそのサーバに u-tokyo.ac.jp を管理するサーバを問い合わせ、最後にそのサーバに www.u-tokyo.ac.jp を管理するサーバの IP アドレスを問い合わせるのである。

尚、ルートサーバというのは DNS の頂点にあるサーバで、現在世界に 13 台ある。

あとがきは第 2 章と同じ。