§2. THÔNG TIN VÀ DỮ LIỆU

1. Khái niệm thông tin và dữ liệu

Thực ra không có sự khác biệt nhiều giữa khái niệm thông tin được hiểu trong đời sống xã hội và khái niệm thông tin trong tin học. Trước mỗi thực thể (sự vật, sự kiện) tồn tại khách quan, con người luôn muốn biết rõ về nó càng nhiều càng tốt. Sự hiểu biết đó càng ít thì con người càng khó xác định thực thể đó. Những hiểu biết có thể có được về một thực thể nào đó được gọi là thông tin về thực thể đó.

Ví dụ, khi đọc lời nhận xét của cô giáo chủ nhiệm: "Em Ngọc Hà ngoan, chăm chỉ và học giỏi" ghi trong "Sổ liên lạc", bố mẹ của Ngọc Hà có thêm thông tin về con mình.

Muốn đưa thông tin vào máy tính, con người phải tìm cách biểu diễn thông tin sao cho máy tính có thể nhận biết và xử lí được. Trong tin học, dữ liệu là thông tin đã được đưa vào máy tính.

2. Đơn vị đo lượng thông tin

Ta không chỉ dừng lại ở một quan niệm định tính về thông tin như trên mà còn cho thông tin một quan niệm định lượng. Mỗi sự vật hay sự kiện đều hàm chứa một lượng thông tin.

Muốn nhận biết một đối tượng nào đó, ta phải biết đủ lượng thông tin về nó. Tương tự, để máy nhận biết một đối tượng nào đó, ta cũng phải cung cấp cho máy đủ lượng thông tin về đối tượng này.

Đơn vị cơ bản đo lượng thông tin là bit. Đó là lượng thông tin vừa đủ để xác định chắc chắn một trạng thái của một sự kiện có hai trạng thái với khả năng xuất hiện như nhau.

Ví dụ, xét việc tung ngẫu nhiên đồng xu có hai mặt hoàn toàn cân xứng với khả năng xuất hiện của mỗi mặt là như nhau. Nếu kí hiệu một mặt của đồng xu

là 1 và mặt kia là 0 thì sự xuất hiện kí hiệu 1 hay 0 sau khi tung đồng xu cho ta một lượng thông tin là 1 bit.

Trong tin học, thuật ngữ bit thường dùng để chỉ phần nhỏ nhất của bộ nhớ máy tính để lưu trữ một trong hai kí hiệu, được sử dụng để biểu diễn thông tin trong máy tính, là 0 và 1.

Ví dụ, giả sử có dãy tám bóng đèn được đánh số từ 1 đến 8, trong đó một số bóng đèn sáng và một số khác tắt, chẳng hạn các bóng đèn thứ hai, ba, năm và tám sáng, các bóng còn lại tắt (h. 2). Nếu ta sử dụng kí hiệu 0 và 1 để biểu diễn tương ứng trạng thái tắt và sáng của mỗi bóng đèn thì thông tin về dãy tám bóng đèn trên được biểu diễn bằng dãy tám bit 01101001.

Hình 2

Để lưu trữ dãy bit đó, ta cần dùng ít nhất tám bit của bộ nhớ máy tính. Ngoài đơn vị bit nói trên, đơn vị đo thông tin thường dùng là byte (đọc là bai) và 1 byte bằng 8 bit. Người ta còn dùng các đơn vị bội của byte như bảng dưới đây:

Kí hiệu	Đọc là	Độ lớn
KB	Ki-lô-bai	1024 byte
MB	Mê-ga-bai	1024 KB
GB	Gi-ga-bai	1024 MB
TB	Tê-ra-bai	1024 GB
PB	Pê-ta-bai	1024 TB

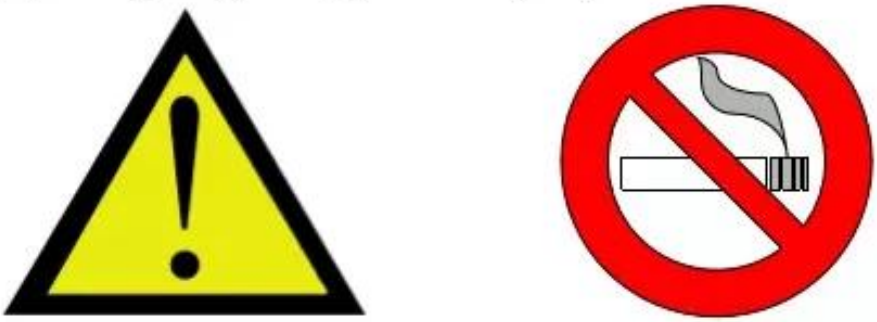
3. Các dạng thông tin

Thế giới quanh ta rất đa dạng nên có nhiều dạng thông tin khác nhau và mỗi dạng có một số cách thể hiện khác nhau. Có thể phân loại thông tin thành loại ***số*** (số nguyên, số thực,...) và loại ***phi số*** (văn bản, hình ảnh, âm thanh,...). Dưới đây là một số dạng thông tin loại phi số thường gặp trong cuộc sống.

a) ***Dạng văn bản:*** Là dạng quen thuộc nhất và thường gặp trên các phương tiện mang thông tin như: Tờ báo, cuốn sách, vở ghi bài, tấm bia,... (h. 3).

Hình 3. *Chữ khắc trên đá ở Mỹ Sơn – thông tin dạng văn bản*

b) ***Dạng hình ảnh:*** Bức tranh vẽ, bức ảnh chụp, bản đồ, băng hình,... là những phương tiện mang thông tin dạng hình ảnh (h. 4).



Hình 4. *Biển báo – thông tin dạng hình ảnh*

c) ***Dạng âm thanh:*** Tiếng nói con người, tiếng sóng biển, tiếng đàn, tiếng chim hót,... là thông tin dạng âm thanh (h. 5). Băng từ, đĩa từ,... có thể dùng làm vật chứa thông tin dạng âm thanh.

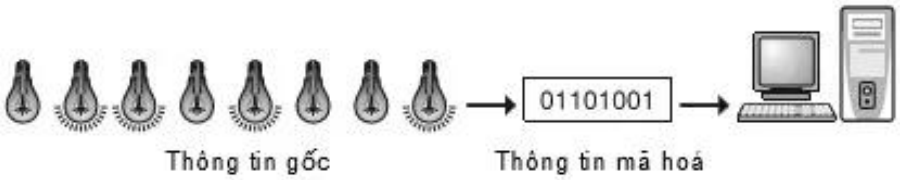
Hình 5. *Tiếng đàn T'rưng – thông tin dạng âm thanh*

Với sự phát triển của khoa học – kĩ thuật, trong tương lai con người sẽ có khả năng thu thập, lưu trữ và xử lí các dạng thông tin mới khác.

4. Mã hoá thông tin trong máy tính

Muốn máy tính xử lí được, thông tin phải được biến đổi thành một dãy bit. Cách biến đổi như vậy được gọi là một cách *mã hoá* thông tin.

Chẳng hạn, thông tin về trạng thái tám bóng đèn trong ví dụ trước được biểu diễn thành dãy tám bit là mã hoá của thông tin đó trong máy tính.



Hình 6. *Mã hoá thông tin trong máy tính*

Ví dụ, xét việc mã hoá thông tin dạng văn bản. Mỗi văn bản là một dãy các kí tự viết liên tiếp theo những quy tắc nào đó. Các kí tự bao gồm các chữ cái thường và hoa như a, b, c,..., z, A, B, C,..., Z; các chữ số thập phân 0, 1, 2,..., 9 và một số kí hiệu khác như các dấu phép toán, các dấu ngắt câu,...

Để mã hoá thông tin dạng văn bản, ta chỉ cần mã hoá các kí tự. Bộ mã ASCII (đọc là A-ski, viết tắt của American Standard Code for Information Interchange – Mã chuẩn của Mĩ dùng trong trao đổi thông tin) sử dụng tám bit để mã hoá kí tự (xem *Phụ lục 1. Bộ mã ASCII cơ sở*). Trong bộ mã này, các kí tự được đánh số từ 0 đến 255 và các số hiệu này được gọi là mã ASCII thập phân của kí tự.

Ví dụ, kí tự "A" có mã ASCII thập phân là 65 và kí tự "a" có mã ASCII thập phân là 97. Mỗi số nguyên trong phạm vi từ 0 đến 255 đều có thể viết trong hệ nhị phân với 8 chữ số (8 bit). Nếu kí tự có mã ASCII thập phân là N, dãy 8 bit biểu diễn N chính là mã hoá của kí tự đó trong máy tính. Ví dụ, mã ASCII của kí tự "A" là 01000001.

Bộ mã ASCII chỉ mã hoá được 256 (= 2^8) kí tự, chưa đủ để mã hoá tất cả các bảng chữ cái của các ngôn ngữ trên thế giới. Do đó với mã ASCII, việc trao đổi thông tin trên toàn cầu còn khó khăn. Bởi vậy, người ta đã xây dựng bộ mã Unicode sử dụng 16 bit để mã hoá. Với bộ mã Unicode ta có thể mã hoá được 65536 (= 2^{16}) kí tự khác nhau, cho phép thể hiện trong máy tính văn bản của tất cả các ngôn ngữ trên thế giới bằng một bộ mã. Hiện nay, nước ta đã chính thức sử dụng bộ mã Unicode như một bộ mã chung để thể hiện các văn bản hành chính.

Để con người có thể biết thông tin được lưu trữ trong máy, máy tính phải biến đổi thông tin đã mã hoá thành dạng quen thuộc như văn bản, âm thanh hoặc hình ảnh.

5. Biểu diễn thông tin trong máy tính

Dữ liệu trong máy tính là thông tin đã được mã hoá thành dãy bit.

Trong mục này, ta tìm hiểu cách biểu diễn thông tin loại số và phi số trong máy tính.

a) Thông tin loại số

- *Hệ đếm*

Hệ đếm được hiểu như tập các kí hiệu và quy tắc sử dụng tập kí hiệu đó để biểu diễn và xác định giá trị các số. Có hệ đếm phụ thuộc vị trí và hệ đếm không phụ thuộc vị trí.

Hệ đếm La Mã là hệ đếm không phụ thuộc vị trí. Tập các kí hiệu trong hệ này gồm các chữ cái: I, V, X, L, C, D, M. Mỗi kí hiệu có một giá trị, cụ thể:

I = 1; V = 5; X = 10; L = 50; C = 100; D = 500; M = 1000.

Trong hệ đếm này, giá trị của kí hiệu không phụ thuộc vị trí của nó trong biểu diễn. Ví dụ, X trong các biểu diễn XI (11) và IX (9) đều có cùng giá trị là 10.

Các hệ đếm thường dùng là các hệ đếm phụ thuộc vị trí. Bất kì một số tự nhiên b nào lớn hơn 1 đều có thể chọn làm cơ số cho một hệ đếm. Trong các hệ đếm này, số lượng các kí hiệu được sử dụng bằng cơ số của hệ đếm đó. Các kí hiệu được dùng cho hệ đếm đó có các giá trị tương ứng: 0, 1,..., $b - 1$.

Hệ thập phân (hệ cơ số 10) sử dụng tập kí hiệu gồm 10 chữ số: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Giá trị của mỗi chữ số phụ thuộc vào vị trí của nó trong biểu diễn. Ví dụ, trong số 545, chữ số 5 ở hàng đơn vị chỉ 5 đơn vị, trong khi đó chữ số 5 ở hàng trăm chỉ 500 đơn vị.

Giá trị số trong hệ thập phân được xác định theo quy tắc: mỗi đơn vị ở một hàng bất kì có giá trị bằng 10 đơn vị của hàng kế cận bên phải.

Ví dụ: $536{,}4 = 5 \times 10^2 + 3 \times 10^1 + 6 \times 10^0 + 4 \times 10^{-1}$.

Trong hệ đếm cơ số b, giả sử số N có biểu diễn:

$$d_n d_{n-1} d_{n-2} \ldots d_1 d_0, \; d_{-1} d_{-2} \ldots d_{-m}$$

trong đó $n + 1$ là số các chữ số bên trái, m là số các chữ số bên phải dấu phân chia phần nguyên và phần phân của số N và các d_i thoả mãn điều kiện $0 \le d_i < b$. Khi đó giá trị của số N được tính theo công thức:

$$N = d_n b^n + d_{n-1} b^{n-1} + ... + d_0 b^0 + d_{-1} b^{-1} + ... + d_{-m} b^{-m}.$$

Ghi chú : Khi cần phân biệt số được biểu diễn ở hệ đếm nào người ta viết cơ số làm chỉ số dưới của số đó. Ví dụ: 101_2 (hệ cơ số 2), 5_{16} (hệ cơ số 16).

• *Các hệ đếm thường dùng trong tin học*

Ngoài hệ thập phân, trong tin học thường dùng hai hệ đếm sau:

Hệ nhị phân (hệ cơ số 2) chỉ dùng hai kí hiệu là chữ số 0 và chữ số 1.

Ví dụ: $101_2 = 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 = 5_{10}$.

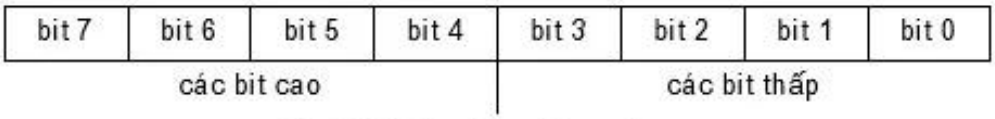
Hệ cơ số mười sáu, còn gọi là hệ hexa, sử dụng các kí hiệu: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F, trong đó A, B, C, D, E, F có các giá trị tương ứng là 10, 11, 12, 13, 14, 15 trong hệ thập phân.

Ví dụ: $1BE_{16} = 1 \times 16^2 + 11 \times 16^1 + 14 \times 16^0 = 446_{10}$.

• *Biểu diễn số nguyên*

Số nguyên có thể có dấu hoặc không dấu. Ta có thể chọn 1 byte, 2 byte, 4 byte,... để biểu diễn số nguyên. Mỗi cách chọn tương ứng với một phạm vi giá trị có thể biểu diễn được.

Xét việc biểu diễn số nguyên bằng một byte. Một byte có 8 bit, mỗi bit là 0 hoặc 1. Các bit của một byte được đánh số từ phải sang trái bắt đầu từ 0. Ta gọi bốn bit số hiệu nhỏ là các bit thấp, bốn bit số hiệu lớn là các bit cao (h. 7).



Hình 7. *Biểu diễn số nguyên*

Một cách biểu diễn số nguyên có dấu: dùng bit cao nhất thể hiện dấu với quy ước 1 là dấu âm, 0 là dấu dương và bảy bit còn lại biểu diễn giá trị tuyệt đối của số viết dưới dạng nhị phân. Theo cách đó, một byte biểu diễn được số nguyên trong phạm vi từ −127 đến 127.

Đối với số nguyên không âm, toàn bộ tám bit được dùng để biểu diễn giá trị số, một byte biểu diễn được các số nguyên không âm trong phạm vi từ 0 đến 255.

• *Biểu diễn số thực*

Cách viết số thực thông thường trong tin học khác với cách viết ta thường dùng trong toán học: dấu phẩy (,) ngăn cách giữa phần nguyên và phần phân được thay bằng dấu chấm (.) và không dùng dấu nào để phân cách nhóm ba chữ số liền nhau. Ví dụ, trong toán ta thường viết 13 456,25 nhưng khi làm việc với máy tính, ta phải viết 13456.25.

Mọi số thực đều có thể biểu diễn được dưới dạng $\pm M \times 10^{\pm K}$ (được gọi là *dạng dấu phẩy động*), trong đó $0,1 \le M < 1$, M được gọi là *phần định trị* và K là một số nguyên không âm được gọi là *phần bậc*.

Ví dụ: Số 13 456,25 được biểu diễn dưới dạng 0.1345625×10^5.

Máy tính sẽ lưu các thông tin gồm dấu của số, phần định trị, dấu của phần bậc và phần bậc.

b) Thông tin loại phi số

• *Văn bản*

Như đã nói ở phần trên, máy tính có thể dùng một dãy bit để biểu diễn một kí tự, chẳng hạn mã ASCII của kí tự đó.

Để biểu diễn một xâu kí tự (dãy các kí tự), máy tính có thể dùng một dãy byte, mỗi byte biểu diễn một kí tự theo thứ tự từ trái sang phải.

Ví dụ, dãy ba byte 01010100 01001001 01001110 biểu diễn xâu kí tự "TIN".

• *Các dạng khác*

Hiện nay, việc tìm cách biểu diễn hiệu quả các dạng thông tin loại phi số như âm thanh, hình ảnh,... rất được quan tâm vì các thông tin loại này ngày càng phổ biến. Để xử lí âm thanh, hình ảnh, ta cũng phải mã hoá chúng thành các dãy bit. Các thành tựu trong lĩnh vực này đã và đang nâng cao chất lượng cuộc sống. Chẳng hạn, hai người ở xa nhau vẫn có thể trò chuyện, thậm chí có thể nhìn thấy hình ảnh của nhau.

Nguyên lí mã hoá nhị phân

Thông tin có nhiều dạng khác nhau như số, văn bản, hình ảnh, âm thanh,... Khi đưa vào máy tính, chúng đều được biến đổi thành dạng chung – dãy bit. Dãy bit đó là mã nhị phân của thông tin mà nó biểu diễn.