MODULE 1. THÔNG TIN VÀ XỬ LÝ THÔNG TIN

1.1. Thông tin

Ngày nay, thuật ngữ "thông tin" (information) được sử dụng khá phổ biến. Người ta có nhu cầu đọc báo, xem truyền hình, giao tiếp với người khác để có thông tin. Thông tin chính là tất cả những gì mang lại hiểu biết cho con người.

Cần đặt thông tin trong mục đích hoạt động. Khi tiếp nhận thông tin, người ta phải "xử lý" để có những quyết định. Một công ty phải luôn luôn tìm hiểu thông tin về thị trường để có chiến lược kinh doanh thích hợp. Một người điều khiển xe máy phải luôn nhìn đường và các đối tượng tham gia giao thông khác để lái tới đích và không bị tai nạn. Thông tin làm tăng thêm hiểu biết của con người, là nguồn gốc của nhận thức và là cơ sở của quyết định.

Thông tin được chuyển tải qua các môi trường vật lý khác nhau như ánh sáng, sóng âm, sóng điện từ. Thông tin được ghi trên các phương tiện hữu hình như các văn bản trên giấy, băng ghi âm hay phim ảnh... Về nguyên tắc, bất kỳ cấu trúc vật chất nào hoặc bất kỳ dòng năng lượng nào cũng có thể mang thông tin. Các vật có thể mang được thông tin được gọi là *giá mang tin* (support*)*. Hình thức vật lý của thông tin được gọi là *tín hiệu* (signal). Thông tin và tín hiệu có một độ độc lập tương đối. Có thể chuyển tải một nội dung thông tin như nhau bằng những tín hiệu khác nhau. Trên sân cỏ, động tác phất cờ của trọng tài biên (hình ảnh), tiếng còi trọng tài chính (âm thanh) có thể cùng mang thông tin báo lỗi. Ngược lại một tín hiệu như nhau trong những hoàn cảnh khác nhau lại có thể thể hiện những thông tin khác nhau. Người nông dân ở đồng bằng Sông Hồng mời khách uống rượu trước để tỏ lòng thành nhưng ở miền Tây Nam bộ thì phải làm ngược lại - chủ phải uống trước.

Thông tin có thể được phát sinh và được lưu lại trong một giá mang tin nào đấy. Thông tin có thể được truyền từ một giá mang này sang một giá mang khác. Như vậy thông tin có thể được nhân bản và khi nhân bản ý nghĩa của thông tin không hề suy giảm.

Dữ liệu (data) là hình thức thể hiện của thông tin trong mục đích lưu trữ và xử lý nhất định. Thuật ngữ "dữ liệu" chúng ta dùng hiện nay có nguồn gốc từ chữ Hán- Việt với ý nghĩa là "cái đã cho". Từ tương ứng trong tiếng Anh (data) là số nhiều của từ datum trong tiếng Latin, tiếng Pháp (donneés), tiếng Nga (данных) cũng đều mang nghĩa là "cái đã cho". Về mặt lịch sử, khái niệm dữ liệu xuất hiện cùng với việc xử lý thông tin bằng máy tính. Vì thế trong nhiều tài liệu người ta định nghĩa dữ liệu là đối tượng xử lý của máy tính. Thông tin luôn mang một ý nghĩa xác định nhưng hình thức thể hiện của thông tin rõ ràng mang tính quy ước. Chẳng hạn kí hiệu "V" trong hệ đếm La mã mang ý nghĩa là 5 đơn vị nhưng trong hệ thống chữ la-tinh nó mang ý nghĩa là chữ cái V. Trong máy tính điện tử (MTĐT), nhóm 8 chữ số 01000001, nếu là số sẽ thể hiện số 65, còn nếu là chữ sẽ là chữ "A".

Tri thức (knowledge) có ý nghĩa khái quát hơn thông tin. Những nhận thức thu nhận được từ nhiều thông tin trong một lĩnh vực cụ thể nào đó, có tính hướng mục đích mới trở thành tri thức. Như vậy tri thức là mục đích của nhận thức trên cơ sở tiếp nhận

thông tin. Quá trình xử lý thông tin chính là quá trình nhận thức để có tri thức.

1.2. Mã hoá thông tin

Có nhiều cách phân loại thông tin. Chúng ta quan tâm đến cách phân loại dựa vào các đặc tính liên tục hay rời rạc của tín hiệu vật lý. Tương ứng, thông tin được chia thành thông tin liên tục và thông tin rời rạc.

Thông tin liên tục đặc trưng cho các đại lượng mà số lượng các giá trị có thể tiếp nhận được là vô hạn như độ dài dịch chuyển cơ học, điện áp... Thông tin rời rạc đặc trưng cho các đại lượng mà số lượng các giá trị có thể kể ra được như số nhà trong dãy phố, số trang của một quyển sách, tên người trong một lớp.

Thông tin rời rạc có thể biểu diễn thông qua các bộ kí hiệu (các chữ số, các chữ cái...) mà ta gọi là bảng chữ. Giả sử, ta có tập đối tượng X cần biểu diễn. Để làm điều này, ta chọn một tập hữu hạn A các kí hiệu làm bảng chữ mà mỗi kí hiệu là một chữ. Chúng ta sẽ gọi mỗi dãy hữu hạn các chữ là một từ trên A. Ví dụ nếu A là tập các chữ số thì mỗi từ chính là một số (cho bằng một dãy số). Mã hoá các thông tin rời rạc của một tập X trên một bảng chữ A chính là cách gán cho mỗi phần tử $x \in X$ một từ y trên A. Phép gán mã phải đảm bảo tính chất: mã của hai đối tượng khác nhau phải khác nhau. Tính chất này đảm bảo khi biết mã có thể tìm được đối tượng tương ứng. Quá trình gán mã được gọi là phép lập mã. Quá trình ngược được gọi là phép giải mã. Ví dụ, nếu X là tập các thí sinh, chọn A là tập các chữ số thì mã của một thí sinh có thể lấy là số báo danh của thí sinh đó. Số báo danh phải cho phép chỉ định duy nhất một thí sinh.

Như đã biết dữ liệu là hình thức biểu diễn thông tin với mục đích xử lý thông tin. Vậy mã hoá chính là con đường chuyển từ thông tin thành dữ liệu. Sau này ta sẽ thấy các thông tin dưới dạng số, văn bản, âm thanh, hình ảnh đều phải chuyển dưới dạng mã phù hợp để máy tính có thể làm việc được.

1.3. Mã hoá nhị phân và đơn vị đo thông tin

Từ lâu người ta đã biết dùng mã Moorse trong truyền tin. Với mã Moorse, mỗi chữ được thể hiện bằng một dãy các kí hiệu chấm và vạch. Khi truyền tin, các điện tín viên nhấn lên cần manip để đóng mạch điện. Để truyền đi một dấu chấm người ta nhấn cần manip rồi nhả ngay. Còn để truyền một vạch người ta nhấn, giữ một chút rồi mới nhả. Tại máy nhận mỗi khi mạch điện được đóng, đầu in áp xuống băng giấy chạy. Mạch điện đóng lâu sẽ tạo ra vạch, đóng nhanh sẽ tạo ra chấm và được tái hiện trên băng giấy.

Mã hoá trên bảng chữ hai kí hiệu được gọi là mã hoá nhị phân. Như vậy mã Moorse là một loại mã nhị phân.

Trong tin học, mã hoá nhị phân được sử dụng rất rộng rãi. Có nhiều lý do trong đó có lý do là máy tính điện tử xây dựng từ các linh kiện vật lý có hai trạng thái như các mạch đóng hoặc ngắt dòng điện. Bảng chữ nhị phân được sử dụng trong tin học chỉ gồm 2 "chữ" là chữ số 0 và chữ số 1. Chính các chữ số này cũng gọi là chữ số nhị phân (binary digit).

Trong một tập hữu hạn đối tượng, để mã hoá nhị phân, cần gán cho mỗi đối tượng một từ nhị phân (mã nhị phân). Ví dụ đối với tập 8 đối tượng ta có thể gán cho mỗi đối tượng một mã khác nhau trong tập mã 3 chữ số nhị phân sau: 000, 001, 010, 011, 100,

101, 110, 111. Một cách tổng quát, nếu dùng các mã k bít sẽ mã hoá được tập đối tượng có tới 2^K đối tượng. Ngược lại, bất cứ một tập n đối tượng sẽ chỉ cần dùng không quá $[\log_2 n] + 1$ chữ số nhị phân để tạo ra các mã đủ phân biệt n đối tượng.

Như vậy, trong mã hoá nhị phân, mỗi một chữ số nhị phân mang một lượng tin nào đó về đối tượng và được xem là một đơn vị thông tin. Đơn vị thông tin đó được gọi là bit do viết tắt từ chính cụm từ " *BI*nary digi*T*". Ta cũng gọi các chữ số 0 hay 1 là một bit. Thông thường để chỉ các lượng tin lớn, người ta không dùng bít mà dùng một số đơn vị bội của bit sau đây:

Bảng 1.1. Các đơn vị đo thông tin

Tên gọi	Viết tắt	Giá trị
Byte	B	8 bit
KiloByte	KB	2^{10} byte (1024 byte)
MegaByte	MB	2^{10} KB
GigaByte	GB	2^{10} MB
TeraByte	TB	2^{10} GB

1.4. Xử lý thông tin

Xử lý thông tin là tìm ra những dạng thể hiện mới của thông tin phù hợp với mục đích sử dụng. Ví dụ, khi cho phương trình $x^2 + bx + c = 0$ ta cần giải (xử lý) để tìm ra hai nghiệm x_1 và x_2. Về mặt thông tin, việc biết b và c hoàn toàn tương đương với biết x_1 và x_2. Tuy nhiên trong mục đích sử dụng thì việc biết x_1 và x_2 khác hẳn với biết b và c. Như vậy xử lý thông tin không làm tăng lượng tin mà chỉ hướng hiểu biết vào những khía cạnh có lợi trong hoạt động thực tiễn. Mục đích của xử lý thông tin là tri thức.

1.5. Xử lý thông tin tự động bằng máy tính điện tử

Quá trình xử lí thông tin trên máy tính điện tử cũng có những bước tương tự như tính toán thủ công.

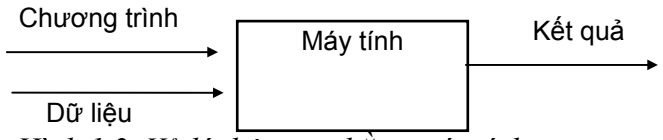
Để mô tả cách thức xử lý, dữ liệu, giữ các kết quả tính toán, con người cần phải sử dụng một số phương tiện ghi nhớ nhất định như giấy, bảng và chính trí nhớ của mình. MTĐT cũng cần có phương tiện nhớ dữ liệu, kết quả và cách xử lý gọi là bộ nhớ.

Con người cần sử dụng một số công cụ nào đó như bàn tính, hay chính bộ não để thực hiện các phép toán. MTĐT cũng sử dụng một số mạch tính toán có khả năng xử lý dữ liệu. Đó chính là bộ số học và logic.

Để xử lý một công việc phức tạp, người ta cần thực hiện nhiều phép xử lý nhỏ theo một trình tự nhất định. Với hiểu biết của mình, tuỳ theo những điều kiện cụ thể, con người tự xác định các phép tính cần thiết và trình tự thực hiện các phép tính. Ví dụ khi giải một phương trình bậc 2, người giải chỉ có thể quyết định giải tiếp để tìm hai nghiệm thực sau khi tính và thấy biệt thức $\Delta \geq 0$. MTĐT thì không thể chủ động như thế. Nó

không thể tự quyết định được, khi nào thì phải làm gì, cộng hay trừ, nhân hay chia, các dữ liệu tham gia xử lí sẽ lấy ở đâu... Để làm được điều đó, người ta phải lập một kịch bản xử lý có đầy đủ mọi tình huống dưới dạng các mệnh lệnh để hướng dẫn MTĐT xử lý công việc theo đúng yêu cầu mong muốn. Tập hợp các mệnh lệnh như vậy được con người soạn thảo bằng một ngôn ngữ mà máy "hiểu" được gọi là chương trình (prrogam). Máy tính cần có phương tiện để lưu chương trình đưa vào và cần có một thiết bị có đảm bảo khả năng tự điều khiển theo chương trình.

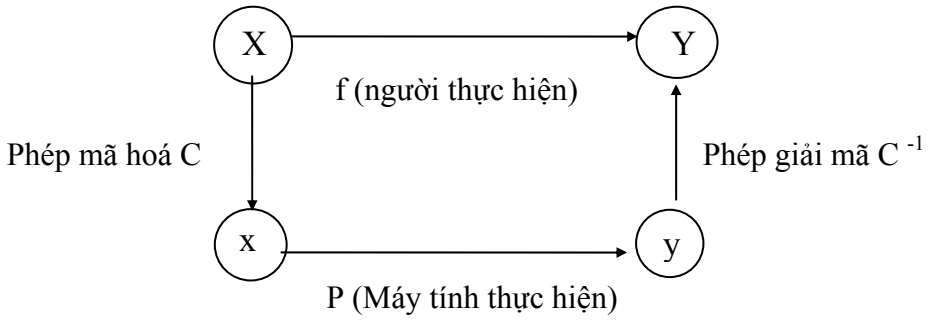
Ta có thể hình dung quá trình xử lí thông tin trên máy tính số bằng sơ đồ ở hình dưới đây:



Hình 1.2. Xử lý thông tin bằng máy tính

Cụ thể hơn, giả sử ta cần xử lí các thông tin X. Bằng một công cụ tính toán nào đó, con người có thể thực hiện tính toán theo một quy trình f để thu nhận được kết quả Y. Với MTĐT, quá trình xử lí đó được tiến hành như sau: mã hóa X nhờ phép mã hoá C để thu được dữ liệu ban đầu x (sau này ta sẽ thấy là máy tính chỉ xử lí trực tiếp với dữ liệu ở mã nhị phân gồm toàn các chữ số 0 và 1). Thay cho quy trình xử lý f, người ta phải lập một chương trình P nạp vào trong máy và giao cho máy tính thực hiện. Sau khi chương trình P thực hiện xong ta thu được kết qủa y (trong dạng nhị phân). Nhờ phép giải mã C^{-1} ta thu được kết qủa phải tìm Y dưới dạng mà con người có thể sử dụng trực tiếp.

Tương ứng giữa hai cách xử lí có thể mô tả như hình dưới đây:



Hình 1.3. Sơ đồ xử lý thông tin bằng máy tính

1.6. Tin học và Công nghệ Thông tin

Bản thân thuật ngữ “tin học” dùng ở Việt Nam có nguồn gốc từ từ “Informatique” trong tiếng Pháp (Xem bài đọc thêm). Informatique được Viện Hàn lâm Khoa học Pháp định nghĩa như sau:

Tin học là ngành khoa học nghiên cứu các phương pháp và qúa trình xử lý thông tin một cách tự động dựa trên các phương tiện kỹ thuật mà hiện nay phương tiện đó là MTĐT.

Như vậy, trong định nghĩa này ta thấy có hai khía cạnh:

- Phần cứng (hardware) là toàn bộ các thiết bị vật lí của MTĐT. Nâng cao tốc độ xử lý, tăng khả năng lưu trữ, tăng độ tin cậy, giảm năng lượng sử dụng, tăng khả năng ghép nối... là những mục tiêu mà công nghệ phần cứng hướng tới.
- Phần mềm (software) là phương pháp xử lý thông tin bao gồm các chương trình có chức năng điều khiển, khai thác phần cứng và để thực hiện các yêu cầu xử lý thông tin. Phần mềm còn nằm ở phương pháp tổ chức dữ liệu tương ứng với chương trình xử lý thông tin. Tìm ra các phương pháp xử lý thông tin có hiệu quả, tổ chức dữ liệu tốt và lập trình thể hiện các phương pháp xử lý đó là vấn đề của phần mềm.

Trước năm 1975, với một nội dung khá thô sơ (chủ yếu là nguyên lý máy và lập trình) ở miền Bắc thường dùng thuật ngữ "Máy tính điện tử" còn ở miền Nam dùng thuật ngữ "Điện toán" với ý nghĩa của tin học. Tất nhiên các thuật ngữ trên đều không phản ánh đầy đủ nội dung của Tin học. Ngay ở Mỹ cho đến nay người ta vẫn dùng thuật ngữ "khoa học máy tính" (Computer Science), "xử lý dữ liệu" (Data Processing), "xử lý thông tin" (Information Processing), "tính toán bằng máy tính" (Computing) để chỉ những mặt nào đó trong tin học. Cũng nên biết rằng tên của Hội Tin học Việt Nam trong tiếng Anh được lấy là VAIP có nguồn gốc từ "Vietnam Association for Information Processing"

Đã từ lâu, nhiều chuyên gia muốn có một tên gọi mới cho ngành khoa học này. Năm 1962, một giáo sư người Pháp tên là Philippe Dreyfus đã đề nghị thuật ngữ *informatique* trên cơ sở hai từ "information" (thông tin) và "automatique" (tự động hoá). Thuật ngữ này được Viện Hàn lâm khoa học Pháp chấp nhận chính thức và công bố ngày 6/4/1966 kèm theo giải thích với nội dung như định nghĩa tin học nêu trên. Trong các hội thảo và các ấn phẩm khoa học, thuật ngữ này được Anh hoá thành từ Informatics (chính tiếng Anh không có từ này). Thuật ngữ này được chấp nhận rộng rãi ở châu Âu nhưng ít được dùng ở Bắc Mỹ.

Cuối những năm 70, một nhóm các nhà khoa học Việt kiều tại Pháp đã dùng thuật ngữ "Tin học" với ý nghĩa của từ "Informatique" và đã sử dụng trong một số hội thảo tại Hà Nội. Từ đó thuật ngữ "Tin học" được chính thức sử dụng tại Việt Nam.

Hiện nay ngay cả trên thế giới cũng có nhiều quan niệm khác nhau về một định nghĩa cho tin học. Sự khác nhau thực chất chỉ ở phạm vi các lĩnh vực được coi là tin học. Ngày nay tin học xâm nhập vào mọi lĩnh vực nên ở một số nơi ranh giới giữa tin học và một số ngành khác không còn rõ nét nữa. Ví dụ viễn thông (telecommunication) ngày nay đã chuyển dịch từ công nghệ tương tự (analog) sang công nghệ số (digital). Phần truyền dẫn ở những môi trường truyền thống còn là tương tự nhưng phần quản lý, chuyển mạch, xử lý dịch vụ...đều do máy tính đảm nhận. Tự động hoá ngày nay cũng thay đổi rất nhiều với những xử lý thông minh qua máy tính trước khi truyền tín hiệu điều khiển đến các cơ cấu chấp hành.

Trong thời gian vừa qua nhiều nhà khoa học đề nghị sử dụng thuật ngữ "Công nghệ Thông tin" (Information Technology) với một nội dung đầy đủ hơn, bao hàm được những lĩnh vực, những nền tảng chủ yếu của khoa học và công nghệ xử lý thông tin dựa trên máy tính. Khi nói đến yếu tố công nghệ, người ta muốn nhấn mạnh đến tính quá

trình, tính tổ chức và phương pháp xử lý thông tin hướng tới ứng dụng. Định nghĩa Công nghệ Thông tin đã được nhóm chuyên gia Việt Nam đứng đầu là Giáo sư Phan Đình Diệu (hiện công tác tại Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội) nghiên cứu, tham khảo từ các định nghĩa của chuyên gia trên thế giới và đã được đưa vào Nghị quyết 49/CP của Chính phủ về phát triển Công nghệ Thông tin của Việt Nam từ 1996 như sau:

"Công nghệ Thông tin là tập hợp các phương pháp khoa học, các phương tiện và công cụ kỹ thuật hiện đại - chủ yếu là kỹ thuật máy tính và viễn thông - nhằm tổ chức và khai thác và sử dụng có hiệu quả nguồn tài nguyên thông tin rất phong phú và tiềm tàng trong mọi lĩnh vực hoạt động của con người và xã hội...Công nghệ thông tin được phát triển trên nền tảng phát triển của các công nghệ Tin học-Điện tử- Viễn thông và Tự động hoá".

Như vậy Công nghệ Thông tin mang một ý nghĩa rộng rãi hơn, nó vừa là khoa học, vừa là công nghệ, vừa là kỹ thuật, bao trùm cả tin học, viễn thông và tự động hoá.

Lưu ý rằng ở nhiều nơi người ta không xem viễn thông là một bộ phận của CNTT. Vì thế thay cho IT (Information Technology) người ta thường dùng ICT (Information - Communication Technology).

Ngày 29 tháng 6 năm 2006 Quốc hội nước Cộng hoà Xã hội Chủ nghĩa Việt Nam đã ban hành Luật Công nghệ Thông tin. Theo đó CNTT được định nghĩa như sau:

Công nghệ thông tin là tập hợp các các phương pháp khoa học, công nghệ và công cụ kỹ thuật hiện đại để sản xuất, truyền đưa, thu thập, xử lý, lưu trữ và trao đổi thông tin số.

Ở đây, thông tin số là thông tin được tạo lập bằng phương pháp dùng tín hiệu số.

Câu hỏi và bài tập

1. Hãy làm rõ mối liên hệ giữa các khái niệm thông tin, tin hiệu, dữ liệu ?

2. Tìm một ví dụ minh hoạ có thông tin nghĩa là giảm độ bất đinh.

3. Một lớp có 48 sinh viên trong đó có 36 nam và 12 nữ. Trong một cuộc thi học sinh giỏi tin học của trường một sinh viên của lớp được giải nhất. Người ta muốn biết người đó là ai. Sau đó người ta được thông báo thêm, người đoạt giải cũng đã từng nhận giải nhì trong một cuộc thi cắm hoa của nữ sinh tổ chức nhân ngày 8/3. Tính lượng tin nhận được trong thông báo trên.

4. Đơn vị đo tin là bít. Nhưng bít chính lại là chữ viết tắt của cụm từ chữ số nhị phân "Binary Digit". Hãy lý giải mối liên hệ giữa hai điều này.

5. Tại sao nói xử lý thông tin không làm tăng lượng tin

6. Hãy nêu vai trò của thông tin trong cuộc sống .