VŨ NGỌC PHÀN

LÝ THUYẾT THÔNG TIN VÀ MÃ HÓA

NHÀ XUẤT BẢN BƯU ĐIỆN

LÝ THUYẾT
THÔNG TIN
VÀ
MÃ HÓA

VŨ NGỌC PHÀN

LÝ THUYẾT THÔNG TIN VÀ MÃ HÓA

NHÀ XUẤT BẢN BƯU ĐIỆN
Hà Nội, tháng 10 - 2006

LỜI NÓI ĐẦU

Ngày nay, các mạng máy tính, mạng điện thoại số hữu tuyến và vô tuyến, mạng truyền hình cáp... đang có xu thế hội tụ thành mạng chung, đa dịch vụ, có khả năng chuyển tải thông tin tích hợp với tốc độ lên đến hàng trăm Mbit/s. Nhìn từ góc độ khoa học, các hệ thống thông tin ngày nay là các hệ thống có độ phức tạp cấu trúc rất lớn và chịu tác động của nhiều loại nhiễu khác nhau. Để các hệ thống đó có thể làm việc ổn định và tin cậy, đáp ứng được yêu cầu của người sử dụng trong việc lưu giữ, chuyển tải và chia sẻ thông tin, không thể không dựa vào kiến thức về lý thuyết thông tin và mã hóa. Có thể nói, không có lý thuyết thông tin và mã hóa thì không thể giải quyết được vấn đề truyền tin chính xác trên các đường truyền dài hàng ngàn km.

*Nhằm đáp ứng nhu cầu tìm hiểu về lý thuyết thông tin và mã hóa, Nhà xuất bản Bưu điện xuất bản cuốn sách "**Lý thuyết thông tin và mã hóa**" của tác giả Vũ Ngọc Phàn giới thiệu đến bạn đọc. Lý thuyết thông tin và mã hóa là công cụ hữu hiệu để giải quyết những vấn đề công nghệ thông tin đang đặt ra như nâng cao tốc độ truyền dẫn theo thời gian thực, nén dữ liệu, bảo mật dữ liệu,... Bố cục cuốn sách gồm 8 chương. Từ chương 1 đến chương 7 trình bày cụ thể các các vấn đề về lý thuyết thông tin rời rạc như: lượng tin và En-tro-py, nguồn rời rạc và kênh rời rạc, mã hóa nguồn và mã hóa kênh, các phương pháp mã hóa và giải mã, mật mã. Đặc biệt chương 8 giới thiệu về lý thuyết thông tin các hệ liên tục được xem như là sự mở rộng của lý thuyết thông tin các hệ rời rạc. Trong khi trình bày những nội dung lý thuyết, cuốn sách đưa ra những ví dụ khá trực quan, giúp người đọc dễ theo dõi. Ngoài ra cuối cuốn sách có phần Phụ lục giới thiệu một số chương trình mô phỏng các thuật toán mã hóa viết trên MatLab giúp cho bạn đọc hiểu một cách trực quan về các mã đã nghiên cứu.*

Cuốn sách là tài liệu tham khảo rất hữu ích cho các chuyên gia, kỹ thuật viên, cũng như cán bộ giảng dạy và đặc biệt học viên ngành viễn thông muốn tìm hiểu những kiến thức cơ bản và nghiên cứu sâu về lý thuyết thông tin và mã hóa.

Nhà xuất bản xin trân trọng giới thiệu đến bạn đọc và rất mong nhận được ý kiến góp ý của quí vị. Mọi ý kiến góp ý xin gửi về Nhà xuất bản Bưu điện - 18 Nguyễn Du, Hà Nội.

Trân trọng cảm ơn./.

Hà Nội, tháng 10 năm 2006

NHÀ XUẤT BẢN BƯU ĐIỆN

Chương 1
MỞ ĐẦU

1.1. KHÁI QUÁT

Sau chiến tranh thế giới thứ hai, có ba lý thuyết ra đời và đã cùng ảnh hưởng rất mạnh mẽ đến sự phát triển khoa học và công nghệ. Đó là *lý thuyết hệ thống* (System Theory), *lý thuyết điều khiển* (Control Theory) và *lý thuyết thông tin* (Information Theory). Trong cuốn sách "Five More Golden Rules" xuất bản năm 2000 tại Mỹ, không phải không có lý khi John L. Casti xếp ba lý thuyết trên cùng với *lý thuyết dây* (Knot Theory) và *giải tích hàm* (Functional Analysis) là năm trong số những lý thuyết lớn của thế kỷ XX. Lúc đầu lý thuyết thông tin được phát triển chủ yếu để phục vụ cho kỹ thuật truyền tin. Nhưng ngay sau đó người ta nhận ra rằng, những ứng dụng của lý thuyết thông tin không chỉ dừng lại ở kỹ thuật truyền tin mà ở cả các lĩnh vực khác như sinh y, kinh tế, ngôn ngữ, âm nhạc, nghệ thuật, hội họa... Lý thuyết thông tin có vai trò hết sức quan trọng trong việc nghiên cứu các *hệ thống tự tổ chức, tự điều chỉnh và tự ổn định*. Lý thuyết thông tin, cùng với lý thuyết hệ thống và lý thuyết điều khiển, đã đặt nền móng cho quá trình chuyển đổi từ cách tiếp cận dựa trên *quan hệ hình thức-nội dung* (Form-Content-Relation) sang cách tiếp cận dựa trên *quan hệ cấu trúc-chức năng* (Structure-Function-Relation). Cách tiếp cận sau đã góp phần tạo ra những thành tựu vô cùng to lớn của nhân loại suốt nửa thế kỷ qua trong việc phân tích và thiết kế hệ thống, đặc biệt là các hệ thống lớn (Large-scale Systems). Vào năm 1948 khi Shannon công bố lý thuyết thông tin của mình qua cuốn sách nổi tiếng "A Mathematical Theory of Communication", đường cáp lớn nhất thế giới lúc bấy giờ mới chỉ cho phép thực hiện đồng thời 1800 cuộc thoại. Dưới tác động của lý thuyết thông tin, 20 năm sau số cuộc thoại đồng thời trên đường truyền đã là 230.000. Năm 2001 đường cáp quang của hãng WaveStar™ đạt con số 64 triệu cuộc thoại cùng một lúc.

Ngày nay người ta đã thừa nhận, bên cạnh quá trình vận động vật chất là quá trình vận động thông tin không kém phần quan trọng. Như đã biết, các nhiễm sắc thể và chất lòng trắng trứng tạo thành bộ nhớ sinh học. Lý thuyết thông tin là một công cụ hữu hiệu để hiểu bản chất của những bộ nhớ sinh học này. Dưới góc độ của lý thuyết thông tin ta thấy hoạt động của tế bào không khác gì hoạt động của một nhà máy, trong đó nhân tế bào là ban giám đốc điều hành toàn bộ quá trình sản xuất, các nhiễm sắc thể là hồ sơ về qui trình công nghệ và kế hoạch sản xuất, tế bào chất là nguyên vật liệu, các en-zim là đội ngũ kỹ sư và công nhân. Nếu vì một lý do nào đó, các thông tin chứa trong nhiễm sắc thể bị sai lệch thì hoạt động của nhà máy tế bào sẽ không đúng kế hoạch và qui trình công nghệ, sản phẩm tạo ra không đảm bảo tiêu chuẩn về chất lượng và số lượng, bệnh tật xuất hiện. Như đã biết, nhiễm sắc thể là một chuỗi rất dài các phân tử chứa thông tin. Trong quá trình phân chia tế bào, các thông tin phải được bảo toàn và chia đều cho cả hai nửa (hai tế bào mới). Các thông tin tồn tại trong nhiễm sắc thể dưới dạng mã được tạo nên từ bốn phần tử cơ sở (lý thuyết thông tin và mã hóa gọi là mã hiệu). Đó là Adenin (ký hiệu là A), Thymin (ký hiệu là T), Guanin (ký hiệu là G) và Cytosin (ký hiệu là C). Theo luật số mũ của sự lựa chọn (Exponential Law of Choice), nếu một nhiễm sắc thể gồm một chuỗi L phân tử nu-clein thì sẽ có 4^L cấu trúc khác nhau. Với L = 100 ta có $4^{100} \approx 1{,}6 \times 10^{60}$ cấu trúc nhiễm sắc thể khác nhau. Trên thực tế, một nhiễm sắc thể có thể bao gồm hàng ngàn phân tử nu-clein. Điều này giải thích đầy đủ tính đa dạng phong phú của thế giới sinh vật.

Trong lĩnh vực thần kinh học, người ta thấy rằng một tế bào thần kinh có thể được mô tả hoàn toàn bởi một ô-tô-mat hữu hạn và hệ thần kinh có thể xem như một mạng các ô-tô-mat hữu hạn. Một nơ-ron bao gồm phần thân và các sy-nap. Các sy-nap bắt đầu từ thân của một nơ-ron này và nối tới một nơ-ron khác. Cứ như vậy, nhiều nơ-ron kết nối với nhau thành một mạng nơ-ron. Mỗi nơ-ron chỉ có thể có hai trạng thái, kích hoạt (active) hoặc không kích hoạt (passive), tương đương hệ nhị phân trong các máy tính và các thiết bị điện tử số thông dụng. Quá trình lưu trữ thông tin trong bộ não được hình thành nhờ hoạt động của

các sy-nap. Đến nay người ta biết rằng, mỗi nơ-ron chứa nhiều hơn 1 bit. Theo Schaefer thì mỗi nơ-ron có dung lượng khoảng 10^2 bit. Như vậy một bộ não trung bình với khoảng 10^{10} nơ-ron có thể lưu giữ 1000 Gbit thông tin.

Lý thuyết thông tin cũng đã trả lời câu hỏi, một hệ thống tự thích nghi hay một hệ thống tự học có thể có bậc bằng bao nhiêu trong môi trường của nó. Dưới góc nhìn của lý thuyết thông tin, người ta có thể giải thích vấn đề này một cách khá thuyết phục. Trước hết chúng ta tạm sử dụng khái niệm *độ dư thừa* và *lượng tin* sẽ được làm sáng tỏ ở các phần sau. Ta đặt:

$$r = 1 - \frac{I}{I_{max}} \tag{1.1-1}$$

Trong biểu thức (1.1-1), r là độ dư thừa, I là lượng tin thực tế của nguồn tin và I_{max} là lượng tin cực đại có thể có. Đối với các hệ thống tự thích nghi hoặc các hệ thống tự học, ta luôn có:

$$\frac{\partial r}{\partial t} > 0 \text{ hay } -\frac{I_{max}\left(\frac{\partial I}{\partial t}\right) - I\left(\frac{\partial I_{max}}{\partial t}\right)}{I_{max}^2} > 0 \tag{1.1-2}$$

Từ (1.1-2) suy ra:

$$I\frac{\partial I_{max}}{\partial t} > I_{max}\frac{\partial I}{\partial t} \tag{1.1-3}$$

Biểu thức (1.1-3) nói lên một cách tổng quát rằng, giá trị của trạng thái thông tin và sự biến thiên của nó liên quan chặt chẽ với nhau. Từ đây ta có thể rút ra kết luận rằng, nếu lượng tin cực đại của một hệ thống tự thích nghi hay hệ thống tự học không thay đổi thì lượng tin thực tế sẽ giảm. Nghĩa là:

$$I_{max} = \text{const} \rightarrow \frac{\partial I}{\partial t} < 0 \tag{1.1-4}$$

Ngược lại, nếu lượng tin thực tế của hệ thống tự thích nghi hay hệ thống tự học không thay đổi thì lượng tin cực đại khả dĩ của nó sẽ tăng. Nghĩa là:

$$I = \text{const} \rightarrow \frac{\partial I_{max}}{\partial t} > 0 \qquad (1.1\text{-}5)$$

Các biểu thức rất đơn giản vừa trình bày trên cho phép giải quyết một vấn đề đã tồn tại rất lâu trong lịch sử nhân loại nhưng chưa được trả lời một cách thỏa đáng, đó là vì sao các cấu trúc sinh học có khả năng thích ứng với môi trường và tự phát triển. Các cấu trúc sinh học khác các cấu trúc vô cơ ở chỗ nó có khả năng chống lại sự tăng en-tro-py. Nhiều nghiên cứu lý thuyết thông tin trên tinh tinh và người đã đi đến kết luận, khả năng nhận biết sự vật là kết quả của một quá trình học rất đa dạng. Các thí nghiệm đã cho thấy rằng, một người mù bẩm sinh có đầu óc tương đối thông minh, sau khi trưởng thành được phẫu thuật mắt và nhìn được, nhưng họ phải cần nhiều tháng để có thể nhận ra các đồ vật hết sức đơn giản, trong khi người bình thường làm được việc này ngay từ cái nhìn đầu tiên. Đối với tinh tinh và trẻ nhỏ, những nghiên cứu đã cho thấy rằng, khi quay một tam giác đi một góc 90 độ thì trẻ nhỏ và tinh tinh cũng phải nghiêng đầu đi 90 độ mới nhận ra được. Tóm lại, không có quá trình học thì không có sự giảm en-tro-py thông tin.

Trong đời sống hàng ngày, con người đã tiếp xúc với rất nhiều hiện tượng mà họ không dự đoán trước được, hoặc chỉ đoán trước được một cách mơ hồ, không cụ thể. Một cơn lốc xoáy ập đến bất ngờ, một trận mưa lụt chưa từng thấy trong lịch sử hàng trăm năm, một con tàu bỗng nhiên mất tích ngoài biển khơi. Các nhà kỹ thuật thường gặp những hiện tượng khó chịu mà họ gọi là tạp âm (noise) hoặc sự thay đổi bất thường (fluctuation). Tất cả những hiện tượng trên ít nhiều liên quan đến khái niệm en-tro-py ta vừa nhắc đến và sẽ được làm rõ dần trong những phần sau của cuốn sách. Xét ở một góc độ nào đó, mục đích cuối cùng của lý thuyết thông tin chính là giúp con người trong việc làm giảm en-tro-py thông tin.

Khái niệm en-tro-py dùng trong lý thuyết thông tin mà chúng ta vừa gọi là en-tro-py thông tin, có nguồn gốc từ khái niệm en-tro-py trong nhiệt động học. Định luật thứ 2 của nhiệt động học chỉ ra rằng, nhiệt chỉ có thể truyền từ nơi có nhiệt độ cao hơn đến nơi có nhiệt độ thấp hơn và không thể ngược lại. Trong tự nhiên, người ta nhận thấy có hai loại quá

trình: *quá trình đảo ngược được* (Reversible Process) và *quá trình không đảo ngược được* (Non-Reversible Process). Quá trình truyền nhiệt là một quá trình không đảo ngược được. Chúng ta sẽ phác họa rất sơ lược khái niệm en-tro-py của nhiệt động học. Gọi S là trạng thái nhiệt của một hệ thống và Q là nhiệt lượng, ta có:

$$\oint \frac{\partial Q}{T} = 0 \qquad (1.1\text{-}6)$$

Lấy tích phân từ trạng thái S_1 đến trạng thái S_2, ta có:

$$\int_{S_1}^{S_2} \frac{\partial Q}{T} = S_2 - S_1 = \Delta S \qquad (1.1\text{-}7)$$

ΔS được gọi là en-tro-py. Năm 1829, Các-nô, một nhà khoa học người Pháp, đã chứng minh rằng, *trong một hệ thống kín, $\Delta S \geq 0$.* Định luật này về sau được mở rộng thành: *trong một hệ thống kín, en-tro-py không tự giảm* theo nghĩa chung nhất. Như vậy nghĩa là, một hệ thống không tiếp xúc với một hệ nào khác (không có quan hệ trao đổi với môi trường của nó), luôn luôn có xu hướng trở về trạng thái xác suất đồng đều, trạng thái có en-tro-py cực đại. Trạng thái xác suất đồng đều là trạng thái hoàn toàn hỗn loạn. Theo định luật này, các hệ thống kín cuối cùng sẽ rơi vào trạng thái hỗn loạn và hủy diệt. Trong kỹ thuật đó là sự hao mòn, trong sinh học đó là sự già cỗi, trong hóa học đó là sự phân hủy, trong xã hội đó là sự phân hóa, trong lịch sử đó là sự suy tàn. Điều này cũng đúng với các hệ thống thông tin mà ở đó sự tăng en-tro-py thông tin sẽ dẫn tới sự bất định hoàn toàn.

1.2. NHỮNG ĐỊNH HƯỚNG CHÍNH CỦA LÝ THUYẾT THÔNG TIN VÀ MÃ HÓA

Lý thuyết thông tin đề cập đến tất cả các hình thái vận động của thông tin như: quá trình hình thành thông tin của một nguồn tin, quá trình thu nhận thông tin, quá trình biến đổi thông tin, quá trình truyền dẫn thông tin, quá trình xử lý thông tin và quá trình lưu trữ thông tin. Những quá trình này có thể diễn ra một cách tường minh như các quá trình thông tin trong kỹ thuật viễn thông, nhưng cũng có khi không tường minh như