Tìm hiểu Lý thuyết Thông tin của Shenon

PGS.TS.Đoàn Phan Tân

Một trong những nét đặc trưng của công nghệ thế kỷ XX là sự phát triển và bùng nổ các phương tiện truyền thông mới. Cùng với sự phát triển của các phương tiện truyền và xử lý thông tin, một ngành khoa học lý thuyết hình thành, phát triển và trở thành đối tượng của nhiều nghiên cứu sâu sắc, đó là Lý thuyết thông tin. Đây là ví dụ điển hình về một lý thuyết được khởi xướng đầu tiên bởi một người, kỹ sư điện người Mỹ Claude E. Shanon, mà ý tưởng đầu tiên của ông được đề xuất trong bài báo "A mathematical theory of communication", đăng trên Bell System Technical Journal (1948).

Nội dung bài báo của Shanon nhằm giải quyết nhu cầu về cơ sở lý thuyết của công nghệ truyền thông. Nhu cầu này nảy sinh do độ phức tạp của quá trình truyền tin trên các kênh truyền thông như các mạng lưới điện thoại, điện báo và truyền thanh. Thuật ngữ thông tin ở đây là để chỉ các thông báo được truyền đi như: tiếng nói và âm nhạc được truyền đi bằng điện thoại hoặc truyền thanh, hình ảnh được truyền đi bằng truyền hình, các dữ liệu số hoá trên các mạng máy tính.

Lý thuyết thông tin (Infomation theory) là lý thuyết liên quan đến các định luật toán học chi phối việc truyền, tiếp nhận và xử lý thông tin. Chính xác hơn, lý thuyết thông tin đề cập tới các vấn đề về đo số lượng thông tin, biểu diễn thông tin (như vấn đề mã hoá) và khả năng của các hệ thống truyền thông có nhiệm vụ truyền, nhận và xử lý thông tin. Việc mã hoá có thể dùng để chuyển các tín hiệu âm thanh và hình ảnh thành tín hiệu điện, điện từ hoặc dùng để bảo mật thông tin. Lý thuyết thông tin còn được ứng dụng trong những lĩnh vực khác nhau như điều khiển học, ngôn ngữ học, tâm lý học...

Các thành phần của một hệ thống truyền thông

Một hệ thống truyền thông bao gồm các thành phần sau đây:

- Nguồn tin sản sinh ra thông tin hay thông báo sẽ được truyền đi (ví dụ một phát thanh viên).

- Vật truyền như điện thoại, micro, máy tăng âm, máy phát thanh, có nhiệm vụ chuyển thông báo thành tín hiệu điện hay điện từ.

- Các tín hiệu này sẽ được truyền đi qua kênh truyền tin như dây dẫn, khí quyển.

- Máy thu có nhiệm vụ chuyển tín hiệu trở về thông báo ban đầu như máy thu thanh, thu hình, tai nghe điện thoại.

- Nơi nhận, ví dụ như một người nghe đài, nghe điện thoại, một bạn xem truyền hình.

Hai vấn đề quan trọng phải giải quyết trong một hệ thống truyền thông là giảm các nhiễu gây ra bởi hệ thống và tăng khả năng sử dụng của kênh truyền.

Đo thông tin

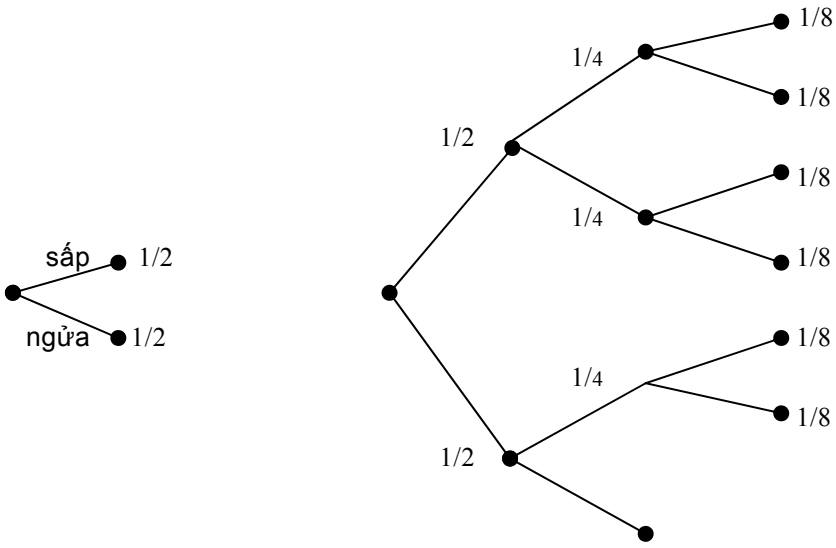
Ý tưởng cơ bản của lý thuyết thông tin là thông tin có thể xử lý như một đại lượng vật lý, cũng như khối lượng hay năng lượng, nó có thể xác định và đo được bằng đại lượng toán học.

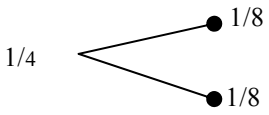
Trước khi xem xét vấn đề thông tin được đo như thế nào, ta cần làm rõ ý nghĩa của "thông tin". Thông thường các thông báo được truyền đi thường mang một ý nghĩa: chúng mô tả hay liên quan tới hiện thực hay các sự kiện có thể nhận thức được. Tuy nhiên không phải bao giờ cũng như vậy. Như trong trường hợp truyền âm nhạc chẳng hạn, ý nghĩa, nếu có sẽ tinh tế hơn trong trường hợp truyền văn bản. Trong một số trường hợp người kỹ sư phải truyền đi một chuỗi các chữ và số hoàn toàn không có nghĩa. Cũng có trường hợp, ý nghĩa không liên quan gì đến vấn đề truyền thông tin.

Xuất phát từ quan điểm truyền tin, khía cạnh ý nghĩa của thông tin là ở chỗ nhận được một thông báo riêng biệt từ một tập hợp các thông báo có thể. Cái mà phải truyền đi là thông báo riêng biệt đã được chọn từ nguồn tin. Như vậy, trong lý thuyết thông tin, thông tin là ý định lựa chọn một thông báo từ một tập hợp các thông báo có thể. Hơn nữa sự lựa chọn này xảy ra với một xác suất nào đó, một vài thông báo có thể xảy ra thường xuyên hơn các thông báo khác.

Sự lựa chọn đơn giản nhất là lựa chọn giữa hai khả năng như nhau, tức là mỗi khả năng có xác suất bằng nhau, bằng 1/2. Đây là tình huống, ví dụ: khi gieo một đồng tiền, khả năng chọn mặt "sấp" hoặc "ngửa" là như nhau và có xác xuất là p = 1/2. Lượng thông tin được tạo ra từ cách lựa chọn như thế được coi là một đơn vị của thông tin và đơn vị này được gọi là bit. Trong ví dụ trên, nếu ký hiệu "sấp" là số 1, "ngửa" là số 0 thì chỉ có một cách chọn để biểu diễn thông báo là 1 hoặc 0. Số 0 và 1 là những chữ số của hệ đếm nhị phân và việc chọn giữa hai ký hiệu đó tương ứng với một đơn vị thông tin nhị phân, đó chính là bit.

Vậy theo quan điểm của lý thuyết thông tin, bit là đơn vị thông tin tương đương với kết quả của sự lựa chọn giữa hai khả năng như nhau, như giữa 1 và 0 trong hệ số nhị phân được sử dụng trong máy tính điện tử. Bit là từ viết tắt của từ "binary digit", nghĩa là số nhị phân. Nó cũng được sử dụng làm đơn vị của bộ nhớ máy tính, tương ứng với khả năng lưu trữ kết quả của sự lựa chọn giữa hai khả năng.





Hình 4. Xác suất của những lựa chọn có khả năng như nhau

Sự lựa chọn với thông tin một bit có thể mô tả bằng sơ đồ trình bày ở hình 4 (trái). Mỗi đường trên (sấp) hoặc đường dưới (ngửa) có thể lựa chọn như nhau với xác suất bằng 1/2.

Nếu tập hợp các thông báo bao gồm N thông báo có khả năng như nhau thì số lượng thông tin, ký hiệu là I được tính bởi công thức:

$$I = \log_2 N.$$

Rõ ràng với N = 2 thì số lượng thông tin sẽ bằng $\log_2 N = \log_2 2 = 1$, cho nên công thức đưa ra phù hợp với đơn vị thông tin đã được lựa chọn ở trên.

Ta biết rằng nếu có N thông báo có khả năng như nhau thì sác xuất để có một thông báo là p = 1/N, suy ra N = 1/p. Do đó để liên kết thông tin với sác xuất, Shanon đưa ra công thức sau, đồng nhất với công thức trên:

$$I = \log_2 1/p$$

Để hiểu rõ ý nghĩa của công thức trên ta xét thêm một ví dụ. Giả sử gieo ba lần liên tiếp một đồng tiền thì với cách ký hiệu mặt "sấp" hoặc "ngửa" như trên, 8 kết quả đồng khả năng (hay thông báo) có thể biểu diễn như sau:

000, 001, 010, 100, 011, 101, 110, 111.

Xác suất của mỗi thông báo này là p =1/8. Sự lựa chọn có thể xảy ra ở ba mức, mỗi mức là một bit. Bit thứ nhất tương ứng lần thứ nhất gieo đồng tiền, được lựa chọn một trong hai khả năng sấp (1) hoặc ngửa (0); bit thứ hai ứng với lần gieo thứ hai, được lựa chọn với bộ-đôi thứ nhất hoặc bộ-đôi thứ hai trong 4 khả năng; bit thứ ba ứng với lần gieo thứ ba, được lựa chọn với bộ-bốn thứ nhất hoặc bộ-bốn thứ hai trong 8 khả năng (hình 4 (phải)). Trong trường hợp này N = 8 và lượng thông tin của nó là:

$$I = \log_2 N = \log_2 1/p = \log_2 8 = 3$$

Đó chính là số bit cần thiết để biểu diễn mỗi thông báo nói trên.

Entropy và thông tin

Nếu xác suất không bằng nhau, các thông báo trong tập hợp có lượng thông tin khác nhau kết hợp với chúng. Giả sử xác suất của các thông báo có thể có trong tập hợp là

$$p_1, p_2, ..., p_N$$

thì lượng thông tin tương ứng kết hợp với chúng tương ứng với

$$\log_2(1/p_1), \log_2(1/p_2), ..., \log_2(1/p_N).$$

Kỳ vọng toán học hay giá trị trung bình của các giá trị này, ký hiệu là H là:

$$H = p_1\log_2(1/p_1) + p_2\log_2(1/p_2) + ... + p_N \log_2(1/p_N)$$

được gọi là entropy, hay thông tin trung bình của tập hợp các thông báo. Công thức trên còn có thể viết dưới dạng:

$$H = - p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ ... \ - p_N \log_2 p_N$$

Entropy là thuật ngữ mượn của nhiệt động học. Đó là đại lượng đặc trưng cho độ bất định của hệ thống. Trong lý thuyết thông tin, entropy của một tập hợp thông báo là lượng thông tin trung bình của nó. Nói cách khác lượng thông tin thu được về việc xảy ra một sự kiện nào đó được xác định là bằng độ bất định của sự kiện đó trước khi biết nó xảy ra. Điều đó cũng có nghĩa là lượng tin càng cao khi độ bất ngờ của sự kiện càng lớn.

Chú ý rằng :

$$\log_2 k = \log_2 10 . \log_{10} k = 3,32 . \lg k$$

nên ta có thể tính loga cơ số 2 thông qua loga cơ số 10.

Ví dụ: Giả sử qua nhiều năm quan sát thời tiết tại một thời điểm người ta thu được kết quả sau:

Thời tiết trong ngày 15 tháng 6 (phép thử α_1):

Các kết cục của phép thử	có mưa	không mưa
Xác suất	0,4	0,6

Thời tiết trong ngày 15 tháng 11 (phép thử α_2):

Các kết cục của phép thử	có mưa	có tuyết	không mưa
Xác suất	0,65	0,15	0,2

Entropy tương ứng của hai phép thử này là:

$H(\alpha_1) = - 0,4 \log_2 0,4 - 0,6 \log_2 0,6 = 0,97$

$H(\alpha_2) = - 0,66 \log_2 0,65 - 0,15 \log_2 0,15 - 0,2 \log_2 0,2 = 1,28$

Vậy $H(\alpha_2) > H(\alpha_1)$, do đó tại khu vực đang xét thời tiết ngày 15 tháng 11 khó dự báo hơn thời tiết ngày 15 tháng 6.

Entropy nhận giá trị nhỏ nhất bằng 0 khi có một thông báo chắc chắn xảy ra (tức là có xác suất bằng 1) còn tất cả các thông báo khác không xảy ra (có xác suất bằng 0). Về mặt trực giác ta cũng thấy rằng sẽ không có thông tin trong một thông báo mà ta đã biết chắc nó xảy ra. Ví dụ khi gieo một đồng tiền mà thông báo "xuất hiện sấp hoặc ngửa" thì coi như chẳng có nội dung thông tin nào. Ngược lại entropy nhận giá trị lớn nhất là $\log_2 N$, khi N thông báo trong tập hợp có khả năng như nhau (trong trường hợp này $p=1/N$ và do đó $N=1/p$).

Ví dụ: Nếu thông báo được truyền đi bao gồm các tổ hợp ngẫu nhiên của 26 chữ cái, một khoảng trống và 5 dấu chấm câu, tổng cộng là 32 ký hiệu và giả sử rằng xác suất của mỗi ký hiệu là như nhau thì entropy của nó là $H = \log_2 32 = 5$. Điều đó có nghĩa là để mã hoá mỗi ký hiệu trong 32 ký hiệu nói trên ít nhất phải cần 5 bit: 00000, 00001, 00010, ..., 11111. Hiệu quả của việc truyền và lưu trữ thông tin đòi hỏi phải rút gọn số các bit dùng để mã hoá.

Như vậy với lý thuyết thông tin, Shanon đã chứng minh được rằng thông tin có thể đo được, tức là với bản tin bất kỳ ta có thể xác định được nó chứa bao nhiêu đơn vị tin tức.

Thông tin có thể đo được. Đó là phát minh cũng có ý nghĩa về sự hiểu biết của con người đối với thế giới khách quan như ý nghĩa về khả năng đo được của năng lượng. Người ta đã chế tạo ra các máy để sản sinh và chế biến được năng lượng và giờ đây người ta cũng chế tạo ra các máy để gia công tin tức, đó là máy tính điện tử.

Những phương pháp của lý thuyết thông tin được sử dụng để mã hoá các thông tin ngữ nghĩa, đưa thông tin vào máy tính để bảo quản và thực hiện các quá trình tìm tin, truyền tin.

Một trong những nhiệm vụ chính của lý thuyết thông tin là nghiên cứu và nâng cao hiệu quả các quá trình truyền tin theo các kênh. Các quá trình này cũng được thông tin học nghiên cứu.