TẠP CHÍ TIN HỌC VÀ ĐIỀU KHIỂN HỌC, TẬP I (1991), Số 1 (15-21)

MỘT SỐ KHÁI NIỆM CỦA CÁC HỆ TÌM KIẾM VÀ LƯU TRỮ THÔNG TIN

NGUYỄN BÁ TƯỜNG
ĐHKT Lê Quí Đôn

Chúng ta đã biết trong [1], [3], [4], [5], [6] các tác giả Babad, Macda, Pawlak, Salton và Wong + Chiang đã định nghĩa và xét các hệ tin theo các mục đích tiếp cận riêng. Trong bài này chúng ta cũng xét các hệ tin theo một hướng tiếp cận thích hợp trong lý thuyết tập hợp và ngôn ngữ hình thức.

Trước hết trong phạm vi bài báo này chúng ta sẽ xét một số tính chất của các hệ tin trong lý thuyết tập hợp. Trên cơ sở các khái niệm đã biết trong lý thuyết tập hợp, khái niệm đẳng cấu, chúng ta sẽ xét được một lớp các bài toán tương đương. Điều này giúp chúng ta trong công tác cài đặt sẽ bớt được nhiều công việc và giảm bớt được các thao tác không cần thiết.

I – CÁC KHÁI NIỆM CƠ BẢN

Thông thường trong lĩnh vực tin học các bài toán trên máy tính có đặc tính chung là quản lý thông tin, lưu trữ và tìm kiếm các đối tượng có một số thuộc tính nào đó.

Loại bài toán này cần lưu trữ:

a - Danh sách các đối tượng, gọi tắt là tập X,

b - Các thuộc tính của các đối tượng lập thành tập A,

c - Với mỗi đối tượng cụ thể biết trước các thuộc tính của nó.

Vậy giữa tập các đối tượng X và tập các thuộc tính A có một qui tắc ràng buộc như hàm đa trị $F : X \to A$.

Sau đây chúng ta hãy xét một vài ví dụ.

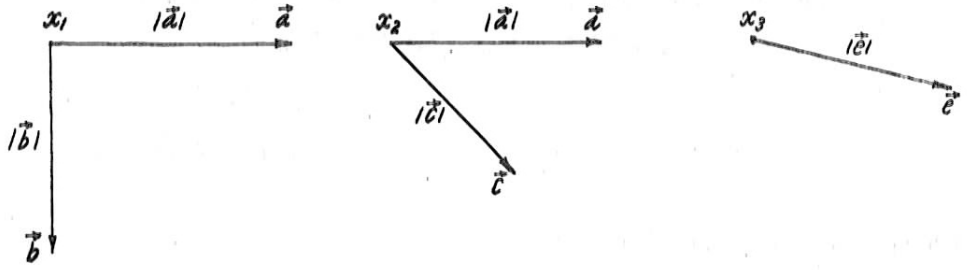
Ví dụ 1: Xét tập hồ sơ cán bộ của một cơ quan. Hồ sơ này gồm danh sách họ tên cán bộ, gọi tắt là tập X. Mỗi cán bộ có thể có các thuộc tính như: kĩ sư, giáo sư, mức lương, dân tộc, nam, nữ, tuổi, v.v... Tập các thuộc tính này gọi là tập A.

Mỗi $x \in X$ ứng với một số thuộc tính của A. Đây là một quan hệ hàm

$$F : X \to 2^A .$$

Ví dụ 2: Xét hệ lực gồm các điểm đặt $x_1, x_2, x_3, \ldots$ thuộc tập X. Mỗi điểm đặt chịu tác động của các lực $\vec{a}$, $\vec{b}$, $\vec{c}$... với các độ lớn tương ứng là $|\vec{a}|$, $|\vec{b}|$, $|\vec{c}|$... Gọi tập hợp tất cả các hướng lực $\vec{a}$, $\vec{b}$, $\vec{c}$ ... và các độ lớn tương ứng $|\vec{a}|$, $|\vec{b}|$, $|\vec{c}|$... là tập A. (Xem hình 1).

Rõ ràng giữa X và A có quan hệ hàm $F : X \to 2^A$.



Hình 1

Chú ý rằng từ ví dụ 2 hàm F có thể viết dưới dạng

$$F = \{x_1 \to \vec{a}\ \vec{b}\ |\vec{a}|\ |\vec{b}|,\ \ x_2 \to \vec{a}\ \vec{c}\ |\vec{a}|\ |\vec{c}|,\ \ x_3 \to \vec{e}\ |\vec{e}|, \ldots\}$$

Trong đó $\vec{a}\ \vec{b}$; $|\vec{a}|$; $|\vec{b}|$ là một xâu của các thuộc tính $\vec{a}$, $\vec{b}$, $|\vec{a}|$, $|\vec{b}|$.

Ví dụ 3: Cho

$$X = \{x, y, z\}$$
$$A = \{a, b, c, d\}$$
$$F = \{x \to a, y \to ab, z \to abcd\}$$

Nếu ta hiểu $y \to ab$ là y có các thuộc tính a, b thì rõ ràng rằng F cũng là một hàm $F : X \to 2^A$.

II – ĐỊNH NGHĨA HỆ TIN

Định nghĩa 1: Hệ tin là bộ 3

$$L = \langle X, A, F\rangle$$

Trong đó X, A là các tập hữu hạn gọi là tập các đối tượng và tập các thuộc tính tương ứng, F là hàm đa trị xác định từ X vào A, tức là $F : X \to 2^A$.

Chú ý:

a - Về sau chúng ta sẽ dùng các chữ bé x, y, z ... để chỉ các phần tử thuộc X, các chữ bé a, b, c ... để chỉ các phần tử của tập A. Còn các chữ lớn N, M, L ... để chỉ các hệ tin, $R(L)$ là họ hệ tin con của L.

b - Không mất tính tổng quát và để hệ tin gắn liền với các bài toán thực tế chúng ta chỉ xét các hệ tin mà mỗi thuộc tính $a \in A$ có ít nhất một phần tử có thuộc tính đó.

III – MỘT SỐ KHÁI NIỆM VÀ TÍNH CHẤT CƠ BẢN

1. Phép ghép. Cho hai hệ tin

$$L = \langle X, A, F\rangle\ ;\ M = \langle Y, B, G\rangle$$

Định nghĩa 2: Phép ghép của hai hệ tin L, M ký hiệu $+$ được định nghĩa như sau:

$$L + M = \langle Z, C, Q \rangle$$

với $Z = X \cup Y$, $C = A \cup B$, $Q : Z \to 2^C$, trong đó

$$Q(x) = \begin{cases} F(x), & x \in Z \setminus X \\ G(x), & x \in Z \setminus Y \\ F(x) \cup G(x), & x \in X \cap Y \end{cases}.$$

Từ định nghĩa ta có ngay các kết quả như sau

a) $\forall L$ thì $L + L = L$;

b) $\forall L, M$ thì $L + M = M + L$,

c) $\forall L, M, N$ thì $L + (M + N) = (L + M) + N$,

Tức là phép ghép có tính giao hoán, kết hợp;

d) Nếu gọi $\Phi = \langle \emptyset, \emptyset, \emptyset \rangle$ là hệ tin Không thì $\forall L : L + \Phi = L$.

Về sau để cho tiện trong phép ghép của nhiều hệ tin L_1, L_2, L_3, ... L_n ta ký hiệu:

$$L_1 + L_2 + L_3 + \ldots + L_n = \sum_{i=1}^{n} L_i$$

2. Hệ tin con. Cho

$$L = \langle X, A, F \rangle$$
$$M = \langle Y, B, G \rangle$$

Định nghĩa 3: Hệ tin M được gọi là hệ tin con của L nếu

a) $Y \subset X$

b) $B \subset A$

c) $G/Y = F/Y$

Khi đó chúng ta ký hiệu $M \subset L$.

Từ định nghĩa chúng ta thấy rằng hệ tin con M của hệ tin L chẳng qua là hệ L sau khi đã bớt đi một số đối tượng của X cùng với các thuộc tính tương ứng của chúng.

Chúng ta cũng dễ dàng thấy ngay rằng

$$\forall L, \quad L \subset L \quad \text{và} \quad \Phi \subset L$$

Về sau những hệ tin con M của L và khác L ta gọi là hệ tin con thực sự và ký hiệu $M \subsetneq L$.

Ví dụ 4: Giả sử

$$X = \{x_1, x_2, x_3, x_4\}$$
$$A = \{a, b, c, d\}$$
$$F = \{x_1 \to a,\ x_2 \to ab,\ x_3 \to cd,\ x_4 \to abcd\}$$

và $L = \langle X, A, F\rangle$. Khi đó $M = \langle Y, B, G\rangle$ với $Y = \{x_1\}$, $B = \{a\}$, $G = \{x_1 \to a\}$.

Rõ ràng rằng M là hệ tin con thực sự của L. Còn $N = \langle \{x_1\}, \{a, b\}, \{x_1 \to b\}\rangle$ không phải là hệ tin con của L.

3. Họ các hệ tin con được sắp

Định nghĩa 4 (xem [2]): Tập A và quan hệ $<$ viết tắt $(A, <)$ được gọi là tập được sắp bộ phận nếu:

a) $\forall a, b \in A$; $a < b$ thì $b < a$,

b) $\forall a, b, c \in A$; $a < b$ và $b < c$ thì $a < c$.

Ngoài ra $(A, <)$ được gọi là được sắp nếu nó là được sắp bộ phận và:

c) $\forall a, b \in A$ thì $a < b$ hoặc $b < a$.

Định nghĩa 5 (xem [2]): $(A, <)$ được gọi là định hướng nếu nó là được sắp bộ phận và

$$\forall a, b \in A,\ \exists c \in A \text{ sao cho } a \leq c \text{ hoặc } b \leq c.$$

Cho hệ tin

$$L = \langle X, A, F\rangle$$

Gọi $R(L)$ là họ các hệ tin con của L.

Hệ quả:

a) Rõ ràng rằng họ $R(L)$ với quan hệ $\subset \neq$ là họ được sắp bộ phận. Tức là $\langle R(L), \subset\rangle$ là được sắp bộ phận.

b) Chúng ta cũng dễ thấy rằng $\langle R(L), \subset\rangle$ là họ định hướng.

Định lý 1. Cho hệ tin $L = \langle X, A, F\rangle$ và $R(L)$ là họ các hệ tin con của L. Khi đó trong $R(L)$ có họ các hệ tin con được sắp.

Chứng minh: Chúng ta có thể xây dựng họ hệ tin con được sắp như sau

$$L_0 = \Phi = \langle \emptyset, \emptyset, \emptyset\rangle$$
$$L_1 = \langle \{x_1\}, A_1, F_1\rangle$$

với $x_1 \in X$, A_1 là tập các thuộc tính của x_1, F_1 là thu hẹp của F trên $\{x_1\}$;

$$L_2 = \langle \{x_1, x_2\}, A_2, F_2\rangle$$

với $x_2 \in X$ và A_2 là tập thuộc tính của $\{x_1, x_2\}$, F_2 là thu hẹp của F trên $\{x_1, x_2\}$.

Tiếp tục mở rộng ra tương tự ta được L_3, L_4 ... Vì X hữu hạn nên sau n bước hữu hạn ta được $L_n = L$ và rõ ràng rằng

$$\Phi \subset L_1 \subset L_2 \subset L_3 \subset \ldots \subset L_n$$

Chúng ta cũng thấy ngay rằng đó là dãy hệ tin con được sắp cực đại của họ $R(L)$. Ở đây khái niệm cực đại chúng ta hiểu là số lớn nhất các hệ tin con trong họ $R(L)$. Về sau nếu không có gì thay đổi ta ký hiệu họ hệ tin con được sắp cực đại là $R_0(L)$.

Ví dụ 5: Chúng ta hãy xét hệ tin trong ví dụ 4

$$L = \langle \{x_1, x_2, x_3, x_4\}, \{a, b, c, d\}, F \rangle$$

với $F = \{x_1 \to a,\ x_2 \to ab,\ x_3 \to cd,\ x_4 \to abcd\}$. Khi đó

$$R_0(L) = \{L_0, L_1, L_2, L_3, L_4\}.$$

Trong đó

$$L_0 = \Phi$$
$$L_1 = \langle \{x_1\}, \{a\}, \{x \to a_1\} \rangle$$
$$L_2 = \langle \{x_1, x_2\}, \{a, b\}, \{x_1 \to a_1,\ x_2 \to ab\} \rangle$$
$$L_3 = \langle \{x_1, x_2, x_3\},\ \{a, b, c, d\},\ \{x_1 \to a,\ x_2 \to ab,\ x_3 \to cd\} \rangle$$
$$L_4 = L$$

4. Lồng nhau. Cho hệ tin $L = \langle X, A, F \rangle$. $R(L)$ là họ các hệ tin con tương ứng.

Định nghĩa 6: Họ hệ tin con $R(L)$ được gọi là tựa lồng nhau nếu: $\exists L_1, L_2 \in R(L)$ sao cho $L_1 \subset L_2$. Ngược lại gọi là rời nhau.

5. Đồng cấu, đẳng cấu

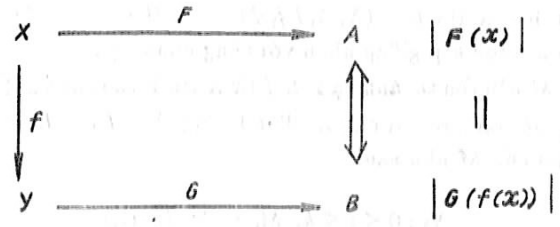
Đồng cấu: Cho $L = \langle X, A, F \rangle$, $M = \langle Y, B, G \rangle$.

Định nghĩa 7: L được gọi là đồng cấu với M nếu có ánh xạ $f : X \to Y$ sao cho $\forall x \in X$ số thuộc tính của x bằng số thuộc tính của $f(x)$. Tức là

$$x \in X : |F(x)| = |G(f(x))|$$

Trong đó: $|F(x)|$ là lực lượng của tập $F(x)$. Khi đó chúng ta ký hiệu $L \sim M$.

Chúng ta có thể hình dung khái niệm đồng cấu như trong hình 2.



Hình 2

Hệ quả: Cho $L = \langle X, A, F\rangle$, $R(L)$ là họ tin con của L; $M \in R(L)$.

a) Chúng ta dễ dàng thấy rằng $M \sim L$.

b) Nếu $M = \langle Y, B, G\rangle$ và $L \sim M$ thì trong M có hệ tin con M_0 sao cho $M_0 \sim M$.

Thật vậy giả sử $L \sim M$ ta xây dựng M_0 như sau:

$$M_0 = \langle \{f(x),\ x \in X\},\ \{G(f(x)),\ x \in X\},\ G_0\rangle$$

Trong đó $G_0 = G/\{F(x),\ x \in X\}$. Từ đây chúng ta dễ dàng thấy rằng $M_0 \sim M$.

Đẳng cấu: Cho $L = \langle X, A, F\rangle$, $M = \langle Y, B, G\rangle$.

Định nghĩa 8: L được gọi là đẳng cấu với M nếu $L \sim M$ và f là ánh xạ lên. Khi đó chúng ta sẽ ký hiệu là $L \simeq M$.

Rõ ràng là quan hệ đẳng cấu là quan hệ tương đẳng trong họ các hệ tin R.

Ví dụ 6: Xét bài toán quản lý học sinh thi vào đại học. Trong trường hợp này ta thấy mỗi học sinh có các đặc trưng sau đây: Nguồn thi vào, ưu tiên, điểm toán, điểm lý, điểm hóa, tổng số điểm.

Mỗi đặc trưng bao gồm các khả năng, gọi là thuộc tính như: Nguồn thi vào: phổ thông (PT), tự do (TD), bộ đội (BĐ). Ưu tiên có thể là không ưu tiên (0), ưu tiên (1), ưu tiên (2). Điểm toán, lý, hóa có thể là từ 0, $\frac{1}{2}$, 1, $1\frac{1}{2}$, 2, ..., 10.

Vậy tập $A = \{PT,\ TD,\ BĐ,\ 0,\ \frac{1}{2},\ \ldots,\ 10\}$ là tập thuộc tính.

Rõ ràng hai tập hồ sơ của học sinh thi vào hai trường đại học là đẳng cấu với nhau.

Nếu số học sinh thi vào các trường trung cấp được lấy từ các học sinh thi vào đại học cùng ngành không đủ điểm vào đại học thì tập hồ sơ của học sinh vào trường trung cấp là hệ tin con của tập hồ sơ học sinh vào trường đại học tương ứng.

Gọi R là họ tất cả các tập hồ sơ vào các trường đại học và trung cấp. Rõ ràng R là họ các hệ tin tựa lồng nhau nếu có ít nhất một trường trung cấp lấy điểm theo cách chọn ở trên. Trong trường hợp này, nếu gọi $L_1, L_2, \ldots, L_k$ là các tập hồ sơ (các hệ tin) của học sinh thi vào các trường đại học thì $L = L_1 + L_2 + \ldots + L_k$ là hồ sơ của các học sinh thi vào các trường đại học được quản lý trên máy tính ở Bộ Đại học.

6. Được sắp giống nhau

Định nghĩa 9 (xem [2]): Hai tập A, B với các quan hệ α, β tương ứng trong chúng gọi là được sắp giống nhau nếu tồn tại ánh xạ 1–1 $f : A \to B$ sao cho

$$\forall x, y \in A;\ x\alpha y \Leftrightarrow f(x)\beta f(y)\ .$$

Định lý 2. Cho hai hệ tin $L = \langle X, A, F\rangle$, $M = \langle Y, B, G\rangle$, $L \simeq M$. Khi đó trong $R(L)$ và $R(M)$ có hai dãy hệ con được sắp giống nhau với cùng quan hệ $\subset$.

Thật vậy vì $L \simeq M$ nên tồn tại ánh xạ 1–1 f từ X lên Y sao cho $\forall x \in X$, $|F(x)| = |G(F(x))|$.

Lấy $R_0(L)$ là họ hệ con cực đại của L. Tức là $R_0(L) = L_0 \subset L_1 \subset L_2 \subset \ldots \subset L_k$. Ta xây dựng họ $R_0(M)$ cực đại của M như sau:

$$\forall i : 0 \le i \le k,\ M_i = \langle Y_i, B_i, G_i\rangle$$

với Y_i là ảnh của X_i qua ánh xạ f,

B_i là tập các thuộc tính tương ứng của Y_i,

$G_i = G/Y_i$.

Ở đây ta hiểu X_i là tập đối tượng của hệ tin con L_i của $R_0(L)$.

Rõ ràng ánh xạ $L_i \to M_i$ là 1–1. Từ đây hai họ hệ tin R và S được sắp giống nhau. Ta ký hiệu $R \equiv S$.

Hệ quả:

a) Họ R là họ các hệ tin được sắp thì được sắp giống nhau với chính nó $R \equiv R$. Trong đó quan hệ trong R là $\subset$.

b) $R \equiv S \Leftrightarrow S \equiv R$.

c) Nếu $R \equiv S$ và $S \equiv T$ thì $R \equiv T$.

Vậy quan hệ được sắp giống nhau là quan hệ tương đẳng trong họ các hệ tin.

Trong phần sau chúng ta sẽ xét một số ứng dụng và cách cài đặt các hệ tin trên máy tính điện tử.

Nhận ngày 18 - 4 - 1991

TÀI LIỆU THAM KHẢO

1. Babad J., A record and file Partinioning model. Comm. ACM, Vol. 20, No. 1, 22–31 (1977).
2. Kuratowski K., Wstep do teorii mnogosci i topologii, Warszawa (1965).
3. Maeda T., A formal treatment of document information system. Information Processing and Management, Vol. 17, No. 6, 319–328 (1981).
4. Pawlak Z., Systemy informacyjne. Podstawy teoretycrne, Warszawa (1983).
5. Salton G.,Automatic information organization and retrieval. McGraw Hill, New York (1968).
6. Wong E., Chiang T. C., Canonical Structure in attribute based file organization. Comm. ACM, Vol. 14, 593–596 (1971).

ABSTRACT

SOME PROPERTIES OF INFORMATION STORAGE AND RETRIEVAL SYSTEMS

In this paper we describe a information system which is some different from previous definition. Problems connected with describing objects by means of an appropriate language of sets as well as efficient methods in the memory of a computer are considered.

Areas for further studies are also suggested.