Tạp chí Tin học và Điều khiển học T. 11, S. 2 (1995) (7 – 15)

NGỮ NGHĨA CỦA DỮ LIỆU TRONG CƠ SỞ DỮ LIỆU VỚI THÔNG TIN KHÔNG ĐẦY ĐỦ

NGUYỄN CÁT HỒ & LÊ THẾ THĂNG

Institute of Information Technology

Summary. In this paper, we consider semantics of data in databases with incomplete information as a pair $S = (r, \alpha)$, where r is a relation with null values and α is an assignment of every tupe t in r to a set of objects in the real world. Some relationships betwwen a database and the real world are also examined.

I. – GIỚI THIỆU

Một vấn đề gây trở ngại trong lý thuyết cơ sở dữ liệu (CSDL) là vấn đề thông tin không đầy đủ. Đã có nhiều cách tiếp cận khác nhau đến vấn đề này (xem [2]-[7]).

Thông tin không đầy đủ xuất hiện trong CSDL do nhiều nguồn gốc khác nhau. Sau đây ta xét một số tình huống mà có thể dẫn đến CSDL với thông tin không đầy đủ. Khi chúng ta tiến hành thu nạp thông tin vào CSDL không phải tất cả các thông tin cần thiết đều đã có, điều này có thể dẫn đến một CSDL với một thông tin không đầy đủ. Một thông tin mới, đến một thời điểm nào đó trở nên có giá trị, theo kiểu từng phần, nếu chúng ta không đưa thông tin này vào CSDL, chúng ta sẽ mất đi một phần thông tin của thế giới thực. Nếu chấp nhận đưa vào, sẽ dẫn đến CSDL với thông tin không đầy đủ. Trong một số trường hợp vì lý do an ninh dữ liệu, chúng ta muốn che một số thông tin của một lớp đối tượng. Điều này cũng đưa đến CSDL với thông tin không đầy đủ. Cũng có khi để đạt được các thông tin cần thiết chúng ta phải chi phí rất lớn hoặc rất khó khăn hoặc không đúng nguyên tắc. Do đó để tránh phí tổn, chúng ta chấp nhận thiếu những thông tin đó trong CSDL. Trong CSDL phân tán, trách nhiệm thu nhập thông tin là thuộc về các địa phương, còn CSDL tổng thể phải bao gồm thông tin của nhiều địa phương gộp lại, do đó sẽ dẫn đến CSDL với thông tin không đầy đủ. Nếu mỗi lớp người sử dụng có một khung nhìn vào CSDL thì việc bỏ đi một số thông tin được lưu trữ trong CSDL ở khung nhìn là hợp lý. Cuối cùng vấn đề cập nhật thông tin trên một khung nhìn thường dẫn đến thông tin không đầy đủ.

Như vậy thông tin không đầy đủ có thể được hiểu như là: thông tin đó đã tồn tại nhưng ta chưa biết hoặc thông tin đó đến nay chưa tồn tại hoặc không xác định được hiện tại nó như thế nào v.v.... Trong CSDL theo mô hình quan hệ, thông tin không đầy

đủ được biểu diễn bằng giá trị null hay nói khác đi, giá trị null trong CSDL thể hiện thông tin không đầy đủ của đối tượng mà ta quan sát. Trong bài này ta ký hiệu giá trị null bằng một ký hiệu đặc biệt và nghiên cứu ngữ nghiã của dữ liệu trong CSDL với thông tin không đầy đủ hay mối quan hệ của CSDL với thế giới thực.

II. – CÁC KHÁI NIỆM CƠ SỞ

Chúng ta xét tập U, hữu hạn, khác rỗng và gọi là tập vũ trụ các thuộc tính. Các phần tử trong U gọi là các thuộc tính và được kí hiệu là những chữ hoa ở đầu bảng các chữ cái như A, B, C,... Các tập con của U, tập các thuộc tính, được kí hiệu là những chữ hoa ở cuối bảng chữ cái như X, Y, Z... Mỗi thuộc tính $A \in U$ sẽ được gán tương ứng với một tập khác rỗng $d(A)$ và gọi là miền giá trị của thuộc tính A. Trong $d(A)$ có một phần tử phân biệt gọi là null và được kí hiệu là $\perp \in d(A)$.

Hợp các miền giá trị của các thuộc tính, được kí hiệu là $D = \cup\{d(A) : A \in U\}$. Giả sử X là một tập con khác rỗng của U. Ánh xạ $t : X \to D$ thoả mãn điều kiện $t(A) \in d(A)$ vơi $\forall A \in X$, được gọi là một bộ trên X và kí hiệu là $\{t(A) : A \in X\}$. Cho t là một bộ trên X, $A \in X$. Nếu $t(A) \neq \perp$ ta viết $t(A)!$. Nếu $t(A) \neq \perp$, $\forall A \in X$ ta viết $t!$, nghiã là t không chứa giá trị null. Giả sử $Y \quad X$, khi đó ta kí hiệu $t[Y] = \{t(A) \quad A \in Y\}$ và gọi là phép chiếu của t lên Y.

Tập tất cả các bộ được kí hiệu là J. Trong J ta đưa ra một quan hệ thứ tự bộ phận $\leq$ như sau: t, $s \in J$, ta nói t ít thông tin hơn s hay s nhiều thông tin hơn t và viết $t \leq s$ hay $s \geq t$, nếu t và s là hai bộ trên tập thuộc tính X và $\forall A \in X$, $t(A)!$ kéo theo $t(A) = s(A)$. Ta viết $t < s$ hay $s > t$ nếu $t \leq s$ và $t \neq s$.

Cho X là tập thuộc tính. Một quan hệ r trên X là một tập các bộ trên X, đôi khi ta viết $r[X]$. Nếu $Y \subseteq X$ tập $\{t[Y] : t \in r\}$ được gọi là phép chiếu của r lên Y và được kí hiệu là $r[Y]$.

Ta xem thế giới thực như là tập O các đối và mối quan hệ giữa chúng. Mối quan hệ của các đối tượng thể hiện ở thông tin về chúng. Giả sử U là tập các thuộc tính mà trên đó ta quan sát các đối tương. Ta giả thiết rằng thông tin về một đối tượng $o \in O$ có thể được biểu diễn bởi nhiều bộ trong J. Tập tất cả các thông tin về đối tượng o, được kí hiệu là infor(o) $\subset J$. Như vậy infor(o) là tập tất cả các bộ biểu diễn một thông tin nào đó về o. Giả định rằng: 1) Có thể phân biệt được đối tượng $o \in O$ với các đối tượng khác bằng tập infor(o), nghĩa là o, $o' \in O$, $o \neq o'$ khi và chỉ khi infor(o') $\neq$ infor(o). 2) infor(o) có tính chất: nếu $t \in$ infor(o), $s \in J$, $s \leq t$ thì $s \in$ infor(o).

III. – NGỮ NGHĨA CỦA DỮ LIỆU VỚI THÔNG TIN KHÔNG ĐẦY ĐỦ

Thông tin của thế giới thực được lưu trữ trong CSDL. Do đó ta có thể xem ngữ nghĩa của dữ liệu trong CSDL như là một cách gán các đối tượng trong thế giới thực với dữ liệu trong CSDL. Mỗi bộ t trong quan hệ r là thông tin của một lớp đối tượng trong thế giới thực O. Ta sẽ hình thức hoá điều đó bằng định nghĩa sau đây.

Định nghĩa 3.1. Giả sử O là tập các đối tượng, U là tập các thuộc tính vũ trụ. P(O) là tập các tập con của O. Ngữ nghĩa dữ liệu hay ngữ nghĩa là bộ $S = (r, \alpha)$ trong đó r là quan hệ trên U, có thể có giá trị null, và α là một ánh xạ từ r vào P(O)

$$\alpha : \ r \to P(O)$$

thỏa các điều kiện sau:

(i) $\forall t \in r, \ \forall o \in \alpha(t) \Rightarrow t \in$ infor(o)

(ii) $\forall t, \ s \in r, \ t \leq s \Rightarrow \alpha(t) \supseteq \alpha(s)$

Điều kiện (i) nói rằng $\alpha(t)$ là tập các đối tượng mà t là một thông tin về nó.

Điều kiện (ii) nói rằng nếu s là một htông tin về o, $t \leq s$ thì t cũng là một thông tin về o.

Cho r là một quan hệ trên $X \subset U$, bằng cách bỏ xung các giá trị null vào các thuộc tính $U \setminus X$ ta có thể xem r là một quan hệ trên U. Dpo vậy: từ đây về sau, ta chỉ xét các quan hệ trên tập thuộc tính vũ trụ U và O là tập các đối tượng mà ta quan sát.

Định nghĩa 3.2. Cho $S = (r, \alpha)$ và $S' = (r', \alpha')$ là hai ngữ nghĩa dữ liệu: Ta nói S ít thông tin hơn S' hay S' nhiều thông tin hơn S và viết $S \leq S'$ nếu $\forall t \in r, \forall o \in \alpha(t)$, tồn tại $t' \in r'$ sao cho $o \in \alpha'(t')$ và $t \leq t'$. Nếu $S \leq S'$ và $S' \leq S$ thì ta nói rằng S và S' có thông tin như nhau và viết $S \equiv S'$.

Cho r là một quan hệ trên U, ta định nghĩa α_r, như sau

$$\alpha_r(t) = \{o \in O : \ t \in \text{ infor(o) } \} \ \text{ với } \forall t \in r.$$

Đặt $S_r = (r, \alpha_r)$. Dễ dàng chứng minh rằng S_r là một ngữ nghĩa dữ liệu và với mọi ngữ nghĩa $S = (r, \alpha)$ ta luân có $S \leq S_r$.

Cho một ngữ nghĩa $S = (r, \alpha)$. Đặt

$$I_r = \{s : \ \exists t \in r, \ s \leq t, \ s \text{ là một bộ trên } \ U\}$$

ta xem I_r là tập các thông tin được suy dẫn từ r.

Xét ánh xạ $\delta: \ I_r \to P(O)$ trong đó

$$\delta(s) = \cup\{\alpha(t): \ t \in r, \ s \leq t\}, \ \forall s \in I_r.$$

Mệnh đề 3.1. *$S_e = (I_r, \delta)$ là một ngữ nghĩa dữ liệu và với hai ngữ nghĩa tùy ý $S = (r, \alpha)$ và $S' = (r', \alpha')$. Khi đó $S \leq S'$ nếu và chỉ nếu $S_e \leq S'_e$.*

Chứng minh Từ lưu ý $s, \ t \in I_r, \ s \leq t$ thì $\delta(s) \supseteq \delta(t)$, ta suy ra S_e là một ngữ nghĩa. Bây giờ giả sử $S \leq S'$. Cho $o \in \delta(s), \ s \in I_r$. Rõ ràng $\exists t \in r, \ s \leq t, \ o \in \alpha(t)$. $S \leq S'$ nên $\exists t' \in r', \ t \leq t', \ o \in \alpha'(t')$. Vậy $o \in \delta'(t'), \ s \leq t'$.

Ngược lại, giả sử $S_e \leq S'_e$. Cho $o \in \alpha(t), \ t \in r$. Ta có: $o \in \delta(t)$ nên $\exists s' \in I_{r'}, t \leq s', \ o \in \delta'(s')$. Do đó $\exists t' \in r', \ s' \leq t', \ o \in \alpha'(t')$ hay $t' \in r', \ t \leq t', \ o \in \alpha'(t')$. $\odot$.

Định nghĩa 3.3. Cho ngữ nghĩa $S = (r, \alpha)$. Một bộ $t \in r$ gọi là mang thông tin nếu $\alpha(t) \neq \cup\{\alpha(s): \ s \in r, \ t < s\}$.

Tập tất cả các bộ mang thông tin trong r được kí hiệu là r_I, α_I ký hiệu là hạn chế của α trên r_I: $\alpha_I = \alpha|_{r_I}$. Ta sẽ gọi $S_I = (r_I, \alpha_I)$ là ngữ nghĩa mang thông tin của S.

Từ định nghĩa suy ra, nếu t mang thông tin thì $\alpha(t) \neq \emptyset$.

Ta sẽ chứng minh rằng định nghĩa 3 là đúng đắn, nghĩa là S_I luân là một ngữ nghĩa. Hơn nữa ta có

Mệnh đề 3.2. *1. Với mọi ngữ nghĩa $S = (r, \alpha)$, $S_I = (r_I, \alpha_I)$ là một ngữ nghiã và ta có $S \equiv S_I$.*

2. Cho hai ngữ nghĩa $S = (r, \alpha)$ và $S' = (r', \alpha')$. Khi đó

(i) $S \leq S'$ nếu và chỉ nếu $S_I \leq S'_I$

(ii) $S \equiv S'$ nếu và chỉ nếu $S_I = S'_I$ nghĩa là $r_I = r'_I, \ \alpha_I = \alpha'_I$.

Chứng minh. 1. Dễ dàng kiểm tra S_I là một ngữ nghĩa.

Hiển nhiên $S_I \leq S$, do đó ta chỉ cần chứng tỏ $S \leq S_I$. Giả sủ $o \in \alpha(t), \ t \in r$. Ta cần chỉ ra rằng $\exists t_I \in r_I, \ t \leq t_I, \ o \in \alpha_I(t)$. Thật vậy, nếu $t \in r_I$, lấy $t_I = t$. Nếu $t \in r \setminus r_I$ điều đó có nghĩa là $\alpha(t) = \cup\{\alpha(s): \ s \in r, \ t < s\}$. Do đó $\exists s_1 \in r, \ t < s_1, \ o \in \alpha(s_1)$. Nếu $s_1 \in r_I$, lấy $t_I = s_1$. Nếu $s_1 \in r \setminus r_I$, khi đó $\exists s_2 \in r, \ s_1 < s_2, \ o \in \alpha(s_2)$... Tiếp tục ta có dãy $t < s_1 < s_2 < ... < s_i < ... < ...$ với $s_i \in r \setminus r_I, \ o \in \alpha(s_i)$. Vì độ dài của dãy này không vượt quá số thuộc tính trong U nên $\exists s_j \in r_I, \ t < s_j, \ o \in \alpha(s_j)$. Lấy $t_I = s_j$.

2) (i) Giả sử $S \leq S'$, theo 1. ta có $S_I \leq S \leq S' \leq S'_I$. Ngược lại nếu $S_I \leq S'_I$ theo 1. ta có $S \leq S_I \leq S'_I \leq S'$

(ii) Nếu $S_I = S'_I$, ta có $S \equiv S_I = S'_I \equiv S'$ hay $S \equiv S'$.

Bây giờ giả sử $S \equiv S'$, tức là theo 1. $S_I \equiv S'_I$.

Đầu tiên ta chứng minh $r_I = r'_I$. Cho $t \in r_I$. Đặt $\beta(t) = \cup\{\alpha(s) : s \in r_I,\ t < s\}$, $\gamma(t) = \alpha(t) - \beta(t)$. Do $\gamma(t) \neq \emptyset$, lấy $o \in \gamma(t)$. Từ $S_I \leq S'_I$ suy ra $\exists t' \in r'_I,\ t \leq t',\ o \in \alpha_I(t')$. Từ $S'_I \leq S_I$ suy ra $\exists s \in r_I,\ t' \leq s,\ o \in \alpha_I(s)$. Như vậy $t \leq t' \leq s,\ o \in \alpha_I(s)$. Do $o \notin \beta(t)$ suy ra $t = t' = s$ hay $t \in r'_I$. Tóm lại $r_I \subseteq r'_I$. Tương tự $r_I \supseteq r'_I$ và ta có $r_I = r'_I$.

Phần còn lại là chứng tỏ $\alpha_I(t) = \alpha'_I(t),\ \forall t \in r_I$. Đặt $\beta_I(t) = \cup\{\alpha_I(s) :\ s \in r_I,\ t < s\}$, $\beta_I(t) \subseteq \alpha_I(t)$. Cho $o \in \alpha_I(t)$. Có hai trường hợp. Nếu $o \notin \beta_I(t)$ tức là $o \in \gamma_I(t) = \alpha_I(t) - \beta_I(t)$. Lập luận tương tự như trên ta có $o \in \alpha'_I(t)$. Nếu $o \in \beta_I(t)$, như vậy $\exists s_1 \in r_I,\ t < s_1,\ o \in \alpha_I(s_1)$. Xét tương quan giữa o và $\beta_I(s_1)$. Nếu $o \notin \beta_I(s_1)$, lập luận như trên ta có $0 \in \alpha'_I(s_1) \subseteq \alpha'_I(t)$ hay $o \in \alpha'_I(t)$. Nếu $o \in \beta_I(s_1)$ thì $\exists s_2 \in r_I,\ s_1 < s_2,\ o \in \alpha_I(s_2)$... Tiếp tục ta có dãy $t < s_1 < s_2 < ... < s_i < ...$ trong đó $s_i \in r_I,\ o \in \beta_I(s_i),\ \forall i$. Độ dài của dãy là hữu hạn nên $\exists s_j \in r_I,\ o \in \gamma_I(s_j)$. Từ đây suy ra $o \in \alpha'_I(t)$.

Trong mọi trường hợp ta đều chứng minh được $o \in \alpha'_I(t)$. Như vậy $\alpha_I(t) \subseteq \alpha'_I(t)$.

Bao hàm ngược lại chứng minh tương tự và $\alpha_I(t) = \alpha'_I(t),\ \forall t \in r_I = r'_I$. ⊙

Quan hệ $\equiv$ giữa các ngữ nghĩa có tính chất phản xạ, bắc cầu nhưng không có tính chất phản xứng. Mệnh đề trên phần (ii), chứng tỏtính phản xứng chỉ xảy ra với các ngữ nghĩa mang thông tin mà thôi.

Tiếp theo chúng ta hãy mở rộng ánh xạ α trên tập P(r)- là tập các tập con của r, như sau

$$\alpha : P(r) \to P(O),\ \forall P \in P(r),\ \alpha(P) = \cap\{\alpha(t) : t \in P\},\ \text{đặc biệt } \alpha(\emptyset) = O.$$

Như vậy, $\alpha(P)$ là tập các đối tượng trong O mà P là thông tin của nó.

Từ định nghĩa suy ra $\alpha(P \cup Q) = \alpha(P) \cap \alpha(Q)$ với $\forall P,\ Q \in P(r)$. Đặt $A = \{\alpha(P) :\ P \in P(r)\}$. Các phần tử trong A được kí hiệu là $a,\ b,\ c,...$ Với mọi $a \in A$, ta kí hiệu

$$a^0 = \cup\{b :\ b \in A; b \subset a\}$$
$$a^- = a - a^0,$$
$$r(a) = \cup\{P \in P(r) :\ \alpha(P) = a\}.$$

Với $o \in O$, kí hiệu $r(o) = \{t \in r :\ o \in \alpha(t)\}$.

Ta thấy rằng $r(a)$ là tập lớn nhất trong các tập $P \in P(r)$ có tính chất $\alpha(P) = a$ vì $\alpha(r(a)) = a$ và nếu $\alpha(P) = a$ thì $P \subseteq r(a)$.

$r(o) \subseteq r$ chính là thông tin của đối tượng o mà ta có trong quan hệ r, các thông tin đó được biểu diễn bằng các bộ. Nói chung $r(\{o\}) \subseteq r(o)$.

Định lý 3.1. *(i) $A = (O, \cap)$ là nửa dàn*

Cho $a,\ b \in A$. Khi đó $a \subseteq b$ nếu và chỉ nếu $r(a) \supseteq r(b)$ như vậy $a = b$ nếu và chỉ nếu $r(a) = r(b)$.

(ii) $\forall a \in A,\ a^- \neq \emptyset,\ \forall o \in a^-$ ta có $r(a) = r(o)$.

Chứng minh. (i) Phần đầu kiểm tra tương đối dễ dàng.

Bây giờ giả sử $a \subseteq b$. khi đó $a = a \cap b = \alpha(r(a)) \cap \alpha(r(b)) = \alpha(r(a) \cup r(b))$. Do nhận xét ở trên $r(a)$ là tập lớn nhất trong các tập P có tính chất $\alpha(P) = a$, ta suy ra $r(a) \cup r(b) \subseteq r(a)$ hay $r(b) \subseteq r(a)$.

Điều ngược lại suy ra từ: $r(b) \subseteq r(a) \Rightarrow \alpha(r(b)) \supseteq \alpha(r(a))$ hay $b \supseteq a$. Phần còn lại suy ra trực tiếp từ phần trên.

(ii) Rõ ràng $o \in a$, suy ra $r(a) \subseteq r(o)$

Ta chứng minh $r(a) \supseteq r(o)$. Cho $t \in r(o)$.

Từ $o \in \alpha(t)$ và $o \in a = \alpha(r(a))$ ta có $o \in \alpha(t) \cap \alpha(r(a) = \alpha(t \cup r(a) \subseteq \alpha(r(a)) = a$. Vì $o \notin a^0 = \{\alpha(P) :\ \alpha(P) \subset a\}$ nên $\alpha(t \cup r(a)) = a$. Như vậy $t \cup r(a) \subseteq r(a)$ hay $t \in r(a)$. Kết luận rằng $t \in r(a)$. $\odot$

Cho $S = (r, \alpha)$ là một ngữ nghĩa. Thực hiện một phân lớp trên O như sau: $o,\ o' \in O$ thuộc cùng một lớp nếu $r(o) = r(o')$, nói một cách khác, hai đối tượng thuộc cùng một lớp nếu thông tin của chúng trong r là như nhau. Một lớp đặc biệt trong sự phân lớp trên là lớp những đối tượng o mà $r(o) = \emptyset$, tức là ta không biết thông tin gì về chúng. Dễ dàng kiểm tra phân lớp trên là một phân hoạch trên O. Kí hiệu $O(S)$ là phân hoạch trên và nói rằng ngữ nghĩa S thực hiện một phân hoạch trên O.

Định lý 3.2. *Cho hai ngữ nghĩa $S = (r, \alpha)$ và $S' = (r', \alpha')$ có thông tin như nhau, khi đó chúng thực hiện cùng một phân hoạch trên O tức là*

$$O(S) = O(S').$$

Chứng minh. Do tính đối xứng ta chỉ cần chứng minh $r(o) = r(o')$ kéo theo $r'(o) = r'(o')$ với $\forall o,\ o' \in O$.

Xét hai trường hợp

1) $r(o) = r(o') = \emptyset$

Khi đó $r'(o) = \emptyset$. Thật vậy, nếu $r'(o) \neq \emptyset$, lấy $t' \in r'(o)$ tức là $t' \in r$ và $o \in \alpha'(t')$, $S \leq S'$ nên $\exists t \in r,\ t' \leq t,\ o \in \alpha(t)$ tức là $r(o) \neq \emptyset$. Điều này mâu thuẫn với 1).

Tương tự $r'(o') = \emptyset$

2.) $r(o) = r(o') \neq \emptyset$

Thero mệnh đề 2, S và S' có cùng một ngữ nghĩa mang thông tin nên $S_I = S'_I = (r_I, \alpha_I)$ trong đó $r_I = \{t \in r : t \text{ mang thông tin }\}$, $\alpha_I = \alpha|r_I = \alpha'|r_I$, $r_I \subseteq r \cap r'$.

Ta có $r'(o) \neq \emptyset$. Thật vậy, $\exists s \in r(o)$, $S \leq S'$ do đó $\exists s' \in r'$, $s \leq s'$, $s' \in r'(o)$. Vậy $r'(o) \neq \emptyset$.

Cho $t' \in r(o)$. Có hai trường hợp

(i) $t' \in r_I$. Ta có $o \in \alpha'(t')$, $\alpha'(t') = \alpha(t') = \alpha_I(t') \Rightarrow o \in \alpha(t')$. Kết hợp với giả thiết $r(o) = r(o')$ ta suy ra $t' \in r(o')$. Vì $\alpha(t') = \alpha'(t')$ nên $o' \in \alpha'(t')$. Do đó $t' \in r'(o')$.

(ii) $t' \in r' \setminus r_I$. Như vậy $o \in \alpha'(t') = \cup\{\alpha'(s') : s' \in r', t' < s'\}$ tức là $\exists s'_1 \in r', t' < s'_1$, $o \in \alpha'(s'_1)$. Nếu $s'_1 \in r_I$, theo (i) $s'_1 \in r'(o')$, do đó $t' \in r'(o')$. Nếu $s'_1 \in r' \setminus r_I$, khi đó $\exists s'_2 \in r', s'_1 < s'_2$, $o \in \alpha'(s'_2)$... Tiếp tục ta xây dựng được dãy $t' < s'_1 < ... < s'_i < ...$ trong đó $s'_i \in r' \setminus r_I$, $o \in \alpha'(s'_i)$ với $\forall i$. Độ dài của dãy này không vượt quá số thuộc tính trong U, nên sẽ $\exists s'_j \in r_I$, $o \in \alpha'(s'_j)$, $t' < s'_j$. Lập luận như trong i) ta có $s'_j \in r'(o')$ và do đó $t' \in r'(o')$.

Trong mọi trường hợp ta đều chứng tỏ được $t' \in r'(o')$. nghĩa là $r'(o) \subseteq r'(o')$. Tương tự thu được $r'(o) \supseteq r'(o')$. Như vậy $r'(o) = r'(o')$. $\odot$

Chúng ta thường biểu diễn thông tin về thế giới thực bằng các bộ trong 1 quan hệ. Vấn đề nảy sinh khi chúng ta phải biểu diễn thông tin không đầy đủ bằng các bộ. Trong trường hợp này một vài dữ liệu trở thành null. Việc chúng ta định nghĩa giá trị null như thế nào, có liên quan mật thiết đến các vấn đề khác trong CSDL. Ngữ nghĩa của các phép toán quan hệ và phụ thuộc dữ liệu hoàn toàn phụ thuộc vào việc định nghĩa giá trị null trong CSDL: (xem [4]-[6]). Phân tích các cách tiếp cận đến giá trị null chúng ta thấy rằng, ngữ nghĩa của các giá trị null và tương tác giữa các phụ thuộc dữ liệu với null được thảo luận độc lập với thế giới thực. Mặt khác ngữ nghĩa của dữ liệu, theo nghĩa tự nhiên là một phép gán, các đối tượng trong thế giới thực với dữ liệu (là tập của các dãy kí hiệu). Phép gán này thể hiện trí thức của chúng ta về thế giới thực. Do đó ta xem xét ngữ nghiã dữ liệu như một cặp $S = (r, \alpha)$ trong đó r là một quan hệ và α là một phép gán. Mỗi ngữ nghiã S xác định một phân hoạch trên tập các đối tượng của thế giới thực O. Phân hoạch này thể hiện trí thức của chúng ta về O. Định lý 3.2 khẳng định rằng hai ngữ nghĩa có cùng thông tin, thực hiện cùng một phân hoạch trên O. Điều này phù hợp với cảm nhận trực giác của chúng ta.

Sau đây ta xét một ví dụ.

Ví dụ. Xem xét tập thuộc tính vũ trụ.

U = { TEN, KHOA, MON-HOC, KHOA HOC }

trong đó TEN là tên sinh viên, KHOA là khoa anh ta theo học, MON-HOC là môn học của anh ta, còn KHOA-HOC là khoá học. $S = (r, \alpha)$ là một ngữ nghĩa được đưa ra như sau:

t_1 = (Nam , Tinhoc, $\perp$, $\perp$) $\rightarrow o_2,\ o_3$
t_2 = (Nam , Tinhoc, CSDL, $\perp$) $\rightarrow o_2$
t_3 = (Nam , Tinhoc, Laptrinh, $\perp$) $\rightarrow o_3$.
t_4 = (Nam , $\perp$, Ngon ngu hinh thuc, $\perp$) $\rightarrow o_4,\ o_2$
t_5 = (Huy , Luật, Kinh tế thế giới, $\perp$) $\rightarrow o_5$

$$\alpha(t_1) = \{o_2, o_3\}, \alpha(t_2) = \{o_2\},\ \alpha(t_3) = \{o_3\},\ \alpha(t_4) = \{o_2, o_4\},\ \alpha(t_5) = \{o_5\},.$$

Quan hệ r gồm 5 bộ trên

Khi đó ta có

$$r(o_2) = \{t_1, t_2, t_4\}$$
$$r(o_3) = \{t_1, t_3\}$$
$$r(o_4) = \{t_4\}$$
$$r(o_5) = \{t_5\}$$

t_1 là bộ không mang thông tin vì $\alpha(t_1) = \alpha(t_2) \cup \alpha(t_3)$ và $t_1 < t_2,\ t_1 < t_3$; $t_2,\ t_3,\ t_4,\ t_5$ là các bộ mang thông tin.

Tiếp theo ta sẽ chứng tỏ rằng mỗi ngữ nghĩa $S = (r, \alpha)$ có thể được đặc trưng bởi một họ các tập con của r. Mỗi đối tượng $o \in O$ mà ta quan sát được gán với một tập $m(o) \subseteq r$ biểu diễn thông tin về đối tượng đó. Những đối tượng không có thông tin được gán đến tập rỗng. Như vậy một cách khác để biểu diễn ngữ nghĩa dữ liệu là đưa ra một ánh xạ từ O vào $P(r)$. Định lý sau sẽ làm sáng tỏ điều đó.

Định lý 3.3 *Cho O là một tập đối tượng, U là tập thuộc tính vũ trụ, r là một quan hệ (có thể có null) trên U, m là một ánh xạ từ O vào $P(r)$.*

m: $O \rightarrow P(r)$ thỏa mãn điều kiện sau:
(i) $\forall o \in O,\ m(o) \subseteq infor(o)$
(ii) Nếu $t \in m(o),\ s \in r,\ s \leq t$ thì $s \in m(o)$
Khi đó, tồn tại duy nhất một ngữ nghĩa $S = (r, \alpha)$ thỏa điều kiện $m(o) = r(o)$

Chứng minh. Lấy $\alpha(t) = \{o \in O,\ t \in m(o)\}$.

Do hai tính chất của m(o), ta dễ dàng kiểm tra để chứng tỏ $S = (r, \alpha)$ là một ngữ nghĩa. Hiển nhiên do cách chọn α của ta, ta có: $r(o) = m(o)$.

Nếu $S' = (r, \alpha')$ là một ngữ nghĩa thỏa điều kiện $r_1(o) = m(o)$ trong đó $r_1(o) = \{t \in r : o \in \alpha'(t)\}$. Ta có $\alpha = \alpha'$. Do đó $S = S'$. ⊙

IV. – KẾT LUẬN

CSDL với thông tin không đầy đủ phản ánh quá trình nhận thức tất yếu của chúng ta về thế giới thực. Chúng ta nắm bắt thông tin của thế giới thực từ ít tới nhiều, từ đơn giản đến phức tạp, từ chưa hoàn tiện đến hoàn thiện. Trong bài báo này chúng ta đã xét ngữ nghĩa của dữ liệu trong CSDL với thông tin không đầy đủ như là xét mối quan hệ của CSDL với thế giới thực. Làm sáng tỏ mối quan hệ đó bằng việc hình thức hoá ngữ nghĩa dữ liệu là cặp $S = (r, \alpha)$ trong đó r là một quan hệ có thể chứa null và α là một phép gán mỗi cặp trong r với tập các đối tượng trong thế giới thực mà nó phản ánh. Chứng minh hình thức một số cảm nhận trực giác của chúng ta về mối quan hệ đó.

TÀI LIỆU THAM KHẢO

1. J.D. Ullman, Priciples of database systems, 2nd ed, Computer science Press, 1982.
2. W. Lipski, On semantic issues connected with incomplete information databases, ACM Trans. Database sys. 4,3, 1979.
3. W. Lipski, On databases with incomplete information, ACM Trans. Database sys. 5, 6, 1981.
4. N.C. Ho, A relational model of databases with context dependent null values, Bull. Pol. Ac. Tech., Vol. 36, N.1-2, 1988.
5. N.C. Ho, Context dependent null values and multivalued dependencies in relational databases, Bull. Pol. Ac. Tech., Vol. 36, N.1-2, 1988.
6. D. Maier, The theory of relational databases, Computer Science, Press, Inc., Rockville, 1983.
7. AM. Keller, M.W. Wilkins, On the use of an extended relational model to handle change incomplete information, IEEE Trans. on Software engineering, Vol. SE-11, N. 7. July,1985.

Trung tâm Khoa học tự nhiên
và Công nghệ Quốc gia
Viện Công nghệ thông tin

Bộ nội vụ