Tạp chí Tin học và Điều khiển học, T.16, S.2 (2000), 41-50

THUẬT TOÁN TỔNG HỢP LƯỢC ĐỒ CƠ SỞ DỮ LIỆU QUAN HỆ DẠNG CHUẨN BA

PHẠM QUANG TRUNG, NGUYỄN XUÂN HUY

Abstract. Designing databases requires translating a relational scheme into a normal form. In this paper we shall present a new algorithm, which synthesizes a relation scheme into third normal form (3NF) with a lossless join and preservation of dependencies. The algorithm is constructed on a notation of compound functional dependencies. In comparison with the well-known algorithms, the algorithm presented here is more simply in implementation.

1. MỞ ĐẦU

Chuẩn hóa là quá trình phân tích một lược đồ quan hệ thành một tập các lược đồ quan hệ phù hợp để có thể tránh được những vấn đề thường nảy sinh khi thiết kế cơ sở dữ liệu [3,4]: 1) Dư thừa dữ liệu; 2) Không nhất quán; 3) Dị thường khi thêm bộ 4) Dị thường khi xóa bộ; 5) Dị thường khi sửa bộ.

Có các dạng chuẩn là: Dạng chuẩn thứ nhất (first normal form - 1NF), dạng chuẩn thứ hai (second normal form - 2NF), dạng chuẩn thứ ba (third normal form - 3NF),...

Trong bài báo này, chúng tôi trình bày một thuật toán tổng hợp lược đồ cơ sở dữ liệu thành dạng chuẩn ba (3NF), thuận tiện trong việc cài đặt và đạt được các yêu cầu sau đây:

- Bảo toàn tập phụ thuộc hàm.
- Các lược đồ con là ở dạng chuẩn 3NF.
- Kết nối không mất thông tin.
- Không tồn tại các lược đồ nào khác có số lượng lược đồ ít hơn mà lại có ba tính chất nêu trên.
- Ngoài ra, thuật toán được thiết kế có sử dụng khái niệm về *phụ thuộc hàm phức hợp*, *phủ dạng hình khuyên* (hay *phủ dạng vành*), những thuật toán cơ sở (xác định bao đóng, loại bỏ thuộc tính dư thừa, loại bỏ phụ thuộc hàm dư thừa,...) được xây dựng trên khái niệm *phụ thuộc hàm*.

2. MỘT SỐ KHÁI NIỆM CƠ BẢN

Các ký hiệu

Theo truyền thống trình bày của lý thuyết cơ sở dữ liệu quan hệ, trong bài báo có sử dụng các ký hiệu sau:

- Quan hệ R trên tập thuộc tính U được ký hiệu là $R(U)$.
- Hợp của hai tập thuộc tính X, Y được viết là XY.
- Với quan hệ R trên tập thuộc tính U, t là một bộ của R, X là tập con của U thì giá trị của t trên X được viết là $t(X)$.
- Phép kết nối tự nhiên được ký hiệu bằng dấu *.

Mục này chỉ nêu một số khái niệm và kết quả liên quan, bạn đọc quan tâm chi tiết hơn đề nghị xem [5,6].

Định nghĩa 1. Cho $R(A_1, A_2, ..., A_n)$ là một lược đồ quan hệ, cho X và Y là các tập con của $\{A_1, A_2, ..., A_n\}$. Chúng ta nói $X \to Y$ (đọc là "X xác định hàm Y" hay "Y phụ thuộc hàm vào X") nếu với mọi quan hệ r là thể hiện của R, thì trong r không thể có hai bộ trùng nhau trên các thành phần của mọi thuộc tính trong tập X mà lại không trùng nhau trên một hay nhiều hơn các thành phần của các thuộc tính của tập hợp Y.

- Chúng ta nói rằng: một quan hệ r *thỏa* phụ thuộc hàm $X \to Y$, nếu cho mọi cặp bộ μ, ν trong r sao cho $\mu[X] = \nu[X]$ thì $\mu[Y] = \nu[Y]$ cũng đúng. Nếu r không thỏa $X \to Y$, thì r vi phạm phụ thuộc đó.

- Cho F là tập phụ thuộc hàm của lược đồ quan hệ R, và cho $X \to Y$ là một phụ thuộc hàm. Chúng ta nói F *suy diễn logic* ra $X \to Y$, viết là $F \models X \to Y$, nếu với mọi quan hệ r của R mà thỏa các phụ thuộc hàm trong F thì cũng thỏa $X \to Y$.

Định nghĩa 2. Bao đóng của tập phụ thuộc hàm F, ký hiệu là F^+, là tập các phụ thuộc hàm được suy diễn logic từ F [1, 2], nghĩa là:

$$F^+ = \{X \to Y \mid F \models X \to Y\}.$$

Định nghĩa 3. Hai tập phụ thuộc hàm F và G trên lược đồ R là *tương đương*, ký hiệu là $F \equiv G$, nếu: $F^+ = G^+$. Nếu $F \equiv G$ thì F là một *phủ* của G.

Định nghĩa 4. Tập phụ thuộc hàm F là *không dư thừa* nếu không có một tập con thực sự F' của F mà $F' \equiv F$. Nếu tồn tại một tập con F' như vậy thì F là *dư thừa*. F được gọi là *phủ không dư* của G nếu F là một phủ của G và F không dư thừa.

Định nghĩa 5. Hai tập thuộc tính X và Y là *tương đương* với nhau trên tập phụ thuộc hàm F, nếu $F \models X \to Y$ và $F \models Y \to X$ (ký hiệu là $X \leftrightarrow Y$).

Giả sử F là tập phụ thuộc hàm trên lược đồ R và tập thuộc tính $X \subseteq R$, giả sử $E_F(X)$ là tập phụ thuộc hàm trong F có các vế trái tương đương với X. Ký hiệu $\overline{E}_F$ là tập hợp:

$$\{E_F(X) \mid X \subseteq R \quad \text{và} \quad E_F(X) \neq \emptyset\}.$$

Nếu trong F không tồn tại phụ thuộc hàm có vế trái tương đương với X thì $E_F(X)$ rỗng. Tập $\overline{E}_F(X)$ là một phân hoạch của tập F [5].

Định nghĩa 6. Giả sử F là tập phụ thuộc hàm trên lược đồ R và $X \to Y$ là phụ thuộc hàm trong F. Phụ thuộc hàm $X \to Y$ được gọi là *rút gọn trái* nếu X không có thuộc tính dư thừa. Phụ thuộc hàm $X \to Y$ được là *rút gọn phải* nếu Y không chứa thuộc tính dư thừa. Phụ thuộc hàm $X \to Y$ được gọi là *rút gọn* nếu nó được rút gọn trái và rút gọn phải và $Y \neq \emptyset$.

Định nghĩa 7. Tập phụ thuộc hàm F được gọi là *rút gọn trái* (rút gọn phải, rút gọn) nếu mỗi phụ thuộc hàm trong F là rút gọn trái (tương ứng là rút gọn phải, rút gọn).

Định nghĩa 8. *Phụ thuộc hàm phức hợp* có dạng $(X_1, X_2, ..., X_k) \to Y$, trong đó $X_1, X_2, ..., X_k$ và Y là các tập con khác nhau của lược đồ R. Quan hệ $r(R)$ thỏa phụ thuộc hàm phức hợp $(X_1, X_2, ..., X_k) \to Y$ nếu nó thỏa các phụ thuộc hàm $X_i \to X_j$ và $X_i \to Y$, $1 \leq i, j \leq k$. Trong phụ thuộc hàm phức này, $(X_1, X_2, ..., X_k)$ được gọi là vế trái, $X_1, X_2, ..., X_k$ là các tập trái, Y là vế phải.

Phụ thuộc hàm phức hợp là cách viết rút gọn hơn tập các phụ thuộc hàm có các vế trái tương đương. Trong trường hợp nếu $Y = \emptyset$, chúng ta có dạng đặc biệt của phụ thuộc hàm phức hợp là $(X_1, X_2, ..., X_k)$.

Định nghĩa 9. Giả sử G là tập các phụ thuộc hàm phức hợp trên R và F là tập các phụ thuộc hàm hay các phụ thuộc hàm phức hợp trên R. Tập G *tương đương* với F, ký hiệu là $G \equiv F$, nếu mỗi quan hệ $r(R)$ thỏa G thì thỏa F và ngược lại.

Định nghĩa 10. Tập F được gọi là *phủ* của G nếu $F \equiv G$, trong đó F và G bao gồm hoặc là tập phụ thuộc hàm, tập các phụ thuộc hàm phức hợp, hoặc là tập hợp chỉ gồm một loại phụ thuộc.

Định nghĩa 11. Tập phụ thuộc hàm F được gọi là *tập đặc trưng* đối với phụ thuộc hàm phức hợp $(X_1, X_2, ..., X_k) \rightarrow Y$ nếu $F \equiv \{(X_1, X_2, ..., X_k) \rightarrow Y\}$. Nếu mỗi tập hợp trái của phụ thuộc hàm phức được sử dụng với tư cách là vế trái của phụ thuộc hàm đúng một lần (nghĩa là F có dạng $\{X_1 \rightarrow Y_1, X_2 \rightarrow Y_2, ..., X_k \rightarrow Y_k\}$), thì F được gọi là *tập phụ thuộc hàm đặc trưng tự nhiên* đối với phụ thuộc hàm phức hợp đã cho.

Định nghĩa 12. Tập phụ thuộc hàm phức hợp F được gọi là *dạng hình khuyên* (hay dạng vành), nếu không có các tập trái X và Z trong các vế trái khác nhau, mà $X \leftrightarrow Z$ trên F.

Bổ đề 1. [5] *Giả sử G là tập phụ thuộc hàm phức hợp dạng vành không dư thừa. Sự hợp nhất các tập đặc trưng tự nhiên của tất cả các phụ thuộc hàm phức hợp trong G tạo thành tập phụ thuộc hàm không dư tương đương với G.*

Chuẩn hóa bằng phép tách [6]

– *Phép tách một lược đồ quan hệ* là việc thay thế một lược đồ R bằng tập các lược đồ con $\rho = \{R_1, R_2, ..., R_k\}$ (các R_i không nhất thiết phải rời nhau) sao cho:

- $R_i \subseteq R$, $i = 1, 2, ..., k$.
- $R = R_1 \cup R_2 \cup ... \cup R_k$.

– Cho lược đồ R được tách thành tập các lược đồ con: $\rho = \{R_1, R_2, ..., R_k\}$. Phép tách ρ là *phép tách có kết nối không mất thông tin* nếu với mọi quan hệ r trên R ta có:

$$r = \prod_{R_1}(r) * \prod_{R_2}(r) * ... * \prod_{R_k}(r)$$

tức là mọi quan hệ r là kết nối tự nhiên của các hình chiếu của nó trên các R_i.

– Cho lược đồ R và tập phụ thuộc hàm F. Phép tách $\rho = \{R_1, R_2, ..., R_k\}$ gọi là *phép tách bảo toàn F*, nếu: $G = F_1 \cup F_2 \cup ... \cup F_k$ suy dẫn ra F (trong đó: $F_i = \{X \rightarrow Y \in F^+ \mid X, Y \subseteq R_i\}$).

Định lý 1. [6] *Nếu $\rho = (R_1, R_2)$ là một phép tách của R và F là tập các phụ thuộc hàm thì ρ là phép tách có kết nối không mất thông tin đối với F khi và chỉ khi: $(R_1 \cap R_2) \rightarrow (R_1 - R_2)$ hoặc $(R_1 \cap R_2) \rightarrow (R_2 - R_1)$. Chú ý rằng các phụ thuộc hàm nêu trên không nhất thiết phải thuộc tập F, nhưng phải thuộc F^+.*

Định lý 2. [6] *Cho ρ là phép tách lược đồ R thành 3NF bảo toàn các phụ thuộc hàm, và cho X là một khóa đối với R. Thế thì $\tau = \rho \cup \{X\}$ là phép tách R thành các lược đồ ở 3NF, phép tách này bảo toàn các phụ thuộc hàm và có tính chất kết nối không mất thông tin.*

Chuẩn hóa bằng phép tổng hợp [5]

Cho trước lược đồ quan hệ R và tập phụ thuộc hàm F trên R. Yêu cầu là cần nhận được các lược đồ cơ sở dữ liệu $\mathcal{R} = \{R_1, R_2, ..., R_k\}$ trên R thỏa mãn bốn tính chất sau đây:

1) Tập F được đặc trưng đầy đủ bởi $\mathcal{R}$, nghĩa là:
$F \equiv \{K \rightarrow R_i | R_i \in \mathcal{R}$ là khóa được chỉ định của $R_i\}$.

2) Mỗi lược đồ quan hệ R_i trong $\mathcal{R}$ là ở 3NF với sự tham chiếu tới F.

3) Không tồn tại các lược đồ cơ sở dữ liệu nào khác có số lượng lược đồ ít hơn $\mathcal{R}$, mà lại có tính chất 1 và 2.

4) Đối với bất kỳ quan hệ $r(R)$ thỏa F đều có

$$r = \prod_{R_1}(r) * \prod_{R_2}(r) * \ldots * \prod_{R_k}(r).$$

Lược đồ cơ sở dữ liệu R thỏa mãn ba tính chất đầu được gọi là lược đồ cơ sở dữ liệu *đầy đủ* (complete) đối với F.

Tính chất 1 đảm bảo bắt buộc F trong $\mathcal{R}$, điều này có thể kiểm chứng được mà không cần tính F^+. Tính chất 1 còn đảm bảo rằng các phụ thuộc hàm được suy dẫn từ các khóa được chỉ định.

Yêu cầu của tính chất 2 là đã rõ. Tính chất 3 để tránh sự dư thừa. Tính chất 4 cho phép biểu diễn một quan hệ trên lược đồ R trung thành với các hình chiếu của nó trên các lược đồ trong $\mathcal{R}$.

Các thuật toán tổng hợp có một số tính chất sau đây.

Bổ đề 2. *Nếu $\mathcal{R}$ là lược đồ cơ sở dữ liệu biểu diễn tập phụ thuộc hàm G, thì trong $\mathcal{R}$ có ít nhất $|\overline{E}_G|$ lược đồ quan hệ. Có nghĩa là số lược đồ ít nhất có trong $\mathcal{R}$ bằng số lượng lớp tương đương trong $\overline{E}_G$.*

Hệ quả 1. *Giả sử F phụ thuộc hàm. Một lược đồ cơ sở dữ liệu $\mathcal{R}$ bất kỳ đặc trưng đầy đủ F, cần phải có ít nhất $|\overline{E}_F|$ lược đồ, trong đó F' là phủ không dư đối với F.*

3. THUẬT TOÁN TỔNG HỢP LƯỢC ĐỒ CƠ SỞ DỮ LIỆU

Thuật toán tổng hợp lược đồ cơ sở dữ liệu thành dạng 3NF: TH-3NF

- *Vào:* Tập U, tập phụ thuộc hàm F trên U.
- *Ra:* Tập lược đồ quan hệ ở dạng chuẩn ba, bảo toàn F, kết nối không mất thông tin, có số lượng lược đồ là ít nhất.
- *Phương pháp:*

- Bổ sung phụ thuộc hàm $U \to @$ vào tập phụ thuộc hàm F (trong đó @ là tên "thuộc tính giả" mà $@ \notin U$).
- Rút gọn vế trái của các phụ thuộc hàm.
- Loại bỏ các phụ thuộc hàm dư thừa. Kết quả của bước này nhận được tập F'.
- Tạo tập phủ dạng vành. Kết quả của bước này nhận được tập G.
- Tạo tập phụ thuộc hàm đặc trưng tự nhiên tương đương với tập phụ thuộc hàm phức hợp đã xây dựng, nhận được tập G_1. Rút gọn vế phải của các phụ thuộc hàm trong G_1. Kết quả của bước này nhận được tập G_2.
- Tạo tập phủ dạng vành G_3.
- Tương ứng với từng phụ thuộc hàm phức hợp đã xây dựng trong G_3, xây dựng các lược đồ quan hệ có tập thuộc tính là tất cả các thuộc tính xuất hiện trong mỗi phụ thuộc hàm phức hợp, tập các khóa được chỉ định của lược đồ là các tập trái của phụ thuộc hàm phức hợp. Thuộc tính giả @ được loại khỏi lược đồ mà nó tham gia.

Thí dụ 1. Cho lược đồ quan hệ gồm tập các thuộc tính $U = \{A, B_1, B_2, C_1, C_2, D_1, D_2\}$, tập các phụ thuộc hàm F:

$$
\begin{aligned}
\{\, & B_1\, B_2 \to A \\
& D_1\, D_2 \to A\, B_1\, B_2 \\
& B_1 \to C_1 \\
& B_2 \to C_2 \\
& D_1 \to A \\
& D_2 \to A \\
& A\, B_1\, C_2 \to D_2 \\
& B_1\, D_1 \to A\, C_1 \\
& A\, B_2\, C_1 \to D_1 \,\}
\end{aligned}
$$

Sau khi bổ sung phụ thuộc hàm $AB_1B_2C_1C_2D_1D_2 \to @$ vào, thực hiện rút gọn vế trái và loại bỏ phụ thuộc hàm dư thừa, chúng ta có tập F':

$$
\begin{aligned}
\{\, & B_1\, B_2 \to A \\
& D_1\, D_2 \to A\, B_1\, B_2 \\
& B_1 \to C_1 \\
& B_2 \to C_2 \\
& D_1 \to A \\
& D_2 \to A \\
& A\, B_1\, C_2 \to D_2 \\
& A\, B_2\, C_1 \to D_1 \\
& B_1\, B_2 \to @\,\}
\end{aligned}
$$

Từ đó chúng ta có phủ dạng vành G:

$$
\begin{aligned}
\{\, & (B_1\, B_2,\, D_1\, D_2) \to A\, @ \\
& (B_1) \to C_1 \\
& (B_2) \to C_2 \\
& (D_1) \to A \\
& (D_2) \to A \\
& (A\, B_1\, C_2) \to D_2 \\
& (A\, B_2\, C_1) \to D_1\,\}
\end{aligned}
$$

Tạo tập phụ thuộc hàm đặc trưng tự nhiên tương đương với G:

$$
\begin{aligned}
G_1 = \{\, & B_1\, B_2 \to A\, D_1\, D_2 @ \\
& D_1\, D_2 \to A\, B_1\, B_2\, @ \\
& B_1 \to C_1 \\
& B_2 \to C_2 \\
& D_1 \to A \\
& D_2 \to A \\
& A\, B_1\, C_2 \to D_2 \\
& A\, B_2\, C_1 \to D_1\,\}
\end{aligned}
$$

Loại bỏ các thuộc tính dư thừa vế phải trong G_1 chúng ta nhận được tập G_2:

$$
\begin{aligned}
G_2 = \{\, & B_1\, B_2 \to D_1\, D_2\, @ \\
& D_1\, D_2 \to B_1\, B_2\, @ \\
& B_1 \to C_1 \\
& B_2 \to C_2 \\
& D_1 \to A \\
& D_2 \to A \\
& A\, B_1\, C_2 \to D_2 \\
& A\, B_2\, C_1 \to D_1\,\}
\end{aligned}
$$

Tiếp theo, tạo tập phủ dạng vành G_3:

$$\begin{aligned}\{\,&(D_1\,D_2,\ B_1\,B_2) \rightarrow @\\ &(B_1) \rightarrow C_1\\ &(B_2) \rightarrow C_2\\ &(D_1) \rightarrow A\\ &(D_2) \rightarrow A\\ &(A\,B_1\,C_2) \rightarrow D_2\\ &(A\,B_2\,C_1) \rightarrow D_1\,\}\end{aligned}$$

Kết quả chúng ta có tập lược đồ quan hệ $\mathcal{R} = \{R_1, R_2, R_3, R_4, R_5, R_6, R_7\}$ sau khi loại thuộc tính giả @ khỏi lược đồ R_1:

- $R_1 :\ B_1\,B_2\,D_1\,D_2$ $\qquad \mathcal{K}_1 = \{B_1\,B_2,\ D_1\,D_2\}$
- $R_2 :\ B_1\,C_1$ $\qquad \mathcal{K}_2 = \{B_1\}$
- $R_3 :\ B_2\,C_2$ $\qquad \mathcal{K}_3 = \{B_2\}$
- $R_4 :\ A\,D_1$ $\qquad \mathcal{K}_4 = \{D_1\}$
- $R_5 :\ A\,D_2$ $\qquad \mathcal{K}_5 = \{D_2\}$
- $R_6 :\ A\,B_1\,C_2\,D_2$ $\qquad \mathcal{K}_6 = \{A\,B_1\,C_2\}$
- $R_7 :\ A\,B_2\,C_1\,D_1$ $\qquad \mathcal{K}_7 = \{A\,B_2\,C_1\}$

Định lý 3. *Tập lược đồ quan hệ $\mathcal{R} = \{R_1, R_2, ..., R_k\}$ được tổng hợp qua thuật toán TH-3NF từ tập các phụ thuộc hàm F thỏa mãn các tính chất sau đây:*

1) *Bảo toàn tập các phụ thuộc hàm.*
2) *Các lược đồ con là ở dạng chuẩn ba.*
3) *Có kết nối không mất thông tin.*
4) *Ngoài ra, không tồn tại tập lược đồ quan hệ nào khác có số lượng ít hơn thỏa mãn các tính chất nêu trên.*

Chứng minh.

1) Qua các bước thực hiện của thuật toán chúng ta có: $F \equiv F' \equiv G_1 \equiv G_2$, chúng ta lại có $G_2 \equiv \{K \rightarrow R_i \mid R_i \in \mathcal{R},\ K$ là khóa được chỉ định của $R_i\}$, như vậy F được đặc trưng hoàn toàn bởi $\mathcal{R}$ (Bổ đề 1).

2) Giả sử tồn tại một lược đồ con R_i có sự phụ thuộc bắc cầu của thuộc tính không khóa $A \in R_i$ vào khóa $K_1^i \in \mathcal{K}_i = \{K_1^i, K_2^i, ..., K_t^i\}$ được chỉ định của R_i qua tập $Z \subset R_i : K_1^i \rightarrow Z,\ Z \not\rightarrow K_1^i$, $Z \rightarrow A$ và $A \notin Z$ (1) (không mất tính tổng quát vì luôn có $K_1^i \leftrightarrow K$, với K là một khóa nào đó của R_i). Cần chỉ ra rằng A là thuộc tính dư thừa trong vế phải của một phụ thuộc hàm thuộc G_2.

Xét lược đồ R_i tổng hợp được từ phụ thuộc hàm phức hợp thứ i, ký hiệu là $FF_i : (K_1^i, K_2^i, ..., K_t^i) \rightarrow Y^i \in G_3$, thuộc tính không khóa của lược đồ R_i chỉ có thể thuộc vế phải của phụ thuộc hàm phức hợp FF_i, do đó $A \in Y^i$. Vì các tập trái $(K_1^i, K_2^i, ..., K_t^i)$ của FF_i tương đương với nhau, nên chỉ cần xét trường hợp A thuộc vế phải Y_1^i của phụ thuộc hàm $K_1^i \rightarrow Y_1^i \in E_{G_2}(X^i) \subseteq G_2$ (với $G_2 = \{E_{G_2}(X^i) \mid X^i \subseteq R_i, i = 1, 2, ..., q\}$; $E_{G_2}(X^i) = \{K_j^i \rightarrow Y_j^i \mid K_j^i \leftrightarrow K_h^i;\ Y^i = \bigcup_{j=1}^{t} Y_j^i; j, h = 1, 2, ..., t\}$). Theo thuật toán thì G_2 chỉ khác G_1 là các phụ thuộc hàm thuộc G_2 đã được loại bỏ các thuộc tính dư thừa ở vế phải, còn số lượng phụ thuộc hàm cũng như các vế trái của các phụ thuộc hàm vẫn giữ nguyên, vì thế phụ thuộc hàm $K_1^i \rightarrow Y_1^i \in G_2$ không có thuộc tính dư thừa vế phải. Nhưng A là thuộc tính dư thừa trong vế phải của $K_1^i \rightarrow Y_1^i$ (vì ta có $A \in Y_1^i$ và theo (1) thì $K_1^i \rightarrow A$ nên $\{G_2 \setminus \{K_1^i \rightarrow Y_1^i\}) \cup \{K_1^i \rightarrow (Y_1^i \setminus A)\}\} \models K_1^i \rightarrow Y_1^i$), điều này chứng tỏ G_2 có phụ thuộc hàm dư thừa vế phải. Mâu thuẫn vì G_2 là G_1 đã được rút gọn vế phải (theo thuật toán) (Lưu ý: chúng ta nhận được tập G_3 chính là tập G đã được rút gọn vế phải; tập phụ thuộc hàm đặc trưng tự nhiên một khi được tạo lại từ G_3 thì không thể xuất hiện một phụ thuộc hàm nào có thuộc tính dư thừa ở vế phải nữa).

3) Việc bổ sung phụ thuộc hàm $U \to @$ vào tập phụ thuộc hàm F (trong đó @ là tên "thuộc tính giả" mà $@ \notin U$), rồi qua các bước biến đổi của thuật toán: rút gọn vế trái, thực chất là loại bớt các thuộc tính khỏi siêu khóa U để nhận được khóa bên vế trái của phụ thuộc hàm bổ sung này, trong thuật toán có thực hiện loại bỏ các phụ thuộc hàm dư thừa nhưng chúng ta dễ thấy rằng duy nhất chỉ có phụ thuộc hàm nói trên có thuộc tính giả @ ở vế phải nên nó không phải là phụ thuộc hàm dư thừa, do đó, việc bổ sung này thực chất là việc bổ sung thêm một khóa K của lược đồ R vào phép tổng hợp (lưu ý rằng thuộc tính giả @ sẽ được loại ra khỏi lược đồ kết quả).

Xét việc bổ sung khóa K như trên vào phép tách bảo toàn các phụ thuộc hàm theo giả thiết của Định lý 2 thì có hai trường hợp có thể xảy ra, hoặc là khóa K là một trong số lược đồ R_i; hoặc là, nếu có một lược đồ R_i nào đó có tập khóa được chỉ định tương đương với K, thậm chí là có thể có một khóa được chỉ định trùng với K (trong các trường hợp như vậy dễ thấy ngay là tập khóa này cũng đồng thời là các khóa của lược đồ R), thì K sẽ được đưa vào trong lược đồ R_i (cụ thể $K \in \mathcal{K}_i$). (Xem thí dụ 1: chúng ta nhận được tập $B_1 B_2$ là khóa của lược đồ R sau khi loại bỏ các thuộc tính dư thừa vế trái của phụ thuộc hàm có thuộc tính giả: $A_1B_1B_2C_1C_2D_1D_2 \to @$, và khóa $B_1 B_2$ nhận được này trùng với khóa được chỉ định $B_1 B_2$ của R_1 nên được bổ sung vào lược đồ R_1). Như vậy, hoàn toàn có thể áp dụng được Định lý 2 để khẳng định tính chất kết nối không mất của tập lược đồ kết quả.

4) Căn cứ theo Bổ đề 2 và Hệ quả 1. □

Chúng ta xét thí dụ 2 để thấy rõ hơn các bước cần thực hiện nêu trong thuật toán TH-3NF nhằm đảm bảo cho tính chất 2 trong Định lý 3.

Thí dụ 2. Sử dụng lược đồ nêu trong Thí dụ 1, xét các bước biến đổi theo thuật toán TH-3NF. Phụ thuộc hàm $B_1 B_2 \to A$ được giữ lại trong tập F' vì nó không phải là phụ thuộc hàm dư thừa và A không phải là thuộc tính dư thừa trong vế phải, nó đóng vai trò không thể thiếu trong việc hình thành phụ thuộc hàm phức hợp đầu tiên trong G, cụ thể là nó cần thiết để suy dẫn ra sự tương đương giữa $B_1 B_2$ và $D_1 D_2$. Tuy nhiên, nếu thuộc tính A vẫn tham gia trong lược đồ R_1 thì dễ dàng thấy rằng R_1 sẽ không ở dạng chuẩn ba, vì có sự phụ thuộc bắc cầu của thuộc tính A qua $X = \{D_1\}$ vào khóa $K = \{B_1 B_2\}$. Nhưng thuộc tính A trong phụ thuộc hàm nêu trên hoàn thành vai trò cần thiết sau khi chúng ta nhận được tập G, trong vế trái của phụ thuộc hàm phức hợp $(D_1 D_2,\ B_1 B_2) \to A@ \in G$ các tập trái là tương đương với nhau tức là $B_1 B_2 \leftrightarrow D_1 D_2$, thuộc tính A không còn cần giữ lại trong vế phải của phụ thuộc hàm phức hợp này để suy dẫn ra điều đó nữa, hơn nữa từ phụ thuộc hàm $D_1 \to A$ (hay $D_2 \to A$) luôn có thể suy dẫn ra phụ thuộc hàm $B_1 B_2 \to A$.

Thực hiện theo bước biến đổi của thuật toán thì từ G tạo được tập phụ thuộc hàm đặc trưng tự nhiên G_1:

$$
\begin{aligned}
\{\, & B_1 B_2 \to A\, D_1\, D_2\, @ \\
& D_1 D_2 \to A\, B_1\, B_2\, @ \\
& B_1 \to C_1 \\
& B_2 \to C_2 \\
& D_1 \to A \\
& D_2 \to A \\
& A\, B_1\, C_2 \to D_2 \\
& A\, B_2\, C_1 \to D_1 \,\}
\end{aligned}
$$

Trong G_1 dễ thấy A là thuộc tính dư thừa ở vế phải của hai phụ thuộc hàm đầu tiên. Do đó việc thực hiện loại bỏ thuộc tính dư thừa ở vế phải của G_1 để nhận được tập G_2 là cần thiết, và tập G_3 chính là tập G đã loại bỏ thuộc tính có thể bỏ được trong vế phải.

Như vậy, việc thực hiện loại bỏ thuộc tính dư thừa ở vế phải của G_1 nhằm đảm bảo tính chất: các lược đồ con được tạo từ G_3 sẽ là các lược đồ ở dạng chuẩn ba.

Thí dụ 3. Xét thí dụ này để thấy rõ hơn việc thực hiện bổ sung phụ thuộc hàm dạng $U \to @$ nêu trong thuật toán TH-3NF nhằm đảm bảo cho tính chất 3 trong Định lý 3.

Giả sử $F = \{A \to C,\ B \to C\}$. Nếu thuật toán TH-3NF không có bước bổ sung phụ thuộc hàm dạng $U \to @$, cụ thể đối với lược đồ ở đây là không bổ sung phụ thuộc hàm $ABC \to @$ vào tập F thì sẽ cho các lược đồ con

- $R_1 : AC$; $\quad K_1 = \{A\}$;
- $R_2 : BC$; $\quad K_2 = \{B\}$;

Tương ứng với các phụ thuộc hàm: $(R_1 \cap R_2) \to (R_1 - R_2)$, $(R_1 \cap R_2) \to (R_2 - R_1)$ là các phụ thuộc hàm $C \to A$, $C \to B$. Dễ thấy ngay là các phụ thuộc hàm này đều không thuộc F^+. Như vậy kết nối của R_1 và R_2 là mất thông tin (theo Định lý 1).

Còn nếu thực hiện đầy đủ các bước đã nêu trong thuật toán TH-3NF(F) tức là có thực hiện bước bổ sung phụ thuộc hàm $ABC \to @$ vào tập F thì sẽ cho các lược đồ con:

- $R_1 : AC$; $\quad K_1 = \{A\}$;
- $R_2 : BC$; $\quad K_2 = \{B\}$;
- $R_3 : AB$; $\quad K_1 = \{AB\}$;

Dễ nhận thấy kết nối của các lược đồ con nêu trên là không mất thông tin (theo Định lý 1).

Bổ đề 3. *Thuật toán TH-3NF có độ phức tạp tính toán theo thời gian là $O(n^2)$, trong đó n là độ dài của dữ liệu vào (với $n = ap$ là độ dài dữ liệu vào, a là số lượng các ký hiệu thuộc tính khác nhau trong F, p là số lượng phụ thuộc hàm trong F).*

Chứng minh. Độ phức tạp tính toán theo thời gian của thuật toán TH-3NF phụ thuộc chủ yếu vào tổng thời gian thực hiện thuật toán loại bỏ thuộc tính dư thừa, thời gian thực hiện thuật toán loại bỏ phụ thuộc hàm dư thừa và thời gian thực hiện thuật toán tạo tập phụ thuộc hàm phức hợp. Trong đó (sử dụng các thuật toán được nêu trong [5]), thuật toán loại bỏ các phụ thuộc hàm dư thừa NONREDUNT có độ phức tạp thời gian $O(np)$; thuật toán loại bỏ các thuộc tính dưthừa REDUCE có độ phức tạp thời gian $O(n^2)$; thuật toán tạo tập phụ thuộc hàm phức hợp là sự thực hiện so sánh các bao đóng của các vế trái (của các phụ thuộc hàm) có độ phức tạp thời gian căn cứ theo thuật toán tính bao đóng của tập thuộc tính với độ phức tạp thời gian là $O(n)$ (với giả thiết chúng ta sử dụng thuật toán LINCLOSURE; Nếu tính bao đóng dùng thuật toán CLOSURE, thì CLOSURE có độ phức tạp thời gian là $O(np)$). Như vậy độ phức tạp thời gian của thuật toán TH-3NF sẽ là $O(n^2)$. □

Một vài điểm lưu ý về thuật toán TH-3NF

1) Chúng ta nhận thấy rằng, khi tạo tập phụ thuộc hàm phức hợp G, nếu tập G chỉ bao gồm có một phụ thuộc hàm phức hợp (tương ứng sẽ chỉ có một lược đồ quan hệ kết quả), thì không cần thực hiện các bước biến đổi tiếp theo nữa, tập G lúc này đóng vai trò của tập G_3, và trực tiếp thực hiện việc tạo lược đồ kết quả từ G.

Thí dụ 4. Cho tập các thuộc tính $U = \{A_1, A_2, A_3, A_4, A_5, A_6\}$, tập các phụ thuộc hàm F:

$$\begin{aligned} \{ A_1 &\to A_2 A_3 A_6 \\ A_2 &\to A_3 A_4 \\ A_3 &\to A_4 A_5 \\ A_5 &\to A_1 A_4 \} \end{aligned}$$

Sau khi bổ sung phụ thuộc hàm $A_1 A_2 A_3 A_4 A_5 A_6 \to @$ vào, thực hiện rút gọn vế trái và vì không có phụ thuộc hàm dư thừa, chúng ta có tập F':

$$\begin{aligned}\{\, & A_1 \rightarrow A_2\, A_3\, A_6 \\ & A_2 \rightarrow A_3\, A_4 \\ & A_3 \rightarrow A_4\, A_5 \\ & A_5 \rightarrow A_1\, A_4 \\ & A_1\, A_4\, A_6 \rightarrow @\,\}\end{aligned}$$

Do đó, chúng ta có tập phủ dạng vành G:

$$\{\,(A_1,\, A_2,\, A_3,\, A_5) \rightarrow A_4\, A_6\, @\,\}$$

Chúng ta thực hiện được ngay việc tạo các lược đồ kết quả từ G, vì tập G bao gồm chỉ có một phụ thuộc hàm phức hợp (tương ứng sẽ chỉ có một lược đồ quan hệ kết quả), thì không cần thực hiện các bước biến đổi tiếp theo nữa, tập G lúc này đóng vai trò của tập G_3, và trực tiếp thực hiện việc tạo lược đồ kết quả từ G. Do đó tập lược đồ kết quả chỉ gồm một lược đồ con:

R_1: $A_1\, A_2\, A_3\, A_4\, A_5\, A_6$; $\quad \mathcal{K}_1 = \{\, A_1,\, A_2,\, A_3,\, A_5\,\}$.

Xem lại lược đồ ban đầu thì rõ ràng lược đồ R đã cho đã ở 3NF.

2) Nếu thực hiện thêm việc rút gọn vế phải trước khi loại bỏ các phụ thuộc hàm dư thừa của tập F thì tập F' sẽ là tập phụ thuộc hàm rút gọn và không dư thừa. Liệu thực hiện việc rút gọn này rồi thì có bỏ được việc rút gọn vế phải của G_1 không? Cũng sử dụng thí dụ 1, chúng ta thấy rằng việc rút gọn vế phải của G_1 là không thể bỏ được. Vì tuy thêm việc rút gọn vế phải của F nhưng hai phụ thuộc hàm đầu tiên trong F' sẽ là: $B_1B_2 \rightarrow A$, $D_1D_2 \rightarrow B_1B_2$ và kết hợp với phụ thuộc hàm cuối cùng là $B_1B_2 \rightarrow @$ được nhóm lại tạo thành phụ thuộc hàm phức hợp $(D_1D_2, B_1B_2) \rightarrow A@$, như nhận xét trong Thí dụ 3, chúng ta thấy rằng lược đồ R_1 tạo từ phụ thuộc hàm phức hợp này không là ở 3NF.

Tuy nhiên, nếu sau khi thực hiện việc rút gọn vế phải trước khi loại bỏ các phụ thuộc hàm dư thừa của tập F để nhận được tập F', chúng ta vẫn thực hiện việc rút gọn vế phải của G_1, thì rõ ràng là sẽ nhận được các lược đồ tổng hợp đều là ở 3NF.

Trong một số trường hợp đặc biệt, việc rút gọn vế phải hai lần sẽ thấy không phải là việc lặp lại không có ích mà sẽ nhờ đó có thể bỏ qua một số bước về sau, đó là trong trường hợp chúng ta nhận được tập G bao gồm các phụ thuộc hàm phức hợp đều không có thuộc tính nào thuộc U có trong vế phải, và thực hiện được ngay việc tạo các lược đồ kết quả từ G (tức là có thể bỏ qua các bước trong thuật toán TH-3NF là: - Tạo tập phụ thuộc hàm đặc trưng tự nhiên tương đương với tập phụ thuộc hàm phức hợp đã xây dựng, nhận được tập G_1. Rút gọn vế phải của các phụ thuộc hàm trong G_1. Kết quả của bước này nhận được tập G_2; - Tạo tập phủ dạng vành G_3).

Thí dụ 5. Cho các tập thuộc tính $U = \{A,\, B_1,\, B_2,\, C_1,\, C_2,\, D\}$, tập các phụ thuộc hàm F:

$$\begin{aligned}\{\, & A\, B_1 \rightarrow B_2\, C_1 \\ & B_2 \rightarrow A\, B_1 \\ & C_1 \rightarrow B_2 \\ & C_2 \rightarrow D \\ & D \rightarrow C_2\,\}\end{aligned}$$

Sau khi bổ sung phụ thuộc hàm $AB_1B_2C_1C_2D \rightarrow @$ vào, thực hiện rút gọn vế trái, vế phải và vì không có phụ thuộc hàm dư thừa, chúng ta có tập F':

$$\begin{aligned}
\{\, & A B_1 \to B_2 C_1 \\
& B_2 \to A B_1 \\
& C_1 \to B_2 \\
& C_2 \to D \\
& D \to C_2 \\
& C_1 D \to @ \,\}
\end{aligned}$$

Do đó chúng ta có tập phủ dạng vành G:

$$\begin{aligned}
\{\, & (A B_1,\ B_2,\ C_1) \\
& (C_2,\ D) \\
& (C_1 D) \to @ \,\}
\end{aligned}$$

Chúng ta thực hiện được ngay việc tạo các lược đồ kết quả từ G, vì tập G bao gồm các phụ thuộc hàm phức hợp đều không có thuộc tính nào thuộc U có trong vế phải, mà chỉ phụ thuộc hàm phức hợp thứ ba có thuộc tính giả @ ở bên vế phải, do đó tập lược đồ kết quả là: $\mathcal{R} = \{R_1,\ R_2,\ R_3\}$ với:

- $R_1 :\ A B_1 B_2 C_1$ $\quad \mathcal{K}_1 = \{A B_1,\ B_2,\ C_1\}$
- $R_2 :\ C_2 D$ $\quad \mathcal{K}_2 = \{C_2,\ D\}$
- $R_3 :\ C_1 D$ $\quad \mathcal{K}_3 = \{C_1 D\}$

TÀI LIỆU THAM KHẢO

[1] Amstrong W. W., Dependency structures of database relationships, *Information Processing* 74, North Holland Pub. Co. Amsterdam, 1974.

[2] Atzeni P., De Antonellis V., *Relational Database Theory*, The Benjamin/Cummings Publishing Company, 1993.

[3] Berri C., Bernstein P. A., Computational problems related to the design of normal form relational schemes, *ACM Translations on Database Systems* **4** (1) (1980).

[4] Codd E. F., A relational model of data for large shared data banks, *CACM* **13** (6) (1970).

[5] Maier D., *The Theory of Relational Databases*, Computer Science Press, 1983.

[6] Ullman J. D., *Principles of Database Systems*, 2nd edition, Computer Science Press, 1982.

Nhận bài ngày 21 - 8 - 1999
Nhận lại sau khi sửa ngày 20 - 12 - 1999

Phạm Quang Trung, Viện Kiểm sát nhân dân tối cao.
Nguyễn Xuân Huy, Viện Công nghệ thông tin.