Phụ thuộc hàm và Chuẩn hóa CSDL

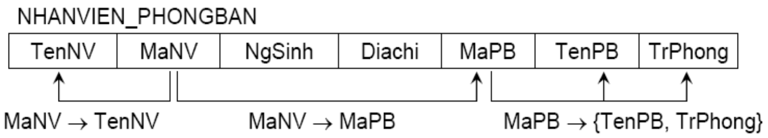
Nội dung

- Phụ thuộc hàm.
- Các dạng chuẩn.
- Một số thuật toán chuẩn hóa.

Phụ thuộc hàm (1)

- Phụ thuộc hàm(PTH) - Functional Dependencies
- Xét lược đồ quan hệ gồm n thuộc tính
  - R(U), U={A_1, A_2,…, A_n}
- PTH giữa hai tập thuộc tính X, Y ⊆ U
  - Ký hiệu: X → Y.
  - ∀r∈R, ∀t_1, t_2 ∈r nếu $t_1[X] = t_2[X]$ thì $t_1[Y] = t_2[Y]$.
- X là vế trái và Y là vế phải của PTH.
- X → Y được gọi là PTH hiển nhiên nếu Y ⊆ X
- X → Y được gọi là PTH nguyên tố (Y PTH đầy đủ vào X nếu nếu ∀X'⊂ X thì X' không → Y

Phụ thuộc hàm (2)



- r ∈R thỏa mãn các PTH gọi là trạng thái hợp lệ của R
- Nhận xét:
  - Các PTH xuất phát từ các ràng buộc trong thế giới thực.
  - ∀r∈R, ∀t∈r, t [X] là duy nhất thì X là một khóa của R.
  - Nếu K là một khóa của R thì K xác định hàm tất cả các tập thuộc tính của R.
  - PTH dùng để đánh giá một thiết kế CSDL

Bao đóng của tập PTH

- F là tập PTH trên R
  - F = {MaNV → TenNV, MaPB → {TenPB, TrPhong}, MaNV → MaPB}.
  - ∀r∈R thỏa F và MaNV → {TenPB, TrPhong} cũng đúng với r thì MaNV → {TenPB, TrPhong} gọi là được suy diễn từ F.
- Bao đóng của F, ký hiệu F^+, gồm
  - F
  - Tất cả các PTH được suy diễn từ F.
- F gọi là đầy đủ nếu $F = F^+$.

Luật suy diễn (1)

- Luật suy diễn dùng để suy diễn một PTH mới từ một tập PTH cho trước.
- Hệ luật suy diễn Armstrong
  - Phản xạ: $Y \subseteq X \Rightarrow X \rightarrow Y$.
  - Tăng trưởng: $X \rightarrow Y \Rightarrow XZ \rightarrow YZ$, với $XZ = X \cup Z$.
  - Bắc cầu: $X \rightarrow Y, Y \rightarrow Z \Rightarrow X \rightarrow Z$.
-
  - Phân rã: $X \rightarrow YZ \Rightarrow X \rightarrow Y, X \rightarrow Z$.
  - Hợp: $X \rightarrow Y, X \rightarrow Z \Rightarrow X \rightarrow YZ$.
  - Bắc cầu giả: $X \rightarrow Y, WY \rightarrow Z \Rightarrow WX \rightarrow Z$.

Luật suy diễn (2)

- Ví dụ 1:
  - Cho $F=\{A \rightarrow B, B \rightarrow C, A \rightarrow D\}$
  - Hãy chứng tỏ PTH $A \rightarrow CD$ suy diễn từ F nhờ luật dẫn Armstrong
  - Cách giải:
    - $A \rightarrow B, B \rightarrow C \Rightarrow A \rightarrow C$ (luật bắc cầu)
    - $A \rightarrow C, A \rightarrow D \Rightarrow A \rightarrow CD$ (luật hợp).
- Ví dụ 2: Cho $F=\{AB\rightarrow E, AG\rightarrow I, BE\rightarrow I, E\rightarrow G, GI\rightarrow H\}$
  - Hãy chứng tỏ PTH $AB \rightarrow GH$ suy diễn từ F nhờ luật dẫn Armstrong

Bao đóng của tập thuộc tính

- Làm thế nào để biết một PTH $X \rightarrow Y$ được suy diễn từ tập PTH F cho trước?
- Bao đóng của tập thuộc tính X đối với F, ký hiệu X^+ là
  - Tập các thuộc tính PTH vào X.
  - $X^+ = \{A \in U \mid X \rightarrow A \in F^+\}$
- Nhận xét:
  - $X \rightarrow Y \in F^+ \Leftrightarrow Y \subseteq X^+$.
  - Nếu K là khóa của R thì $K^+ = U$.

Thuật toán tìm X^+

- Input: U, F và $X \subseteq U$
- Output: X^+
- Thuật toán
  - *B1*: $X^+ = X$;
  - *B2*: Nếu tồn tại $Y \rightarrow Z \in F$ và $Y \subseteq X^+$ thì
    - $X^+ = X^+ \cup Z$;
    - tiếp tục *B2*.
  - Ngược lại qua *B3*.
  - *B3*: output X^+

Ví dụ tìm X^+

- Input:
  - $F = \{AB \rightarrow C, BC \rightarrow D, D \rightarrow EG\}$
  - $X = BD$
- Output: X^+
- Thuật toán
  - $X^+ = BD$.
  - Lặp 1:
    - Tìm các PTH có vế trái là tập con của $X^+ = BD$
      - $D \rightarrow EG$, thêm EG vào X^+ ta được $X^+ = BDEG$.
  - Lặp 2:
    - Tìm các PTH có vế trái là tập con của $X^+ = BDEG$
      - Không có PTH nào.
  - Vậy $X^+ = BDEG$.

Ví dụ tìm X^+

- VD2: Cho lược đồ quan hệ Q(ABCDEGH) và tập PTH F
  - $F = \{ B \rightarrow A, DA \rightarrow CE, D \rightarrow H, GH \rightarrow C, AC \rightarrow D \}$
  - Tìm bao đóng của tập X={AC} dựa trên F

- VD3: Cho lược đồ quan hệ Q(ABCDEGH) và tập PTH F
  - $F = \{A \rightarrow C, A \rightarrow EG, B \rightarrow D, G \rightarrow E\}$
  - Xác định X^+
    - X= {AB}
    - X={CGD}

Các tập PTH tương đương

- Tập PTH F được nói là phủ tập PTH G nếu $G \subset F^+$
- Hai tập PTH F và G là tương đương nếu
  - F phủ G và
  - G phủ F
- Nhận xét
  - $\forall X \rightarrow Y \in G$, nếu $Y \subseteq X_{F^+}$ thì F phủ G.
  - F và G tương đương nếu và chỉ nếu $F^+ = G^+$

Kiểm tra PTH suy diễn

- Cho F = {AB → C, A → D, D → E, AC → B}
  - Hai PTH AB → E và D → C có được suy diễn từ F hay không?

Tập PTH tối thiểu (1)

- Thừa PTH
  - {A → B, B → C, A → C}, vì A → C được suy diễn từ {A → B, B → C}
  - A → B, B → C ⇒ A → C (luật bắc cầu).
- Thừa thuộc tính
  - {A → B, B → C, A → CD}, vì A → CD được suy diễn từ {A → B, B → C, A → D}
    - A → B, B → C ⇒A → C (luật bắc cầu)
    - A → C, A → D ⇒ A → CD (luật hợp).
  - {A → B, B → C, AC → D}, vì AC → D được suy diễn từ {A → B, B → C, A → D}
    - A → B, A → D ⇒ A → BD (luật hợp)
    - A → BD ⇒AC → BCD (luật tăng trưởng)
    - AC → BCD ⇒ AC → D (luật phân rã).

Tập PTH tối thiểu (2)

- Tập PTH F là tối thiểu nếu thỏa các điều kiện sau:
  - Mọi PTH của F chỉ có một thuộc tính ở vế phải.
  - Không thể thay $X \rightarrow A$ thuộc F bằng $Y \rightarrow A$ với $Y \subset X$ mà tập mới tương đương với F.
  - Nếu bỏ đi một PTH bất kỳ trong F thì tập PTH còn lại không tương đương với F.
- Phủ tối thiểu của tập PTH E là tập PTH tối thiểu F tương đương với E.
- Nhận xét
  - Mọi tập PTH có ít nhất một phủ tối thiểu.

Thuật toán tìm tập PTH tối thiểu

- Input: tập PTH E.
- Output: phủ tối thiểu F của E.
- Thuật toán:
  - *B1*: $F = \emptyset$
  - *B2*: Với mọi $X \rightarrow Y \in E$, $Y = \{A_1, \ldots, A_k\}$, $A_i \in U$
    - $F = F \cup \{X \rightarrow \{A_i\}\}$
  - *B3*: Với mỗi $X \rightarrow \{A\} \in F$, $X = \{B_1, \ldots, B_l\}$, $B_i \in U$
    - Với mỗi B_i, nếu $A \in (X - \{B_i\})_F^+$ thì
      - $F = (F - \{X \rightarrow \{A\}\}) \cup \{(X - \{B_i\}) \rightarrow \{A\}\}$
  - *B4*: Với mỗi $X \rightarrow \{A\} \in F$
    - $G = F - \{X \rightarrow \{A\}\}$
    - Nếu $A \in X_G+$ thì $F = F - \{X \rightarrow \{A\}\}$.

Ví dụ tìm tập PTH tối thiểu

- Tìm phủ tối thiểu của
  $E = \{A \to BC, A \to B, B \to C, AB \to C\}$
  - *B1*: $F = \varnothing$.
  - *B2*: $F = \{A \to B, A \to C, B \to C, AB \to C\}$.
  - *B3*: Xét $AB \to C$
    - $(B)_F^+ = C$
    - $F = \{A \to B, A \to C, B \to C\}$.
  - *B4*: $A \to C$ thừa.
  - $F = \{A \to B, B \to C\}$.

Siêu khóa và Khóa

- Cho R(U)
  - $S \subseteq U$ là siêu khóa nếu $\forall r \in R$, $\forall t_1, t_2 \in r$, $t_1 \neq t_2$ thì $t_1[S] \neq t_2[S]$.
  - $K \subseteq U$ là khóa nếu K là siêu khóa nhỏ nhất.
    - $A \in K$ được gọi là thuộc tính khóa.
- Nhận xét
  - S xác định hàm tất cả các thuộc tính của R.
  - R có thể có nhiều khóa.

Xác định khóa của lược đồ

- Input: tập PTH F xác định trên lược đồ R(U).
- Output : khóa K của R.
- Thuật toán
  - *B1*:
    - $K = U = \{A_1, \ldots, A_n\}$
    - $i = 1;$
  - *B2*:
    - Nếu $U \subseteq (K - \{A_i\})_F^+$ thì $K = K - \{A_i\}$.
    - $i = i + 1;$
    - Nếu $i > n$ thì sang *B3*. Ngược lại, tiếp tục *B2*.
  - *B3*:
    - Output K.

Ví dụ tìm khóa của lược đồ

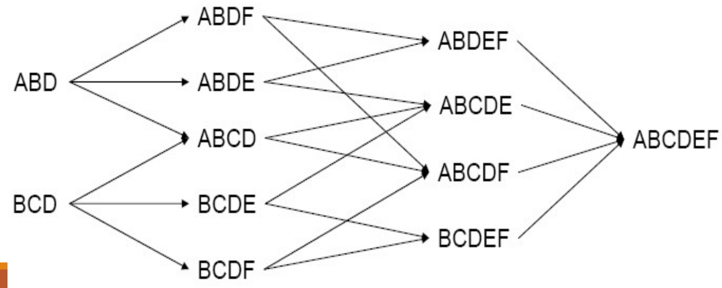
- Cho R(U), U = {A, B, C, D, E, F, G}.
  - $F = \{B \rightarrow A, D \rightarrow C, D \rightarrow BE, DF \rightarrow G\}$.
- Tìm khóa của R
  - B1:
    - K = ABCDEFG.
  - B2:
    - Lặp 1: $(BCDEFG)_F^+ = BCDEFGA \Rightarrow K = BCDEFG$.
    - Lặp 2: $(CDEFG)_F^+ = CDEFGBA \Rightarrow K = CDEFG$.
    - Lặp 3: $(DEFG)_F^+ = DEFGCBA \Rightarrow K = DEFG$.
    - Lặp 4: $(EFG)_{F+} = EFG$.
    - Lặp 5: $(DFG)_F^+ = DFGCBEA \Rightarrow K = DFG$.
    - Lặp 6: $(DG)_F^+ = DGCBEA$.
    - Lặp 7: $(DF)_F^+ = DFCBEAG \Rightarrow K = DF$.
  - B3:
    - Khóa là K = DF.

Xác định tất cả khóa của lược đồ

- Input: tập PTH F xác định trên lược đồ R(U).
- Output: tất cả khóa của R.
- Thuật toán
  - *B1*:
    - Xây dựng 2^n tập con của $U = \{A_1, \ldots, A_n\}$
    - $S = \{\}$;
  - *B2*:
    - Với mỗi tập con $X \subseteq U$
    - Nếu $U \subseteq X_F^+$ thì $S = S \cup \{X\}$
  - *B3*:
    - $\forall X, Y \in S$, nếu $X \subset Y$ thì $S = S - \{Y\}$
  - *B4*:
    - S là tập các khóa của R

Ví dụ tìm tất cả khóa của lược đồ

- Cho R(U), U = {A, B, C, D, E, F}.
  - $F = \{AE \rightarrow C, CF \rightarrow A, BD \rightarrow F, AF \rightarrow E\}$.
- Tìm tất cả khóa của R
  - Tập siêu khóa
    - S = {ABD, BCD, ABCD, ABDE, BCDE, ABCDE, ABDF, BCDF, ABCDF, ABDEF, BCDEF, ABCDEF}.



Chuẩn hóa dữ liệu

- Giới thiệu về chuẩn hóa?
- Các dạng chuẩn
  - Dạng 1 (1 Normal Form - 1NF)
  - Dạng 2 (2 Normal Form - 2NF)
  - Dạng 3 (3 Normal Form - 3NF).
  - Dạng Boyce - Codd (Boyce - Codd Normal Form - BCNF)

Các dạng chuẩn

- Dạng 1 (1 Normal Form - 1NF)
- Dạng 2 (2 Normal Form - 2NF)
- Dạng 3 (3 Normal Form - 3NF).
- Dạng Boyce - Codd (Boyce - Codd Normal Form - BCNF)

Dạng chuẩn 1 (1)

- Định nghĩa
  - Lược đồ quan hệ R được gọi là thuộc dạng chuẩn 1 khi và chỉ khi mọi thuộc tính của R là thuộc tính đơn.
- Ví dụ

PHONGBAN

TENPB	MAPB	TrPhong	CacTruso
Hành chính	5	22221	Đống Đa, Hoàng Mai
Nghiên cứu	2	21113	Ba Đình

Không thuộc dạng chuẩn 1

PHONGBAN

TENPB	MAPB	TrPhong	CacTruso
Hành chính	5	22221	Đống Đa
Hành chính	5	22221	Hoàng Mai
Nghiên cứu	2	21113	Ba Đình

Thuộc dạng chuẩn 1

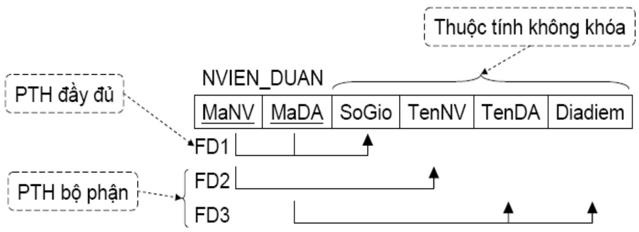
Dạng chuẩn 2 (1)

- Định nghĩa
  - Lược đồ quan hệ R được gọi là thuộc dạng chuẩn 2 khi và chỉ khi:
    - R ở dạng chuẩn 1
    - Mọi thuộc tính không khóa đều phụ thuộc hàm đầy đủ vào khóa chính.
  - R(U), $K \subseteq U$ là khóa chính của R
    - $A \in U$ là thuộc tính không khóa nếu $A \notin K$.
    - $X \rightarrow Y$ là PTH đầy đủ nếu $\forall A \in X$ thì $(X - \{A\}) \rightarrow Y$ không đúng trên R.

      Ngược lại $X \rightarrow Y$ là PTH bộ phận.

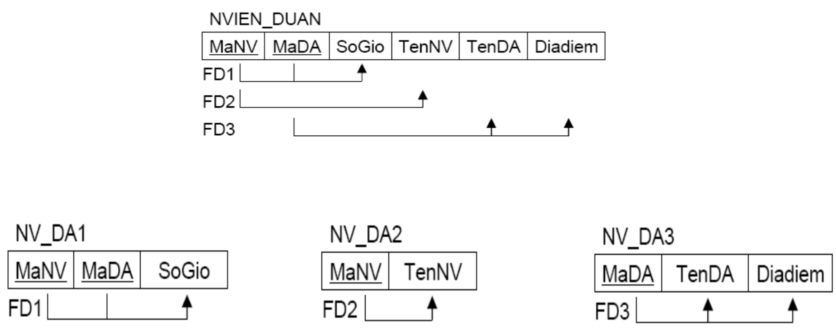
Dạng chuẩn 2 (2)

- Ví dụ 1:



Dạng chuẩn 2 (3)

- Ví dụ 2:

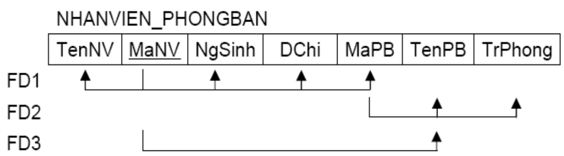


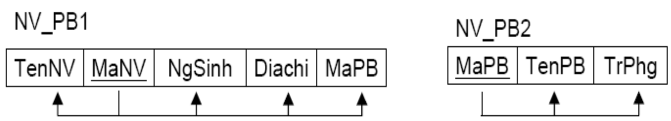
Dạng chuẩn 3 (1)

- **Định nghĩa**
  - Lược đồ quan hệ R được gọi là thuộc dạng chuẩn 3 khi và chỉ khi:
    - R ở dạng chuẩn 2
    - Mọi thuộc tính không khóa đều không phụ thuộc hàm bắc cầu vào khóa chính.
  - R(U)
    - $X \rightarrow Y$ là PTH bắc cầu nếu $\exists Z \subseteq U$, Z không là khóa và cũng không là tập con của khóa của R mà $X \rightarrow Z$ và $Z \rightarrow Y$ đúng trên R.

Dạng chuẩn 3 (2)

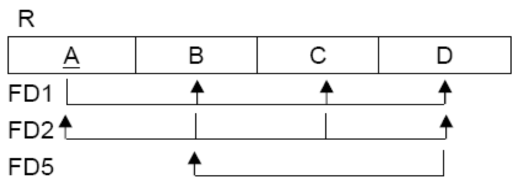
- **Ví dụ:**
  - FD3 là PTH bắc cầu





Dạng chuẩn Boyce Codd (1)

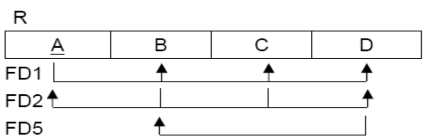
- Định nghĩa
  - Lược đồ quan hệ R được gọi là thuộc dạng chuẩn BCNF khi và chỉ khi:
    - PTH không hiển nhiên X → Y đúng trên R thì X là siêu khóa của R.
- Ví dụ
  - Cho lược đồ quan hệ R(ABCD)



R ở dạng chuẩn nào?

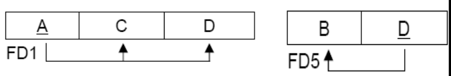
Dạng chuẩn Boyce Codd (2)

A	B	C	D
1	a	a	1
2	a	b	1
3	b	a	2
4	b	b	2



A	C	D
1	a	1
2	b	1
3	a	2
4	b	2

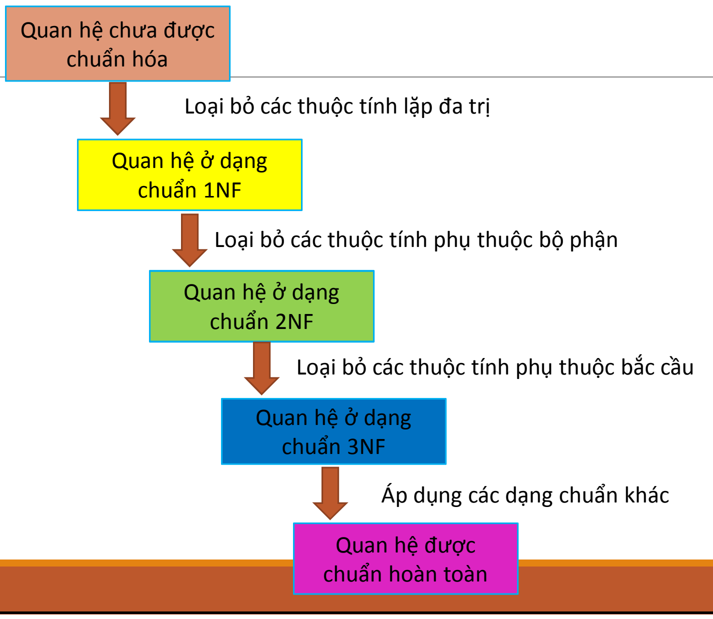
D	B
1	a
2	b



Dạng chuẩn Boyce Codd (3)

- Nhận xét:
  - Mọi quan hệ thuộc dạng chuẩn BCNF cũng thuộc dạng chuẩn 3
  - Dạng chuẩn BCNF đơn giản và chặt chẽ hơn chuẩn 3
  - Mục tiêu của quá trình chuẩn hóa là đưa lược đồ quan hệ về dạng chuẩn 3 hoặc chuẩn BCNF.

Các dạng chuẩn



Mục đích của chuẩn hóa dữ liệu

Xác định được 1 tập các lược đồ quan hệ cho phép tìm kiếm thông tin một cách dễ dàng, đồng thời tránh được dư thừa dữ liệu

Giải pháp:

Tách các lược đồ quan hệ "có vấn đề" thành những lược đồ quan hệ "chuẩn hơn"

Phép tách các lược đồ quan hệ

Mục đích

Thay thế một sơ đồ quan hệ $R(A_1, A_2, \ldots, A_n)$ bằng một tập các sơ đồ con $\{R_1, R_2, \ldots, R_k\}$ trong đó $R_i \subseteq R$ và $R = R_1 \cup R_2 \cup \ldots \cup R_k$

Yêu cầu của phép tách

- Bảo toàn thuộc tính, ràng buộc
- Bảo toàn dữ liệu

Phép tách không mất mát thông tin (Lossless join)

Định nghĩa: Cho lược đồ quan hệ R(U) phép tách R thành các sơ đồ con $\{R_1, R_2, \dots, R_k\}$ được gọi là phép tách không mất mát thông tin đối với một tập phụ thuộc hàm F nếu với mọi quan hệ r xác định trên R thỏa mãn F thì:

$$r = \Pi_{R1}(r) \bowtie \Pi_{R2}(r) \bowtie \dots \bowtie \Pi_{Rk}(r)$$

Ví dụ:
Supplier(sid, sname, pname,colour,quantity)

S1(sid, sname, city)
SP1(sid,pname,colour,quantity)

Kiểm tra tính không mất mát thông tin

Vào: $R(A_1, A_2, \dots, A_n)$, F, phép tách $\{R_1, R_2, \dots, R_k\}$
Ra: phép tách là mất mát thông tin hay không
Thuật toán
B.1. Thiết lập một bảng k hàng, n cột
Nếu A_j là thuộc tính của R_i thì điền a_j vào ô (i,j).
Nếu không thì điền b_{ij}.
B.i. Xét $f = X \rightarrow Y \in F$.
Nếu $\exists$ 2 hàng t1, t2 thuộc bảng : t1[X] = t2[X]
thì t1[Y] = t2[Y], ưu tiên đồng nhất về giá trị a
Lặp cho tới khi không thể thay đổi được giá trị nào trong bảng
B.n. Nếu bảng có 1 hàng gồm các kí hiệu $a_1, a_2, \dots, a_n$
thì phép tách là không mất mát thông tin.
ngược lại, phép tách không bảo toàn thông tin.

Ví dụ

R(MONHOC, SOTIET, LOP, GV, HOCVI, DC)

Kiểm tra: R_1(MONHOC, SOTIET, LOP, GV), R_2(GV, HOCVI, DC)

F = {MONHOC → SOTIET; MONHOC, LOP → GV; GV → HOCVI, DC}

	MONHOC	SOTIET	LOP	GV	HOCVI	DC
R1	a_1	a_2	a_3	a_4	b_{15}	b_{16}
R2	b_{21}	b_{22}	b_{23}	a_4	a_5	a_6

GV → HOCVI, DC

	MONHOC	SOTIET	LOP	GV	HOCVI	DC
R1	a_1	a_2	a_3	a_4	a_5	a_6
R2	b_{21}	b_{22}	b_{23}	a_4	a_5	a_6

Phép tách bảo toàn tập phụ thuộc hàm

Hình chiếu của tập phụ thuộc hàm

Cho sơ đồ quan hệ R, tập phụ thuộc hàm F, phép tách $\{R_1, R_2, \dots, R_k\}$ của R trên F.

Hình chiếu F_i của F trên R_i là tập tất cả $X \to Y \in F+$:

$$XY \subseteq R_i .$$

Phép tách sơ đồ quan hệ R thành $\{R_1, R_2, \dots, R_k\}$ là một phép tách bảo toàn tập phụ thuộc hàm F nếu

$$(F_1 \cup F_2 \dots \cup F_k)+ = F+$$

hay hợp của tất cả các phụ thuộc hàm trong các hình chiếu của F lên các sơ đồ con sẽ suy diễn ra các phụ thuộc hàm trong F.

Tách bảo toàn tập phụ thuộc hàm về 3NF

Vào: R(U), F (giả thiết F là phủ tối thiểu)

Ra: Phép tách bảo toàn tập phụ thuộc hàm về 3NF

Thuật toán

B1. Với các $A_i \in U$, $A_i \notin F$ thì loại A_i khỏi R và lập 1 quan hệ mới cho các A_i

B2. Nếu $\exists$ f $\in$ F, f chứa tất cả các thuộc tính của R thì kết quả là R

B3. Ngược lại, với mỗi $X \rightarrow A \in F$, xác định một quan hệ $R_i(XA)$.

Nếu $\exists$ $X \rightarrow A_i$, $X \rightarrow A_j$ thì tạo một quan hệ chung $R'(XA_iA_j)$

Ví dụ

Cho R = {A,B,C,D,E,F,G}

F = {A→B, ACD→E, EF→G}

- Xác định phép tách bảo toàn tập phụ thuộc hàm về 3NF

  B1. không lập được quan hệ nào mới.

  B2. !$\exists$ f $\in$ F: f chứa tất cả các thuộc tính của R

  B3. A→B ⇨ R1(AB)

  ACD→E ⇨ R2(ACDE)

  EF→G ⇨ R3(EFG)

Tách không mất mát thông tin và bảo toàn tập phụ thuộc hàm về 3NF

Yêu cầu:

- Bảo toàn tập phụ thuộc hàm (như thuật toán trên)
- Đảm bảo là có một lược đồ con chứa khóa của lược đồ được tách

Các bước tiến hành

- B1. Tìm một khóa tối thiểu của lược đồ quan hệ R đã cho
- B2. Tách lược đồ quan hệ R theo phép tách bảo toàn tập phụ thuộc hàm
- B3. Nếu 1 trong các sơ đồ con có chứa khóa tối thiểu thì kết quả của B2 là kết quả cuối cùng.
  Ngược lại, thêm vào kết quả đó một sơ đồ quan hệ được tạo bởi khóa tối thiểu tìm được ở 1.

Ví dụ

Cho R(A,B,C,D,E,F,G).
F = {A–>B, ACD->E, EF->G}

B1. Khóa tối thiểu cần tìm là ACDF
B2. Phép tách bảo toàn tập phụ thuộc hàm R cho 3 sơ đồ con R_1(AB), R_2(ACDE), R_3(EFG)
B3. Do khóa ACDF không nằm trong bất kỳ một sơ đồ con nào trong 3 sơ đồ con trên, ta lập một sơ đồ con mới R_4(ACDF)

Kết quả cuối cùng ta có phép tách R thành 4 sơ đồ con {R_1, R_2, R_3, R_4} là một phép tách không mất mát thông tin và bảo toàn tập phụ thuộc hàm

Tách không mất mát thông tin về BCNF

Vào: Sơ đồ quan hệ R, tập phụ thuộc hàm F.

Ra: phép tách không mất mát thông tin bao gồm một tập các sơ đồ con ở BCNF với các phụ thuộc hàm là hình chiếu của F lên sơ đồ đó.

Cách tiến hành

B1. KQ = {R},

B2. Với mỗi S $\in$ KQ, S không ở BCNF, xét $X \rightarrow A \in S$, với điều kiện X không chứa khóa của S và $A \notin X$. Thay thế S bởi S1, S2 với $S1 = A \cup \{X\}$, $S2 = \{S\} \setminus A$.

B3. Lặp (B2) cho đến khi $\forall S \in KQ$ đều ở BCNF

KQ gồm các sơ đồ con của phép tách yêu cầu