

Chuẩn hóa dữ liệu

Biên soạn: TS. Nguyễn Quốc Tuấn
Bm. Mạng và Các HTTT

Mục đích của chuẩn hóa

- 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”

Nội dung

- Phép tách các lược đồ quan hệ
- Các dạng chuẩn

Phép tách các lược đồ quan hệ

- **Mục đích**

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

- **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:

$$r = \Pi_{R_1}(r) \bowtie \Pi_{R_2}(r) \bowtie \dots \bowtie \Pi_{R_k}(r)$$

- **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 $t1[Y] \neq t2[Y]$, 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

- $R(\text{MONHOC}, \text{SOTIET}, \text{LOP}, \text{GV}, \text{HOCVI}, \text{DC})$
 - Kiểm tra: $R_1(\text{MONHOC}, \text{SOTIET}, \text{LOP}, \text{GV})$, $R_2(\text{GV}, \text{HOCVI}, \text{DC})$
- $F = \{\text{MONHOC} \rightarrow \text{SOTIET}; \text{MONHOC}, \text{LOP} \rightarrow \text{GV}; \text{GV} \rightarrow \text{HOCVI}, \text{DC}\}$

	MONHOC	SOTIET	LOP	GV	HOCVI	DC
R ₁	a ₁	a ₂	a ₃	a ₄	b ₁	b ₂
R ₂	b ₃	b ₄	b ₅	a ₄	a ₅	a ₆

$\text{GV} \rightarrow \text{HOCVI}, \text{DC}$

	MONHOC	SOTIET	LOP	GV	HOCVI	DC
R ₁	a ₁	a ₂	a ₃	a ₄	a ₅	a ₆
R ₂	b ₃	b ₄	b ₅	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 $X \rightarrow Y \in F^+$:

$$XY \subseteq R_i.$$

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 .

Các dạng chuẩn

- Mục đích:
 - Mỗi dạng chuẩn đảm bảo ngăn ngừa (giảm thiểu) một số các dạng dư thừa hay dị thường dữ liệu
- Các dạng chuẩn hay sử dụng:
 - Dạng chuẩn 1 (1NF)
 - Dạng chuẩn 2 (2NF)
 - Dạng chuẩn 3 (3NF)
 - Dạng chuẩn Boyce Codd (BCNF)

Các dạng chuẩn

- Dạng chuẩn 1:
 - 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.
- Dạng chuẩn 2:
 - 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.
- Dạng chuẩn 3:
 - 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.
- Dạng chuẩn BCNF
 - Lược đồ quan hệ R được gọi là thuộc dạng chuẩn BCNF khi và chỉ khi:
 - PTH không hiển nhiên $X \rightarrow Y$ đúng trên R thì X là siêu khóa của R.

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

$F = \{A \rightarrow B, ACD \rightarrow E, EF \rightarrow G\}$

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

B1. không lập được quan hệ nào mới.

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

B3. $A \rightarrow B \quad \Rightarrow R1(AB)$

$ACD \rightarrow E \quad \Rightarrow R2(ACDE)$

$EF \rightarrow G \quad \Rightarrow R3(EFG)$

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 $R(A,B,C,D,E,F,G)$.
 $F = \{A \rightarrow B, ACD \rightarrow E, EF \rightarrow G\}$

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

Ví dụ