

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

CƠ SỞ DỮ LIỆU PHÂN TÁN

THIẾT KẾ
CƠ SỞ DỮ LIỆU PHÂN TÁN
Phân mảnh dọc

Ts. Phạm Thế Quế

Phân mảnh dọc

Định nghĩa

Phân mảnh dọc quan hệ R sinh ra các mảnh R_1, R_2, \dots, R_r , sao cho mỗi mảnh chứa một tập con các thuộc tính của quan hệ R và khoá của nó.

Mục đích

Phân chia quan hệ R thành các mảnh nhỏ hơn là để cho nhiều ứng dụng có thể thực hiện trên một mảnh tối ưu, giảm thiểu thời gian thực hiện ứng dụng. Nâng cao hiệu năng xử lý đồng thời.

Tối ưu ?

Một phân mảnh tối ưu là phân mảnh sinh ra một lược đồ phân mảnh cho phép giảm tối đa thời gian thực thi các ứng dụng chạy trên phân mảnh đó

Ma trận giá trị sử dụng thuộc tính

- ❑ $R(A_1, A_2, \dots, A_n)$ quan hệ toàn cục
- ❑ $Q = \{q_1, q_2, \dots, q_m\}$ tập các ứng dụng
- ❑ Ma trận giá trị sử dụng thuộc tính định nghĩa như sau:

$$A = (\text{use}(q_i, A_j))_{m \times n}$$

$$\text{use}(q_i, A_j) = \begin{cases} 1 & \text{Nếu } q_i \text{ tham chiếu đến thuộc tính } A_j \\ 0 & \text{Ngược lại} \end{cases}$$

$$i = 1..m \text{ và } j = 1..n$$

$$n = |\Omega| \text{ và } m = |Q|$$

Mã trận giá trị sử dụng thuộc tính

A =

	A1	A2	A_n
q1	Use(q1,A1)	Use(q1,A2)		Use(q1,A_n)
q2	Use(q2,A1)	Use(q2,A2)		Use(q2,A_n)
....
q_m	Use(q_m,A1)	Use(q_m,A2)		Use(q_m,A_n)

Ví dụ ma trận giá trị sử dụng thuộc tính

Quan hệ: PROJ (PNO, PNAME, BUDGET, LOC)

Tập các ứng dụng:

q1: Kinh phí của dự án khi biết mã dự án

SELECT BUDGET FROM PROJ WHERE PNO = Value

q2: Tên và kinh phí của tất các dự án

SELECT PNAME, BUDGET FROM PROJ

q3: Tìm tên các dự án khi biết thành phố

SELECT PNAME FROM PROJ WHERE LOC = Value

q4: Tổng kinh phí của các dự án tại mỗi thành phố

SELECT SUM(BUDGET) FROM PROJ WHERE LOC = Value

Ví dụ ma trận giá trị sử dụng thuộc tính

Ký hiệu: A1= PNO, A2=PNAME, A3=BUDGET, A4=LOC

q1: SELECT A3 FROM PROJ WHERE A1= Value

q2: SELECT A2, A3 FROM PROJ

q3: SELECT A2 FROM PROJ WHERE A4 = Value

q4: SELECT SUM(A3) FROM PROJ WHERE A4= Value

$$\mathbf{A} = \begin{matrix} & \begin{matrix} A_1 & A_2 & A_3 & A_4 \end{matrix} \\ \begin{matrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{matrix} & \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{matrix}$$

Ma trận lực hút AA(Attribute Affinity Matrix)

- ❑ R (A_1, A_2, \dots, A_n) quan hệ toàn cục
- ❑ $Q = \{q_1, q_2, \dots, q_m\}$ tập các ứng dụng
- ❑ Bảng tần số ứng dụng trên các site: $S = \{S_1, S_2, \dots, S_t\}$
- ❑ Khi đó $AA = (aff(A_i, A_j))_{n \times n}$ Ma trận lực hút

$$aff(A_i, A_j) = \sum_{k: [(use(q_k, A_i) \wedge (use(q_k, A_j)) = 1 \forall S_l]} \sum_{l} ref_l(q_k) acc_l(q_k)$$

Trong đó:

- ❑ $ref_l(q_k)$ là tần số truy suất của q_k trên (A_i, A_j) tại vị trí S_l
- ❑ $acc_l(q_k)$ là tần số truy suất của q_k tại vị trí S_l

$Ma\ tr\grave{a}n\ l\grave{y}c\ h\grave{u}t\ AA(Attribute\ Affinity\ Matrix)$

$AA =$

	A1	A2	An
A1	$aff(A1,A1)$	$aff(A1,A2)$		$aff(A1,A_n)$
A2	$aff(A2,A1)$	$aff(A2,A2)$		$aff(A2,A_n)$
....
A_n	$aff(A_n,A1)$	$aff(A_n,A2)$		$aff(A_n,A_n)$

Ví dụ ma trận lực hút AA

- ❑ Giả sử $\text{ref}_l(q_k) = 1$ cho tất cả q_k và S_l
- ❑ Giả sử tần số các ứng dụng trên các Site là:

Site1

$$\text{acc}_1(q_1)=15$$

$$\text{acc}_1(q_2)=5$$

$$\text{acc}_1(q_3)=25$$

$$\text{acc}_1(q_4)=3$$

Site2

$$\text{acc}_2(q_1)=20$$

$$\text{acc}_2(q_2)=0$$

$$\text{acc}_2(q_3)=25$$

$$\text{acc}_2(q_4)=0$$

Site3

$$\text{acc}_3(q_1)=10$$

$$\text{acc}_3(q_2)=0$$

$$\text{acc}_3(q_3)=25$$

$$\text{acc}_3(q_4)=0$$

Ví dụ ma trận lực hút AA

Site1

$$acc_1(q_1)=15$$

$$acc_1(q_2)=5$$

$$acc_1(q_3)=25$$

$$acc_1(q_4)=3$$

Site2

$$acc_2(q_1)=20$$

$$acc_2(q_2)=0$$

$$acc_2(q_3)=25$$

$$acc_2(q_4)=0$$

Site3

$$acc_3(q_1)=10$$

$$acc_3(q_2)=0$$

$$acc_3(q_3)=25$$

$$acc_3(q_4)=0$$

$$aff(A_1, A_3) = \sum_{k=1}^l \sum_{l=1}^3 acc_1(q_k) = acc_1(q_1) + acc_2(q_1) + acc_3(q_1) = 45$$

$$\mathbf{A} = \begin{matrix} & \mathbf{A}_1 & \mathbf{A}_2 & \mathbf{A}_3 & \mathbf{A}_4 \\ \mathbf{q}_1 & \begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix} \\ \mathbf{q}_2 & \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix} \\ \mathbf{q}_3 & \begin{bmatrix} 0 & 1 & 0 & 1 \end{bmatrix} \\ \mathbf{q}_4 & \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix} \end{matrix}$$

$$\mathbf{AA} =$$

	\mathbf{A}_1	\mathbf{A}_2	\mathbf{A}_3	\mathbf{A}_4
\mathbf{A}_1	45	0	45	0
\mathbf{A}_2	0	80	5	75
\mathbf{A}_3	45	5	53	3
\mathbf{A}_4	0	75	3	78

Ví dụ ma trận lực hút AA

AA =

	A₁	A₂	A₃	A₄
A₁	45	0	45	0
A₂	0	80	5	75
A₃	45	5	53	3
A₄	0	75	3	78

Thuật toán tụ nhóm

- Số đo đóng góp của thuộc tính A_k khi đặt vào A_i và A_j
 A_k, A_i, A_j ; A_i, A_k, A_j ; A_i, A_j, A_k

$$cont(A_i, A_k, A_j) = 2[bond(A_i, A_k) + bond(A_k, A_j) - bond(A_i, A_j)]$$

- Cầu nối (bond) 2 thuộc tính A_x và A_y định nghĩa như sau:

$$bond(A_x, A_y) = \sum_{z=1}^n aff(A_z, A_x) aff(A_z, A_y)$$

- Độ đo cầu nối giữa hai thuộc tính được tính là tổng của tích 2 phần tử cùng hàng của hai cột. Vì ma trận AA đối xứng, có thể thực hiện tương tự theo hàng.

Thuật toán tụ nhóm

$$\text{cont}(A_1, A_4, A_2) = 2[\text{bond}(A_1, A_4) + \text{bond}(A_4, A_2) - \text{bond}(A_1, A_2)]$$

$$\begin{aligned}\text{bond}(A_1, A_4) = & \text{aff}(A1, A1) * \text{aff}(A1, A4) + \\ & \text{aff}(A2, A1) * \text{aff}(A2, A4) + \\ & \text{aff}(A3, A1) * \text{aff}(A3, A4) + \\ & \text{aff}(A4, A1) * \text{aff}(A4, A4)\end{aligned}$$

$$\text{bond}(A_1, A_4) = 135$$

$$\text{bond}(A_4, A_2) = 11865$$

$$\text{bond}(A_1, A_2) = 225$$

	A1	A2	A3	A4
A1	45	0	45	0
A2	0	80	5	75
A3	45	5	53	3
A4	0	75	3	78

$$\text{cont}(A_1, A_4, A_2) = 2 * 135 + 2 * 11865 - 2 * 225 = 23550$$

Thuật toán tụ nhóm BEA (Bond Energy Algorithm)

Nhóm các thuộc tính của quan hệ toàn cục bằng cách hoán vị các hàng và các cột của ma trận AA, sao cho số đo hấp dẫn **cont()** là lớn nhất. Kết quả sẽ là một ma trận tụ hấp dẫn CA (Cluster Affinity). Thuật toán gồm 3 bước:

- ❑ *Bước 1:* Đặt cột 1 và 2 của AA vào cột 1&2 trong CA.
- ❑ *Bước 2:* Giả sử có i cột đã được đặt vào CA. Lấy lần lượt một trong $(n-i)$ cột còn lại của AA, đặt vào cột thứ $(i+1)$ của CA, sao cho số đo AM tại vị trí đó là lớn nhất.
- ❑ *Bước 3:* Sắp thứ tự hàng theo thứ tự cột

Thuật toán tụ nhóm BEA (Bond Energy Algorithm)

Input: AA:ma trận hấp dẫn thuộc tính

Output: CA : Ma trận hấp dẫn tụ nhóm

Begin

(1) $CA(.,1) \leftarrow AA(.,1)$

(2) $CA(.,2) \leftarrow AA(.,2)$

(3) $index \leftarrow 3$

(4) **While** $index \leq n$ **do** {chọn vị trí tốt nhất cho thuộc tính AA_{index} }

(5) **Begin**

(6) **For** i **from** 1 **to** $index-1$ **by** 1 **do**

(7) Tính $cont(A_{i-1}, A_{index}, A_i)$

(8) **End-for**

(9) Tính $cont(A_{index-1}, A_{index}, A_{index+1})$

(10) $loc \leftarrow$ vị trí được đặt bởi giá trị $cont$ lớn nhất

(11) **For** j **from** $index$ **to** loc **By** -1 **do**

(12) $CA(.,j) \leftarrow AA(.,j-1)$

(13) **End-for**

(14) $CA(.,loc) \leftarrow AA(.,index)$

(15) $index \leftarrow index+1$

(16) **End-While** |

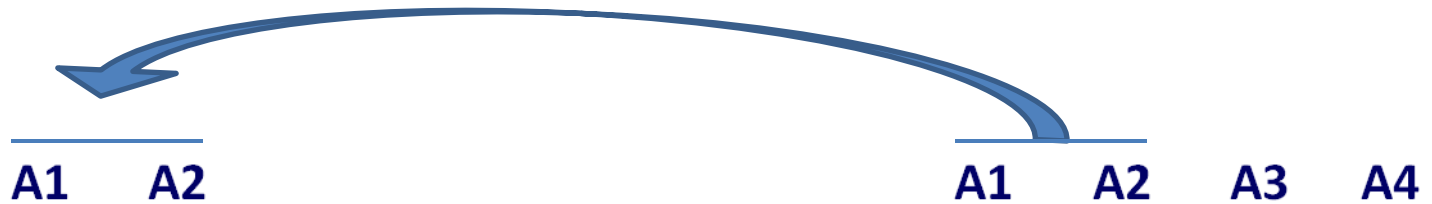
End {BEA}

Ví dụ

Chép cột 1 và cột 2 ma trận AA vào ma trận CA

$$(1) CA(*,1) \leftarrow AA(*,1)$$

$$(2) CA(*,2) \leftarrow AA(*,2)$$



CA =

	A1	A2		
A1	45	0		
A2	0	80		
A3	45	5		
A4	0	75		

AA =

	A1	A2	A3	A4
A1	45	0	45	0
A2	0	80	5	75
A3	45	5	53	3
A4	0	75	3	78

Ví dụ

index=3

While *index* $\leq n$ **do**

index ≤ 4 {thỏa mãn}

For *i* **from** 1 **to** *index* – 1 **by** 1 **do**

Tính $\text{cont}(A_{i-1}, A_{\text{index}}, A_i)$

i=1 thứ tự (0-3-1): $\text{cont}(A_0, A_3, A_1) = 8820$

i=2 thứ tự (1-3-2): $\text{cont}(A_1, A_3, A_2) = 10150$

End – for


Điều kiện biên, thứ tự (2-3-4): $\text{cont}(A_2, A_3, A_4) = 1780$

loc =2 thứ tự (1-3-2) có $\text{cont} = 10150$ lớn nhất

For *j* **from** *index* **to** *Loc* **by** – 1 **do** {xáo trộn hai ma trận}

$CA(*, j) := AA(*, j-1)$

Ví dụ



CA =

	A1	A3	A2	
A1	45	45	0	
A2	0	5	80	
A3	45	53	5	
A4	0	3	75	

AA =

	A1	A2	A3	A4
A1	45	0	45	0
A2	0	80	5	75
A3	45	5	53	3
A4	0	75	3	78

Đặt A_3 giữa A_1 và A_2

index=4

While *index* $\leq n$ **do**

index ≤ 4 {thỏa mãn}

For *i* **from** 1 **to** *index* - 1 **by** 1 **do**

Tính $\text{cont}(A_{i-1}, A_{\text{index}}, A_i)$

i=1 thứ tự (0-4-1): $\text{cont}(A_0, A_4, A_1) = 270$

i=2 thứ tự (1-4-2): $\text{cont}(A_1, A_4, A_3) = - 5208$

i=3 thứ tự (2-4-3): $\text{cont}(A_3, A_4, A_2) = 23698$

End – for

Điều kiện biên, thứ tự (3-4-5): $\text{cont}(A_2, A_4, A_0) = 23730$

loc =4 thứ tự (1-3-2) có cont =10150 lớn nhất

CA =

	A1	A3	A2	<u>A4</u>
A1	45	45	0	0
A2	0	5	80	75
A3	45	53	5	3
A4	0	3	75	78

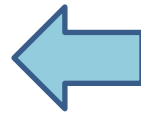
AA =

	A1	A2	A3	<u>A4</u>
A1	45	0	45	0
A2	0	80	5	75
A3	45	5	53	3
A4	0	75	3	78

Đặt A_4 bên phải A_2

CA =

	A1	A3	A2	A4
A1	45	45	0	0
A3	45	53	5	3
A2	0	5	80	75
A4	0	3	75	78



CA =

	A1	A3	A2	A4
A1	45	45	0	0
A2	0	5	80	75
A3	45	53	5	3
A4	0	3	75	78

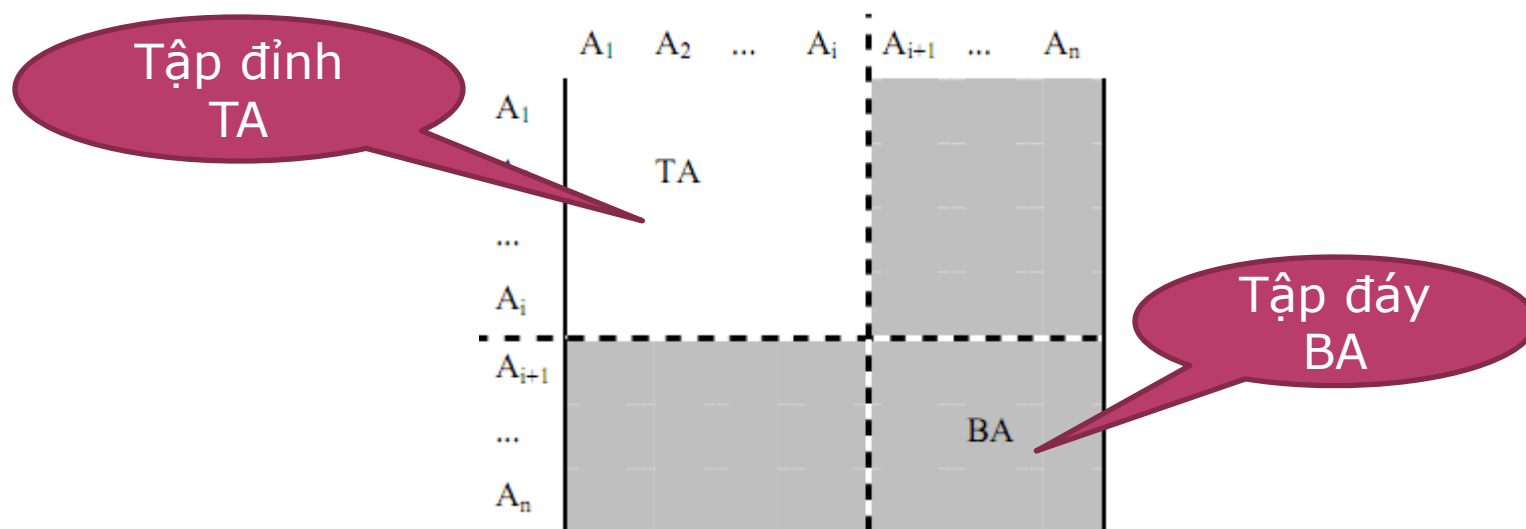
Thuật toán phân mảnh dọc

- ❑ Độ đo cầu nối giữa hai thuộc tính được tính là tổng của tích 2 phần tử cùng hàng của hai cột. Vì ma trận AA đối xứng, có thể thực hiện tương tự theo hàng.
- ❑ Trong bước khởi gán, cột 1 và 2 được đặt vào vị trí 1&2 trong CA, vì A2 có thể đặt ở bên trái hoặc phải của A1.
- ❑ Nếu A_j là thuộc tính tận trái trong ma trận CA, kiểm tra đóng góp khi đặt thuộc tính A_k vào bên trái của A_j , khi đó $\text{bond}(A_0, A_k) = \text{bond}(A_0, A_j) = 0$,
- ❑ Nếu A_j là thuộc tính tận phải đã được đặt trong ma trận CA và đang kiểm tra đóng góp khi đặt thuộc tính A_k vào bên phải của A_j , Khi đó $\text{bond}(A_j, A_{k+1}) = \text{bond}(A_k, A_{k+1}) = 0$.

Thuật toán phân mảnh dọc

Xét ma trận tự lực CA

- ❑ $TA = \{A_1, A_2, \dots, A_i\}$ ở góc trái cao nhất gọi là tập đỉnh (Top)
- ❑ $BA = \{A_{i+1}, A_{i+2}, \dots, A_n\}$ ở góc phải thấp nhất gọi là tập đáy (Bottom)



Thuật toán phân mảnh dọc

Ký hiệu	Ý nghĩa
$Q = \{q_1, q_2, \dots, q_n\}$	Tập các ứng dụng.
$AQ(q_i) = \{A_j \mid \text{use}(q_i, A_j) = 1\}$	Tập các thuộc tính được truy xuất bởi ứng dụng q_i
$TQ = \{q_i \mid AQ(q_i) \subseteq TA\}$	Tập các ứng dụng chỉ truy xuất trên các thuộc tính TA
$BQ = \{q_i \mid AQ(q_i) \subseteq BA\}$	Tập các ứng dụng chỉ truy xuất trên các thuộc tính BA
$OQ = Q - \{TQ \cup BQ\}$	Tập các ứng dụng truy xuất trên cả BA và TA

Thuật toán phân mảnh dọc

Ký hiệu

Ý nghĩa

$$CQ = \sum_{q_i \in \Omega} \sum_{\forall S_j} ref_j(q_i) acc_j(q_i)$$

Tổng chi phí truy xuất của tất cả các ứng dụng trên tất cả các vị trí

$$CTQ = \sum_{qi \in TQ} \sum_{\forall S_j} ref_j(q_i).acc_j(q_i)$$

Tổng số các truy cập đến các thuộc tính bởi các ứng dụng chỉ truy cập TA

$$CBQ = \sum_{qi \in BQ} \sum_{\forall S_j} ref_j(q_i).acc_j(q_i)$$

Tổng số các truy cập đến các thuộc tính bởi các ứng dụng chỉ truy cập BA

$$COQ = \sum_{qi \in OQ} \sum_{\forall S_j} ref_j(q_i).acc_j(q_i)$$

Tổng số các truy cập đến các thuộc tính bởi ứng dụng truy cập cả TA và BA

Thuật toán phân mảnh dọc

Bài toán tối ưu hóa phân mảnh chính là bài toán xác định định một điểm : $1 \leq z \leq n$ sao cho :

$$z = \text{CTQ}^* \text{CBQ} - \text{COQ}^2 \text{ là lớn nhất}$$

Thuật toán phân mảnh dọc

CA =

	A1	A3	A2	A4
A1	45	45	0	0
A3	45	53	5	3
A2	0	5	80	75
A4	0	3	75	78

A=

	A ₁	A ₂	A ₃	A ₄
q ₁	1	0	1	0
q ₂	0	1	1	0
q ₃	0	1	0	1
q ₄	0	0	1	1

Vị trí 1:

$$TA = \{ A_1 \},$$

$$BA = \{ A_3, A_2, A_4 \},$$

$$TQ = \{ \},$$

$$BQ = \{ q_2, q_3, q_4 \},$$

$$OQ = \{ q_1 \}$$

$$CTQ = 0;$$

$$CBQ = acc_1(q_2) + acc_2(q_2) + acc_3(q_2) + acc_1(q_3) + acc_2(q_3) + acc_3(q_3) + acc_1(q_4) + acc_2(q_4) + acc_3(q_4) = 83$$

$$COQ = acc_1(q_1) + acc_2(q_1) + acc_3(q_1) = 45$$

$$Z = CTQ * CBQ - COQ^2 = -2025$$

Site1

$$acc_1(q_1)=15$$

$$acc_1(q_2)=5$$

$$acc_1(q_3)=25$$

$$acc_1(q_4)=3$$

Site2

$$acc_2(q_1)=20$$

$$acc_2(q_2)=0$$

$$acc_2(q_3)=25$$

$$acc_2(q_4)=0$$

Site3

$$acc_3(q_1)=10$$

$$acc_3(q_2)=0$$

$$acc_3(q_3)=25$$

$$acc_3(q_4)=0$$

Thuật toán phân mảnh dọc

Vị trí 2:

CA =

	A1	A3	A2	A4
A1	45	45	0	0
A3	45	53	5	3
A2	0	5	80	75
A4	0	3	75	78

	A ₁	A ₂	A ₃	A ₄
q ₁	1	0	1	0
q ₂	0	1	1	0
q ₃	0	1	0	1
q ₄	0	0	1	1

$$TA = \{A_1, A_3\}, TQ = \{q_1\},$$

$$BA = \{A_2, A_4\}, BQ = \{q_3\},$$

$$OQ = \{q_2, q_4\}$$

$$CTQ_2 = acc_1(q_1) + acc_2(q_1) + acc_3(q_1) = 45$$

$$CBQ_2 = acc_1(q_3) + acc_2(q_3) + acc_3(q_3) = 75$$

$$COQ_2 = acc_1(q_2) + acc_2(q_2) + acc_3(q_2) + acc_1(q_4) + acc_2(q_4) + acc_3(q_4) = 8$$

$$Z = CTQ * CBQ - COQ^2 = 3311$$

Site1

$$acc_1(q_1)=15$$

$$acc_1(q_2)=5$$

$$acc_1(q_3)=25$$

$$acc_1(q_4)=3$$

Site2

$$acc_2(q_1)=20$$

$$acc_2(q_2)=0$$

$$acc_2(q_3)=25$$

$$acc_2(q_4)=0$$

Site3

$$acc_3(q_1)=10$$

$$acc_3(q_2)=0$$

$$acc_3(q_3)=25$$

$$acc_3(q_4)=0$$

Thuật toán phân mảnh dọc

CA =

	A1	A3	A2	A4
A1	45	45	0	0
A3	45	53	5	3
A2	0	5	80	75
A4	0	3	75	78

	A ₁	A ₂	A ₃	A ₄
q ₁	1	0	1	0
q ₂	0	1	1	0
q ₃	0	1	0	1
q ₄	0	0	1	1

Vị trí 3:

$$TA = \{A_1, A_3, A_2\}, \quad TQ = \{q_2, q_1\},$$

$$BA = \{A_4\}, \quad BQ = \{\},$$

$$OQ = \{q_4, q_3\}$$

$$CTQ_3 = acc_1(q_1) + acc_2(q_1) + acc_3(q_1)$$

$$acc_1(q_2) + acc_2(q_2) + acc_3(q_2) = 50$$

$$CBQ_3 = 0$$

$$COQ_3 = acc_1(q_3) + acc_2(q_3) + acc_3(q_3) +$$

$$acc_1(q_4) + acc_2(q_4) + acc_3(q_4) = 78$$

$$Z = CTQ * CBQ - COQ^2 = -6084$$

Site1

$$acc_1(q_1)=15$$

$$acc_1(q_2)=5$$

$$acc_1(q_3)=25$$

$$acc_1(q_4)=3$$

Site2

$$acc_2(q_1)=20$$

$$acc_2(q_2)=0$$

$$acc_2(q_3)=25$$

$$acc_2(q_4)=0$$

Site3

$$acc_3(q_1)=10$$

$$acc_3(q_2)=0$$

$$acc_3(q_3)=25$$

$$acc_3(q_4)=0$$

Thuật toán phân mảnh dọc

❑ Vị trí 1: $Z = -2025$

❑ Vị trí 2: $Z = 3311$

❑ Vị trí 3: $Z = -6084$

❑ Như vậy vị trí 2 có chi phí là lớn nhất

❑ Quan hệ PROJ chia thành 2 mảnh:

$PROJ_1 \{A_1, A_3\} = PROJ_1 \{\underline{PNO}, BUDGET\}$

$PROJ_2 \{A_1, A_2, A_4\} = PROJ_2 \{\underline{PNO}, PNAME, LOC\}$

Thuật toán phân mảnh dọc

PROJ

PNO	PNAME	BUDGET	LOG
P1	Instrumentation	150000	Montreal
P2	Database Develop	135000	NewYork
P3	CAD/CAM	250000	NewYork
P4	Maintenance	310000	Paris

PROJ1

PNO	BUDGET
P1	150000
P2	135000
P3	250000
P4	310000

PROJ2

PNO	PNAME	LOG
P1	Instrumentation	Montreal
P2	Database Develop	NewYork
P3	CAD/CAM	NewYork
P4	Maintenance	Paris