บทที่ 4

ดัชนีบิตแมปแบบคู่กัน

ในงานวิจัยเกี่ยวกับเรื่องดัชนีบิตแมปที่ผ่านมา เราจะเห็นว่ามีการพัฒนาขั้นตอนการสร้างดัชนีบิตแมปต่อ ๆ กันมาเพื่อให้สอดคล้องกับความต้องการ และให้มีประสิทธิภาพมากกว่าเดิมทั้งในเรื่องของพื้นที่ที่ใช้ในการจัดเก็บดัชนีและเวลาที่ใช้ในการสอบถาม ดัชนีบิตแมปแบบพื้นฐานใช้เวลาในการสอบถามแบบค่าเท่ากันน้อยที่สุด แต่ใช้พื้นที่ในการจัดเก็บดัชนีมากที่สุด ในทางกลับกัน ดัชนีบิตแมปแบบเข้ารหัสใช้พื้นที่ในการจัดเก็บดัชนีน้อยที่สุด แต่ใช้เวลาในการสอบถามแบบค่าเท่ากันมากที่สุด ต่อมาจึงมีผู้คิดค้นดัชนีบิตแมปแบบช่วงขึ้น ซึ่งใช้พื้นที่เพียงครึ่งหนึ่งของดัชนีบิตแมปแบบพื้นฐาน และใช้เวลาในการสอบถามแบบค่าเท่ากันน้อยกว่าดัชนีบิตแมปแบบเข้ารหัส กล่าวคือ ในการสอบถามแบบค่าเท่ากันบนดัชนีบิตแมปแบบช่วงมีการอ่าน 1 ถึง 2 บิตแมปเวกเตอร์และใช้การดำเนินการตรรกะ 0 ถึง 2 ครั้ง แต่ดัชนีบิตแมปแบบเข้ารหัสต้องอ่านทุกบิตแมปเวกเตอร์ และดำเนินการตรรกะหลายครั้ง อย่างไรก็ตาม ดัชนีบิตแมปแบบช่วงยังใช้พื้นที่มากเกินไป ต่อมาจึงได้มีผู้คิดค้นดัชนีบิตแมปแบบกระจายขึ้นมา ซึ่งสามารถลดพื้นที่ลงจากดัชนีบิตแมปแบบช่วงได้อีก โดยที่ในการสอบถามแบบค่าเท่ากันบนดัชนีบิตแมปแบบกระจายมีการอ่าน 2 บิตแมปเวกเตอร์ และดำเนินการตรรกะครั้งเดียวเสมอ

ในงานวิจัยนี้ ผู้วิจัยนำเสนอเทคนิคในการทำดัชนีบิตแมปแบบใหม่ ซึ่งใช้พื้นที่ในการจัดเก็บดัชนีน้อยกว่าดัชนีบิตแมปแบบกระจาย โดยที่การสอบถามแบบค่าเท่ากันยังคงอ่าน 2 บิตแมปเวกเตอร์ และดำเนินการตรรกะเพียงครั้งเดียวเช่นเดิม ดัชนีบิตแมปแบบใหม่ที่คิดค้นขึ้นนี้มีชื่อเรียกว่า ดัชนีบิตแมปแบบคู่กัน(Dual Bitmap Index) แนวคิดในการหาสูตรทั่วไปสำหรับลงรหัสและสอบถามแบบค่าเท่ากัน[21,22,23,26] สามารถพิจารณาได้ดังหลักการที่จะกล่าวต่อไป

4.1 นิยามคำศัพท์และสัญลักษณ์ที่เกี่ยวข้อง

1) สำหรับจำนวนเต็ม $n \geq 0$ n แฟกทอเรียล (n factorial) เขียนแทนด้วย $n!$ ถูกกำหนดให้ดังนี้ [7]

$$0! = 1$$

$$n! = n(n-1)!$$

2) กำหนดให้ $C(n,r)$ แทน จำนวนวิธีในการเลือกของ r สิ่งจากของทั้งหมด n สิ่ง[7] จะได้ว่า สำหรับ $0 \le r \le n$

$$C(n,r) = \frac{n!}{(n-r)!r!}$$

บางครั้งเราอาจเขียนแทนสัญลักษณ์ $C(n,r)$ ด้วย ${}^{n}C_{r}$ หรือ $\binom{n}{r}$ ซึ่งสัญลักษณ์ $\binom{n}{r}$ เป็นที่นิยมใช้กันมาก

3) $\lceil \mathbf{x} \rceil$ คือ จำนวนเต็มที่น้อยที่สุดที่มากกว่าหรือเท่ากับ $\mathbf{x}$

4) **บิตแมปเวกเตอร์** คือ แถวของข้อมูลที่แต่ละสมาชิกแต่ละตัวมีค่าเป็น ‘0’ หรือ ‘1’ ซึ่งบนตารางดัชนีบิตแมปจะประกอบด้วยหลายบิตแมปเวกเตอร์ แต่ละบิตแมปเวกเตอร์จะมีความยาวเท่ากับจำนวนแถวทั้งหมดของข้อมูล เช่น ข้อมูลมีทั้งหมด 1,000,000 แถว เมื่อนำแอทริบิวต์มาสร้างเป็นดัชนีบิตแมป แล้วแต่ละบิตแมปเวกเตอร์จะมีความยาวเท่ากับ 1,000,000 บิต

5) **คาร์ดินอลิตี้** คือ จำนวนค่าที่เป็นไปได้ของแอทริบิวต์ เช่น แอทริบิวต์เพศมีค่าที่เป็นไปได้ คือ ชายและหญิง จึงได้ว่า แอทริบิวต์เพศมีคาร์ดินอลิตี้เท่ากับ 2

4.2 การคิดค้นดัชนีบิตแมปแบบคู่กัน

จากการศึกษางานวิจัยที่ผ่านมาเราจะเห็นว่า ดัชนีบิตแมปแบบช่วงและดัชนีบิตแมปแบบกระจายเป็นเทคนิคที่คิดค้นขึ้น เพื่อลดพื้นการจัดเก็บดัชนีลงจากพื้นที่ที่ใช้ในการจัดเก็บดัชนีบิตแมปแบบพื้นฐาน และเพิ่มประสิทธิภาพในการสอบถามให้มากกว่าดัชนีบิตแมปแบบเข้ารหัส ซึ่งการพัฒนาจากดัชนีบิตแมปแบบช่วงมาเป็นดัชนีบิตแมปแบบกระจาย มีการลดจำนวนบิตแมปที่ใช้แทนแต่ละค่าของแอทริบิวต์ โดยแต่ละค่าจะถูกแทนด้วย 2 บิตแมปเวกเตอร์ ทำให้การสอบถามมีการอ่าน 2 บิตแมปเวกเตอร์ แต่รูปแบบการลงรหัสดัชนีบิตแมปแบบกระจายยังมีข้อด้อย ในส่วนของบางบิตแมปเวกเตอร์ที่ใช้ประโยชน์ไม่คุ้มค่า โดยเฉพาะ Z^0 ดังนั้น ในการคิดค้นเทคนิคการลงรหัสดัชนีบิตแมปแบบคู่กัน จึงใช้หลักการที่ว่า แต่ละค่าของแอทริบิวต์ถูกแทนด้วย 2 บิตแมปเวกเตอร์ และทำให้ทุกบิตแมปเวกเตอร์ถูกใช้อย่างคุ้มค่ามากที่สุด

จากหลักการที่กล่าวไว้ข้างต้น ทำให้เราได้สูตรทั่วไปในการหาจำนวนบิตแมปเวกเตอร์ที่ใช้ในการสร้างดัชนีบิตแมปแบบคู่กัน ดังทฤษฎีบท 4.1

ทฤษฎีบท 4.1 ให้ C แทนคาร์ดินอลิตี้ของแอทริบิวต์ที่เลือกมาทำดัชนี และ n แทนจำนวนบิตแมปเวกเตอร์ที่น้อยที่สุดที่ใช้ในการสร้างดัชนีบิตแมปแบบคู่กัน(Dual Bitmap Index) ซึ่งแต่ละค่าของแอทริบิวต์ถูกแทนด้วย 2 บิตแมปเวกเตอร์ แล้ว

$$n = \left\lceil \sqrt{2C + \frac{1}{4}} + \frac{1}{2} \right\rceil \qquad \text{...............(4.1)}$$

<u>พิสูจน์</u> ภายใต้เงื่อนไขที่ว่า ต้องการให้มีการอ่าน 2 บิตแมปเวกเตอร์และดำเนินการตรรกะเพียงครั้งเดียวในการสอบถามแบบค่าเท่ากันแต่ละครั้ง แต่ละค่าของแอทริบิวต์จึงถูกแทนด้วย 2 บิตแมปเวกเตอร์ จากทั้งหมด n บิตแมปเวกเตอร์ ดังนั้นจึงต้องมีวิธีในการเลือก 2 บิตแมปเวกเตอร์ จาก n บิตแมปเวกเตอร์ มากกว่าหรือเท่ากับ C วิธี ดังอสมการ

$$\binom{n}{2} \geq C$$

$$\frac{n!}{(n-2)!2!} \geq C$$

$$\frac{n(n-1)}{2} \geq C$$

$$n(n-1) \geq 2C$$

$$n^2 - n \geq 2C$$

$$n^2 - 2\cdot\frac{1}{2}n + \left(\frac{1}{2}\right)^2 \geq 2C + \frac{1}{4}$$

$$\left(n - \frac{1}{2}\right)^2 \geq 2C + \frac{1}{4}$$

$$n - \frac{1}{2} \leq -\sqrt{2C + \frac{1}{4}} \quad \text{หรือ} \quad n - \frac{1}{2} \geq \sqrt{2C + \frac{1}{4}}$$

$$n \leq -\sqrt{2C + \frac{1}{4}} + \frac{1}{2} \quad \text{หรือ} \quad n \geq \sqrt{2C + \frac{1}{4}} + \frac{1}{2}$$

เนื่องจาก เราต้องการจำนวนบิตแมปเวกเตอร์ที่เป็นจำนวนเต็มบวกที่น้อยที่สุด
จึงได้ว่า

$$n = \left\lceil \sqrt{2C + \frac{1}{4}} + \frac{1}{2} \right\rceil \qquad \square$$

<u>ตัวอย่างที่ 1</u> สมมติให้ แอทริบิวต์ที่เลือกมาทำดัชนีบิตแมปมีคาร์ดินอลิตี้ C = 15
จากสมการ (1) แทนค่า C = 15

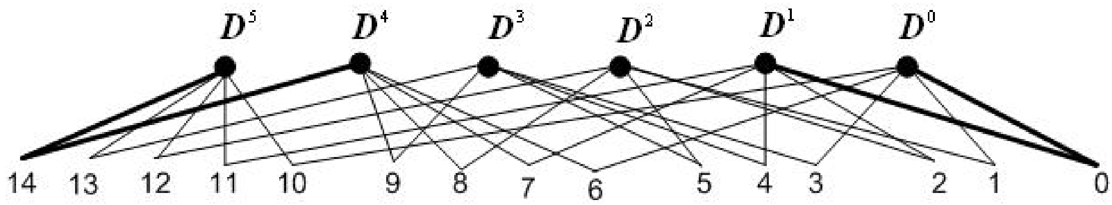
จะได้

$$n = \left\lceil \sqrt{(2\times15)+\frac{1}{4}}+\frac{1}{2} \right\rceil$$

$$n = \left\lceil \sqrt{30.25}+0.5 \right\rceil$$

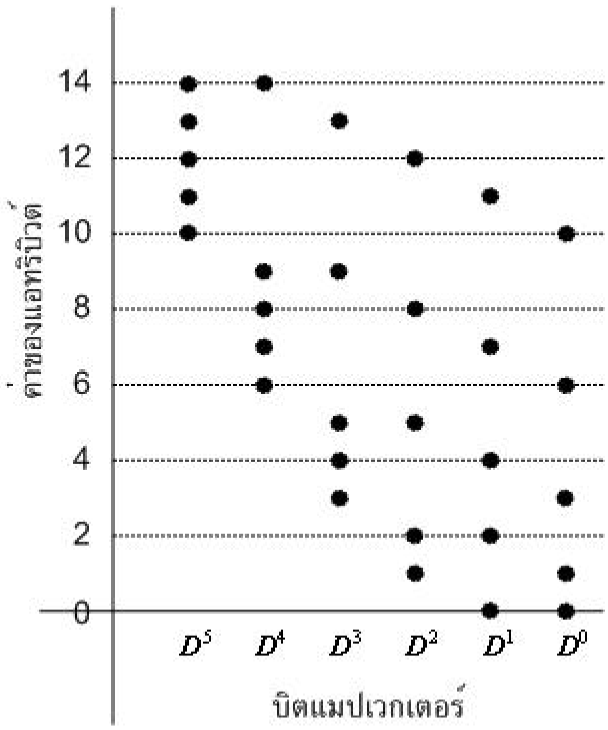
$$n = \lceil 5.5+0.5 \rceil = \lceil 6 \rceil = 6$$

นั่นคือ ในการทำดัชนีบิตแมปแบบคู่กันบนแอทริบิวต์ที่มีคาร์ดินอลิตี้ 15 ใช้ 6 บิตแมปเวกเตอร์ ซึ่งสามารถเขียนความสัมพันธ์ระหว่างบิตแมปเวกเตอร์กับค่าข้อมูลของแอทริบิวต์ได้ดังภาพประกอบ 4-1



ภาพประกอบ 4-1 แผนภาพแสดงความสัมพันธ์ระหว่างบิตแมปเวกเตอร์กับค่าข้อมูลของแอทริบิวต์ที่มีคาร์ดินอลิตี้ 15 ใช้ 6 บิตแมปเวกเตอร์

จากภาพประกอบ 4-1 แอทริบิวต์ที่นำมาสร้างดัชนีมีคาร์ดินอลิตี้ 15 ค่า คือ 0,1,2,...,14 จะเห็นว่า แต่ละค่าของแอทริบิวต์ถูกแทนด้วย 2 บิตแมปเวกเตอร์ เช่น ค่าของ 14 ถูกแทนด้วย D^5 และ D^4, ค่าของ 0 ถูกแทนด้วย D^1 และ D^0 เป็นต้น จากความสัมพันธ์ดังกล่าวนี้ นำไปสู่รูปแบบการลงรหัสข้อมูลของดัชนีบิตแมปแบบคู่กัน พิจารณาได้ดังภาพประกอบ 4-2



ภาพประกอบ 4-2 แผนภาพแสดงการลงรหัสดัชนีบิตแมปแบบคู่กัน เมื่อ $C = 15$ ใช้ 6 บิตแมปเวกเตอร์

จากภาพประกอบ 4-2 แสดงให้เห็นว่า แต่ละค่าข้อมูลถูกแทนค่าด้วย 2 บิตแมปเวกเตอร์เสมอ เช่น ค่าข้อมูล 14 ถูกแทนด้วยบิตแมปเวกเตอร์ D^5 และ D^4 เป็นต้น และแต่ละบิตแมปเวกเตอร์ถูกใช้แทนค่าข้อมูล 5 ค่า เช่น บิตแมปเวกเตอร์ D^5 ถูกใช้แทนค่าข้อมูล 10, 11, 12, 13 และ 14 เป็นต้น กำหนดให้จุดสีดำหมายถึงการลงรหัสข้อมูลด้วย 1 และไม่มีจุดใด ๆ หมายถึงการลงรหัสข้อมูลด้วย 0 ซึ่งการลงรหัสข้อมูลสำหรับแต่ละค่าของแอททริบิวต์สามารถพิจารณาได้ดังภาพประกอบ 4-3

ค่าข้อมูล	ลงรหัสข้อมูล					
	D^5	D^4	D^3	D^2	D^1	D^0
0	0	0	0	0	1	1
1	0	0	0	1	0	1
2	0	0	0	1	1	0
3	0	0	1	0	0	1
4	0	0	1	0	1	0
5	0	0	1	1	0	0
6	0	1	0	0	0	1
7	0	1	0	0	1	0
8	0	1	0	1	0	0
9	0	1	1	0	0	0
10	1	0	0	0	0	1
11	1	0	0	0	1	0
12	1	0	0	1	0	0
13	1	0	1	0	0	0
14	1	1	0	0	0	0

ภาพประกอบ 4-3 แสดงการลงรหัสข้อมูลของแอทริบิวต์ในการสร้างดัชนีบิตแมปแบบคู่กัน เมื่อ $C = 15$ ใช้ 6 บิตแมปเวกเตอร์

ต่อไปเราจะหาหลักการทั่วไปในการสร้างดัชนีบิตแมปแบบคู่กัน กำหนดรูปแบบการลงรหัสดังนี้

$$D^{n-1} \ldots D^2 D^1 D^0$$

เมื่อ n เป็นจำนวนบิตแมปเวกเตอร์ที่ใช้

ตัวอย่างเช่น การลงรหัสข้อมูลของแอทริบิวต์ที่มีคาร์ดินอลิตี้เท่ากับ 15 ใช้ 6 บิตแมปเวกเตอร์ มีรูปแบบการลงรหัสดัชนีเป็น $D^5D^4D^3D^2D^1D^0$ จากภาพประกอบ 4-2 จะเห็นว่าค่าข้อมูล 14 ถูกแทนด้วยบิตแมปเวกเตอร์ D^5 และ D^4 จึงได้ว่าค่าข้อมูล 14 มีการลงรหัสเป็น 110000 ดังภาพประกอบ 4-3

จากทฤษฎีบท 4.1 เราหาจำนวนบิตแมปเวกเตอร์ (ค่าของ n) ที่ใช้ในการทำดัชนีบิตแมปแบบคู่กัน โดยใช้หลักการที่ว่า แต่ละค่าของแอทริบิวต์ถูกแทนด้วย 2 บิตแมปเวกเตอร์ และทำให้ทุกบิตแมปเวกเตอร์ถูกใช้อย่างคุ้มค่ามากที่สุด ขั้นตอนต่อไป เราต้องการทราบว่า แต่และค่าของแอทริบิวต์จะมีการลงรหัสอย่างไร และจะให้ 2 บิตแมปเวกเตอร์ใด จากทั้งหมด n บิตแมปเวกเตอร์ เป็นตัวแทนแต่ละค่าของแอทริบิวต์ หรือกล่าวว่า จะให้ 2 บิตแมปเวกเตอร์ใดถูกลงรหัส ‘1’ นั่นเอง

กำหนดให้ D^r หมายถึง บิตแมปเวกเตอร์แรก (ที่ถูกลงรหัสเป็น '1')

D^s หมายถึง บิตแมปเวกเตอร์ที่สอง (ที่ถูกลงรหัสเป็น '1')

v หมายถึง ค่าของแอทริบิวต์ที่เลือกมาทำดัชนี ($v \in \{0,1,2,...,C-1\}$)

C หมายถึง คาร์ดินอลิตี้ของแอทริบิวต์ที่เลือกมาทำดัชนี

เงื่อนไขสำหรับบิตแมปเวกเตอร์ D^r และ D^s คือ $0 \le s < r$

ทฤษฎีบทและหลักการที่จะกล่าวต่อไปนี้ จะนำไปสู่วิธีการลงรหัสดัชนีบิตแมปแบบคู่กัน

ทฤษฎีบท 4.2 ถ้าให้ n คือ จำนวนบิตแมปเวกเตอร์ที่ใช้ในการทำดัชนีบิตแมปแบบคู่กัน, $S_r = \left\{\left(D^r, D^s\right) \mid 0 \le s < r\right\}$ และ $S = S_1 \cup S_2 \cup ... \cup S_{n-1}$ แล้ว $|S| = \binom{n}{2}$

<u>พิสูจน์</u> สมมติให้ n คือ จำนวนบิตแมปเวกเตอร์ที่ใช้ในการทำดัชนีบิตแมปแบบคู่กัน, $S_r = \left\{\left(D^r, D^s\right) \mid 0 \le s < r\right\}$ และ $S = S_1 \cup S_2 \cup ... \cup S_{n-1}$

จากที่กำหนดให้ $S_r = \left\{\left(D^r, D^s\right) \mid 0 \le s < r\right\}$ จะได้ว่า $|S_r| = r$

พิจารณา $S_1 = \left\{\left(D^1, D^s\right) \mid 0 \le s < 1\right\}$ จะได้ $|S_1| = 1$

$S_2 = \left\{\left(D^2, D^s\right) \mid 0 \le s < 2\right\}$ จะได้ $|S_2| = 2$

$S_3 = \left\{\left(D^3, D^s\right) \mid 0 \le s < 3\right\}$ จะได้ $|S_3| = 3$

$\vdots \qquad \vdots$

$S_{n-1} = \left\{\left(D^{n-1}, D^s\right) \mid 0 \le s < n-1\right\}$ จะได้ $|S_{n-1}| = n-1$

จะเห็นได้ว่า สมาชิกใน $S_1, S_2, ..., S_{n-1}$ ไม่ซ้ำกันเลย

ดังนั้น
$$\begin{aligned} |S| &= |S_1| + |S_2| + ... + |S_{n-1}| \\ &= 1 + 2 + ... + (n-1) \\ &= \frac{n(n-1)}{2} \\ &= \binom{n}{2} \end{aligned}$$
□

ทฤษฎีบท 4.2 เป็นการแสดงให้เห็นว่า เซตของคู่ลำดับ (D^r, D^s) มีจำนวนสมาชิก $\binom{n}{2}$ ตัว นั่นหมายความว่า การลงรหัสข้อมูลด้วยการกำหนดให้ D^r เป็นบิตแมป

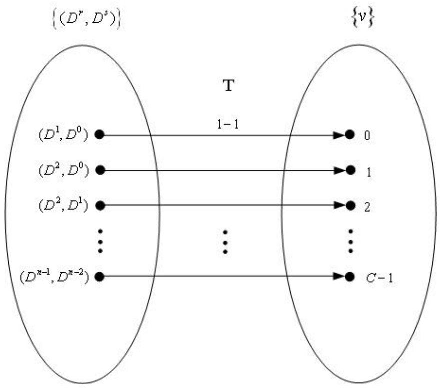
เวกเตอร์แรก และ D^s เป็นบิตแมปเวกเตอร์ที่สอง โดยที่ $0 \le s < r,\ 1 \le r \le n-1$ เป็นวิธีที่สามารถใช้ลงรหัสดัชนีบิตแมปแบบคู่กันบนแอทริบิวต์ที่มีคาร์ดินอลลิตี้ C ซึ่งใช้ n บิตแมปเวกเตอร์ได้ครบถ้วน $\left(\because\ C \le \binom{n}{2}\right)$ ตัวอย่างเช่น จากภาพประกอบ 4-3 เป็นการลงรหัสข้อมูลของแอทริบิวต์ในการสร้างดัชนีบิตแมปแบบคู่กัน โดยที่แอทริบิวต์ที่เลือกมาทำดัชนีมีคาร์ดินอลลิตี้เท่ากับ 15 ใช้ 6 บิตแมปเวกเตอร์ สามารถพิจารณาบิตแมปเวกเตอร์ D^r (บิตแมปเวกเตอร์แรกที่ถูกลงรหัส ‘1’ , $1 \le r \le 5$) และ D^s (บิตแมปเวกเตอร์ที่สองที่ถูกลงรหัส ‘1’ , $0 \le s < r$) สำหรับแต่ละค่าของแอทริบิวต์ได้ดังภาพประกอบ 4-4

กลุ่มที่	บิตแมปเวกเตอร์แรก	บิตแมปเวกเตอร์ที่สอง	ค่าข้อมูล
①	D^1	D^0	0
②	D^2 D^2	D^0 D^1	1 2
③	D^3 D^3 D^3	D^0 D^1 D^2	3 4 5
④	D^4 D^4 D^4 D^4	D^0 D^1 D^2 D^3	6 7 8 9
⑤	D^5 D^5 D^5 D^5 D^5	D^0 D^1 D^2 D^3 D^4	10 11 12 13 14

ภาพประกอบ 4-4 แสดงบิตแมปเวกเตอร์แรกและบิตแมปเวกเตอร์ที่สองที่ถูกลงรหัส ‘1’ บนดัชนีบิตแมปแบบคู่กัน เมื่อ $C = 15$

ต่อไปเราต้องการทราบว่าสำหรับแต่ละค่าของแอทริบิวต์ จะมีบิตแมปเวกเตอร์ใดเป็นบิตแมปเวกเตอร์แรก (D^r) และบิตแมปเวกเตอร์ที่สอง (D^s) กำหนดความสัมพันธ์ดังต่อไปนี้

$$\mathrm{T}\ :\ \{(D^r, D^s)\} \xrightarrow{1-1} \{v\}$$
$$\mathrm{T}(D^r, D^s) \ = \ v$$



โดยที่ $0 \le v < C$ และ $0 < s < r \le n-1$,
$\mathrm{T}(D^{r-1}, D^s) \ < \ \mathrm{T}(D^r, D^s)$,
$\mathrm{T}(D^r, D^{s-1}) \ < \ \mathrm{T}(D^r, D^s)$ และ
$\mathrm{T}(D^1, D^0) \ = \ 0$ ....(4.2)

ความสัมพันธ์ T เป็นฟังก์ชันหนึ่งต่อหนึ่งที่มีการส่งจากโดเมนที่เป็นเซตของคู่ลำดับของบิตแมปเวกเตอร์แรกและบิตแมปเวกเตอร์ที่สอง (D^r, D^s) ไปยังเรนจ์ที่เป็นเซตของค่าข้อมูล v ตัวอย่างเช่น จากภาพประกอบ 4-4 ค่าข้อมูล 14 มีบิตแมปเวกเตอร์แรก คือ D^5 และบิตแมปเวกเตอร์ที่สอง คือ D^4 จะได้ว่า $\mathrm{T}(D^5, D^4) = 14$ เป็นต้น

ความหมายของเงื่อนไขของความสัมพันธ์ T คือ เราแบ่งกลุ่มค่าข้อมูลโดยพิจารณาจากโดเมนที่เป็นบิตแมปเวกเตอร์แรก โดยที่กลุ่มที่ r มี D^r เป็นบิตแมปเวกเตอร์แรก กล่าวคือ การแบ่งกลุ่มความสัมพันธ์เริ่มจากกลุ่มที่ 1 มี D^1 เป็นบิตแมปเวกเตอร์แรก เรียงต่อไปจนถึงกลุ่มที่ $n-1$ มี D^{n-1} เป็นบิตแมปเวกเตอร์แรก ส่วนค่าข้อมูล (v) จะถูกนับเรียงต่อกันไป และถูกกำหนดค่าต่อจากกลุ่มที่อยู่ก่อนหน้า โดยกลุ่มแรกเริ่มจากค่า 0 ซึ่งค่าข้อมูลในแต่ละกลุ่มแปรผันตามบิตแมปเวกเตอร์ที่สอง (D^s) จากภาพประกอบ 4-4 สามารถพิจารณากลุ่มของความสัมพันธ์ได้ดังนี้

- กลุ่มที่ 1 $\mathrm{T}(D^1, D^0) = 0$
- กลุ่มที่ 2 $\mathrm{T}(D^2, D^0) = 1$
  $\mathrm{T}(D^2, D^1) = 2$
- กลุ่มที่ 3 $\mathrm{T}(D^3, D^0) = 3$
  $\mathrm{T}(D^3, D^1) = 4$
  $\mathrm{T}(D^3, D^2) = 5$
- กลุ่มที่ 4 $\mathrm{T}(D^4, D^0) = 6$
  $\mathrm{T}(D^4, D^1) = 7$

$\mathrm{T}(D^4, D^2) = 8$
$\mathrm{T}(D^4, D^3) = 9$

- กลุ่มที่ 5 $\mathrm{T}(D^5, D^0) = 10$
$\mathrm{T}(D^5, D^1) = 11$
$\mathrm{T}(D^5, D^2) = 12$
$\mathrm{T}(D^5, D^3) = 13$
$\mathrm{T}(D^5, D^4) = 14$

ต่อไปเราจะพิจารณาความสัมพันธ์ระหว่างโดเมนและเรนจ์ในแต่ละกลุ่ม ดังนี้

กำหนดให้ S_r เป็นเซตของคู่อันดับบิตแมปเวกเตอร์แรกและบิตแมปเวกเตอร์ที่สอง โดยมีบิตแมปเวกเตอร์แรกเป็น D^r และบิตแมปเวกเตอร์ที่สองเป็น D^s เมื่อ $1 \le r \le n-1$ และ $0 \le s < r$

V_r เป็นเซตของค่าข้อมูล (v) บนความสัมพันธ์ T ซึ่งส่งจากโดเมน S_r เมื่อ $1 \le r \le n-1$

จากเงื่อนไขของ S_r และ V_r ที่กล่าวมา สามารถเขียนแทนด้วยสัญลักษณ์ได้ดังนี้

$$S_r = \left\{ \left(D^r, D^s \right) \middle| \ 0 \le s < r \right\} \quad \text{โดยที่ } 1 \le r \le n-1$$

$$V_r = \left\{ v \ \middle| \ \mathrm{T}(D^r, D^s) = v,\ (D^r, D^s) \in S_r \right\}$$

จะได้ว่า $|S_r| = r$

และ $|V_r| = r$

เนื่องจากค่าข้อมูลในแต่ละกลุ่ม (V_r) มีการนับเรียงต่อ ๆ กันไป โดยกลุ่มแรกเริ่มจากค่าข้อมูล 0 ซึ่งค่าข้อมูลของกลุ่มใด ๆ จะมีการนับต่อจากข้อมูลที่อยู่ในกลุ่มก่อนหน้า และจำนวนสมาชิกของค่าข้อมูลในแต่ละกลุ่ม ($|V_r|$) มีค่าเท่ากับ r จึงได้ว่า

$$V_r = \left\{ v \ \middle| \ v \in \left[|V_1| + |V_2| + \ldots + |V_{r-1}|,\ |V_1| + |V_2| + \ldots + |V_{r-1}| + r \right) \right\}$$

เนื่องจาก ค่าข้อมูล v เป็นจำนวนเต็ม จึงได้ว่า

$$V_r = \left\{ v \ \middle| \ v \in \left[|V_1| + |V_2| + \ldots + |V_{r-1}|,\ |V_1| + |V_2| + \ldots + |V_{r-1}| + r - 1 \right] \right\}$$

$$= \left\{ v \ \middle| \ v \in \left[1 + 2 + \ldots + (r-1),\ 1 + 2 + \ldots + (r-1) + r - 1 \right] \right\}$$

$$= \left\{ v \;\middle|\; v \in \left[\sum_{i=1}^{r-1} i,\ \left(\sum_{i=1}^{r} i \right) - 1 \right] \right\}$$

$$= \left\{ v \;\middle|\; \sum_{i=1}^{r-1} i \le v \le \left(\sum_{i=1}^{r} i \right) - 1 \right\}$$

พิจารณาเงื่อนไข $$\sum_{i=0}^{r-1} i \quad \le \quad v \quad \le \quad \left(\sum_{i=0}^{r} i \right) - 1 \qquad \text{.............(4.3)}$$

พิจารณาอสมการทางซ้าย : $$\sum_{i=0}^{r-1} i \quad \le \quad v$$

$$\frac{r(r-1)}{2} \quad \le \quad v$$

จะได้ $$r(r-1) \quad \le \quad 2v$$

$$r^2 - r \quad \le \quad 2v$$

$$r^2 - 2 \cdot \frac{r}{2} + \left(\frac{1}{2} \right)^2 \quad \le \quad 2v + \left(\frac{1}{2} \right)^2$$

$$r^2 - 2 \cdot \frac{r}{2} + \left(\frac{1}{2} \right)^2 \quad \le \quad 2v + \left(\frac{1}{2} \right)^2$$

$$\left(r - \frac{1}{2} \right)^2 \quad \le \quad 2v + \frac{1}{4}$$

$$-\sqrt{2v + \frac{1}{4}} \quad \le \quad r - \frac{1}{2} \quad \le \quad \sqrt{2v + \frac{1}{4}}$$

$$-\sqrt{2v + \frac{1}{4}} + \frac{1}{2} \quad \le \quad r \quad \le \quad \sqrt{2v + \frac{1}{4}} + \frac{1}{2} \qquad \text{.............(4.4)}$$

พิจารณาอสมการข้างขวา : $$v \quad \le \quad \left(\sum_{i=1}^{r} i \right) - 1$$

$$v \quad \le \quad \frac{r(r+1)}{2} - 1$$

$$2v \quad \le \quad r(r+1) - 2$$

$$2v+2 \leq r^2+r$$

$$2v+2+\left(\frac{1}{2}\right)^2 \leq r^2+2\cdot\frac{r}{2}+\left(\frac{1}{2}\right)^2$$

$$2v+\frac{9}{4} \leq \left(r+\frac{1}{2}\right)^2$$

$$r+\frac{1}{2} \leq -\sqrt{2v+\frac{9}{4}} \quad \text{หรือ} \quad r+\frac{1}{2} \geq \sqrt{2v+\frac{9}{4}}$$

$$r \leq -\sqrt{2v+\frac{9}{4}}-\frac{1}{2} \quad \text{หรือ} \quad r \geq \sqrt{2v+\frac{9}{4}}-\frac{1}{2} \quad \text{.............(4.5)}$$

ต่อไปเราต้องการแสดงให้เห็นว่า $\sqrt{2v+\frac{9}{4}}-\frac{1}{2} \leq \sqrt{2v+\frac{1}{4}}+\frac{1}{2}$ เพื่อยืนยันว่ามีค่าของ r ที่สอดคล้องกับเงื่อนไข (4.3)

เนื่องจาก
$$0 \leq v$$

$$0 \leq 2v$$

$$\frac{1}{4} \leq 2v+\frac{1}{4}$$

$$\frac{1}{2} \leq \sqrt{2v+\frac{1}{4}}$$

$$1 \leq 2\sqrt{2v+\frac{1}{4}}$$

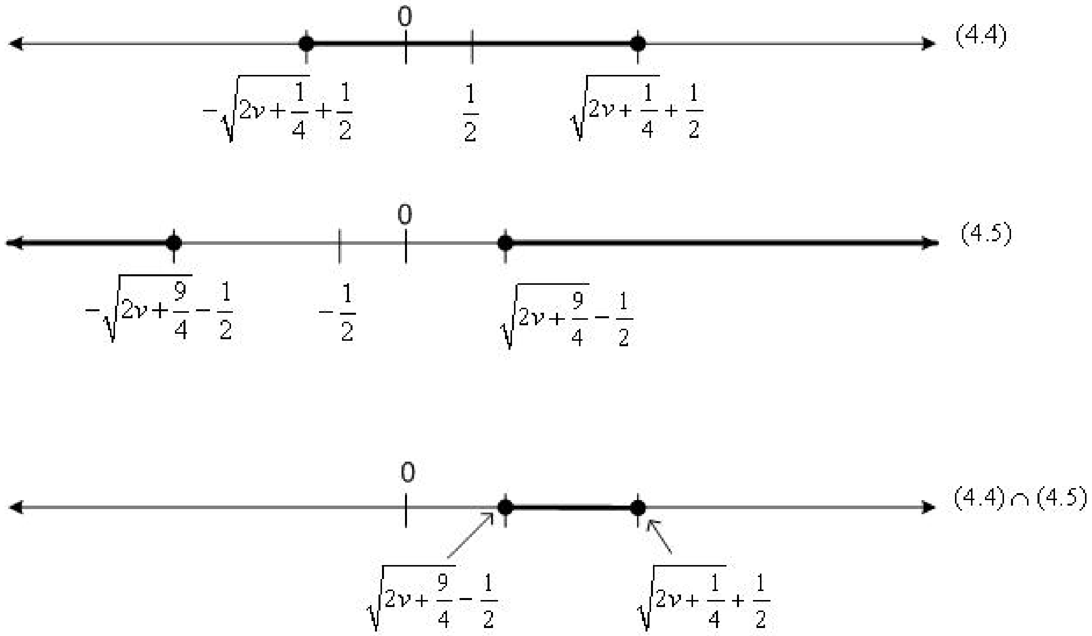
$$\left(\sqrt{2v+\frac{1}{4}}\right)^2+1+1 \leq \left(\sqrt{2v+\frac{1}{4}}\right)^2+2\sqrt{2v+\frac{1}{4}}+1$$

$$2v+\frac{9}{4} \leq \left(\sqrt{2v+\frac{1}{4}}+1\right)^2$$

$$\sqrt{2v+\frac{9}{4}} \leq \sqrt{2v+\frac{1}{4}}+1$$

$$\sqrt{2v+\frac{9}{4}}-\frac{1}{2} \leq \sqrt{2v+\frac{1}{4}}+\frac{1}{2}$$

ดังนั้น จาก (4.4) และ (4.5) เราสามารถเขียนค่าของ r ซึ่งเป็นคำตอบของอสมการได้ภาพประกอบ 4-5



ภาพประกอบ 4-5 แสดงค่าของ r ซึ่งเป็นคำตอบของอสมการ $\sum_{i=0}^{r-1} i \le v \le \left(\sum_{i=0}^{r} i\right) - 1$

จากภาพประกอบ 4-5 จะได้ว่าค่าของ r ที่เป็นคำตอบของอสมการ คือ ค่าที่อยู่ในช่วงปิด $\sqrt{2v+\frac{9}{4}}-\frac{1}{2}$ ถึง $\sqrt{2v+\frac{1}{4}}+\frac{1}{2}$

เนื่องจาก $\sqrt{2v+\frac{1}{4}}+\frac{1}{2}-\left(\sqrt{2v+\frac{9}{4}}-\frac{1}{2}\right) < \sqrt{2v+\frac{1}{4}}+\frac{1}{2}-\left(\sqrt{2v+\frac{1}{4}}-\frac{1}{2}\right) = 1$

ดังนั้น $\sqrt{2v+\frac{9}{4}}-\frac{1}{2}$ และ $\sqrt{2v+\frac{1}{4}}+\frac{1}{2}$ มีค่าต่างกันไม่ถึง 1

นั่นคือ มีจำนวนเต็มเพียงจำนวนเดียวเท่านั้นที่อยู่ในช่วงปิด $\sqrt{2v+\frac{9}{4}}-\frac{1}{2}$ ถึง $\sqrt{2v+\frac{1}{4}}+\frac{1}{2}$

เนื่องจากเราต้องการ r ที่เป็นจำนวนเต็ม

จึงได้ว่ามีค่าของ r ที่เป็นไปได้มีเพียงค่าเดียว คือ $r = \left\lceil \sqrt{2v+\frac{9}{4}}-\frac{1}{2} \right\rceil = \left\lfloor \sqrt{2v+\frac{1}{4}}+\frac{1}{2} \right\rfloor$

ซึ่งในที่นี้เราเลือกใช้รูปแบบ $r = \left\lceil \sqrt{2v+\frac{9}{4}} - \frac{1}{2} \right\rceil$ □

ดังนั้น V_r ซึ่งเป็นเซตของค่าข้อมูลที่ถูกส่งมาจากโดเมนที่มี D^r เป็นบิตแมปเวกเตอร์แรก สามารถเขียนแทนด้วย

$$V_r = \left\{ v \;\middle|\; r = \left\lceil \sqrt{2v+\frac{9}{4}} - \frac{1}{2} \right\rceil \right\}$$

นั่นหมายความว่า สำหรับการลงรหัสให้กับค่าข้อมูล (v) ใด ๆ แล้ว D^r เป็นบิตแมปเวกเตอร์แรก ก็ต่อเมื่อ $r = \left\lceil \sqrt{2v+\frac{9}{4}} - \frac{1}{2} \right\rceil$

ต่อไปเราจะพิจารณาว่าบิตแมปเวกเตอร์ที่สอง$\left(D^s\right)$ คือ บิตแมปเวกเตอร์ใด จาก (4.2) เงื่อนไขบนความสัมพันธ์ T กำหนดไว้ว่า $0 \le s < r$
ดังนั้น ค่าของ s ที่เป็นไปได้ คือ $0,1,2,...,r-1$

เนื่องจาก

$$V_r = \left\{ v \;\middle|\; v \in \left[\sum_{i=1}^{r-1} i,\ \left(\sum_{i=1}^{r-1} i\right) + r - 1 \right] \right\}$$

$$= \left\{ \left(\sum_{i=1}^{r-1} i\right)+0, \left(\sum_{i=1}^{r-1} i\right)+1, \left(\sum_{i=1}^{r-1} i\right)+2, ..., \left(\sum_{i=1}^{r-1} i\right)+r-1 \right\}$$

$$= \left\{ \left(\sum_{i=1}^{r-1} i\right)+s \;\middle|\; s \in \{0,1,2,...,r-1\} \right\}$$

จาก (4.2) เงื่อนไขบนความสัมพันธ์ T กำหนดไว้ว่า $\mathrm{T}(D^r, D^{s-1}) < \mathrm{T}(D^r, D^s)$ หมายความว่า ค่าข้อมูล v ในกลุ่มที่มี D^r เป็นบิตแมปเวกเตอร์แรก แปรผันตาม D^s จึงได้ว่า

$$v = \left(\sum_{i=0}^{r-1} i\right) + s$$

$$v = \frac{r(r-1)}{2} + s$$

$$\mathrm{s} = v - \frac{r(r-1)}{2}$$

นั่นหมายความว่า สำหรับการลงรหัสให้กับค่าข้อมูล (v) ใด ๆ ที่มี D^r เป็นบิตแมปเวกเตอร์แรก แล้ว D^s เป็นบิตแมปเวกเตอร์ที่สอง ก็ต่อเมื่อ $s = v - \frac{r(r-1)}{2}$ □

จากที่กล่าวมาทั้งหมดทำให้เราได้ทฤษฎีบทในการลงรหัสข้อมูลดังต่อไปนี้

ทฤษฎีบท 4.3 สำหรับค่าข้อมูล v ใด ๆ แล้ว D^r เป็นบิตแมปเวกเตอร์แรก และ D^s เป็นบิตแมปเวกเตอร์ที่สอง ก็ต่อเมื่อ $r = \left\lceil \sqrt{2v + \frac{9}{4}} - \frac{1}{2} \right\rceil$ และ $s = v - \frac{r(r-1)}{2}$

4.3 ขั้นตอนวิธีการสร้างดัชนีบิตแมปแบบคู่กัน

จากที่กล่าวมาข้างต้น สรุปเป็นขั้นตอนวิธีการทำดัชนีบิตแมปแบบคู่กัน ได้ดังนี้

สัญลักษณ์ที่ใช้

C แทน คาร์ดินอลิตี้(จำนวนค่าที่เป็นไปได้)ของแอทริบิวต์ที่นำมาสร้างดัชนีบิตแมปแบบคู่กัน

n แทน จำนวนบิตแมปเวกเตอร์ที่ใช้ในการทำดัชนีบิตแมปแบบคู่กัน

D^j แทน บิตแมปเวกเตอร์ที่ j ของดัชนีบิตแมปแบบคู่กัน โดยที่ $j = 0, 1, 2, ..., n-1$

v แทน ค่าของแอทริบิวต์ที่เลือกมาทำดัชนีบิตแมปแบบคู่กัน $\left(v = \{0, 1, 2, ..., C-1\}\right)$

ขั้นตอนวิธีการสร้างดัชนีบิตแมปแบบคู่กัน

1. หาจำนวนบิตแมปเวกเตอร์ที่ใช้ในการสร้างดัชนี (ค่าของ n)

$$n = \left\lceil \sqrt{2C + 0.25} + 0.5 \right\rceil$$

2. ลงรหัสค่าของแอทริบิวต์ (v) เรคอร์ดที่ i บนแต่ละบิตแมปเวกเตอร์ โดยให้บิตที่ i ของบิตแมปเวกเตอร์ D^j เมื่อ $0 \le j \le n-1$ มีรูปแบบการลงรหัสดังนี้

$$D^j = \begin{cases} 1 & \text{ถ้า } j = r \text{ หรือ } j = s \\ 0 & \text{กรณีอื่น ๆ} \end{cases}$$

โดยที่ $r = \left\lceil \sqrt{2v + 2.25} - 0.5 \right\rceil$ และ $s = v - \frac{r(r-1)}{2}$ ....(4.6)

ตัวอย่างเช่น ต้องการสร้างดัชนีบิตแมปแบบคู่กันบนแอทริบิวต์ที่มีคาร์ดินอลิตี้เท่ากับ 15 มีขั้นตอนดังต่อไปนี้

1. หาจำนวนบิตแมปเวกเตอร์ที่ใช้ในการสร้างดัชนี (ค่าของ n)

$$n = \left\lceil \sqrt{(2\times 15)+0.25}+0.5 \right\rceil = 6$$

2. ลงรหัสค่าของแอทริบิวต์ (v) เรคอร์ดที่ i บนแต่ละบิตแมปเวกเตอร์ โดยให้บิตที่ i ของบิตแมปเวกเตอร์ D^j เมื่อ $0 \le j \le 5$ มีการลงรหัสตามรูปแบบ (4.6) พิจารณาตัวอย่างดัชนีบิตแมปแบบคู่กันบนแอทริบิวต์ A ซึ่งมีคาร์ดินอลิตี้เท่ากับ 15 ดังภาพประกอบ 4-6

	$\pi_A(R)$
1	3
2	11
3	1
4	2
5	7
6	10
7	14
8	6
9	0
10	5
11	4
12	2

(ก) แอทริบิวต์ A

D^5	D^4	D^3	D^2	D^1	D^0
0	0	1	0	0	1
1	0	0	0	1	0
0	0	0	1	0	1
0	0	0	1	1	0
0	1	0	0	1	0
0	1	0	0	0	1
1	1	0	0	0	0
0	1	0	0	0	1
0	0	0	0	1	1
0	0	1	1	0	0
0	0	1	0	1	0
0	0	0	1	1	0

(ข) ดัชนีบิตแมปแบบคู่กันบนแอทริบิวต์ A

ภาพประกอบ 4-6 แสดงดัชนีบิตแมปแบบคู่กันบนแอทริบิวต์ A เมื่อ $C=15$ ใช้ 6 บิตแมปเวกเตอร์

4.4 ขั้นตอนวิธีการสอบถามแบบค่าเท่ากันบนดัชนีบิตแมปแบบคู่กัน

การสอบถามแบบค่าเท่ากันมีรูปแบบ คือ "$A=v$" หมายถึง การสอบถามว่าบนแอทริบิวต์ A มีเรคอร์ดใดบ้างที่มีค่าเท่ากับ v สำหรับการสอบถามแบบค่าเท่ากันบนดัชนีบิตแมปแบบคู่กัน ประกอบด้วย การอ่าน 2 บิตแมปเวกเตอร์ (เขียนแทนด้วย D^r และ D^s) และการดำเนินการตรรกะ AND ระหว่างบิตแมปเวกเตอร์ 1 ครั้ง ดังนั้นการสอบถามแบบค่าเท่ากันบนดัชนีบิตแมปแบบคู่กันมีรูปแบบทั่วไปคือ

$$"A=v" = D^r \wedge D^s$$

โดยที่ $r = \left\lceil \sqrt{2v+2.25}-0.5 \right\rceil$ และ $\mathrm{s} = v-\frac{r(r-1)}{2}$

ตัวอย่างเช่น จากภาพประกอบ 4-6 ต้องการทราบว่าบนแอทริบิวต์ A มีเรคอร์ดใดบ้างที่มีค่าเท่ากับ 3 สามารถทำได้ตามขั้นตอนดังนี้

1. ในที่นี้ v มีค่าเท่ากับ 3
2. คำนวณหาค่าของ r ซึ่งเป็นตัวกำหนดบิตแมปเวกเตอร์แรกที่จะต้องอ่าน
   จากสูตร $r = \left\lceil \sqrt{2v+2.25}-0.5 \right\rceil$
   แทนค่า $v=3$ จะได้ $r = \left\lceil \sqrt{(2\times 3)+2.25}-0.5 \right\rceil = 3$
3. คำนวณหาค่าของ s ซึ่งเป็นตัวกำหนดบิตแมปเวกเตอร์ที่สองที่จะต้องอ่าน
   จากสูตร $\mathrm{s} = v-\frac{r(r-1)}{2}$
   แทนค่า $v=3$ และ $r=3$ จะได้ $\mathrm{s} = 3-\frac{3(3-1)}{2} = 0$
4. อ่านบิตแมปเวกเตอร์ D^3 และ D^0
5. ดำเนินการตรรกะ AND บิตต่อบิตระหว่างบิตแมปเวกเตอร์ D^3 และ D^0
6. ผลลัพธ์ที่ได้ พบว่า บิตที่ 1 ของบิตแมปเวกเตอร์ที่ได้จากข้อ 5 มีค่าเท่ากับ 1 ส่วนบิตอื่น ๆ มีค่าเท่ากับ 0 จึงไว้ว่า เรคอร์ดที่มีค่าของแอทริบิวต์เป็น 3 คือเรคอร์ดที่ 1 ดังภาพประกอบ 4-7

	$\pi_A(R)$	D^5	D^4	D^3	D^2	D^1	D^0	D^3	D^0	$D^3 \wedge D^0$
1	3	0	0	1	0	0	1	1	1	1
2	11	1	0	0	0	1	0	0	0	0
3	1	0	0	0	1	0	1	0	1	0
4	2	0	0	0	1	1	0	0	0	0
5	7	0	1	0	0	1	0	0	0	0
6	10	0	1	0	0	0	1	0	1	0
7	14	1	1	0	0	0	0	0	0	0
8	6	0	1	0	0	0	1	0	1	0
9	0	0	0	0	0	1	1	0	1	0
10	5	0	0	1	1	0	0	1	0	0
11	4	0	0	1	0	1	0	1	0	0
12	2	0	0	0	1	1	0	0	0	0

(ก) แอทริบิวต์ A (ข) ดัชนีบิตแมปแบบคู่กันบนแอทริบิวต์ A (ค) การสอบถาม "$A = 3$"

ภาพประกอบ 4-7 แสดงการสอบถามแบบค่าเท่ากันบนดัชนีบิตแมปแบบคู่กัน กรณีการสอบถาม "$A = 3$"

4.5 ข้อดีของดัชนีบิตแมปแบบคู่กัน

ดัชนีบิตแมปแบบคู่กันใช้พื้นที่ในการจัดเก็บดัชนีน้อยกว่าดัชนีบิตแมปแบบพื้นฐาน แบบช่วง และแบบกระจาย เหมาะสำหรับการสอบถามแบบค่าเท่ากัน เพราะมีการอ่าน 2 บิตแมปเวกเตอร์ และดำเนินการตรรกะ AND ระหว่างบิตแมปเวกเตอร์เพียงครั้งเดียวเท่านั้น จึงทำให้ดัชนีบิตแมปแบบค่าเท่ากันใช้เวลาในการสอบถามน้อยกว่าดัชนีบิตแมปแบบเข้ารหัสซึ่งจะต้องมีการอ่านทุกบิตแมปเวกเตอร์และดำเนินการตรรกะระหว่างบิตแมปเวกเตอร์หลายครั้ง

4.6 ข้อจำกัดของดัชนีบิตแมปแบบคู่กัน

ดัชนีบิตแมปแบบคู่กันใช้พื้นที่ในการจัดเก็บดัชนีมากกว่าดัชนีบิตแมปแบบเข้ารหัส ดัชนีบิตแมปแบบคู่กันจึงมีความสามารถในการสร้างดัชนีบิตแมปบนแอทริบิวต์ที่มีคาร์ดินอลิตี้สูง ๆ และมีจำนวนเรคอร์ดมาก ๆ ได้น้อยกว่าดัชนีบิตแมปแบบเข้ารหัส