

## สารบัญ

	หน้า
สารบัญ .....	(6)
รายการตาราง.....	(10)
รายการภาพประกอบ .....	(11)
บทที่	
1. บทนำ.....	1
1.1 ความเป็นมา .....	1
1.2 การตรวจเอกสารและงานวิจัยที่เกี่ยวข้อง .....	2
1.2.1 การตรวจเอกสาร .....	2
1.2.2 งานวิจัยที่เกี่ยวข้อง.....	2
1.3 วัตถุประสงค์.....	3
1.4 วิธีการดำเนินการวิจัย .....	3
1.5 ขอบเขตของงานวิจัย.....	4
1.6 ขั้นตอนการดำเนินงาน .....	4
1.7 ระยะเวลาการดำเนินงาน .....	4
1.8 เครื่องมือและอุปกรณ์ที่ใช้ .....	5
1.9 ประโยชน์ที่คาดว่าจะได้รับ.....	6
2. คลังข้อมูลและการค้นหาข้อมูล .....	7
2.1 คลังข้อมูล .....	7
2.1.1 การกำหนดทิศทางหรือมุ่งเน้นไปที่หัวข้อ .....	7
2.1.2 การรวมเข้าด้วยกันเป็นหนึ่งเดียว .....	7
2.1.3 มีเวลาเข้ามาเกี่ยวข้อง .....	8
2.1.4 ไม่เปลี่ยนแปลงได้ง่ายหรือข้อมูลมีความเสถียร.....	8
2.2 ฐานข้อมูลดำเนินการและคลังข้อมูล.....	9
2.2.1 ฐานข้อมูลดำเนินการ.....	9
2.2.2 คลังข้อมูล .....	9
2.3 การสร้างคลังข้อมูลเชิงแนวคิด .....	11
2.3.1 แผนผังแบบดาว .....	11
2.3.2 แผนผังแบบผลึกหิมะ .....	12
2.3.3 แผนผังแบบกลุ่มกาแลคซีของ Fact .....	13

## สารบัญ (ต่อ)

	หน้า
2.4 การค้นหาข้อมูล.....	14
2.4.1 แอส ..... 14	14
2.4.2 ดัชนี..... 18	18
3. ดัชนีแบบบิตแมป..... 21	21
3.1 ดัชนีบิตแมปแบบพื้นฐาน ..... 21	21
3.1.1 การค้นหาข้อมูลแบบค่าเท่ากัน ..... 23	23
3.1.2 การค้นหาข้อมูลแบบความเป็นสมาชิก ..... 23	23
3.1.3 การดำเนินการตรรกะระดับบิตระหว่างบิตแมปเวกเตอร์ของต่างแตริบิวต์ .... 25	25
3.1.4 ข้อดีของดัชนีบิตแมปแบบพื้นฐาน ..... 27	27
3.1.5 ข้อจำกัดของดัชนีบิตแมปแบบพื้นฐาน ..... 27	27
3.2 ดัชนีบิตแมปแบบ Range ..... 27	27
3.2.1 การค้นหาข้อมูลแบบค่าเท่ากัน ..... 29	29
3.2.2 การค้นหาข้อมูลแบบความเป็นสมาชิก ..... 29	29
3.2.3 การดำเนินการตรรกะระดับบิตระหว่างบิตแมปเวกเตอร์ของต่างแตริบิวต์ .... 33	33
3.2.4 ข้อดีของดัชนีบิตแมปแบบ Range ..... 36	36
3.2.5 ข้อจำกัดของดัชนีบิตแมปแบบ Range ..... 36	36
3.3 ดัชนีบิตแมปแบบช่วง ..... 36	36
3.3.1 การค้นหาข้อมูลแบบค่าเท่ากัน ..... 37	37
3.3.2 การค้นหาข้อมูลแบบความเป็นสมาชิก ..... 38	38
3.3.3 การดำเนินการตรรกะระดับบิตระหว่างบิตแมปเวกเตอร์ของต่างแตริบิวต์ .... 40	40
3.3.4 ข้อดีของดัชนีบิตแมปแบบช่วง ..... 42	42
3.3.5 ข้อจำกัดของดัชนีบิตแมปแบบช่วง ..... 42	42
3.4 ดัชนีบิตแมปแบบเข้ารหัส ..... 42	42
3.4.1 การค้นหาข้อมูลแบบค่าเท่ากัน ..... 43	43
3.4.2 การค้นหาข้อมูลแบบความเป็นสมาชิก ..... 43	43
3.4.3 การดำเนินการตรรกะระดับบิตระหว่างบิตแมปเวกเตอร์ของต่างแตริบิวต์ .... 44	44
3.4.4 ข้อดีของดัชนีบิตแมปแบบเข้ารหัส ..... 44	44
3.4.5 ข้อจำกัดของดัชนีบิตแมปแบบเข้ารหัส ..... 44	44
4. ดัชนีบิตแมปแบบกระจาย ..... 46	46
4.1 ขั้นตอนวิธีการสร้าง (Algorithm) ..... 47	47

## สารบัญ (ต่อ)

	หน้า
4.2 การหาค่าที่เหมาะสมที่สุดของจำนวนสมาชิกภายในกลุ่ม Z (ค่า m).....	60
4.3 ขั้นตอนวิธีการค้นหาข้อมูลแบบค่าเท่ากัน .....	64
4.4 ขั้นตอนวิธีการค้นหาข้อมูลแบบความเป็นสมาชิก .....	65
4.5 การดำเนินการตรรกะระดับบิตระหว่างบิตแมปเวกเตอร์ของต่างแอทธิบิวต์ .....	67
4.6 ข้อดีของดัชนีบิตแมปแบบกระจาย .....	70
4.7 ข้อจำกัดของดัชนีบิตแมปแบบกระจาย .....	70
5. การวิเคราะห์และผลการทดลอง.....	71
5.1 ค่าใช้จ่ายจากการวิเคราะห์ .....	71
5.1.1 พื้นที่ที่ใช้ในการจัดเก็บดัชนี .....	71
5.1.2 เวลาที่ใช้ในการค้นหาข้อมูล .....	73
5.2 ค่าใช้จ่ายจากผลการทดลอง .....	74
5.2.1 พื้นที่ที่ใช้ในการจัดเก็บดัชนี .....	74
5.2.2 เวลาที่ใช้ในการค้นหาข้อมูล .....	75
5.2.3 ความสัมพันธ์ระหว่างพื้นที่ที่ใช้ในการจัดเก็บดัชนีและเวลาที่ใช้ใน การค้นหาข้อมูล .....	81
6. บทสรุปและข้อเสนอแนะ.....	92
6.1 บทสรุป.....	92
6.2 อุปสรรคและปัญหา.....	94
6.2.1 ปัญหาการจัดเตรียมข้อมูล .....	95
6.2.2 ปัญหาการตรวจสอบความถูกต้องของการค้นหาข้อมูล .....	95
6.3 ข้อเสนอแนะและงานในอนาคต .....	95
บรรณานุกรม.....	97
ภาคผนวก.....	101
ก .....	102
ก.1 โครงสร้างของการวัดเปรียบเทียบสมรรถนะของ TPC-H .....	102
ก.2 ตัวอย่างการสอบถามของการวัดเปรียบเทียบสมรรถนะของ TPC-H.....	105
ก.3 ตัวอย่างการกรองข้อมูลเฉพาะแอทธิบิวต์ที่จะนำมาสร้างดัชนี .....	106
ก.4 ตัวอย่างการสอบถามข้อมูลแบบซับซ้อน (Complex Queries) และ .....	107
การแก้ปัญหาโดยใช้ดัชนีแบบบิตแมป	

## สารบัญ (ต่อ)

	หน้า
ข .....	111
ข.1 แผนผังโปรแกรมการค้นหาข้อมูลของดัชนีบิตแมปแบบช่วง .....	111
ข.2 แผนผังโปรแกรมน้อยการเลือกไฟล์ (bf_rw2m) .....	117
ข.3 แผนผังโปรแกรมน้อยการเปิดอ่านไฟล์ (getFile).....	117
ข.4 แผนผังโปรแกรมน้อยการแสดงผลข้อมูล (dispData) .....	118
ข.5 แผนผังโปรแกรมน้อยการรวมบิตแมปเวกเตอร์ย่อยเข้าด้วยกัน(merge1Data)	119
ข.6 แผนผังโปรแกรมน้อยการรวมบิตแมปเวกเตอร์ย่อยเข้าด้วยกัน .....	120
เพื่อค้นหาค่าแบบความเป็นสมาชิก (merge2Data)	
ข.7 แผนผังโปรแกรมน้อยการย้ายไปยังโหนดแรก (movFirst).....	121
ข.8 แผนผังโปรแกรมน้อยการสร้างโหนดใหม่ (newNode).....	121
ข.9 แผนผังโปรแกรมน้อยการย้ายไปยังโหนดถัดไป (nextNode).....	122
ข.10 แผนผังโปรแกรมน้อยการเคลียร์ลิสต์ (clearList) .....	122
ประวัติผู้เขียน .....	123

## รายการตาราง

ตาราง		หน้า
1-1	ตารางแสดงระยะเวลาดำเนินงาน	5
2-1	ตารางเปรียบเทียบฐานข้อมูลดำเนินการและคลังข้อมูล	9
3-1	ตารางสรุปข้อดีข้อจำกัดของดัชนีบิตแมปแต่ละชนิด	44
4-1	ตารางสรุปคุณสมบัติที่สำคัญของดัชนีบิตแมปแบบพื้นฐาน แบบ Range แบบช่วง และแบบเข้ารหัส	46
4-2	ตารางสรุปคุณสมบัติที่สำคัญของดัชนีบิตแมปแบบพื้นฐาน แบบ Range แบบช่วง แบบเข้ารหัส และแบบกระจาย	70
5-1	ตารางการวิเคราะห์ค่าใช้จ่ายในเรื่องพื้นที่ที่ใช้ในการจัดเก็บดัชนีบิตแมปทั้ง 5 แบบ	71
5-2	ตารางการวิเคราะห์เวลาที่ใช้ในการค้นหาข้อมูลแบบค่าเท่ากันของดัชนีบิตแมปทั้ง 5 แบบ	74
6-1	ตารางสรุปการใช้พื้นที่ในการจัดเก็บดัชนี	93
6-2	ตารางสรุปเวลาที่ใช้ในการค้นหาข้อมูลแบบค่าเท่ากันและแบบความเป็นสมาชิก	93

## รายการภาพประกอบ

ภาพประกอบ	หน้า	
2-1	สถาปัตยกรรมคลังข้อมูล	8
2-2	ตัวอย่างแผนผังแบบดาว	12
2-3	ตัวอย่างแผนผังแบบผลึกหิมะ	13
2-4	ตัวอย่างแผนผังแบบกลุ่มกาแลคซีของ Fact	14
2-5	ตัวอย่างการแฮชแบบเปิด โดยตารางแฮชมี 10 สล็อต ใช้ฟังก์ชันแฮช $h(K) = K \bmod 10$ จัดเก็บเลข 8 ค่า	15
2-6	ตัวอย่างการแฮชแบบปิด โดยตารางแฮชมี 5 บั๊กเก็ต ใช้ฟังก์ชันแฮช $h(K) = K \bmod 5$ จัดเก็บเลข 8 ค่า	16
2-7	ตัวอย่างการแฮชแบบปิด โดยตารางแฮชมี 11 สล็อต ใช้ฟังก์ชันแฮช $(h(K)+i) \bmod M$ จัดเก็บเลข 8 ค่า ( $M = 11$ )	17
2-8	โครงสร้างดัชนีแบบต้นไม้	19
2-9	ตัวอย่างดัชนี B+ Tree ที่โหนดภายในมีลูกระหว่าง 2 ถึง 4 และลีฟโหนดจัดเก็บข้อมูล 3 ถึง 5 เรคอร์ด	20
3-1	ตัวอย่างแผนภาพการทำดัชนีบิตแมปแบบพื้นฐานบนแอทริบิวต์ A (มีค่า $C = 15$ ) ใช้ 15 บิตแมปเวกเตอร์	22
3-2	ดัชนีบิตแมปแบบพื้นฐานบนแอทริบิวต์ A ใช้ 15 บิตแมปเวกเตอร์	23
3-3	การดำเนินการตรรกะระดับบิตของแอทริบิวต์ A ของดัชนีบิตแมปแบบพื้นฐาน	24
3-4	ดัชนีบิตแมปแบบพื้นฐานบนแอทริบิวต์ brand ใช้ 20 บิตแมปเวกเตอร์	25
3-5	การดำเนินการตรรกะระดับบิตของแอทริบิวต์ type และ brand ของดัชนีบิตแมปแบบพื้นฐาน	26
3-6	ตัวอย่างแผนภาพการทำดัชนีบิตแมปแบบ Range บนแอทริบิวต์ A (มีค่า $C = 15$ ) ใช้ 14 บิตแมปเวกเตอร์	28
3-7	ดัชนีบิตแมปแบบ Range บนแอทริบิวต์ A ใช้ 14 บิตแมปเวกเตอร์	28
3-8	ดัชนีบิตแมปแบบ Range บนแอทริบิวต์ brand ใช้ 19 บิตแมปเวกเตอร์	33
3-9	การดำเนินการตรรกะระดับบิตของแอทริบิวต์ type และ brand ของดัชนีบิตแมปแบบ Range	35
3-10	ตัวอย่างแผนภาพการทำดัชนีบิตแมปแบบช่วง บนแอทริบิวต์ A (มีค่า $C = 15$ ) ใช้ 8 บิตแมปเวกเตอร์	36
3-11	ดัชนีบิตแมปแบบช่วง บนแอทริบิวต์ A ใช้ 8 บิตแมปเวกเตอร์	37
3-12	ดัชนีบิตแมปแบบช่วง บนแอทริบิวต์ brand ใช้ 10 บิตแมปเวกเตอร์	40

## รายการภาพประกอบ (ต่อ)

ภาพประกอบ	หน้า
3-13 การดำเนินการตรรกะระดับบิตของแตริบิวต์ type และ brand ของดัชนีบิตแมปแบบช่วง	41
3-14 ตัวอย่างดัชนีบิตแมปแบบเข้ารหัส บนแตริบิวต์ A (มีค่า $C = 15$ )	42
4-1 รูปแบบทั่วไปของดัชนีบิตแมปแบบกระจาย บนแตริบิวต์ A (มีค่า $C = 15$ ) โดยมีจำนวนสมาชิกภายในกลุ่ม Z เท่ากับ 5 ( $m = 5$ )	47
4-2 ตัวอย่างแผนภาพการทำดัชนีบิตแมปแบบกระจายบนแตริบิวต์ A (มีค่า $C = 15, m = 5$ ) ใช้ 8 บิตแมปเวกเตอร์	48
4-3 รูปแบบของดัชนีบิตแมปแบบกระจาย เมื่อ $C = 15, m = 2$	49
4-4 ดัชนีบิตแมปแบบกระจายบนแตริบิวต์ A เมื่อ $C = 15$ และ $m = 2$ ใช้ 16 บิตแมปเวกเตอร์	52
4-5 รูปแบบของดัชนีบิตแมปแบบกระจาย เมื่อ $C = 15, m = 3$	53
4-6 ดัชนีบิตแมปแบบกระจายบนแตริบิวต์ A เมื่อ $C = 15$ และ $m = 3$ ใช้ 10 บิตแมปเวกเตอร์	55
4-7 รูปแบบของดัชนีบิตแมปแบบกระจาย เมื่อ $C = 15, m = 4$	55
4-8 ดัชนีบิตแมปแบบกระจายบนแตริบิวต์ A เมื่อ $C = 15$ และ $m = 4$ ใช้ 8 บิตแมปเวกเตอร์	57
4-9 รูปแบบของดัชนีบิตแมปแบบกระจาย บนแตริบิวต์ A เมื่อ $C = 15, m = 7$	58
4-10 ดัชนีบิตแมปแบบกระจายบนแตริบิวต์ A เมื่อ $C = 15$ และ $m = 7$ ใช้ 9 บิตแมปเวกเตอร์	59
4-11 ดัชนีบิตแมปแบบกระจายบนแตริบิวต์ A เมื่อ $C = 15$ และ $m = 5$ (ค่าที่เหมาะสมที่สุด) ใช้ 8 บิตแมปเวกเตอร์	64
4-12 รูปแบบของดัชนีบิตแมปแบบกระจาย เมื่อ $C = 20, m = 6$	67
4-13 ดัชนีบิตแมปแบบกระจายบนแตริบิวต์ brand เมื่อ $C = 20$ และ $m = 6$ (ค่าที่เหมาะสมที่สุด) ใช้ 9 บิตแมปเวกเตอร์	68
4-14 การดำเนินการตรรกะระดับบิตของแตริบิวต์ type และ brand ของดัชนีบิตแมปแบบกระจาย	69
5-1 กราฟแสดงการเปรียบเทียบการประเมินประสิทธิภาพในเรื่องพื้นที่ที่ใช้ในการจัดเก็บดัชนีบิตแมปทั้ง 5 แบบ	72
5-2 กราฟแสดงการเปรียบเทียบการประเมินประสิทธิภาพในเรื่องพื้นที่ที่ใช้ในการจัดเก็บดัชนีบิตแมปแบบช่วง แบบเข้ารหัส และแบบกระจาย	73

## รายการภาพประกอบ (ต่อ)

ภาพประกอบ	หน้า
5-3 กราฟแสดงพื้นที่ที่ใช้ในการจัดเก็บดัชนีบิตแมปทั้ง 4 แบบ (ข้อมูลทดสอบนำมาจากการวัดเปรียบเทียบสมรรถนะของ TPC-H)	75
5-4 กราฟแสดงเวลาที่ใช้ในการค้นหาข้อมูลแบบค่าเท่ากันของดัชนีบิตแมปทั้ง 4 แบบ	76
5-5 กราฟแสดงเวลาที่ใช้ในการค้นหาข้อมูลแบบความเป็นสมาชิกของแอทริบิวต์ที่มีค่า $C = 5$ ของดัชนีบิตแมปทั้ง 4 แบบ	77
5-6 กราฟแสดงเวลาที่ใช้ในการค้นหาข้อมูลแบบความเป็นสมาชิกของแอทริบิวต์ที่มีค่า $C = 25$ ของดัชนีบิตแมปทั้ง 4 แบบ	78
5-7 กราฟแสดงเวลาที่ใช้ในการค้นหาข้อมูลแบบความเป็นสมาชิกของแอทริบิวต์ที่มีค่า $C = 50$ ของดัชนีบิตแมปทั้ง 4 แบบ	79
5-8 กราฟแสดงเวลาที่ใช้ในการค้นหาข้อมูลแบบความเป็นสมาชิกของดัชนีบิตแมปทั้ง 4 แบบ โดยที่ข้อมูลที่ต้องการค้นหาอยู่ในกลุ่ม $Z$ เดียวกัน ( $C = 5, m = 4$ )	80
5-9 กราฟแสดงเวลาที่ใช้ในการค้นหาข้อมูลแบบความเป็นสมาชิกของดัชนีบิตแมปทั้ง 4 แบบ โดยที่ข้อมูลที่ต้องการค้นหาอยู่ในกลุ่ม $Z$ เดียวกัน ( $C=25, m=6$ )	81
5-10 กราฟแสดงความสัมพันธ์ระหว่างพื้นที่ที่ใช้ในการจัดเก็บดัชนีและเวลาที่ใช้ในการค้นหาข้อมูล กรณีที่ $C = 5$ และเป็นการค้นหาข้อมูลแบบค่าเท่ากัน (ข้อมูลทดสอบนำมาจากการวัดเปรียบเทียบสมรรถนะของ TPC-H)	82
5-11 กราฟแสดงความสัมพันธ์ระหว่างพื้นที่ที่ใช้ในการจัดเก็บดัชนีและเวลาที่ใช้ในการค้นหาข้อมูล กรณีที่ $C = 25$ และเป็นการค้นหาข้อมูลแบบค่าเท่ากัน (ข้อมูลทดสอบนำมาจากการวัดเปรียบเทียบสมรรถนะของ TPC-H)	83
5-12 กราฟแสดงความสัมพันธ์ระหว่างพื้นที่ที่ใช้ในการจัดเก็บดัชนีและเวลาที่ใช้ในการค้นหาข้อมูล กรณีที่ $C = 50$ และเป็นการค้นหาข้อมูลแบบค่าเท่ากัน (ข้อมูลทดสอบนำมาจากการวัดเปรียบเทียบสมรรถนะของ TPC-H)	84
5-13 กราฟแสดงความสัมพันธ์ระหว่างพื้นที่ที่ใช้ในการจัดเก็บดัชนีและเวลาที่ใช้ในการค้นหาข้อมูล กรณี $C = 5$ การสอบถามข้อมูลแบบความเป็นสมาชิก โดย $n$ คือจำนวนค่าที่ต้องการค้นหา (ข้อมูลทดสอบนำมาจากการวัดสมรรถนะของ TPC-H)	85

## รายการภาพประกอบ (ต่อ)

ภาพประกอบ	หน้า
5-14 กราฟแสดงความสัมพันธ์ระหว่างพื้นที่ที่ใช้ในการจัดเก็บดัชนีและเวลาที่ใช้ในการค้นหาข้อมูล กรณี $C = 25$ การสอบถามข้อมูลแบบความเป็นสมาชิก โดย $n$ คือ จำนวนค่าที่ต้องการค้นหา (ข้อมูลทดสอบนำมาจากประวัติสมรรถนะของ TPC-H)	86
5-15 กราฟแสดงความสัมพันธ์ระหว่างพื้นที่ที่ใช้ในการจัดเก็บดัชนีและเวลาที่ใช้ในการค้นหาข้อมูล กรณี $C = 50$ การสอบถามข้อมูลแบบความเป็นสมาชิก โดย $n$ คือ จำนวนค่าที่ต้องการค้นหา (ข้อมูลทดสอบนำมาจากประวัติสมรรถนะของ TPC-H)	87
5-16 กราฟแสดงความสัมพันธ์ระหว่างพื้นที่ที่ใช้ในการจัดเก็บดัชนีและเวลาที่ใช้ในการค้นหาข้อมูลแบบความเป็นสมาชิก กรณีที่ค่าที่ต้องการค้นหาอยู่ในกลุ่ม $Z$ เดียวกัน ( $C = 5, n = 4$ ) (ข้อมูลทดสอบนำมาจากประวัติสมรรถนะของ TPC-H)	89
5-17 กราฟแสดงความสัมพันธ์ระหว่างพื้นที่ที่ใช้ในการจัดเก็บดัชนีและเวลาที่ใช้ในการค้นหาข้อมูลแบบความเป็นสมาชิก กรณีที่ค่าที่ต้องการค้นหาอยู่ในกลุ่ม $Z$ เดียวกัน ( $C = 25, n = 6$ ) (ข้อมูลทดสอบนำมาจากประวัติสมรรถนะของ TPC-H)	90