สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง

การพัฒนาประสิทธิภาพของอัลกอริทึมเอไพรออริ

A Development of the Apriori Algorithm



นางสาวสุภาวดี ยี่รงค์

นายสุรจักษ์ คูมณีจิรกุล

ตรหมู่.....
เลขทะเบียน.....117243
วัน,เดือน,ปี.....19 ก.ค. 2554

b.....
i.....

โครงงานพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2553

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อปัญหาพิเศษ	การพัฒนาประสิทธิภาพของอัลกอริทึมเอไพรออริ
	A Development of the Apriori Algorithm
ชื่อนักศึกษา	นางสาวสุภาวดี ยี่รงค์ 50050228
	นายสุรจักษ์ คูมณีจิรกุล 50050229
ปริญญา	วิทยาศาสตรบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษา	ดร.อนันตพร ศรีสวัสดิ์

คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง อนุมัติให้โครงงานพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร วิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ประจำปีการศึกษา 2553

คณะกรรมการสอบ		ลายมือชื่อ
ประธานกรรมการ	ผศ.ดร.นันทิกา เบญจเทพานันท์	
กรรมการ	อ.สันธนะ อู่อุดมยิ่ง	
กรรมการและอาจารย์ที่ปรึกษา	ดร.อนันตพร ศรีสวัสดิ์	

ลิขสิทธิ์ของคณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อปัญหาพิเศษ	การพัฒนาประสิทธิภาพของอัลกอริทึมเอไพรออริ
ชื่อนักศึกษา	นางสาวสุภาวดี ยี่รงค์ 50050228
	นายสุรจักษ์ คูมณีจิรกุล 50050229
ปริญญา	วิทยาศาสตรบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
ปีการศึกษา	2553
อาจารย์ที่ปรึกษา	ดร.อนันตพร ศรีสวัสดิ์

บทคัดย่อ

ปัญหาพิเศษนี้ได้ทำการนำเสนอการพัฒนาอัลกอริทึมในการค้นหาไอเท็มเซตปรากฏบ่อยให้มีจำนวนครั้งในการอ่านข้อมูลจากฐานข้อมูลน้อยลง จากการทดลองเปรียบเทียบเวลาการทำงานในการค้นหาไอเท็มเซตปรากฏบ่อยระหว่างอัลกอริทึมเอไพรออริกับอัลกอริทึมที่ได้พัฒนาขึ้น พบว่าอัลกอริทึมที่ได้พัฒนาขึ้นใช้เวลาในการค้นหาไอเท็มเซตปรากฏบ่อยน้อยกว่าอัลกอริทึมเอไพรออริในกรณีที่กำหนดค่าสนับสนุนขั้นต่ำที่สูง และพบว่าผลต่างระหว่างเวลาในการทำงานของอัลกอริทึมที่พัฒนาขึ้นกับอัลกอริทึมเอไพรออริมีค่ามากขึ้นเมื่อมีจำนวนรายการเพิ่มขึ้น นอกจากนี้ได้ทำการสร้างหน้าจอโปรแกรมสำเร็จรูปที่ใช้สำหรับค้นหากฎความสัมพันธ์ของข้อมูลด้วย

คำสำคัญ : อัลกอริทึมเอไพรออริ ไอเท็มเซตปรากฏบ่อย กฎความสัมพันธ์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Title	The Development of Apriori Algorithm	
Students	Ms.Supawadee Yeeronk	50050228
	Mr.Surajuk Koomaneejirakul	50050229
Degree	Bachelor of Science	
Program	Computer Science	
Academic Year	2010	
Advisor	Dr.Anantaporn Srisawat	

ABSTRACT

This special problem presents a new algorithm to generate frequent itemsets. The objective of this algorithm is to reduce time for scanning a database. In the experiments, we compared the running time between the proposed algorithm and the Apriori algorithm for finding frequent itemsets. The results show that the proposed algorithm spends less time on fiding frequent itemsets than Apriori when a minimum support is high. In addition, the difference between run time of these two algorithms is greater when a number of transactions is increased. Finally, we also implemented an application software for association rule mining.

Keywords : Apriori algorithm, Frequent Itemset, Association rule

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

การทำปัญหาพิเศษหัวข้อการพัฒนาประสิทธิภาพของอัลกอริทึมเอไพรออริทั้งในส่วนของอัลกอริทึม และหน้าจอโปรแกรมสำหรับผู้ใช้งานสำเร็จลุล่วงไปได้ด้วยดี ด้วยการช่วยเหลือและสนับสนุนจากบุคคลากรหลายท่าน ดังนี้

บิดา มารดา และบุคคลในครอบครัวทุกท่าน ผู้ซึ่งมีพระคุณอย่างมากที่ได้ให้กำเนิดเลี้ยงดู อบรม ส่งเสริมให้ได้รับและกระทำใน สิ่งที่ดีมาโดยตลอด รวมทั้งเป็นกำลังใจและคอยรับฟังและเสนอแนวทางแก้ไขปัญหาตลอดมา

ดร.อนันตพร ศรีสวัสดิ์ที่เป็นอาจารย์ที่ปรึกษาปัญหาพิเศษที่คอยให้คำชี้แนะ คำปรึกษา การเสนอวิธีแก้ไขปัญหา และปรับปรุงรูปแบบรายงานด้วยดีตลอดการทำปัญหาพิเศษนี้เพื่อให้ผลงานออกมาอย่างดีที่สุด

ผศ.ดร.นันทิกา เบญจเทพานันท์ และ อ.สันธนะ อู่อุดมยิ่งที่เป็นประธาน และกรรมการการสอบที่ให้คำชี้แนะถึงขั้นตอนวิธีในการทำปัญหาพิเศษและแก้ไขปรับปรุงรูปแบบรายงานในครั้งนี้

อาจารย์ภาควิชาวิทยาการคอมพิวเตอร์ทุกๆท่านที่ได้ประสิทธิ์ประสาทวิชาความรู้ทั้งในภาคทฤษฎีและภาคปฏิบัติแก่ผู้จัดทำตลอดเวลาทั้ง 4 ปีจนกระทั่งปัญหาพิเศษสัมฤทธิ์ผลได้ด้วยดีทุกประการ

บุคคลที่ร่วมทำงานปัญหาพิเศษด้วยกัน รวมถึงเพื่อนๆ รุ่นพี่ รุ่นน้องทุกคนที่คอยเป็นกำลังใจในยามท้อแท้ และคอยรับฟังกรณีมีปัญหา รวมถึงช่วยแก้ไขปัญหาตลอดมา และบุคคลอื่นที่อาจมิได้กล่าวถึงในที่นี้ซึ่งมีส่วนร่วมและคอยช่วยเหลือในการทำปัญหาพิเศษในครั้งนี้

นางสาวสุภาวดี ยี่รงค์
นายสุรจักษ์ คูมณีจิรกุล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	i
บทคัดย่อภาษาอังกฤษ	ii
กิตติกรรมประกาศ	iii
สารบัญ	iv
สารบัญตาราง	vi
สารบัญรูป	viii
บทที่ 1 บทนำ	1
1.1 ที่มาและความสำคัญของปัญหา	1
1.2 วัตถุประสงค์ของการวิจัย	2
1.3 ขอบเขตของการวิจัย	2
1.4 ผลที่คาดว่าจะได้รับ	2
1.5 ขั้นตอนในการดำเนินงาน	2
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง	3
2.1 ทฤษฎีที่เกี่ยวข้อง	3
2.1.1 การทำเหมืองข้อมูล	3
2.1.2 การค้นหากฎความสัมพันธ์	3
2.1.2.1 การหาไอเท็มเซตปรากฏบ่อย	5
2.1.2.2 การหากฎความสัมพันธ์ของข้อมูล	8
2.1.2.2.1 ประโยชน์ของกฎความสัมพันธ์	10
2.1.2.2.2 การวัดความน่าสนใจของกฎความสัมพันธ์	10
2.1.3 ความจริงเกี่ยวกับค่าสนับสนุนและค่าความเชื่อมั่น	11
2.2 งานวิจัยที่เกี่ยวข้อง	12
2.2.1 การเพิ่มประสิทธิภาพและปรับปรุงอัลกอริทึมเอไพรออริ	12
2.2.1.1 แนวคิด	12
2.2.1.2 ขั้นตอนการทำงาน	13
2.2.2 การปรับปรุงประสิทธิภาพอัลกอริทึมเอไพรออริด้วยเมทริกซ์	16
2.2.2.1 แนวคิด	16
2.2.2.2 ตัวอย่างการดำเนินการตามอัลกอริทึม	18

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ (ต่อ)

	หน้า
บทที่ 3 การพัฒนาอัลกอริทึมค้นหาไอเท็มเซตปรากฏบ่อย	21
3.1 หลักการทำงาน	21
3.2 ตัวอย่างการทำงาน	23
3.3 ส่วนโปรแกรมติดต่อกับผู้ใช้	33
3.4 รูปแบบของไฟล์ที่รองรับ	43
บทที่ 4 ผลการทดลอง	44
4.1 แหล่งที่มาและรายละเอียดของข้อมูล	44
4.1.1 ข้อมูล Data Generator	44
4.1.2 ข้อมูล Mushroom	45
4.2 ผลการทดลอง	46
บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ	52
5.1 สรุปผลการวิจัย	52
5.2 ข้อเสนอแนะ	53
เอกสารอ้างอิง	54
ภาคผนวก การใช้งานโปรแกรม	56

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

ตารางที่	หน้า
2.1 ข้อมูลในการหาค่าสหสัมพันธ์	10
2.2 ข้อมูลรายการซื้อสินค้า	18
2.3 เมทริกซ์ตั้งต้น	18
2.4 เมทริกซ์ที่ใช้ในการหาไอเท็มเซตปรากฏบ่อย	19
2.5 การ AND กันของ 2 3 และ 5	19
3.1 ฐานข้อมูล	23
3.2 ข้อมูลที่อยู่ในรูปของเมทริกซ์	24
3.3 การ AND กันของ I1 I2 และ I5	25
3.4 เมทริกซ์ที่ได้หลังจากลบ T100	25
3.5 การ AND กันของ I2 และ I4	26
3.6 เมทริกซ์ที่ได้หลังจากลบ T200	26
3.7 การ AND กันของ I2 และ I3	27
3.8 เมทริกซ์ที่ได้หลังจากลบ T300	27
3.9 การ AND กันของ I1 I2 และ I4	28
3.10 การ AND กันของ I1 และ I4	28
3.11 เมทริกซ์ที่ได้หลังจากลบ T400	28
3.12 การ AND กันของ I1 และ I3	29
3.13 เมทริกซ์ที่ได้หลังจากลบ T500	29
3.14 เมทริกซ์ที่ได้หลังจากลบ T600	30
3.15 เมทริกซ์ที่ได้หลังจากลบ T700	30
3.16 การ AND กันของ I1 I2 และ I3	31
3.17 การ AND กันของ I1 I2 และ I3	31
3.18 การ AND กันของ I1 I3 และ I5	32
3.19 การ AND กันของ I3 และ I5	32
3.20 เมทริกซ์ที่ได้หลังจากลบ T800	32
3.21 Use case description เลือกแฟ้มข้อมูลมาทำการวิเคราะห์	35
3.22 Use case description แสดงผลลัพธ์ไอเท็มเซตปรากฏบ่อย	36
3.23 Use case description แสดงผลลัพธ์ในรูปแบบกราฟ	37

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง (ต่อ)

ตารางที่	หน้า
3.24 Use case description บันทึกกราฟ	38
3.25 Use case description แสดงกฎความสัมพันธ์	39
3.26 Use case description บันทึกกฎความสัมพันธ์	40
3.27 Use case description ดูคำแนะนำการใช้งาน	41
4.1 แสดงลักษณะของข้อมูล	44
4.2 เวลาเฉลี่ยในการหาไอเท็มเซตปรากฏบ่อยของอัลกอริทึมเอไพรออริกับ อัลกอริทึมที่พัฒนาขึ้นในข้อมูล Data Generator	46
4.3 เวลาเฉลี่ยในการค้นหาไอเท็มเซตปรากฏบ่อยของในชุดข้อมูล Data Generator ในจำนวนรายการต่างๆ ที่ค่าสนับสนุนขั้นต่ำ 40%	47
4.4 เวลาเฉลี่ยในการค้นหาไอเท็มเซตปรากฏบ่อยของในชุดข้อมูล Data Generator ในจำนวนรายการต่างๆ ที่ค่าสนับสนุนขั้นต่ำ 50%	48
4.5 เวลาเฉลี่ยในการค้นหาไอเท็มเซตปรากฏบ่อยของในชุดข้อมูล Data Generator ในจำนวนรายการต่างๆ ที่ค่าสนับสนุนขั้นต่ำ 60%	49
4.6 เวลาเฉลี่ยในการหาไอเท็มเซตปรากฏบ่อยของอัลกอริทึมเอไพรออริกับอัลกอริทึมที่พัฒนาขึ้นในข้อมูล Mushroom	50

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูปภาพ

รูปที่	หน้า
2.1 Pseudo code ขั้นตอนการหาไอเท็มเซตปรากฏบ่อย	5
2.2 การรวม (Joining) และการตัดไอเท็มเซตปรากฏบ่อยบางตัวออก (Pruning)	6
2.3 การสร้างแคนดิเดตไอเท็มเซต C_3 จาก L_2	6
2.4 ตัวอย่างการรวม และการตัดไอเท็มเซตปรากฏบ่อยบางตัวออก	6
2.5 การทำงานของอัลกอริทึมเอไพรออริ	7
2.6 อัลกอริทึมเอไพรออริที่มีการปรับปรุง	14
2.7 ตัวอย่างการดำเนินการตามอัลกอริทึมเอไพรออริ	14
2.8 อัลกอริทึมเอไพรออริแบบเมทริกซ์ในรูปแบบของโปรแกรม	17
3.1 ลักษณะของเมทริกซ์ขนาด m x n	21
3.2 ขั้นตอนการทำงานของอัลกอริทึมในการค้นหาไอเท็มเซตปรากฏบ่อย	23
3.3 Use Case Diagram แสดงการทำงานของโปรแกรม	34
3.4 กิจกรรมในการใช้งานระบบของผู้ใช้	44
4.1 กราฟแสดงเวลาการทำงานในการหาไอเท็มเซตปรากฏบ่อยของอัลกอริทึมในชุดข้อมูล Data Generator	47
4.2 เวลาเฉลี่ยในการค้นหาไอเท็มเซตปรากฏบ่อยของในชุดข้อมูล Data Generator ในจำนวนรายการต่างๆ ที่ค่าสนับสนุนขั้นต่ำ 40%	48
4.3 เวลาเฉลี่ยในการค้นหาไอเท็มเซตปรากฏบ่อยของในชุดข้อมูล Data ในจำนวนรายการต่างๆ ที่ค่าสนับสนุนขั้นต่ำ 50%	49
4.4 เวลาเฉลี่ยในการค้นหาไอเท็มเซตปรากฏบ่อยของในชุดข้อมูล Data Generator ในจำนวนรายการต่างๆ ที่ค่าสนับสนุนขั้นต่ำ 60%	50
4.5 กราฟแสดงเวลาการทำงานในการหาไอเท็มเซตปรากฏบ่อยของอัลกอริทึมในชุดข้อมูล Mushroom	51
ก.1 หน้าจอหลักของโปรแกรม	56
ก.2 แสดงคำอธิบายของไอคอนเมื่อผู้ใช้นำเมาส์ไปวางบนไอคอน	56
ก.3 หน้าต่างเลือกไฟล์ที่ผู้ใช้ต้องการเปิด	57
ก.4 หน้าต่างสำหรับระบุค่า minimum support	58
ก.5 ตารางแสดงผลลัพธ์ในการหาไอเท็มเซตปรากฏบ่อย	58

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูปภาพ (ต่อ)

รูปที่	หน้า
ก.6 หน้าต่างสำหรับระบุค่า minimum support และ minimum confidence	59
ก.7 ตารางแสดงผลลัพธ์ในการหากฎความสัมพันธ์	60
ก.8 กราฟเปรียบเทียบเวลาและค่าสนับสนุนขั้นต่ำระหว่างอัลกอริทึมทั้งสอง	61
ก.9 หน้าต่างบันทึกกราฟ	62
ก.10 หน้าต่างคำอธิบายการใช้งานโปรแกรม	62

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในปัจจุบันข้อมูลมีความสำคัญต่อการพัฒนาประเทศเป็นอย่างมาก แต่การรวบรวมข้อมูลนั้นมีทั้งข้อมูลที่มีประโยชน์และไม่มีประโยชน์ จึงได้มีการใช้วิธีการทำเหมืองข้อมูล (Data Mining) นำมาค้นหาข้อมูลที่เป็นประโยชน์และมีความน่าสนใจจากฐานข้อมูลที่มีขนาดใหญ่ โดยการนำข้อมูลที่มีมาวิเคราะห์แล้วดึงสิ่งที่สำคัญออกมา เพื่อนำมาใช้ในการวิเคราะห์หรือทำการทำนายสิ่งต่างๆ ที่จะเกิดประโยชน์ โดยใช้เทคนิคต่างๆ มากมาย ซึ่งเทคนิคที่น่าสนใจ คือการสร้างกฎความสัมพันธ์ (Association Rules) ซึ่งเป็นการค้นหาความสัมพันธ์ของข้อมูลจากฐานข้อมูลขนาดใหญ่ที่มีอยู่ เพื่อนำมาวิเคราะห์และทำนายปรากฏการณ์ต่างๆ โดยเฉพาะการนำไปใช้กับห้างสรรพสินค้าที่มีการเก็บข้อมูลการซื้อสินค้าของลูกค้า เป็นต้น

ขั้นตอนการสร้างกฎความสัมพันธ์ของข้อมูล ประกอบด้วยสองขั้นตอน คือขั้นตอนในการหาไอเท็มเซตปรากฏบ่อย (Frequent Itemset) ในฐานข้อมูล ซึ่งเป็นการหาเซตของไอเท็ม (Itemset) ที่มีค่าสนับสนุน (support) มากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ (Minimum support) และขั้นตอนการสร้างกฎความสัมพันธ์ (Association Rules) จากไอเท็มเซตปรากฏบ่อยที่ได้ โดยขั้นตอนวิธีในการหาไอเท็มเซตปรากฏบ่อยที่นิยมนำมาใช้ คืออัลกอริทึมเอไพรออริ (Apriori Algorithm) ซึ่งทำงานโดยการอ่านข้อมูลจากฐานข้อมูลเพื่อสร้างไอเท็มเซตปรากฏบ่อย (Frequent Itemset) แต่หลักการสำคัญของอัลกอริทึมเอไพรออริยังคงมีปัญหาอยู่ คือการอ่านข้อมูลจากฐานข้อมูลทุกครั้งที่มีการหาค่าสนับสนุนของไอเท็มเซต ทำให้ใช้เวลานานในการหาไอเท็มเซตปรากฏบ่อย และมีการสร้างไอเท็มเซตปรากฏบ่อยจำนวนมาก จึงทำให้มีจำนวนกฎความสัมพันธ์จำนวนมากด้วย และยากแก่การตีความจากกฎความสัมพันธ์ที่ได้

งานวิจัยนี้จึงจัดทำขึ้นเพื่อช่วยลดปัญหาในการอ่านข้อมูลจากฐานข้อมูล ในขั้นตอนการหาไอเท็มเซตปรากฏบ่อย ให้มีรอบของการอ่านข้อมูลจากฐานข้อมูลน้อยลง ซึ่งจะทำให้เวลาในการประมวลผล ของอัลกอริทึมเอไพรออริมีความรวดเร็วขึ้น โดยวิธีการปรับโครงสร้างของข้อมูลเพื่อให้สะดวกในการค้นหาข้อมูลแทนที่จะทำการค้นหาโดยการอ่านจากฐานข้อมูลซึ่งมีขนาดใหญ่และซับซ้อนกว่า และใช้หลักการ AND ทางตรรกะคณิตศาสตร์ในการพิจารณาการเกิดร่วมกันของไอเท็ม โดยการนำข้อมูลของไอเท็มที่ปรากฏในแต่ละรายการมาทำการ AND เพื่อใช้ในการค้นหาไอเท็มเซตปรากฏบ่อย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.2 วัตถุประสงค์ของการวิจัย

พัฒนาอัลกอริทึมที่ใช้ในการหาไอเท็มเซตปรากฏบ่อย ให้มีจำนวนครั้งในการอ่านข้อมูลจากฐานข้อมูลน้อยลง เพื่อลดเวลาในการหาไอเท็มเซตปรากฏบ่อย

1.3 ขอบเขตของการวิจัย

1. พัฒนาอัลกอริทึมที่ใช้ในการหาไอเท็มเซตปรากฏบ่อย ที่มีสมาชิกในเซตน้อยกว่าหรือเท่ากับสี่ไอเท็ม (Frequent 4-items)

2. เปรียบเทียบเวลาที่ใช้ในการหาไอเท็มเซตปรากฏบ่อยระหว่างอัลกอริทึมที่พัฒนาขึ้นกับอัลกอริทึมเอไพรออริ

3. พัฒนาโปรแกรมสำหรับผู้ใช้เพื่อใช้ในการค้นหากฎความสัมพันธ์

1.4 ผลที่คาดว่าจะได้รับ

1. ได้อัลกอริทึมที่ลดการอ่านข้อมูลจากฐานข้อมูล ทำให้ใช้เวลาในการหาไอเท็มเซตปรากฏบ่อยน้อยลง

2. โปรแกรมสำเร็จรูปที่ใช้สำหรับค้นหากฎความสัมพันธ์

1.5 ขั้นตอนในการดำเนินงาน

1. ศึกษาทฤษฎีอัลกอริทึมเอไพรออริ (Apriori Algorithm)
2. ออกแบบอัลกอริทึมการค้นหาไอเท็มเซตปรากฏบ่อย
3. รวบรวมและสร้างข้อมูลตัวอย่างเพื่อทดสอบอัลกอริทึมใหม่ที่ได้พัฒนาขึ้น
4. จัดทำโปรแกรมด้วยภาษาจาวาตามที่ได้ออกแบบไว้
5. ทดสอบและเปรียบเทียบอัลกอริทึมเอไพรออริกับอัลกอริทึมที่ได้พัฒนาขึ้น
6. ปรับปรุงและตกแต่งโปรแกรมในส่วนของการติดต่อกับผู้ใช้ให้ใช้งานได้ง่าย
7. จัดทำเอกสาร

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 การทำเหมืองข้อมูล

การทำเหมืองข้อมูล (Data Mining) คือการค้นหาข้อมูลจากฐานข้อมูลขนาดใหญ่ (Knowledge Discovery in Database) เป็นการค้นหาสิ่งที่สำคัญของข้อมูล โดยไม่ทราบมาก่อน และมีแนวโน้มว่าจะมีประโยชน์ โดยทำการค้นหาข้อมูลในฐานข้อมูล ซึ่งข้อมูลที่จะนำมาทำเหมืองข้อมูล จะต้องเป็นข้อมูลที่ไม่มีการเปลี่ยนแปลงตลอดการทำเหมืองข้อมูล ถ้าข้อมูลมีการเปลี่ยนแปลงในช่วงเวลานั้น จะทำให้ผลลัพธ์ที่ได้จากการทำเหมืองข้อมูล ถูกต้องแค่ช่วงเวลาหนึ่งเท่านั้น การทำเหมืองข้อมูลมีการใช้เทคนิคที่แตกต่างกันออกไปมากมาย เช่น การสร้างกฎความสัมพันธ์ (Association Rules) และการแบ่งกลุ่มข้อมูล (Clustering) หรือจำแนกข้อมูล (Classification) เป็นต้น

2.1.2 การค้นหากฎความสัมพันธ์

การค้นหากฎความสัมพันธ์ (Association Rules Discovery)[1] คือการค้นหาความสัมพันธ์ของข้อมูลจากฐานข้อมูลขนาดใหญ่ที่มีอยู่เพื่อช่วยในการวิเคราะห์ และตัดสินใจในธุรกิจตัวอย่างของการค้นหากฎความสัมพันธ์ เช่น ในการวิเคราะห์การซื้อสินค้าของลูกค้า (Market Basket Analysis) ซึ่งการสร้างกฎความสัมพันธ์ของข้อมูลได้แบ่งขั้นตอนออกเป็น 2 ขั้นตอน คือ ขั้นตอนในการหาไอเท็มเซตปรากฏบ่อย และการสร้างกฎความสัมพันธ์ของข้อมูล ซึ่งวิธีหนึ่งที่นิยมนำมาใช้ในการหาไอเท็มเซตปรากฏบ่อย คืออัลกอริทึมเอไพรออริ (Apriori Algorithm) ซึ่งเสนอโดย Agrawal ในปี ค.ศ. 1993

นิยามและคำศัพท์ต่างๆ ที่เกี่ยวข้อง

1. $I = \{i_1, i_2, \dots, i_m\}$ เป็นเซตของตัวอักษร หรือตัวเลข (Literals) และเรียกในแต่ละสมาชิกของ I ว่าไอเท็ม (Item)

2. $D = \{t_1, t_2, \dots, t_n\}$ เป็นฐานข้อมูลซึ่งประกอบด้วยเซตของรายการ (Transaction) และแต่ละรายการ t_i ประกอบด้วยหมายเลขรายการ (TID) และไอเท็มเซต $I_t \subseteq I$

3. TID หมายถึงหมายเลข หรือรหัสประจำตัวแต่ละรายการที่ไม่ซ้ำกัน (unique identifier)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4. ไอเท็มเซต (Itemset) หมายถึงเซตที่มีสมาชิกเป็นไอเท็ม K ตัว จะถูกเรียกว่าเป็น K-Itemset

5. ค่าสนับสนุน (Support) หมายถึงจำนวนของรายการ (t_i) ที่มีไอเท็มทั้งหมดปรากฏอยู่ร่วมกันเทียบกับจำนวนรายการทั้งหมดในฐานข้อมูล เช่น ค่าสนับสนุนของสินค้า A กับ B เท่ากับ 35% ซึ่งหมายความว่าจากการซื้อสินค้าทั้งหมด 100 รายการ นั้นจะมีการซื้อสินค้า A กับ B ด้วยกันอยู่ 35 รายการ

6. ค่าสนับสนุนขั้นต่ำ (Minimum Support) เป็นค่าขีดแบ่ง (Threshold) ที่ใช้ในการวัดค่าสนับสนุนของไอเท็มเซตโดยที่ผู้ใช้เป็นคนกำหนดค่าซึ่งนิยมระบุเป็นร้อยละ หรือเปอร์เซ็นต์

7. ค่าความเชื่อมั่น (Confidence) คือ จำนวนของรายการที่เมื่อมีไอเท็มหนึ่งเกิดขึ้นแล้วมักจะมีอีกไอเท็มหนึ่งเกิดขึ้นร่วมด้วย เทียบกับจำนวนรายการที่มีไอเท็มแรกทั้งหมด เช่น ค่าความเชื่อมั่นของสินค้า A กับ B ($A \rightarrow B$) หาได้จาก จำนวนรายการที่มีสินค้า A และ B เทียบกับ จำนวนรายการที่มีสินค้า A

8. ความเชื่อมั่นขั้นต่ำ (Minimum confidence) เป็นค่าขีดแบ่งที่ใช้ในการวัดค่าความเชื่อมั่นโดยที่ผู้ใช้เป็นคนกำหนดค่าซึ่งนิยมระบุเป็นร้อยละ หรือเปอร์เซ็นต์

9. ไอเท็มเซตปรากฏบ่อย (Frequent Itemset) หมายถึงไอเท็มเซตที่มีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำในกรณีที่ไอเท็มเซตปรากฏบ่อยมีจำนวนสมาชิกเท่ากับ k จำนวนเซตนี้จะถูกเรียกว่าเป็น k- Frequent Itemset มักถูกแทนด้วยสัญลักษณ์ L_k

10. แคนดิเดตไอเท็มเซต (Candidate Itemset) หมายถึงเซตของ k-Itemset ที่มีโอกาสเป็นไอเท็มเซตปรากฏบ่อย ซึ่งเป็นไอเท็มเซตที่ยังไม่ได้ทำการหาค่าสนับสนุน และยังไม่ได้ตรวจสอบว่าเป็นไอเท็มเซตปรากฏบ่อย หรือไม่มักถูกแทนด้วยสัญลักษณ์ C_k

11. กฎความสัมพันธ์ที่มีความแข็งแกร่ง (Strong association rules) คือ กฎความสัมพันธ์ที่มีค่าสนับสนุนมากกว่าค่าสนับสนุนขั้นต่ำ และมีค่าความเชื่อมั่นมากกว่าค่าความเชื่อมั่นขั้นต่ำ ซึ่งค่าสนับสนุนขั้นต่ำ และค่าความเชื่อมั่นขั้นต่ำนั้นผู้ใช้เป็นผู้กำหนด

12. ค่าสหสัมพันธ์ (Correlation) เป็นค่าทางสถิติที่ใช้หาความสัมพันธ์ของไอเท็มที่อยู่ในกฎความสัมพันธ์นั้น ตัวอย่างเช่นกฎความสัมพันธ์ $A \rightarrow C$ เป็นการหาว่าสินค้า A และสินค้า C มีความสัมพันธ์กันในเชิงบวก เชิงลบ หรือเป็นอิสระต่อกัน เป็นต้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

คุณสมบัติของอัลกอริทึมเอไพรออริ

คุณสมบัติของอัลกอริทึมเอไพรออริ คือ ทุกๆ เซตย่อย (subset) ของไอเท็มเซตปรากฏบ่อย (Frequent Itemset) ต้องเป็นไอเท็มเซตปรากฏบ่อย (Frequent Itemset)

2.1.2.1 การหาไอเท็มเซตปรากฏบ่อย

อัลกอริทึมเอไพรออริ (Apriori Algorithm) เป็นขั้นตอนที่นิยมใช้ในการหาไอเท็มเซตปรากฏบ่อย ซึ่งจะนำไอเท็มเซตปรากฏบ่อยนี้ไปสร้างกฎความสัมพันธ์ของข้อมูลต่อไป โดยสามารถหาไอเท็มเซตปรากฏบ่อยได้จากแคนดิเดตไอเท็มเซต ด้วยขั้นตอนดังรูปที่ 2.1

```
(1)  L1 = {large 1- itemsets};
(2)  for (k = 2; Lk-1 ≠ ∅; k++) do begin
(3)       Ck = apriori-gen(Lk-1);
(4)       for all transactions t∈D do begin
(5)  Ct = subset(Ck,t);
(6)                  for all candidates c∈Ct do
(7)             c.count++;
(8)       end
(9)       Lk = {c∈Ck |c.count≥minsup}
(10) end
(11) Answer =∪kLk;
```

รูปที่ 2.1 Pseudo code ขั้นตอนการหาไอเท็มเซตปรากฏบ่อย

ในขั้นแรกของอัลกอริทึมเอไพรออริจะทำการอ่านข้อมูลจากฐานข้อมูลโดยนับไอเท็มที่ปรากฏอยู่ในฐานข้อมูล เพื่อหาไอเท็มเซตปรากฏบ่อยที่มีสมาชิก 1 ไอเท็ม (Frequent 1-Itemset) หลังจากนั้นจะถูกนำมาใช้ในการสร้างแคนดิเดตไอเท็ม (Candidate Itemset) ในรอบต่อๆ ไป

แคนดิเดตไอเท็มเซตในแต่ละรอบสร้างจากการรวม (Joining) กันของไอเท็มเซตปรากฏบ่อยในระดับก่อนหน้า (L_{k-1}* L_{k-1}) มาเชื่อมต่อกันเป็นไอเท็มเซตชุดเดียวกัน และมีการตัดไอเท็มเซตปรากฏบ่อยบางตัวออก (Pruning) โดยพิจารณาจากเซตย่อย (k-1 subset) ที่ไม่ได้อยู่ในไอเท็มเซตปรากฏบ่อย (L_{k-1}) ซึ่งเป็นไปตามคุณสมบัติของอัลกอริทึมเอไพรออริเพื่อหาแคนดิเดตไอเท็มเซตที่แท้จริง ซึ่งขั้นตอนการรวม และการตัดไอเท็มเซตปรากฏบ่อยบางตัวออกแสดงดังรูปที่ 2.2 และรูปที่ 2.3

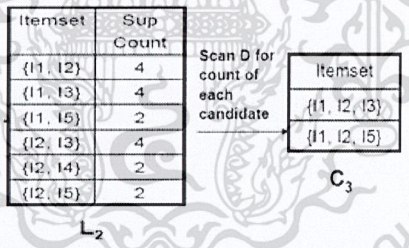
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Step 1 : join L_{k-1} with L_{k-1}
(1) Insert into C_k
(2) Select $p.item_1, p.item_2, \ldots, p.item_{k-1}, q.item_{k-1}$,
From L_{k-1} p, L_{k-1} q
Where $p.item_1 = q.item_1, \ldots, p.item_{k-2} = q.item_{k-2}, p.item_{k-1} < q.item_{k-1}$;
Step 2 : prune
(3) For all itemsets $C \in C_k$ do
(4) forall (K-1)-sunsets s of c do
(5) if ($s \in L_{k-1}$) then
(6) delete c from C_k

รูปที่ 2.2 การรวม (Joining) และการตัดไอเท็มเซตปรากฏบ่อยบางตัวออก (Pruning)

(1) Insert into C_3
(2) select $c.item_1, p.item_2, q.item_2$,
from L_2p, L_2q
where $p.item_1 = q.item_1, p.item_2 < q.item_2$;

รูปที่ 2.3 การสร้างแคนดิเดตไอเท็มเซต C_3 จาก L_2



รูปที่ 2.4 ตัวอย่างการรวม และการตัดไอเท็มเซตปรากฏบ่อยบางตัวออก

จากตัวอย่าง แสดงขั้นตอนการรวมและการตัดไอเท็มเซตปรากฏบ่อยบางตัวออกได้ดังนี้

สมมุติ L_2 = {{I1, I2}, {I1, I3}, {I1, I5}, {I2, I3}, {I2, I4}, {I2, I5}}

การรวม : $L_2 * L_2$

ได้ {I1, I2, I3} (จาก {I1, I2} และ {I1, I3})

{I1, I2, I5} (จาก {I1, I2} และ {I1, I5})

{I1, I3, I5} (จาก {I1, I3} และ {I1, I5})

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

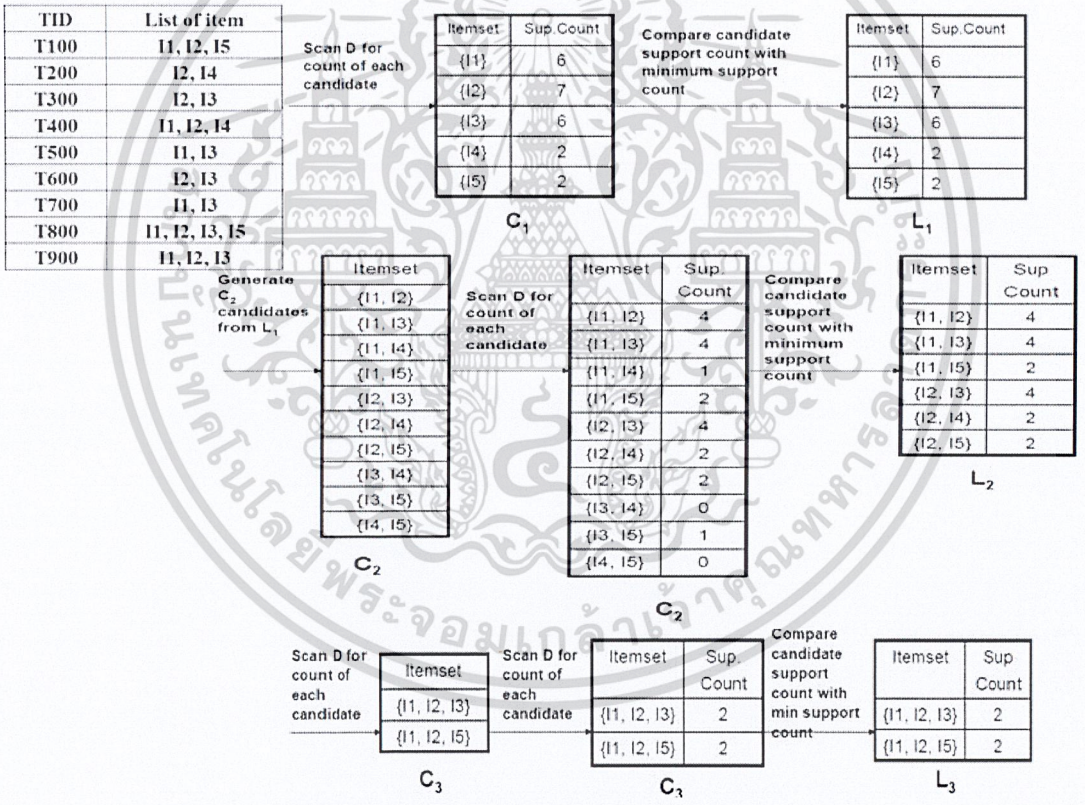
{I2, I3, I4} (จาก {I2, I3} และ {I2, I4})

{I2, I3, I5} (จาก {I2, I3} และ {I2, I5})

{I2, I4, I5} (จาก {I2, I4} และ {I2, I5})

การตัดออก : {I1, I3, I5}, {I2, I3, I4}, {I2, I3, I5}, {I2, I4, I5} และ {I2, I4, I5} ถูกตัดออกเนื่องจากเซตย่อยบางตัวของไอเท็มเซตปรากฏบ่อย เหล่านี้ ไม่ได้ปรากฏอยู่ใน L_2 เช่น {I3, I5} ซึ่งเป็นเซตย่อยของ {I1, I3, I5} และ {I2, I3, I5} แต่ไม่ปรากฏอยู่ใน L_2

C_3 = {I1, I2, I3}, {I1, I2, I5}



รูปที่ 2.5 การทำงานของอัลกอริทึมเอไพรออริ

จากรูปที่ 2.5 แสดงการทำงานของอัลกอริทึมเอไพรออริกับตัวอย่างของกลุ่มข้อมูล ซึ่งกำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 2 หลังจากอ่านข้อมูลจากฐานข้อมูลครั้งแรก ก็จะได้ C_1 (1 – Candidate itemset) แล้วอ่านข้อมูลจากฐานข้อมูลเพื่อนับค่าสนับสนุนแต่ละเซตของ C_1 แล้วนำมา

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เปรียบเทียบกับค่าสนับสนุนขั้นต่ำ ถ้าเซตของ C_1 ใดมีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ แสดงว่า เซตของ C_1 นั้น เป็นเซตของ L_1 ด้วย

ด้วยวิธีเดียวกันนั้นทำการสร้างแคนดิเดตไอเท็มเซตตัวถัดไปด้วยการรวม และการตัดไอเท็มเซตปรากฏบ่อยบางตัวออก แล้วอ่านข้อมูลจากฐานข้อมูลเพื่อหาค่าสนับสนุนของแคนดิเดตไอเท็มเซต ถ้าแคนดิเดตไอเท็มเซตนั้นมีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำแสดงว่าแคนดิเดตไอเท็มเซตนั้นเป็นไอเท็มเซตปรากฏบ่อย และจะทำการเช่นนี้ซ้ำไปจนกว่าจะไม่สามารถสร้างแคนดิเดตไอเท็มเซตได้แล้ว

2.1.2.2 การหากฎความสัมพันธ์ของข้อมูล

การวิเคราะห์เพื่อหาความสัมพันธ์ของข้อมูลนั้นสามารถหาความสัมพันธ์ของข้อมูลได้ด้วยการหากฎความสัมพันธ์ของข้อมูล ซึ่งจะเขียนอยู่ในรูปความสัมพันธ์จากไอเท็มเซตหนึ่งไปยังอีกไอเท็มเซต ดังนิยามต่อไปนี้

นิยามที่ 1 กำหนดให้ I เป็นเซตของตัวอักษร หรือตัวเลขที่แสดงรายการทั้งหมดในฐานข้อมูล D และกำหนดให้ X และ Y เป็นไอเท็มเซตนั้นคือ $X \subseteq I$ และ $Y \subseteq I$ และ $X \cap Y = \emptyset$ กฎความสัมพันธ์ของข้อมูลจาก X ไปยัง Y ซึ่งจะหมายถึงความสัมพันธ์ของการปรากฏของไอเท็มทั้งหมดใน X ซึ่งจะมีผลต่อการปรากฏของไอเท็มทั้งหมดใน Y ในรายการเดียวกันในฐานข้อมูล

ในการสร้างกฎความสัมพันธ์จะต้องทำการหาไอเท็มเซตปรากฏบ่อยที่มีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำที่ผู้ใช้กำหนดก่อนจากนั้น นำแต่ละ ไอเท็มเซตปรากฏบ่อยที่หาได้ที่มีขนาดตั้งแต่สองรายการขึ้น ไปมาทำการสร้างกฎความสัมพันธ์ของข้อมูลแต่ในการสร้างกฎแต่ละกฎจะมีความน่าเชื่อถือมากหรือน้อยเพียงใดนั้น จะต้องมีค่าขีดแบ่งที่ใช้วัดค่าความเชื่อมั่นของกฎโดยค่าความเชื่อมั่นของกฎความสัมพันธ์นั้นสามารถนิยามได้ดังนี้

นิยามที่ 2 ค่าความเชื่อมั่น (Confidence) ของกฎความสัมพันธ์จาก $X \rightarrow Y$ หมายถึงอัตราค่าสนับสนุนของการเกิดไอเท็มใน X และ Y พร้อมกันเทียบค่าสนับสนุนของ X ซึ่งคำนวณได้จากสมการที่ 2.1

$$\text{ค่าความเชื่อมั่น } (X \rightarrow Y) = \frac{\text{Support}(X \sqcup Y)}{\text{Support}(X)} \qquad (2.1)$$

นิยามที่ 3 ค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence) เป็นค่าขีดแบ่งที่ในการวัดความเชื่อมั่นของกฎโดยที่ผู้ใช้จะเป็นคนกำหนดค่าซึ่งจะนิยมระบุเป็นร้อยละเพื่อให้กฎที่หาได้มีความ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ถูกต้องสมบูรณ์พร้อมนำไปใช้ในงานต่างๆ ได้ค่าความเชื่อมั่นของกฎนั้นๆ จะต้องมีค่ามากกว่าหรือเท่ากับค่าความเชื่อมั่นขั้นต่ำซึ่งทั้งค่าสนับสนุน และค่าความเชื่อมั่นของกฎสามารถเขียนอยู่ในรูป

$$X \rightarrow Y \text{ [Support = S, Confidence = C]}$$

นั่นหมายถึงกฎความสัมพันธ์ระหว่างเซตของรายการ X กับเซตของรายการ Y มีค่าสนับสนุนเท่ากับ S (เขียนแทนด้วยค่าสนับสนุน support($X \cup Y$)) และมีค่าความเชื่อมั่นเท่ากับ C (เขียนแทนด้วย confidence ($X \rightarrow Y$)) โดยค่าสนับสนุนของกฎสามารถหาได้จากสมการที่ 2.2

$$\text{จาก Support } (X \cup Y) = \frac{\text{จำนวนรายการที่ประกอบด้วยเซตของไอเท็ม X และ Y}}{\text{จำนวนรายการทั้งหมดในฐานข้อมูล}} \quad (2.2)$$

ตัวอย่างกำหนดให้เซตของไอเท็มเซตปรากฏบ่อย {1, 2, 3} สามารถนำมาแตกเป็นกฎความสัมพันธ์ทั้งหมดได้ดังต่อไปนี้

Antecedance		Consequence
{1, 2}	→	{3}
{1, 3}	→	{2}
{2, 3}	→	{1}
{1}	→	{2, 3}
{2}	→	{1, 3}
{3}	→	{1, 2}

จากกฎข้างต้นสามารถสังเกตได้ว่าไอเท็มเซตทางด้านซ้าย (Antecedent) เมื่อนำมารวมกับไอเท็มเซตทางด้านขวา (Consequent) ต้องเท่ากับไอเท็มเซตปรากฏบ่อยที่พิจารณาโดยขนาดของไอเท็มเซตทางด้านซ้าย และด้านขวาสามารถเพิ่ม หรือลดได้แต่เมื่อจำนวนข้อมูลของ ด้านใดด้านหนึ่งเพิ่มจำนวนข้อมูลอีกด้านหนึ่งจะต้องลดลง และไอเท็มเซตทางด้านซ้าย และด้านขวาจะต้องไม่เป็นเซตว่าง จากนั้นนำกฎที่ได้มาหาค่าความเชื่อมั่น แล้วจึงนำมาเปรียบเทียบกับค่าความเชื่อมั่นขั้นต่ำ หากค่าความเชื่อมั่นของกฎมีค่ามากกว่าหรือเท่ากับค่าความเชื่อมั่นขั้นต่ำซึ่งจะทำการคัดเลือกกฎนั้นมาเก็บไว้ใช้งาน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.1.2.2.1 ประโยชน์ของกฎความสัมพันธ์

กฎความสัมพันธ์ที่ได้มานั้นบอกถึงความสัมพันธ์ของสินค้าที่ลูกค้ามักจะซื้อพร้อมกัน ซึ่งจะเป็นประโยชน์ต่อผู้ขายในการวางแผนการตลาด วางแผนการโฆษณาสินค้า ออกแบบการจัดวางสินค้าในร้านและวางแผนกลดราคาสินค้า

2.1.2.2.2 การวัดความน่าสนใจของกฎความสัมพันธ์

กฎความสัมพันธ์ที่ได้นั้นสามารถบอกได้ว่าไอเท็มนั้นมีความสัมพันธ์กัน เช่น {A} →{C} สามารถบอกได้ว่า เมื่อมีไอเท็ม A เกิดขึ้น มักมีไอเท็ม C เกิดตามมา แต่ไม่สามารถบอกได้ว่าความสัมพันธ์ที่เกิดขึ้นนั้นอยู่ในรูปแบบใด เช่น ขึ้นต่อกันเชิงบวก เชิงลบ หรือเป็นอิสระต่อกัน เป็นต้น ซึ่งค่าที่ใช้ในการวัดเพื่อจัดประเภทของกฎความสัมพันธ์ เรียกว่าค่าสหสัมพันธ์ โดยค่าสหสัมพันธ์แบ่งได้ 3 ประเภท ดังนี้

1. สหสัมพันธ์เชิงบวก (Positive Correlation) หมายถึง เมื่อมีไอเท็มหนึ่งเกิดขึ้นก็จะมีอีกไอเท็มหนึ่งเกิดขึ้นด้วย โดยเมื่อคำนวณออกมาแล้วค่าที่ได้ต้องมากกว่า 1
2. สหสัมพันธ์เชิงลบ (Negative Correlation) หมายถึง เมื่อมีไอเท็มหนึ่งเกิดขึ้น อีกไอเท็มหนึ่งจะไม่เกิดขึ้นด้วย โดยเมื่อคำนวณออกมาแล้วค่าที่ได้ต้องน้อยกว่า 1
3. เป็นอิสระต่อกัน (Zero Correlation) หมายถึง ไอเท็มทั้งสองไม่มีความสัมพันธ์ซึ่งกันและกัน โดยเมื่อคำนวณออกมาแล้วค่าที่ได้ต้องเท่ากับ 1

การหาค่าสหสัมพันธ์ของกฎความสัมพันธ์ คำนวณได้ดังสมการที่ 2.3

$$\text{Correlation}(A, B) = \frac{P(A \cup B)}{P(A)P(B)} \qquad (2.3)$$

ตารางที่ 2.1 ข้อมูลในการหาค่าสหสัมพันธ์

X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	0	1	1	1	1	1	1	1

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากข้อมูลในตารางที่ 2.1 สามารถหาค่าสหสัมพันธ์ของไอเท็ม X และไอเท็ม Y ได้ดังนี้

$$\text{Correlation}(X, Y) = \frac{P(X \cup Y)}{P(X)\,P(Y)}$$

$$= \frac{2/8}{(1/2)\,(1/4)}$$

$$= 2$$

เนื่องจากค่าสหสัมพันธ์ที่คำนวณได้มีค่ามากกว่า 1 ดังนั้นไอเท็ม X และไอเท็ม Y มีความสัมพันธ์กันเชิงบวก หมายความว่า เมื่อมีไอเท็ม X เกิดขึ้นก็จะมีไอเท็ม Y เกิดขึ้นด้วย

2.1.3 ความจริงเกี่ยวกับค่าสนับสนุนและค่าความเชื่อมั่น

1. การค้นหากฎความสัมพันธ์อาจล้มเหลวถ้ากำหนดค่าสนับสนุนขั้นต่ำและค่าความเชื่อมั่นขั้นต่ำสูงเกินไป
2. ถ้ากำหนดค่าสนับสนุนขั้นต่ำและค่าความเชื่อมั่นขั้นต่ำ ต่ำเกินไป อาจได้ความสัมพันธ์ระหว่างผลิตภัณฑ์หลากหลายเกินกว่าที่ต้องการ
3. กฎความสัมพันธ์ที่มีค่าสนับสนุนและค่าความเชื่อมั่นสูง แสดงระดับความเกี่ยวข้อง (degree of relevance) มากกว่ากฎความสัมพันธ์ที่มีค่าสนับสนุนและค่าความเชื่อมั่นต่ำ
4. ในบางครั้งรายการที่มีค่าสนับสนุนต่ำอาจเป็นประโยชน์ต่อการค้นหากฎความสัมพันธ์บางอย่าง
5. ค่าความเชื่อมั่นเพียงอย่างเดียวไม่อาจบอกได้ว่าการเกิดร่วมกันของผลิตภัณฑ์ A และ B เกิดขึ้นโดยบังเอิญหรือไม่ ควรจะสนใจกฎความสัมพันธ์ระหว่างผลิตภัณฑ์ที่ไม่ได้เกิดขึ้นโดยความบังเอิญมากกว่า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.2 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้อง เป็นงานวิจัยที่ศึกษาเกี่ยวกับการคำนวณหาไอเท็มเซตปรากฏบ่อย เพื่อปรับปรุงอัลกอริทึมเอไพรออริในการหาไอเท็มเซตปรากฏบ่อย ให้มีจำนวนครั้งในการการอ่านข้อมูลจากฐานข้อมูลน้อยลง สำหรับงานวิจัยนี้ มีงานวิจัยที่เกี่ยวข้องดังนี้

2.2.1 การเพิ่มประสิทธิภาพและปรับปรุงอัลกอริทึมเอไพรออริ

สำหรับงานวิจัยการเพิ่มประสิทธิภาพและปรับปรุงอัลกอริทึมเอไพรออริ ที่พัฒนาโดย Feng WANG และ Yong-hua LI [2] ได้ตระหนักถึงข้อบกพร่องของอัลกอริทึมเอไพรออริอยู่ 3 ประการคือ

1. จากการที่มีรายการเป็นจำนวนมากในฐานข้อมูล ทำให้การคำนวณความถี่ของแคนดิเดตไอเท็มเซตต้องทำการอ่านข้อมูลจากฐานข้อมูลเป็นประจำซึ่งเป็นการเสียเวลา

2. จากการที่แคนดิเดตไอเท็มเซตถูกสร้างจากไอเท็มเซตปรากฏบ่อยสองเซตที่มีการทำการรวม และการตัดออกเสมอ ดังนั้นถ้าสามารถลดกระบวนการทำงานที่เหมือนๆ กันในขั้นตอนดังกล่าวนี้ลงก็จะทำให้อัลกอริทึมเอไพรออริมีประสิทธิภาพที่ดีขึ้น

3. มีการใช้เวลา และพื้นที่ในการเก็บไอเท็มเซตปรากฏบ่อยมากเกินไป

ดังนั้นงานวิจัยนี้จึงเสนอแนวคิดในการลบรายการภายในฐานข้อมูลที่ไม่ได้ใช้ประโยชน์ออกเพื่อลดขนาดของฐานข้อมูล

2.2.1.1 แนวคิด

ในงานวิจัยนี้ได้มีการปรับปรุงอัลกอริทึมเอไพรออริดังนี้

1. ในการหาไอเท็มเซตปรากฏบ่อยที่มีสมาชิก k ไอเท็ม ขณะที่ทำการอ่านข้อมูลจากฐานข้อมูลเพื่อหาค่าสนับสนุนของแคนดิเดตไอเท็มเซตนั้นก็จะทำเครื่องหมาย (mark deleting tag) ในรายการที่มีขนาดของไอเท็มน้อยกว่าค่า k

2. ถ้าแคนดิเดตไอเท็มเซตไม่ได้ปรากฏอยู่ในรายการ t ก็จะทำเครื่องหมายไว้ที่รายการนั้น

ดังนั้นในการอ่านข้อมูลครั้งถัดไปก็จะข้ามรายการที่ได้ทำเครื่องหมายไว้ การลดขนาดของฐานข้อมูล หลังจากมีการสร้างแคนดิเดตไอเท็มเซตที่มีสมาชิก k+1 ไอเท็ม จากไอเท็มเซตปรากฏบ่อยที่มีสมาชิก k ไอเท็มแล้ว ถ้าขนาดของรายการมีค่าน้อยกว่าค่า k ก็อาจกล่าวได้ว่ารายการนั้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ไม่มีประโยชน์ในการนำมาสร้างแคนดิเดตไอเท็มเซตที่มีสมาชิก k+1 ไอเท็ม และแคนดิเดตไอเท็มเซตไม่ได้ปรากฏอยู่ในรายการนั้น ก็จะทำเครื่องหมายไว้เพื่อไม่ให้มีการอ่านข้อมูลรายการนี้ในรอบถัดไป

2.2.1.2 ขั้นตอนการทำงาน

1. อ่านข้อมูลจากฐานข้อมูลเพื่อหาแคนดิเดตไอเท็มเซตที่มีสมาชิก 1 ไอเท็ม โดยกำหนดค่าสนับสนุนขั้นต่ำที่เหมาะสม เพื่อนำไปเปรียบเทียบกับค่าสนับสนุนของแต่ละแคนดิเดตไอเท็มเซต เพื่อหาไอเท็มเซตปรากฏบ่อยที่มีสมาชิก 1 ไอเท็ม

2. สร้างแคนดิเดตไอเท็มเซตที่มีสมาชิก k+1 ไอเท็ม จากการรวมของไอเท็มเซตปรากฏบ่อย 2 ชุด โดยไอเท็มที่ m ของไอเท็มเซตปรากฏบ่อยชุดแรกรวมกับไอเท็มที่ m+1 ของไอเท็มเซตปรากฏบ่อยในชุดที่สอง

3. ถ้าขนาดของรายการน้อยกว่าค่า k ก็จะทำเครื่องหมายไว้ที่รายการนั้นในกรณีไม่น้อยกว่าก็จะตรวจสอบต่อว่าแคนดิเดตไอเท็มเซตปรากฏอยู่ในรายการ t หรือไม่

4. ตรวจสอบแคนดิเดตไอเท็มเซตในรายการ ถ้าแคนดิเดตไอเท็มเซตไม่ได้ปรากฏอยู่ในรายการก็จะทำเครื่องหมายไว้ที่รายการ แล้วไปที่รายการถัดไป

5. ถ้าค่าสนับสนุนของแคนดิเดตไอเท็มเซตมีค่ามากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำก็จะเพิ่มแคนดิเดตไอเท็มเซตนั้น ลงในไอเท็มเซตปรากฏบ่อย

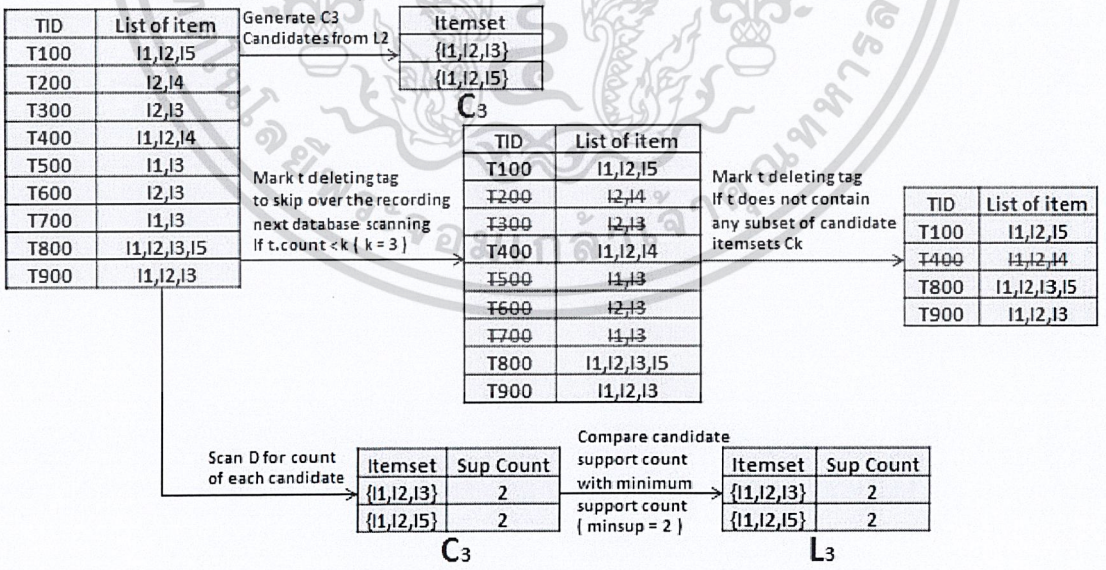
6. ถ้าไอเท็มเซตปรากฏบ่อยที่มีสมาชิก k-1 ไอเท็มเป็นเซตว่างก็จะจบการทำงาน หรือถ้าไอเท็มเซตปรากฏบ่อยที่มีสมาชิก k-1 ไอเท็มไม่เป็นเซตว่างก็จะเพิ่มค่า k ขึ้นทีละ 1 และกระทำคำสั่งต่อๆ ไป จนไอเท็มเซตปรากฏบ่อยที่มีสมาชิก k-1 ไอเท็มจะเป็นเซตว่างก็จะจบการทำงาน

ซึ่งสามารถแสดงกระบวนการในการทำงานของอัลกอริทึมเอไพรออริในรูปแบบของโปรแกรม ดังรูปที่ 2.6

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
1)      L1 = {large 1-itemsets};
2)      for (k = 2; Lk-1 ≠ ∅; k++)
3)      {
4)      Ck = apriori-gen(Lk-1); //generate new candidate itemsets
5)      for all transactions t ∈ D and t.delete = 0
6)          {
7)              if  t.count< k then //    if the size of transaction t is less than k, t is
                                          useless for Ck  generated
8)              t.delete = 1 //      mark t the deleting tag to skip over the recording
                                     next database scanning
9)      else
10)             {
11)               Ct = subset(Ck, t); //candidate itemsets contained in transaction t
12)             if Ct = ∅then // if t does not contain any subset of candidate
                                     itemsetsCk, mark the deleting tag
13)     t.delete = 1;
14)     else
15)             {
16)     for all candidates c ∈ Ct
17)     c.count++;
18)             }
19)         }
20)       }
21)     Lk = {c ∈ Ck | c.count≥minsup}
22)     }
23)     Answer = ∪kLk;
```

รูปที่ 2.6 อัลกอริทึมเอไพรออริที่มีการปรับปรุง



รูปที่ 2.7 ตัวอย่างการดำเนินการตามอัลกอริทึมเอไพรออริ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูปที่ 2.7 หลังจากที่ได้ทำการสร้าง C_3 (3–Candidate Itemset) จากการ join(L_3*L_3)แล้ว ในอัลกอริทึมเอไพรออริจะมีการตรวจสอบแต่ละรายการ (Transaction) ภายในฐานข้อมูลด้วยการทำเครื่องหมายลงในรายการ (mark deleting tag) ตามเงื่อนไขที่กำหนดซึ่งในการอ่านข้อมูลจากฐานข้อมูลในรอบถัดไปก็จะข้ามรายการที่ถูกทำเครื่องหมายไว้ สำหรับเงื่อนไขในการตรวจสอบได้แก่

1. ทำการตรวจสอบขนาดของแต่ละรายการด้วยการเปรียบเทียบกับค่า k ซึ่งเป็นจำนวนสมาชิกในแต่ละ itemset ถ้าขนาดของรายการมีค่าน้อยกว่า k รายการนั้นๆ ก็จะถูกทำเครื่องหมายไว้
2. ทำการตรวจสอบในแต่ละรายการว่ามีแคนดิเดตไอเท็มเซตปรากฏอยู่ในรายการหรือไม่ ถ้าแคนดิเดตไอเท็มเซตไม่ได้ปรากฏอยู่ในรายการ รายการนั้นๆ ก็จะถูกทำเครื่องหมายไว้

จากในรูปที่ 2.7 จะทำการตรวจสอบทีละรายการตามเงื่อนไขที่กำหนด อย่างเช่น

- T100 ประกอบด้วยสมาชิกในรายการคือ {I1, I2, I5} ซึ่งมีขนาดของไอเท็มเป็น 3 จะถูกนำมาเปรียบเทียบกับค่า k ซึ่งมีค่าเป็น 3 เท่ากัน ดังนั้นจึงผ่านเกณฑ์ในของเงื่อนไขที่ถูกกำหนดไว้ในการตรวจสอบชนิดแรก จากนั้น T100 จะถูกตรวจสอบตามเงื่อนไขถัดมา พบว่าสมาชิกของ T100 ปรากฏอยู่ในแคนดิเดตไอเท็มเซตจึงผ่านเงื่อนไขในการตรวจสอบจากนั้นจึงไปทำคำสั่งในขั้นตอนการนับค่าสนับสนุนของแคนดิเดตไอเท็มเซตต่อไป
- T200 ประกอบด้วยสมาชิกในรายการคือ {I2, I4} ซึ่งมีขนาดของไอเท็มเป็น 2 จะถูกนำมาเปรียบเทียบกับค่า k ซึ่งมีค่าเป็น 3 พบว่ามีขนาดน้อยกว่าค่า k ดังนั้น T200 จะถูกทำเครื่องหมายไว้ และทำการตรวจสอบรายการถัดไปทันที ซึ่งในการอ่านข้อมูลจากฐานข้อมูลในรอบถัดไปก็จะข้ามรายการนี้ไป
- T400 ประกอบด้วยสมาชิกในรายการคือ {I1, I2, I4} ซึ่งมีขนาดของไอเท็มเป็น 3 จะถูกนำมาเปรียบเทียบกับค่า k ซึ่งมีค่าเป็น 3 เท่ากัน ดังนั้นจึงผ่านเกณฑ์ของเงื่อนไขที่ถูกกำหนดไว้ในการตรวจสอบข้อแรก จากนั้น T400 จะถูกตรวจสอบตามเงื่อนไขถัดมา พบว่าสมาชิกของ T400 ไม่ปรากฏอยู่ในแคนดิเดตไอเท็มเซต ดังนั้น T400 จะถูกทำเครื่องหมายไว้ และทำการตรวจสอบรายการถัดไปทันที ซึ่งในการอ่านข้อมูลจากฐานข้อมูลในรอบถัดไปก็จะข้ามรายการนี้ไป เป็นต้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.2.2 การปรับปรุงประสิทธิภาพอัลกอริทึมเอไพรออริด้วยเมทริกซ์

สำหรับงานวิจัยการปรับปรุงประสิทธิภาพอัลกอริทึมเอไพรออริด้วยเมทริกซ์ ที่พัฒนาโดย Yiwu Xie และ Yutong Li และ Chunli Wang และ Mingyu Lu [3] มีวัตถุประสงค์ในการแก้ปัญหาของอัลกอริทึมเอไพรออริ ในเรื่องของการลดจำนวนครั้งในการอ่านข้อมูลจากฐานข้อมูล และลดจำนวนแคนดิเดตไอเท็มเซต โดยอัลกอริทึมนี้ใช้เมทริกซ์แทนแต่รายการในฐานข้อมูล ดังนั้นจึงใช้เมทริกซ์ในการนับค่าสนับสนุนของแคนดิเดตไอเท็มเซตด้วย และไม่ต้องมีการอ่านข้อมูลจากฐานข้อมูลหลายครั้ง ทำให้มีการนับค่าสนับสนุนที่รวดเร็ว โดยการ AND ข้อมูลที่สนใจในเมทริกซ์

2.2.2.1 แนวคิด

สำหรับแนวคิดนี้มีขั้นตอนการทำงานหลักๆ อยู่ 3 ขั้นตอนคือ

1. แปลงข้อมูลให้อยู่ในรูปของเมทริกซ์

เปลี่ยนฐานข้อมูลให้อยู่ในรูปของเมทริกซ์ที่มี I ไอเท็ม (แถว) และ t รายการ (คอลัมน์) และนับค่าสนับสนุนของแต่ละไอเท็ม ถ้าไอเท็มใดมีค่าสนับสนุนน้อยกว่าค่าสนับสนุนขั้นต่ำ ก็แสดงว่าไม่ใช่ไอเท็มเซตปรากฏบ่อย จึงลบแถวนั้นออกได้เลยซึ่งจะง่ายต่อการดำเนินการในขั้นตอนต่อๆ ไป

2. หาไอเท็มเซตปรากฏบ่อยที่ใหญ่ที่สุด

ไอเท็มเซตปรากฏบ่อยที่ใหญ่ที่สุดคือ ไอเท็มเซตปรากฏบ่อยที่มีสมาชิกในเซตจำนวนมากที่สุดหาไอเท็มเซตปรากฏบ่อยที่ใหญ่ที่สุดเพื่อที่จะได้หาไอเท็มเซตปรากฏบ่อยที่เล็กกว่าได้ง่าย โดยมีขั้นตอนในการหาไอเท็มเซตปรากฏบ่อยที่ใหญ่ที่สุดดังต่อไปนี้

(a) หารายการที่มีจำนวนไอเท็มมากที่สุด แล้วทำเครื่องหมายไว้ว่าที่มากที่สุดนั้นเป็น k ถ้าจำนวนไอเท็มที่อยู่ในรายการ มีไม่น้อยกว่าค่า k แต่น้อยกว่าค่าสนับสนุนขั้นต่ำ นั่นหมายความว่าไอเท็มเซตปรากฏบ่อยที่ใหญ่ที่สุดไม่ได้มีสมาชิก k ตัว ดังนั้นมาพิจารณารายการที่มีสมาชิก k – 1 ตัว แล้วก็ทำเครื่องหมายไว้ว่าตอนนี้ k เท่ากับk – 1 แล้ว ใช้ขั้นตอนวิธีเดิมจนกระทั่งได้ k ที่มีขนาดใหญ่ที่สุดที่มีค่าสนับสนุนไม่น้อยกว่าค่าสนับสนุนขั้นต่ำ

(b) สร้างเมทริกซ์ขึ้นมาใหม่โดยเมทริกซ์นี้มีเฉพาะรายการที่มีจำนวน ไอเท็มไม่น้อยกว่า k แล้วพิจารณาทีละรายการในเมทริกซ์ใหม่ โดยเลือก k ไอเท็มในรูปของไอเท็มเซตแล้วใช้การ AND เพื่อนับค่าสนับสนุนของไอเท็มเซตที่ได้เลือกไว้ถ้าค่าสนับสนุนไม่น้อยกว่าค่าสนับสนุนขั้นต่ำ แสดงว่าไอเท็มเซตนั้นเป็นไอเท็มเซตปรากฏบ่อย และถ้าไอเท็มเซตนั้นมีค่าสนับสนุนน้อยกว่าค่าสนับสนุนขั้นต่ำแสดงว่าไม่ใช่ไอเท็มเซตปรากฏบ่อย หลังจากนั้นเลือก k ไอเท็มใน t ที่ไม่เหมือนกันกับชุดเดิม แล้วหาค่าสนับสนุนอีกใช้วิธีนี้จนกว่าจะหาเซตที่มีสมาชิก k ไอเท็มที่แตกต่าง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งไม่อนุญาตให้นำไปดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารนี้ทุกครั้งที่มีการนำไปใช้

สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง

กันทั้งหมดใน t (รายการที่กำลังหาอยู่) แล้วพิจารณาที่รายการต่อไป เมื่อพิจารณาหมดทุกรายการ ก็จะได้ไอเท็มเซตปรากฏบ่อยที่ใหญ่ที่สุด

3. หาไอเท็มเซตปรากฏบ่อยตัวอื่นๆ

หาไอเท็มเซตปรากฏบ่อยทั้งหมด ตั้งแต่ไอเท็มเซตปรากฏบ่อยที่มีสมาชิก 2 ไอเท็ม จนถึงไอเท็มเซตปรากฏบ่อยที่มีสมาชิก k-1 ไอเท็ม ด้วยอัลกอริทึมเอไพรออริ แต่จะนับค่าสนับสนุนได้ง่ายขึ้น โดยใช้เมทริกซ์ที่สร้างไว้ และเนื่องจากรู้ไอเท็มเซตปรากฏบ่อยที่ใหญ่ที่สุด ก็จะได้ไอเท็มเซตปรากฏบ่อยที่มีสมาชิกตั้งแต่ 1 ถึง k - 1 ไอเท็มนั่นคือสามารถลดจำนวนแคนดิเดตไอเท็มเซตได้ และช่วยลดเวลาในการนับค่าสนับสนุน เชื่อมต่อฐานข้อมูล และการตัดออกซึ่งเป็นข้อเสียของอัลกอริทึมเอไพรออริซึ่งสามารถแสดงกระบวนการในการทำงานของอัลกอริทึมเอไพรออริที่มีการปรับปรุงด้วยเมทริกซ์ในรูปแบบของโปรแกรม ดังรูปที่ 2.8

```
1)   Mat ← Create (D, Minsupport);
2)   If Mat = NULL
3)         End, no frequent itemsets
4)   MaxK ← Max (Mat, 0)   [MaxK is the number of items in
                            an affair which contain the most items]
5)   While MaxK > 0,   repeat the following code
6)         NewMat  ←  SimplifyMat (Mat,MaxK,Minsupport) [simplify the Mat]
7)   If NewMat ≠ NULL
8)         L_MaxK ← Find_Max(NewMat, MaxK) [find out the MaxK-frequent itemsets]
9)               If L_MaxK ≠ NULL [if find out the MaxK-frequent itemsets, then break]
10)                    Break
11)  MaxK ← Max (Mat, MaxK)
12)        If L_MaxK = NULL
13)              End no frequent itemsets
14)  L1 ← Find_1 (Mat) [the collection of items in the Mat is 1-frequent itemsets]
15)  While K from 2 to MaxK,   execute
16)  Mat ← SimplifyMat(Mat, K, Minsupport) [simplify the Mat]
17)  Ck ← GenC( Mat , L_MaxK , L_K-1 )      [find out the candidates of frequent itemsets]
18)  LK ← GenL( Mat , Ck , Minsupport ) [find out frequent itemsets]
19)  Answer ← L1UL2U...ULKUL_MaxKUL [L is subset of LMaxK]
20)  END
```

รูปที่ 2.8 อัลกอริทึมเอไพรออริแบบเมทริกซ์ในรูปแบบของโปรแกรม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้
117243

2.2.2.2 ตัวอย่างการดำเนินการตามอัลกอริทึม

ในตารางที่ 2.2 แสดงข้อมูล และกำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 50% ซึ่งเท่ากับ 2 รายการจากทั้งหมด 4 รายการ

ตารางที่ 2.2 ข้อมูลรายการซื้อสินค้า

TID	Itemsets
T100	1, 3, 4
T200	2, 3, 5
T300	1, 2, 3, 5
T400	2, 5

ขั้นตอนที่ 1 แปลงข้อมูลให้อยู่ในรูปของเมทริกซ์ และตัดไอเท็ม 4 ที่มีค่าสนับสนุนน้อยกว่าค่าสนับสนุนขั้นต่ำเรียกเมทริกซ์นี้ว่า เมทริกซ์ตั้งต้น

ตารางที่ 2.3 เมทริกซ์ตั้งต้น

	T100	T200	T300	T400
1	1	0	1	0
2	0	1	1	1
3	1	1	1	0
5	0	1	1	1

จากขั้นตอนนี้ทำให้ทราบด้วยว่า ไอเท็มเซตปรากฏบ่อยที่มีสมาชิก 1 ไอเท็ม คือ {1}, {2}, {3} และ {5}

ขั้นตอนที่ 2 หาไอเท็มเซตปรากฏบ่อยที่ใหญ่ที่สุด

เนื่องจากจำนวนไอเท็มมากที่สุดที่เกิดขึ้นคือ 4 แต่มีแค่ T300 เพียงรายการเดียว ที่มี 4 ไอเท็ม ดังนั้นรายการ T300 ที่ประกอบไปด้วย 4 ไอเท็มนี้ จึงมีค่าสนับสนุนน้อยกว่าค่า สนับสนุนขั้นต่ำจึงพิจารณารายการที่มีจำนวนไอเท็มตั้งแต่ 3 ไอเท็ม คือ T200 และ T300

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 2.4 เมทริกซ์ที่ใช้ในการหาไอเท็มเซตปรากฏบ่อย

	T200	**T300**
1	0	1
2	1	1
3	1	1
5	1	1

ตารางที่ 2.5 การ AND กันของ 2 3 และ 5

2	1	1
3	1	1
5	1	1
AND	1	1

จากตารางที่ 2.4 ในรายการที่ T200 ประกอบด้วย 3 ไอเท็ม {2, 3, 5} เมื่อทำการ AND กัน ดังตารางที่ 2.5 แล้วให้นำผลลัพธ์ที่ได้จากการ AND มาบวกกัน โดยผลลัพธ์ที่ได้จากการบวกมีค่าเท่ากับ 2 ซึ่งไม่น้อยกว่าค่าสนับสนุนขั้นต่ำ ดังนั้นไอเท็มเซต {2, 3, 5} จึงเป็น ไอเท็มเซตปรากฏบ่อยที่ใหญ่ที่สุด และเนื่องจาก T200 มีเพียงเซตเดียวที่ให้ไอเท็มเซตที่ประกอบด้วยสมาชิก 3 ไอเท็ม ดังนั้นจึงพิจารณาที่รายการถัดไป ในทำนองเดียวกันนี้ เลือกไอเท็มเซตของรายการถัดไปที่มีสมาชิก 3 ไอเท็มที่ไม่ซ้ำกัน แล้วทำการ AND เพื่อเปรียบเทียบกับค่าสนับสนุนขั้นต่ำ แล้วดูว่าผลบวกของผลลัพธ์จากการ AND มีค่ามากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ แสดงว่าไอเท็มเซตนั้นเป็นไอเท็มเซตปรากฏบ่อยที่ใหญ่ที่สุด แต่ถ้าน้อยกว่าก็แสดงว่าไม่เป็นไอเท็มเซตปรากฏบ่อยที่ใหญ่ที่สุด

จากการหาไอเท็มเซตปรากฏบ่อยที่ใหญ่ที่สุดในขั้นตอนนี้ จะได้ไอเท็มเซตปรากฏบ่อยที่ใหญ่ที่สุด คือ {2, 3, 5}

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ขั้นตอนที่ 3 หาไอเท็มเซตปรากฏบ่อยตัวอื่นๆ

เนื่องจาก {2, 3, 5} เป็นไอเท็มเซตปรากฏบ่อยที่ใหญ่ที่สุด ดังนั้นจะได้ไอเท็มเซตปรากฏบ่อยที่มีสมาชิก 2 ไอเท็ม คือ {2, 3}, {2, 5} และ {3, 5}

จากขั้นตอนที่ 1 ไอเท็มเซตปรากฏบ่อยที่มีสมาชิก 1 ไอเท็ม คือ {1}, {2}, {3} และ {5} เมื่อทำการรวม และจะได้แคนดิเดตไอเท็มเซตที่มีสมาชิก 2 ไอเท็ม คือ {1, 2}, {1, 3}, {1, 5}, {2, 3}, {2, 5} และ {3, 5} จะเห็นว่า ไอเท็มเซตปรากฏบ่อยที่มีสมาชิก 2 ไอเท็ม คือ {2, 3}, {2, 5} และ {3, 5} เป็นไอเท็มเซตปรากฏบ่อยแน่ๆ (จากการหาไอเท็มเซตปรากฏบ่อยที่ใหญ่ที่สุด) ดังนั้นต้องทำการหาว่า {1, 2}, {1, 3} และ {1,5} เป็นไอเท็มเซตปรากฏบ่อยหรือไม่ โดยการนำข้อมูลจากเมทริกซ์ตั้งต้นมา AND กันแล้วเปรียบเทียบกับค่าสนับสนุนขั้นต่ำ ถ้าผลบวกของผลลัพธ์ที่ได้จากการ AND มีค่ามากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ แสดงว่าไอเท็มเซตนั้นเป็นไอเท็มเซตปรากฏบ่อย แต่ถ้าน้อยกว่าก็แสดงว่าไม่เป็นไอเท็มเซตปรากฏบ่อย

จะได้ไอเท็มเซตปรากฏบ่อยที่มีสมาชิก 2 ไอเท็ม คือ {1, 3}, {2, 3}, {2, 5} และ {3, 5}

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3

การพัฒนาอัลกอริทึมค้นหาไอเท็มเซตปรากฏบ่อย

อัลกอริทึมในการค้นหาไอเท็มเซตปรากฏบ่อยที่ได้พัฒนาขึ้นมานั้น มีการนำเมทริกซ์และหลักการทางตรรกะคณิตศาสตร์ คือการ AND มาช่วยในการทำงาน มีการตัดไอเท็มที่มีค่าสนับสนุนน้อยกว่าค่าสนับสนุนขั้นต่ำออก ก่อนการคำนวณหาไอเท็มเซตปรากฏบ่อย เนื่องจากไอเท็มเซตปรากฏบ่อยต้องมีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ ดังนั้นจึงไม่จำเป็นต้องพิจารณาไอเท็มที่มีค่าสนับสนุนน้อยกว่าค่าสนับสนุนขั้นต่ำ

3.1 หลักการทำงาน

ขั้นตอนการทำงานในการค้นหาไอเท็มเซตปรากฏบ่อยของอัลกอริทึมที่ได้พัฒนาขึ้น มี 2 ขั้นตอนหลักๆ คือ ขั้นตอนในการแปลงข้อมูลให้อยู่ในรูปของเมทริกซ์ และขั้นตอนในการหาไอเท็มเซตปรากฏบ่อย ซึ่งแต่ละขั้นตอนมีการทำงานดังต่อไปนี้

ขั้นตอนที่ 1 แปลงข้อมูลให้อยู่ในรูปของเมทริกซ์

แปลงข้อมูลให้อยู่ในรูปของเมทริกซ์ A_{ij} ให้ชื่อว่า matrixT โดยที่ i แทนไอเท็ม (แถว) และ j แทนรายการ (คอลัมน์) ซึ่งเมทริกซ์นี้มีขนาด m x n เมื่อ m คือ จำนวนไอเท็มและ n คือ จำนวนรายการ จากนั้นจะพิจารณาว่าแต่ละรายการมีไอเท็มใดปรากฏอยู่บ้าง โดยกำหนดให้

$$a_{ij} = \begin{cases} 1 & \text{หมายความว่า มีไอเท็ม i ปรากฏอยู่ในรายการ j} \\ 0 & \text{หมายความว่า ไม่มีไอเท็ม i ปรากฏอยู่ในรายการ j} \end{cases}$$

ซึ่งลักษณะของเมทริกซ์แสดงดังรูปที่ 3.1

$$A = A_{m\times n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdots & a_{2n} \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & \cdots & a_{mn} \end{bmatrix}$$

รูปที่ 3.1 ลักษณะของเมทริกซ์ขนาด m x n

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สำหรับในขั้นตอนนี้จะทำการนับค่าสนับสนุนของแต่ละไอเท็ม ($count_{Items}$) ด้วย โดยนับจำนวนของ "1" ที่ปรากฏอยู่ในไอเท็มนั้นๆ จากนั้นทำการตัดแถวของไอเท็มที่มีค่าสนับสนุนน้อยกว่าค่าสนับสนุนขั้นต่ำออกไปจากเมทริกซ์ เนื่องจากไอเท็มเหล่านั้นไม่จำเป็นต้องนำมาคำนวณอีก แล้วจะได้ไอเท็มเซตปรากฏบ่อยที่มีสมาชิก 1 ไอเท็มจากขั้นตอนนี้ด้วย

ขั้นตอนที่ 2 หาไอเท็มเซตปรากฏบ่อย

เมื่อได้เมทริกซ์ที่ทำการตัดไอเท็มที่มีค่าสนับสนุนน้อยกว่าค่าสนับสนุนขั้นต่ำแล้ว ให้พิจารณาทีละรายการ โดยพิจารณาว่าในรายการนั้นมีไอเท็มใดปรากฏอยู่บ้าง นำข้อมูลของไอเท็มที่ปรากฏอยู่มาทำการ AND กันในแนวคอลัมน์ เพื่อหาการเกิดร่วมกันของไอเท็มเหล่านั้น แล้วนำผลลัพธ์ที่ได้จากการ AND มาบวก จากนั้นนำผลลัพธ์จากการบวกมาเปรียบเทียบกับค่าสนับสนุนขั้นต่ำ ถ้าผลลัพธ์ที่ได้จากการบวก มีค่ามากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำแสดงว่า เซตของไอเท็มนั้นเป็นไอเท็มเซตปรากฏบ่อย และในทางกลับกันถ้าผลลัพธ์ที่ได้จากการบวก มีค่าน้อยกว่าค่าสนับสนุนขั้นต่ำ แสดงว่าเซตของไอเท็มนั้นไม่เป็นไอเท็มเซตปรากฏบ่อย จึงต้องพิจารณาเซตย่อยของไอเท็มเซตนั้นด้วย เมื่อพิจารณารายการใดแล้ว ให้ทำการลบรายการนั้น (คอลัมน์) ออกจากเมทริกซ์ จากคุณสมบัติของอัลกอริทึมเอไพรออริที่กล่าวว่า "ทุกๆ เซตย่อย (Subset) ของไอเท็มเซตปรากฏบ่อย (Frequent Itemset) ต้องเป็นไอเท็มเซตปรากฏบ่อยด้วย" นั่นคือเมื่อได้ไอเท็มเซตปรากฏบ่อยที่ได้จากการหาในแต่ละรายการแล้ว ก็จะได้ว่าทุกๆ เซตย่อยของไอเท็มเซตปรากฏบ่อยที่ได้ ก็จะเป็นไอเท็มเซตปรากฏบ่อยด้วย

อัลกอริทึมที่ได้พัฒนาขึ้นมาเพื่อใช้ในการค้นหาไอเท็มเซตปรากฏบ่อยนั้น สามารถแสดงเป็นอัลกอริทึมได้ดังนี้ รูปที่ 3.2

```
Input  : - Data, D;
         - Minimum Support, min_sup.
Output :  All frequent itemsets from L1 to L4, Lk when k = 1 to 4.

(1)    generate matrixT from D
(2)    count support of each item (count_items)
(3)    if support of each item(count_items) less than min_sup
(4)          delete row of that item in matrixT
```

รูปที่ 3.2(ก) ขั้นตอนการทำงานของอัลกอริทึมในการค้นหาไอเท็มเซตปรากฏบ่อย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
        [Find all frequent itemsets]
(5)     For each transaction
(6)             Generate the  candidate itemset (Ck)
(7)             If the candidate itemset Ck never happened before
(8)                     count support of the candidate itemset Ck
(9)                     If support >= min_sup
(10)                            Add the itemset Ck and its subsets into Lk
(11)                            mark the itemset Ck and its subset are considered
(12)                    If support < min_sup
(13)                            Find all subsets of the itemset Ck
(14)                               For each  subsets
(15)                                    repeat in line 7 to 16 until consider all subsets
(16)    Answer  = ∪K Lk
```

รูปที่ 3.2(ข) ขั้นตอนการทำงานของอัลกอริทึมในการค้นหาไอเท็มเซตปรากฏบ่อย

3.2 ตัวอย่างการทำงาน

ในการแสดงตัวอย่างการทำงาน ได้มีการยกตัวอย่างฐานข้อมูลดังตารางที่ 3.1 ซึ่งประกอบไปด้วยรายการทั้งหมด 9 รายการ ไอเท็มที่แตกต่างกัน 5 ไอเท็ม และกำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 25% ซึ่งคิดเป็น 2 รายการ จากทั้งหมด 9 รายการ แล้วนำข้อมูลนี้ไปดำเนินการตามขั้นตอนต่อไป

ตารางที่ 3.1 ฐานข้อมูล

Tid	Items
T100	I1 , I2 , I5
T200	I2 , I4
T300	I2 , I3
T400	I1 , I2 , I4
T500	I1 , I3
T600	I2 , I3
T700	I1 , I3
T800	I1 , I2 , I3 , I5
T900	I1 , I2 , I3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ขั้นตอนที่ 1 แปลงข้อมูลให้อยู่ในรูปของเมทริกซ์

จากข้อมูลในตารางที่ 3.1 นำข้อมูลนั้นมาแปลงให้อยู่ในรูปของเมทริกซ์ (A_{ij}) โดยแทนตำแหน่งที่มีข้อมูลเป็น "1" และตำแหน่งที่ไม่มีข้อมูลเป็น "0"

ตารางที่ 3.2 ข้อมูลที่อยู่ในรูปของเมทริกซ์

Items \ Tid	T100	T200	T300	T400	T500	T600	T700	T800	T900	$count_{Items}$
I1	1	0	0	1	1	0	1	1	1	6
I2	1	1	1	1	0	1	0	1	1	7
I3	0	0	1	0	1	1	1	1	1	6
I4	0	1	0	1	0	0	0	0	0	2
I5	1	0	0	0	0	0	0	1	0	2
SUP_{item}	3	2	2	3	2	2	2	4	3	

จากตารางที่ 3.2 จะเห็นว่าในรายการ T100 มีไอเท็ม I1, I2 และ I5 ปรากฏอยู่ จึงแทนตำแหน่งของ I1, I2 และ I5 ใน T100 ด้วย "1" ตามลำดับ และในตำแหน่งของ I3 และ I4 ด้วย "0" แล้วดำเนินการในรายการอื่นๆ ด้วยวิธีเดียวกัน

หลังจากที่แปลงข้อมูลให้อยู่ในรูปของเมทริกซ์เรียบร้อยแล้วให้ทำการหาค่าสนับสนุนของแต่ละไอเท็ม ($count_{items}$) โดยนับจำนวนรายการที่มีไอเท็มนั้นปรากฏอยู่ และผลรวมของไอเท็มในรายการ (SUP_{item}) โดยนับจำนวนไอเท็มที่ปรากฏในรายการนั้น แล้วทำการเปรียบเทียบค่าสนับสนุนของแต่ละไอเท็มกับค่าสนับสนุนขั้นต่ำ ถ้าค่าสนับสนุนของไอเท็มใด มีค่าน้อยกว่า ค่าสนับสนุนขั้นต่ำ ให้ทำการลบแถวนั้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ขั้นตอนที่ 2 หาไอเท็มเซตปรากฏบ่อย

จากตารางที่ 3.2 จะทำการพิจารณาทีละรายการ โดยการนำไอเท็มที่ปรากฏในเมทริกซ์ในรายการที่กำลังพิจารณาอยู่ มาทำการ AND กันในแนวคอลัมน์ หลังจากนั้นทำการหาผลบวกของผลลัพธ์จากการ AND แล้วเปรียบเทียบผลบวกของผลลัพธ์จากการ AND กับค่าสนับสนุนขั้นต่ำ ดังนี้

<u>พิจารณา T100</u> จะเห็นว่าใน T100 มีไอเท็ม I1 I2 และ I5 ปรากฏอยู่จึงนำข้อมูล I1 I2 และ I5 ในเมทริกซ์มาทำการ AND กันในแนวคอลัมน์ และเมื่อทำการ AND แล้วให้นำผลของการ AND มาบวกกันในทุกรายการ แล้วจึงไปพิจารณาในรายการถัดไป พร้อมทั้งลบคอลัมน์ของรายการ T100 ออก

ตารางที่ 3.3 การ AND กันของ I1 I2 และ I5

I1	1	0	0	1	1	0	1	1	1
I2	1	1	1	1	0	1	0	1	1
I5	1	0	0	0	0	0	0	1	0
AND	1	0	0	0	0	0	0	1	0

จากตารางที่ 3.3 เมื่อทำการ AND กันระหว่าง I1 I2 และ I5 แล้วให้นำผลที่ได้จากการ AND มาบวกกันในทุกรายการ ในที่นี้ได้เท่ากับ 2 ซึ่งมีค่าไม่น้อยกว่าค่าสนับสนุนขั้นต่ำ ดังนั้น {I1, I2, I5} จึงเป็นไอเท็มเซตปรากฏบ่อยและจากคุณสมบัติของอัลกอริทึมเอไพรออริ คือ ทุกๆ เซตย่อย (subset) ของไอเท็มเซตปรากฏบ่อยต้องเป็นไอเท็มเซตปรากฏบ่อย ก็จะได้ว่า {I1, I2}, {I1, I5} และ {I2, I5} เป็น ไอเท็มเซตปรากฏบ่อยด้วย ดังนั้น L_k = {{I1, I2}, {I1, I5}, {I2, I5}, {I1, I2, I5}}

ตารางที่ 3.4 เมทริกซ์ที่ได้หลังจากลบ T100

Tid / Items	**T200**	**T300**	**T400**	**T500**	**T600**	**T700**	**T800**	**T900**
I1	0	0	1	1	0	1	1	1
I2	1	1	1	0	1	0	1	1
I3	0	1	0	1	1	1	1	1
I4	1	0	1	0	0	0	0	0
I5	0	0	0	0	0	0	1	0

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากตารางที่ 3.4 หลังจากที่ลบรายการ T100 ที่พิจารณาแล้วออกจากเมทริกซ์ ทำให้ทราบว่ารายการที่ต้องพิจารณาถัดไป คือ T200

พิจารณา T200 จะเห็นว่าใน T200 มีไอเท็ม I2 และ I4 ปรากฏอยู่จึงนำข้อมูลของ I2 และ I4 ในเมทริกซ์มาทำการ AND กันในแนวคอลัมน์ และเมื่อทำการ AND แล้วให้นำผลของการ AND มาบวกกันในทุกรายการ แล้วจึงไปพิจารณาในรายการถัดไป พร้อมทั้งลบคอลัมน์ของรายการ T200 ออก

ตารางที่ 3.5 การ AND กันของ I2 และ I4

I2	1	1	1	0	1	0	1	1
I4	1	0	1	0	0	0	0	0
AND	1	0	1	0	0	0	0	0

จากตารางที่ 3.5 เมื่อทำการ AND กันระหว่าง I2 และ I4 แล้วให้นำผลที่ได้จากการ AND มาบวกกันในทุกรายการ ในที่นี้ได้เท่ากับ 2 ซึ่งมีค่าไม่น้อยกว่าค่าสนับสนุนขั้นต่ำ {I2, I4} จึงเป็นไอเท็มเซตปรากฏบ่อย ดังนั้น L_k = {{I1, I2}, {I1, I5}, {I2, I4}, {I2, I5}, {I1, I2, I5}}

ตารางที่ 3.6 เมทริกซ์ที่ได้หลังจากลบ T200

Tid / Items	T300	T400	T500	T600	T700	T800	T900
I1	0	1	1	0	1	1	1
I2	1	1	0	1	0	1	1
I3	1	0	1	1	1	1	1
I4	0	1	0	0	0	0	0
I5	0	0	0	0	0	1	0

จากตารางที่ 3.6 หลังจากที่ลบรายการ T200 ที่พิจารณาแล้วออกจากเมทริกซ์ ทำให้ทราบว่ารายการที่ต้องพิจารณาถัดไป คือ T300

พิจารณา T300 จะเห็นว่าใน T300 มีไอเท็ม I2 และ I3 ปรากฏอยู่จึงนำข้อมูลของ I2 และ I3 ในเมทริกซ์มาทำการ AND กันในแนวคอลัมน์ และเมื่อทำการ AND แล้วให้นำผลของการ AND มา

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บวกกันในทุกรายการ แล้วจึงไปพิจารณาในรายการถัดไป พร้อมทั้งลบคอลัมน์ของรายการ T300 ออก

ตารางที่ 3.7 การ AND กันของ I2 และ I3

I2	1	1	0	1	0	1	1
I3	1	0	1	1	1	1	1
AND	1	0	0	1	0	1	1

จากตารางที่ 3.7 เมื่อทำการ AND กันระหว่าง I2 และ I3 แล้วให้นำผลที่ได้จากการ AND มาบวกกันในทุกรายการ ในที่นี้ได้เท่ากับ 4 ซึ่งมีค่าไม่น้อยกว่าค่าสนับสนุนขั้นต่ำ {I2, I3} จึงเป็นไอเท็มเซตปรากฏบ่อย ดังนั้น L_k = {{I1, I2}, {I1, I5}, {I2, I3}, {I2, I4}, {I2, I5}, {I1, I2, I5}}

ตารางที่ 3.8 เมทริกซ์ที่ได้หลังจากลบ T300

Items \ Tid	T400	T500	T600	T700	T800	T900
I1	1	1	0	1	1	1
I2	1	0	1	0	1	1
I3	0	1	1	1	1	1
I4	1	0	0	0	0	0
I5	0	0	0	0	1	0

จากตารางที่ 3.8 หลังจากที่ลบรายการ T300 ที่พิจารณาแล้วออกจากเมทริกซ์ ทำให้ทราบว่ารายการที่ต้องพิจารณาถัดไป คือ T400

พิจารณา T400 จะเห็นว่าใน T400 มีไอเท็ม I1 I2 และ I4 ปรากฏอยู่จึงนำข้อมูล I1 I2 และ I4 ในเมทริกซ์มาทำการ AND กันในแนวคอลัมน์ และเมื่อทำการ AND แล้วให้นำผลของการ AND มาบวกกันในทุกรายการ แล้วจึงไปพิจารณาในรายการถัดไป พร้อมทั้งลบคอลัมน์ของรายการ T400 ออก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.9 การ AND กันของ I1 I2 และ I4

I1	1	1	0	1	1	1
I2	1	0	1	0	1	1
I4	1	0	0	0	0	0
AND	1	0	0	0	0	0

จากตารางที่ 3.9 เมื่อทำการ AND กันระหว่าง I1 I2 และ I4 แล้วให้นำผลที่ได้จากการ AND มาบวกกันในทุกรายการ ในที่นี้ได้เท่ากับ 1 ซึ่งมีค่าน้อยกว่าค่าสนับสนุนขั้นต่ำ ดังนั้น {I1, I2, I4} จึงไม่เป็นไอเท็มเซตปรากฏบ่อย

เนื่องจาก {I1, I2, I4} ไม่เป็นไอเท็มเซตปรากฏบ่อย จึงต้องพิจารณาเซตย่อยของไอเท็มเซตนี้ด้วย โดยพิจารณาไอเท็มเซตที่มีสมาชิกน้อยกว่าเพียง 1 ตัวก่อน ดังนี้ {I1, I2}, {I1, I4}, {I2, I4} แต่ได้พิจารณา {I1, I2,}, {I2, I4} ไปแล้ว จึงพิจารณาเพียง {I1, I4}

ตารางที่ 3.10 การ AND กันของ I1 และ I4

I1	1	1	0	1	1	1
I4	1	0	0	0	0	0
AND	1	0	0	0	0	0

จากตารางที่ 3.10 เมื่อทำการ AND กันระหว่าง I1 และ I4 แล้วให้นำผลที่ได้จากการ AND มาบวกกันในทุกรายการ ในที่นี้ได้เท่ากับ 1 ซึ่งมีค่าน้อยกว่าค่าสนับสนุนขั้นต่ำ ดังนั้น {I1, I4} จึงไม่เป็นไอเท็มเซตปรากฏบ่อย

ตารางที่ 3.11 เมทริกซ์ที่ได้หลังจากลบ T400

Tid / Items	**T500**	**T600**	**T700**	**T800**	**T900**
I1	1	0	1	1	1
I2	0	1	0	1	1
I3	1	1	1	1	1
I4	0	0	0	0	0
I5	0	0	0	1	0

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากตารางที่ 3.11 หลังจากที่ลบรายการ T400 ที่พิจารณาแล้วออกจากเมทริกซ์ ทำให้ทราบว่ารายการที่ต้องพิจารณาถัดไป คือ T500

พิจารณา T500 จะเห็นว่าใน T500 มีไอเท็ม I1 และ I3 ปรากฏอยู่จึงนำข้อมูลของ I1 และ I3 ในเมทริกซ์มาทำการ AND กันในแนวคอลัมน์ และเมื่อทำการ AND แล้วให้นำผลของการ AND มาบวกกันในทุกรายการ แล้วจึงไปพิจารณาในรายการถัดไป พร้อมทั้งลบคอลัมน์ของรายการ T500 ออก

ตารางที่ 3.12 การ AND กันของ I1 และ I3

I1	1	0	1	1	1
I3	1	1	1	1	1
AND	1	0	1	1	1

จากตารางที่ 3.12 เมื่อทำการ AND กันระหว่าง I1 และ I3 แล้วให้นำผลที่ได้จากการ AND มาบวกกันในทุกรายการ ในที่นี้ได้เท่ากับ 4 ซึ่งมีค่าไม่น้อยกว่าค่าสนับสนุนขั้นต่ำ {I1, I3} จึงเป็นไอเท็มเซตปรากฏบ่อย ดังนั้น L_k = {{I1, I2}, {I1, I3}, {I1, I5}, {I2, I3}, {I2, I4}, {I2, I5}, {I1, I2, I5}}

ตารางที่ 3.13 เมทริกซ์ที่ได้หลังจากลบ T500

Items \ Tid	**T600**	**T700**	**T800**	**T900**
I1	0	1	1	1
I2	1	0	1	1
I3	1	1	1	1
I4	0	0	0	0
I5	0	0	1	0

จากตารางที่ 3.13 หลังจากที่ลบรายการ T500 ที่พิจารณาแล้วออกจากเมทริกซ์ ทำให้ทราบว่ารายการที่ต้องพิจารณาถัดไป คือ T600

พิจารณา T600 จะเห็นว่าใน T600 มีไอเท็ม I2 และ I3 ปรากฏอยู่ แต่เหมือนกับ T300 ซึ่งได้มีการพิจารณาไปแล้ว ดังนั้นจึงพิจารณาที่รายการถัดไป พร้อมลบคอลัมน์ของรายการ T600 ออก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.14 เมทริกซ์ที่ได้หลังจากลบ T600

Items \ Tid	T700	T800	T900
I1	1	1	1
I2	0	1	1
I3	1	1	1
I4	0	0	0
I5	0	1	0

จากตารางที่ 3.14 หลังจากที่ลบรายการ T600 ที่พิจารณาแล้ว ออกจากเมทริกซ์ทำให้ทราบว่ารายการที่ต้องพิจารณาถัดไป คือ T700

พิจารณา T700 จะเห็นว่าใน T700 มีไอเท็ม I1 และ I3 ปรากฏอยู่ แต่เหมือนกับ T500 ซึ่งได้มีการพิจารณาไปแล้วดังนั้นจึงลบรายการนี้แล้วพิจารณาที่รายการถัดไป พร้อมลบคอลัมน์ของรายการ T700 ออก

ตารางที่ 3.15 เมทริกซ์ที่ได้หลังจากลบ T700

Items \ Tid	T800	T900
I1	1	1
I2	1	1
I3	1	1
I4	0	0
I5	1	0

จากตารางที่ 3.15 หลังจากที่ลบรายการ T700 ที่พิจารณาแล้วออกจากเมทริกซ์ ทำให้ทราบว่ารายการที่ต้องพิจารณาถัดไป คือ T800

พิจารณา T800 จะเห็นว่าใน T800 มีไอเท็ม I1 I2 I3 และ I5 ปรากฏอยู่จึงนำข้อมูล I1 I2 I3 และ I5 ในเมทริกซ์มาทำการ AND กันในแนวคอลัมน์ และเมื่อทำการ AND แล้วให้นำผลของการ AND มาบวกกันในทุกรายการ แล้วจึงไปพิจารณาในรายการถัดไป พร้อมทั้งลบคอลัมน์ของรายการ T800 ออก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.16 การ AND กันของ I1 I2 และ I3

I1	1	1
I2	1	1
I3	1	1
I5	1	0
AND	1	0

จากตารางที่ 3.16 เมื่อทำการ AND กันระหว่าง I1 I2 I3 และ I5 แล้วให้นำผลที่ได้จากการ AND มาบวกกันในทุกรายการ ในที่นี้ได้เท่ากับ 1 ซึ่งมีค่าน้อยกว่าค่าสนับสนุนขั้นต่ำ ดังนั้น {I1, I2, I3, I5} จึงไม่เป็นไอเท็มเซตปรากฏบ่อย

เนื่องจาก {I1, I2, I3, I5} ไม่เป็นไอเท็มเซตปรากฏบ่อย จึงต้องพิจารณาเซตย่อยของไอเท็มเซตนี้ด้วย โดยพิจารณาไอเท็มเซตที่มีสมาชิกน้อยกว่าเพียง 1 ตัวก่อน ดังนี้ {I1, I2, I3}, {I1, I2, I5}, {I1, I3, I5} แต่ได้พิจารณา {I1, I2, I5} ไปแล้ว จึงพิจารณาเพียง {I1, I2, I3} และ {I1, I3, I5}

ตารางที่ 3.17 การ AND กันของ I1 I2 และ I3

I1	1	1
I2	1	1
I3	1	1
AND	1	1

จากตารางที่ 3.17 เมื่อทำการ AND กันระหว่าง I1 I2 และ I3 แล้วให้นำผลที่ได้จากการ AND มาบวกกันในทุกรายการ ในที่นี้ได้เท่ากับ 2 ซึ่งมีค่าไม่น้อยกว่าค่าสนับสนุนขั้นต่ำ ดังนั้น {I1, I2, I3} จึงเป็นไอเท็มเซตปรากฏบ่อย และจากคุณสมบัติของอัลกอริทึมเอไพรออริ คือ ทุกๆ เซตย่อย (subset) ของไอเท็มเซตปรากฏบ่อยต้องเป็นไอเท็มเซตปรากฏบ่อย ก็จะได้ว่า {I1, I2}, {I1, I3} และ {I2, I3} เป็น ไอเท็มเซตปรากฏบ่อยด้วย ดังนั้น L_k = {{I1, I2}, {I1, I3}, {I1, I5}, {I2, I3},{I2, I4}, {I2, I5}, {I1, I2, I3}, {I1, I2, I5}}

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.18 การ AND กันของ I1 I3 และ I5

I1	1	1
I3	1	1
I5	1	0
AND	1	0

จากตารางที่ 3.18 เมื่อทำการ AND กันระหว่าง I1 I3 และ I5 แล้วให้นำผลที่ได้จากการ AND มาบวกกันในทุกรายการ ในที่นี้ได้เท่ากับ 1 ซึ่งมีค่าน้อยกว่าค่าสนับสนุนขั้นต่ำ ดังนั้น {I1, I3, I5} จึงไม่เป็นไอเท็มเซตปรากฏบ่อย

เนื่องจาก {I1, I3, I5} ไม่เป็นไอเท็มเซตปรากฏบ่อย จึงต้องพิจารณาเซตย่อยของไอเท็มเซตนี้ด้วย โดยพิจารณาไอเท็มเซตที่มีสมาชิกน้อยกว่าเพียง 1 ตัวก่อน ดังนี้ {I1, I3}, {I1, I5}, {I3, I5} แต่ได้พิจารณา {I1, I3}, {I1, I5} ไปแล้ว จึงพิจารณาเพียง {I3, I5}

ตารางที่ 3.19 การ AND กันของ I3 และ I5

I3	1	1
I5	1	0
AND	1	0

จากตารางที่ 3.19 เมื่อทำการ AND กันระหว่าง I3 และ I5 แล้วให้นำผลที่ได้จากการ AND มาบวกกันในทุกรายการ ในที่นี้ได้เท่ากับ 1 ซึ่งมีค่าน้อยกว่าค่าสนับสนุนขั้นต่ำ ดังนั้น {I3, I5} จึงไม่เป็นไอเท็มเซตปรากฏบ่อย

ตารางที่ 3.20 เมทริกซ์ที่ได้หลังจากลบ T800

Items \ Tid	**T900**
I1	1
I2	1
I3	1
I4	0
I5	0

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากตารางที่ 3.20 จะเห็นได้ว่ามีเพียงรายการเดียวที่ต้องพิจารณา ดังนั้นจึงไม่จำเป็นต้องทำการหาต่อไป เนื่องจากผลบวกของผลลัพธ์จากการ AND จะมีค่าน้อยกว่าค่าสนับสนุนขั้นต่ำที่เท่ากับ 2 เสมอ ดังนั้นจึงสรุปว่า จากการดำเนินการในขั้นตอนที่ 3 เสร็จเรียบร้อยแล้ว ก็จะได้ไอเท็มเซตปรากฏบ่อยทั้งหมด ซึ่งประกอบด้วยไอเท็มเซตดังนี้

ไอเท็มเซตปรากฏบ่อยที่มีสมาชิก 1 ไอเท็ม: {I1}, {I2}, {I3}, {I4} และ {I5}

ไอเท็มเซตปรากฏบ่อยที่มีสมาชิก 2 ไอเท็ม: {I1, I2}, {I1, I3}, {I1, I5}, {I2, I3}, {I2, I4} และ {I2, I5}

ไอเท็มเซตปรากฏบ่อยที่มีสมาชิก 3 ไอเท็ม: {I1, I2, I3} และ {I1, I2, I5}

3.3 ส่วนโปรแกรมติดต่อกับผู้ใช้

สำหรับอัลกอริทึมที่ได้พัฒนาขึ้นมานั้น จะนำเสนอในรูปแบบที่มีการติดต่อกับผู้ใช้งานได้ด้วย เพื่อความสะดวกในการทดสอบการทำงานของอัลกอริทึมนี้ ได้จัดทำโปรแกรมเพื่อใช้ในการทดสอบขึ้นมา โดยโปรแกรมได้มีการนำอัลกอริทึมเอไพรออริมาทำการทดสอบ ร่วมด้วย

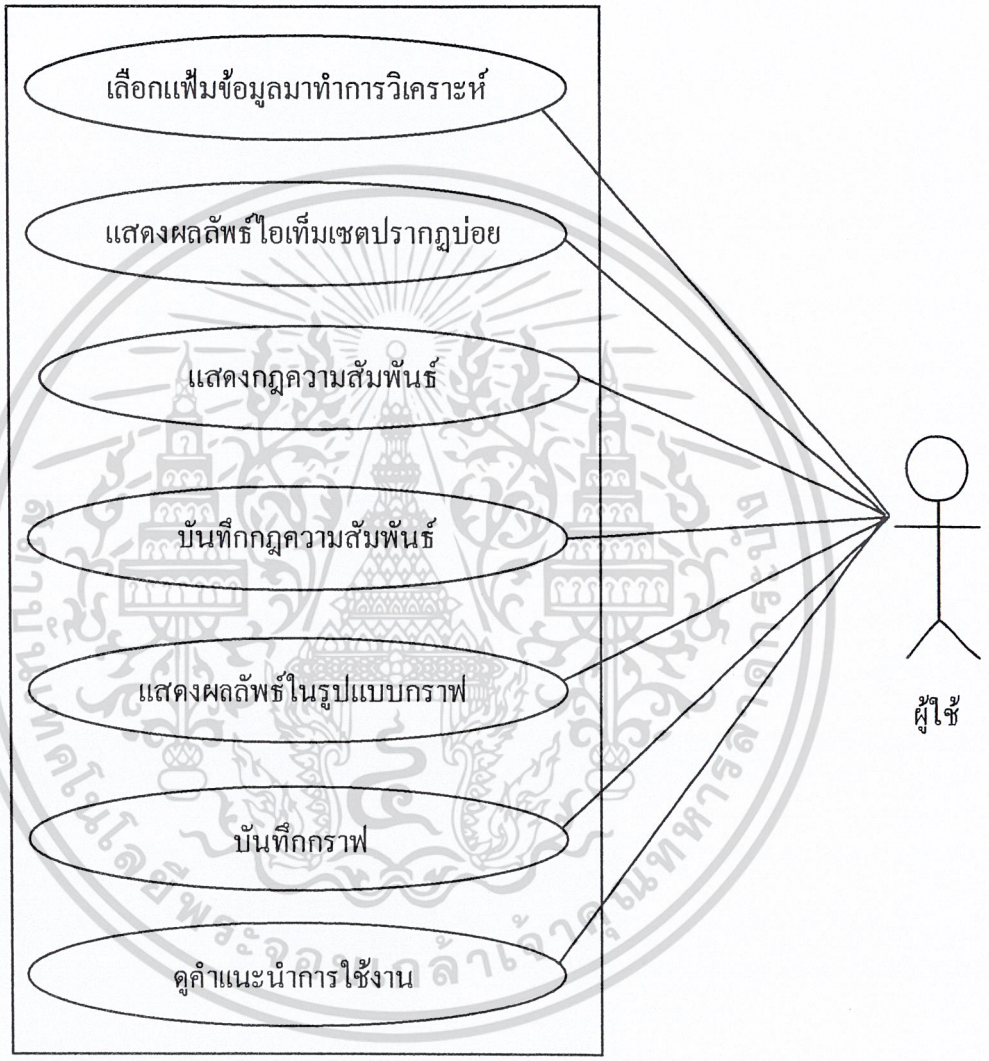
ขั้นตอนในการใช้งานโปรแกรม คือ ผู้ใช้ต้องเข้าสู่หน้าโปรแกรมสำหรับทดสอบ ซึ่งในหน้าจอโปรแกรมนี้ผู้ใช้ต้องทำการเลือกแฟ้มข้อมูลที่นำมาวิเคราะห์ และป้อนค่าสนับสนุนขั้นต่ำ เพื่อให้โปรแกรมจะได้ทำการค้นหาไอเท็มเซตปรากฏบ่อย ที่มีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุน ขั้นต่ำออกมา นอกจากนั้นแล้วผู้ใช้ยังสามารถทำการค้นหากฎความสัมพันธ์โดยการป้อนค่าความเชื่อมั่นขั้นต่ำ โปรแกรมจะทำการค้นหากฎความสัมพันธ์ และแสดงผลลัพธ์ออกสู่หน้าจอ

หลังจากที่ผู้ใช้ได้ทำการกำหนดค่าสนุบสนุนขั้นต่ำแล้ว โปรแกรมจะเริ่มทำการทดสอบข้อมูลกับอัลกอริทึมเอไพรออริและอัลกอริทึมที่ได้พัฒนาขึ้น โดยขั้นตอนการทดสอบโปรแกรมจะมีการหาไอเท็มเซตปรากฏบ่อยจากค่าสนับสนุนขั้นต่ำตามที่ผู้ใช้ป้อนไว้และค่าสนับสนุนขั้นต่ำ 2 ค่า โดยโปรแกรมจะทำการสุ่มมาให้โดยอ้างอิงจากค่ากลางจากที่ผู้ใช้ป้อน จากนั้นโปรแกรมจะแสดงผลลัพธ์การทดสอบออกมาเป็นทุกเซตของไอเท็มเซตปรากฏบ่อย สามารถแสดงกราฟของค่าสนับสนุนขั้นต่ำกับเวลาที่ใช้ในการทำงานเปรียบเทียบระหว่างอัลกอริทึมเอไพรออริกับอัลกอริทึมที่ได้พัฒนาขึ้น และกราฟแสดงเวลาที่ใช้ในการทำงานในการค้นหาไอเท็มเซตปรากฏบ่อย ที่เปรียบเทียบระหว่างอัลกอริทึมเอไพรออริกับอัลกอริทึมที่ได้พัฒนาขึ้น และผู้ใช้สามารถบันทึก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ข้อมูลของกฎความสัมพันธ์ในรูปแบบไฟล์ .xls และกราฟแสดงการเปรียบเทียบเวลาการทำงานในการค้นหาไอเท็มเซตปรากฏบ่อยในรูปแบบไฟล์ .png

การทำงานของโปรแกรมทดสอบ แสดงเป็น Use Case Diagram ได้ดังรูปที่ 3.3



รูปที่ 3.3 Use Case Diagram แสดงการทำงานของโปรแกรม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.21 Use case description เลือกแฟ้มข้อมูลมาทำการวิเคราะห์

Use Case Name	เลือกแฟ้มข้อมูลมาทำการวิเคราะห์	
Scenario	ผู้ใช้เลือกแฟ้มข้อมูลมาทำการวิเคราะห์เพื่อหาไอเท็มเซตปรากฏบ่อย	
Triggering Event	ผู้ใช้ทำการกดปุ่มเลือกแฟ้มข้อมูล	
Brief Description	ผู้ใช้เลือกแฟ้มข้อมูลที่ต้องการหาไอเท็มเซตปรากฏบ่อย มาทำการวิเคราะห์โดยการคลิกที่ปุ่ม Browse ซึ่งแฟ้มที่นำมาวิเคราะห์ต้องเป็นนามสกุลที่โปรแกรมรองรับ (*.txt, *.dat)	
Actors	ผู้ใช้	
Precondition	-	
Postcondition	-	
Flow of Events	**ผู้ใช้งาน**	**ระบบ**
	1. เลือกไฟล์ที่จะนำมาวิเคราะห์	2.ทำการอ่านแฟ้มข้อมูลที่ถูกเลือก
Alternative flow	-	

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.22 Use case description แสดงผลลัพธ์ไอเท็มเซตปรากฏบ่อย

Use Case Name	แสดงผลลัพธ์ไอเท็มเซตปรากฏบ่อย	
Scenario	แสดงผลลัพธ์ของการทำงานโดยผลที่ได้เป็นไอเท็มเซตปรากฏบ่อย	
Triggering Event	เมื่อผู้ใช้กดปุ่มสร้างไอเท็มเซตปรากฏบ่อย	
Brief Description	โปรแกรมจะแสดงผลลัพธ์เป็นไอเท็มเซตปรากฏบ่อย แต่ถ้าหากไม่มีไอเท็มเซตใดเลยที่มีค่ามากกว่าค่าสนับสนุนขั้นต่ำตามที่กำหนด โปรแกรมจะแสดง “ไม่มีไอเท็มเซตปรากฏบ่อย”	
Actors	ระบบ	
Preconditions	เลือกแฟ้มข้อมูลมาทำการวิเคราะห์	
Postcondition	-	
Flow of Events	**ผู้ใช้งาน**	**ระบบ**
	1. กดปุ่ม Generate Frequent-Itemset 2. ทำการเลือกค่าสนับสนุนขั้นต่ำตามที่ต้องการ	3. สร้างเมทริกซ์รายการ จากไฟล์ข้อมูลที่ผู้ใช้เลือกมาทำการวิเคราะห์ 4. หาผลรวมของค่าสนับสนุนของแต่ละไอเท็ม และผลรวมของไอเท็มในแต่ละรายการ 5. เปรียบเทียบค่าสนับสนุนของไอเท็มกับค่าสนับสนุนขั้นต่ำ 6. ทำการค้นหาไอเท็มเซตปรากฏบ่อย 7. แสดงผลลัพธ์ ไอเท็มเซตปรากฏบ่อย
Alternative flow	1.แจ้งข้อความเตือนกรณีผู้ใช้ยังไม่ได้เลือกไฟล์ในการนำมาวิเคราะห์	

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.23 Use case description แสดงผลลัพธ์ในรูปแบบกราฟ

Use Case Name	แสดงผลลัพธ์ในรูปแบบกราฟ	
Scenario	แสดงผลการทำงานในรูปแบบกราฟ	
Triggering Event	เมื่อผู้ใช้กดปุ่มแสดงกราฟ	
Brief Description	กดที่ปุ่ม View Graph โปรแกรมจะแสดงกราฟเส้นของค่าสนับสนุนขั้นต่ำกับเวลาที่ใช้ในการค้นหาไอเท็มเซตปรากฏบ่อยระหว่างอัลกอริทึมเอไพรออริกับอัลกอริทึมที่ได้พัฒนาขึ้น	
Actors	ผู้ใช้	
Preconditions	1.แสดงผลลัพธ์ไอเท็มเซตปรากฏบ่อย 2.บันทึกผลลัพธ์ไอเท็มเซตปรากฏบ่อย	
Postconditions	บันทึกกราฟ	
Flow of Events	**ผู้ใช้งาน**	**ระบบ**
	1. กดปุ่ม View Graph	2. โปรแกรมแสดงผลการทำงานในรูปแบบกราฟเส้นเปรียบเทียบเวลาที่ใช้ในการทำงานกับค่าสนับสนุนขั้นต่ำต่างๆ
Alternative flow	1.แจ้งข้อความเตือนกรณีผู้ใช้ยังไม่ได้สั่งค้นหาไอเท็มเซตปรากฏบ่อย	

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.24 Use case description บันทึกกราฟ

Use Case Name	บันทึกกราฟ	
Scenario	บันทึกรูปภาพของกราฟที่ได้จากการประมวลผล	
Triggering Event	เมื่อผู้ใช้กดปุ่มบันทึกกราฟ	
Brief Description	บันทึกรูปภาพของกราฟที่ได้จากการเปรียบเทียบเวลาการทำงานในการค้นหาไอเท็มเซตปรากฏบ่อยกับค่าสนับสนุนขั้นต่ำต่างๆ โดยโปรแกรมจะบันทึกภาพกราฟเป็นแฟ้มนามสกุล .png	
Actors	ผู้ใช้	
Preconditions	แสดงผลลัพธ์ในรูปแบบกราฟ	
Postconditions	-	
Flow of Events	**ผู้ใช้งาน**	**ระบบ**
	1. คลิกขวาที่กราฟ 2. เลือก Save as 2. ป้อนชื่อไฟล์	3. โปรแกรมจะทำการบันทึกข้อมูลกราฟ เป็นแฟ้มนามสกุล .png
Alternative flow	-	

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.25 Use case description แสดงกฎความสัมพันธ์

Use Case Name	แสดงกฎความสัมพันธ์	
Scenario	ค้นหากฎความสัมพันธ์ และแสดงกฎความสัมพันธ์	
Triggering Event	เมื่อผู้ใช้กดปุ่มแสดงกฎความสัมพันธ์	
Brief Description	โปรแกรมทำการแสดงกฎความสัมพันธ์ตามค่าความเชื่อมั่นขั้นต่ำที่ผู้ใช้ระบุ	
Actors	ผู้ใช้	
Preconditions	แสดงผลลัพธ์ไอเท็มเซตปรากฏบ่อย	
Postconditions	บันทึกกฎความสัมพันธ์	
Flow of Events	**ผู้ใช้งาน**	**ระบบ**
	1. กดปุ่ม Generate Rules 2. ป้อนค่าความเชื่อมั่นขั้นต่ำ	3. ระบบทำการประมวลผลเพื่อค้นหากฎความสัมพันธ์
Alternative flow	1.แจ้งข้อความเตือนกรณีผู้ใช้ยังไม่ได้เลือกไฟล์ในการนำมาวิเคราะห์	

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.26 Use case description บันทึกกฎความสัมพันธ์

Use Case Name	บันทึกกฎความสัมพันธ์	
Scenario	บันทึกบันทึกกฎความสัมพันธ์ที่ได้จากการประมวลผล	
Triggering Event	เมื่อผู้ใช้กดปุ่มบันทึกกฎความสัมพันธ์	
Brief Description	บันทึกผลลัพธ์ที่ได้จากการหากฎความสัมพันธ์ โดยโปรแกรมจะบันทึกเป็นแฟ้มนามสกุล .xls	
Actors	ผู้ใช้	
Preconditions	แสดงกฎความสัมพันธ์	
Postconditions	-	
Flow of Events	ผู้ใช้งาน	ระบบ
	1. กดปุ่ม Save File 2. ป้อนชื่อไฟล์	3. โปรแกรมจะทำการบันทึกข้อมูลกฎความสัมพันธ์ ค่าสนับสนุน ค่าความเชื่อมั่นและค่าความสหสัมพันธ์ตามที่ได้จากการประมวลผล
Alternative flow	-	

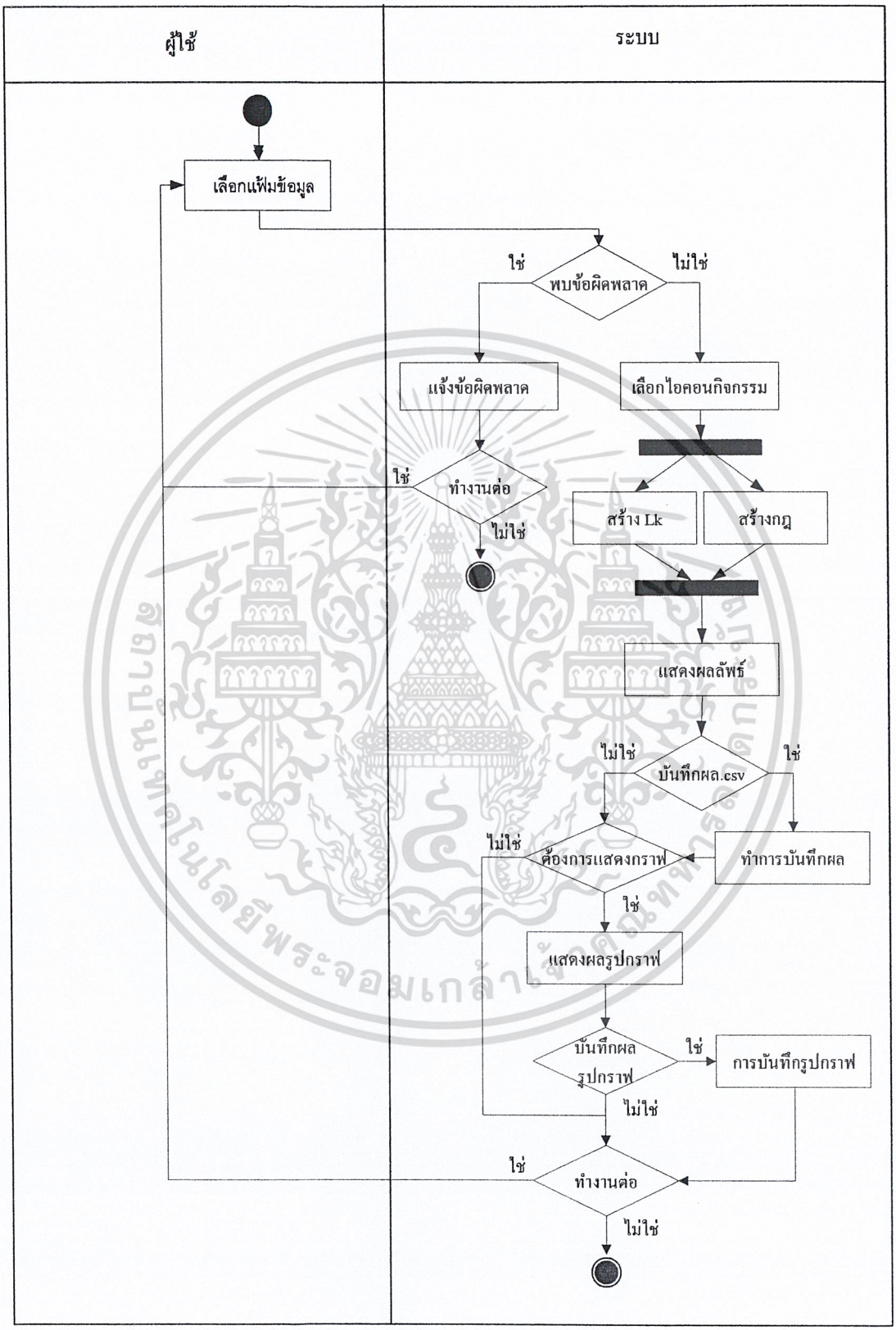
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.27 Use case description ดูคำแนะนำการใช้งาน

Use Case Name	ดูคำแนะนำการใช้งาน	
Scenario	ผู้ใช้เลือกดูคำแนะนำการใช้งานเบื้องต้นสำหรับโปรแกรม	
Triggering Event	เมื่อผู้ใช้กดปุ่มดูคำแนะนำการใช้งาน	
Brief Description	ผู้ใช้เลือกดูคำแนะนำเบื้องต้นในการใช้งานต่างๆของโปรแกรม	
Actors	ผู้ใช้	
Precondition	-	
Postcondition	-	
Flow of Events	**ผู้ใช้งาน**	**ระบบ**
	1. กดปุ่ม Click to see an instruction	2.แสดงคำแนะนำในการใช้งานต่างๆของโปรแกรม
Alternative flow	-	

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

แผนภาพกิจกรรม



รูปที่ 3.4 กิจกรรมในการใช้งานระบบของผู้ใช้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำ ให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูปที่ 3.4 เมื่อผู้ใช้ต้องการใช้งานโปรแกรมสำเร็จรูปเพื่อหากฎความสัมพันธ์ ขั้นแรกผู้ใช้จะต้องทำการเปิดไฟล์ข้อมูลก่อน โปรแกรมจะทำการตรวจสอบว่าประเภทไฟล์ที่ผู้ใช้เปิดถูกต้องหรือไม่ ถ้าพบว่าถูกต้อง ผู้ใช้สามารถดำเนินการต่อโดยสามารถเลือกกิจกรรมที่จะทำต่อได้ คือ สร้างไอเท็มเซตปรากฏบ่อย หรือ สร้างกฎความสัมพันธ์ หลังจากที่ผู้ใช้ระบุกิจกรรมที่ต้องการทำแล้วระบบจะทำการแสดงผลลัพธ์ หลังจากนั้นหากผู้ใช้ต้องการบันทึกผลลัพธ์ทันทีหรือแสดงกราฟแล้วจึงบันทึกผลก็สามารถกระทำได้ หรือผู้ใช้ต้องการออกจากโปรแกรมก็จะเป็นการสิ้นสุดการทำงานของระบบ

3.4 รูปแบบของไฟล์ที่รองรับ

รูปแบบของไฟล์ที่รองรับ ประกอบด้วย text (.txt) และ Data file (.dat)

ในการออกแบบรูปแบบการเก็บไฟล์ของข้อมูลที่จะนำมาใช้ในการวิเคราะห์ จะให้แต่ละบรรทัดแทนด้วยแต่ละรายการ และในแต่ละบรรทัดจะประกอบไปด้วยไอเท็มที่ปรากฏอยู่ในรายการ ซึ่งแต่ละไอเท็มนั้นจะถูกคั่นด้วยช่องว่าง ดังตัวอย่าง

```
1 2          // ไอเท็มที่ปรากฏในรายการที่ 1
1 3 4        // ไอเท็มที่ปรากฏในรายการที่ 2
2 3 4 5      // ไอเท็มที่ปรากฏในรายการที่ 3
3 5          // ไอเท็มที่ปรากฏในรายการที่ 4
1 2 4        // ไอเท็มที่ปรากฏในรายการที่ 5
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4

ผลการทดลอง

4.1 แหล่งที่มาและรายละเอียดของข้อมูล

ในการทดลองได้นำข้อมูลที่มีลักษณะที่แตกต่างกัน 2 ชุดข้อมูล มาทำการทดลอง โดยชุดข้อมูลแรกมีลักษณะเป็นข้อมูลรายการ (Transactional Data) คือชุดข้อมูลที่ได้จากการสังเคราะห์ด้วยโปรแกรม Data Generator และอีกชุดข้อมูลมีลักษณะเป็นชุดข้อมูลจำแนกประเภท (Categorical Data) คือชุดข้อมูล Mushroom ซึ่งมีรายละเอียดของแต่ละชุดข้อมูล ดังนี้

4.1.1 ชุดข้อมูล Data Generator

ข้อมูลที่ใช้ในการทดลองเพื่อทดสอบประสิทธิภาพการทำงานของอัลกอริทึมที่ได้พัฒนาขึ้นสังเคราะห์จากโปรแกรมสร้างข้อมูลเสมือนจริงที่ชื่อว่า IBM Quest Market-Basket Synthetic Data Generator ดาวน์โหลดโปรแกรมได้ที่ http://www.almaden.ibm.com/cs/disciplines/iis/#assocSynData หรือ http://www.cs.loyola.edu/~cgiannel/assoc_gen.html เป็นโปรแกรมสร้างข้อมูลที่มีลักษณะคล้ายกับข้อมูลจริงที่ได้พัฒนาโดยทีมวิจัยของบริษัท IBM โดยในการสร้างข้อมูลนั้นผู้ใช้สามารถกำหนดลักษณะของข้อมูลเองได้ จากค่าต่างๆ 3 ค่า ได้แก่

1. จำนวนรายการ
2. ค่าเฉลี่ยของไอเท็มในแต่ละรายการ
3. จำนวนไอเท็มที่แตกต่างกันทั้งหมด

สำหรับในงานวิจัยนี้ได้ทำการสุ่มชุดข้อมูลตัวอย่าง 6 ชุดข้อมูล สำหรับใช้ทดสอบประสิทธิภาพการทำงานของอัลกอริทึมที่ได้พัฒนาขึ้น โดยกำหนดลักษณะของชุดข้อมูลดังต่อนี้

ตารางที่ 4.1 แสดงลักษณะของข้อมูล

	จำนวนรายการ	ค่าเฉลี่ยของไอเท็มในรายการ	จำนวนไอเท็ม
ชุดข้อมูลที่ 1	500	10	20
ชุดข้อมูลที่ 2	1000	10	20
ชุดข้อมูลที่ 3	3000	10	20
ชุดข้อมูลที่ 4	5000	10	20
ชุดข้อมูลที่ 5	7000	10	20
ชุดข้อมูลที่ 6	9000	10	20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.1.2 ชุดข้อมูล Mushroom

ชุดข้อมูลทดสอบอีกชุดที่ใช้ในการทดลองเพื่อทดสอบประสิทธิภาพการทำงานของอัลกอริทึมที่ได้พัฒนาขึ้นนั้น ได้ทำการคัดเลือกชุดข้อมูลมาจากแหล่งข้อมูลของมหาวิทยาลัยแห่งรัฐแคลิฟอร์เนีย เมืองเออร์ไวย์ ประเทศสหรัฐอเมริกา (http://archive.ics.uci.edu/ml/) (Blake and Merz, 1998) ซึ่งได้รวบรวมชุดข้อมูลสำหรับทดสอบประสิทธิภาพการทำงานของอัลกอริทึมต่างๆ ของการทำเหมืองข้อมูล โดยชุดข้อมูลที่ได้คัดเลือกมานั้น คือชุดข้อมูล Mushroom

ชุดข้อมูล Mushroom เป็นชุดข้อมูลที่เก็บข้อมูลเกี่ยวกับเห็ด ซึ่งมีทั้งหมด 22 คุณลักษณะ แบ่งออกเป็น 3 กลุ่ม คือ สามารถทานได้ ไม่สามารถทานได้และไม่แนะนำให้ทาน ประกอบด้วยรายการทั้งหมด 8124 รายการ โดยมีคุณลักษณะต่างๆ ดังนี้

- รูปร่างของหมวกเห็ด
- ผิวของหมวกเห็ด
- สีของหมวกเห็ด
- รอยช้ำของเห็ด
- กลิ่นของเห็ด
- สิ่งที่ติดมากับครีบเห็ด
- ช่องว่างใต้หมวกเห็ด
- ขนาดใต้หมวกเห็ด
- สีใต้หมวกเห็ด
- ประเภทสิ่งปกคลุม
- สีของสิ่งปกคลุม
- จำนวนของวงแหวน
- ประเภทของวงแหวน
- รูปร่างลำต้น
- ลำต้นของราก
- ผิวด้านบนของลำต้น
- ผิวด้านล่างของลำต้น
- สีด้านบนของลำต้น
- สีด้านล่างของลำต้น
- สีของสปอร์
- จำนวนประชากร
- แหล่งที่อยู่อาศัย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2 ผลการทดลอง

การทดลองวัดประสิทธิภาพการทำงานของอัลกอริทึมที่ได้ปรับปรุงขึ้นนั้น ได้ทำการพัฒนาด้วยโปรแกรม NetBeans IDE 6.9.1 (Build 201007282301) โดยใช้ Compiler คือ Java : 1.6.0_21 ; Java HotSpot(TM) Client VM 17 [5]

4.2.1 เปรียบเทียบเวลาการทำงานในชุดข้อมูล Data Generator

การทดลองทำโดยวิเคราะห์เปรียบเทียบประสิทธิภาพการทำงานของอัลกอริทึมที่พัฒนาขึ้นกับอัลกอริทึมเอไพรออริ โดยเปรียบเทียบและวิเคราะห์ผลด้านเวลาการทำงานในการค้นหาไอเท็มเซตปรากฏบ่อย ในชุดข้อมูล Data Generator ที่มีจำนวนรายการ 500 รายการ ระหว่างอัลกอริทึมเอไพรออริกับอัลกอริทึมที่ได้พัฒนาขึ้นซึ่งผลลัพธ์ที่ได้แสดงเป็นตารางที่ 4.2

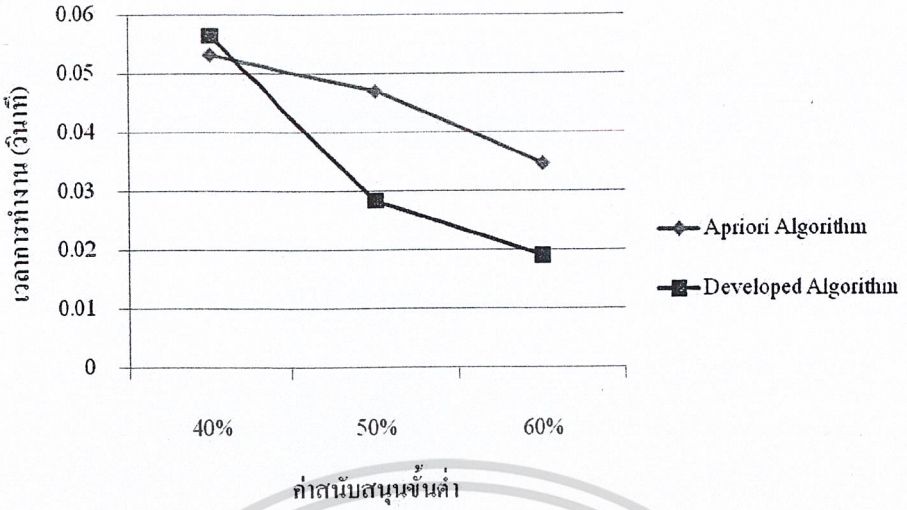
ตารางที่ 4.2 เวลาเฉลี่ยในการหาไอเท็มเซตปรากฏบ่อยของอัลกอริทึมเอไพรออริกับอัลกอริทึมที่พัฒนาขึ้นในข้อมูล Data Generator

อัลกอริทึม	เวลาการทำงาน (วินาที)		
	40%	50%	60%
Apriori Algorithm	0.0532	0.0470	0.0348
Developed Algorithm	0.0564	0.0284	0.0190

จากตารางที่ 4.2 แสดงผลลัพธ์การทำงานของอัลกอริทึมเอไพรออริ และอัลกอริทึมที่พัฒนาขึ้น โดยแสดงผลลัพธ์ในรูปแบบของเวลา (วินาที) จากค่าสนับสนุนขั้นต่ำที่แตกต่างกัน จะเห็นได้ว่าทั้งอัลกอริทึมเอไพรออริและอัลกอริทึมที่พัฒนาขึ้น เมื่อค่าสนับสนุนขั้นต่ำมีค่ามากขึ้น การทำงานก็จะเร็วขึ้นตามไปด้วย แต่เมื่อเปรียบเทียบประสิทธิภาพระหว่างอัลกอริทึมทั้งสอง จะเห็นได้ว่าเมื่อค่าสนับสนุนขั้นต่ำน้อย อัลกอริทึมเอไพรออริจะทำงานได้เร็วกว่า แต่เมื่อค่าสนับสนุนขั้นต่ำเยอะขึ้น อัลกอริทึมที่พัฒนาขึ้นจะทำงานได้เร็วกว่า

จากข้อมูลในตารางที่ 4.2 เมื่อนำข้อมูลที่ได้มาสร้างเป็นกราฟเวลาในการทำงานเทียบกับค่าสนับสนุนขั้นต่ำ สามารถแสดงได้ดังรูปที่ 4.1

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.1 กราฟแสดงเวลาการทำงานในการหาไอเท็มเซตปรากฏบ่อยของอัลกอริทึม ในชุดข้อมูล Data Generator

สำหรับการเปรียบเทียบเวลาการทำงานในชุดข้อมูล Data Generator ในจำนวนรายการต่างๆ ทำการทดลองเพื่อวิเคราะห์ถึงเวลาการทำงานในการหาไอเท็มเซตปรากฏบ่อย ของอัลกอริทึมที่ได้พัฒนาขึ้นกับอัลกอริทึมเอไพรออริในชุดข้อมูลที่มีจำนวนรายการต่างๆ กันได้แก่ 1000 3000 5000 7000 9000 รายการ โดยทำการวิเคราะห์ที่ค่าสนับสนุนขั้นต่ำ 40% 50% และ 60%

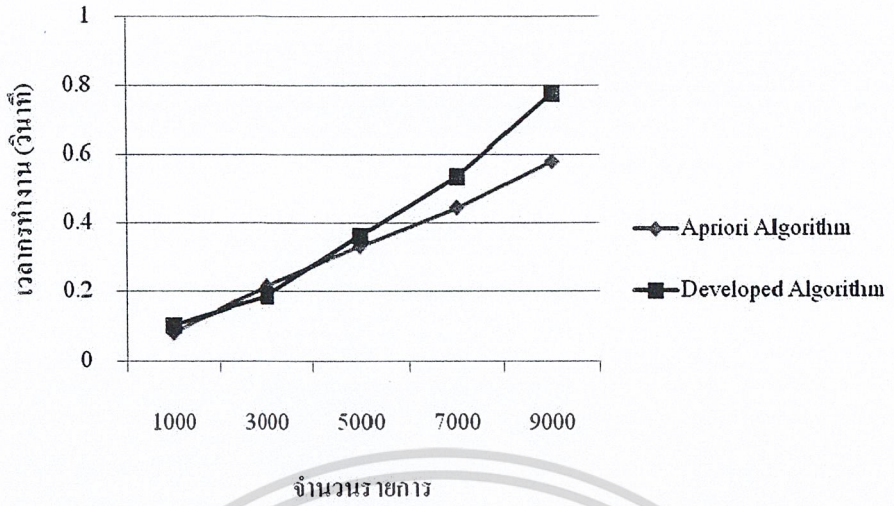
การทดลองเปรียบเทียบเวลาการทำงานในการหาไอเท็มเซตปรากฏบ่อย ของอัลกอริทึมที่ได้พัฒนาขึ้นกับอัลกอริทึมเอไพรออริ ที่จำนวนรายการต่างๆ โดยกำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 40% ได้ผลลัพธ์ดังตารางที่ 4.3

ตารางที่ 4.3 เวลาเฉลี่ยในการค้นหาไอเท็มเซตปรากฏบ่อยของในชุดข้อมูล Data Generator ในจำนวนรายการต่างๆ ที่ค่าสนับสนุนขั้นต่ำ 40%

อัลกอริทึม	เวลาการทำงาน (วินาที)				
	1000	3000	5000	7000	9000
Apriori Algorithm	0.0814	0.2156	0.3312	0.4440	0.5780
Developed Algorithm	0.1036	0.1878	0.3626	0.5346	0.7752

จากข้อมูลในตารางที่ 4.3 เมื่อนำข้อมูลที่ได้มาสร้างเป็นกราฟแสดงเวลาในการทำงานเทียบกับจำนวนรายการ สามารถแสดงได้ดังรูปที่ 4.2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.2 เวลาเฉลี่ยในการค้นหาไอเท็มเซตปรากฏบ่อยของในชุดข้อมูล Data Generator ในจำนวนรายการต่างๆ ที่ค่าสนับสนุนขั้นต่ำ 40%

จากตารางที่ 4.3 และรูปที่ 4.2 แสดงเวลาในการทำงานของอัลกอริทึมเอไพรออริและอัลกอริทึมที่พัฒนาขึ้น โดยแสดงผลลัพธ์ในรูปแบบของเวลา (วินาที) เทียบกับจำนวนรายการที่แตกต่างกัน จะเห็นได้ว่าอัลกอริทึมเอไพรออริและอัลกอริทึมที่พัฒนาขึ้นใช้เวลาในการทำงานที่มากขึ้นเมื่อมีจำนวนรายการมากขึ้น และในค่าสนับสนุนขั้นต่ำที่ 40% อัลกอริทึมที่ได้พัฒนาขึ้นใช้เวลาการทำงานในการหาไอเท็มเซตปรากฏบ่อยมากกว่าอัลกอริทึมเอไพรออริในหลายจำนวนรายการ

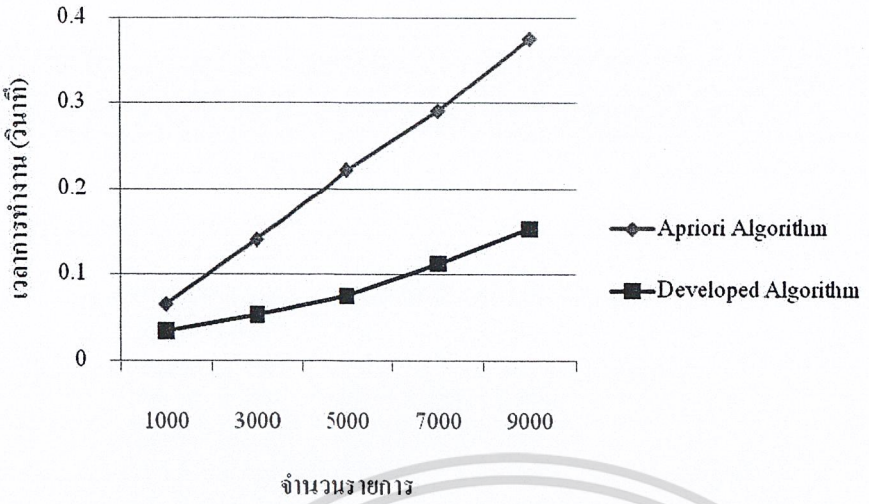
การทดลองเปรียบเทียบเวลาการทำงานในการหาไอเท็มเซตปรากฏบ่อย ของอัลกอริทึมที่ได้พัฒนาขึ้นกับอัลกอริทึมเอไพรออริ ที่จำนวนรายการต่างๆ โดยกำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 50% ได้ผลลัพธ์ดังตารางที่ 4.4

ตารางที่ 4.4 เวลาเฉลี่ยในการค้นหาไอเท็มเซตปรากฏบ่อยของในชุดข้อมูล Data Generator ในจำนวนรายการต่างๆ ที่ค่าสนับสนุนขั้นต่ำ 50%

อัลกอริทึม	เวลาการทำงาน (วินาที)				
	1000	3000	5000	7000	9000
Apriori Algorithm	0.0656	0.1406	0.2220	0.2906	0.3750
Developed Algorithm	0.0344	0.0530	0.0752	0.1126	0.1532

จากข้อมูลในตารางที่ 4.4 เมื่อนำข้อมูลที่ได้มาสร้างเป็นกราฟแสดงเวลาในการทำงานเทียบกับจำนวนรายการ สามารถแสดงได้ดังรูปที่ 4.3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.3 เวลาเฉลี่ยในการค้นหาไอเท็มเซตปรากฏบ่อยของในชุดข้อมูล Data Generator ในจำนวนรายการต่างๆ ที่ค่าสนับสนุนขั้นต่ำ 50%

จากตารางที่ 4.4 และรูปที่ 4.3 แสดงเวลาในการทำงานของอัลกอริทึมเอไพรออริและอัลกอริทึมที่พัฒนาขึ้น โดยแสดงผลลัพธ์ในรูปแบบของเวลา (วินาที) เทียบกับจำนวนรายการที่แตกต่างกัน จะเห็นได้ว่าอัลกอริทึมเอไพรออริและอัลกอริทึมที่พัฒนาขึ้นจะใช้เวลาในการทำงานที่มากขึ้นเมื่อมีจำนวนรายการที่มากขึ้น และในค่าสนับสนุนขั้นต่ำที่ 50% อัลกอริทึมที่ได้พัฒนาขึ้นใช้เวลาในการหาไอเท็มเซตปรากฏบ่อยน้อยกว่าอัลกอริทึมเอไพรออริในทุกจำนวนรายการ

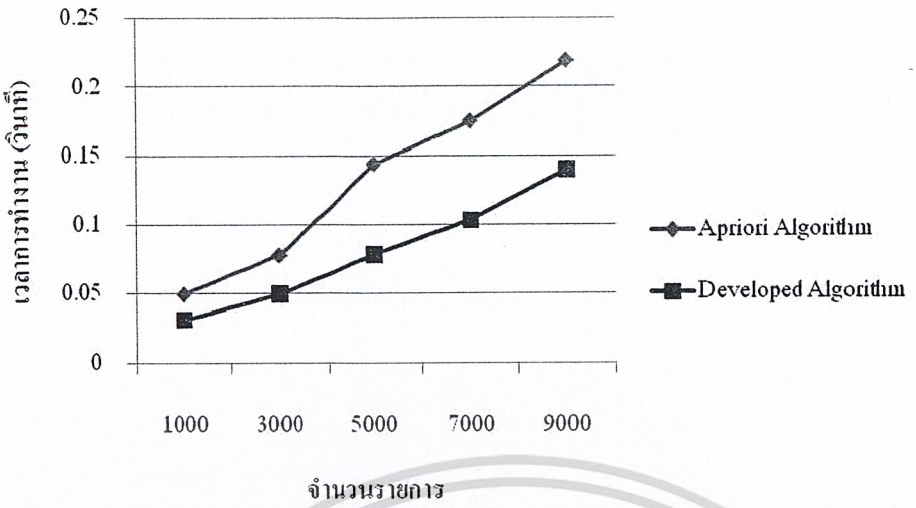
การทดลองเปรียบเทียบเวลาการทำงานในการหาไอเท็มเซตปรากฏบ่อย ของอัลกอริทึมที่ได้พัฒนาขึ้นกับอัลกอริทึมเอไพรออริ ที่จำนวนรายการต่างๆ โดยกำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 60% ได้ผลลัพธ์ดังตารางที่ 4.5

ตารางที่ 4.5 เวลาเฉลี่ยในการค้นหาไอเท็มเซตปรากฏบ่อยของในชุดข้อมูล Data Generator ในจำนวนรายการต่างๆ ที่ค่าสนับสนุนขั้นต่ำ 60%

อัลกอริทึม	เวลาการทำงาน (วินาที)				
	1000	3000	5000	7000	9000
Apriori Algorithm	0.0500	0.0780	0.1440	0.1750	0.2188
Developed Algorithm	0.0312	0.0500	0.0782	0.1032	0.1400

จากข้อมูลในตารางที่ 4.5 เมื่อนำข้อมูลที่ได้มาสร้างเป็นกราฟแสดงเวลาในการทำงานเทียบกับจำนวนรายการ สามารถแสดงได้ดังรูปที่ 4.4

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.4 เวลาเฉลี่ยในการค้นหาไอเท็มเซตปรากฏบ่อยของในชุดข้อมูล Data Generator ในจำนวนรายการต่างๆ ที่ค่าสนับสนุนขั้นต่ำ 60%

จากตารางที่ 4.5 และรูปที่ 4.4 แสดงเวลาในการทำงานของอัลกอริทึมเอไพรออริและอัลกอริทึมที่พัฒนาขึ้น โดยแสดงผลลัพธ์ในรูปแบบของเวลา (วินาที) เทียบกับจำนวนรายการที่แตกต่างกัน จะเห็นได้ว่าอัลกอริทึมเอไพรออริและอัลกอริทึมที่พัฒนาขึ้นใช้เวลาในการทำงานที่เพิ่มขึ้นเมื่อจำนวนรายการมากขึ้น และในค่าสนับสนุนขั้นต่ำที่ 60% อัลกอริทึมที่ได้พัฒนาขึ้นใช้เวลาในการหาไอเท็มเซตปรากฏบ่อยน้อยกว่าอัลกอริทึมเอไพรออริในทุกจำนวนรายการ

4.2.2 เปรียบเทียบเวลาการทำงานในชุดข้อมูล Mushroom

การทดลองทำโดยวิเคราะห์เปรียบเทียบประสิทธิภาพการทำงานของอัลกอริทึมที่พัฒนาขึ้นกับอัลกอริทึมเอไพรออริ โดยเปรียบเทียบ และวิเคราะห์ผลด้านเวลาการทำงานในการค้นหาไอเท็มเซตปรากฏบ่อยในชุดข้อมูล Mushroom ซึ่งผลลัพธ์ที่ได้แสดงเป็นตารางที่ 4.6

ตารางที่ 4.6 เวลาเฉลี่ยในการหาไอเท็มเซตปรากฏบ่อยของอัลกอริทึมเอไพรออริกับอัลกอริทึมที่พัฒนาขึ้นในข้อมูล Mushroom

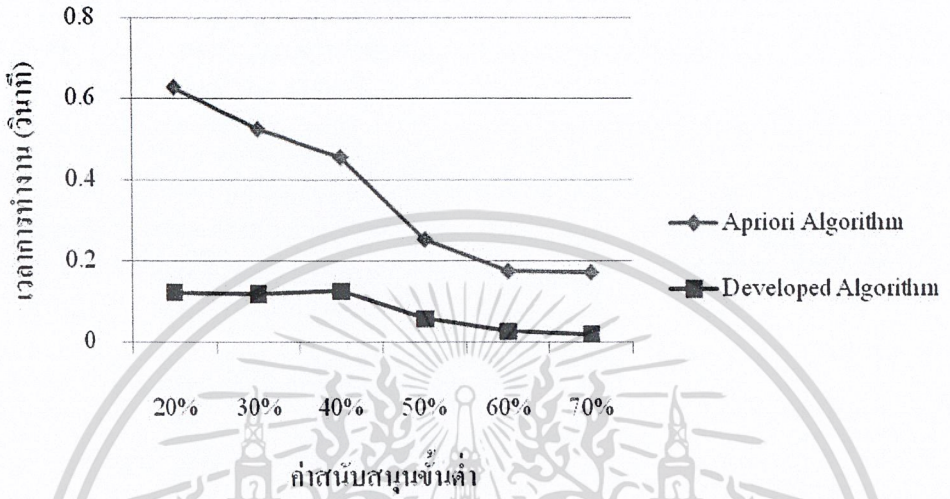
อัลกอริทึม	เวลาการทำงาน (วินาที)					
	20%	30%	40%	50%	60%	70%
Apriori Algorithm	0.6280	0.5248	0.4560	0.2532	0.1750	0.1718
Developed Algorithm	0.1218	0.1186	0.1250	0.0564	0.0250	0.0188

จากตารางที่ 4.6 แสดงผลลัพธ์การทำงานของอัลกอริทึมเอไพรออริ และอัลกอริทึมที่พัฒนาขึ้น โดยแสดงผลลัพธ์ในรูปแบบของเวลา (วินาที) พบว่าค่าสนับสนุนขั้นต่ำที่มีค่ามากขึ้นส่งผลต่อความเร็วในการทำงานของอัลกอริทึมเอไพรออริและอัลกอริทึมที่พัฒนาขึ้น โดยจะมีการ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ทำงานเร็วขึ้นและในชุดข้อมูล Mushroom อัลกอริทึมที่ได้พัฒนาขึ้นทำงานได้เร็วกว่าอัลกอริทึมเอไพรออริในทุกค่าสนับสนุนขั้นต่ำ

จากข้อมูลในตารางที่ 4.6 เมื่อนำข้อมูลที่ได้มาสร้างเป็นกราฟแสดงเวลาในการทำงานเทียบกับค่าสนับสนุนขั้นต่ำสามารถแสดงได้ดังรูปที่ 4.5



รูปที่ 4.5 กราฟแสดงเวลาการทำงานในการหาไอเท็มเซตปรากฏบ่อยของอัลกอริทึมในชุดข้อมูล Mushroom

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

ในการหากฎความสัมพันธ์ของข้อมูล แบ่งออกเป็น 2 ขั้นตอนคือ ขั้นตอนการหาไอเท็มเซตปรากฏบ่อยซึ่งนิยมใช้อัลกอริทึมเอไพรออริ และขั้นตอนการสร้างกฎความสัมพันธ์ โดยอัลกอริทึมเอไพรออรินั้นยังมีปัญหาคือ มีการอ่านข้อมูลจากฐานข้อมูลหลายครั้งและมีการสร้างแคนดิเดตไอเท็มเซตจำนวนมาก จึงทำให้การหากฎความสัมพันธ์ของข้อมูลนั้นใช้เวลานาน งานวิจัยนี้มีจุดมุ่งหมายที่จะพัฒนาอัลกอริทึมในการค้นหาไอเท็มเซตปรากฏบ่อย โดยขั้นตอนการวิจัยเริ่มจากศึกษาค้นคว้าจากการทำงานของอัลกอริทึมเอไพรออริและหาข้อบกพร่องของอัลกอริทึมมาปรับ เช่น ความล่าช้าในการวนอ่านข้อมูลจากฐานข้อมูล ของขั้นตอนการหาไอเท็มเซตปรากฏบ่อย และได้ศึกษาขั้นตอนการคำนวณทางตรรกะคณิตศาสตร์เพื่อนำมาใช้ร่วมในการหาไอเท็มเซตปรากฏบ่อย

การพัฒนาอัลกอริทึมในงานวิจัยนี้ได้นำข้อบกพร่องของอัลกอริทึมเอไพรออริมาปรับใหม่ คือการวนอ่านข้อมูลในฐานข้อมูลทุกครั้งที่มีการหาไอเท็มเซตปรากฏบ่อย ซึ่งทำให้ใช้เวลาประมวลผลนาน จึงได้พัฒนาให้มีการอ่านข้อมูลจากฐานข้อมูลเพียงครั้งเดียว โดยอ่านทีละรายการแล้วแปลงข้อมูลให้อยู่ในรูปของไบนารีเมทริกซ์ หลังจากนั้นทำการตัดไอเท็มที่มีค่าสนับสนุนน้อยกว่าค่าสนับสนุนขั้นต่ำออก แล้วพิจารณาไอเท็มเซตในแต่ละรายการว่าเป็นไอเท็มเซตปรากฏบ่อยหรือไม่ ถ้าไอเท็มเซตนั้นเป็นไอเท็มเซตปรากฏบ่อยเซตย่อยของไอเท็มเซตปรากฏบ่อยนั้นจะเป็นไอเท็มเซตปรากฏบ่อยด้วย แต่ถ้าไอเท็มเซตในรายการนั้นไม่ใช่ไอเท็มเซตปรากฏบ่อย ต้องพิจารณาเซตย่อยของไอเท็มเซตนั้นว่าเป็นไอเท็มเซตปรากฏบ่อยหรือไม่ เนื่องจากเซตย่อยของไอเท็มเซตนั้นมีโอกาสเป็นไอเท็มเซตปรากฏบ่อยด้วย

ในการเปรียบเทียบประสิทธิภาพการทำงานของอัลกอริทึมที่พัฒนาขึ้นจะนำมาเปรียบเทียบกับอัลกอริทึมเอไพรออริด้วยชุดข้อมูลรายการเสมือนจริง และข้อมูลจำแนกประเภท ซึ่งเกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพจะพิจารณาจากความถูกต้องและความเร็วในการทำงานของขั้นตอนการหาไอเท็มเซตปรากฏบ่อย

จากผลการทดลองแสดงให้เห็นว่า การเปลี่ยนวิธีการอ่านข้อมูลใหม่โดยอ่านเพียงครั้งเดียวแล้วเก็บข้อมูลในรูปแบบเมทริกซ์ และใช้หลักทางตรรกะคณิตศาสตร์มาร่วมในการคำนวณ ทำให้เวลาที่ใช้ในการหาไอเท็มเซตปรากฏบ่อยของอัลกอริทึมที่พัฒนาขึ้นสามารถทำงานได้เร็วกว่า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

อัลกอริทึมเอไพรออริในค่าสนับสนุนขั้นต่ำที่มีค่าสูง ตั้งแต่ค่าสนับสนุนขั้นต่ำ 40% ขึ้นไปในชุดข้อมูลแบบ Transactional Data และพบว่าผลต่างของเวลาในการทำงานระหว่างอัลกอริทึมที่ได้พัฒนาขึ้นกับอัลกอริทึมเอไพรออริมีค่ามากขึ้นเมื่อจำนวนรายการมีมากขึ้น สำหรับชุดข้อมูลแบบ Categorical Data นั้น อัลกอริทึมที่ได้พัฒนาขึ้น สามารถทำงานได้เร็วกว่าอัลกอริทึมเอไพรออริในทุกค่าสนับสนุนขั้นต่ำ เนื่องจากข้อมูลประเภทนี้มีรายการที่มีไอเท็มเหมือนกันอยู่มาก ซึ่งมีลักษณะที่เหมาะกับอัลกอริทึมที่ได้พัฒนาขึ้น

5.2 ข้อเสนอแนะ

สำหรับอัลกอริทึมเอไพรออริซึ่งเป็นอัลกอริทึมที่ใช้ในขั้นตอนหนึ่งในการค้นหากฎความสัมพันธ์ของข้อมูลนั้น ได้เป็นที่ยอมรับและมีการใช้งานจริงอย่างแพร่หลาย อีกทั้งมีผู้พัฒนาอัลกอริทึมในการหาไอเท็มเซตปรากฏบ่อยหลายวิธี ดังนั้นจึงเป็นการยากที่จะพัฒนาอัลกอริทึมขึ้นใหม่ให้แตกต่างและมีประสิทธิภาพที่ดีกว่าเดิม ในงานวิจัยนี้ได้พบข้อบกพร่องและข้อเสียของอัลกอริทึมที่ได้พัฒนาขึ้น ซึ่งอาจจะเป็นประโยชน์ต่อนักวิจัยท่านอื่นๆ ในการปรับปรุงและพัฒนาอัลกอริทึมค้นหาไอเท็มเซตปรากฏบ่อยต่อไป โดยข้อจำกัดของอัลกอริทึมที่ได้พัฒนาขึ้น มีดังนี้

1. อัลกอริทึมที่ได้ปรับปรุงขึ้นนั้นได้ทำการแปลงข้อมูลให้อยู่ในรูปของเมทริกซ์ ซึ่งในขั้นตอนการแปลงให้เป็นเมทริกซ์หากข้อมูลมีปริมาณมาก ก็จะใช้เวลาในการแปลงข้อมูลนานมากตามไปด้วย อีกทั้งการทำให้ข้อมูลอยู่ในรูปของเมทริกซ์นี้ยังเป็นการใช้ทรัพยากรของเครื่องคอมพิวเตอร์มากด้วย และหากข้อมูลมีจำนวนเยอะเกินไปก็ไม่สามารถนำมาทดสอบได้

2. ในการหาไอเท็มเซตปรากฏบ่อยโดยการพิจารณาไอเท็มเซตทีละรายการนั้น ถ้าไอเท็มเซตในรายการใดที่ไม่ได้เป็นไอเท็มเซตปรากฏบ่อย จำเป็นต้องทำการพิจารณาเซตย่อยของไอเท็มเซตนั้นด้วยว่ามีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำหรือไม่ ทำให้เสียเวลาในการพิจารณาเซตย่อย เหล่านี้ด้วย

3. ในการเปรียบเทียบเวลาโดยใช้โปรแกรมที่ได้สร้างขึ้นมานั้น โปรแกรมจะทำงานติดต่อกัน 18 ครั้ง (จากการเปรียบเทียบค่าสนับสนุนขั้นต่ำ 3 ค่า ระหว่างอัลกอริทึมเอไพรออริกับอัลกอริทึมที่ได้ปรับปรุง และแต่ละค่านั้นมาจากการเฉลี่ย 3 ครั้ง) ทำให้เครื่องคอมพิวเตอร์ต้องทำงานหนัก จึงอาจทำให้ผลการทดลองคลาดเคลื่อนได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารอ้างอิง

[1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases In: Proc. of the l993 ACM on Management of Data, Washington, D.C, May 1993.

[2] Yiwu Xie, Yutong Li, Chunli Wang, Mingyu Lu, "The Optimization and Improvement of the Apriori Algorithm ",Inf. Sci. Technol. Dept., Dalian Maritime Univ., Dalian, 2008.

[3] Feng WANG, Yong-hua LI, "An improved Apriori Algorithm based on the matrix", Wuhan University of Technology Wuhan, China, 2008.

[4] IBM Almaden Group, "Quest Synthetic Data Generation Code", http://www.almaden.ibm.com/cs/ projects/iis/hdb/Projects/data_mining/datasets/syndata.html

[5] วีระศักดิ์ ซึงถาวร. 2549. Java Programming VolumeII. พิมพ์ครั้งที่ 2. กรุงเทพฯ : ซีเอ็ดยูเคชั่น

[6] วรเศรษฐ สุวรรณิก. 2551. Java GUI using NetBeans. พิมพ์ครั้งที่ 1. กรุงเทพฯ : ซีเอ็ดยูเคชั่น

[7] [Online].Available : http://download.oracle.com

[8] [Online].Available : http://www.java2s.com

[9] [Online].Available : http://code.function.in.th/java-swing

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภาคผนวก

การใช้งานโปรแกรม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

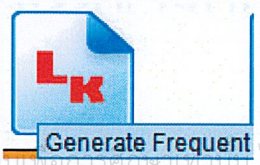
Association Rules Mining

Association Rules Mining เป็น Windows Application ที่ถูกสร้างขึ้นจากภาษา Java เพื่อใช้ในการสร้างไอเท็มเซตปรากฏบ่อย (Frequent itemset) สร้างกฎความสัมพันธ์และสร้างกราฟสำหรับเปรียบเทียบผลในการทดลอง



รูปที่ ก.1 หน้าจอหลักของโปรแกรม

หน้าหลักของโปรแกรม (Main Menu) มีลักษณะดังรูป ก.1 ประกอบด้วยกลุ่มไอคอน (Icon) ที่ผู้ใช้ สามารถกดเพื่อควบคุมการทำงานต่างๆ ของโปรแกรม เมื่อผู้ใช้นำเมาส์ไปชี้ที่ไอคอนจะปรากฏคำอธิบายของไอคอนออกมา ดังรูป ก.2



เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งไม่อนุญาตให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

รูปที่ ก.2 แสดงคำอธิบายของไอคอนเมื่อผู้ใช้นำเมาส์ไปวางบนไอคอน

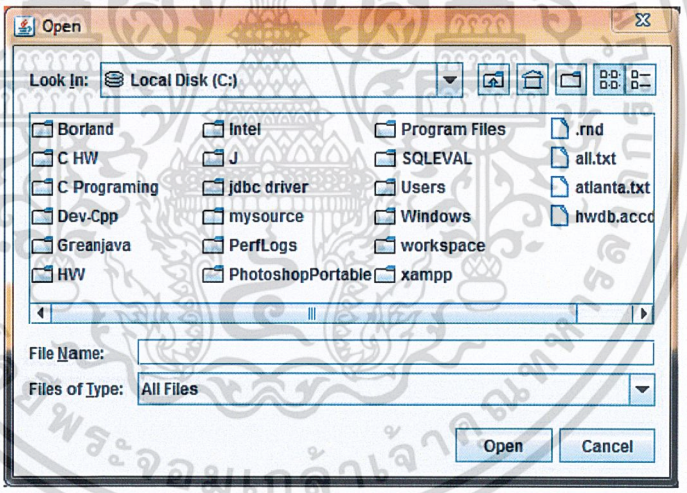
โดยไอคอนทั้งหมด 5 ปุ่ม ประกอบด้วย

1. ไอคอน Browse File
2. ไอคอน Generate Frequent itemset
3. ไอคอน Generate Rules
4. ไอคอน View Graph
5. ไอคอน Instruction

สำหรับแต่ละไอคอนสามารถอธิบายได้ดังนี้

1. ไอคอน Browse File

เป็นไอคอนที่ใช้สำหรับเปิดไฟล์ที่ผู้ใช้ต้องการนำมาประมวลผลก่อนที่จะเข้าสู่กระบวนการใช้งานอื่นๆ ของโปรแกรม โดยไฟล์ที่โปรแกรมรองรับเป็นไฟล์ที่มีนามสกุล .txt หรือ .dat เท่านั้น เมื่อผู้ใช้ทำการคลิกจะปรากฏหน้าต่าง pop up ขึ้นมาบนหน้าจอเพื่อให้ผู้ใช้ทำการเลือกไฟล์ที่ต้องการเปิดใช้งาน ดังรูป ก.3



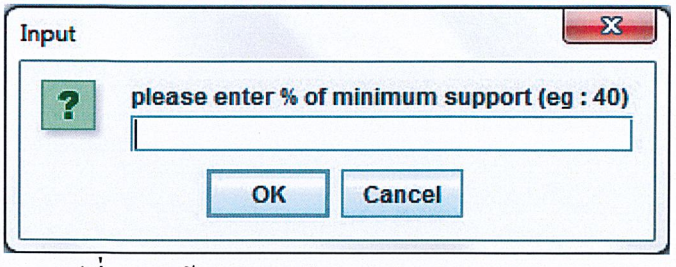
รูปที่ ก.3 หน้าต่างเลือกไฟล์ที่ผู้ใช้ต้องการเปิด

2. ไอคอน Generate Frequent itemset

เป็นไอคอนที่ใช้ในการสร้างไอเท็มเซตปรากฏบ่อย จากไฟล์ที่ผู้ใช้ได้ทำการเปิด ดังนั้นไอคอนนี้ผู้ใช้จะสามารถเข้าใช้ได้ก็ต่อเมื่อผู้ใช้ได้ทำการเปิดไฟล์แล้วเท่านั้น สำหรับใช้งานหลังจากที่ผู้ใช้ได้ทำการเปิดไฟล์แล้ว เมื่อคลิกก็จะปรากฏหน้าต่าง pop up ขึ้นมาเพื่อให้ผู้ใช้ทำการกรอกค่า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

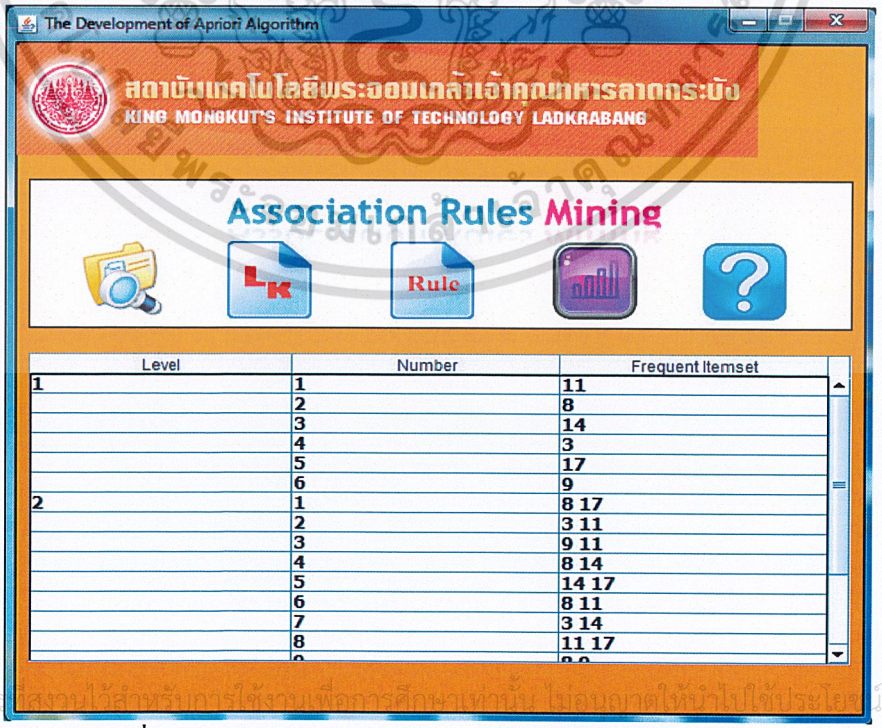
minimum support เพื่อนำมาใช้ในการประมวลผล ดังรูป ก.4 โดยค่า minimum support ที่ผู้ใช้สามารถกรอกได้จะอยู่ในรูปแบบตัวเลขเป็นเปอร์เซ็นต์ (%) ซึ่งมีช่วงตั้งแต่ 1-100



รูปที่ ก.4 หน้าต่างสำหรับระบุค่า minimum support

เมื่อผู้ใช้กรอกค่า minimum support เรียบร้อยและกดปุ่ม ok เพื่อยืนยันแล้ว ระบบก็จะทำการประมวลผลทันที และแสดงผลลัพธ์ไอเท็มเซตปรากฏบ่อยที่ได้จากการประมวลผลบนหน้าจอในรูปแบบตาราง ดังรูป ก.5 โดยตารางแสดงผลลัพธ์ในการหาไอเท็มเซตปรากฏบ่อยประกอบด้วย 3 คอลัมน์ คือ

1. คอลัมน์ Level อธิบายถึงระดับของแต่ละไอเท็มเซตปรากฏบ่อย
2. คอลัมน์ Number อธิบายถึงลำดับของไอเท็มเซตปรากฏบ่อยแต่ละตัวที่ปรากฏในแต่ละระดับ
3. คอลัมน์ Frequent itemset อธิบายถึงไอเท็มเซตปรากฏบ่อยแต่ละตัวที่ปรากฏในแต่ละระดับ

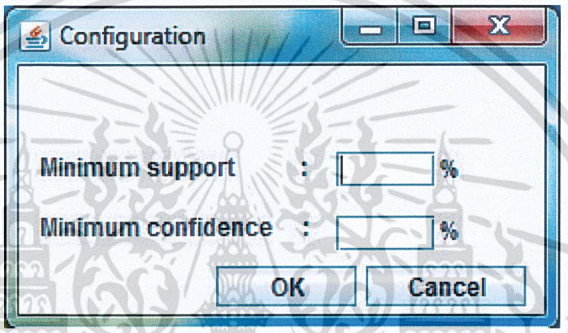


รูปที่ ก.5 ตารางแสดงผลลัพธ์ในการหาไอเท็มเซตปรากฏบ่อย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3. ไอคอน Generate Rules

เป็นไอคอนที่ใช้ในการสร้างกฎความสัมพันธ์จากไฟล์ที่ผู้ใช้ได้ทำการเปิด ดังนั้นไอคอนนี้ผู้ใช้จะสามารถเข้าใช้ได้ก็ต่อเมื่อผู้ใช้ได้ทำการเปิดไฟล์แล้วเท่านั้น สำหรับใช้งานหลังจากที่ผู้ใช้ได้ทำการเปิดไฟล์แล้ว เมื่อคลิกก็จะปรากฏหน้าต่าง pop up ขึ้นมาเพื่อให้ผู้ใช้ทำการเปิด text file (*.txt,*.dat) ที่เก็บรายชื่อของ itemset และระบุค่า minimum support และ minimum confidence เพื่อนำมาใช้ในการประมวลผล โดยค่าทั้งสองจะอยู่ในรูปแบบเปอร์เซ็นต์ (%) ซึ่งผู้ใช้ต้องกรอกในข้อมูลรูปแบบตัวเลขตั้งแต่ 1-100 การค้นหากฎความสัมพันธ์ของข้อมูลที่มีอยู่ในไฟล์ดังรูป ก.6



รูปที่ ก.6 หน้าต่างสำหรับระบุค่า minimum support และ minimum confidence

เมื่อกรอกค่า minimum support และ minimum confidence เรียบร้อยแล้วระบบจะทำการประมวลผลทันที และแสดงกฎความสัมพันธ์ที่ได้จากการประมวลผลออกมาแสดงบนหน้าจอ หน้าจอ เมื่อผู้ใช้กดที่ปุ่ม refresh โปรแกรมจะแสดงกฎความสัมพันธ์ที่ได้จากการประมวลผลออกมาแสดงบนหน้าจอให้ผู้ใช้เห็นในรูปแบบตาราง ดังรูป ก.7 โดยตารางที่แสดงผลลัพธ์ในการสร้างกฎความสัมพันธ์ ประกอบด้วย คอลัมน์ต่างๆ ดังนี้

1. คอลัมน์ Rules แสดงถึงกฎความสัมพันธ์แต่ละกฎ
2. คอลัมน์ Support แสดงถึงค่า support ของกฎความสัมพันธ์
3. คอลัมน์ Confidence แสดงถึงค่า confidence ของกฎความสัมพันธ์
4. คอลัมน์ Correlation แสดงถึงค่า correlation ของกฎความสัมพันธ์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



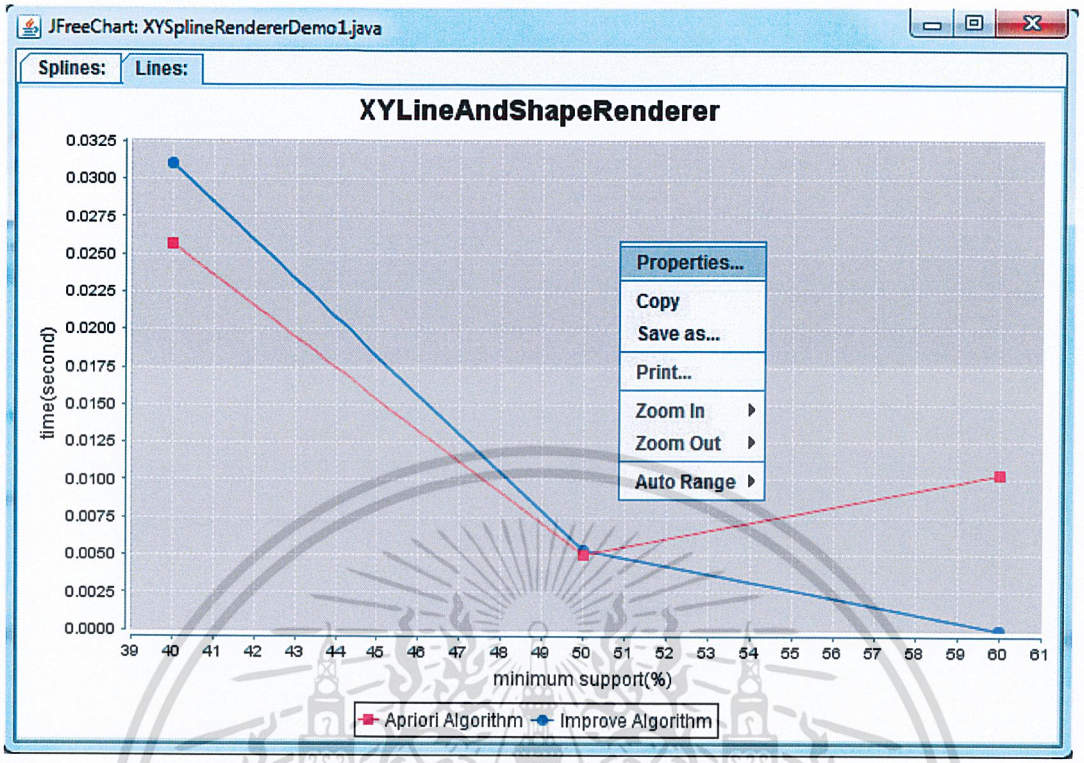
รูปที่ ก.7 ตารางแสดงผลลัพธ์ในการหากฎความสัมพันธ์

ผู้ใช้สามารถทำการบันทึกข้อมูลกฎความสัมพันธ์ โดยคลิกที่ปุ่ม Save ในมุมด้านขวาล่างของ Main Menu โปรแกรมจะแสดงหน้าต่างให้ผู้ใช้เลือก directory และตั้งชื่อไฟล์ในการบันทึก สำหรับการบันทึกกฎความสัมพันธ์ ข้อมูลจะถูกบันทึกในรูปแบบไฟล์ Excel (*.xls)

4. ไอคอน View Graph

เป็นไอคอนที่ใช้สำหรับการดูกราฟในการเปรียบเทียบเวลาและ minimum support ระหว่างอัลกอริทึมเอไพรออริกับอัลกอริทึมที่พัฒนาขึ้น โดยผู้ใช้จะต้องทำการสร้างไอเท็มเซตปรากฏบ่อยก่อน จึงจะสามารถดูกราฟได้ ตัวอย่างกราฟแสดงให้เห็นดังรูป ก.8

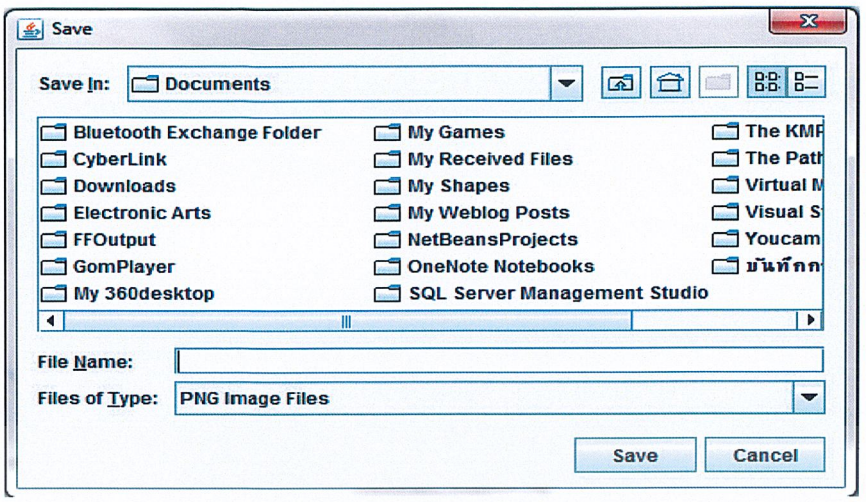
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ ก.8 กราฟเปรียบเทียบเวลาและค่าสนับสนุนขั้นต่ำระหว่างอัลกอริทึมทั้งสอง

ผู้ใช้สามารถกำหนดรูปแบบของกราฟที่ใช้ในการแสดงผล โดยแบ่งออกเป็น 2 ประเภท ได้แก่ กราฟเส้นโค้ง (Splines) และกราฟเส้นตรง (Lines) นอกจากนี้เมื่อผู้ใช้คลิกเมาส์ขวาที่หน้าต่างกราฟจะปรากฏเมนูให้ผู้ใช้เลือกตามรูป ก.8 ผู้ใช้สามารถกำหนดลักษณะทั่วไปของการแสดงผลเช่นชื่อกราฟ สีของเส้นกราฟได้จากเมนู Properties และสามารถคัดลอกรูปภาพไปวางในแฟ้มแอกสารได้โดยคลิก copy และสั่งพิมพ์รูปกราฟได้โดยตรงจากเมนู print นอกจากนี้ผู้ใช้สามารถบันทึกกราฟได้ โดยคลิกขวาที่รูปกราฟแล้วเลือก Save as จะปรากฏหน้าต่างดังรูป ก.9 เพื่อให้ผู้ใช้ระบุ directory สำหรับบันทึกรูปกราฟ รูปกราฟที่ถูกบันทึกจะเป็นไฟล์นามสกุล .PNG นอกจากนี้ผู้ใช้ยังสามารถทำการย่อหรือขยายขนาดของกราฟได้จากเมนู Zoom In และ Zoom Out รวมทั้งระบุช่วงกราฟที่ผู้ใช้ต้องการดูได้ด้วยเมนู Auto Range

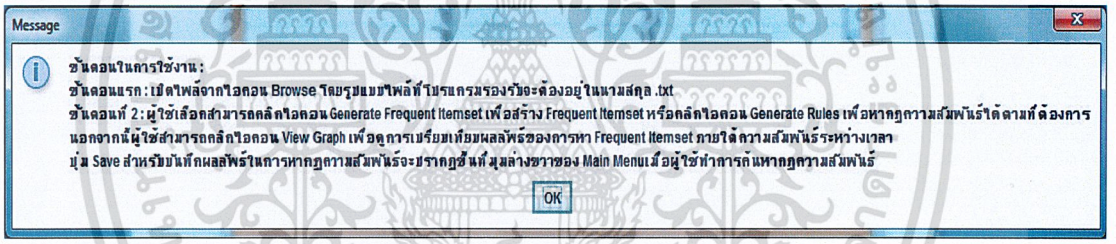
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ ก.9 หน้าต่างบันทึกกราฟ

5. ไอคอน Instruction

เมื่อคลิกจะปรากฏหน้าต่าง pop up ขึ้นมายังหน้าจอเพื่อให้คำอธิบายคร่าวๆในการใช้งานในส่วนต่างๆของโปรแกรม ดังรูป ก.10



รูปที่ ก.10 หน้าต่างคำอธิบายการใช้งานโปรแกรม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้