บทที่ 4

การทดลองและผลการทดลอง

4.1 การวัดประสิทธิภาพการทำงานของโมเดล

การวัดประสิทธิภาพการทำงานของโมเดล ได้ใช้ตัววัดประสิทธิภาพที่ใช้ในงานวิจัยทั่วไป คือการวัดความเร็วในการค้นหากฎความสัมพันธ์ด้วย min_sup ที่แตกต่างกัน โดยจะทำการวัดประสิทธิภาพการค้นหากฎความสัมพันธ์ระหว่างข้อมูลเทียบระหว่างอัลกอริทึมที่เคยมีและอัลกอริทึม probability-based incremental association rule discovery algorithm

การสร้างข้อมูลเพื่อใช้ในการทดลอง ด้วย Synthetic data generation

การทดลองจะใช้ Synthetic data ซึ่งเป็นเทคนิคเดียวกับ Agrawal and Srikant 1994 ซึ่งใช้หลักทางสถิติมาประยุกต์ใช้เพื่อสร้างข้อมูลเลียนแบบข้อมูลจริงที่เกิดจากการเลือกซื้อสินค้าในลักษณะที่เรียกว่า market-basket

ชุดข้อมูลที่ใช้ในการทดลองจะได้จาก synthetic data ที่ใช้คือ T10.I4.D10K (ขนาดเฉลี่ยของ transaction = 10, ขนาดเฉลี่ยของ maximal potentially large item sets = 4 และจำนวนของ transaction = 10,000) ประกอบด้วย 2 ส่วน ส่วนแรกเป็นข้อมูลที่ได้จากการสุ่มจำนวน 100,000 transaction เพื่อเป็นข้อมูลที่ในการค้นหากฎความสัมพันธ์ของฐานข้อมูลเดิม (original database) และส่วนที่ 2 เป็นส่วนของ incremental database จำนวน 100,000 transaction โดยจะทำการเพิ่มจำนวน incremental database ไปทีละ 10 % ของข้อมูลเดิมจนกว่าจะครบ 100% ขนาดของข้อมูลที่ใช้เพิ่มเข้าไปในฐานข้อมูลเดิมนี้จะนำมาใช้ในการพิจารณาความถูกต้องของกฎความสัมพันธ์ใหม่ที่ได้จากการปรับปรุงฐานข้อมูล และวัดประสิทธิภาพของอัลกอริทึมเมื่อนำไปใช้ในการค้นหากฎความสัมพันธ์เมื่อมีการเพิ่มข้อมูลใหม่เข้าไป

วัตถุประสงค์ในการทดลอง

1. เพื่อทดสอบการไมน์นิ่ง Frequent itemset ที่ถูกต้องหลังจากเพิ่ม incremental database เข้าไปปรับปรุง original database โดยมีการนำความรู้ที่ได้จากการไมน์นิ่งใน original database มาใช้ให้เกิดประโยชน์เพื่อลดจำนวนครั้งและจำนวน itemset ในการสแกน original database
2. เพื่อวัดประสิทธิภาพในการเพิ่มขยายการค้นหากฎความสัมพันธ์ของอัลกอริทึม(ใหม่) ในกรณีของการเพิ่ม transaction เข้าไปในฐานข้อมูล

วิธีการทดลอง

1. **ทดสอบความถูกต้องของอัลกอริทึม**

ในการทดสอบความถูกต้องของข้อมูลที่ได้จากการเพิ่ม transaction ใหม่เข้าไปในฐานข้อมูล โดยเปรียบเทียบกับกฎความสัมพันธ์ที่ได้จากการ rerun Apriori เมื่อมีการเพิ่ม transaction ใหม่เข้าไปในฐานข้อมูล

2. **ทดสอบประสิทธิภาพของอัลกอริทึม**

การทดสอบประสิทธิภาพของอัลกอริทึม ซึ่งทำการทดสอบเปรียบเทียบผลกับ 3 อัลกอริทึม FUP, Border,Probability-based โดยใช้ตัวชี้วัด 3 ตัว คือ เวลา, ขนาดของข้อมูลที่เพิ่ม และจำนวน candidate itemset

minimum support threshold

การค้นหากฎความสัมพันธ์ โดยกำหนดค่า Minimum support ในระดับต่างๆกัน ระหว่างค่า minimum support threshold 1.0 – 3.0%

Effect of size of updates

วัดความเร็วเมื่อมีการเพิ่มขนาดของ Incremental database เข้าไปในฐานข้อมูลจาก 10 - 100 % ด้วยค่า minimum support threshold ที่คงที่ โดยพิจารณาจาก execution time ที่ใช้ในการ run เพื่อหากฎความสัมพันธ์ที่ได้จากการปรับปรุง

Varying the number of added transactions independently

การทดลองนี้จะใช้สำหรับหาขนาดของ Transaction ที่เพิ่มเข้าไปใน original database ที่มีผลกับ performance ของอัลกอริทึม โดยใช้ T10.I4.D10 +10 ในการทดลองจะกำหนด initial database 10,000 transactions และทำการเพิ่มทีละ 1,000 transactions เข้าไปในฐานข้อมูล

4.2 ชุดข้อมูลที่ใช้ในการทดลองและผลการทดลอง

การประเมินประสิทธิภาพของ Probability-based incremental association rule discovery algorithm ได้ทำการทดลองด้วยเครื่องคอมพิวเตอร์ Pentium 4 หน่วยความจำหลัก 1 GB ชุดข้อมูลที่ใช้ในการทดลองคือ synthetic dataset จำนวน 110,000 transactions ที่ประกอบด้วย unique items 100 items ด้วยค่าเฉลี่ยของ items ที่ปรากฏในแต่ละ transaction คือ 10 items และ maximal size itemset คือ 4 itemsets นำสุ่มเลือกเพื่อสร้างเป็น original database 1 ชุด ขนาด 10000 transactions และ incremental database จำนวน 1000 ชุด ขนาด 1000 transactions

การทดลองเริ่มจากขั้นตอนการค้นหากฎความสัมพันธ์จาก original database จำนวน 10000 transactions ดังตารางที่ 4.1 แสดงจำนวน itemset ของ original database ที่จัดเก็บในแต่ละอัลกอริทึม

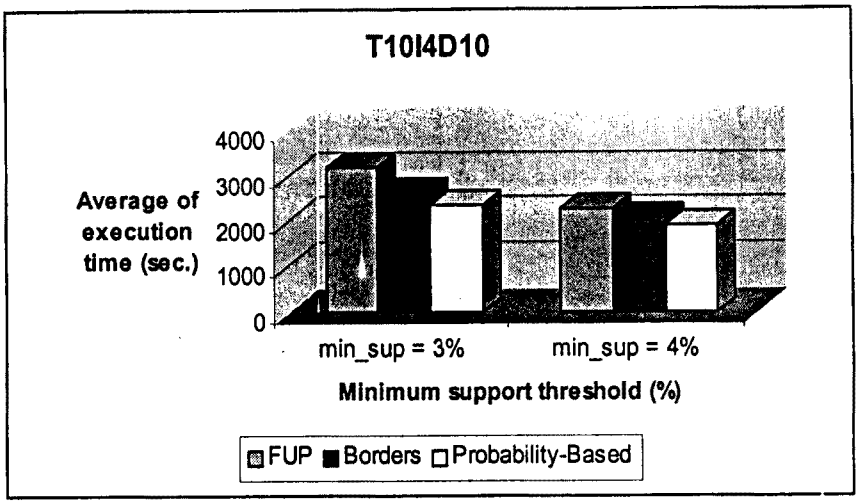
จากนั้นทำการเพิ่ม incremental database ในขนาด 10% ของ original database ในที่นี้คือ 1000 transactions เข้าไปใน original database และทำการทดลองจำนวน 100 ครั้ง โดยเปรียบเทียบการทดลองกับ 2 อัลกอริทึมคือ FUP และ Border ด้วยค่า minimum support 3% และ 4% ดังแสดงใน ตารางที่ 4.2 และรูปที่ 4.1

ตารางที่ 4.1 จำนวน itemset ของ original database สำหรับแต่ละอัลกอริทึม

Min_sup (%)	Algorithm	Number of Frequent k-itemset			Number of Infrequent k-itemset		
		k=1	k=2	k=3	k=1	k=2	k=3
3%	FUP	97	263	9	0	0	0
	Borders	97	263	9	3	4393	371
	Probability-Based	97	263	9	0	25	5
4%	FUP	93	74	1	0	0	0
	Borders	93	74	1	7	4204	26
	Probability-Based	93	74	1	1	8	0

ตารางที่ 4.2 ค่าเฉลี่ยของเวลาที่ใช้ในการประมวลผล

	Average of Execution time for 100 trials		
min_sup (%)	FUP	Borders	Probability- Based
3%	3195.6939	2660.9297	2354.24533
4%	2274.4477	2087.8636	1931.19147



รูปที่ 4.1 กราฟเปรียบเทียบเวลาที่ใช้ในการประมวลผลของอัลกอริทึม FUP, Borders และ Probability-based ด้วยค่า minimum support 3% และ 4%

จากผลการทดลองพบว่า Probability-base incremental association rule discovery algorithm มีผลการทำงานที่ดีกว่าอัลกอริทึม FUP และ Border