



เทคนิคการเก็บ ไอเท็มเซตที่เกิดขึ้นบ่อย

โดยพิจารณาค่าความเชื่อมั่นขั้นต่ำ เพื่อรองรับการเพิ่มของข้อมูล

กฤษฎากร กังอุบล* ธนาวิทย์ รักธรรมานนท์* และ กฤษณะ ไวยมัย*

บทคัดย่อ

กฎความสัมพันธ์ (Association rule) ที่มีความสำคัญ คือ กฎที่ผ่านค่าสนับสนุนขั้นต่ำ (Minimum support) และค่าความเชื่อมั่นขั้นต่ำ (Minimum confidence) งานวิจัยในอดีตจึงมีการเสนอเทคนิคที่มีประสิทธิภาพสำหรับค้นหากฎที่มีความสำคัญในฐานข้อมูลขนาดใหญ่ แต่ยังคงมีปัญหของการปรับปรุงกฎเมื่อมีการเพิ่มของข้อมูล เนื่องจากการเพิ่มของข้อมูลไม่เพียงแต่ทำให้ไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำเปลี่ยนเป็นไม่ผ่านค่าสนับสนุนขั้นต่ำ แต่กลับทำให้ไอเท็มเซตที่ไม่ผ่านค่าสนับสนุนขั้นต่ำเปลี่ยนเป็นผ่านค่าสนับสนุนขั้นต่ำ

งานวิจัยชิ้นนี้จึงมีการเสนอเทคนิคการเก็บไอเท็มเซตที่เกิดขึ้นบ่อย เพื่อรองรับการเพิ่มของข้อมูล (Fast Update of Frequent Itemsets, FUFU) เทคนิค FUFU จะพิจารณาค่าความเชื่อมั่นขั้นต่ำและค่าขอบเขตล่าง (Lower bound) เป็นเกณฑ์ในการเลือกไอเท็มเซตที่มีโอกาสมากที่สุดที่จะผ่านค่าสนับสนุนขั้นต่ำในรอบถัดไป ไอเท็มเซตที่มีโอกาสมากที่สุดที่จะผ่านค่าสนับสนุนขั้นต่ำในรอบถัดไปคือไอเท็มเซตที่เป็นตัวแทนของไอเท็มเซตทั้งหมดที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำในรอบถัดไป ดังนั้นเทคนิคการคำนวณของ FUFU จึงลดขั้นตอนการค้นหาในฐานข้อมูลขนาดใหญ่เมื่อมีการเพิ่มของข้อมูล และยังคงสามารถปรับปรุง เพิ่มเติม ไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำเมื่อมีข้อมูลเพิ่มเข้ามาจากการทดลองกับชุดข้อมูลที่เกาะกลุ่ม (Dense datasets) แสดงให้เห็นว่าเทคนิค FUFU มีประสิทธิภาพด้านเวลาประมวลผลและการลดจำนวนไอเท็มเซตทั้งหมดที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำในรอบถัดไป เมื่อเทียบกับเทคนิค FUP และ NUWEP

คำสำคัญ: ค่าสนับสนุนขั้นต่ำ ค่าความเชื่อมั่นขั้นต่ำ ฟรีควเอนไอเท็มเซต การค้นหากฎความสัมพันธ์

1. บทนำ

การค้นพบกฎความสัมพันธ์ (Association Rule Discovery) คือเทคนิคอย่างหนึ่งของการทำเหมืองข้อมูล ที่ได้จากการนับความถี่ของชุดข้อมูลที่เกิดในดาต้าเซต (Datasets) เรียกว่า ไอเท็มเซต (itemsets) แล้วนำไอเท็มเซตเหล่านี้ มาสร้างกฎที่มีความสำคัญ

ในการสร้างกฎที่มีความสำคัญ จะใช้ไอเท็มเซตที่มีความถี่ผ่านค่าสนับสนุนขั้นต่ำ เรียกว่า ฟรีควเอนไอเท็มเซต (Frequent itemsets) เป็นตัวสร้างกฎที่มีความสำคัญ ฉะนั้นสิ่งสำคัญในการหากฎความสัมพันธ์คือ การนับค่าความถี่หรือค่าสนับสนุน (support) ของไอเท็มเซต ซึ่งเป็นข้อมูลสำคัญที่จะใช้พิจารณาว่ากฎใด เป็นกฎที่มีความสำคัญ เทคนิคแรกที่ใช้ค้นหาฟรีควเอน ไอเท็มเซต จะรู้จักในชื่อของอัลกอริทึม Apriori [1, 2, 3, 4]

การค้นหากฎความสัมพันธ์เมื่อมีการเพิ่มของข้อมูล จะต้องมีการปรับปรุงค่าสนับสนุนของไอเท็มเซตให้สอดคล้องกับการเพิ่มของข้อมูล โดยการปรับปรุงค่าสนับสนุนของไอเท็มเซตเมื่อมีการเพิ่มของข้อมูล จะต้องเข้าไปนับค่าสนับสนุนใหม่ในข้อมูลเก่า ซึ่งมีข้อเสียสำหรับการนับค่าสนับสนุนในข้อมูลเก่า คือใช้เวลามากเนื่องจากข้อมูลเก่ามีขนาดใหญ่ และเมื่อมีการเพิ่มของข้อมูล จะต้องมีการนับค่าสนับสนุนในข้อมูลเก่าทุกครั้ง

แนวทางเริ่มแรก มุ่งเน้นที่จะลดจำนวนไอเท็มเซตของข้อมูลที่เพิ่มเข้ามา ซึ่งจะนำไปเพื่อปรับปรุงค่าสนับสนุนในข้อมูลเก่า หรือจำนวนไอเท็มเซตของข้อมูลเก่าที่ไม่ได้เก็บไว้ เนื่องจากไม่ผ่านค่าสนับสนุนขั้นต่ำ แต่กลับกลายเป็นไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำเมื่อมีการเพิ่มของข้อมูล จึงจำเป็นต้องนับค่าสนับสนุนใหม่ เพื่อนำไปปรับปรุงค่าสนับสนุนล่าสุดของข้อมูลเก่า แต่แนวทางนี้มีข้อเสียคือจะต้องเข้าไปนับค่าสนับสนุนของข้อมูลเก่าซึ่งมีขนาดใหญ่ในทุกๆ ความยาว

* ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

ของไอเท็มเซต ซึ่งใช้เวลามาก จะรู้จักในชื่อของอัลกอริทึม FUP [5], FUP2 [6]

แนวทางถัดมา มุ่งเน้นที่จะลดจำนวนครั้งที่จะนับค่าสนับสนุนในข้อมูลเก่าเมื่อมีการเพิ่มของข้อมูล โดยมีการเก็บไอเท็มเซตที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำเมื่อมีการเพิ่มของข้อมูลโดยใช้ค่าขอบเขตล่าง (Lower bound) สำหรับการนับครั้งแรก และเมื่อมีการเพิ่มของข้อมูล จะนำไอเท็มเซตเหล่านี้ที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำไปใช้ปรับปรุงค่าสนับสนุนของฐานข้อมูลเก่า โดยไม่ต้องกลับมานับค่าสนับสนุนในข้อมูลเก่าอีกครั้ง ซึ่งมีขนาดใหญ่และใช้เวลามาก แต่ข้อเสียของวิธีนี้ จะมีการเก็บไอเท็มเซตที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำเมื่อมีการเพิ่มของข้อมูล ไว้เป็นจำนวนมาก ทำให้เพิ่มเวลาในการนับค่าสนับสนุนของการนับครั้งแรก วิธีการนี้จะรู้จักในชื่อของ Negative border [7], NUWEP [8]

จากข้อจำกัดที่ผ่านมา งานวิจัยชิ้นนี้จึงเสนอวิธีการคัดเลือกจำนวนไอเท็มเซตที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำเมื่อมีการเพิ่มของข้อมูลให้น้อยลง โดยใช้ค่าขอบเขตล่างและค่าความเชื่อมั่นขั้นต่ำ (Minimum confidence) เป็นเกณฑ์ในการเลือกไอเท็มเซตที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำเมื่อมีการเพิ่มของข้อมูล สำหรับการนับครั้งแรก เพื่อลดเวลาสำหรับการนับและจำนวนไอเท็มเซตที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำเมื่อมีการเพิ่มของข้อมูล โดยจากผลการทดลองกับข้อมูลสังเคราะห์ ชุดข้อมูลที่เกาะกลุ่ม (dense datasets) C10D10K.dat [9] การใช้ค่าขอบเขตล่าง และค่าความเชื่อมั่นขั้นต่ำ เป็นเกณฑ์ในการเลือกไอเท็มเซตที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำเมื่อมีการเพิ่มของข้อมูล จำนวนไอเท็มเซตและเวลาในการนับค่าสนับสนุนในข้อมูลเก่าสำหรับการนับครั้งแรกจะลดลง และไอเท็มเซตที่ถูกคัดเลือกยังคงครอบคลุมไอเท็มเซตทุกตัวที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำในการนับครั้งถัดไป

2. ปัญหาและงานวิจัยที่เกี่ยวข้อง

2.1 ปัญหา

การปรับปรุงค่าสนับสนุนของไอเท็มเซตเมื่อมีการเพิ่มของข้อมูล จะต้องมีการปรับปรุงค่าสนับสนุนของไอเท็มเซตให้สอดคล้องกับการเพิ่มของข้อมูล โดยการปรับปรุงค่า

สนับสนุนของไอเท็มเซตเมื่อมีการเพิ่มของข้อมูลในข้อมูลเก่าจะต้องเข้าไปนับค่าสนับสนุนใหม่ในข้อมูลเก่า ซึ่งมีข้อเสียสำหรับการนับค่าสนับสนุนในข้อมูลเก่า คือใช้เวลามากเนื่องจากข้อมูลเก่ามีขนาดใหญ่ และเมื่อมีการเพิ่มของข้อมูลจะต้องมีการนับค่าสนับสนุนในข้อมูลเก่าทุกครั้ง เพื่อง่ายต่อการเข้าใจ ดูตารางที่ 1 และตารางที่ 2 ประกอบ ถ้ามีฐานข้อมูลเก่า (DB) และข้อมูลใหม่ที่เพิ่มเข้ามา (db) การจะหาค่าสนับสนุนใหม่ของไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำจะต้องหาจากฐานข้อมูลเก่าและข้อมูลใหม่ที่เพิ่มเข้ามารวมกัน

การเข้าไปค้นหาค่าสนับสนุนจาก ฐานข้อมูลเก่า ทุกครั้งจะใช้เวลามากเพราะว่าฐานข้อมูลเก่ามีขนาดใหญ่ เมื่อเทียบกับการเข้าไปค้นหาค่าสนับสนุนจากข้อมูลใหม่ที่เพิ่มเข้ามาซึ่งมีขนาดเล็ก จึงเกิดวิธีที่เก็บค่าสนับสนุนของไอเท็มเซตในฐานข้อมูลเก่า ที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำ (minsup) ไว้ก่อนแล้วค่อยมานำรวมกับค่าสนับสนุนที่ได้จากข้อมูลใหม่ที่เพิ่มเข้ามา แต่วิธีที่จะเก็บค่าสนับสนุนของไอเท็มเซตของฐานข้อมูลเก่าไว้ก่อนเพื่อจะไปรวมค่าสนับสนุนไอเท็มเซตใหม่ของข้อมูลที่เพิ่มเข้ามากับจะเกิดปัญหาการปรับปรุงค่าสนับสนุนได้หลายกรณี ดังแสดงไว้ในตารางที่ 1

จากตารางที่ 1 สามารถพิจารณากรณีค่าสนับสนุนของไอเท็มเซตที่เกิดใน ฐานข้อมูลเก่า ข้อมูลใหม่ที่เพิ่มเข้ามา และผลลัพธ์ของค่าสนับสนุนใหม่ในฐานข้อมูลรวม

พิจารณาจากกรณีที่ 1: ไอเท็มเซต สามารถนำค่าสนับสนุนของ (DB) และ (db) มารวมเป็นค่าสนับสนุนใหม่ที่ผ่านค่าสนับสนุนขั้นต่ำ

พิจารณาจากกรณีที่ 2: ไอเท็มเซต สามารถนำค่าสนับสนุนของ (DB) และ (db) มารวมเป็นค่าสนับสนุนใหม่ซึ่งอาจจะผ่านหรือไม่ผ่านค่าสนับสนุนขั้นต่ำ

พิจารณาจากกรณีที่ 3: ไอเท็มเซตนี้ ค่าสนับสนุนใหม่ไม่ผ่านค่าสนับสนุนขั้นต่ำ จึงไม่ต้องสนใจ

พิจารณาจากกรณีที่ 4: ไอเท็มเซต ไม่สามารถนำค่าสนับสนุนของ (DB) มารวมกับ (db) เพราะค่าสนับสนุนของไอเท็มเซตไม่ได้เก็บไว้ จึงต้องมีการเข้าไปค้นหาในฐานข้อมูลเก่า อีกครั้งจากกรณีที่ 3 จึงเกิดวิธีการเก็บไอเท็มเซตที่ไม่ผ่านค่าสนับสนุนของฐานข้อมูล ไว้ก่อน ซึ่งจะกล่าวในหัวข้อถัดไป



ตารางที่ 1 แสดงกรณีของค่าสนับสนุนของ ไอเท็มเซตที่เป็นไปได้

กรณี	ฐานข้อมูลเก่า (DB)	ข้อมูลใหม่ที่เพิ่มเข้ามา (db)	ฐานข้อมูลรวม (DBdb)
1	Itemset.sup ≥ minsup	Itemset.sup ≥ minsup	Itemset.sup ≥ minsup
2	Itemset.sup ≥ minsup	Itemset.sup < minsup	Itemset.sup ≥ minsup / Itemset.sup < minsup
3	Itemset.sup < minsup	Itemset.sup ≥ minsup	Itemset.sup ≥ minsup / Itemset.sup < minsup
4	Itemset.sup < minsup	Itemset.sup < minsup	Itemset.sup < minsup

หมายเหตุ :

✓ Itemset.sup ≥ minsup คือ ไอเท็มเซต ที่ผ่านค่าสนับสนุนขั้นต่ำ

✓ Itemset.sup < minsup คือ ไอเท็มเซต ที่ไม่ผ่านค่าสนับสนุนขั้นต่ำ

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 FUP

อัลกอริทึมนี้ [5] มุ่งเน้นที่จะลดจำนวนไอเท็มเซตของข้อมูลที่เพิ่มเข้ามา ซึ่งจะนำไปเพื่อปรับปรุงค่าสนับสนุนในข้อมูลเก่าหรือจำนวนไอเท็มเซตของข้อมูลเก่าที่ไม่ได้เก็บไว้เนื่องจากไม่ผ่านค่าสนับสนุนขั้นต่ำ แต่กลับกลายเป็นไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำเมื่อมีการเพิ่มของข้อมูล โดยการทำงานของอัลกอริทึม นี้จะแบ่งเป็น 2 ส่วน คือ

1) การสร้างไอเท็มเซตความยาวหนึ่ง ที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลรวม โดยจะเข้าไปนับค่าสนับสนุนในฐานข้อมูลเก่า และข้อมูลใหม่ที่เพิ่มเข้ามา แล้วแยกไอเท็มเซตที่เหมือนกัน และไม่เหมือนกันของฐานข้อมูลเก่า และข้อมูลใหม่ที่เข้ามา เพื่อพิจารณาสร้างไอเท็มเซตความยาวหนึ่ง ที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลรวม

2) การสร้างไอเท็มเซตความยาวมากกว่าหนึ่ง ที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลรวม จะทำการตรวจสอบค่าสนับสนุนของไอเท็มเซตฐานข้อมูลเก่า และข้อมูลใหม่ โดยไอเท็มเซตของข้อมูลใหม่ที่เหมือนกับไอเท็มเซตของข้อมูลเก่า และยังไม่ทราบค่าสนับสนุน ก็จะไปนับค่าสนับสนุนจากข้อมูลใหม่ที่เข้ามา หรือไอเท็มเซตของข้อมูลใหม่ที่เข้ามา ไม่เหมือนกับไอเท็มเซตของข้อมูลเก่า ก็จะไปนับค่าสนับสนุน

จากข้อมูลใหม่และฐานข้อมูลเก่าด้วย

แต่แนวทางนี้มีข้อเสีย คือเมื่อมีข้อมูลใหม่เพิ่มเข้ามา จะทำแคสต์ไอเท็มเซตที่ต้องใช้ไปนับในฐานข้อมูลเก่า ไม่ได้ลดจำนวนครั้งของการเข้าไปนับค่าสนับสนุนในฐานข้อมูลเก่าซึ่งใช้เวลามาก ทำให้เวลาโดยรวมเมื่อมีข้อมูลใหม่เข้ามาหลาย ๆ ครั้งยังสูงอยู่มาก

2.2.2 NUWEP

อัลกอริทึมนี้ [8] มุ่งเน้นจะใช้ความรู้จากฐานข้อมูลเดิม โดยเก็บไอเท็มเซตที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำเมื่อมีข้อมูลใหม่เข้ามา ในการนับครั้งแรก โดยใช้ค่าขอบเขตล่างสำหรับไอเท็มเซตที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำเมื่อมีข้อมูลใหม่เข้ามา

ในกรณีปกติ ถ้าเก็บไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำเมื่อข้อมูลใหม่เข้ามา ค่าสนับสนุนของข้อมูลใหม่อาจจะทำให้ไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลเก่า ไม่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลทั้งหมด หรือค่าสนับสนุนของข้อมูลใหม่ที่เข้ามาอาจจะทำให้ไอเท็มเซตที่ไม่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลเก่า ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลทั้งหมด

การแก้ปัญหาของการไอเท็มเซตเหล่านี้ ทำได้โดยใช้ค่าขอบเขตล่าง ซึ่งหาได้จากจำนวน transaction ของค่าสนับสนุนขั้นต่ำของฐานข้อมูลรวมลบด้วยจำนวน transactions เฉลี่ยที่เข้ามา กลายเป็นค่าขอบเขตล่าง ก็สามารถที่จะเก็บไอเท็มเซตที่ไม่ผ่านค่าสนับสนุนขั้นต่ำหรืออินฟริควเอนไอเท็มเซต (infrequent itemsets) ไว้ก่อน เพื่อให้สามารถนำมาใช้ได้ในการปรับปรุงค่าสนับสนุนในครั้งต่อไป

เทคนิค NUWEP ที่มีการใช้ค่าขอบเขตล่าง เพื่อเก็บตัวไอเท็มเซตที่ไม่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลเก่าไว้ก่อน จะมีข้อเสียคือไอเท็มเซตจำนวนมากที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำเมื่อมีการเพิ่มของข้อมูล มีโอกาสน้อยมากที่จะกลายเป็น ไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำเมื่อมีการเพิ่มของข้อมูล ทำให้เสียเวลาในการนับและเก็บไอเท็มเซตที่ไม่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลเก่า

เทคนิค NUWEP ที่ประสบปัญหาการเก็บไอเท็มเซตจำนวนมากที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำเมื่อมีการเพิ่มของข้อมูล แต่มีโอกาสน้อยมากที่จะกลายเป็น ไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำเมื่อมีการเพิ่มของข้อมูล งานวิจัยชิ้นนี้จึงเสนอวิธีการคัดเลือก ไอเท็มเซตที่มีแนวโน้มว่าจะกลายเป็นไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำเมื่อมีข้อมูลใหม่เข้ามา

โดยใช้ค่าขอบเขตล่างและค่าความเชื่อมั่นขั้นต่ำ เป็นเกณฑ์ในการเลือกเก็บไอเท็มเซตที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำ เมื่อมีข้อมูลใหม่เข้ามา สำหรับการนับครั้งแรก เพื่อให้จำนวนไอเท็มเซตและเวลาในการนับค่าสนับสนุนในข้อมูลเก่าสำหรับการนับครั้งแรก ลดลง และไอเท็มเซตที่ถูกคัดเลือกยังคงครอบคลุมไอเท็มเซตทุกตัวที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำในการนับครั้งถัดไป

3. Fast Update of Frequent Itemsets (FUFU)

3.1 แนวคิด

3.1.1 การเก็บ infrequent itemsets

การเก็บไอเท็มเซตที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำ เมื่อมีข้อมูลใหม่เข้ามา จะเก็บจาก ไอเท็มเซตที่ไม่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลเก่า แต่ผ่านค่าขอบเขตล่าง (จำนวน transaction ของค่าสนับสนุนขั้นต่ำของฐานข้อมูลรวม (DBdb) ลบ ด้วยจำนวน transaction (db) ของข้อมูลใหม่ที่เข้ามา) ยกตัวอย่างเช่น $|DB|=10000$, $|db|=1000$, ค่าสนับสนุนขั้นต่ำ (min sup) = 20% ฉะนั้นค่าขอบเขตล่าง (Lower bound) = จำนวน transaction ของค่าสนับสนุนขั้นต่ำของฐานข้อมูลรวม $(10000+1000)*20\% = 2200$ ลบด้วยจำนวน transaction ของข้อมูลใหม่ (db) 1000 ฉะนั้น ค่าขอบเขตล่างเพื่อเก็บ infrequent item set คือ 1200

เมื่อได้ขอบเขตล่างที่จะเก็บ ไอเท็มเซตที่ไม่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลเก่าแต่ผ่านค่าขอบเขตล่าง (ไอเท็มเซตที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำเมื่อมีข้อมูลใหม่เข้ามา) จะมีปัญหากรณี ไอเท็มเซตจำนวนมากที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำเมื่อมีข้อมูลใหม่เข้ามา แต่มีโอกาสน้อยมากที่จะกลายเป็น ไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำเมื่อมีข้อมูลใหม่เข้ามา ทำให้เสียเวลาในการนับและเก็บไอเท็มเซตที่ไม่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลเก่า แต่ผ่านค่าขอบเขตล่างไว้เป็นจำนวนมาก

งานวิจัยชิ้นนี้จึงเสนอวิธีพิจารณาเลือกไอเท็มเซตที่ไม่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลเก่า แต่ผ่านค่าขอบเขตล่างที่มีแนวโน้มว่าจะกลายเป็นไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลทั้งหมดและลดไอเท็มเซตที่ไม่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลเก่า แต่ผ่านค่าขอบเขตล่างที่มีโอกาสน้อยมากที่จะกลายเป็น ไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลทั้งหมดเมื่อมีข้อมูลใหม่เข้ามา โดยใช้เกณฑ์ของค่าขอบเขตล่างและค่าความเชื่อมั่นขั้นต่ำ

(Minimum confidence) ซึ่งจะกล่าวถึงในหัวข้อต่อไป

3.1.2 การกำหนดค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence)

การพิจารณาว่าไอเท็มเซตใดๆ มีค่าความเชื่อมั่นผ่านค่าความเชื่อมั่นขั้นต่ำ สามารถคำนวณค่าความเชื่อมั่นได้จากสมการความเชื่อมั่น และค่าความเชื่อมั่นขั้นต่ำกำหนดโดยผู้ใช

$$\text{Confidence (A} \rightarrow \text{B)} = P(\text{B}|\text{A})$$

$$\text{ยกตัวอย่าง AB.support} = 10; \text{B.support} = 5;$$

$$\text{Confidence (A} \rightarrow \text{B)} = \text{B.support} / \text{AB.support} = 0.5$$

จากสมการของค่าเชื่อมั่น จะเห็นว่า ความน่าจะเป็นที่เกิด ไอเท็มเซต AB ขึ้นอยู่กับความน่าจะเป็นของไอเท็มเซต A หมายความว่าน่าจะเป็นของไอเท็มเซตที่มีความยาว n+1 จะขึ้นอยู่กับความน่าจะเป็นของไอเท็มเซตที่มีความยาว n เมื่อ n คือจำนวนเต็มบวก จากค่าเชื่อมั่นขั้นต่ำ (Minimum confidence) ไอเท็มเซตที่ไม่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลเก่า แต่ผ่านค่าสนับสนุนขอบเขตล่าง ในแต่ละความยาวของ ไอเท็มเซต และมีค่าความเชื่อมั่นสูง แสดงว่า ไอเท็มเซตนั้นมีความน่าจะเป็นที่จะกลายเป็น ไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลทั้งหมด

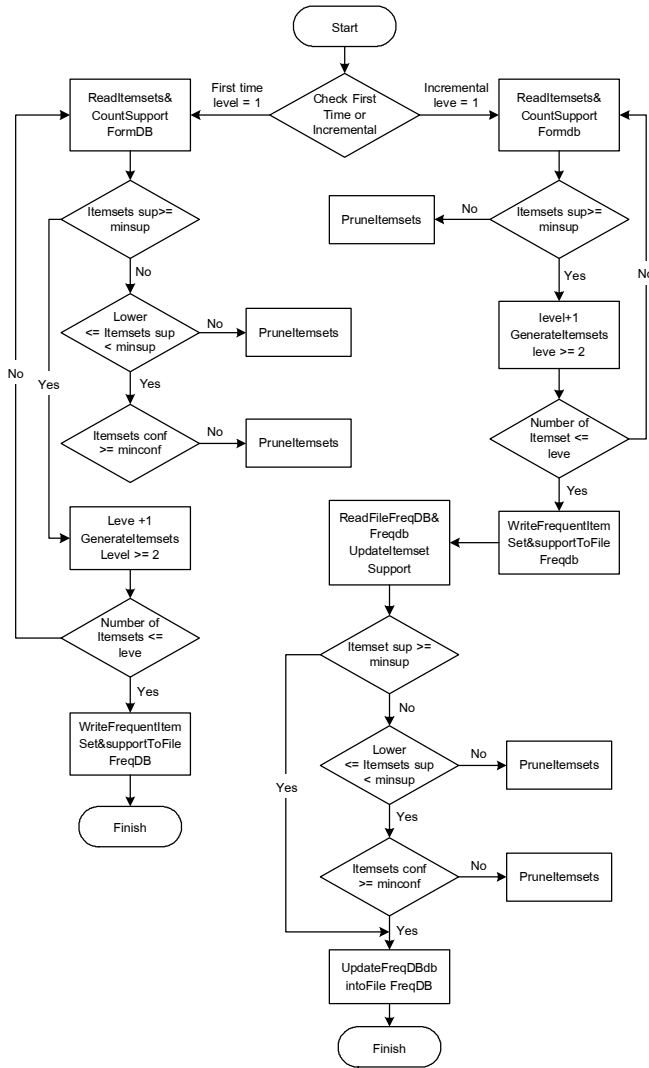
3.2 Algorithm FUFU (Fast Update of Frequent Itemsets)

จากภาพที่ 1 แสดงขั้นตอนการทำงานของอัลกอริทึม FUFU ทั้งหมด และสัญลักษณ์ที่ใช้ในภาพที่ 1 สามารถดูได้จากตารางที่ 2

จากขั้นตอนการทำงานของอัลกอริทึม FUFU จะแยกการทำงานเป็น 2 ทาง คือ

3.2.1 ประมวลผลฐานข้อมูลครั้งแรก (First Time Processing)

- อ่านและนับค่าสนับสนุนของ itemsets จาก DB
- ถ้าค่าsupport ของ itemset $\geq$ minsup จริง ก็ให้ไปสร้าง itemsets level ต่อไป ถ้าเท็จก็เช็ค itemset.sup < minsup และ $\geq$ lower
- ถ้าค่า support ของ itemset < minsup และ $\geq$ lower จริง ก็ให้ไปเช็ค minconf ถ้าเท็จก็ตัดทิ้ง
- ถ้าค่า confidence ของ itemset > minconf จริง ก็ให้ไปสร้าง itemsets level ต่อไป เท็จก็ตัดทิ้ง
- สร้าง itemsets สำหรับ level ถัดไป
- ทำซ้ำตั้งแต่ขั้นตอนแรกถึงขั้นตอนล่าสุด ก็ต่อเมื่อจำนวน



ภาพที่ 1 แสดงขั้นตอนการทำงานของอัลกอริทึม FUF

itemsets $\leq$ level ไม่จริง

- เขียน Frequent Items และค่า support ของ items ลงไฟล์

3.2.2 ประมวลผลเมื่อมีข้อมูลเข้ามา (Incremental Processing)

- อ่านและนับค่าสนับสนุนของ itemsets จาก db
- ถ้าค่าsupport ของ itemset $\geq$ minsup จริง ก็ให้ไปสร้าง itemsets level ต่อไป เท็จก็ให้ตัดทิ้ง
- สร้าง itemsets สำหรับ level ถัดไป
- ทำซ้ำตั้งแต่ขั้นตอนแรกถึงขั้นตอนล่าสุด ก็ต่อเมื่อจำนวน itemsets $\leq$ level ไม่จริง
- เขียน Frequent Items และค่า support ของ items ใน db ลงไฟล์
- อ่านและปรับปรุงค่าสนับสนุนของ frequent itemsets จาก ไฟล์ frequent itemset ของ DB และ db

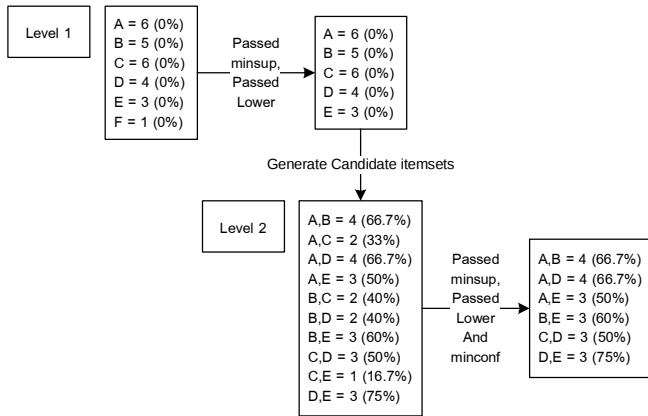
- ถ้าค่าsupport ของ itemset $\geq$ minsup จริง ก็ให้เขียน Frequent Items และค่า support ของ items ใน DBdb ลงไฟล์ FreqDB เป็นเท็จ ก็ใช้ itemset $<$ minsup และ $\geq$ lower
- ถ้าค่า support ของ itemset $<$ minsup และ $\geq$ lower จริง ก็ให้ไปเช็ค minconf ถ้าเท็จก็ตัดทิ้ง
- ถ้าค่า confidence ของ itemset $>$ minconf จริง ก็ให้ไปเขียน Frequent Items และค่า support ของ items ใน DBdb ลงไฟล์ FreqDB

ตารางที่ 2 ตารางค่าพารามิเตอร์

ตัวแปร	คำอธิบาย
Level	ความยาวของไอเท็มเซต
Itemset.sup	ค่าสนับสนุนของไอเท็มเซต
Itemset.conf	ค่าความเชื่อมั่นของไอเท็มเซต
Minsup	ค่าสนับสนุนขั้นต่ำ
Minconf	ค่าความเชื่อมั่นขั้นต่ำ
Lower	ค่าขอบเขตล่าง ($ DBdb * minsup / db $)
DB	ฐานข้อมูลเก่า
db	ข้อมูลใหม่ที่เพิ่มเข้ามา
DBdb	ฐานข้อมูลรวม
Frequent itemsets	ไอเท็มเซตที่ผ่านค่าความเชื่อมั่นขั้นต่ำ
FreqDB	ไอเท็มเซตที่ผ่านค่าความเชื่อมั่นขั้นต่ำในฐานข้อมูลเดิม
Freqdb	ไอเท็มเซตที่ผ่านค่าความเชื่อมั่นขั้นต่ำในฐานข้อมูลใหม่

3.3 ตัวอย่างการทำงาน

เพื่อให้ง่ายต่อการเข้าใจ ดูภาพที่ 2 และสัญลักษณ์จาก ตารางที่ 2 แสดงถึงตัวอย่างการทำงานของอัลกอริทึม FUF สำหรับการประมวลผลฐานข้อมูลครั้งแรกในฐานข้อมูล DB กำหนดให้ $|DB|=10$ transactions $|db|=6$ transactions $Minsup=5$ transactions (50%) $Minconf=50\%$ $Lower=2$ transactions ยกตัวอย่าง A=6 (0%) หมายความว่า ไอเท็มเซต A มีค่าสนับสนุน 6 transactions มีค่าความเชื่อมั่น 0 %



ภาพที่ 2 แสดงตัวอย่างการทำงานเบื้องต้นของ FUFU

จากภาพที่ 2

Level 1

- ในการคำนวณจะเริ่มต้นจาก itemsets level 1 เพื่อนับค่าสนับสนุนและค่าความเชื่อมั่นของแต่ละ items
- ส่วนของ Passed minsup, Passed lower จะกำจัด itemsets ที่ไม่ผ่าน minsup หรือ ไม่ผ่าน lower
- สร้าง แคนดิเดต ไอเท็มเซต (candidate itemsets) ที่มี ความยาว 2

Level 2

- จะนำผลลัพธ์จาก candidate itemsets จาก level 1 มาเพื่อนับค่าสนับสนุนและค่าความเชื่อมั่นของแต่ละ items
- ส่วนของ Passed minsup, Passed lower and minconf จะกำจัด itemsets ที่ไม่ผ่าน minsup หรือ ผ่าน lower แต่ไม่ผ่าน minconf

o จากรูปจะมี itemsets บางตัว เช่น CE ไม่ผ่าน min sup และ lower ซึ่งจะถูกตัดทิ้งไป

o จากรูปจะมี itemsets บางตัว เช่น BD ไม่ผ่าน min sup ผ่าน lower แต่ไม่ผ่าน minconf ก็จะถูกตัดทิ้งไป

หลังจากนั้น Itemsets ที่ผ่านการตรวจสอบ Passed minsup, Passed minconf จะถูกนำมาสร้างเป็น candidate itemsets ของ level ถัดไปเพื่อทำต่อไป จนกว่าไม่สามารถสร้าง candidate itemsets ได้

4. ผลการดำเนินงาน

4.1 ผลการทดลอง

งานวิจัยนี้ทำการวิเคราะห์และเปรียบเทียบประสิทธิภาพในการเลือกไอเท็มเซตที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำเมื่อมีข้อมูลใหม่เข้ามา เพื่อใช้ปรับปรุงค่าสนับสนุน เมื่อมี

การเพิ่มข้อมูลใหม่เข้ามา โดยศึกษาเวลาในการประมวลผล จำนวนการเก็บไอเท็มเซตที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำเมื่อมีข้อมูลใหม่เข้ามา และไอเท็มเซตที่ถูกคัดเลือกยังคงครอบคลุมไอเท็มเซตทุกตัวที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำในการนับครั้งถัดไปหรือไม่

ในการทดลองนี้ใช้ฐานข้อมูลสังเคราะห์ ชุดข้อมูลที่เกาะกลุ่ม (Dense datasets) C10D10K.dat [10] มาเป็น ฐานข้อมูลเดิม (DB) จำนวน 10000 transactions และจำนวนไอเท็ม = 10 ไอเท็มต่อ 1 transaction และข้อมูลใหม่ที่เพิ่มเข้ามา (db) ครั้งละ 100 transactions และจำนวนไอเท็ม = 10 ไอเท็มต่อ 1 transaction จำนวน 5 ครั้ง

การทดลองนี้ ทดสอบอยู่บนเครื่อง Notebook CPU 1.5 GHz, RAM 1256 MB บนระบบปฏิบัติการ windows XP SP2 ใช้ Microsoft Visual Studio 2005 ในการ compile และประมวลผล

ตารางที่ 3 ตารางเปรียบเทียบจำนวน frequent item set และเวลาในการประมวลผล เมื่อมีการ re-scan ฐานข้อมูลเก่า ไม่ re-scan ฐานข้อมูลเก่าและใช้ min conf ในการพิจารณา เมื่อค่า minimum support = 10 %

Round (number transaction)	Minimum support=10%							
	Rerun Apriori		NUWEP		FUFU		FUFU	
	Time (ms)	# frequent	Time (ms)	# frequent	Time (ms)	# frequent	Time (ms)	# frequent
1 (10000)	79824	487	90369	580	88597	552	89308	532
2 (10100)	79924	487	2062	556	2052	540	2052	525
3 (10200)	80355	487	2052	552	2022	536	2022	524
4 (10300)	81106	487	2042	552	2022	536	2022	524
5 (10400)	82058	475	2032	548	2032	532	2032	520
6 (10500)	81977	475	2032	536	2032	528	2032	516
Total	485244	2898	100589	3324	98757	3224	99468	3141



ภาพที่ 3 แสดงเวลาประมวลผลในในรอบของการเพิ่มข้อมูล เมื่อมีการ re-scan ฐานข้อมูล ไม่ re-scan ฐานข้อมูลและใช้ minimum confidence ในการพิจารณา



เพื่อถ่ายทอดการเข้าใจผลการทดลอง ตารางที่ 3 และกราฟจากภาพที่ 3 ประกอบ แล้วพิจารณาผลการทดลองดังหัวข้อต่อไป

4.1.1 การเก็บไอเท็มเซตที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำเมื่อมีข้อมูลใหม่เข้ามา

ประเด็นที่ต้องพิจารณาสำหรับการวิจัยกรณีนี้คือการเก็บไอเท็มเซตที่ไม่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูล (DB) แต่ผ่านค่าสนับสนุนขอบเขตล่าง ที่คาดว่าจะกลายเป็นไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลทั้งหมด (DB+db) เมื่อข้อมูลใหม่เข้ามา เวลารวมและเวลาในการทำงานแต่ละครั้งเมื่อข้อมูลเข้ามาจะดีกว่ากรณีไม่เก็บไอเท็มเซตที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำเมื่อมีข้อมูลใหม่เข้ามาไว้ แต่ต้องเข้าไปนับค่าสนับสนุนในฐานข้อมูลเก่า (DB) ทุกครั้งที่มีข้อมูลใหม่เข้ามา (db) หรือไม่ ถ้าสังเกตที่ NUWEP, minconf=0% และ rerun Apriori จะเห็นว่ารอบแรก rerun Apriori จะเหนือกว่า เพราะเก็บไอเท็มเซตน้อยกว่า เนื่องจากไม่ได้ใช้ค่าขอบเขตล่าง แต่สำหรับ NUWEP, minconf=0% จะใช้ค่าขอบเขตล่าง เพื่อเก็บไอเท็มเซตที่คาดว่าจะกลายเป็นไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลทั้งหมด (DB+db) จำนวนมากกว่าทำให้เวลา NUWEP, minconf=0% มากกว่า rerun Apriori ในรอบแรก แต่หลังจากรอบที่หนึ่งเป็นต้นไป NUWEP, minconf=0% ได้นำไอเท็มเซตที่คาดว่าจะกลายเป็นไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลทั้งหมด (DBdb) มาช่วยปรับปรุงค่าสนับสนุน ทำให้ลดเวลาการปรับปรุงค่าสนับสนุนได้มาก และเหนือกว่า rerun Apriori อย่างมากมาย

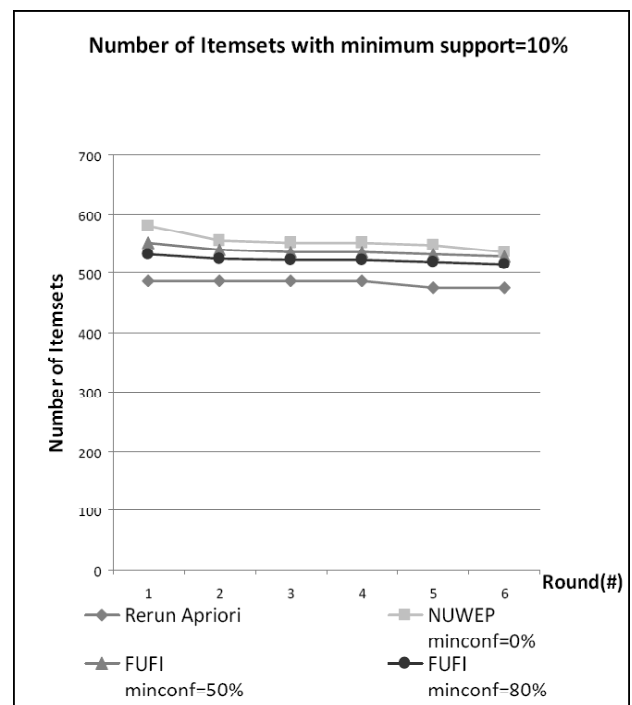
4.1.2 เงื่อนไข minimum confidence

ประเด็นที่ต้องพิจารณาต่อมาก็คือ กรณีเพิ่มเงื่อนไข Minimum confidence เข้าไปเพื่อสร้างเกณฑ์ในการเก็บไอเท็มเซตที่คาดว่าจะกลายเป็นไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลทั้งหมด (DBdb) จะทำให้ประสิทธิภาพของเทคนิคนี้ลดลงหรือไม่ จาก FUFU, minconf=50% และ FUFU, minconf=80% เวลาที่ใช้ประมวลผลในรอบที่ 0-6 แต่ละรอบและเวลารวม ก็ยังน้อยกว่าเวลาของ rerun Apriori และ NUWEP, minconf=0% จากข้อความที่กล่าวมา จึงถือว่าการเพิ่ม minconf ในการพิจารณา ไม่ได้ทำเวลาในการประมวลผลเพิ่มขึ้น แต่กลับช่วยลดเวลาการประมวลผลเมื่อเทียบ rerun Apriori และ NUWEP, minconf=0%

4.1.3 เวลาประมวลผลและการคัดเลือกไอเท็มเซต

ประเด็นที่ ต้องพิจารณาคือ การวิจัยนี้มุ่งหวังที่จะลดเวลาประมวลผลและจำนวนไอเท็มเซต ที่คาดว่าจะกลายเป็นไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลทั้งหมด (DBdb) ให้น้อยลงสำหรับการนับครั้งแรก

จากตารางที่ 3 เวลาการประมวลผลของ FUFU, minconf=50% และ FUFU, minconf=80% ในรอบที่ 1 จะน้อยกว่า NUWEP, minconf=0% ค่อนข้างมาก แต่ในรอบที่ 2-6 เวลาการประมวลผลของ FUFU, minconf=50%; FUFU, minconf=80% และ minconf=0% น้อยกว่าเพียงเล็กน้อย ดังนั้นสามารถสรุปได้ว่าเทคนิค FUFU ช่วยลดเวลาในการประมวลผลได้ดีกว่าเทคนิค NUWEP



ภาพที่ 4 แสดงจำนวนไอเท็มเซตที่ผ่านเกณฑ์ของแต่ละอัลกอริทึม ในแต่ละรอบของการเพิ่มข้อมูล เมื่อ minimum support = 10%

จากภาพที่ 4 พิจารณาจำนวนไอเท็มเซตที่คัดเลือกโดยเทคนิค FUFU, minconf 50%; FUFU, minconf 80% มีจำนวนไอเท็มเซตน้อยกว่า เทคนิคของ NUWEP ในทุกรอบของการเพิ่มของข้อมูล แต่ยังมากกว่าเทคนิค Rerun Apriori เพียงเล็กน้อย

อย่างไรก็ตามถ้าเทียบเวลาการประมวลผลระหว่าง FUFU กับ Rerun Apriori จากตารางที่ 3 จะพบว่า FUFU จะใช้เวลาประมวลผลในทุกรอบของการเพิ่มข้อมูลน้อยกว่า Rerun

Apriori อย่างมากมาย จึงถือจำนวนไอเท็มเซตที่เพิ่มขึ้นมาสำหรับ FUFU มีข้อเสียเพียงเล็กน้อย เมื่อเทียบกับเวลาประมวลผลที่ลดลงไปอย่างมาก

และจากตารางที่ 4 พบว่าความผิดพลาดของเทคนิค NUWEP, FUFU ในการคัดเลือกไอเท็มเซตที่คาดว่าจะกลายเป็นไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลทั้งหมด ทุกรอบของการเพิ่มของข้อมูล มีค่าเป็นศูนย์

ฉะนั้นจึงสามารถสรุปได้ว่าเทคนิค FUFU สามารถคัดเลือกไอเท็มเซตที่ยังคงครอบคลุมไอเท็มเซตทุกตัวที่คาดว่าจะผ่านค่าสนับสนุนขั้นต่ำในการนับครั้งถัดไป

ตารางที่ 4 ตารางเปรียบเทียบความผิดพลาดในการคัดเลือกไอเท็มเซตที่คาดว่าจะกลายเป็นไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลทั้งหมด และค่า $minimum\ support = 10\%$

Round (number transactions)	Minimum support=10%						
	Rerun Apriori	NUWEP minconf=0%		FUFU minconf=50%		FUFU minconf=80%	
	# frequent	# frequent	Error of pruned frequent itemset in next round	# frequent	Error of pruned frequent itemset in next round	# frequent	Error of pruned frequent itemset in next round
1 (10000)	487	580	0	552	0	532	0
2 (10100)	487	556	0	540	0	525	0
3 (10200)	487	552	0	536	0	524	0
4 (10300)	487	552	0	536	0	524	0
5 (10400)	475	548	0	532	0	520	0
6 (10500)	475	536	0	528	0	516	0
Total	2898	3324	0	3224	0	3141	0

4.2 สรุป

เมื่อมีข้อมูลใหม่เข้ามา ค่าสนับสนุนของแต่ละไอเท็มเซตต้องมีการปรับปรุงให้สอดคล้องกับข้อมูลใหม่ แต่การปรับปรุงก็ต้องเข้าค้นหาในฐานข้อมูลซึ่งมีขนาดใหญ่ ทำให้เสียเวลาในการปรับปรุง จึงมีการเสนอวิธีว่า เมื่อมีการหาค่าสนับสนุนในฐานข้อมูลครั้งแรกให้เก็บไอเท็มเซตที่คาดว่าจะกลายเป็นไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลทั้งหมด (DBdb) ในอนาคตไว้ เพื่อจะนำมาใช้ปรับปรุงค่าสนับสนุนในครั้งถัดไป

แต่การเก็บไอเท็มเซตที่คาดว่าจะกลายเป็นไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลทั้งหมด (DBdb) ในอนาคตไว้ ก็ใช้เวลามาก และไอเท็มเซตที่คาดว่าจะกลายเป็นไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลทั้งหมด (DBdb) บางตัวก็ไม่ได้ใช้เลย การวิจัยนี้จึงเริ่มจากเข้าไปปรับปรุงวิธีการเลือกไอเท็มเซตที่คาดว่าจะกลายเป็นไอเท็มเซตที่ผ่านค่าสนับสนุนขั้นต่ำของฐานข้อมูลทั้งหมด (DBdb) ให้มี

ประสิทธิภาพ โดยเสนอ ค่าความเชื่อมั่นขั้นต่ำ (minimum confidence) ในการคัดเลือก จากการทดลองกับฐานข้อมูลสังเคราะห์ dense datasets ผลลัพธ์แสดงถึงว่า การใช้ค่าความเชื่อมั่นขั้นต่ำในการพิจารณา ไม่ทำเวลาในการประมวลผลเพิ่มขึ้น แต่กลับช่วยเวลาในการประมวลผลให้น้อยลงได้

แนวคิดของงานวิจัยชิ้นนี้สามารถนำไปประยุกต์ใช้กับเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ (Associative classification) [10] เพราะว่าเมื่อข้อมูลใหม่มีการเพิ่มเข้ามา เทคนิค Associative classification จะเกิดปัญหาต้องนับค่าสนับสนุนใหม่ในฐานข้อมูลเก่าทุกครั้ง ทำให้สามารถที่จำแนกของงานวิจัยชิ้นนี้ไปแก้ปัญหาและสามารถนำแนวคิดงานวิจัยชิ้นนี้ไปพัฒนาเป็นเทคนิค Incremental associative classification เพราะแนวคิดงานวิจัยชิ้นนี้เสนอแนวทางในการนับและปรับปรุงค่าสนับสนุนใหม่ในฐานข้อมูลเก่า

5. เอกสารอ้างอิง

- [1] R. Srikant, R. Agrawal: "Mining Generalized Association Rules", Proc. of the 21st Int'l Conference on Very Large Databases, Zurich, Switzerland, Sep. 1995. Expanded version available as IBM Research Report RJ 9963, June 1995.
- [2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A. I. Verkamo: "Fast Discovery of Association Rules", Advances in Knowledge Discovery and Data Mining, Chapter 12, AAAI/MIT Press, 1995.
- [3] R. Agrawal, R. Srikant: "Fast Algorithms for Mining Association Rules", Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, Sept. 1994. Expanded version available as IBM Research Report RJ9839, June 1994.
- [4] R. Agrawal, T. Imielinski, A. Swami: "Mining Associations between Sets of Items in Massive Databases", Proc. of the ACM-SIGMOD 1993 Int'l Conference on Management of Data, Washington D.C., May 1993, 207-216.
- [5] David W. Cheung, J. Han, V. Ng, and C.Y. Wong : "Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Techniques",



- Proc. 12th IEEE International Conference on Data Engineering (ICDE-96), New Orleans, Louisiana, U.S.A., March, 1996.
- [6] David Wai-Lok Cheung, Sau Dan Lee, and Benjamin Kao : "A general incremental technique for maintaining discovered association rules", In Proceedings of the 5 th Intl. Conf. on Database Systems for Advanced Applications (DASFAA'97), Melbourne, Australia, April 1997.
- [7] Shiby Thomas, Sreenath Bodagala, Khaled Alsabti, Sanjay Ranka: "An Efficient Algorithm for the Incremental Updation of Association Rules in Large Databases", In Proceedings of the 3rd International conference on Knowledge Discovery and Data Mining (KDD 97), New Port Beach, California. August 1997.
- [8] Susan P. Imberman, Abdullah Uz Tansel, Eric Pacuit: "An Efficient Method For Finding Emerging Large Itemsets" , The Third Workshop on Mining Temporal and Sequential Data, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2004.
- [9] R. J. Bayardo, R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense databases. Proc. of the 15th ICDE Conference, pages 188-197, March 1999.
- [10] Bing Liu, Wynne Hsu, and Yiming Ma : "Integrating classification and association rule mining", In Knowledge Discovery and Data Mining, pages 80—86, 1998.
-