การทำเหมืองข้อมูลเพื่อหากฎความสัมพันธ์แบบหลายลำดับชั้นที่ปรากฏขึ้นบ่อยสุดเคอันดับแรก

นายสรพล ชมไพศาล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2556

MINING TOP-K MULTI-LEVEL ASSOCIATION RULES

Mr. Sorapol Chompaisal

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2013

หัวข้อวิทยานิพนธ์	การทำเหมืองข้อมูลเพื่อหากฎความสัมพันธ์แบบหลายลำดับชั้นที่ปรากฏขึ้นบ่อยสุดเคอันดับแรก
โดย	นายสรพล ชมไพศาล
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.อรรถสิทธิ์ สุรฤกษ์
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	อาจารย์ ดร.โกเมศ อัมพวัน

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.บัณฑิต เอื้ออาภรณ์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.เศรษฐา ปานงาม)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.อรรถสิทธิ์ สุรฤกษ์)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(อาจารย์ ดร.โกเมศ อัมพวัน)

.....กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.อานนท์ รุ่งสว่าง)

สรพล ชมไพศาล : การทำเหมืองข้อมูลเพื่อหากฎความสัมพันธ์แบบหลายลำดับชั้นที่ปรากฏขึ้นบ่อยสุดเคอันดับแรก. (MINING TOP-K MULTI-LEVEL ASSOCIATION RULES) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร.อรรถสิทธิ์ สุรฤกษ์, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: อ. ดร.โกเมศ อัมพวัน, 64 หน้า.

การค้นหากฎความสัมพันธ์ข้อมูลเป็นกระบวนการหนึ่งในการทำเหมืองข้อมูลที่ใช้สำหรับวิเคราะห์หาความสัมพันธ์ระหว่างสิ่งของหรือเหตุการณ์ที่เกิดขึ้นร่วมกันตั้งแต่สองสิ่งขึ้นไปภายใต้ค่าขีดแบ่งสนับสนุนและ/หรือค่าขีดแบ่งความเชื่อมั่นที่ผู้ใช้เป็นผู้กำหนด อย่างไรก็ดี สิ่งของหรือเหตุการณ์อาจถูกจัด/รวบรวมเป็นกลุ่มหรือหมวดหมู่ต่างๆที่ซึ่งเป็นเหตุให้มีการคิดค้นการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นที่จะทำให้ทราบถึงความสัมพันธ์ระหว่างหมวดหมู่หรือกลุ่มข้อมูลเพิ่มเติม อันนำมาซึ่งการได้รับองค์ความรู้จากผลลัพธ์มากขึ้นจากเดิมที่ได้ผลลัพธ์เพียงกฎความสัมพันธ์แบบลำดับชั้นเดียว ในการค้นหากฎความสัมพันธ์หลายลำดับชั้น ผู้ใช้อาจทำการกำหนดค่าขีดแบ่งสนับสนุนและ/หรือค่าขีดแบ่งความเชื่อมั่นในแต่ละลำดับชั้นที่เหมือนกันหรือแตกต่างกัน ในกรณีที่ผู้ใช้กำหนดค่าขีดแบ่งดังกล่าวไม่เหมาะสมอาจทำให้ได้ผลลัพธ์ที่มีจำนวนน้อยเกินไปหรือมากเกินไปจนผู้ใช้ไม่สามารถนำผลลัพธ์ที่ได้ไปสังเคราะห์องค์ความรู้ให้เกิดประโยชน์ได้ ดังนั้น งานวิจัยนี้จึงเสนอการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจที่สุดตามจำนวนที่ผู้ใช้ต้องการ และเสนอขั้นตอนวิธีการหารูปแบบเซตแต่ละลำดับชั้นที่ปรากฏบ่อยที่สุดตามจำนวนที่ต้องการ โดยเริ่มพิจารณาจากสิ่งของที่มีความถี่สูงสุดจากต้นไม้แสดงรูปแบบการเกิดขึ้นเป็นลำดับแรก จากนั้นทำการสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุดจากรูปแบบเซตที่ปรากฏบ่อยโดยคำนวณจากค่าถ่วงน้ำหนักความน่าสนใจของค่าสนับสนุนและค่าความเชื่อมั่นที่ผู้ใช้เป็นผู้กำหนด จากผลการทดสอบพบว่า ขั้นตอนวิธีที่นำเสนอสามารถค้นหาผลลัพธ์ได้อย่างมีประสิทธิภาพทั้งในเชิงเวลาและเชิงหน่วยความจำ เมื่อทำการกำหนดจำนวนผลลัพธ์ที่แตกต่างกัน

ภาควิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต .....
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์	ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....
ปีการศึกษา	2556	ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม .....

5470407021 : MAJOR COMPUTER SCIENCE
KEYWORDS: DATA MINING / MULTI-LEVEL ASSOCIATION RULES / N-MOST INTERESTING PATTERNS / TOP-K ASSOCIATION RULES

SORAPOL CHOMPAISAL: MINING TOP-K MULTI-LEVEL ASSOCIATION RULES. ADVISOR: ASST. PROF. ATHASIT SURARERKS, Ph.D., CO-ADVISOR: KOMATE AMPHAWAN, Ph.D., 64 pp.

Mining association rules is one interesting area of data mining used to discover correlation co-occurrence of items or events with user-given support/confidence thresholds. In addition, items or events can be grouped into categories. Thus, mining multi-level association rules is introduced. It can help users to gain more knowledge about correlation of items among their hierarchical categories. To discover the results, users have to assign one or more different support/confidence thresholds for items in all levels of concept hierarchy. If the defined thresholds to mine the results are not suitable, it may give more or less results that users cannot take the advantage from results. Therefore, this thesis introduces an alternative approach to mine the most interesting multi-level association rules which allow the users give a number of desired results. It starts to find the highest support of itemsets at each level by first considering the most frequent item from FP-tree, and then generate rules by calculate interesting values of these from weight of interesting values. The extensive performance studies show that the proposed approach have high performance and scalable in terms of time and memory on the various number of desired results.

Department:	Computer Engineering	Student's Signature .....
Field of Study:	Computer Science	Advisor's Signature .....
Academic Year:	2013	Co-Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์สำเร็จลุล่วงไปได้ด้วยความดี ด้วยความช่วยเหลือจากผู้ช่วยศาสตราจารย์ ดร. อรรถสิทธิ์ สุรฤกษ์ อาจารย์ที่ปรึกษาหลัก ที่ให้คำปรึกษาและข้อแนะนำต่างๆ ตลอดเวลาที่ผู้วิจัยได้ศึกษาในสถาบันแหล่งนี้

ขอขอบคุณอาจารย์ ดร.โกเมศ อัมพวัน อาจารย์ที่ปรึกษาร่วม ที่ให้คำแนะนำ ข้อคิด ข้อเสนอแนะในการทำวิจัย ตลอดจนตรวจทานแก้ไขวิทยานิพนธ์ฉบับนี้ให้มีความสมบูรณ์

ขอขอบพระคุณ คณะกรรมการสอบวิทยานิพนธ์ทุกท่านเป็นอย่างสูง ได้แก่ ผู้ช่วยศาสตราจารย์ ดร.เศรษฐา ปานงาม และรองศาสตราจารย์ ดร.อานนท์ รุ่งสว่าง ที่กรุณาเสียสละให้คำแนะนำให้วิทยานินพธ์ฉบับนี้ให้มีความสมบูรณ์มากขึ้น รวมถึงขอขอบพระคุณคณาจารย์ประจำภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย ทุกท่านที่ประสิทธิประสาทวิชาความรู้ที่มีคุณค่ากับผู้วิจัย

ขอขอบพระคุณ มารดา ญาติพี่น้องทุกคน ที่สนับสนุนทั้งด้านกำลังใจ ทุนทรัพย์ และคำภาวนาสำหรับช่วงเวลาที่ศึกษาอยู่นี้ ขอบคุณเพื่อนพี่น้องสมาชิกห้องปฏิบัติการทางวิศวกรรมระบบนับได้เชิงทฤษฎี (elite) ที่ช่วยเหลือ ให้คำแนะนำต่างๆ ที่เป็นประโยชน์ในการทำวิจัย และช่วยให้วิทยานิพนธ์ฉบับนี้เสร็จได้ด้วยดี ขอบคุณเพื่อนๆ CS41 และเพื่อนทุกคนที่ให้ความสนุกสนานฉันท์เพื่อน ซึ่งเป็นกำลังใจที่ดี นอกนั้นยังให้คำแนะนำดีๆ สำหรับการทำวิจัยด้วย สุดท้ายขอขอบคุณทุกคนที่มิได้กล่าวไว้ ณ ที่นี้ ที่มีส่วนทำให้วิทยานิพนธ์นี้สำเร็จลุล่วงไปได้ด้วยดี

สารบัญ

หน้า

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การหากฎความสัมพันธ์ข้อมูลจากชุดข้อมูล (Association rules mining) เป็นแขนงหรือสาขาหนึ่งของการทำเหมืองข้อมูล (Data Mining) ที่ซึ่งสามารถอธิบายความสัมพันธ์ระหว่างสิ่งของหรือเหตุการณ์ตั้งแต่สองเหตุการณ์ขึ้นไปที่เกิดขึ้นร่วมกันในฐานข้อมูล [1, 2] โดยในการหากฎความสัมพันธ์ของข้อมูลได้ถูกนำไปประยุกต์ใช้ในธุรกิจหรือกระบวนการต่างๆ มากมาย เช่น ด้านธุรกิจการค้าปลีก (retail marketing) จะประยุกต์ใช้กฎความสัมพันธ์สำหรับการวิเคราะห์พฤติกรรมการอุปโภคของลูกค้าซึ่งจะช่วยให้ทางบริษัทหรือร้านค้าสามารถเข้าใจถึงพฤติกรรมของผู้บริโภคมากขึ้น ในส่วนของบริษัทที่มีเว็บไซต์หรือบริษัทผู้สร้างเว็บไซต์สามารถประยุกต์ใช้กฎความสัมพันธ์ไปใช้ในการติดตามการเข้าชมเว็บไซต์ของลูกค้าหรือผู้ใช้บริการอันนำมาซึ่งการปรับปรุงเว็บไซต์ให้สามารถใช้งานได้ง่ายขึ้น สำหรับด้านวงการแพทย์ก็นำกฎความสัมพันธ์ไปใช้เพื่อช่วยวิเคราะห์หรือวินิจฉัยโรคได้ เป็นต้น จากที่กล่าวมาข้างต้น เราสามารถสังเกตได้ว่าการหากฎความสัมพันธ์ได้เข้ามามีบทบาทในการดำรงชีวิตประจำวัน ซึ่งรูปแบบของกฎความสัมพันธ์ของข้อมูลสามารถแสดงในรูปแบบ $X \rightarrow Y$ ที่ซึ่งหมายถึง เมื่อ X เกิดขึ้นก็จะเกิด Y ขึ้นด้วย

ขั้นตอนการค้นหากฎความสัมพันธ์ สามารถแบ่งได้เป็นสองขั้นตอนย่อย คือ 1) การหารูปแบบเซตที่ปรากฏขึ้นบ่อยในฐานข้อมูลหรือชุดข้อมูล (Frequent Itemsets) โดยรูปแบบใดก็ตามจะเป็นรูปแบบเซตที่ปรากฏบ่อยได้นั้นจะต้องเป็นรูปแบบที่มีค่าสนับสนุนไม่น้อยไปกว่าค่าขีดแบ่งสนับสนุนซึ่งผู้ใช้จะกำหนดขึ้น (Minimum Support Threshold) และ 2) การสร้างกฎความสัมพันธ์ของข้อมูลที่น่าสนใจจากรูปแบบเซตที่ปรากฏบ่อย ซึ่งได้ทำการหาไว้ในขั้นตอนแรก โดยกฎที่ถูกสร้างจะต้องมีค่าความเชื่อมั่นไม่น้อยกว่าค่าขีดแบ่งความเชื่อมั่น (Minimum Confidence Threshold) ซึ่งผู้ใช้เป็นผู้กำหนดเช่นกัน

จากแนวคิดเริ่มต้นของการหากฎความสัมพันธ์ของข้อมูล ซึ่งเป็นการค้นหากฎความสัมพันธ์แบบลำดับชั้นเดียว (Single-level Association Rules) เช่น ขนมปัง$\rightarrow$สบู่ ซึ่งหมายถึง ลูกค้าซื้อขนมปังแล้วจะซื้อสบู่ด้วย โดยที่ขนมปังจะถูกจัดอยู่ในกลุ่มอาหารและสบู่ก็มักจะจัดอยู่ในกลุ่มของใช้ทำความสะอาดร่างกาย ซึ่งในหลายองค์กรอาจต้องการข้อมูลสารสนเทศของกฎความสัมพันธ์เพิ่มขึ้นจากเดิมที่หาเฉพาะความสัมพันธ์ของสินค้านั้นๆ กลายเป็นการหาความสัมพันธ์ของหมวดหมู่สินค้าด้วย เช่น ขนมปัง$\rightarrow$สบู่ เราสามารถหาได้กฎความสัมพันธ์เป็น อาหาร$\rightarrow$ของใช้ทำความสะอาดร่างกาย จากความต้องการดังกล่าวจึงมีแนวคิดในการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้น (Multi-level Association Rules) [3] ที่ซึ่งสามารถบ่งบอกสารสนเทศของกฎความสัมพันธ์ได้มากขึ้นกว่าแบบเดิม ข้อมูลของปัญหานี้เป็นข้อมูลที่ถูกจัดแบ่งเป็นลำดับชั้น (Concept Hierarchy) จากการกำหนดของผู้ใช้ สำหรับการหากฎความสัมพันธ์แบบหลายลำดับชั้น สามารถกำหนดค่าขีดแบ่งสนับสนุนและ/หรือค่าขีดแบ่งความเชื่อมั่นได้สองแบบ คือ 1). คือ กำหนดค่าขีดแบ่งทุกลำดับชั้นข้อมูลเหมือนกันซึ่งอาจให้ผลลัพธ์ที่ไม่สามารถใช้เป็นองค์ความรู้ได้ หากกำหนดค่าขีดแบ่งมากไปอาจ

ได้กฎความสัมพันธ์เฉพาะลำดับชั้นบนๆ (ความสัมพันธ์ระหว่างหมวดหมู่ข้อมูล) หรือ หากกำหนดค่าขีดแบ่งน้อยไปก็อาจให้กฎความสัมพันธ์ที่ดีในลำดับชั้นล่างๆ (ความสัมพันธ์ระหว่างข้อมูล) แต่ก็ให้ผลลัพธ์ที่มากเกินไปสำหรับลำดับชั้นบนๆ ด้วย 2). การกำหนดค่าขีดแบ่งในแต่ละลำดับชั้นที่ต่างกัน โดยลดค่าขีดแบ่งจากลำดับชั้นบนสู่ลำดับชั้นล่าง ซึ่งจะช่วยให้การค้นหากฎความสัมพันธ์แบบหลายลำดับชั้น ได้ผลลัพธ์ที่เหมาะสมในทุกๆ ลำดับชั้น นอกจากการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นแล้ว เราสามารถหากฎความสัมพันธ์แบบข้ามลำดับชั้น (Cross-Level Association Rules) ได้ด้วย เพื่อแก้ปัญหาผลลัพธ์ในกรณีที่รูปแบบเซตข้อมูลที่ข้อมูลไม่ได้ปรากฏบ่อย แต่รูปแบบเซตของหมวดหมู่/กลุ่ม/ประเภทข้อมูลปรากฏขึ้นบ่อย โดยสมาชิกในรูปแบบข้อมูลที่ปรากฏแบบข้ามลำดับชั้นจะไม่มีข้อมูลใดถูกจัดอยู่กลุ่ม/ประเภทข้อมูลเดียวกัน เช่น ขนมปัง → ของใช้ทำความสะอาดร่างกาย ซึ่งคำตอบกฎความสัมพันธ์แบบข้ามลำดับชั้นจะไม่สนใจถึงลำดับชั้นข้อมูล

อย่างไรก็ตาม การกำหนดค่าขีดแบ่งสนับสนุน (Minimum Support Threshold) เพื่อทำการหารูปแบบเซตที่ปรากฏขึ้นบ่อยอาจกำหนดได้ยาก เนื่องจากผู้ใช้ไม่ทราบลักษณะข้อมูลที่เกิดขึ้นในฐานข้อมูล โดยหากกำหนดค่าขีดแบ่งสนับสนุนน้อยเกินไป จำนวนรูปแบบที่ปรากฏขึ้นบ่อยตามค่าขีดแบ่งก็จะมาก จนอาจไม่สามารถนำคำตอบที่ได้มาใช้ให้เป็นประโยชน์ได้ ในทางกลับกันหากกำหนดค่าขีดแบ่งสนับสนุนมากเกินไป จำนวนรูปแบบที่ปรากฏขึ้นบ่อยอาจน้อยตาม จนอาจไม่ได้คำตอบจากการหารูปแบบนั้นเลย ดังนั้นจึงมีแนวคิดการหาเซตของรูปแบบที่ปรากฏบ่อยสุดจำนวนเคอันดับแรก (Mining Top-K Frequent Patterns) [4] เพื่อหลีกเลี่ยงการกำหนดค่าขีดแบ่งสนับสนุน จากนั้น [5, 6] เสนอการหาเซตข้อมูลที่มีค่าสนับสนุนมากสุดจำนวนเคอันดับแรกโดยไม่มีรูปแบบเซตใดที่มีขนาดใหญ่กว่าที่มีค่าสนับสนุนเท่ากับรูปแบบเซตนั้น (Frequent Closed Pattern) นอกจากนี้ [7] ได้เสนอการหารูปแบบเซตที่ปรากฏบ่อยสุดเคอันดับแรก โดยที่มีรายการข้อมูลเข้ามาอย่างต่อเนื่อง (Data Streams) ต่อมา [8] ได้เสนอการค้นหาเซตของรูปแบบที่ปรากฏขึ้นบ่อยอย่างสม่ำเสมอจำนวนเคอันดับแรกด้วย (Regular Frequent Pattern)

จากที่กล่าวมาทั้งหมดในข้างต้น งานวิจัยนี้จึงได้นำเสนอการหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจสุดตามจำนวนที่ผู้ใช้ต้องการ เพื่อลดปัญหาการกำหนดค่าขีดแบ่งสนับสนุนและ/หรือค่าขีดแบ่งความเชื่อมั่นของผู้ใช้ในแต่ละลำดับชั้น โดยนำเสนอขั้นตอนวิธีการสร้างต้นไม้แสดงรูปแบบการเกิดขึ้นจากฐานข้อมูล การค้นหารูปแบบเซตที่ปรากฏบ่อยสุดตามจำนวนผลลัพธ์ในแต่ละลำดับชั้น การสร้างต้นไม้แสดงรูปแบบ ณ ลำดับชั้นบนๆ จากต้นไม้แสดงรูปแบบการเกิดขึ้นของลำดับชั้นก่อนหน้า และการสร้างกฎความสัมพันธ์ที่น่าสนใจตามจำนวนผลลัพธ์ที่ต้องการ โดยคำนวณหาค่าความน่าสนใจแต่ละกฎจากค่าสนับสนุนและค่าความเชื่อมั่นของกฎนั้นๆ

1.2 วัตถุประสงค์ของงานวิจัย

ทำการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจสุด ที่สามารถลดทอนความยุ่งยากในการกำหนดค่าขีดแบ่งสนับสนุนและ/หรือค่าขีดแบ่งความเชื่อมั่น โดยให้ผู้ใช้กำหนดจำนวนผลลัพธ์ที่ต้องการได้

1.3 ขอบเขตของงานวิจัย

1. ชุดข้อมูลที่ทำการหากฎความสัมพันธ์แบบหลายลำดับชั้น จะเป็นชุดข้อมูลแบบรายการ (Transactional database) โดยในแต่ละข้อมูลจะถูกจัดแบ่งเป็นลำดับชั้น
2. ผู้ใช้ต้องกำหนดจำนวนผลลัพธ์ที่ต้องการ (ค่า k) ก่อนการประมวลผล
3. การหากฎความสัมพันธ์ของข้อมูลแบบหลายลำดับชั้นที่น่าสนใจสุด ผู้ใช้จะต้องกำหนดค่าถ่วงน้ำหนักความน่าสนใจของค่าสนับสนุนและค่าความเชื่อมั่น
4. ชุดข้อมูลที่ใช้ในการทดลองเป็นชุดข้อมูลที่มีการปรากฏขึ้นของข้อมูลหรือกลุ่ม/ประเภทข้อมูลแต่ละรายการ โดยไม่คำนึงถึงจำนวนหรือความถี่การปรากฏของข้อมูลหรือกลุ่ม/ประเภทข้อมูลนั้น

1.4 วิธีดำเนินการวิจัย

1. ศึกษางานวิจัยทางด้านการหากฎความสัมพันธ์ การหากฎความสัมพันธ์แบบหลายลำดับชั้น และการหารูปแบบที่ปรากฏบ่อยสุดจำนวนเคอันดับแรก
2. วิเคราะห์ปัญหาของงานวิจัยต่างๆ ที่มีความสอดคล้องกับงานวิจัยนี้
3. ออกแบบและนำเสนอขั้นตอนวิธีการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจสุดตามจำนวนที่ผู้ใช้ต้องการ
4. วิเคราะห์ผลการทดลอง
5. สรุปผลและจัดทำวิทยานิพนธ์

1.5 ประโยชน์ที่คิดว่าจะได้รับ

1. ได้ขั้นตอนวิธีสำหรับการหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจสุด ที่ซึ่งสามารถให้คำตอบเป็นเซตของกฎความสัมพันธ์ในรูปแบบลำดับชั้นต่อลำดับชั้น
2. สามารถลดทอนความยุ่งยากให้แก่ผู้ใช้ในการกำหนดค่าขีดแบ่งที่ใช้ในการหาผลลัพธ์

1.6 งานวิจัยที่ได้รับการตีพิมพ์

"Mining N-most interesting multi-level frequent itemsets without support threshold" โดย สรพล ชมไพศาล, โกเมศ อัมพวัน และอรรถสิทธิ์ สุรฤกษ์ ในงานประชุมวิชาการ The 10th international conference of computing and information technology (IC2IT 2014)

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงนิยามเบื้องต้นเกี่ยวกับการค้นหากฎความสัมพันธ์จากฐานข้อมูลขนาดใหญ่ ซึ่งจะกล่าวถึงขั้นตอนวิธีสำหรับการค้นหารูปแบบเซตด้วย นอกนั้นยังกล่าวถึงการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้น ซึ่งข้อมูลนั้นอาจถูกจัดแบ่งเป็นหมวดหมู่หรือกลุ่มข้อมูล และการหารูปแบบเซตที่ปรากฏบ่อยสุดตามที่ผู้ใช้ต้องการด้วย

2.1 การค้นหากฎความสัมพันธ์ของข้อมูล

กระบวนการค้นหากฎความสัมพันธ์ของฐานข้อมูลชุดหนึ่ง มาจากการหารูปแบบเซต (Itemsets) ที่ปรากฏขึ้นบ่อยในฐานข้อมูล (Transaction Database) ซึ่งในฐานข้อมูลรายการจะประกอบไปด้วยเซตรายการข้อมูลและเซตข้อมูลที่ปรากฏในรายการข้อมูลนั้น ซึ่ง [1] ได้นิยามการค้นหากฎความสัมพันธ์ของข้อมูลดังนี้

นิยามที่ 2.1 กำหนดให้ $I = \{i_1, i_2, i_3, \ldots, i_n\}$ แทนเซตของรายการทั้งหมดที่ปรากฏขึ้นในฐานข้อมูล

นิยามที่ 2.2 กำหนดให้ t เป็นเซตรายการข้อมูลซึ่งประกอบด้วยรหัสรายการข้อมูล (Transaction-ID) และเซตของข้อมูล $I_t \subseteq I$

นิยามที่ 2.3 กำหนดให้ $D = \{t_1, t_2, t_3, \ldots, t_m\}$ แทนเซตที่ประกอบด้วยรายการข้อมูลทั้งหมดในฐานข้อมูล ถ้าเซต A หรือ B เป็นรูปแบบเซตใดๆ โดยที่ $A, B \subseteq I$ แล้ว ดังนั้นจึงสามารถให้นิยามกฎความสัมพันธ์ได้ว่า กฎความสัมพันธ์คือ การอุปนัยของข้อมูลในรูปแบบ $A \rightarrow B$ เมื่อ $A, B \subset I$ และ $A \cap B = \emptyset$

มาตรวัดในการค้นหากฎความสัมพันธ์ของข้อมูล มีมาตรวัดสองส่วนที่สำคัญคือ 1) ค่าสนับสนุน (Support value) กล่าวคือ กฎความสัมพันธ์ $A \rightarrow B$ จะมีค่าสนับสนุนเท่ากับจำนวนการปรากฏขึ้นร่วมกันของ A และ B หารด้วยจำนวนรายการทั้งหมดที่ปรากฏในฐานข้อมูล ดังสมการที่ 1

$$supp(A \rightarrow B) = p(A \cup B) = \frac{f(A \cup B)}{|D|} \quad (1)$$

และ 2) ค่าความเชื่อมั่น (Confidence value) กล่าวคือ กฎความสัมพันธ์ $A \rightarrow B$ จะมีค่าความเชื่อมั่นเท่ากับจำนวนการปรากฏขึ้นร่วมกันของ A กับ B หารด้วยจำนวนปรากฏขึ้นของ A ดังสมการที่ 2

$$conf(A \rightarrow B) = \frac{p(A \cup B)}{p(A)} \quad (2)$$

ตัวอย่างที่ 2.1 กฎความสัมพันธ์ $A \rightarrow B$ มีค่าสนับสนุน 20% และค่าความเชื่อมั่น 70% หมายถึงเซต A กับเซต B ปรากฏขึ้นร่วมกัน 20% จากจำนวนรายการข้อมูลทั้งหมดในฐานข้อมูล และถ้า A ปรากฏขึ้นในฐานข้อมูล แล้ว B ก็จะปรากฏขึ้นด้วยกัน 70% จากจำนวนที่ A ปรากฏขึ้น

เกณฑ์ที่วัดความน่าสนใจของกฎความสัมพันธ์มีเกณฑ์วัดสองเกณฑ์หลักๆ ได้แก่ 1) ค่าขีดแบ่งสนับสนุน คือเกณฑ์ที่วัดว่าจำนวนการปรากฏขึ้นร่วมกันของข้อมูลชุดหนึ่งมีการปรากฏขึ้นบ่อย

หรือไม่ เมื่อเทียบกับค่าขีดแบ่งสนับสนุน และ 2) ค่าขีดแบ่งความเชื่อมั่น คือเกณฑ์การวัดเมื่อจำนวนเหตุการณ์ทางด้านซ้ายของกฎปรากฏขึ้นแล้วจำนวนเหตุการณ์ทางด้านขวาของกฎก็จะปรากฏขึ้นด้วย เมื่อเทียบกับค่าขีดแบ่งความเชื่อมั่น ซึ่งทั้งสองเกณฑ์นี้ผู้ใช้จะเป็นผู้กำหนด โดยมักกำหนดเป็นร้อยละ ดังนั้นกฎความสัมพันธ์หนึ่งจะเป็นกฎความสัมพันธ์ที่น่าสนใจได้ ก็ต่อเมื่อ กฎความสัมพันธ์นั้นมีค่าสนับสนุนไม่น้อยกว่าค่าขีดแบ่งสนับสนุน และมีค่าความเชื่อมั่นไม่น้อยกว่าค่าขีดแบ่งความเชื่อมั่น

ขั้นตอนการหากฎความสัมพันธ์ของข้อมูลแบ่งขั้นตอนออกเป็นสองขั้น คือ 1) หารูปแบบเซตที่ปรากฏบ่อย โดยที่ค่าสนับสนุนของรูปแบบเซตที่ปรากฏนั้นจะต้องไม่น้อยกว่าค่าขีดแบ่งสนับสนุน และ 2) สร้างกฎความสัมพันธ์ของข้อมูลจากรูปแบบเซตที่ปรากฏขึ้นบ่อย โดยที่ค่าความเชื่อมั่นของกฎความสัมพันธ์นั้นๆ มีค่าไม่น้อยกว่าค่าขีดแบ่งความเชื่อมั่น ในขั้นตอนการหารูปแบบเซตที่ปรากฏบ่อยนั้น ถือเป็นปัญหาในการค้นหา เนื่องจากรูปแบบเซตทั้งหมดที่เป็นไปได้ที่ทำการค้นหามีจำนวนถึง 2^I เซต โดยที่ I แทนจำนวนสมาชิกของเซตข้อมูล ซึ่งหากจำนวนข้อมูลในฐานข้อมูลมีขนาดเพิ่มขึ้น จะทำให้การค้นหารูปแบบเซตนั้นมีขนาดเพิ่มขึ้นแบบเลขยกกำลังด้วย

ดังนั้นในปีค.ศ. 1994 งานวิจัยโดย Agrawal R. และ Srikant R. ได้เสนอขั้นตอนวิธีที่ชื่อว่าเอไพโอริ (Apriori algorithm) [2] เป็นขั้นตอนวิธีที่ลดปริภูมิการค้นหา โดยที่รูปแบบเซตของข้อมูลใดที่ไม่ปรากฏบ่อย รูปแบบเซตของข้อมูลเดิมที่มีขนาดใหญ่กว่าก็ย่อมไม่ปรากฏขึ้นบ่อยด้วย ขั้นตอนวิธีดังกล่าวนี้ใช้วิธีการค้นหาแบบแนวกว้าง (Breath First Search) กล่าวคือ เริ่มจากอ่านฐานข้อมูล เพื่อนับค่าสนับสนุนรูปแบบเซตขนาดหนึ่งตัว (รูปแบบเซตที่มีจำนวนสมาชิกเท่ากับหนึ่ง) โดยหากรูปแบบเซตใดที่มีค่าสนับสนุนไม่น้อยกว่าค่าขีดแบ่งสนับสนุน จะทำการเก็บรูปแบบเซตนั้นไว้ในรูปแบบเซตที่ปรากฏบ่อยขนาดหนึ่ง ในทางตรงกันข้าม หากรูปแบบเซตใดที่มีค่าสนับสนุนน้อยกว่าค่าขีดแบ่งก็จะทำการตัดรูปแบบเซตนั้นทิ้งไป ขั้นตอนต่อมานำรูปแบบเซตที่ปรากฏบ่อยมาสร้างรูปแบบเซตที่เป็นไปได้ (Candidate Itemsets) ขนาดสอง โดยนำรูปแบบเซตมารวมกันเป็นคู่ๆ แล้วทำการนับค่าสนับสนุนของรูปแบบเซตนั้นจากฐานข้อมูล ทำการเทียบค่าสนับสนุนของรูปแบบเซตขนาดสองกับค่าขีดแบ่งสนับสนุน หากรูปแบบเซตนั้นมีค่าไม่น้อยกว่าก็ทำการเก็บในรูปแบบเซตที่ปรากฏบ่อยขนาดสอง ขั้นตอนต่อไป ทำการสร้างรูปแบบเซตที่เป็นไปได้ขนาดสาม โดยที่สมาชิกในรูปแบบเซตสองเซตที่นำมารวมกันในตำแหน่งหนึ่งจนถึงตำแหน่งรองสุดท้าย ต้องเป็นสมาชิกตัวเดียวกัน และสับเซตจากรูปแบบเซตที่ถูกสร้างที่มีขนาดเท่ากับจำนวนสมาชิกของรูปแบบเซตที่พิจารณาลบหนึ่งทุกสับเซตจะต้องเป็นรูปแบบเซตที่ปรากฏบ่อยด้วย หากสับเซตใดไม่เป็นรูปแบบเซตที่ปรากฏบ่อย ก็จะทำการตัดรูปแบบเซตที่น่าจะเป็นไปได้นั้นทิ้งไป จากนั้นทำการนับค่าสนับสนุนของรูปแบบเซตที่เป็นไปได้จากฐานข้อมูล แล้วทำการเทียบกับค่าขีดแบ่งสนับสนุน โดยที่รูปแบบเซตใดที่มีค่าไม่น้อยกว่าค่าขีดแบ่ง ก็จะนำรูปแบบเซตนั้นไปสร้างรูปแบบเซตที่เป็นไปได้ขนาดใหญ่กว่า ทำการหารูปแบบเซตที่ปรากฏบ่อยขนาดต่างๆ จนกระทั่งไม่สามารถหารูปแบบเซตที่ปรากฏบ่อยได้

<u>ตัวอย่างที่ 2.2</u> กำหนดฐานข้อมูลที่มี 8 รายการข้อมูล ดังตารางที่ 2-1 และกำหนดค่าขีดแบ่งสนับสนุนเท่ากับ 3 จะสามารถหารูปแบบเซตที่ปรากฏบ่อยสุดตามขั้นตอนวิธีเอไพโอริได้ดังนี้

ตารางที่ 2-1 ตัวอย่างฐานข้อมูลที่มีแปดรายการข้อมูล

TID	Items
T01	a, c, f, g, i
T02	a, e, g, j
T03	a, d, f, i
T04	b, d, f, i, l, m
T05	b, c, g, h, j
T06	a, b, d, i, k
T07	a, e, g, j, n
T08	d, f, h, j, l

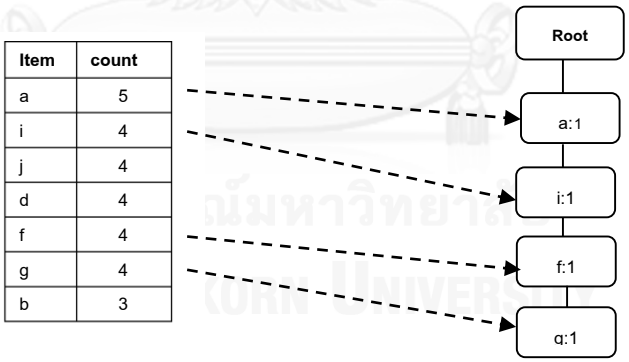
การหารูปแบบเซตที่ปรากฏบ่อยสุดโดยขั้นตอนวิธีเอไพโอริ เริ่มต้นจากอ่านฐานข้อมูล แล้วนับรูปแบบเซตขนาดหนึ่ง จะได้ {(a:5), (b:3), (c:2), (d:4), (e:3), (f:4), (g:4), (h:2), (i:4), (j:4), (k:1), (l:2), (m:1), (n:1)} จากนั้นทำการคัดรูปแบบเซตที่มีค่าสนับสนุนน้อยกว่าค่าขีดแบ่งสนับสนุน ซึ่งจะได้รูปแบบเซตที่ปรากฏบ่อยสุดขนาดหนึ่ง คือ{(a:5), (b:3), (d:4), (f:4), (g:4), (i:4), (j:4)} ขั้นตอนต่อมาจับคู่รูปแบบเซตที่ปรากฏบ่อยขนาดหนึ่ง ให้เป็นรูปแบบเซตที่เป็นไปได้ขนาดสอง ซึ่งจะได้ {(ab), (ad), (ae), (af), (ag), (ai), (aj), (bd), (be), (bf), (bg), (bi), (bj), (de), (df), (dg), (di), (dj), (ef), (eg), (ei), (ej), (fg), (fi), (fj), (gi), (gj), (ij)} จากนั้นอ่านฐานข้อมูล เพื่อนับค่าสนับสนุนของรูปแบบเซตเหล่านี้ ซึ่งจะได้ {(ab:1), (ad:2), (ae:3), (af:1), (ag:2), (ai:3), (aj:2), (bd:2), (be:0), (bf:1), (bg:1), (bi:2), (bj:1), (de:1), (df:3), (dg:0), (di:3), (dj:1), (ef:1), (eg:2), (ei:1), (ej:2), (fg:1), (fi:3), (fj:2), (gi:1), (gj:3), (ij:0)} แล้วคัดรูปแบบเซตขนาดสองที่มีค่าสนับสนุนที่ไม่น้อยกว่าค่าขีดแบ่ง ซึ่งจะได้รูปแบบเซตที่ปรากฏบ่อยขนาดสอง {(ai:3), (df:3), (di:3), (fi:3), (gj:3)} ต่อมาทำการสร้างรูปแบบเซตที่เป็นไปได้ขนาดสาม โดยการจับรูปแบบเซตที่สมาชิกที่เหมือนกันตั้งแต่ตัวหน้าสุดจนตัวรองสุดท้าย พบว่าได้รูปแบบเซต {(dfi)} จากนั้นนับค่าสนับสนุนของรูปแบบเซตดังกล่าว จะได้ {(dfi:2)} ซึ่งพบว่า เมื่อเทียบกับค่าขีดแบ่งสนับสนุน ไม่มีรูปแบบเซตที่ปรากฏบ่อยขนาดสาม ดังนั้นขั้นตอนการทำงานของเอไพโอริจะสิ้นสุดเท่านี้ โดยผลลัพธ์ที่ได้มีดังนี้ รูปแบบเซตที่ปรากฏบ่อยขนาดหนึ่งคือ {(a:5), (b:3), (d:4), (f:4), (g:4), (i:4), (j:4)} และรูปแบบเซตที่ปรากฏบ่อยขนาดสองคือ {(ai:3), (df:3), (di:3), (fi:3), (gj:3)}

จากนั้นปีค.ศ. 2000 มีงานวิจัยที่พัฒนาขั้นตอนวิธีการหารูปแบบเซตที่ปรากฏบ่อย ซึ่งสามารถลดเวลาในการสร้างรูปแบบเซตที่เป็นไปได้ และทำอ่านฐานข้อมูลเพียงสองครั้ง ขั้นตอนวิธีนี้มีชื่อว่า เอฟพี-โกรธ (Frequent Pattern Growth: FP-growth) [9] ซึ่งขั้นตอนวิธีนี้ใช้โครงสร้างต้นไม้ที่ชื่อ เอฟพี-ทรี (Frequent Pattern tree: FP-tree) ในการค้นหา ขั้นตอนวิธีนี้เริ่มต้นจากอ่านฐานข้อมูล เพื่อนับค่าสนับสนุนรูปแบบเซตขนาดหนึ่ง หากข้อมูลใดมีค่าสนับสนุนน้อยกว่าค่าขีดแบ่งสนับสนุนก็ทำการตัดทิ้ง จากนั้นทำการเรียงค่าสนับสนุนของข้อมูลจากมากไปน้อย นำมาเก็บไว้ที่ตารางแจกแจงความถี่ของข้อมูล (Header table) ต่อมาทำการเรียงข้อมูลแต่ละรายการข้อมูลตามตารางแจกแจงความถี่ แล้วทำการอ่านฐานข้อมูลแต่ละรายการข้อมูลอีกครั้ง เพื่อสร้างต้นไม้ที่แสดงรูปแบบการเกิดขึ้นของรายการข้อมูล หากข้อมูลใดที่ปรากฏซ้ำๆ จะนับค่าสนับสนุนเพิ่มทีละหนึ่ง

ขั้นตอนต่อมาทำการหารูปแบบเซตที่ปรากฏบ่อย โดยพิจารณาข้อมูลลำดับสุดท้ายจากตารางแจกแจงความถี่เป็นลำดับแรก จากนั้นทำการสำรวจหากิ่งที่มีข้อมูลที่ทำการพิจารณาจากต้นไม้ แล้วนำเส้นทางของข้อมูลที่อยู่ก่อนหน้า (ข้อมูลที่เกิดขึ้นร่วมกัน) ทุกตัว เก็บไว้ในคอนดิชันนอลแพทเทินเบซ (Conditional Base Patterns) จากนั้นสร้างต้นไม้จากเส้นทางของข้อมูลก่อนหน้ากับข้อมูลที่พิจารณา (Conditional FP-tree) แล้วทำการนับข้อมูลที่เกิดขึ้นร่วมกันทุกๆ ตัว หากข้อมูลที่เกิดขึ้นร่วมกันใดที่มีค่าสนับสนุนไม่น้อยกว่าค่าขีดแบ่งสนับสนุนที่กำหนดไว้ ก็ถือว่าข้อมูลที่เกิดขึ้นร่วมกันนั้นกับข้อมูลที่พิจารณาเป็นรูปแบบเซตที่ปรากฏบ่อย ทำการพิจารณาข้อมูลทุกตัวในตารางแจกแจงความถี่จนถึงข้อมูลลำดับแรกสุด

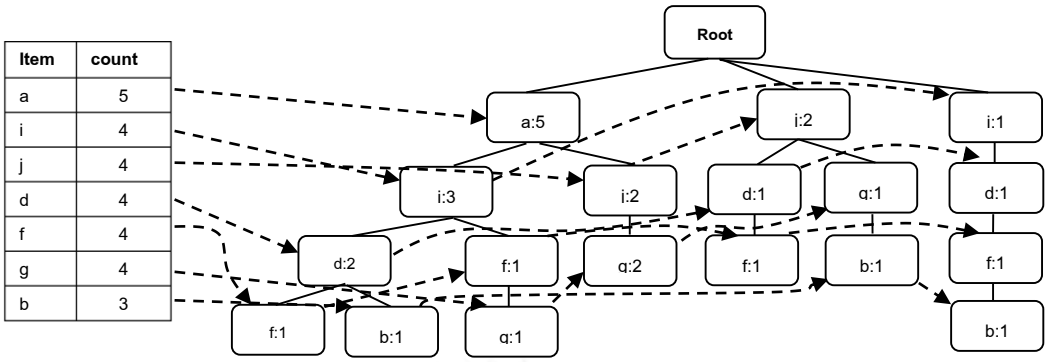
ตัวอย่างที่ 2.3 จากฐานข้อมูลที่มี 8 รายการข้อมูล ดังตารางที่ 2-1 และกำหนดค่าขีดแบ่งสนับสนุนเท่ากับ 3 จะสามารถหารูปแบบเซตที่ปรากฏบ่อยสุดตามขั้นตอนวิธีเอฟพี-โกรธ โดยเริ่มต้นจากอ่านฐานข้อมูลแต่ละรายการข้อมูล เพื่อนับค่าสนับสนุนแต่ละข้อมูล (รูปแบบเซตขนาดหนึ่ง) จากนั้นนำไปเทียบกับค่าขีดแบ่งสนับสนุน ซึ่งจะได้ข้อมูลที่มีค่าไม่น้อยกว่าค่าขีดแบ่งดังนี้ {(a:5),(b:3),(d:4),(f:4),(g:4),(i:4),(j:4)} ต่อมาทำการเรียงค่าสนับสนุนของข้อมูลจากมากไปน้อย คือ {(a:5),(i:4),(j:4),(d:4),(f:4),(g:4), (b:3)} แล้วเก็บรูปแบบเซตเหล่านี้ในตารางแจกแจงความถี่ของข้อมูล

จากนั้นทำการอ่านฐานข้อมูลแต่ละรายการอีกครั้ง โดยตัดข้อมูลที่ไม่ปรากฏในตารางแจกแจงความถี่ออก ต่อมาทำการเรียงข้อมูลแต่ละรายการตามตารางแจกแจงความถี่ แล้วสร้างเป็นต้นไม้แสดงรายการการเกิดขึ้น เช่น รายการที่ 1 {a, c, f, g, i} เมื่อตัดข้อมูลและเรียงลำดับใหม่จะได้ {a, i, f, g} จากนั้นนำมาสร้างต้นไม้ ดังภาพ 2-1 โดยบัพ (node) แรกสุดที่ต้องสร้างคือ บัพราก (root node)



ภาพที่ 2- 1 ต้นไม้แสดงรูปแบบการเกิดขึ้น เมื่ออ่านรายการข้อมูลแรก

เมื่ออ่านทุกรายการข้อมูลที่ทำการตัดและเรียงข้อมูลตามตารางแจกแจงความถี่ จะได้ต้นไม้แสดงรายการการเกิดขึ้นของข้อมูล ดังภาพที่ 2-2



ภาพที่ 2- 2 ต้นไม้แสดงรูปแบบการเกิดขึ้น เมื่ออ่านทุกรายการข้อมูล

ขั้นตอนหลังจากได้ต้นไม้แสดงรายการแล้ว คือ หารูปแบบเซตที่ปรากฏบ่อย โดยเริ่มพิจารณาจากข้อมูลอันดับสุดท้ายในตารางแจกแจงความถี่ คือ ‘b’ จากนั้นสำรวจกิ่งที่มีบัพ ‘b’ ปรากฏอยู่ จะได้ {(b,d,i,a:1), (b,g,j:1), (b,f,d,i:1)} แล้วเก็บไว้ในคอนดิชันนอลแพทเทินเบซ ต่อมานับค่าสนับสนุนของข้อมูลที่เกิดขึ้นร่วมกัน ‘b’ แล้วเทียบกับค่าขีดแบ่ง หากข้อมูลใดที่มีค่าสนับสนุนน้อยกว่าค่าขีดแบ่งก็ทำการตัดทิ้ง ซึ่งไม่มีข้อมูลใดที่เกิดขึ้นร่วมกับ ‘b’ ที่มีค่าสนับสนุนไม่น้อยกว่า 3 จึงไม่จำเป็นที่จะสร้างต้นไม้ใหม่ ต่อมาพิจารณาข้อมูล ‘g’ สำรวจหากิ่งที่มีบัพ ‘g’ พบ {(g,f,i,a:1), (g,j,a:2), (g,j:1)} จากนั้นหาข้อมูลที่เกิดขึ้นร่วมกันที่มีค่าสนับสนุนไม่น้อยกว่าค่าขีดแบ่ง พบ ‘dj’ มีค่าสนับสนุนเท่ากับ 3 เมื่อสร้างต้นไม้แล้ว จะพบเพียงหนึ่งเส้นทาง ก็ถือว่า ‘dj’ เป็นรูปแบบเซตที่ปรากฏบ่อยด้วย

ทำการพิจารณารูปแบบเซตที่ปรากฏบ่อย จนถึงข้อมูลอันดับแรกสุดในตารางแจกแจงความถี่ เมื่อทำครบแล้ว จะได้ผลลัพธ์ตามขั้นตอนวิธีเอฟพี-โกรธ ดังตารางที่ 2-2

ตารางที่ 2-2 รูปแบบเซตที่ปรากฏบ่อยสุดจากขั้นตอนวิธีเอฟพี-โกรธ

Item	Frequent itemsets (itemset:support)
b	{(b:3)}
g	{(g:4),(gj:3)}
f	{(f:4),(fd:3),(fi:3)}
d	{(d:4),(di:3)}
j	{(j:4)}
i	{(i:4),(ai:3)}
a	{(a:5)}

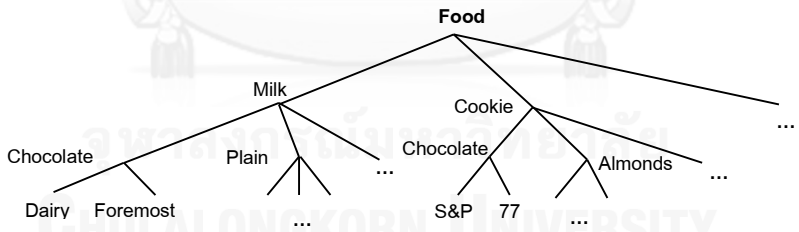
นอกจากนี้มีงานวิจัยที่เสนอการหารูปแบบเซตที่ปรากฏบ่อยจากต้นไม้แสดงรูปแบบการเกิดขึ้น เพื่อลดการสร้างต้นไม้ซ้ำๆ (non-recursive) เช่น งานวิจัย [10] เสนอขั้นตอนวิธีชื่อว่าโคพี-ทรี (COFI-tree) โดยสร้างต้นไม้ข้อมูลที่เกิดขึ้นร่วมกันกับข้อมูลที่พิจารณาเพียงครั้งเดียว ต้นไม้ดังกล่าวเป็นต้นไม้แบบสองทิศทาง ซึ่งข้อดีก็คือ ลดการสร้างต้นไม้ซ้ำๆ สำหรับการหารูปแบบเซตที่ปรากฏบ่อย อีกทั้งยังลดหน่วยความจำในการค้นหาผลลัพธ์ด้วย และ [11] เสนอขั้นตอนวิธีชื่อว่า ซีที-โปร (CT-PRO) ซึ่งทำการย่อยขนาดของต้นไม้แสดงรูปแบบการเกิดขึ้น โดยทำการจับข้อมูลมาเทียบกับตัวชี้ข้อมูล (index of item) และ [12] ได้เสนอขั้นตอนวิธีชื่อ เอช-ไมน (H-mine) ซึ่งพัฒนาโครงสร้าง

ข้อมูลคือ เอช-ชทรัค (H-struct) โดยเริ่มจากอ่านข้อมูลเพียงครั้งเดียว เพื่อนับความถี่ของข้อมูล จากนั้นนำรายการข้อมูลที่มีข้อมูลที่กรองจากค่าขีดแบ่ง มาเก็บไว้ในหน่วยความจำ แล้วทำการโปรเจคชัน (Projection) เพื่อหารูปแบบเซตที่ปรากฏบ่อย ข้อดีคือ สามารถประมาณการใช้หน่วยความจำในการค้นหาผลลัพธ์ได้ และเวลาที่ใช้ในการหาผลลัพธ์น้อยลง

2.2 การหากฎความสัมพันธ์แบบหลายลำดับชั้น

สิ่งของหรือเหตุการณ์ต่างๆ ที่ปรากฏในฐานข้อมูลสามารถถูกจัดแบ่งเป็นหมวดหมู่หรือกลุ่มของข้อมูล จึงทำให้ Han, J. และ Fu, Y. [3] เสนอการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้น โดยผลลัพธ์ที่ได้นอกจากให้ความสัมพันธ์ระหว่างสิ่งของที่เกิดขึ้นร่วมกันแล้ว ยังสามารถให้ความสัมพันธ์ระหว่างกลุ่มข้อมูลที่เกิดขึ้นร่วมกันด้วย ซึ่งผลลัพธ์นี้สามารถบ่งบอกหรือให้องค์ความรู้สารสนเทศมากขึ้นจากเดิมที่หากฎความสัมพันธ์แบบลำดับชั้นเดียว (Single-level Association Rules) ตัวอย่างเช่น กฎความสัมพันธ์ นมจืด → ขนมปังสังขยา จะทำให้ได้กฎความสัมพันธ์ระหว่างกลุ่มข้อมูลด้วย คือ นม → ขนมปัง มากกว่านั้น การหากฎความสัมพันธ์แบบหลายลำดับชั้นยังสามารถลดปัญหาของการหากฎความสัมพันธ์แบบลำดับชั้นเดียวได้ กล่าวคือ รูปแบบเซตที่ปรากฏบ่อยนั้น อาจไม่สามารถสร้างกฎความสัมพันธ์ได้ แต่หากพิจารณากลุ่มของข้อมูล จะทำให้สร้างกฎความสัมพันธ์ของระหว่างกลุ่มข้อมูลนั้นได้ อีกประการหนึ่ง ผลลัพธ์ที่จากการหากฎความสัมพันธ์แบบลำดับชั้นเดียวหลายๆ กฎอาจจะมีความคล้ายคลึงกันในแง่ความสัมพันธ์ของข้อมูล ผู้ใช้อาจจะสามารถจับกลุ่มความสัมพันธ์ระหว่างข้อมูลนั้นๆ ซึ่งทำให้ได้องค์ความรู้สารสนเทศใหม่จากผลลัพธ์ นอกนั้นยังสามารถลดกฎความสัมพันธ์ที่ซ้ำซ้อนออกได้

สิ่งของหรือเหตุการณ์ต่างๆ นั้นมักถูกจัดแบ่งตามโครงสร้างการจัดลำดับชั้นของข้อมูล ดังภาพที่ 2-3



ภาพที่ 2- 3 โครงสร้างลำดับชั้นข้อมูล

จากภาพที่ 2-3 ซึ่งเป็นโครงสร้างการจัดแบ่งลำดับชั้นข้อมูลของผลิตภัณฑ์อาหาร พบว่า “Food” ถูกจัดอยู่ลำดับชั้นบนสุด (ลำดับชั้นศูนย์) ถือเป็นหมวดหมู่ใหญ่ที่ใช้ในการพิจารณาข้อมูลย่อยๆ “milk” ถูกจัดอยู่ลำดับชั้นหนึ่งในโครงสร้างลำดับชั้นข้อมูล ซึ่งแทนด้วยประเภทของอาหาร “chocolate milk” ถูกจัดอยู่ลำดับชั้นสอง ซึ่งแทนด้วยรสชาติของอาหาร และ “foremost chocolate milk” ถูกจัดอยู่ลำดับชั้นล่างสุด (ลำดับชั้นสาม) ถือว่าเป็นยี่ห้อสินค้าของอาหาร ผู้ใช้สามารถแทนรหัสสิ่งของของข้อมูลที่ปรากฏในฐานข้อมูลได้ กล่าวคือ “milk” แทนด้วย “1**” ถือเป็นรหัสสิ่งของ ณ ลำดับชั้นหนึ่ง “chocolate milk” แทนด้วย “11*” ถือเป็นรหัสข้อมูล ณ ลำดับ

ชั้นสอง และ “foremost chocolate milk” แทนด้วย “112” ถือเป็นรหัสสิ่งของ ณ ลำดับชั้นล่างสุด

นอกจากการจัดแบ่งข้อมูลในลักษณะโครงสร้างการจัดแบ่งลำดับชั้นข้อมูลแล้ว งานวิจัย [13] เสนอการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้น โดยใช้ข้อมูลที่แทนในรูปแบบเชิงวัตถุ (Object-Orient Model) เพื่อเน้นการค้นหาความคล้ายคลึงกันของข้อมูลให้มีง่ายขึ้น มากกว่านั้นสามารถปรับรูปแบบลำดับชั้นได้ง่ายกว่าแบบเดิม

การกำหนดเกณฑ์ค่าขีดแบ่งสนับสนุนและ/หรือค่าขีดแบ่งความเชื่อมั่นสำหรับการกฎความสัมพันธ์แบบหลายลำดับชั้น มีการกำหนดเกณฑ์ค่าขีดแบ่งอยู่ด้วยกันสองแบบ คือ 1) การกำหนดค่าขีดแบ่งสนับสนุนค่าเดียว (Uniform minimum support/confidence thresholds) คือ ทุกลำดับชั้นใช้เกณฑ์ค่าขีดแบ่งเพื่อหากฎความสัมพันธ์ที่น่าสนใจเพียงค่าเดียว จะพบว่า เมื่อกำหนดค่าขีดแบ่งที่สูงเกินไป จะทำให้ได้กฎความสัมพันธ์ที่น่าสนใจเฉพาะลำดับชั้นบนๆ ซึ่งอาจจะไม่มีผลลัพธ์ที่น่าสนใจที่ลำดับชั้นล่างๆ เลย ในทางกลับกัน เมื่อกำหนดค่าขีดแบ่งที่น้อยเกินไป อาจจะได้กฎความสัมพันธ์ที่น่าสนใจในลำดับชั้นล่างๆ แต่อาจจะได้กฎความสัมพันธ์ที่ไม่น่าสนใจในลำดับชั้นบน ซึ่งสามารถแก้ปัญหานี้ได้ด้วย 2) การกำหนดค่าขีดแบ่งหลายค่า (Reduced minimum support/confidence thresholds) กล่าวคือ แต่ละลำดับชั้นข้อมูล ใช้เกณฑ์กำหนดค่าขีดแบ่งสำหรับการหากฎความสัมพันธ์ที่น่าสนใจที่แตกต่างกัน โดยลดเกณฑ์ค่าขีดแบ่งแต่ละลำดับชั้นจากลำดับชั้นบนสู่ลำดับชั้นล่าง

งานวิจัยที่เกี่ยวข้องกับการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้น มีการเสนอขั้นตอนวิธีซึ่งสามารถแบ่งออกเป็นสองวิธี วิธีแรกคือ การหากฎความสัมพันธ์จากลำดับชั้นบนสู่ลำดับชั้นล่าง (Top-down Progressively) ซึ่งการค้นหาแบบนี้มักนำขั้นตอนวิธีเอไพโอริมาประยุกต์ใช้เพื่อหากฎความสัมพันธ์แบบหลายลำดับชั้น งานวิจัย [3] เสนอขั้นตอนวิธี T2L1 ซึ่งเริ่มจากอ่านฐานข้อมูลเดิม (T1) เพื่อนับค่าสนับสนุนของข้อมูล ณ ลำดับชั้นบนสุด (ลำดับชั้นหนึ่ง) จากนั้นทำการเก็บข้อมูลที่มีค่าสนับสนุนไม่น้อยกว่าค่าขีดแบ่ง ณ ลำดับชั้นบนสุดไว้ในตารางข้อมูลที่ปรากฏบ่อย ส่วนข้อมูลที่มีค่าสนับสนุนน้อยกว่าค่าขีดแบ่งก็จะถูกตัดทิ้ง โดยข้อมูลเหล่านี้จะถูกตัดทิ้งออกจากฐานข้อมูลเดิมด้วย จากนั้นสร้างฐานข้อมูลใหม่ (T2) ซึ่งนำมาจากฐานข้อมูลเดิมที่มีเฉพาะข้อมูลที่ปรากฏขึ้นบ่อยสุดจากตารางเก็บข้อมูลที่ปรากฏบ่อย ขั้นตอนต่อไปทำการหารูปแบบเซตที่ปรากฏบ่อยสุดขนาดสอง ณ ลำดับชั้นบนสุด โดยการจับคู่ข้อมูลจากตารางข้อมูลที่ปรากฏบ่อย นับค่าสนับสนุนของรูปแบบเซตนั้นแล้วเทียบกับค่าขีดแบ่งสนับสนุน ณ ลำดับชั้นบนสุด หากรูปแบบเซตที่มีค่าสนับสนุนมากกว่าค่าขีดแบ่ง ก็ทำการหารูปแบบเซตที่ปรากฏบ่อยขนาดมากกว่าสองจนกระทั่งไม่สามารถหารูปแบบเซตที่ปรากฏบ่อย ณ ลำดับชั้นบนสุดได้ ต่อมาพิจารณาข้อมูล ณ ลำดับชั้นถัดมา (ลำดับชั้นสอง)โดยทำการนับค่าสนับสนุนของข้อมูลจากฐานข้อมูลใหม่ แล้วเทียบค่าสนับสนุนของข้อมูลกับค่าขีดแบ่งของลำดับชั้นเดียวกัน หากข้อมูลใดที่มีค่าสนับสนุนมากกว่าค่าขีดแบ่ง ก็ทำการหารูปแบบเซตที่ปรากฏบ่อย ณ ลำดับชั้นสองขนาดตั้งแต่สองขึ้นไป ทำจนกระทั่งไม่สามารถหารูปแบบเซตที่ปรากฏบ่อย ณ ลำดับชั้นนั้นแล้ว จากนั้นก็ทำการรูปแบบเซตที่ปรากฏบ่อย ณ ลำดับชั้นอื่นๆ ซึ่งทำการหาผลลัพธ์เช่นเดียวกับการหารูปแบบเซตที่ปรากฏบ่อย ณ ลำดับชั้นสอง ทำการค้นหาผลลัพธ์จนถึงลำดับชั้นล่างสุดของข้อมูล ในงานวิจัยเดียวกันนี้ได้เสนอขั้นตอนวิธีอื่นๆ ด้วย คือ T1LA, T2LA, T1LA โดยดัดแปลงมา

จากขั้นตอนวิธี T2L1 เพื่อเปรียบเทียบประสิทธิภาพการทำงานของแต่ละขั้นตอนวิธี โดยทดสอบจากการกำหนดค่าขีดแบ่งสนับสนุนที่ต่างกัน จำนวนข้อมูลที่แตกต่างกัน และอื่นๆ

ต่อมางานวิจัย [14] เสนอการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นด้วยฟังก์ชั่น (Multi-Level Fuzzy Mining) โดยฟังก์ชั่นที่กำหนดมานั้น เป็นฟังก์ชั่นที่กำหนดช่วงค่าสนับสนุนสำหรับรูปแบบเซตที่มีจำนวนการปรากฏขึ้นของข้อมูลในแต่ละรายการข้อมูล ซึ่งจะนำมาใช้สำหรับการคำนวณหาค่าสนับสนุนที่แท้จริงของรูปแบบเซตเหล่านั้น จากนั้นจึงนำค่าสนับสนุนที่ได้จากการคำนวณ มาเทียบกับค่าขีดแบ่งสนับสนุนแต่ละลำดับชั้น เพื่อหารูปแบบเซตที่ปรากฏบ่อยต่อไป มากกว่านั้น [15] เสนอการค้นหาแบบเดียวกัน โดยที่กำหนดค่าขีดแบ่งสนับสนุนแต่ละข้อมูล จากนั้นทำการค้นหารูปแบบเซตของกลุ่มข้อมูล จากนั้นใช้ฟังก์ชั่นช่วงค่าสนับสนุนมาคำนวณกับจำนวนข้อมูลที่ปรากฏในแต่ละรายการ เพื่อหาค่าสนับสนุนที่แท้จริงของรูปแบบเซตเหล่านั้น แล้วนำค่าสนับสนุนที่ได้มาเทียบค่าขีดแบ่งของกลุ่มข้อมูลที่มีค่าน้อยสุดที่ถูกจัดอยู่ในกลุ่ม/หมวดหมู่เดียวกัน จากนั้นก็ทำการหารูปแบบเซตที่ปรากฏบ่อยสุดต่อไป

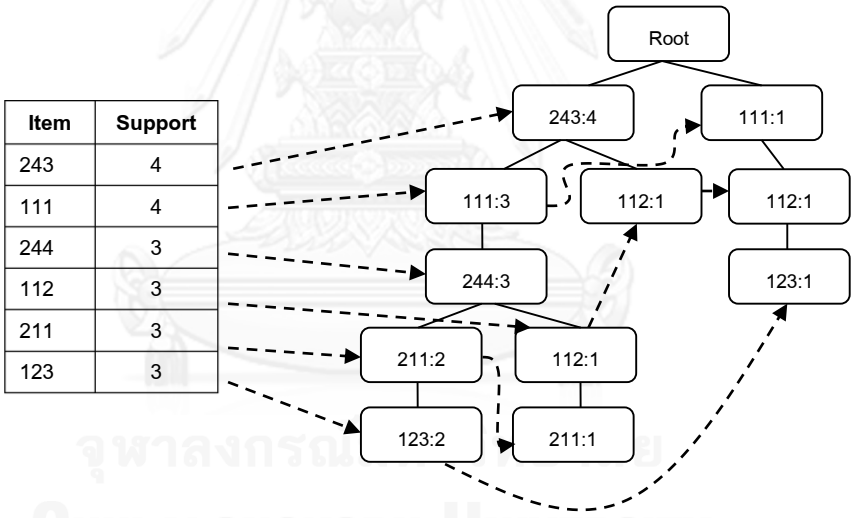
งานวิจัย [16] ได้เสนอการหากฎความสัมพันธ์แบบหลายลำดับชั้น ด้วยเกณฑ์ค่าสหสัมพันธ์ (Multi-Level Correlation) แทนการหากฎความสัมพันธ์ด้วยค่าความเชื่อมั่น ซึ่งค่าสหสัมพันธ์ของกฎความสัมพันธ์สามารถบ่งบอกว่าความสัมพันธ์ของกฎนั้นๆ มีความสอดคล้องกันหรือไม่ กล่าวคือ ถ้าค่าสหสัมพันธ์ของกฎความสัมพันธ์นั้นมีค่ามากกว่าหนึ่ง กฎความสัมพันธ์นั้นย่อมมีความสอดคล้องกันในเชิงบวก แต่ถ้าหากค่าสหสัมพันธ์ของกฎความสัมพันธ์มีค่าน้อยกว่าหนึ่ง กฎความสัมพันธ์นั้นมีความสอดคล้องกันในเชิงลบ ซึ่งประโยชน์ของค่าสหสัมพันธ์ของกฎความสัมพันธ์ สามารถช่วยให้นักวิเคราะห์ข้อมูลตัดสินใจได้ว่า กฎความสัมพันธ์ที่หาได้เป็นกฎที่มีความสอดคล้องกันของข้อมูลหรือกลุ่มข้อมูลหรือไม่

ส่วนวิธีที่สอง คือ การค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นจากลำดับชั้นล่างไปลำดับชั้นบน (Button-up Progressively) [17, 18] ในงานวิจัยกลุ่มนี้ เช่น เสนอการหารูปแบบเซตที่ปรากฏบ่อยแบบหลายลำดับชั้น (Multi-Level Frequent Pattern) โดยนำขั้นตอนวิธีเอฟพี-โกรธมาประยุกต์ใช้ ขั้นตอนการหาผลลัพธ์นี้เริ่มจากการอ่านฐานข้อมูล เพื่อนับค่าสนับสนุนของข้อมูลในแต่ละลำดับชั้น จากนั้นเรียงค่าสนับสนุนของข้อมูลจากมากไปน้อย และตัดข้อมูลที่มีค่าสนับสนุนน้อยกว่าค่าขีดแบ่งสนับสนุนที่กำหนดในแต่ละลำดับชั้น หากกลุ่มข้อมูลลำดับชั้นบนสุดใดที่มีค่าสนับสนุนน้อยกว่าค่าขีดแบ่ง ณ ลำดับชั้นเดียวกัน ข้อมูลย่อยก็จะถูกตัดทิ้งไปด้วย จากนั้นเรียงข้อมูลแต่ละรายการข้อมูลตามค่าสนับสนุนจากมากไปน้อยและอ่านฐานข้อมูลอีกครั้ง เพื่อสร้างต้นไม้แสดงรูปแบบการเกิดขึ้นของฐานข้อมูล ณ ลำดับชั้นล่างสุด แล้วทำการค้นหารูปแบบเซตที่ปรากฏบ่อยจากต้นไม้ที่ถูกสร้าง ขั้นตอนต่อมาทำการหารูปแบบเซตที่ปรากฏบ่อย ณ ลำดับชั้นถัดไป (Level $l-1$) โดยการสร้างต้นไม้แสดงรูปแบบการเกิดขึ้นจากต้นไม้ ณ ลำดับชั้นก่อนหน้า (Level l) ซึ่ง [17] พิจารณาว่ากิ่งจากต้นไม้กิ่งใด มีข้อมูลที่สามารถจัดอยู่ในกลุ่มเดียวกันได้ (รหัสข้อมูลที่เหมือนกันตั้งแต่ตัวแรกจนถึงตัวรองสุดท้าย) จะนำค่าสนับสนุนของข้อมูลเหล่านั้นมาบวกกัน เมื่อจัดกลุ่มบวกค่าสนับสนุนของกลุ่มข้อมูลจากทุกกิ่งในต้นไม้เรียบร้อยแล้ว ทำการเรียงจากค่าสนับสนุนของข้อมูลลำดับชั้นที่พิจารณาจากมากไปน้อย และสร้างต้นไม้แสดงรูปแบบการเกิดขึ้นในลำดับชั้นถัดไป ส่วน [18] พิจารณาว่ากิ่งจากต้นไม้กิ่งใด เมื่อแปลงรหัสข้อมูล ณ ลำดับชั้นที่พิจารณาแล้ว ข้อมูลใดที่มีรหัสข้อมูลตรงกัน ให้เก็บ

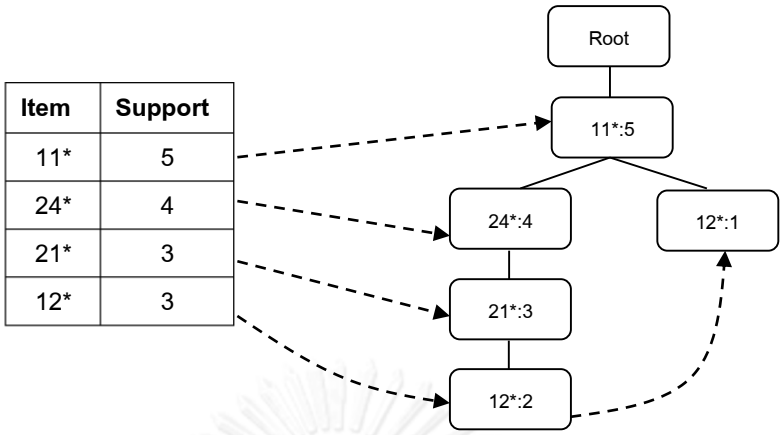
เฉพาะกลุ่มข้อมูลที่มีค่าสนับสนุนมากสุด และนับค่าสนับสนุนของกลุ่มข้อมูลนั้น ส่วนข้อมูลอื่นทำการตัดทิ้ง เมื่อพิจารณาทุกๆ ข้อมูลที่เหมือนกันในทุกกิ่งแล้ว ทำการเรียงค่าสนับสนุนของกลุ่มข้อมูลจากมากไปน้อย แล้วสร้างต้นไม้แสดงรูปแบบการเกิดขึ้น ณ ลำดับชั้นนั้น ทั้งนี้งานวิจัยนี้ได้เสนอทั้งการสำรวจแต่ละกิ่งของต้นไม้จากใบสู่ราก และการสำรวจแต่ละกิ่งของต้นไม้จากรากสู่ใบ

สังเกตได้ว่า การหากฎความสัมพันธ์จากลำดับชั้นบนสู่ลำดับชั้นล่าง จะต้องทำอ่านฐานข้อมูลทุกครั้งเมื่อพิจารณาในแต่ละลำดับชั้นข้อมูล การนับค่าสนับสนุนของรูปแบบเซตที่พิจารณาจะต้องอ่านฐานข้อมูลทุกครั้งด้วยเช่นกัน ส่วนการหากฎความสัมพันธ์จากลำดับชั้นล่างสู่ลำดับชั้นบนยังคงทำการอ่านฐานข้อมูลเพียงแค่สองครั้งตามเดิม

ตัวอย่างที่ 2.4 จากภาพที่ 2-4 ต้นไม้แสดงรูปแบบการเกิดขึ้น ณ ลำดับชั้นล่างสุด (ลำดับชั้นสาม) สามารถสร้างต้นไม้แสดงรูปแบบการเกิดขึ้น ณ ลำดับชั้นสองได้ โดยการแปลงข้อมูลจากลำดับชั้นสามสู่ลำดับชั้นสอง จับกลุ่มแต่ละข้อมูลที่มีรหัสข้อมูลตรงกันในแต่ละกิ่ง จากนั้นทำการนับค่าสนับสนุนของข้อมูล แล้วเรียงข้อมูลตามค่าสนับสนุนจากมากไปน้อย สุดท้ายทำการสร้างต้นไม้ ณ ลำดับชั้นสองจากแต่ละกิ่งในลำดับชั้นสาม ดังภาพที่ 2-5 ซึ่งจะเห็นได้ว่า สามารถช่วยลดการอ่านฐานข้อมูลในทุกๆ ครั้งที่มีการพิจารณาข้อมูลลำดับชั้นใหม่ได้



ภาพที่ 2-4 ต้นไม้แสดงรูปแบบการเกิดขึ้น ณ ลำดับชั้นล่างสุด



ภาพที่ 2-5 ต้นไม้แสดงรูปแบบการเกิดขึ้น ณ ลำดับชั้นสอง

นอกจากนี้งานวิจัย [19, 20] ได้เสนอการหากฎความสัมพันธ์แบบข้ามลำดับชั้น (Mining Cross-Level Association Rules) จากรูปแบบเซตที่ปรากฏบ่อยโดยไม่คำนึงถึงลำดับชั้นของข้อมูล โดยรูปแบบเซตแบบข้ามลำดับชั้นที่ปรากฏบ่อยนั้น จะพิจารณาจากค่าสนับสนุนของรูปแบบเซตนั้นมีค่าไม่น้อยกว่าค่าขีดแบ่งสนับสนุนที่ลำดับชั้นล่างสุดที่ข้อมูลนั้นปรากฏในรูปแบบเซต ตัวอย่างเช่น กำหนดค่าขีดแบ่งสนับสนุนของลำดับชั้นหนึ่งและลำดับชั้นสองคือ 50% และ 30% ตามลำดับ รูปแบบเซต {2**, 32*} จะเป็นเซตที่ปรากฏขึ้นบ่อยได้ ก็ต่อเมื่อค่าสนับสนุนของรูปแบบเซตดังกล่าวไม่น้อยกว่า 30% (ใช้เกณฑ์ค่าขีดแบ่ง ณ ลำดับชั้นสอง) นอกนั้นงานวิจัย [20] ได้เสนอการหารูปแบบเซตแบบข้ามลำดับชั้น โดยทำการอ่านฐานข้อมูล เพื่อหาข้อมูลในแต่ละรายการข้อมูลที่สามารถจัดอยู่กลุ่มเดียวกันได้ มาใส่เพิ่มในรายการข้อมูลนั้น และการกำหนดค่าขีดแบ่งสนับสนุนของงานวิจัยนี้จะแตกต่างจากงานวิจัยอื่นๆ คือ กำหนดค่าขีดแบ่งสนับสนุนที่ลำดับชั้นบน น้อยกว่าค่าขีดแบ่งสนับสนุนที่ลำดับชั้นล่าง หลังจากนั้นนำข้อมูลทุกลำดับชั้นที่คัดจากค่าขีดแบ่งสนับสนุนมาเรียงกันแล้วสร้างเป็นต้นไม้แสดงรูปแบบการเกิดขึ้น และทำการหารูปแบบเซตที่ปรากฏบ่อยแบบข้ามลำดับชั้นต่อไป

2.3 รูปแบบเซตที่ปรากฏบ่อยสุดจำนวนเคอันดับแรก

เนื่องจากการกำหนดค่าขีดแบ่งสนับสนุน เพื่อหากฎความสัมพันธ์ของข้อมูลหรือรูปแบบเซตที่น่าสนใจที่เหมาะสม และตรงกับความต้องการของผู้ใช้อาจกำหนดได้ยาก โดยหากกำหนดค่าขีดแบ่งสนับสนุนน้อยเกินไป ผลลัพธ์ที่ได้อาจมีจำนวนมากเกินความจำเป็น ซึ่งผู้ใช้ไม่สามารถนำผลลัพธ์ที่ได้มาใช้ประโยชน์หรือเป็นองค์ความรู้ได้ หรือถ้าหากกำหนดค่าขีดแบ่งสนับสนุนมากเกินไป อาจทำให้ไม่มีผลลัพธ์เลย ดังนั้นจึงมีงานวิจัยต่างๆ ที่เสนอการหากฎความสัมพันธ์หรือรูปแบบเซตที่ปรากฏบ่อยสุดจำนวนเคอันดับแรก (ตามจำนวนที่ผู้ใช้ต้องการ) โดยผู้ใช้เป็นผู้กำหนดค่าจำนวนผลลัพธ์ที่ต้องการ ซึ่งช่วยผู้ใช้กำหนดค่าสำหรับการค้นหาผลลัพธ์ได้ง่ายขึ้น อีกทั้งช่วยให้ผู้ใช้หลีกเลี่ยงการกำหนดเกณฑ์ค่าขีดแบ่งสนับสนุน

ดังนั้นเราจึงสามารถหารูปแบบเซตที่มีค่าสนับสนุนมากสุดจำนวนเคอันดับแรก ตามบทนิยาม

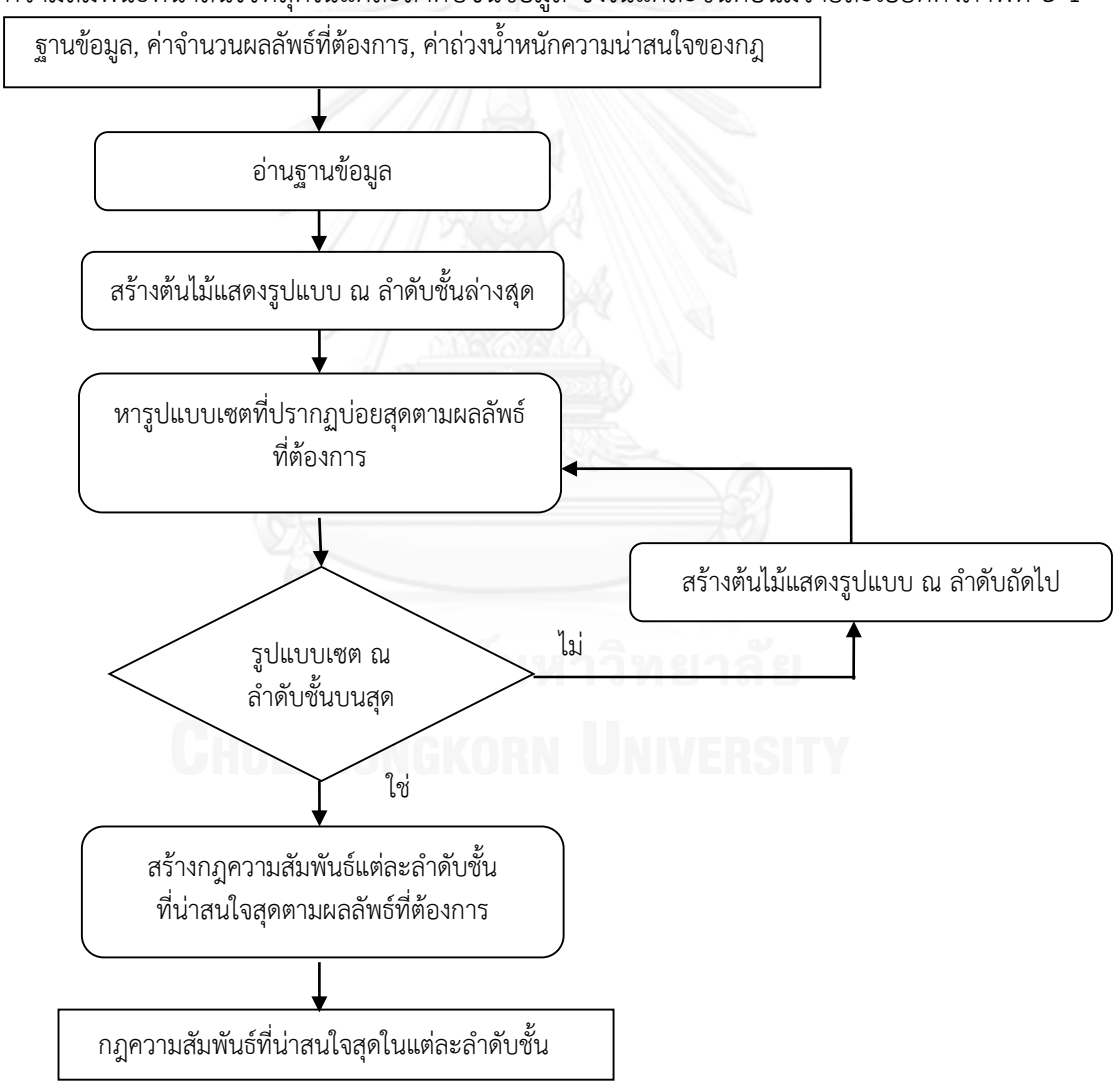
นิยามที่2.4 รูปแบบเซตหนึ่ง (itemset X) เป็นรูปแบบเซตที่ปรากฏบ่อยสุดเคอันดับแรก ถ้าไม่มีรูปแบบเซตใดที่มีค่าสนับสนุนของรูปแบบนั้นมีค่ามากกว่าค่าสนับสนุนของรูปแบบเซตที่พิจารณา

งานวิจัย [4] ได้เสนอการหารูปแบบเซตที่ปรากฏบ่อยสุดจำนวน n เซต โดยมีการกำหนดจำนวนรูปแบบเซตขนาดไม่เกิน m โดยในงานวิจัยนี้ได้เสนอ 2 ขั้นตอนวิธีคือ Itemset-Loop และ Itemset-iLoop โดย Itemset-Loop เริ่มจากนับค่าสนับสนุนของรูปแบบเซตขนาดหนึ่ง แล้วมาเรียงค่าสนับสนุนของรูปแบบเซตนั้นจากมากไปน้อย จากนั้นพิจารณารูปแบบเซตที่คาดว่าน่าจะเป็นไปได้ขนาดสองตัวขึ้นไป หากค่าสนับสนุนของรูปแบบเซตที่มีขนาดตั้งแต่ 2 ขึ้นไป มีค่ามากกว่าค่าสนับสนุนของรูปแบบเซตในอันดับสุดท้าย จะมีเปลี่ยนอันดับรูปแบบเซตที่น่าสนใจ ทำการหาไปเรื่อยๆ จนถึงรูปแบบเซตขนาด m ส่วนขั้นตอนวิธี Itemset-iLoop มีขั้นตอนวิธีที่คล้ายกับ Itemset-Loop แต่จะเริ่มพิจารณาจากรูปแบบเซตขนาด m ก่อน งานวิจัย [5] ได้เสนอการค้นหารูปแบบเซตที่ปรากฏบ่อยสุดเคอันดับแรกโดยที่ไม่มีรูปแบบเซตใดที่มีขนาดมากกว่ารูปแบบเซตนั้นมีค่าสนับสนุนเท่ากัน (Mining Top-K Frequent Closed Pattern) การค้นหาผลลัพธ์นี้ มีการกำหนดค่าขนาดรูปแบบเซตขั้นต่ำ (ขนาดของรูปแบบเซตที่ปรากฏบ่อยสุดจะต้องไม่น้อยกว่าค่าที่กำหนดไว้) งานวิจัยนี้ได้เสนอขั้นตอนวิธีทีเอฟพี (TFP algorithm) ซึ่งเริ่มจากตัดรายการข้อมูลในฐานข้อมูลที่มีขนาดของรายการข้อมูลน้อยกว่าค่าขนาดรูปแบบเซตขั้นต่ำ จากนั้นทำการสร้างต้นไม้แสดงรูปแบบที่เกิดขึ้นจากฐานข้อมูลที่ตัดรายการข้อมูลแล้ว ต่อมาพิจารณาต้นไม้แสดงรูปแบบโดยทำการแบ่งระดับชั้นของต้นไม้ด้วยค่าขนาดรูปแบบเซตขั้นต่ำ แล้วทำการหาค่าขีดแบ่งสนับสนุนจากผลรวมค่าสนับสนุนจากข้อมูลที่อยู่ภายใต้การแบ่งขนาดของรูปแบบในต้นไม้ จากนั้นหารูปแบบเซตที่ปรากฏบ่อยสุด ซึ่งพิจารณาจากข้อมูลที่มีความถี่สูงสุดก่อน โดยที่รูปแบบเซตใดมีค่าสนับสนุนมากกว่ารูปแบบเซตอันดับสุดท้ายที่ปรากฏบ่อย ก็ทำการแทรกรูปแบบเซตนั้น และตัดรูปแบบเซตอันดับสุดท้ายทิ้ง ทำการหารูปแบบเซตที่มีขนาดไม่น้อยกว่าค่าขนาดรูปแบบเซตขั้นต่ำ จนกระทั่งไม่มีรูปแบบเซตใดมีค่าสนับสนุนมากกว่ารูปแบบเซตอันดับสุดท้าย ต่อมางานวิจัย [21] ได้เสนอการหารูปแบบเซตที่ปรากฏบ่อยสุด โดยใช้ต้นไม้ที่พิจารณาข้อมูลที่เกิดขึ้นร่วมกันในการหารูปแบบเซตที่ปรากฏบ่อย ซึ่งผลลัพธ์ที่ได้ให้รูปแบบเซตที่ปรากฏบ่อยทุกๆขนาด (จำนวนสมาชิก) ทั้งหมดที่หาได้ จากนั้น งานวิจัย [22] ได้เสนอการหารูปแบบเซตที่ปรากฏบ่อยสุดที่มีขนาดของรูปแบบเซตมากสุดจำนวนเคอันดับแรก (Mining Top-K Maximum Frequent Pattern) ซึ่งขั้นตอนวิธีนี้ค้นหาคำตอบด้วยการหาอัตราส่วนของการปรากฏขึ้นร่วมกันระหว่างข้อมูลสองข้อมูล จากนั้นนำอัตราส่วนนี้ไปสร้างเป็นโครงสร้างข้อมูลแบบกราฟ และนำอัตราส่วนที่มีค่ามากในกราฟที่ถูกสร้าง ทำการหารูปแบบเซตที่ปรากฏบ่อยที่มีจำนวนสมาชิกในเซตมากสุดจำนวนเคอันดับแรก และ [8] ได้เสนอการหารูปแบบเซตที่ปรากฏขึ้นสม่ำเสมอจำนวนเคอันดับแรก (Mining Top-K Regular Frequent Pattern) โดยผู้ใช้จะต้องกำหนดค่าความสม่ำเสมอขั้นต่ำ งานวิจัยนี้พิจารณาด้วยว่ารูปแบบเซตที่ปรากฏบ่อยสุดมีการปรากฏขึ้นอย่างสม่ำเสมอหรือไม่ โดยที่นับระยะห่างจากรายการข้อมูลล่าสุดที่รูปแบบเซตนั้นปรากฏจนถึงรายการข้อมูลต่อไปที่รูปแบบเซตเดียวกันนั้นปรากฏอยู่ ซึ่งขั้นตอนในการหาผลลัพธ์นี้ได้ใช้โครงสร้างข้อมูลแบบลิงค์ลิตส์ (Link-List) โดยอ่านฐานข้อมูลทีละรายการข้อมูลเพื่อ

เก็บรูปแบบเซต ค่าสนับสนุนของรูปแบบเซต ระยะห่างที่รูปแบบเซตนั้นปรากฏในแต่ละรายการข้อมูล และเซตของรหัสรายการข้อมูลที่รูปแบบเซตนั้นปรากฏ จากนั้นคัดเลือกข้อมูลที่มีค่าสนับสนุนมากสุดจำนวนเคอันดับแรก และค่าความสม่ำเสมอของรูปแบบเซตที่ปรากฏจะต้องไม่น้อยกว่าค่าความสม่ำเสมอที่กำหนดไว้ จากนั้นจึงนำรูปแบบเซตที่ได้แต่ละตัวมาจับคู่กัน และทำการหารหัสรายการข้อมูลที่ปรากฏซ้ำกัน เพื่อหารูปแบบเซตปรากฏขึ้นบ่อยอย่างสม่ำเสมอขนาดตั้งแต่สองต่อไป และ [23] ได้เสนอการหากฎความสัมพันธ์ที่ปรากฏบ่อยที่สุด (Mining Top-K Association rules) โดยผู้ใช้กำหนดเพียงค่าจำนวนผลลัพธ์ที่ต้องการ เพื่อหากฎความสัมพันธ์ที่ปรากฏบ่อยที่สุดตามจำนวนที่ต้องการ

บทที่ 3
การค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจตามจำนวนที่ผู้ใช้ต้องการ

ในบทนี้จะกล่าวถึงการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจมากสุดตามจำนวนที่ผู้ใช้ต้องการ ซึ่งได้นำแนวคิดจากการจัดจำแนกหมวดหมู่ของข้อมูล และการค้นหารูปแบบเซตที่ปรากฏบ่อยสุดจำนวนเคอันดับแรกรวมเข้าด้วยกัน โดยการค้นหาคำตอบนี้ได้นำโครงสร้างต้นไม้แสดงรูปแบบการเกิดขึ้นของข้อมูลมาประยุกต์ใช้ ซึ่งเราสามารถแบ่งขั้นตอนการค้นหาออกเป็น 4 ขั้นตอน คือ การสร้างต้นไม้ที่แสดงรูปแบบการเกิดขึ้นจากฐานข้อมูล ณ ลำดับชั้นล่างสุด (ลำดับชั้นเดิมของฐานข้อมูล) การค้นหารูปแบบเซตที่ปรากฏบ่อยสุดตามจำนวนที่ผู้ใช้ต้องการจากต้นไม้ที่ถูกสร้างขึ้น การสร้างต้นไม้แสดงรูปแบบการเกิดขึ้นของข้อมูล ณ ลำดับชั้นสูงกว่า และการค้นหากฎความสัมพันธ์ที่น่าสนใจที่สุดในแต่ละลำดับชั้นข้อมูล ซึ่งในแต่ละขั้นตอนมีรายละเอียดดังภาพที่ 3-1



ภาพที่ 3-1 ขั้นตอนการทำงานการหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจที่สุด

3.1 การสร้างต้นไม้แสดงรูปแบบการเกิดขึ้นจากฐานข้อมูล

สำหรับส่วนนี้จะกล่าวถึงขั้นตอนสำหรับการสร้างต้นไม้ที่แสดงรูปแบบการเกิดขึ้นจากฐานข้อมูล ในลักษณะคล้ายกับขั้นตอนวิธีเอฟพี-โกรธ [9] เริ่มจากการสร้างตารางแจกแจงความถี่ของข้อมูล ซึ่งเก็บข้อมูลที่ปรากฏในฐานข้อมูลและจำนวนความถี่ที่ปรากฏของข้อมูลนั้น จากนั้นทำการอ่านฐานข้อมูลทีละรายการเพื่อนับความถี่ของข้อมูลที่ปรากฏขึ้นจากรายการนั้น ทำการนับความถี่ของข้อมูลจนกระทั่งถึงรายการสุดท้าย ต่อมาทำการเรียงข้อมูลใหม่ด้วยความถี่ที่ปรากฏขึ้นจากมากไปน้อย แล้วทำการอ่านฐานข้อมูลทีละรายการที่ถูกเรียงลำดับข้อมูลแล้วอีกครั้งเพื่อทำการสร้างต้นไม้แสดงรูปแบบการเกิดขึ้น หากข้อมูลใดที่ปรากฏในต้นไม้แล้วให้นับความถี่เพิ่มทีละหนึ่ง

ตัวอย่างที่ 3.1 จากฐานข้อมูลในตาราง 2-1 แต่ละข้อมูลในฐานข้อมูลทำการแปลงรหัสข้อมูลซึ่งถูกจัดแบ่งเป็นกลุ่มข้อมูลออกเป็น 3 ลำดับชั้น ตามตาราง 3-1 จะได้ฐานข้อมูลที่ทำการแปลงข้อมูลแล้วตามตาราง 3-2 โดยในตัวอย่างนี้ผู้ใช้ต้องการผลลัพธ์ที่น่าสนใจสุดจำนวน 5 กฎแรก

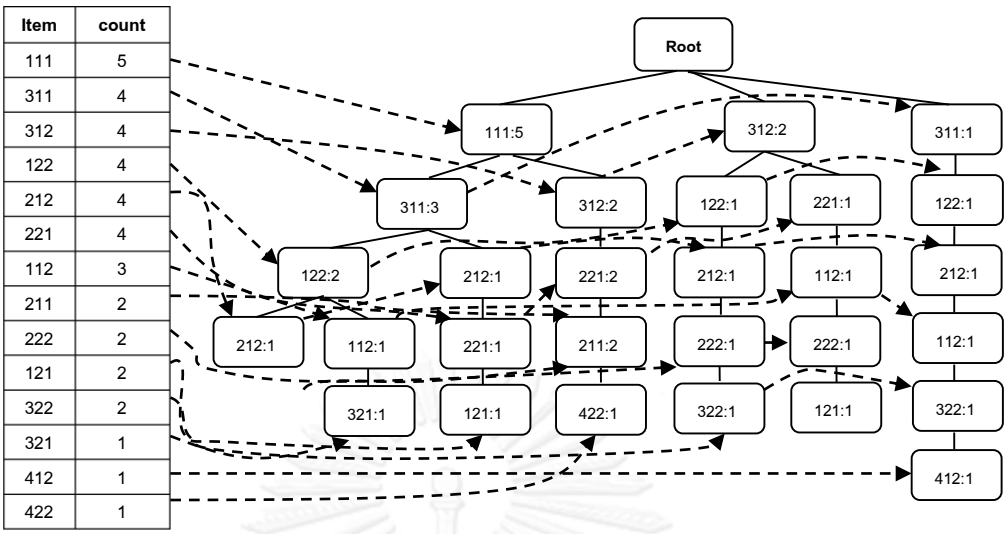
ตารางที่ 3-1 การแปลงรหัสข้อมูลของแต่ละข้อมูล

Item	Encoded item	Item	Encoded item
a	111	h	222
b	112	i	311
c	121	j	312
d	122	k	322
e	211	l	322
f	212	m	412
g	221	n	422

ตารางที่ 3-2 ฐานข้อมูลที่ทำการแปลงรหัสข้อมูลแล้ว

TID	Items
T01	111, 121, 212, 221, 311
T02	111, 211, 221, 312
T03	111, 122, 212, 311
T04	112, 122, 212, 311, 322, 412
T05	112, 121, 221, 222, 312
T06	111, 112, 122, 311, 321
T07	111, 211, 221, 312, 422
T08	122, 212, 222, 312, 322

เมื่ออ่านฐานข้อมูลจากตาราง 3-2 ทั้งสองรอบแล้ว ต้นไม้แสดงรูปแบบที่เกิดขึ้นของข้อมูล ณ ลำดับชั้นล่างสุดถูกสร้างขึ้น ตามภาพ 3-2



ภาพที่ 3-2 ต้นไม้แสดงรูปแบบการเกิดขึ้นจากฐานข้อมูล ณ ลำดับชั้นล่างสุด

3.2 การหารูปแบบเซตที่ปรากฏบ่อยสุดจากต้นไม้ที่แสดงรูปแบบการเกิดขึ้น

เมื่อได้ต้นไม้แสดงรูปแบบการเกิดขึ้นของฐานข้อมูล ณ ลำดับชั้นต่างๆ แล้ว ขั้นตอนต่อมาคือหารูปแบบเซตที่ปรากฏบ่อยสุดตามจำนวนที่ผู้ใช้ต้องการ ณ ลำดับชั้นข้อมูลที่จะพิจารณา โดยขั้นตอนนี้สามารถแบ่งออกเป็นสองขั้นตอนย่อย คือ การสร้างตารางสำหรับการเก็บรูปแบบเซตที่ปรากฏบ่อยสุด การพิจารณาข้อมูลจากต้นไม้ที่ถูกสร้างขึ้นเพื่อหาข้อมูลที่น่าจะเกิดขึ้นร่วมกันบ่อยกับข้อมูลที่พิจารณา ซึ่งในแต่ละขั้นตอนย่อยมีรายละเอียดดังนี้

3.2.1 การสร้างตารางสำหรับเก็บรูปแบบเซตที่ปรากฏขึ้นบ่อย

ทำการสร้างตารางสำหรับเก็บรูปแบบเซตที่ปรากฏบ่อยสุด ณ ลำดับชั้นที่กำลังจะพิจารณาเพื่อเก็บรูปแบบเซตที่ปรากฏบ่อยและค่าสนับสนุนของรูปแบบเซตนั้นตามจำนวนที่ผู้ใช้ต้องการ (N)

3.2.2 การหารูปแบบเซตที่ปรากฏขึ้นบ่อย

ขั้นตอนต่อมาทำการค้นหารูปแบบเซตที่ปรากฏขึ้นบ่อย (เซตที่มีค่าสนับสนุนมากๆ) โดยพิจารณาจากข้อมูลที่มีความถี่สูงสุดอันดับสองจนถึงข้อมูลอันดับสุดท้ายในตารางแจกแจงความถี่ของข้อมูล จากนั้นทำการค้นหาข้อมูลที่เกิดขึ้นร่วมกัน (‘Y’) และอยู่เหนือกว่าข้อมูลที่กำลังพิจารณา (‘X’) ในแต่ละกิ่งของต้นไม้ที่แสดงรูปแบบการเกิดขึ้น (ข้อมูลนั้นมีค่าความถี่ไม่น้อยกว่าค่าสนับสนุนของข้อมูลที่พิจารณาอยู่) แล้วทำการนับค่าสนับสนุนของข้อมูลแต่ละตัวที่เกิดขึ้นร่วมกับข้อมูลที่พิจารณา เมื่อนับความถี่ของข้อมูลที่เกิดขึ้นร่วมกันทุกกิ่งแล้ว หากค่าสนับสนุนของข้อมูลที่เกิดขึ้นร่วมกันกับข้อมูลที่พิจารณา (‘XY’) ไม่น้อยกว่าค่าสนับสนุนของรูปแบบเซตอันดับสุดท้าย (อันดับที่ N) ในตารางเก็บรูปแบบเซตที่ปรากฏบ่อยแล้ว ก็ทำการแทรกรูปแบบเซตข้อมูลที่เกิดขึ้นร่วมกับข้อมูลที่พิจารณาและทำการลบรูปแบบเซตอันดับสุดท้ายทิ้ง และถ้าหากรูปแบบเซตใหม่ (‘XY’) มีจำนวนมากกว่า 1 รูปแบบ จะทำการสร้างต้นไม้ที่แสดงการเกิดขึ้นร่วมกันกับข้อมูลที่พิจารณา แล้วทำการค้นหารูปแบบ

เซตขนาดใหญ่กว่า (ขนาดของเซตที่มากกว่าสอง) ที่มีค่าสนับสนุนไม่น้อยกว่ารูปแบบเซตอันดับสุดท้ายของตารางเก็บรูปแบบที่ปรากฏบ่อยสุด โดยทำการสำรวจจากต้นไม้ที่ถูกสร้างขึ้น

ตัวอย่างที่ 3.2 การค้นหารูปแบบเซตที่ปรากฏบ่อยสุดที่ลำดับชั้นล่างสุด

จากตารางแจกแจงความถี่จากต้นไม้แสดงรูปแบบการเกิดขึ้น ณ ลำดับชั้นที่สาม ตามภาพที่ 3-2 ผู้ใช้ต้องการผลลัพธ์ที่น่าสนใจมากสุด 5 ผลลัพธ์แรก ค้นหารูปแบบเซตที่ปรากฏบ่อยสุดโดยเริ่มต้นพิจารณาข้อมูล ‘311’ ซึ่งเป็นข้อมูลที่มีค่าสนับสนุนมากสุดอันดับสองในตารางเก็บข้อมูล จากนั้นทำการสำรวจทุกๆกิ่งที่มีข้อมูล ‘311’ ปรากฏอยู่ในต้นไม้ เพื่อค้นหาข้อมูลที่เกิดขึ้นร่วมกันและอยู่เหนือกว่าบัพ ‘311’ และนับค่าสนับสนุนของข้อมูลนั้นๆ เราพบกิ่ง {111, 311} มีค่าสนับสนุนเท่ากับ 3 และกิ่ง {311} มีค่าสนับสนุนเท่ากับ 1 ซึ่งทำให้เราทราบว่าข้อมูล ‘111’ เกิดขึ้นร่วมกับ ‘311’ สามครั้ง จากนั้นนำค่าสนับสนุนของรูปแบบเซต {(111, 311)} ไปเทียบกับค่าสนับสนุนอันดับสุดท้ายในตารางเก็บรูปแบบเซตที่ปรากฏบ่อยสุด เราพบว่าไม่มีรูปแบบเซตใดที่เก็บในตาราง ดังนั้นเราสามารถแทรกรูปแบบเซตในตารางได้

ต่อมาพิจารณาข้อมูล ‘312’ เมื่อทำการสำรวจทุกๆ กิ่งที่มี ‘312’ ปรากฏอยู่ เราได้กิ่ง {111,312} มีการปรากฏสองครั้งและกิ่ง {312} มีการปรากฏหนึ่งครั้ง ซึ่งเราได้รูปแบบเซต {(111, 312)} เพียงรูปแบบเดียว จากนั้นเทียบค่าสนับสนุนของรูปแบบเซต {(111, 312)} กับรูปแบบเซตอันดับสุดท้ายในตาราง เราพบว่ามีเพียงรูปแบบเซตที่ที่ถูกเก็บในตาราง ดังนั้นเราจึงสามารถแทรกรูปแบบเซต {(111, 312)} โดยเรียงตามค่าสนับสนุนจากมากไปน้อย

N-most table on considering item ‘311’

Itemset:support
(111,311:3)

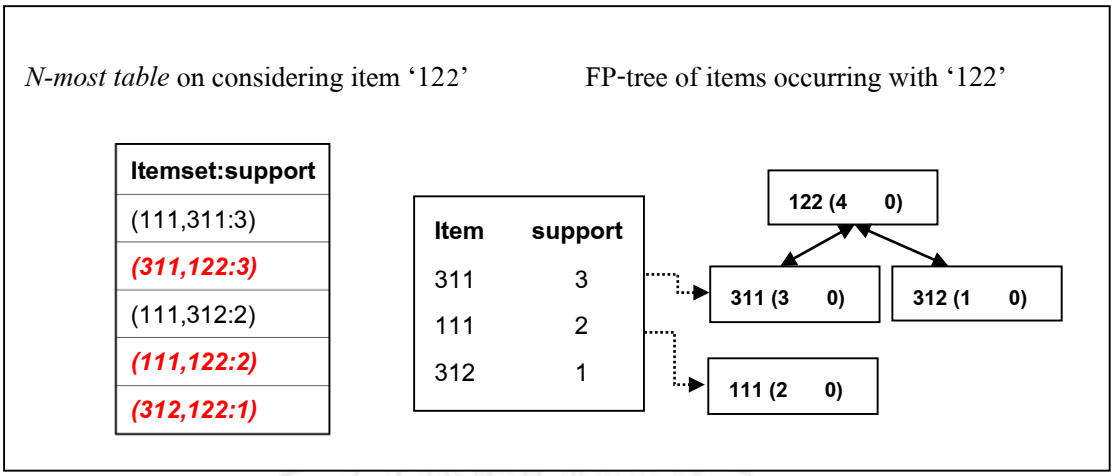
N-most table on considering item ‘312’

Itemset:support
(111,311:3)
(111,312:2)

ภาพที่ 3-3 ตารางรูปแบบเซตที่ปรากฏบ่อยเมื่อพิจารณาข้อมูล ‘311’ และ ‘312’

ต่อมาเมื่อพิจารณาและสำรวจทุกๆ กิ่งของข้อมูล ‘122’ เราพบกิ่ง {111, 311, 122} ปรากฏสองครั้ง {312, 122} ปรากฏหนึ่งครั้ง และ {311, 122} ปรากฏหนึ่งครั้ง ซึ่งพบว่ารูปแบบเซต {(111, 122)} มีค่าสนับสนุนเท่ากับ 2, รูปแบบเซต {(311, 122)} มีค่าสนับสนุนเท่ากับ 3 และรูปแบบเซต {(312, 122)} มีค่าสนับสนุนเท่ากับ 1 ซึ่งเรานำรูปแบบเซตทั้งหมดไปแทรกลงในตารางเก็บรูปแบบเซตที่ปรากฏบ่อยได้ ในกรณีนี้เราพบว่า มีข้อมูลที่เกิดขึ้นร่วมกับ ‘122’ มากกว่าหนึ่งข้อมูล ดังนั้นเราจึงสร้างต้นไม้แสดงการเกิดขึ้นร่วมกันกับข้อมูล ‘122’ พร้อมกับสร้างตารางค่าสนับสนุนของข้อมูลที่เกิดขึ้นร่วมกันกับ ‘122’ ดังภาพ 3-4

จากนั้นทำการสำรวจแต่ละกิ่งจากต้นไม้ที่ถูกสร้างเพื่อเก็บค่าสนับสนุนรูปแบบเซตที่มีขนาดมากกว่าสอง เราพบว่ารูปแบบเซต {(111, 311, 122)} มีค่าสนับสนุนเท่ากับ 2 ซึ่งเมื่อเทียบกับค่าสนับสนุนรูปแบบเซตอันดับสุดท้ายในตาราง พบว่ารูปแบบเซต {(111, 311, 122)} มีค่ามากกว่ารูปแบบเซตอันดับสุดท้าย ดังนั้นเราลบรูปแบบเซตอันดับสุดท้ายออกจากตาราง แล้วแทรกรูปแบบเซตข้างต้นแทน โดยแทรกตามค่าสนับสนุนจากมากไปน้อย



ภาพที่ 3-4 รูปแบบเซตที่ปรากฏบ่อยที่สุดเมื่อพิจารณาข้อมูล '122'

ทำการพิจารณาข้อมูลอื่นๆ ในตารางเก็บข้อมูลจนถึงข้อมูลอันดับสุดท้าย เพื่อค้นหารูปแบบเซตที่ปรากฏบ่อยสุดตามจำนวนที่ผู้ใช้ต้องการ ณ ลำดับชั้นที่สาม

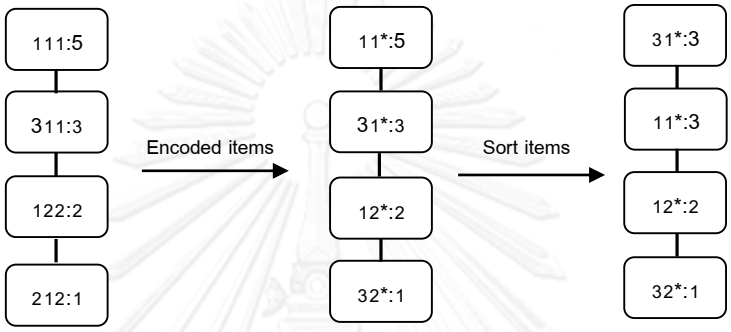
3.3 การสร้างต้นไม้แสดงรูปแบบการเกิดขึ้นของข้อมูลลำดับชั้นที่สูงกว่า

ในขั้นตอนต่อไปจะกล่าวถึงการสร้างต้นไม้แสดงรูปแบบการเกิดขึ้นของข้อมูลลำดับชั้นที่สูงกว่า (L-1) จากต้นไม้แสดงรูปแบบการเกิดขึ้นของข้อมูล ณ ลำดับชั้นเดิม (L) ด้วยการสร้างตารางที่เก็บข้อมูลลำดับชั้นใหม่ เพื่อเก็บค่าสนับสนุนข้อมูลแต่ละตัว จากนั้นทำการสำรวจแต่ละกิ่งของต้นไม้ ณ ลำดับชั้นเดิม โดยแปลงรหัสข้อมูลในแต่ละบัพข้อมูลเป็นรหัสข้อมูลลำดับชั้นใหม่ (เช่น '111' ถูกแปลงเป็น '11*' หรือ '21*' ก็ถูกแปลงเป็น '2**') เพื่อนับค่าสนับสนุนของข้อมูลนั้นลงในตารางที่เก็บข้อมูลลำดับชั้นใหม่ ถ้าหากบัพข้อมูลใดที่ถูกแปลงรหัสข้อมูลแล้วมีการปรากฏซ้ำกันอยู่ในกิ่งเดียวกัน จะนับเฉพาะค่าสนับสนุนที่มากที่สุดของข้อมูลนั้น จนกระทั่งเมื่อนับค่าสนับสนุนของข้อมูลลำดับชั้นใหม่ในทุกกิ่งแล้ว ทำการเรียงข้อมูลในตารางตามค่าสนับสนุนจากมากไปน้อย จากนั้นทำการสำรวจแต่ละกิ่งของต้นไม้ลำดับชั้นเดิมอีกครั้ง เพื่อเรียงบัพข้อมูลที่ถูกแปลงรหัสข้อมูลแล้วตามการเรียงข้อมูลในตารางเก็บข้อมูล หากบัพข้อมูลใดที่แปลงรหัสข้อมูลเหมือนกัน จะพิจารณาเฉพาะบัพที่มีค่าสนับสนุนของข้อมูลมากสุด เมื่อเรียงบัพแต่ละกิ่งแล้ว ก็จะได้กิ่งใหม่สำหรับสร้างต้นไม้แสดงรูปแบบการเกิดขึ้นของข้อมูล ณ ลำดับที่สูงกว่า และทำการลบกิ่งของข้อมูลลำดับชั้นเดิมทิ้ง

ตัวอย่างที่ 3.3 การสร้างต้นไม้แสดงรูปแบบการเกิดขึ้นของข้อมูล ณ ลำดับชั้นสอง

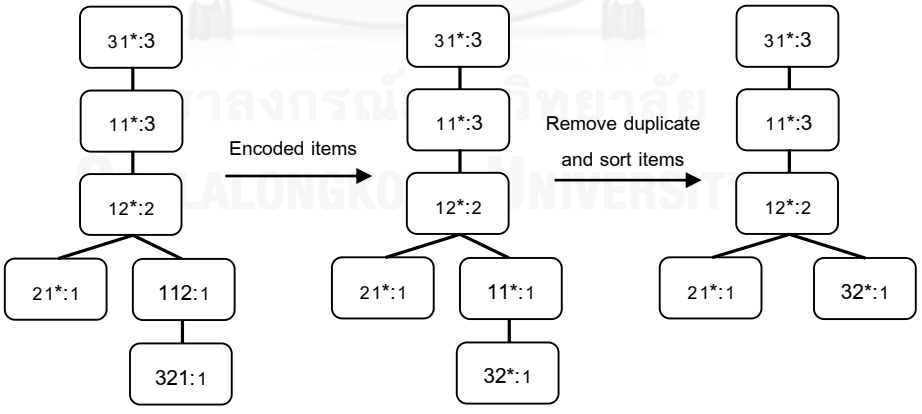
จากต้นไม้แสดงรูปแบบการเกิดขึ้นของข้อมูล ณ ลำดับชั้นล่างสุด (ลำดับชั้นที่สาม) ตามภาพที่ 3-2 เราจะสร้างต้นไม้แสดงรูปแบบการเกิดขึ้นของข้อมูล ณ ลำดับชั้นที่สอง โดยเริ่มแรกกิ่งด้านซ้ายสุดจะถูกสำรวจและแปลงรหัสข้อมูลแต่ละบัพข้อมูล จะได้ {11*:5, 31*:3, 12*:3, 21*:1} จากนั้นทำ

การนับค่าความถี่แต่ละข้อมูลที่ถูกแปลงลงในตารางเก็บข้อมูล ต่อมาพิจารณากิ่งถัดไปที่ถูกแตกออกจากบัพ ‘122’ ซึ่งจะได้กิ่งย่อย {11*:1, 32*:1} ในกรณีนี้เราพบว่า ‘11*’ ปรากฏซ้ำในกิ่งเดียวกัน เราจึงนับเฉพาะข้อมูลที่ปรากฏครั้งแรกเท่านั้น ทำการสำรวจทุกๆกิ่งจากต้นไม้ เพื่อนับค่าความถี่แต่ละข้อมูล จากนั้นทำการเรียงข้อมูลที่ลำดับชั้นใหม่ตามค่าความถี่จากมากไปน้อย ขั้นตอนต่อมาทำการสำรวจแต่ละกิ่งจากต้นไม้อีกครั้ง เริ่มพิจารณาจากกิ่งด้านซ้ายสุด {111:5, 311:3, 122:2, 212:1} จะถูกแปลงรหัสข้อมูล เรียงบัพข้อมูลตามตารางเก็บข้อมูลที่ลำดับชั้นใหม่และตัดบัพข้อมูลที่ซ้ำกันออก จะได้ได้กิ่งใหม่คือ {31*3, 11*:3, 12*:2, 21*1} ตามภาพ 3-5

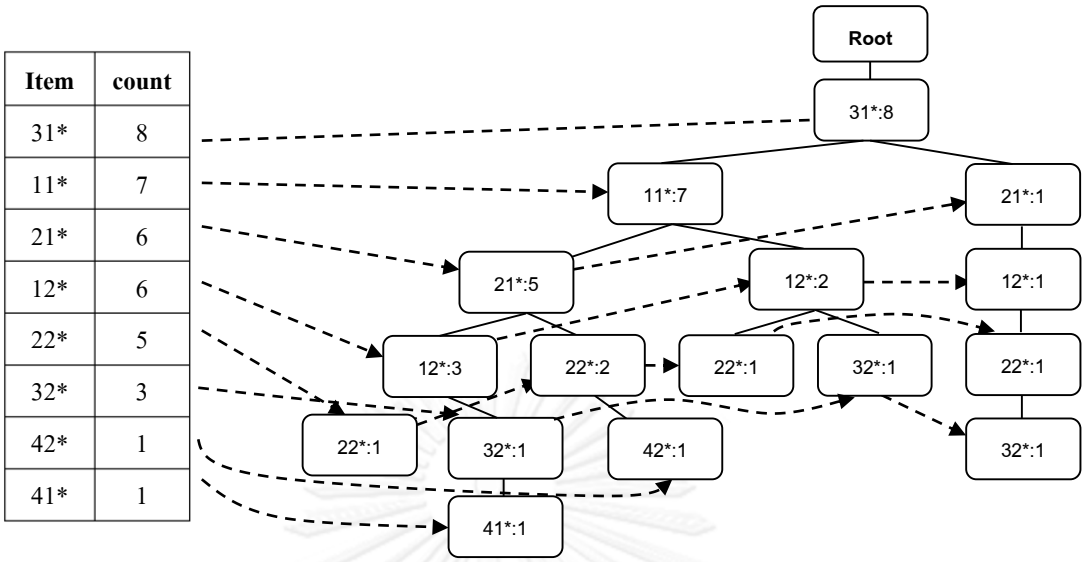


ภาพที่ 3-5 การสร้างกิ่งแรกของต้นไม้แสดงรูปแบบ ณ ลำดับชั้นสอง

จากนั้นนำกิ่งใหม่ที่ได้ไปสร้างต้นไม้ ณ ลำดับชั้นใหม่ แล้วทำการลบกิ่งที่พิจารณาเดิมทิ้ง ต่อมาพิจารณากิ่งย่อย {112:1, 321:1} ซึ่งเป็นกิ่งที่ถูกแตกย่อยจาก {111:5, 311:3, 122:2} พบว่าบัพ ‘112’ เมื่อแปลงรหัสข้อมูลแล้ว จะซ้ำกัน บัพ ‘11*’ จึงตัดบัพดังกล่าวทิ้งได้ ส่วนบัพ ‘321’ เมื่อแปลงรหัสข้อมูลแล้วจะได้บัพ ‘32*’ ซึ่งเป็นบัพใหม่ ดังนั้นเราบัพดังกล่าวถูกสร้างในกิ่งย่อยใหม่ของ {31*:3, 11*:3, 12*:2} ดังภาพ 3-6 ทำการสำรวจทุกๆกิ่งจากต้นไม้เดิม จนกระทั่งได้ต้นไม้แสดงรูปแบบการเกิดขึ้นของข้อมูล ณ ลำดับชั้นใหม่ (ลำดับชั้นที่สอง) ตามภาพ 3-7



ภาพที่ 3-6 การสร้างกิ่งต้นไม้ที่สองของต้นไม้แสดงรูปแบบ ณ ลำดับชั้นสอง



ภาพที่ 3-7 ต้นไม้แสดงรูปแบบการเกิดขึ้น ณ ลำดับชั้นสอง

เมื่อได้ต้นไม้แสดงรูปแบบการเกิดขึ้นของข้อมูลลำดับชั้นใหม่ (ลำดับชั้นที่สูงกว่า) แล้วทำการค้นหารูปแบบเซตที่ปรากฏขึ้นบ่อยสุดตามจำนวนที่ผู้ใช้ต้องการ ณ ลำดับชั้นนั้นๆ ตามข้อ 3.1 ทำการสร้างต้นไม้แสดงรูปแบบ ณ ลำดับชั้นถัดไปและค้นหารูปแบบเซตที่ปรากฏขึ้นบ่อยสุดจนกระทั่งถึงข้อมูลลำดับชั้นบนสุด จากภาพ 3-8 แสดงรูปแบบเซตที่ปรากฏบ่อยสุดในแต่ละลำดับชั้นจำนวน 5 อันดับแรกตามขั้นตอนในข้อ 3.2 และ 3.3

N-most interesting itemsets for all levels

Level 3
Itemset:support
111,311:3
122,311:3
122,212:3
212,311:3
221,312:3

Level 2
Itemset:support
31*,11*:7
31*,21*:6
31*,12*:6
11*,21*:5
31*,11*,21*:5

Level 1
Itemset:support
1**,3**:8
1**,2**:7
3**,2**:7
1**,3**,2**:7
1**,4**:2

ภาพที่ 3-8 รูปแบบเซตที่ปรากฏบ่อยสุดห้าอันดับแรกในทุกลำดับชั้น

3.4 การหากฎความสัมพันธ์ที่น่าสนใจจากรูปแบบเซตที่ปรากฏบ่อยสุด

ขั้นตอนนี้กล่าวถึงการหากฎความสัมพันธ์ที่น่าสนใจที่สุดตามจำนวนที่ผู้ใช้ต้องการจากรูปแบบเซตที่ปรากฏบ่อยสุด ตามภาพ 3-8 โดยที่ผู้ใช้เป็นผู้กำหนดค่าถ่วงน้ำหนักที่น่าสนใจของค่าสนับสนุนและ/หรือค่าความเชื่อมั่นของกฎความสัมพันธ์แต่ละกฎ

3.4.1 การสร้างกฎความสัมพันธ์

ทำการสร้างลิสต์เพื่อเก็บกฎความสัมพันธ์ที่น่าสนใจในแต่ละลำดับชั้นข้อมูลตามจำนวนที่ผู้ใช้ต้องการ จากนั้นทำการสร้างกฎความสัมพันธ์แต่ละลำดับชั้นจากรูปแบบเซตที่ปรากฏขึ้นบ่อยสุดที่ถูกเก็บในตารางรูปแบบเซตที่ปรากฏบ่อยสุด ณ ลำดับชั้นเดียวกัน โดยทำการสร้างกฎความสัมพันธ์ทีละรูปแบบเซตในตารางรูปแบบเซตที่ปรากฏบ่อยสุด ซึ่งสามารถสร้างในรูปแบบ x → I-x โดยที่ x เป็นสับเซตของ I และ x ไม่ใช่เซตว่าง จากนั้นหาค่าความเชื่อมั่นของกฎความสัมพันธ์นั้นๆ

3.4.2 การหาค่าความน่าสนใจของกฎความสัมพันธ์

หลังจากที่สร้างกฎความสัมพันธ์แต่ละกฎแล้ว ต่อมาจึงคำนวณหาค่าความน่าสนใจของกฎความสัมพันธ์นั้น โดยสามารถคำนวณได้จากสมการที่

$$I_R = w_c \times Conf^R + w_s \times Supp^R$$

โดย I_R คือค่าความน่าสนใจของกฎความสัมพันธ์ w_c คือค่าถ่วงน้ำหนักความน่าสนใจของค่าความเชื่อมั่น ซึ่งค่านี้มีค่าตั้งแต่ 0 ถึง 1 เท่านั้น และ w_s คือค่าถ่วงน้ำหนักความน่าสนใจของค่าสนับสนุน โดย $w_s = 1 - w_c$

หลังจากคำนวณค่าความน่าสนใจของกฎความสัมพันธ์แต่ละกฎแล้ว หากกฎความสัมพันธ์ใดมีค่าความน่าสนใจของกฎมากกว่าค่าความน่าสนใจของกฎอันดับสุดท้ายในลิสต์ของกฎแล้ว ก็ทำการแทรกกฎความสัมพันธ์ใหม่เรียงตามค่าความน่าสนใจของกฎ และลบกฎที่น่าสนใจอันดับสุดท้ายทิ้งไป ทำจนกระทั่งกฎความสัมพันธ์ที่ถูกสร้างจากรูปแบบเซตอันดับสุดท้ายในตารางรูปแบบเซตที่ปรากฏบ่อยสุด

<u>ตัวอย่างที่ 3.4</u> จากตารางเก็บรูปแบบที่ปรากฏบ่อย ณ ลำดับชั้นล่างสุด (ลำดับชั้นที่สาม) ตามภาพที่ 3-8 รูปแบบเซตแรกในตารางเก็บรูปแบบที่ปรากฏบ่อย {(111, 311)} ที่มีค่าสนับสนุนเท่ากับ 0.375 (3/8) เราสามารถสร้างกฎความสัมพันธ์ได้ ‘111’ ⟶ ’311’ และ ‘311’ ⟶ ‘111’ ซึ่งแต่ละกฎมีค่าความเชื่อมั่น 0.6 (3/5) และ 0.75 (3/4) ตามลำดับ จากนั้นนำค่าสนับสนุนและค่าความเชื่อมั่นของแต่ละกฎ ไปคำนวณหาค่าความน่าสนใจของกฎ โดยกำหนดให้ค่าถ่วงน้ำหนักความสนใจของค่าสนับสนุนเท่ากับ 0.8 (ค่าถ่วงน้ำหนักความสนใจของค่าความเชื่อมั่นเท่ากับ 0.2) เมื่อคำนวณหาค่าความน่าสนใจของกฎความสัมพันธ์ทั้งสองจะได้ 0.42 และ 0.46 ตามลำดับ ต่อมาเทียบค่าความน่าสนใจของแต่ละกฎกับค่าความน่าสนใจของกฎอันดับสุดท้ายในรายการเก็บกฎที่น่าสนใจ (N-most rules) ซึ่งพบว่าไม่มีกฎใดที่ถูกเก็บในรายการนี้ จึงสามารถแทรกกฎความสัมพันธ์ทั้งสองลงในรายการกฎที่น่าสนใจ โดยเรียงตามค่าความน่าสนใจของกฎจากมากไปน้อย

พิจารณารูปแบบเซตต่อมา {(122, 311)} มีค่าสนับสนุนเท่ากับ 0.375 ซึ่งสามารถสร้างกฎได้ ‘122’ ⟶ ‘311’ มีค่าความเชื่อมั่นเท่ากับ 0.75 และ ‘311’ ⟶ ‘122’ มีค่าความเชื่อมั่นเท่ากับ 0.75 จากนั้นคำนวณหาค่าความน่าสนใจของกฎ ซึ่งแต่ละกฎได้ค่าความน่าสนใจเท่ากับ 0.45 และ 0.45 ตามลำดับ แล้วนำค่าความเชื่อมั่นไปเทียบกับกฎอันดับสุดท้าย เราพบว่าในรายการเก็บกฎที่น่าสนใจ

เก็บกฎความสัมพันธ์เพียงสองกฎ ดังนั้นเราจึงสามารถแทรกกฎความสัมพันธ์ทั้งสองในรายการเก็บกฎที่น่าสนใจ ดังภาพ 3-9

N-most rules on considering itemset '111,311'

[rule] supp; conf; interesting
[311 → 111] 0.375; 0.75; 0.45
[111 → 311] 0.375; 0.6; 0.42

N-most rules on considering itemset '122,311'

[rule] supp; conf; interesting
[311 → 111] 0.375; 0.75; 0.45
[122 → 311] 0.375; 0.75; 0.45
[311 → 122] 0.375; 0.75; 0.45
[111 → 311] 0.375; 0.6; 0.42

ภาพที่ 3-9 การสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุดจากรูปแบบเซต '111, 311' และ '122, 311'

ต่อมารูปแบบเซต {(122, 212)} สามารถผลิตกฎ '122' ⟶ '212' มีค่าความเชื่อมั่น 0.75 และค่าความน่าสนใจของกฎ 0.45 และกฎ '212' ⟶ '122' มีค่าความเชื่อมัน 0.75 ค่าความน่าสนใจของกฎ 0.45 นำค่าความน่าสนใจของทั้งสองกฎไปเทียบกับกฎอันดับสุดท้ายในรายการเก็บกฎ พบว่าทั้งสองกฎมีค่าความน่าสนใจของกฎมากกว่าค่าความสนใจอันดับสุดท้าย ดังนั้นจึงสามารถกฎทั้งสองแทรกลงในรายการเก็บกฎที่น่าสนใจได้ แล้วกฎ '111' ⟶ '311' ก็ถูกตัดทิ้ง

N-most rules on considering itemset '122,212'

[rule] supp; conf; interesting
[311 → 111] 0.375; 0.75; 0.45
[122 → 311] 0.375; 0.75; 0.45
[311 → 122] 0.375; 0.75; 0.45
[122 → 212] 0.375; 0.75; 0.45
[212 → 122] 0.375; 0.75; 0.45

N-most rules on considering itemset '212,311'

[rule] supp; conf; interesting
[311 → 111] 0.375; 0.75; 0.45
[122 → 311] 0.375; 0.75; 0.45
[311 → 122] 0.375; 0.75; 0.45
[122 → 212] 0.375; 0.75; 0.45
[212 → 311] 0.375; 0.75; 0.45

ภาพที่ 3-10 การสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุดจากรูปแบบเซต '122, 212'และ '212, 311'

รูปแบบเซต {(122, 311)} สามารถสร้างกฎ '212' ⟶ '311' มีค่าความเชื่อมั่นเท่ากับ 0.75 และค่าความน่าสนใจของกฎ 0.45 และกฎ '311' ⟶ '212' มีค่าความเชื่อมั่น 0.75 และค่าความน่าสนใจของกฎ 0.45 เมื่อเทียบค่าความน่าสนใจของกฎทั้งสองกับกฎอันดับสุดท้ายพบว่า ทั้งสองกฎมีค่าความน่าสนใจเท่ากันกับกฎอันดับสุดท้ายในรายการเก็บกฎที่น่าสนใจ ดังนั้นกฎ '212' ⟶ '311' กับ '311' ⟶ '212' จะถูกเก็บด้วย ดังภาพ 3-10

พิจารณารูปแบบเซตถัดไปจนถึงรูปแบบเซตอันดับสุดท้ายในตารางเก็บรูปแบบที่ปรากฏบ่อยสุด เพื่อหากฎความสัมพันธ์ที่น่าสนใจตามจำนวนที่ผู้ใช้ต้องการในทุกๆ ลำดับชั้นของข้อมูล ดังภาพ 3-11

N-most interesting association rules for all levels

Level 3	Level 2	Level 1
[rule] supp; conf; interesting	**[rule] supp; conf; interesting**	**[rule] supp; conf; interesting**
[311→111] 0.375; 0.75; 0.45	[11*→31*] 0.875; 1; 0.9	[1**→3**] 1; 1; 1
[122→311] 0.375; 0.75; 0.45	[31*→11*] 0.875; 0.875; 0.875	[3**→1**] 1; 1; 1
[311→122] 0.375; 0.75; 0.45	[21*→31*] 0.75; 1; 0.8	[2**→1**] 0.875; 1; 0.9
[122→212] 0.375; 0.75; 0.45	[12*→31*] 0.75; 1; 0.8	[2**→3**] 0.875; 1; 0.9
[212→311] 0.375; 0.75; 0.45	[31*→21*] 0.75; 0.75; 0.75	[2**→1**,3**] 0.875; 1; 0.9

ภาพที่ 3-11 กฎความสัมพันธ์ที่น่าสนใจที่สุดทุกๆ ลำดับชั้นจำนวนห้าอันดับแรก

บทที่ 4
การทดลองและผลการทดลอง

ในบทนี้จะกล่าวถึงการทดสอบการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจที่สุดตามจำนวนผลลัพธ์ที่ผู้ใช้ต้องการ ซึ่งได้ทำการทดสอบประสิทธิภาพด้วยขั้นตอนวิธีเอ็นเอ็มแอลเอฟพี (NMLFP algorithm - NMLFP) จากบทที่ 3 เทียบกับขั้นตอนวิธีการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นภายใต้ค่าขีดแบ่งสนับสนุนที่ชื่อว่าเอฟพีเอ็ม-ที (FPM-T) [18] โดยในงานวิจัยนี้ได้ทดสอบกับขั้นตอนวิธีเอ็นเอ็มแอลเอฟพีก่อน จากนั้นจึงนำค่าสนับสนุนและค่าความเชื่อมั่นอันดับสุดท้ายที่ได้จากผลลัพธ์มาเทียบกับการทดสอบโดยขั้นตอนวิธีเอฟพีเอ็ม-ที ในการทดสอบงานวิจัยนี้ใช้ชุดข้อมูลจำลองแบบกระจายตัว (Synthetic dataset) [24] จำนวนสามชุดสำหรับการทดลอง คือ ชุดข้อมูล T10I4D100K, T20I6D100K และ T40I10D100K ซึ่งทั้งสามชุดข้อมูลประกอบไปด้วยจำนวนข้อมูลทั้งหมด 1000 ตัว มีรายการข้อมูลทั้งหมด 100,000 รายการ โดยแต่ละชุดข้อมูลมีการจัดลำดับชั้นของข้อมูลเป็น 4 ลำดับชั้นสองแบบ คือ 8-5-5-5 และ 15-6-3-4 และแบ่งเป็น 3 ลำดับชั้นคือ 10-10-10 [3] การจัดแบ่งตามลำดับชั้นนี้เพื่อให้เห็นถึงความแตกต่างของผลลัพธ์ที่ได้ในการทดสอบ ซึ่งทั้งหมดมีการแบ่งการจัดลำดับชั้นข้อมูล ดังตารางที่ 4-1

ตารางที่ 4-1 ตารางแจกแจงการจัดลำดับชั้นข้อมูลแต่ละชุดข้อมูล

ชื่อชุดข้อมูล	|T|	|I|	ลำดับชั้น	ข้อมูลลำดับชั้น 1	กระจายข้อมูลจากลำดับชั้น 1 ไป 2	กระจายข้อมูลจากลำดับชั้น 2 ไป 3	กระจายข้อมูลจากลำดับชั้น 3 ไป 4
DB1	10	4	4	8	5	5	5
DB2	10	4	4	15	6	3	4
DB3	10	4	3	10	10	10	
DB4	20	6	4	8	5	5	5
DB5	20	6	4	15	6	3	4
DB6	20	6	3	10	10	10	
DB7	40	10	4	8	5	5	5
DB8	40	10	4	15	6	3	4
DB9	40	10	3	10	10	10	

โดย |T| แทนค่าเฉลี่ยของจำนวนข้อมูลที่ปรากฏในแต่ละรายการข้อมูลในฐานข้อมูล

|I| แทนค่าเฉลี่ยสูงสุดของความยาวของรูปแบบเซตที่น่าจะเป็นรูปแบบเซตที่ปรากฏบ่อย

สำหรับการทดสอบขั้นตอนวิธีในการวิจัยนี้ ใช้เครื่องมือที่มีหน่วยประมวลผล (CPU) 2.5 GHz หน่วยความจำ (Ram) ขนาด 4 กิกะไบต์ และใช้ภาษาโปรแกรม JAVA สำหรับทั้งสองขั้นตอนวิธี โดย

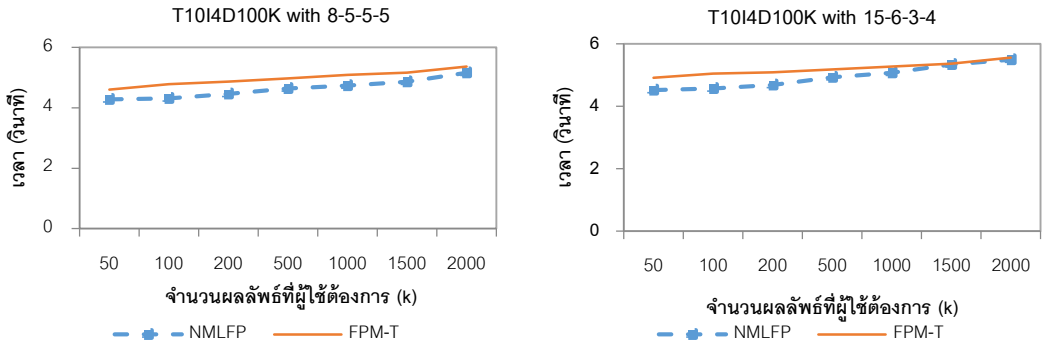
การทดสอบงานวิจัยนี้ได้แบ่งออกเป็น 2 ส่วนคือ การทดสอบวัดประสิทธิภาพเชิงเวลา และการทดสอบวัดประสิทธิภาพเชิงหน่วยความจำ นอกจากนี้ได้วิเคราะห์ผลลัพธ์การค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจสุดตามจำนวนที่ผู้ใช้ต้องการด้วย

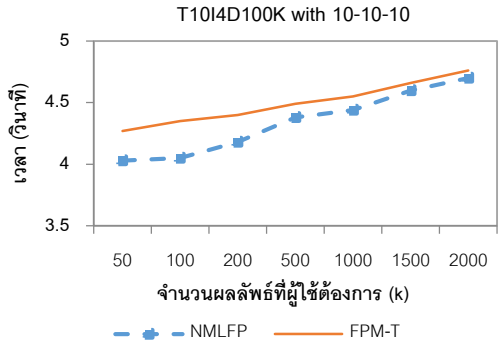
4.1 การทดสอบวัดประสิทธิภาพเชิงเวลา

สำหรับการทดสอบวัดประสิทธิภาพเชิงเวลาของงานวิจัยนี้ ได้วัดการใช้เวลาสำหรับการประมวลผลของทั้งสองขั้นตอนวิธีเทียบกันตั้งแต่ขั้นตอนแรกจนถึงได้ผลลัพธ์ออกมา โดยการทดสอบเชิงเวลานี้ได้แบ่งการทดสอบย่อยออกเป็น 2 ส่วนคือ การทดสอบขั้นตอนการค้นหารูปแบบเซตที่ปรากฏบ่อยสุด และการทดสอบขั้นตอนการสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุด ในแต่ละส่วนย่อยมีรายละเอียดดังนี้

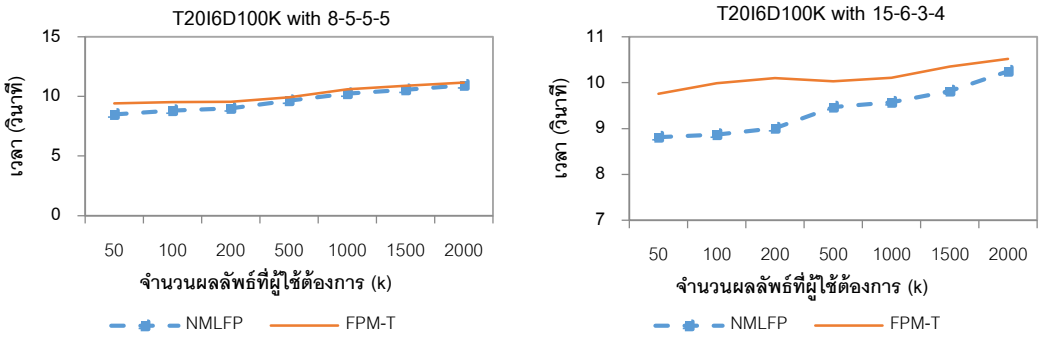
4.1.1 การทดสอบการค้นหารูปแบบเซตที่ปรากฏบ่อยสุดตามจำนวนผลลัพธ์ที่ต้องการ

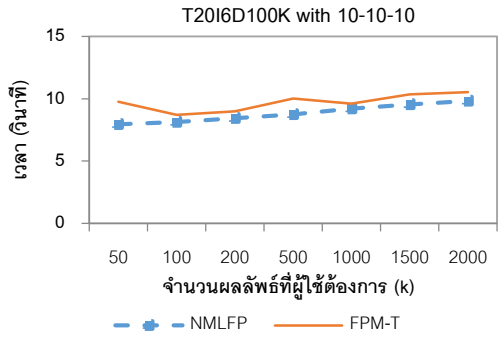
ทำการทดสอบด้วยขั้นตอนวิธีทั้งสองกับชุดข้อมูลแต่ละชุด โดยกำหนดจำนวนผลลัพธ์ตั้งแต่ 50 จนถึง 2000 พบว่าการหารูปแบบเซตที่ปรากฏบ่อยด้วยขั้นตอนวิธีเอ็นเอ็มแอลเอฟพีใช้เวลาเร็วกว่าการค้นหารูปแบบเซตด้วยขั้นตอนวิธีเอฟพีเอ็ม-ทีเล็กน้อย เนื่องจากขั้นตอนวิธีเอ็นเอ็มแอลเอฟพีเป็นค้นหาที่พิจารณาข้อมูลที่มีค่าความถี่สูงสุดก่อน จึงทำให้ได้ผลลัพธ์ที่ปรากฏบ่อยสุดเสมอ อีกทั้งมีหน้าที่เพิ่มเกณฑ์ค่าสนับสนุนเมื่อได้รูปแบบเซตตามจำนวนผลลัพธ์แรก ซึ่งสามารถตัดรูปแบบเซตที่ไม่ปรากฏขึ้นบ่อยได้ด้วย นอกจากนี้เมื่อจำนวนผลลัพธ์มีค่ามากขึ้น เวลาที่ใช้ในการค้นหารูปแบบเซตที่ปรากฏบ่อยสุดย่อมใช้เวลาเพิ่มขึ้นตามด้วย ซึ่งแสดงด้วยภาพที่ 4-1 ถึง 4-3



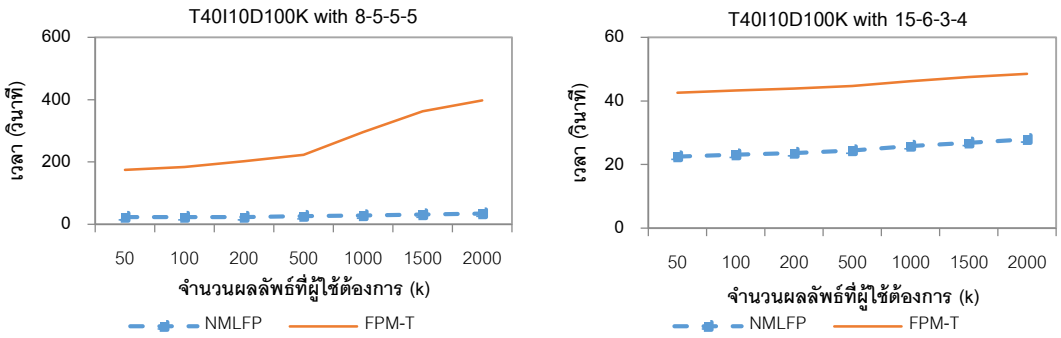


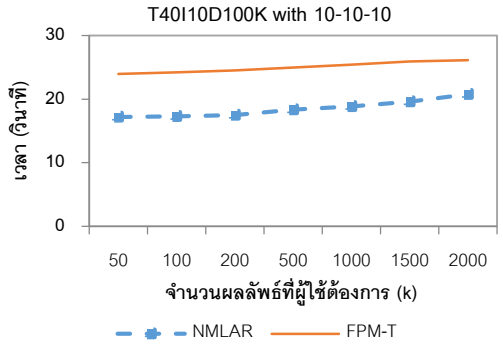
ภาพที่ 4-1 การทดสอบหารูปแบบเซตที่ปรากฏบ่อยที่สุดทุกลำดับชั้นกับชุดข้อมูล T10I4D100K





ภาพที่ 4-2 การทดสอบหารูปแบบเซตที่ปรากฏบ่อยที่สุดทุกลำดับชั้นกับชุดข้อมูล T20I6D100K

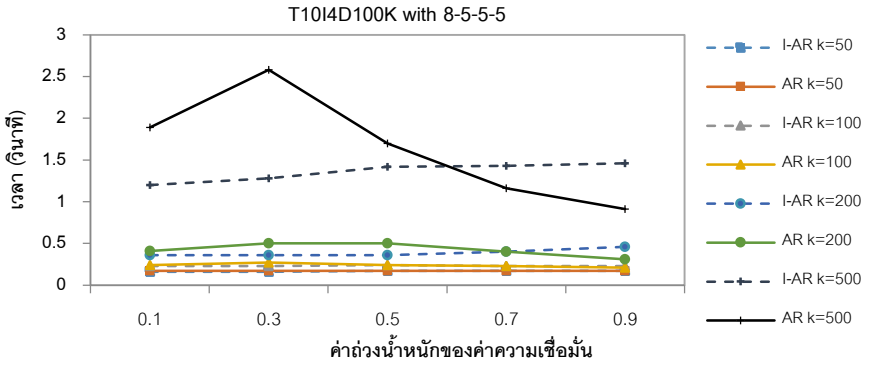


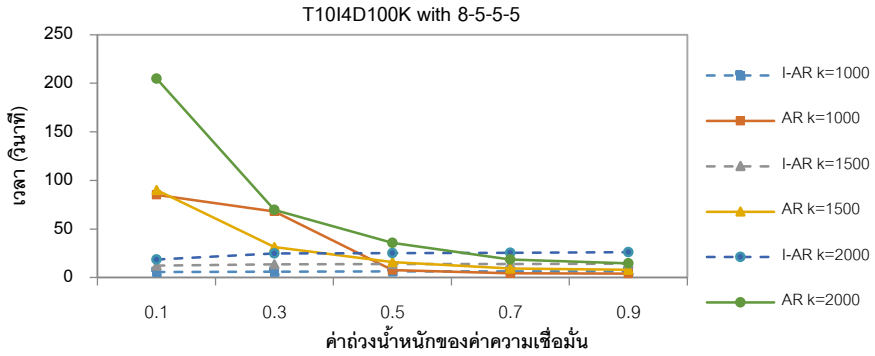


ภาพที่ 4-3 การทดสอบหารูปแบบเซตที่ปรากฏบ่อยที่สุดทุกลำดับชั้นกับชุดข้อมูล T40I10D100K

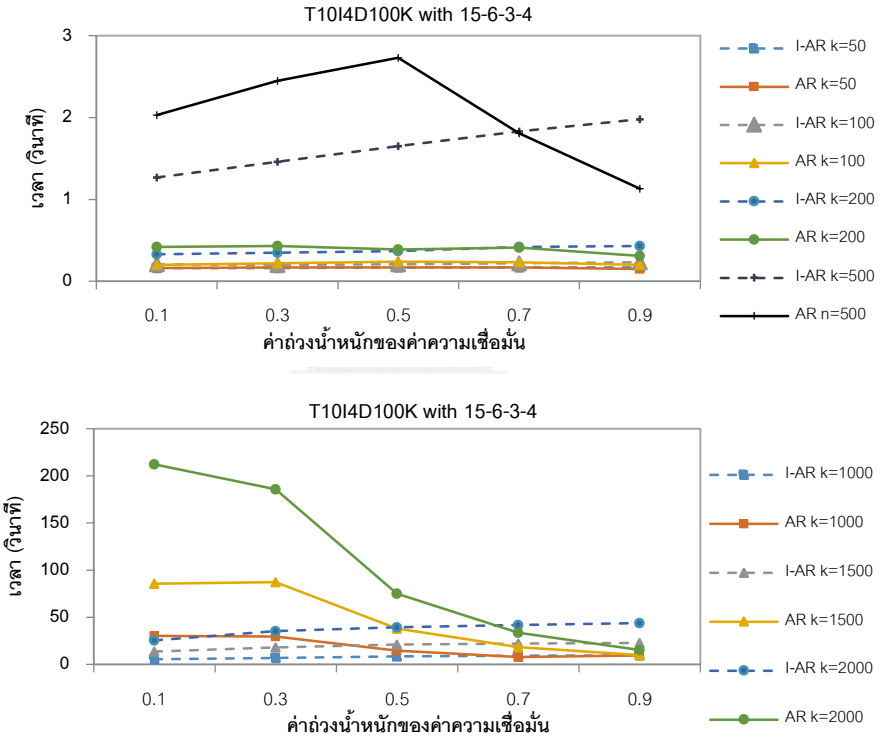
4.1.2 การทดสอบการสร้างกฎความสัมพันธ์ที่น่าสนใจสุดตามจำนวนผลลัพธ์ที่ต้องการ

ขั้นตอนนี้ทำการทดสอบกับชุดข้อมูลต่างๆ และจำนวนผลลัพธ์ที่แตกต่างกัน โดยกำหนดค่าถ่วงน้ำหนักความสนใจของผลลัพธ์ในช่วง 0.1 ถึง 0.9 ผลการทดสอบคือ เมื่อทำการทดสอบกับชุดข้อมูล T10I4D100K (DB1 ถึง DB3) จากภาพที่ 4-4 ถึง 4-6 โดยกำหนดค่าถ่วงน้ำหนักความน่าสนใจของค่าความเชื่อมั่นมีค่ามากขึ้นและกำหนดจำนวนผลลัพธ์คงที่ พบว่าการสร้างกฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจที่สุด (I-AR) จะใช้เวลามากขึ้นตามค่าถ่วงน้ำหนักความน่าสนใจของผลลัพธ์ที่เพิ่มขึ้น แต่เมื่อเทียบกับการสร้างกฎความสัมพันธ์แบบหลายลำดับชั้นด้วยค่าขีดแบ่ง (AR) จะสังเกตได้ว่า เมื่อเทียบโดยกำหนดค่าถ่วงน้ำหนักที่ 0.1 ถึง 0.5 การสร้างกฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจสุดจะใช้เวลาเร็วกว่า การสร้างกฎความสัมพันธ์แบบหลายลำดับชั้นด้วยค่าขีดแบ่ง เนื่องจากจำนวนผลลัพธ์ที่ได้จากการค้นหาด้วยค่าขีดแบ่งมีจำนวนมาก แต่เมื่อเทียบโดยกำหนดค่าถ่วงน้ำหนักที่ 0.7 ถึง 0.9 การสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุด จะใช้เวลามากกว่าการสร้างกฎความสัมพันธ์ด้วยค่าขีดแบ่ง เนื่องจากจำนวนกฎความสัมพันธ์ที่ได้จากการกำหนดค่าขีดแบ่งมีจำนวนผลลัพธ์ที่ใกล้เคียงกันกับการสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุด ซึ่งแสดงให้เห็นค่าขีดแบ่งที่ได้นั้นเป็นค่าที่เหมาะสมสำหรับจำนวนผลลัพธ์ตามที่ผู้ใช้ต้องการ

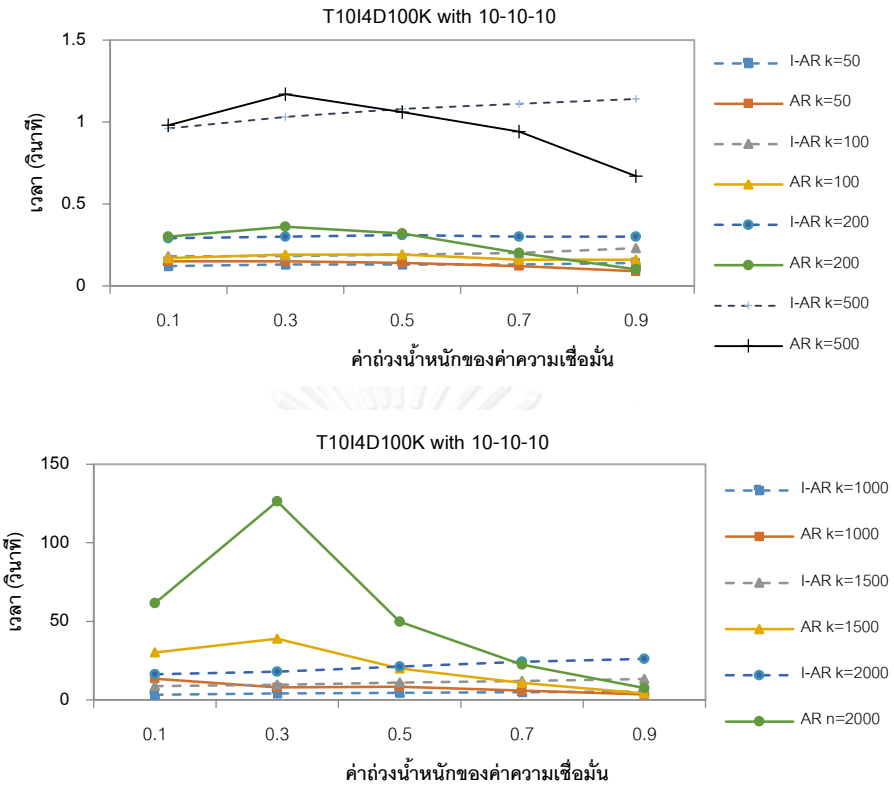




ภาพที่ 4-4 การทดสอบการสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุดกับชุดข้อมูล DB1

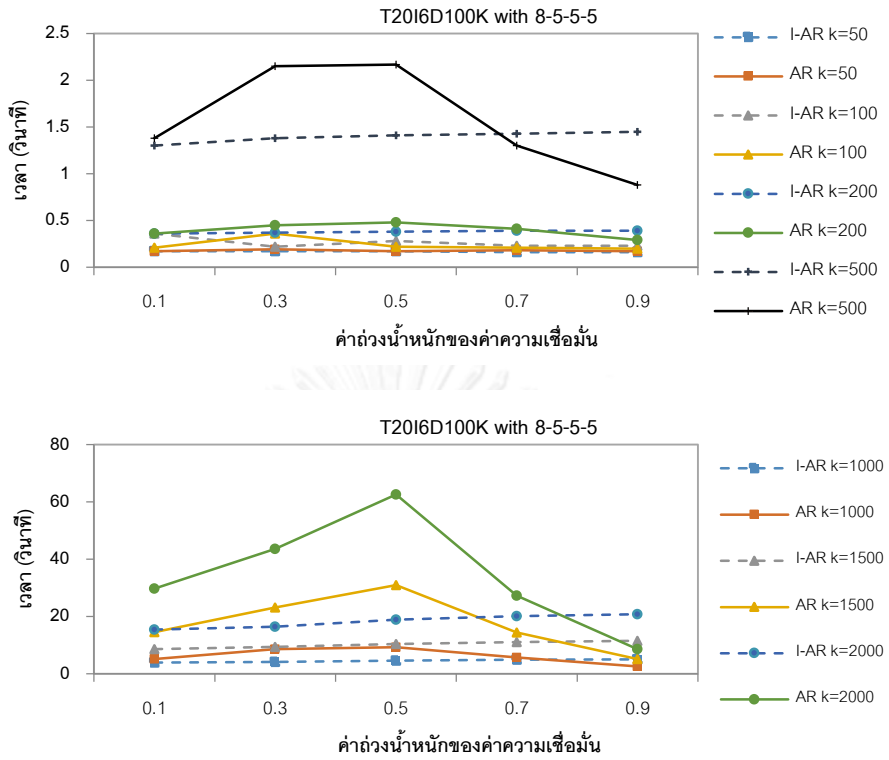


ภาพที่ 4-5 การทดสอบการสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุดกับชุดข้อมูล DB2

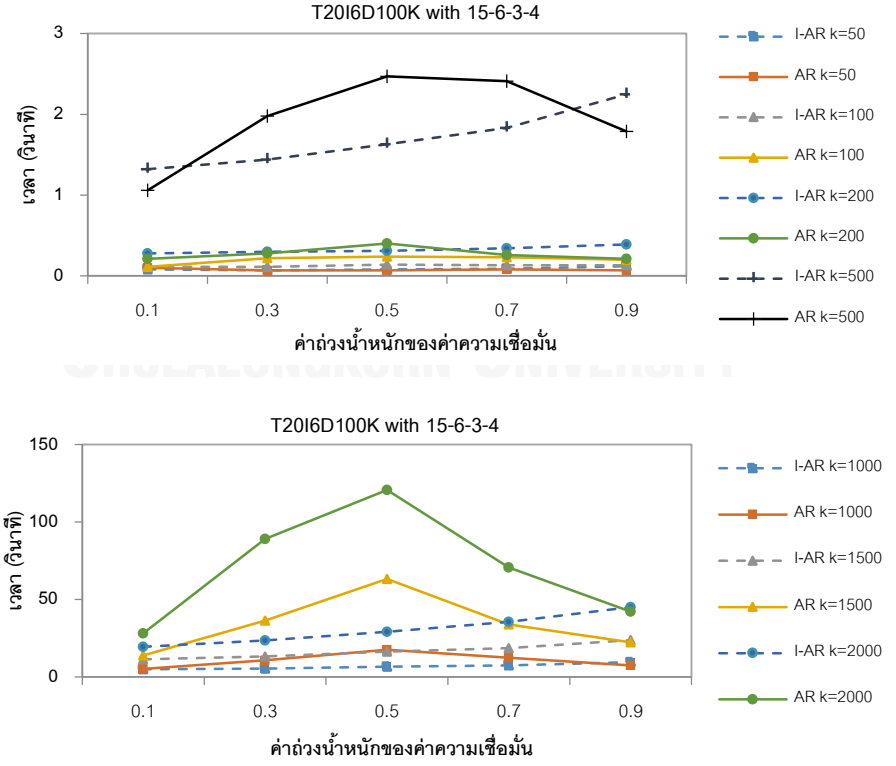


ภาพที่ 4-6 การทดสอบการสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุดกับชุดข้อมูล DB3

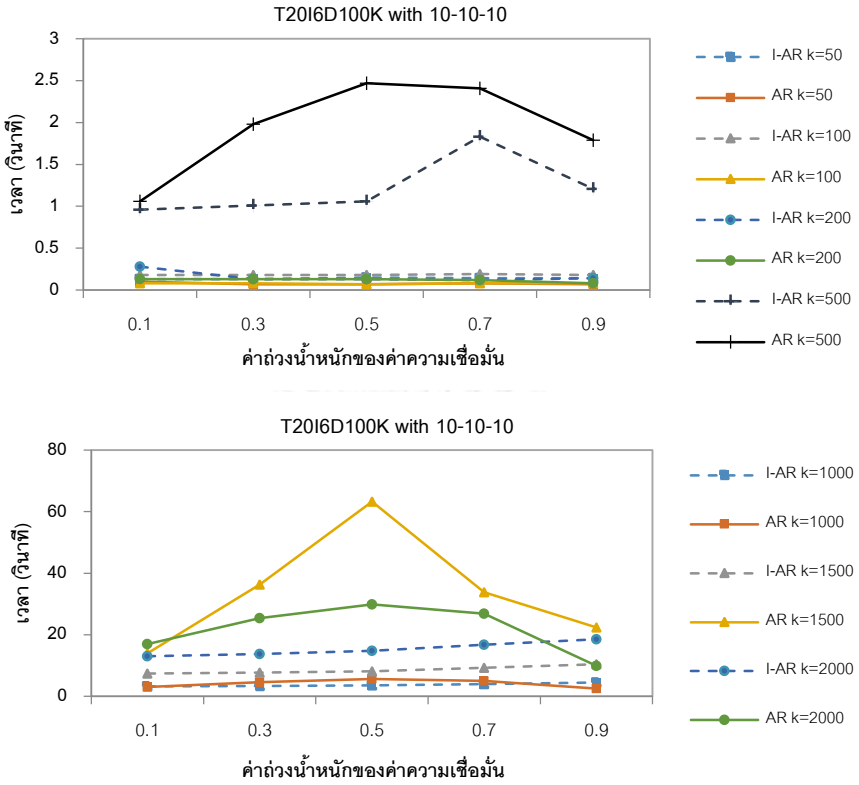
เมื่อทำการทดสอบกับชุดข้อมูล T20I6D100K (DB4 ถึง DB6) ดังภาพที่ 4-7 ถึง 4-9 เมื่อกำหนดค่าถ่วงน้ำหนักของค่าความเชื่อมั่นแต่ละค่าและกำหนดจำนวนผลลัพธ์คงที่ พบว่าการสร้างกฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจที่สุดจะใช้เวลามากขึ้น ตามค่าถ่วงน้ำหนักที่มีค่าเพิ่มขึ้น แต่เมื่อเทียบกับการสร้างกฎความสัมพันธ์แบบหลายลำดับชั้นด้วยค่าขีดแบ่ง จะสังเกตได้ว่า เมื่อเทียบโดยกำหนดค่าถ่วงน้ำหนักที่ 0.1 ถึง 0.7 การสร้างกฎความสัมพันธ์ที่น่าสนใจสุดจะใช้เวลาน้อยกว่าการสร้างกฎความสัมพันธ์ด้วยค่าขีดแบ่ง เนื่องจากจำนวนผลลัพธ์ที่ได้จากการค้นหาด้วยค่าขีดแบ่งมีจำนวนผลลัพธ์มากเกินกว่าจำนวนที่ต้องการ แต่เมื่อเทียบโดยกำหนดค่าถ่วงน้ำหนักที่ 0.9 การสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุดตามจำนวนผลลัพธ์จะใช้เวลานานกว่า การสร้างกฎความสัมพันธ์ด้วยค่าขีดแบ่ง เนื่องจากจำนวนผลลัพธ์ที่ได้จากการกำหนดค่าขีดแบ่งมีจำนวนผลลัพธ์ที่ใกล้เคียงกันกับการสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุด ซึ่งค่าขีดแบ่งที่ได้นั้นเป็นค่าขีดแบ่งที่เหมาะสมสำหรับการสร้างกฎความสัมพันธ์แบบหลายลำดับชั้นที่มีการกำหนดค่าขีดแบ่ง



ภาพที่ 4-7 การทดสอบการสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุดกับชุดข้อมูล DB4

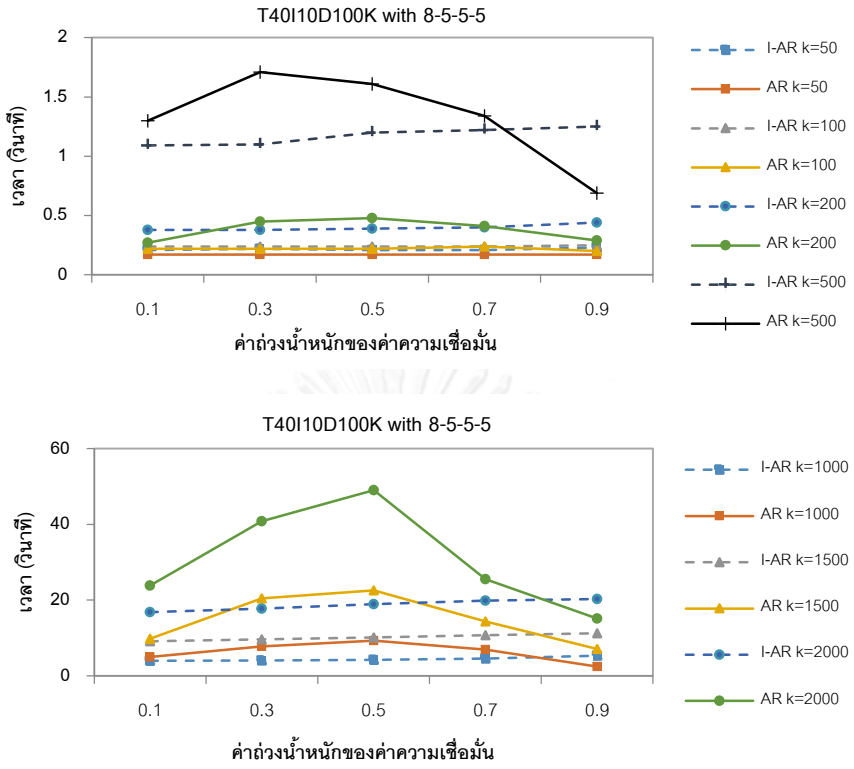


ภาพที่ 4-8 การทดสอบการสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุดกับชุดข้อมูล DB5

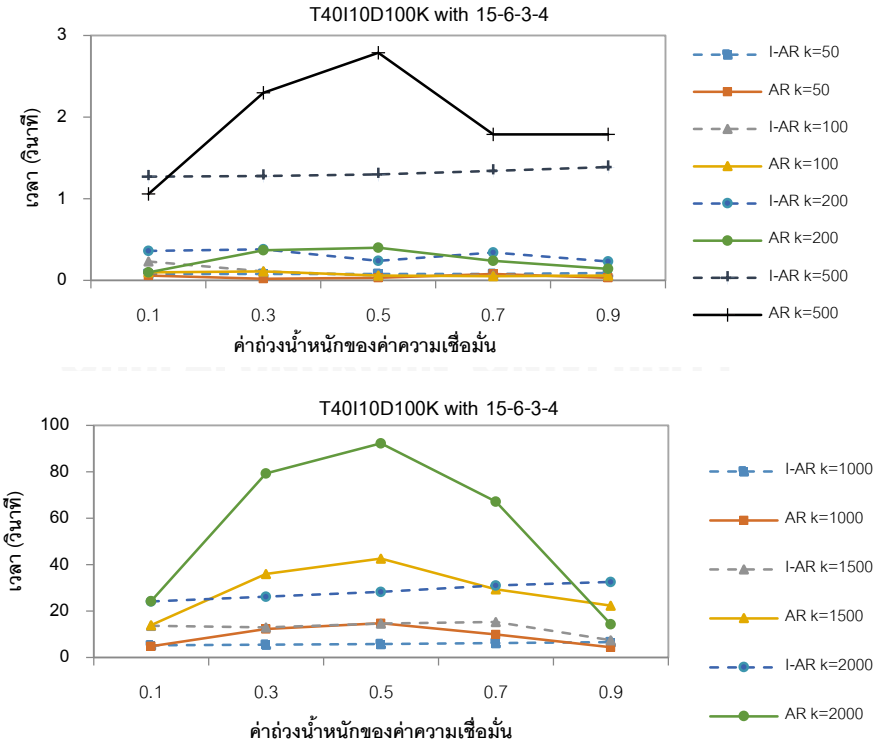


ภาพที่ 4-9 การทดสอบการสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุดกับชุดข้อมูล DB6

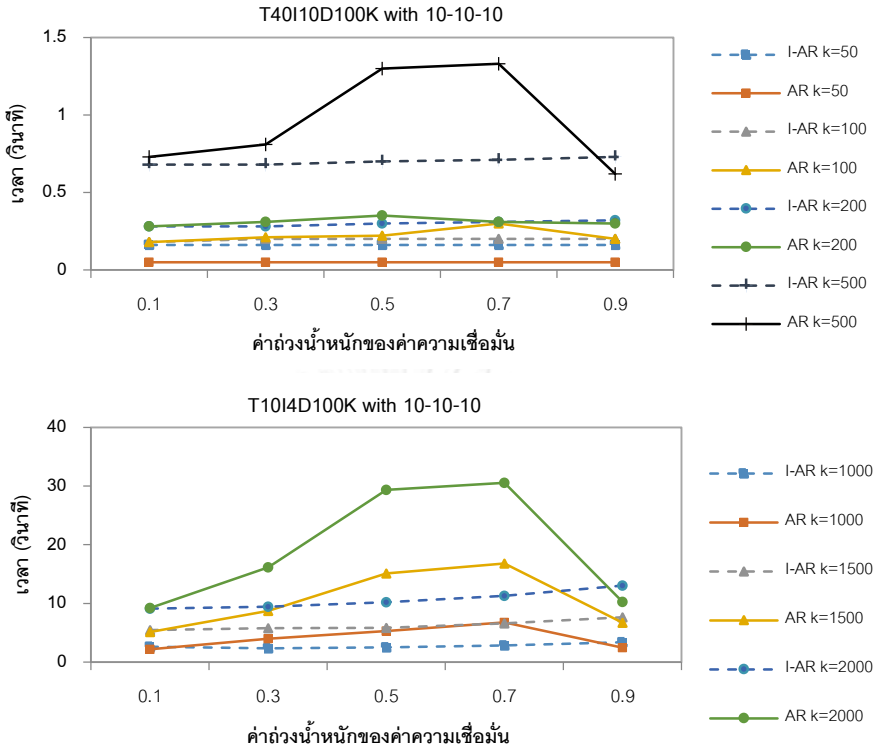
เมื่อทำการทดสอบกับชุดข้อมูล T40I10D100K (DB7 ถึง DB9) ดังภาพที่ 4-10 ถึง 4-12 โดยกำหนดค่าถ่วงน้ำหนักค่าความเชื่อมั่นแต่ละค่าและกำหนดจำนวนผลลัพธ์คงที่ พบว่าการสร้างกฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจที่สุดจะใช้เวลามากขึ้น เมื่อค่าถ่วงน้ำหนักความน่าสนใจของผลลัพธ์มีค่าเพิ่มขึ้น แต่เมื่อเทียบกับการสร้างกฎความสัมพันธ์แบบหลายลำดับชั้นด้วยค่าขีดแบ่งสนับสนุนและค่าขีดแบ่งความเชื่อมั่น จะสังเกตได้ว่า เมื่อเทียบโดยกำหนดค่าถ่วงน้ำหนักที่ 0.1 การสร้างกฎความสัมพันธ์ที่น่าสนใจสุดจะใช้เวลาใกล้เคียงกันกับการสร้างกฎความสัมพันธ์ด้วยกำหนดค่าขีดแบ่ง แต่เมื่อเทียบโดยกำหนดค่าถ่วงน้ำหนักที่ 0.3 ถึง 0.7 การสร้างกฎความสัมพันธ์ที่น่าสนใจสุดจะใช้เวลาเร็วกว่าการสร้างกฎความสัมพันธ์ด้วยค่าขีดแบ่ง เนื่องจากจำนวนผลลัพธ์ที่ได้จากการค้นหาด้วยค่าขีดแบ่งมีจำนวนมากเกินกว่าจำนวนที่ต้องการ แต่เมื่อเทียบโดยกำหนดค่าถ่วงน้ำหนักที่ 0.9 การสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุดจะใช้เวลานานกว่า การสร้างกฎความสัมพันธ์ด้วยค่าขีดแบ่งเนื่องจากจำนวนผลลัพธ์ที่ได้จากการสร้างกฎความสัมพันธ์ด้วยค่าขีดแบ่งให้จำนวนผลลัพธ์ที่ใกล้เคียงกันกับจำนวนที่ต้องการ ซึ่งค่าขีดแบ่งที่ได้นั้นเป็นค่าขีดแบ่งที่เหมาะสมสำหรับการสร้างกฎความสัมพันธ์แบบหลายลำดับชั้นที่มีการกำหนดค่าขีดแบ่ง



ภาพที่ 4-10 การทดสอบการสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุดกับชุดข้อมูล DB7



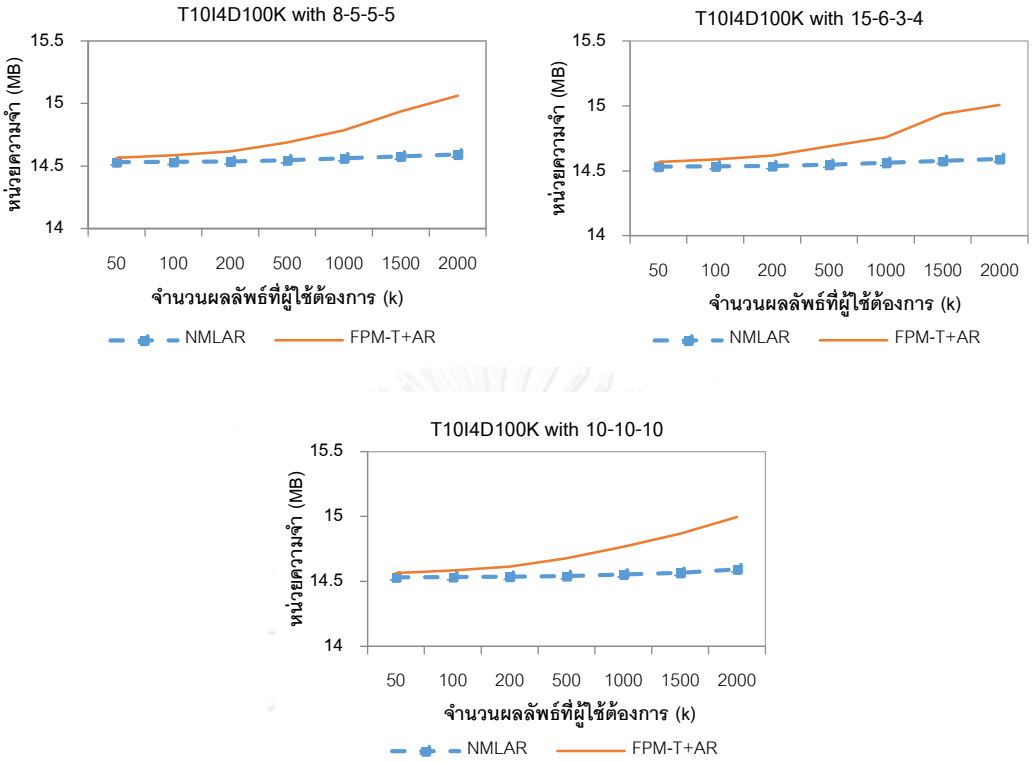
ภาพที่ 4-11 การทดสอบการสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุดกับชุดข้อมูล DB8



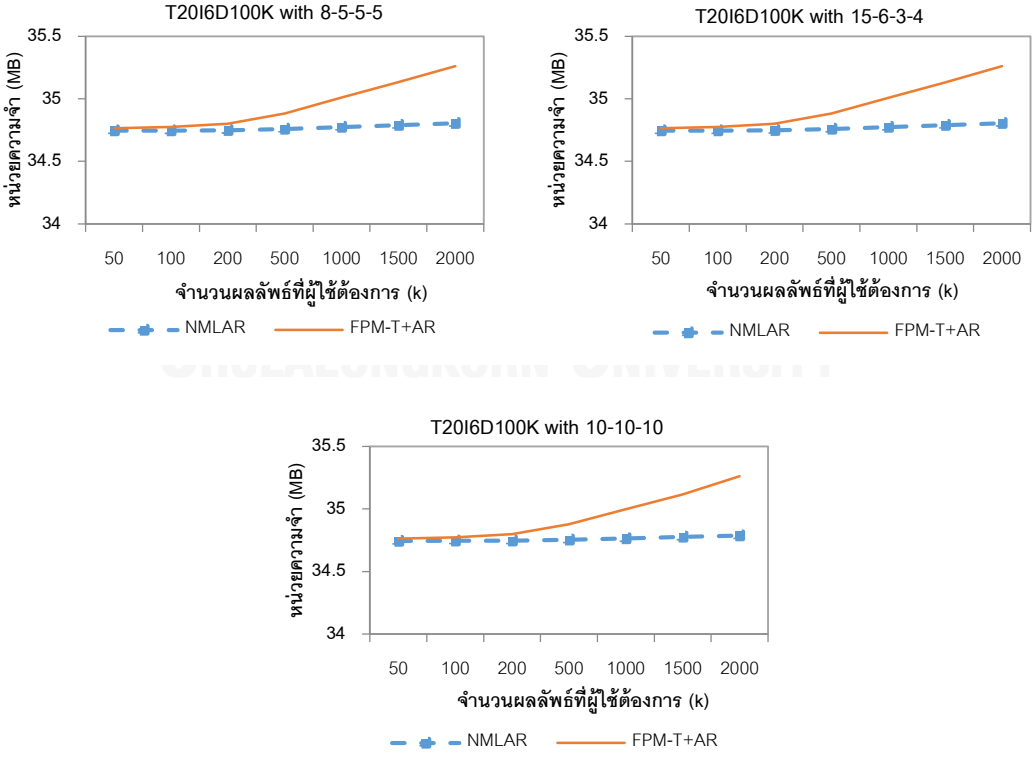
ภาพที่ 4-12 การทดสอบการสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุดกับชุดข้อมูล DB9

4.2 การทดสอบวัดประสิทธิภาพเชิงหน่วยความจำ

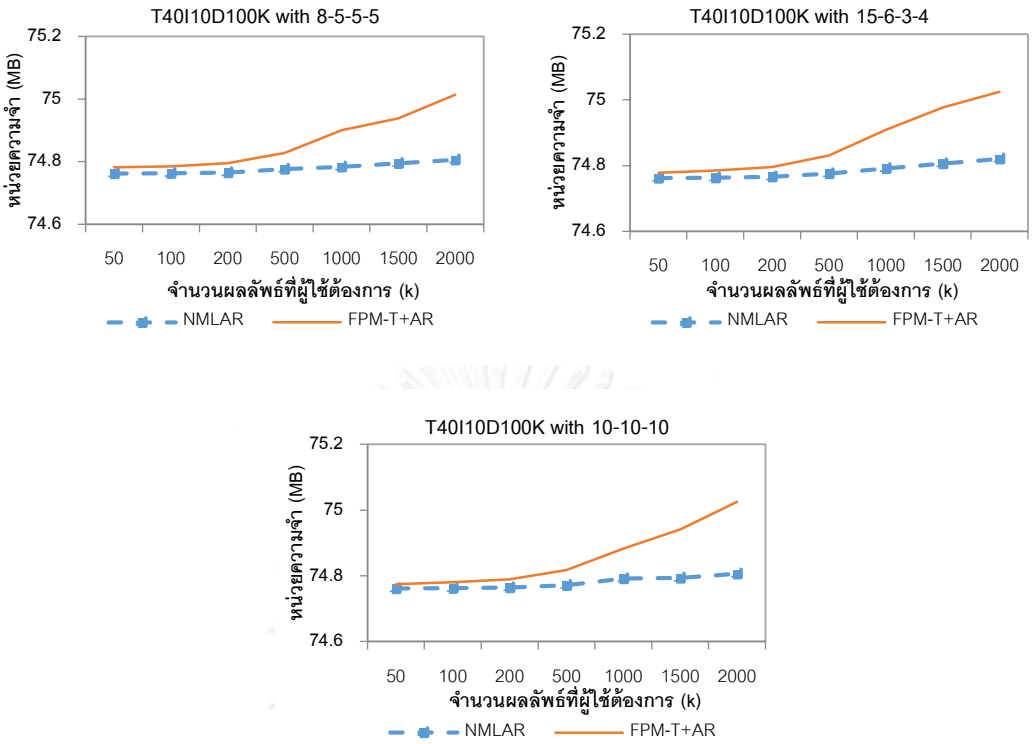
การทดสอบวัดประสิทธิภาพเชิงหน่วยความจำของงานวิจัยนี้ วัดจากหน่วยความจำที่ใช้สำหรับต้นไม้แสดงรูปแบบการเกิดขึ้นของฐานข้อมูล และหน่วยความจำที่ใช้สำหรับการเก็บผลลัพธ์ที่น่าสนใจที่สุดตามที่ผู้ใช้ต้องการจากขั้นตอนวิธีเอ็นเอ็มแอลเอฟพีและเอฟทีเอ็ม-ที โดยการทดสอบนี้ได้ทดสอบกับทุกชุดข้อมูล จากภาพที่ 4-13 ถึง 4-15 พบว่า ขั้นตอนเอ็นเอ็มแอลเอฟพีใช้เนื้อที่ในหน่วยความจำที่น้อยกว่าขั้นตอนวิธีเอฟทีเอ็ม-ที เนื่องจากขั้นตอนวิธีเอ็นเอ็มแอลเอฟพีจะเก็บเฉพาะรูปแบบเซตที่น่าสนใจที่สุดตามจำนวนที่ต้องการ แต่ขั้นตอนวิธีเอฟทีเอ็ม-ทีทำการเก็บทุกๆ รูปแบบเซตที่ปรากฏขึ้นบ่อย นอกจากนั้นเมื่อกำหนดจำนวนผลลัพธ์ตามที่ผู้ใช้ต้องการมากขึ้น การใช้เนื้อที่หน่วยความจำก็เพิ่มขึ้นตามจำนวนผลลัพธ์ด้วย ทั้งนี้สรุปได้ว่าการใช้เนื้อที่สำหรับการค้นหาผลลัพธ์ขึ้นกับจำนวนผลลัพธ์ตามที่ผู้ใช้ต้องการ



ภาพที่ 4-13 การทดสอบหาผลลัพธ์ที่น่าสนใจที่สุดเชิงหน่วยความจำกับชุดข้อมูล T10I4D100K



ภาพที่ 4-14 การทดสอบหาผลลัพธ์ที่น่าสนใจที่สุดเชิงหน่วยความจำกับชุดข้อมูล T20I6D100K



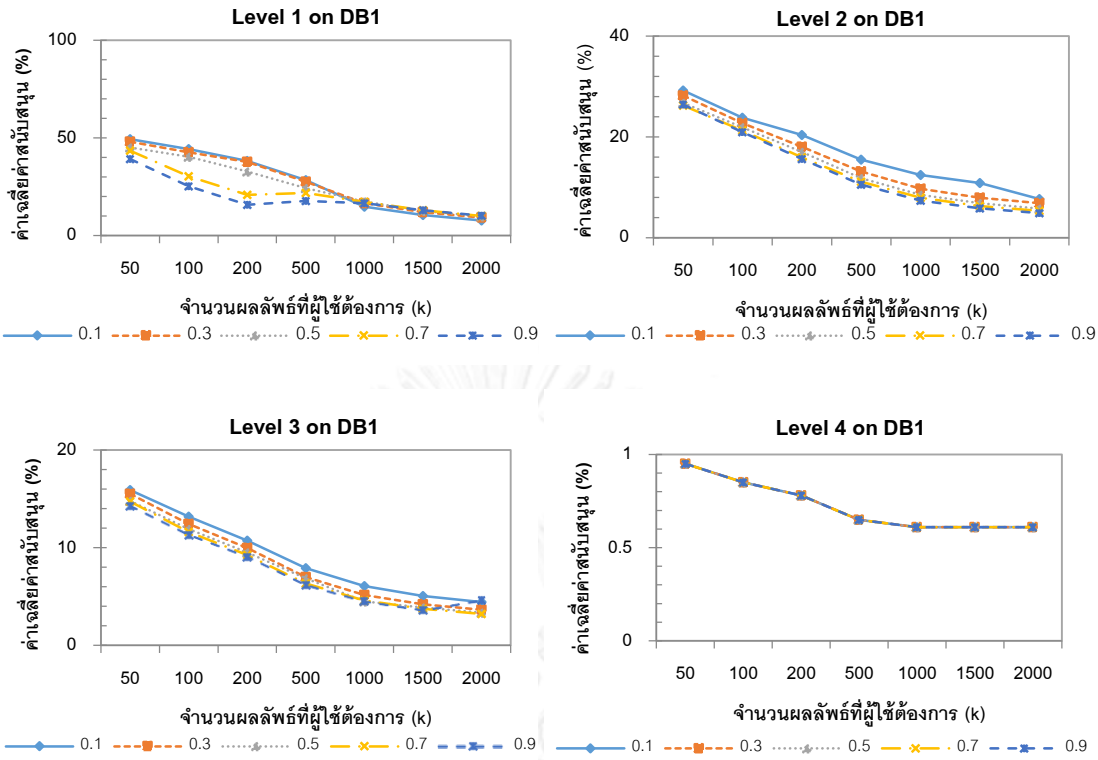
ภาพที่ 4-15 การทดสอบหาผลลัพธ์ที่น่าสนใจที่สุดเชิงหน่วยความจำกับชุดข้อมูล T40I10D100K

4.3 การวิเคราะห์ผลลัพธ์จากการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจที่สุด

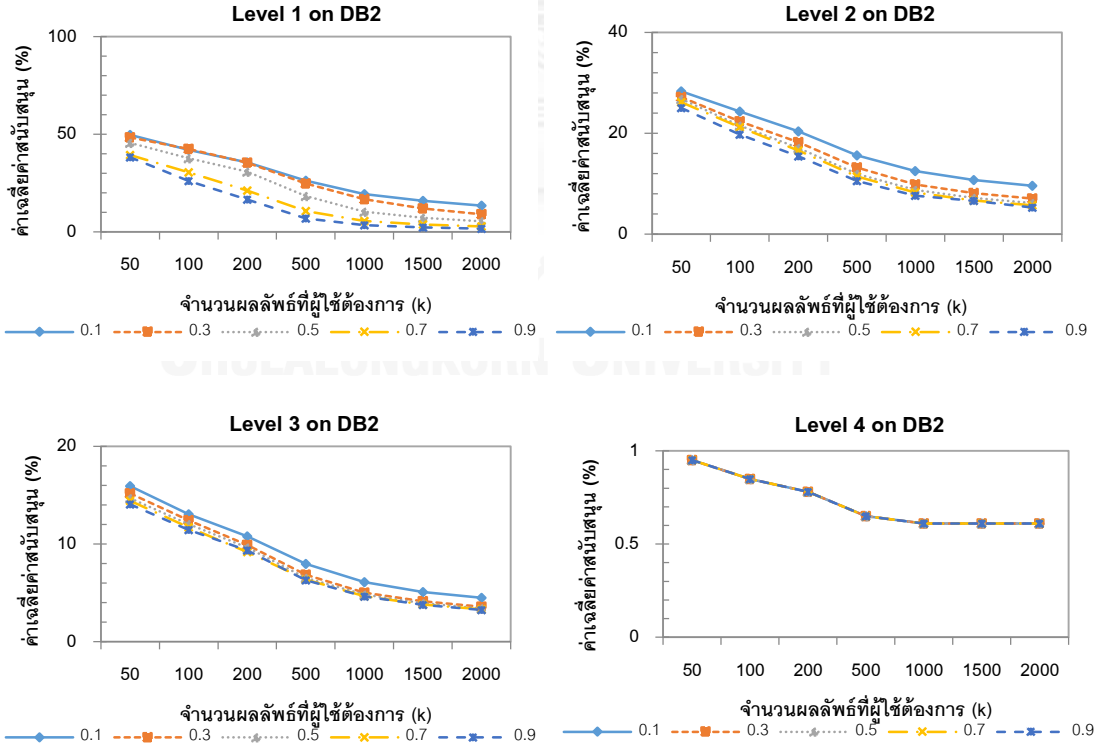
ในส่วนนี้จะกล่าวถึงผลการวิเคราะห์ผลลัพธ์จากการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจที่สุดตามจำนวนที่ผู้ใช้ต้องการ ซึ่งในงานวิจัยนี้ได้แบ่งการวิเคราะห์ผลลัพธ์ออกเป็น 6 ส่วนย่อย ดังนี้

4.3.1 ผลวิเคราะห์ค่าเฉลี่ยค่าสนับสนุนของกฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจสุดโดยจำแนกแต่ละลำดับชั้น

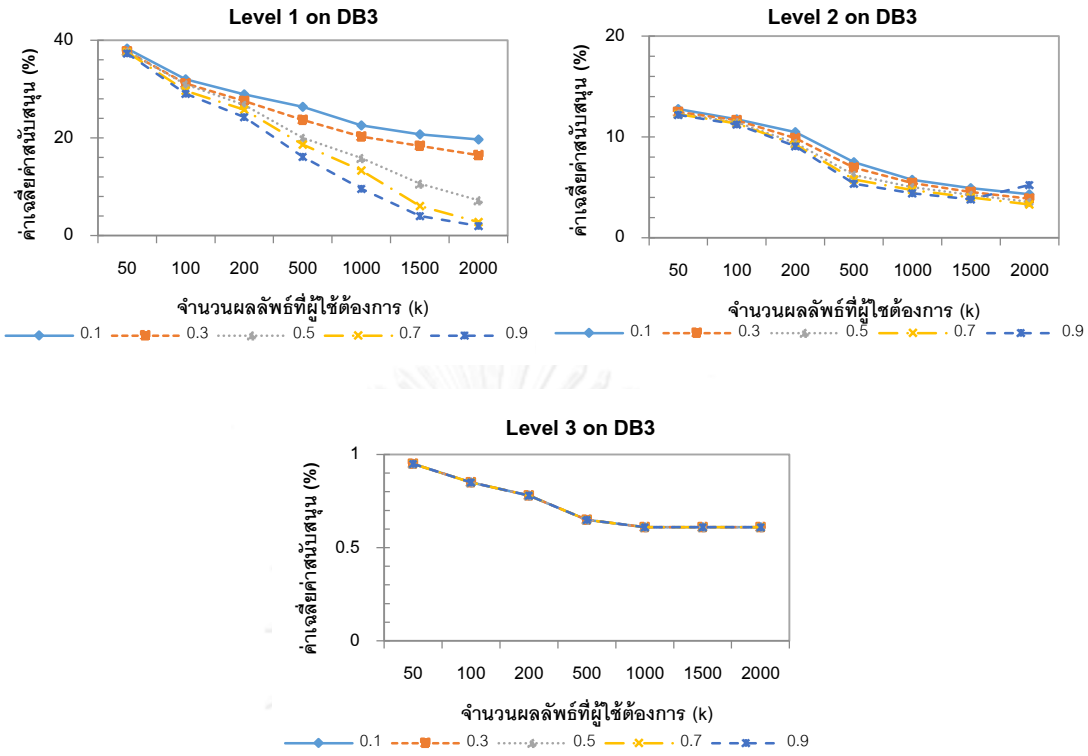
จากภาพ 4-16 ถึง 4-24 ที่วิเคราะห์หาค่าเฉลี่ยสนับสนุนของผลลัพธ์แต่ละลำดับชั้นโดยทำการทดสอบกับทุกชุดข้อมูล พบว่า ในแต่ลำดับชั้นหากผู้ใช้ต้องการจำนวนผลลัพธ์ที่เพิ่มขึ้น จะมีผลทำให้ค่าเฉลี่ยค่าสนับสนุนในแต่ละลำดับชั้นนั้นมีค่าน้อยลง เนื่องจากค่าสนับสนุนของกฎความสัมพันธ์ที่ได้มีค่าน้อยลงตามจำนวนผลลัพธ์ที่ต้องการมากขึ้น นอกจากนั้นหากให้ค่าถ่วงน้ำหนักความน่าสนใจของค่าความเชื่อมั่นมีค่ามากขึ้น ก็ทำให้ค่าเฉลี่ยค่าสนับสนุนมีค่าน้อยลง เพราะว่าการคำนวณหาค่าความน่าสนใจของกฎความสัมพันธ์ ได้ให้ค่าถ่วงน้ำหนักความสนใจไปทางค่าความเชื่อมั่นมากกว่าที่ให้ค่าถ่วงน้ำหนักไปทางค่าสนับสนุนของผลลัพธ์



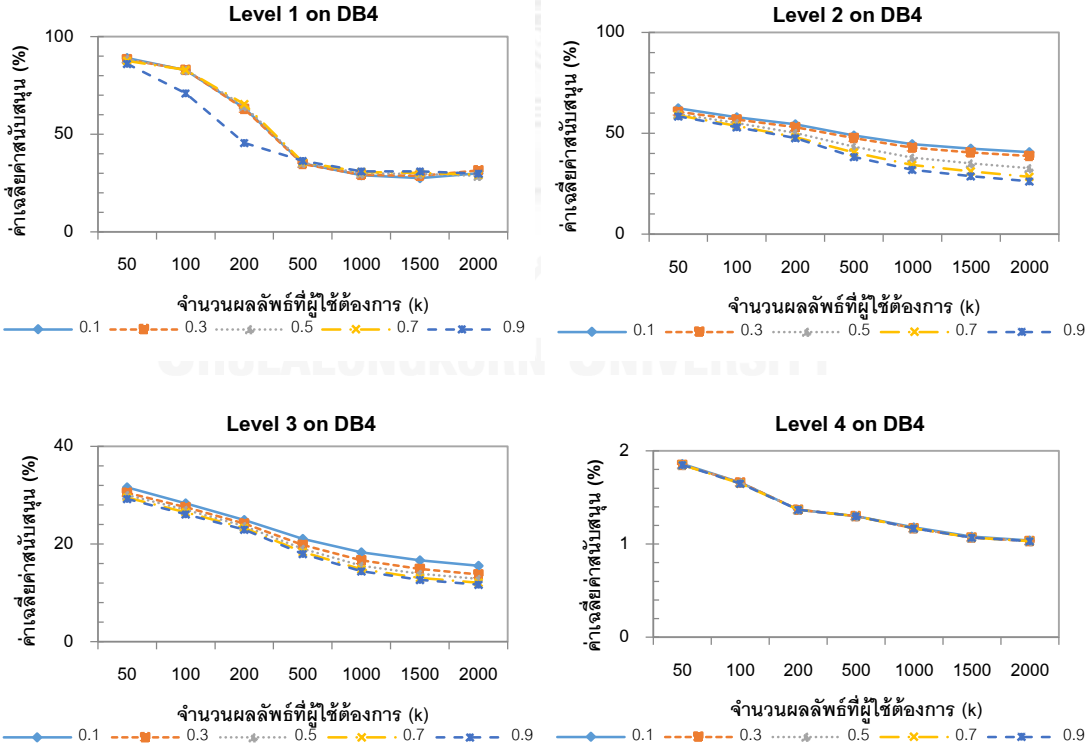
ภาพที่ 4-16 ค่าเฉลี่ยค่าสนับสนุนของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB1



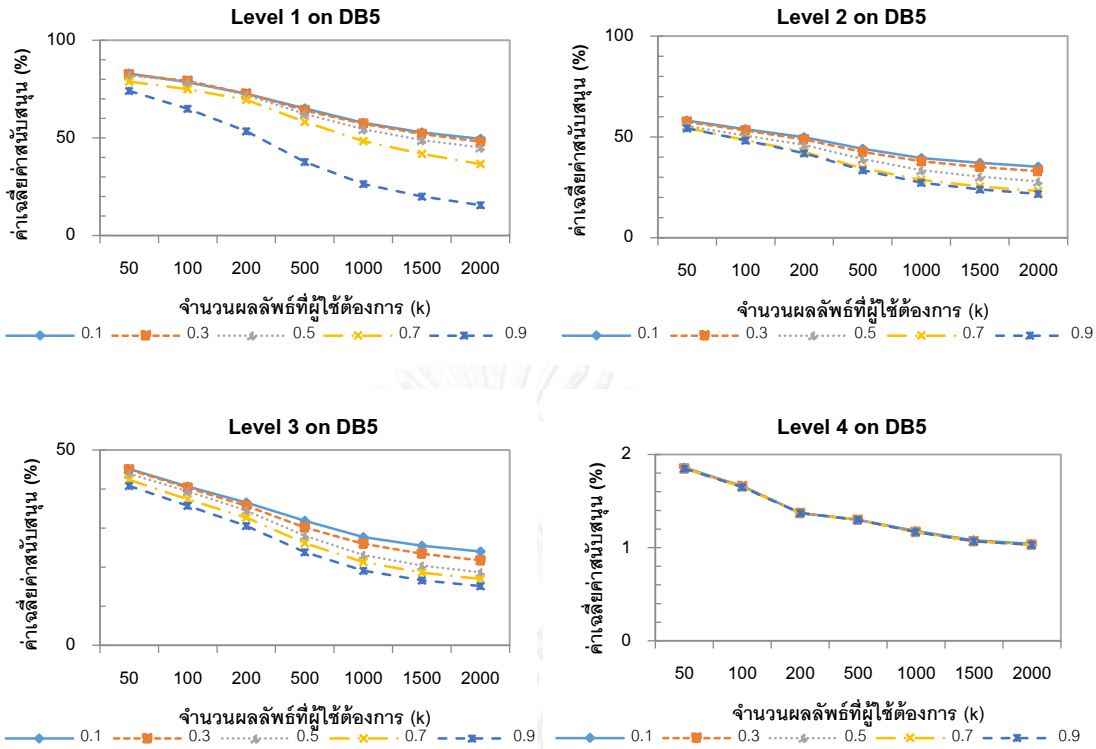
ภาพที่ 4-17 ค่าเฉลี่ยค่าสนับสนุนของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB2



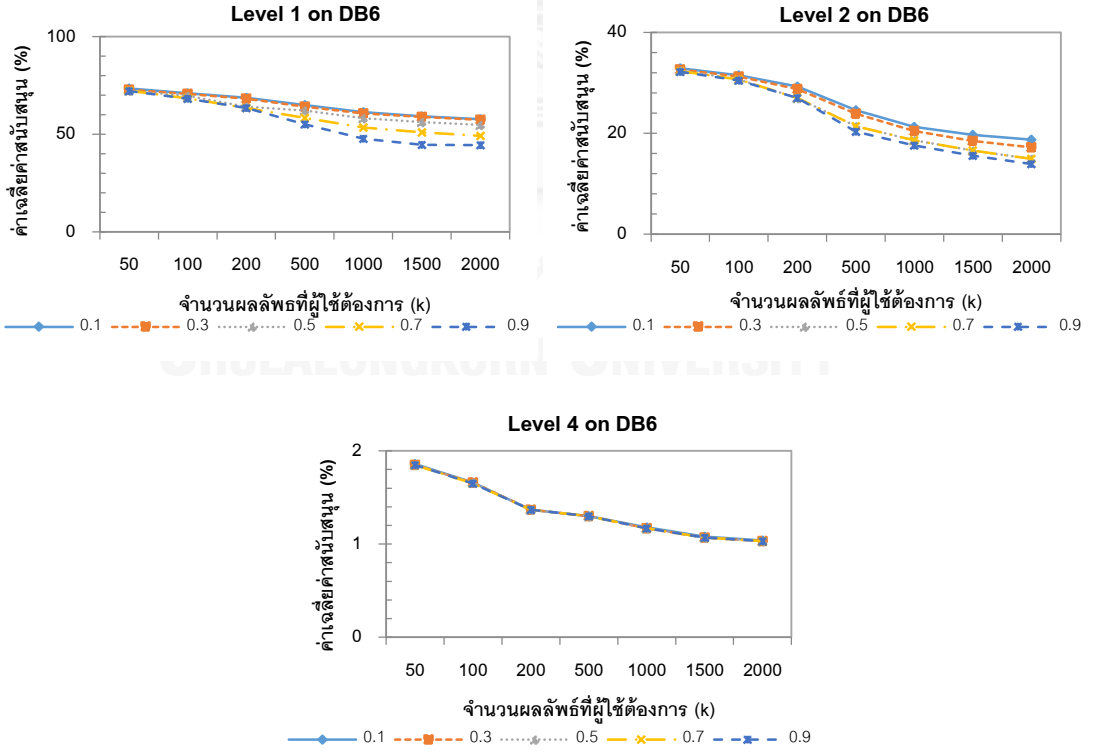
ภาพที่ 4-18 ค่าเฉลี่ยค่าสนับสนุนของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB3



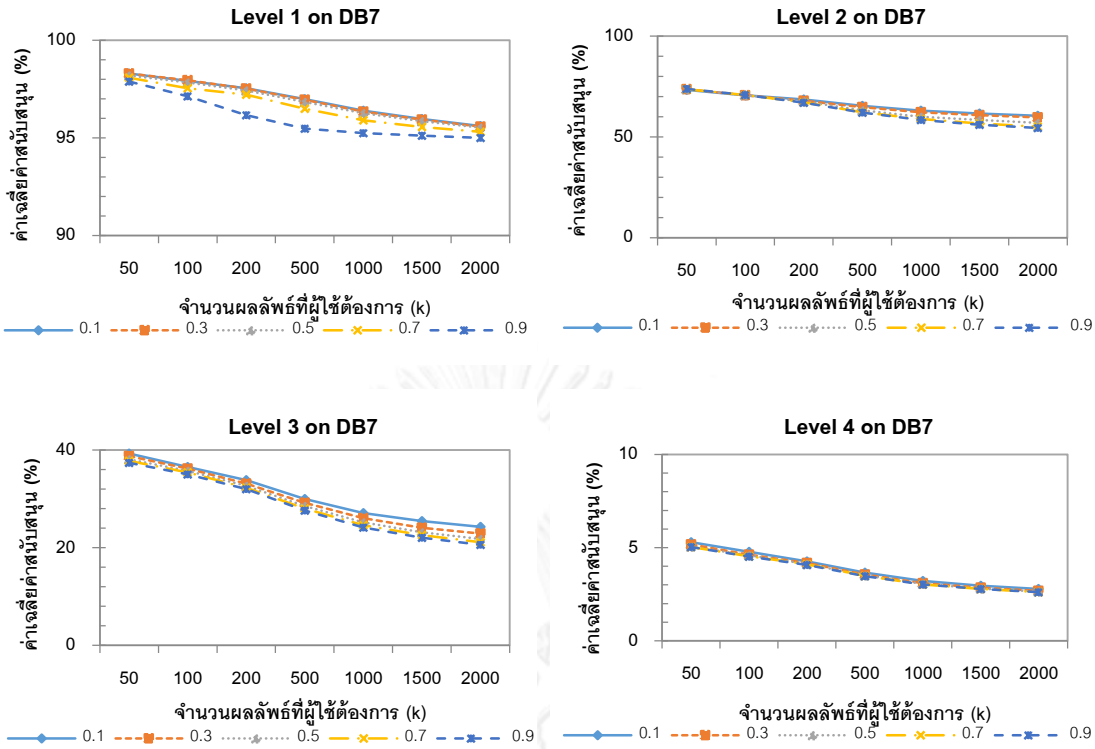
ภาพที่ 4-19 ค่าเฉลี่ยค่าสนับสนุนของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB4



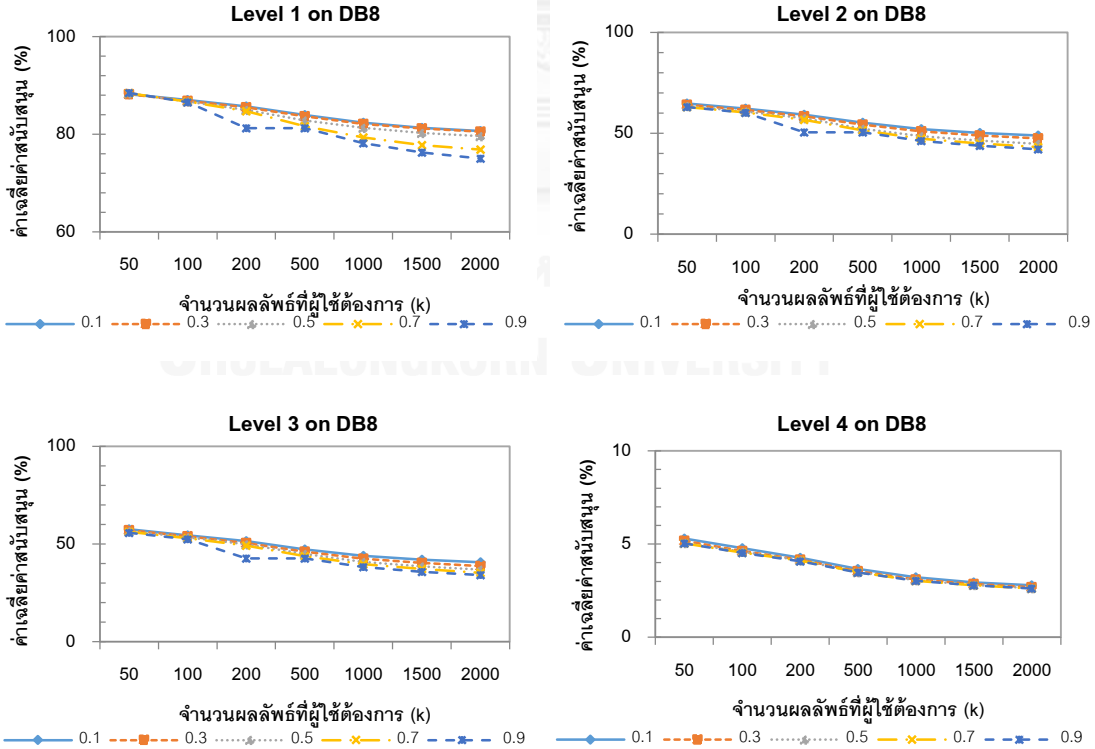
ภาพที่ 4-20 ค่าเฉลี่ยค่าสนับสนุนของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB5



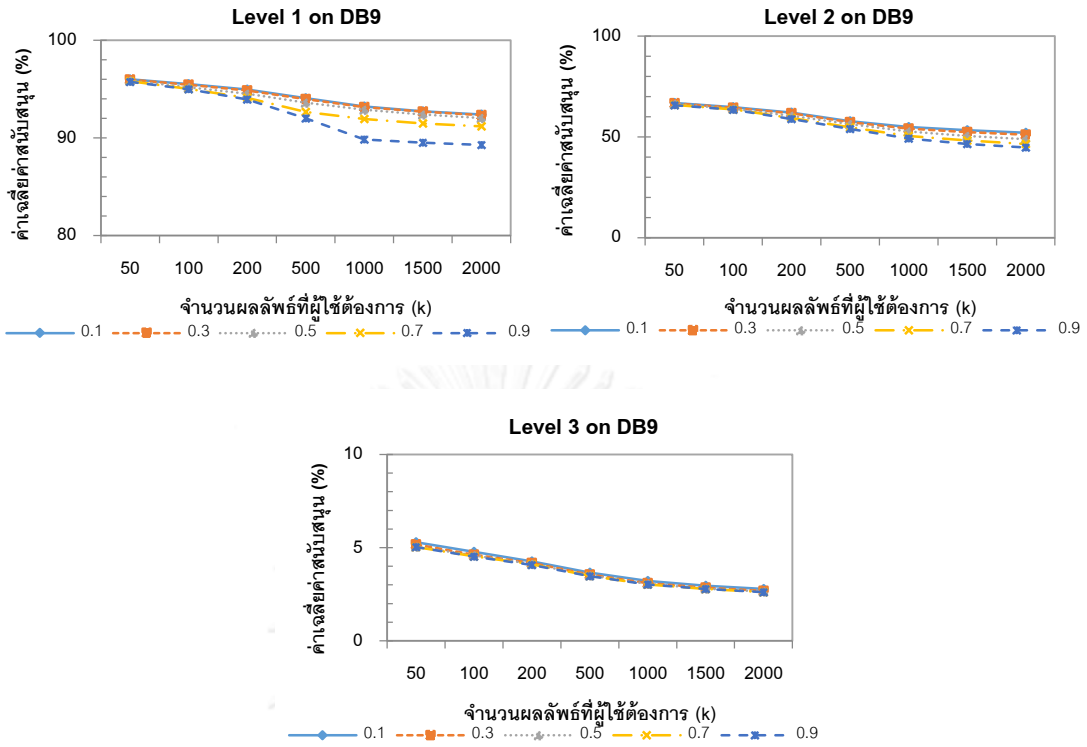
ภาพที่ 4-21 ค่าเฉลี่ยค่าสนับสนุนของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB6



ภาพที่ 4-22 ค่าเฉลี่ยค่าสนับสนุนของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB7



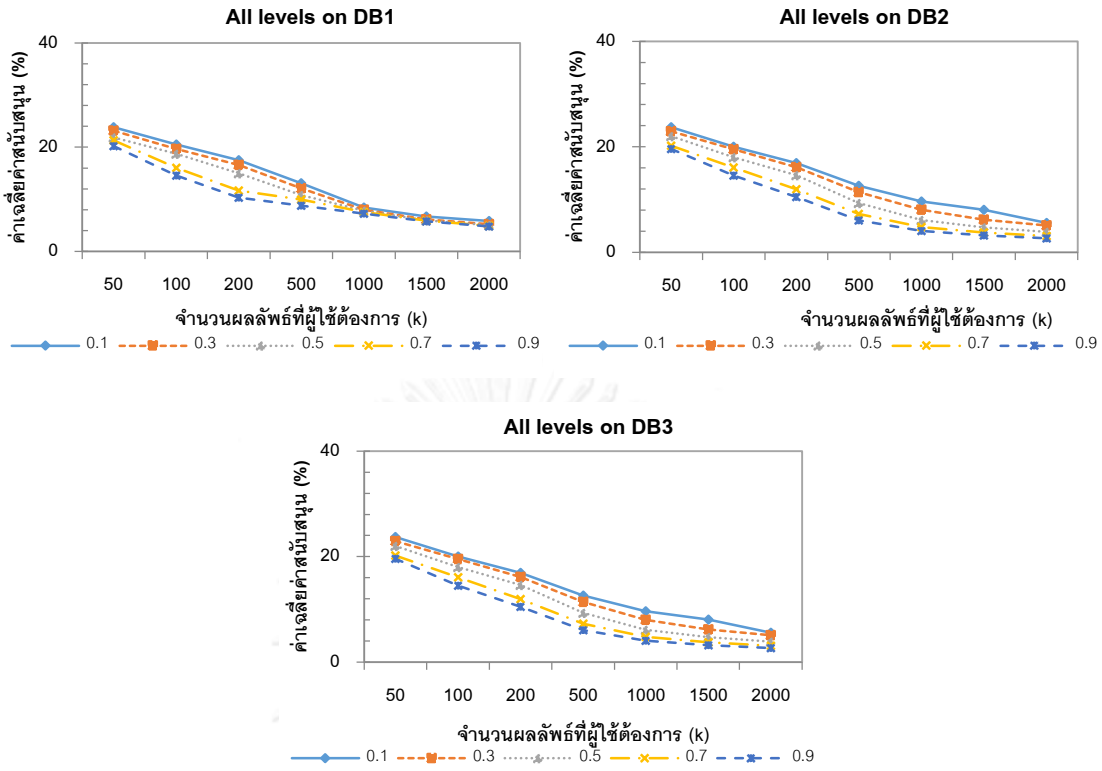
ภาพที่ 4-23 ค่าเฉลี่ยค่าสนับสนุนของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB8



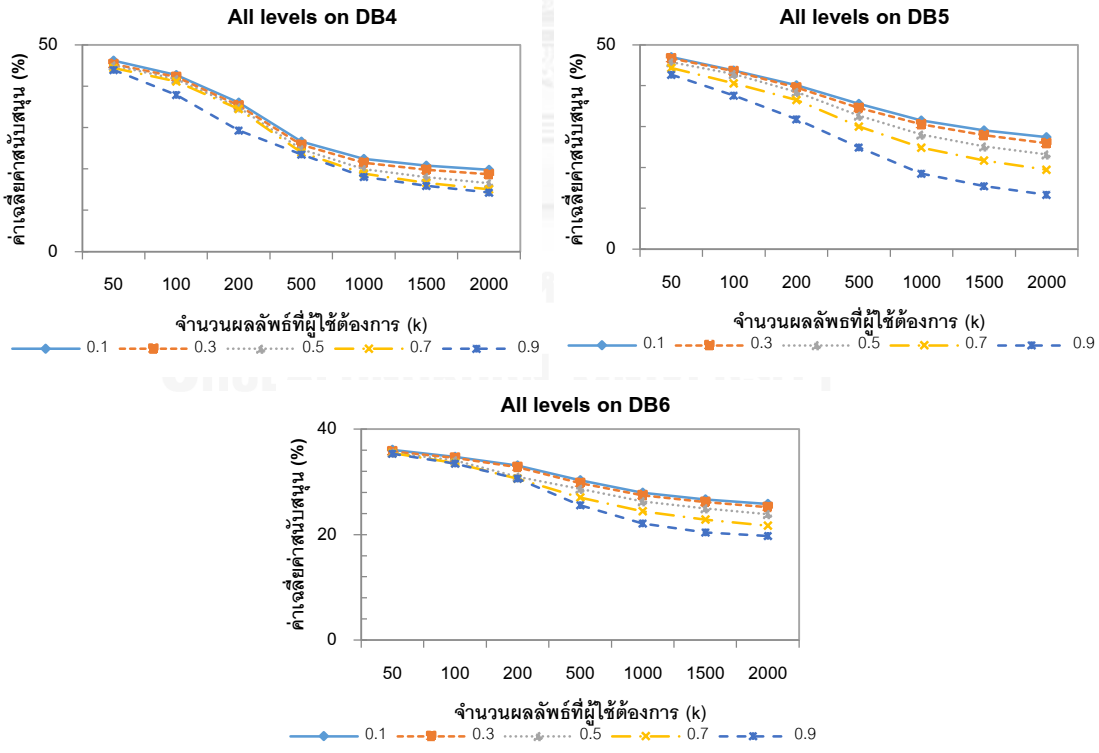
ภาพที่ 4-24 ค่าเฉลี่ยค่าสนับสนุนของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB1

4.3.2 ผลการวิเคราะห์หาค่าเฉลี่ยค่าสนับสนุนของกฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจสุดทุกๆ ลำดับชั้น

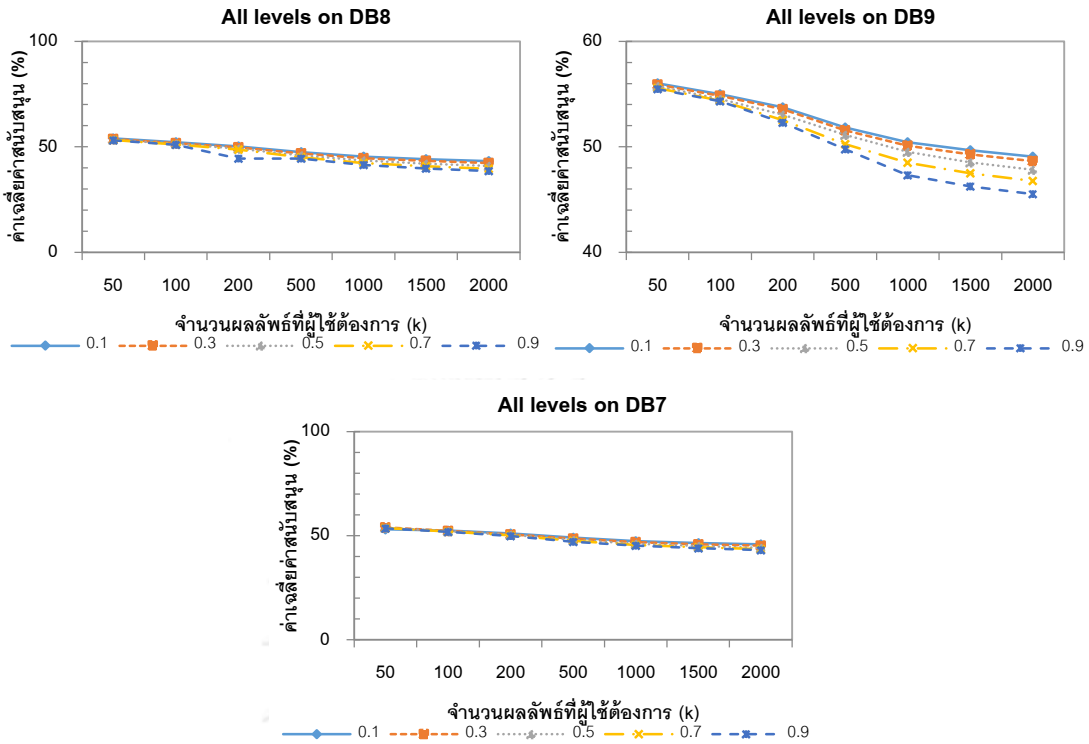
จากภาพที่ 4-25 ถึง 4-27 ที่วิเคราะห์หาค่าเฉลี่ยสนับสนุนของผลลัพธ์ทุกลำดับชั้นโดยทำการทดสอบกับทุกชุดข้อมูล พบว่า หากกำหนดจำนวนผลลัพธ์ที่ต้องการมากขึ้น จะทำให้ค่าเฉลี่ยค่าสนับสนุนของผลลัพธ์ที่ค้นหาในทุกๆ ลำดับชั้นได้มีค่าน้อยลง ซึ่งมีผลมาจากค่าเฉลี่ยค่าสนับสนุนในแต่ละลำดับชั้นที่ได้มีค่าน้อยลง มากกว่านั้นถ้ากำหนดจำนวนผลลัพธ์ที่คงที่ และค่าถ่วงน้ำหนักความน่าสนใจของค่าความเชื่อมั่นมีค่าเพิ่มขึ้น จะสังเกตว่าค่าเฉลี่ยค่าสนับสนุนของผลลัพธ์ที่น่าสนใจสุดในทุกลำดับชั้นมีค่าน้อยลงเช่นกัน เพราะหากให้ค่าถ่วงน้ำหนักความสนใจของผลลัพธ์มาก มีผลทำให้ค่าถ่วงน้ำหนักของค่าสนับสนุนของกฎความสัมพันธ์มีค่าน้อยลง ซึ่งมีผลต่อการคำนวณหาค่าความน่าสนใจของกฎความสัมพันธ์ที่เน้นผลลัพธ์ที่น่าสนใจที่สุดไปทางค่าความเชื่อมั่น



ภาพที่ 4-25 ค่าเฉลี่ยค่าสนับสนุนของผลลัพธ์ทุกลำดับชั้นกับชุดข้อมูล T10I4D100K (DB1 ถึง DB3)



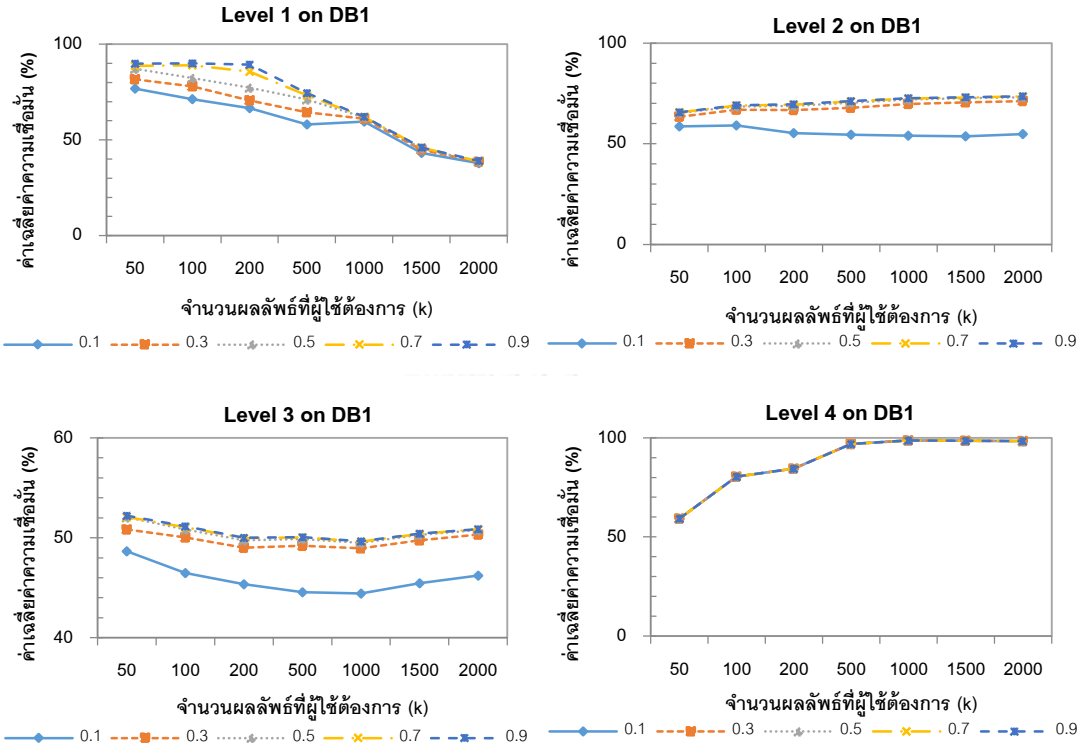
ภาพที่ 4-26 ค่าเฉลี่ยค่าสนับสนุนของผลลัพธ์ทุกลำดับชั้นกับชุดข้อมูล T20I6D100K (DB4 ถึง DB6)



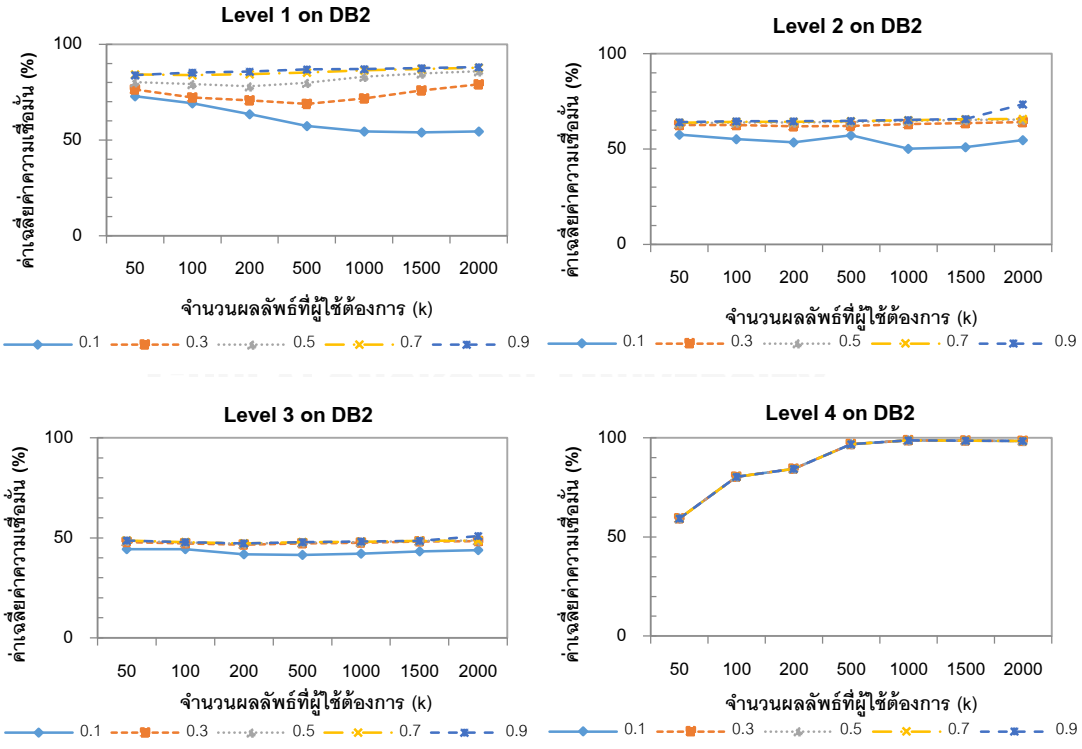
ภาพที่ 4-27 ค่าเฉลี่ยค่าสนับสนุนของผลลัพธ์ทุกลำดับชั้นกับชุดข้อมูล T40I10D100K (DB7 ถึง DB9)

4.3.3 ผลวิเคราะห์ค่าเฉลี่ยค่าความเชื่อมั่นของกฎความสัมพันธ์ที่น่าสนใจสุดโดยจำแนกแต่ละลำดับชั้น

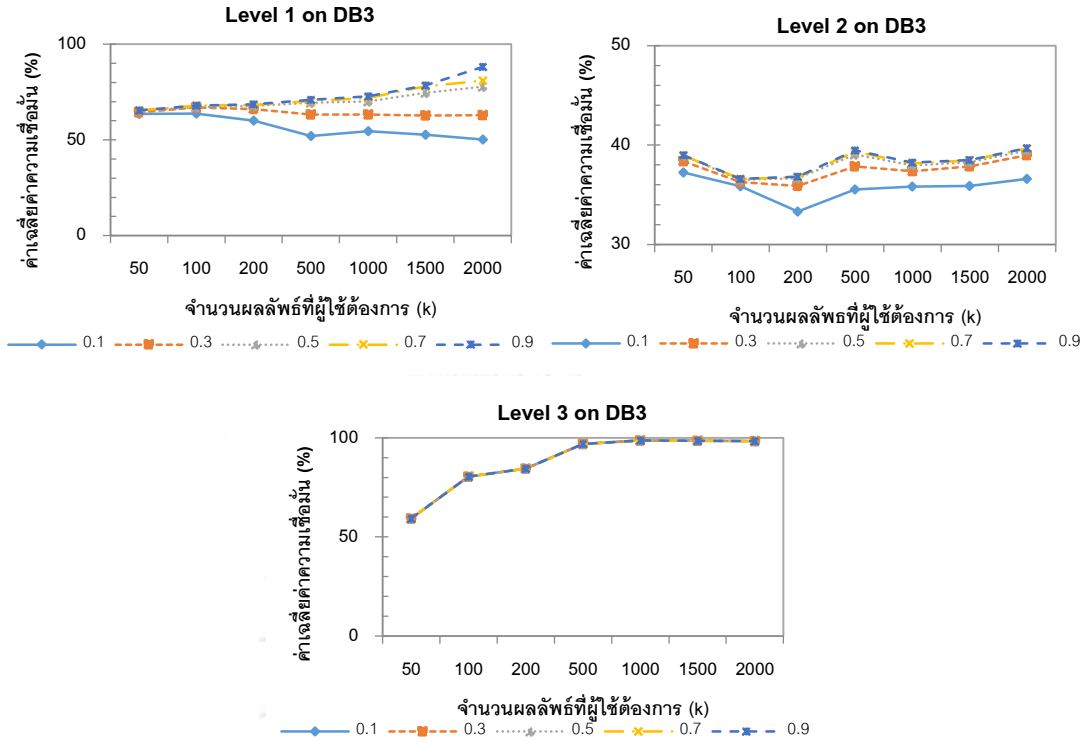
จากภาพที่ 4-28 ถึง 4-36 ที่วิเคราะห์หาค่าเฉลี่ยความเชื่อมั่นของผลลัพธ์ที่น่าสนใจสุดแต่ละลำดับชั้นกับทุกชุดข้อมูลที่ทำการทดสอบ พบว่า เมื่อกำหนดจำนวนผลลัพธ์ในแต่ละลำดับชั้นคงที่ และถ้ากำหนดค่าถ่วงน้ำหนักความน่าสนใจของค่าความเชื่อมั่นให้มีค่าเพิ่มขึ้น จะทำให้ค่าเฉลี่ยค่าความเชื่อมั่นของผลลัพธ์ที่น่าสนใจสุดมีค่ามากขึ้นด้วย เพราะว่าการคำนวณหาค่าความน่าสนใจของกฎความสัมพันธ์นั้น คำนวณค่าถ่วงน้ำหนักเน้นไปทางค่าความเชื่อมั่น จึงทำให้ผลลัพธ์ที่น่าสนใจที่สุดนั้น มีค่าเฉลี่ยค่าความเชื่อมั่นที่สูงด้วย



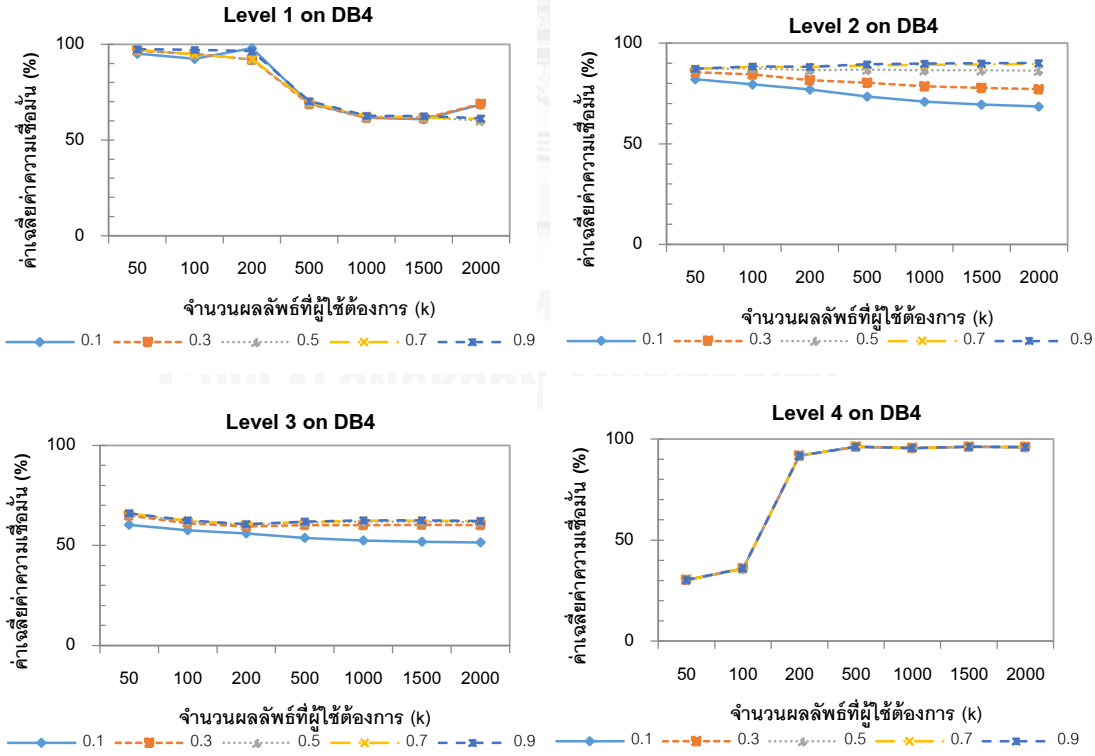
ภาพที่ 4-28 ค่าเฉลี่ยค่าความเชื่อมั่นของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB1



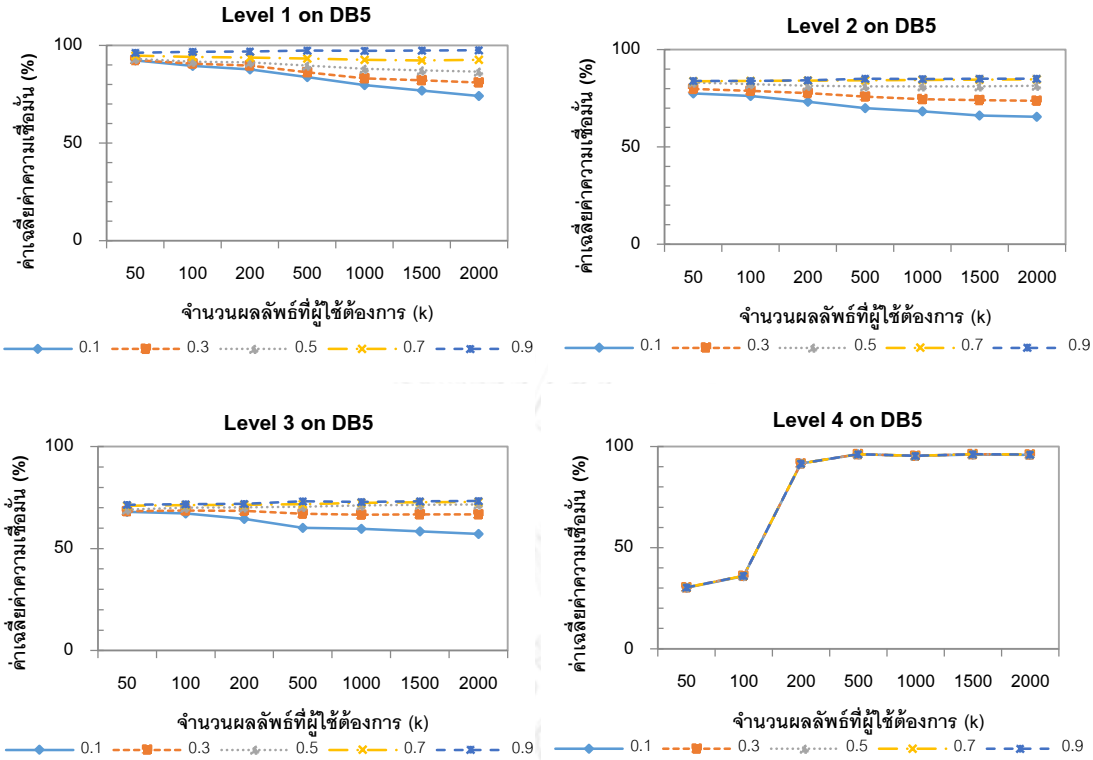
ภาพที่ 4-29 ค่าเฉลี่ยค่าความเชื่อมั่นของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB2



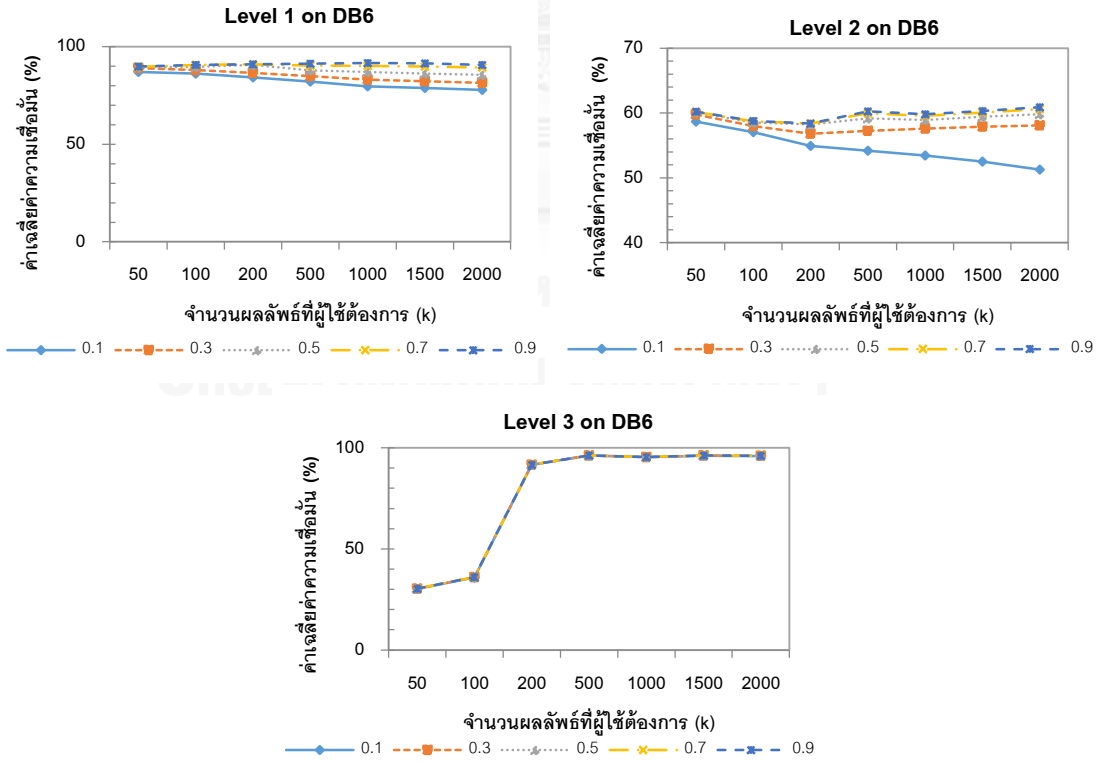
ภาพที่ 4-30 ค่าเฉลี่ยค่าความเชื่อมั่นของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB3



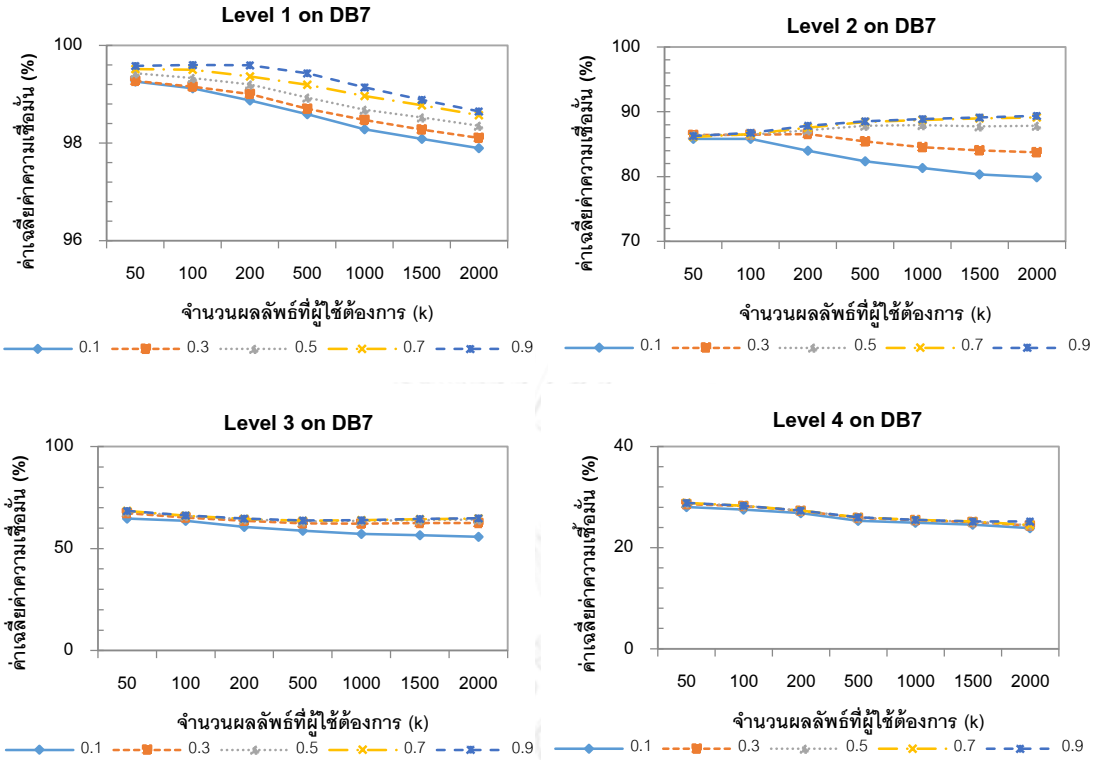
ภาพที่ 4-31 ค่าเฉลี่ยค่าความเชื่อมั่นของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB4



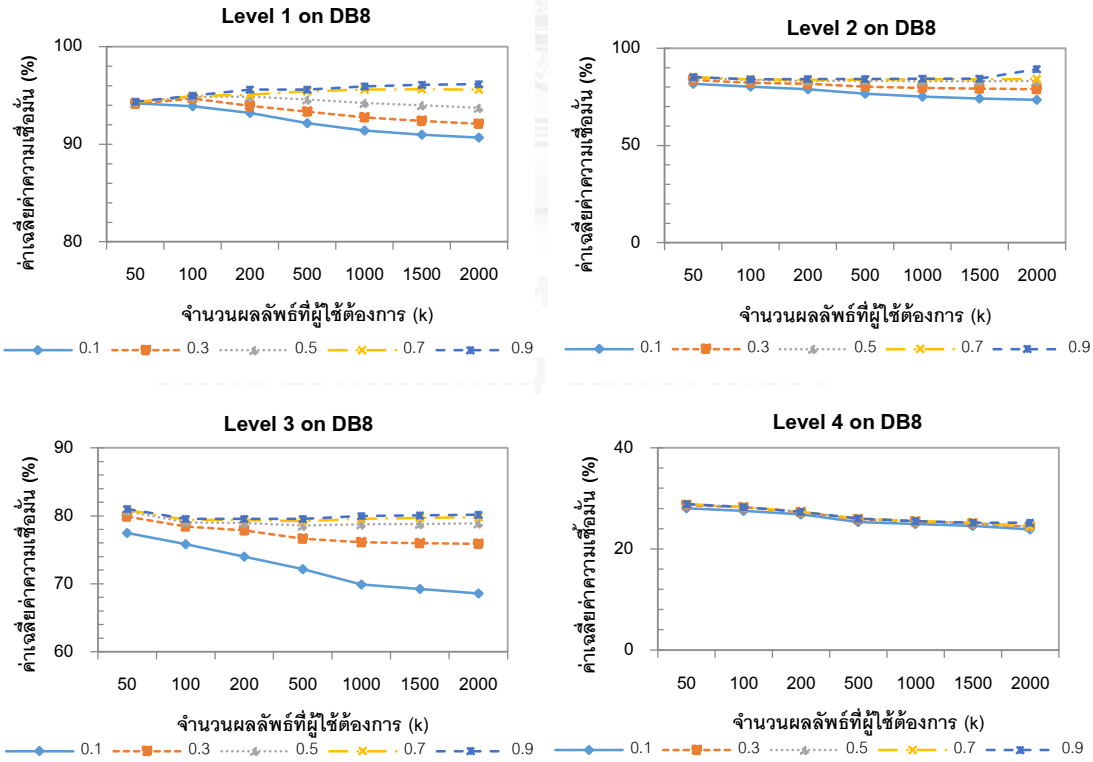
ภาพที่ 4-32 ค่าเฉลี่ยค่าความเชื่อมั่นของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB5



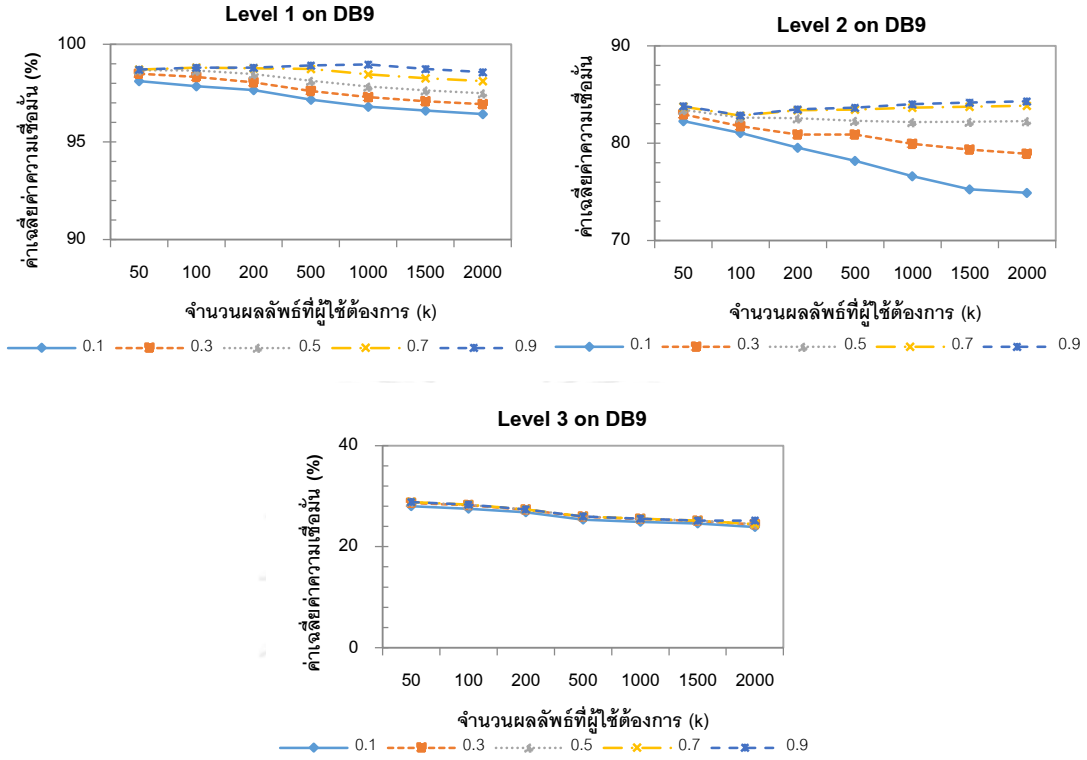
ภาพที่ 4-33 ค่าเฉลี่ยค่าความเชื่อมั่นของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB6



ภาพที่ 4-34 ค่าเฉลี่ยค่าความเชื่อมั่นของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB7



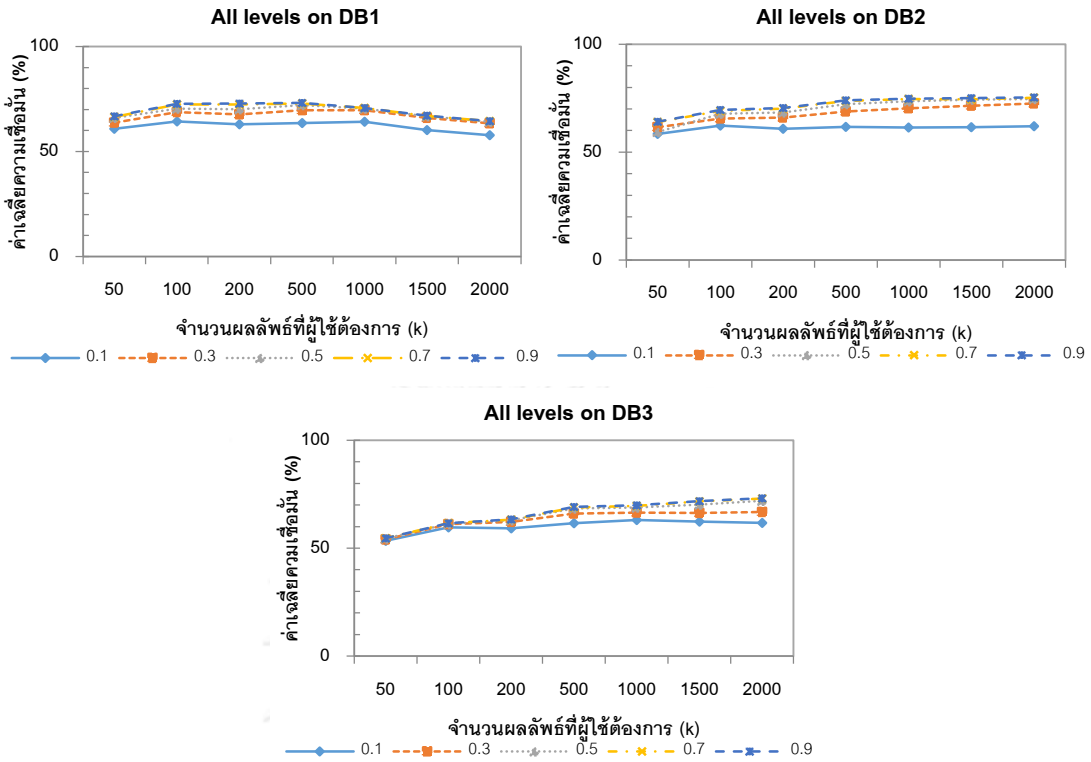
ภาพที่ 4-35 ค่าเฉลี่ยค่าความเชื่อมั่นของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB8



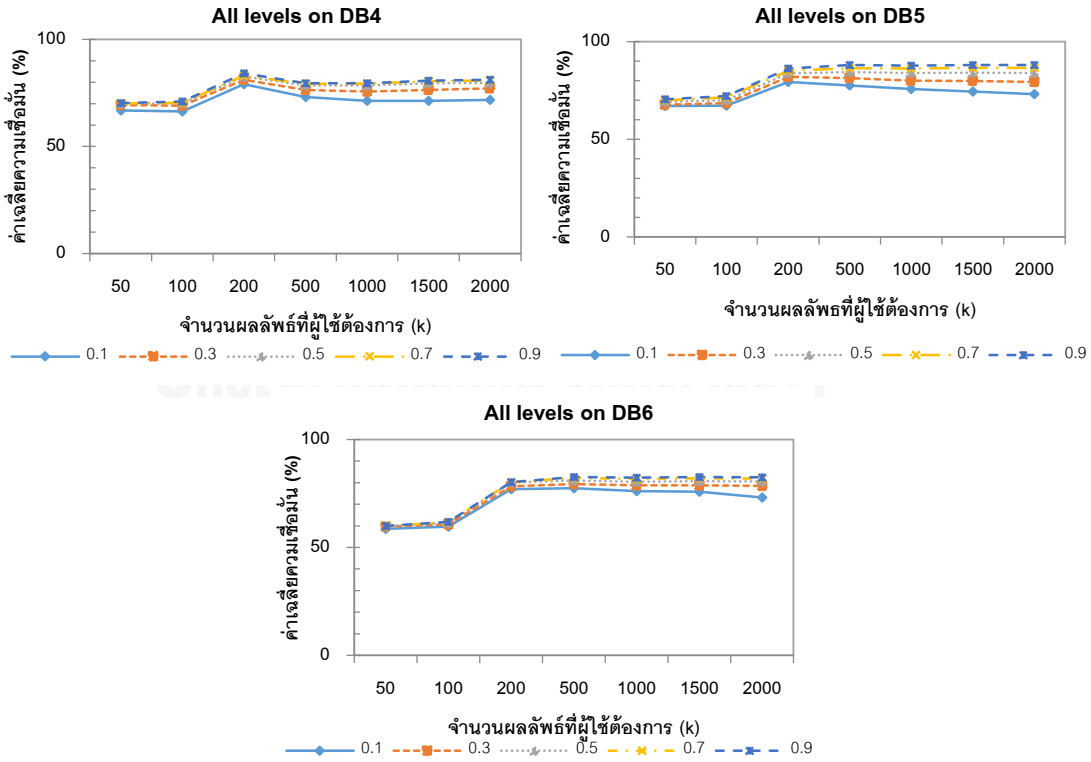
ภาพที่ 4-36 ค่าเฉลี่ยค่าความเชื่อมั่นของผลลัพธ์แต่ละลำดับชั้นกับชุดข้อมูล DB9

4.3.4 ผลการวิเคราะห์หาค่าเฉลี่ยค่าความเชื่อมั่นของกฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจสุดในทุกๆ ลำดับชั้น

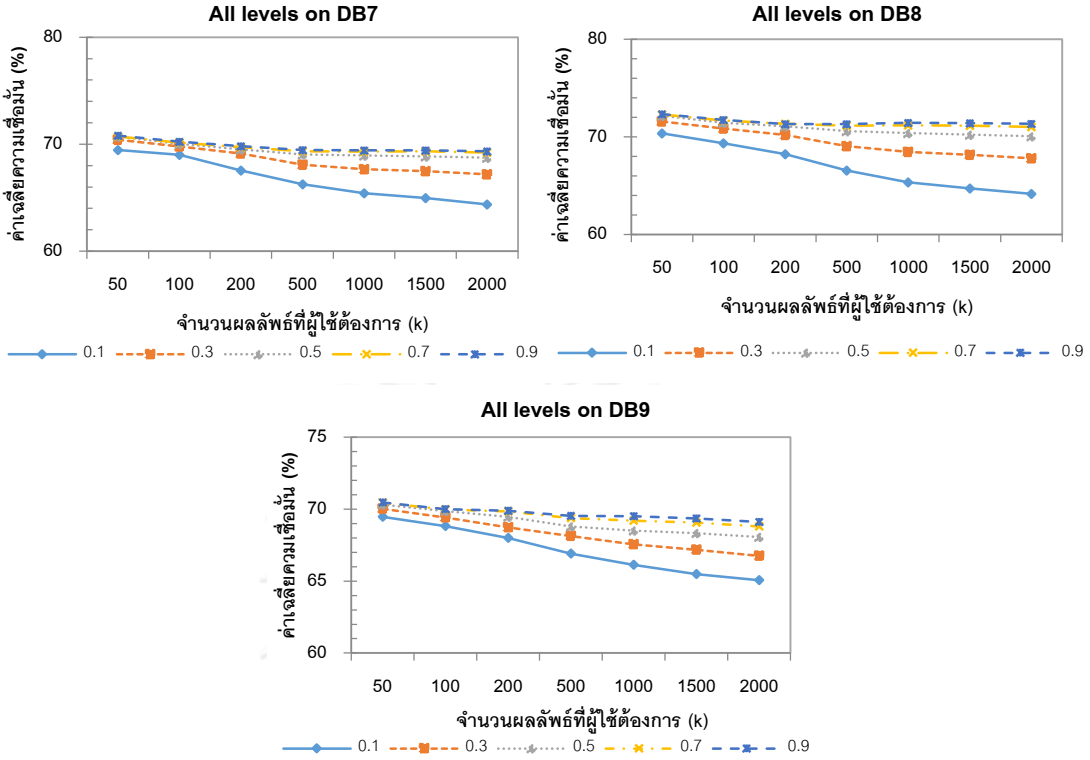
จากภาพที่ 4-37 ถึง 4-39 ที่วิเคราะห์หาค่าเฉลี่ยความเชื่อมั่นของผลลัพธ์ทุกลำดับชั้นกับทุกชุดข้อมูลที่ทำการทดสอบ พบว่า หากจำนวนผลลัพธ์มีค่ามากขึ้น ค่าเฉลี่ยค่าความเชื่อมั่นเฉลี่ยทุกลำดับชั้นจะมีค่าที่ใกล้เคียงกัน ซึ่งแสดงว่า ไม่ว่าจำนวนจะมากเพียงใด ผลลัพธ์ที่ได้จะมีค่าความเชื่อมั่นที่มีค่าใกล้เคียงกัน และหากกำหนดจำนวนผลลัพธ์คงที่ และกำหนดค่าถ่วงน้ำหนักความน่าสนใจของค่าความเชื่อมั่นมีค่าที่เพิ่มขึ้น ทำให้ค่าเฉลี่ยค่าความเชื่อมั่นของผลลัพธ์ที่น่าสนใจที่สุดในทุกๆ ลำดับชั้นนั้นมีค่ามากขึ้นด้วย เนื่องจากในแต่ละลำดับชั้นได้ค่าเฉลี่ยค่าความเชื่อมั่นที่สูงขึ้น โดยหากจำนวนผลลัพธ์เปลี่ยนไป ก็ทำให้ค่าเฉลี่ยค่าความเชื่อมั่นของผลลัพธ์นั้นเปลี่ยนแปลงไปเล็กน้อยด้วย



ภาพที่ 4-37 ค่าเฉลี่ยค่าความเชื่อมั่นของผลลัพธ์ทุกลำดับชั้นกับข้อมูล T10I4D100K (DB1 ถึง DB3)



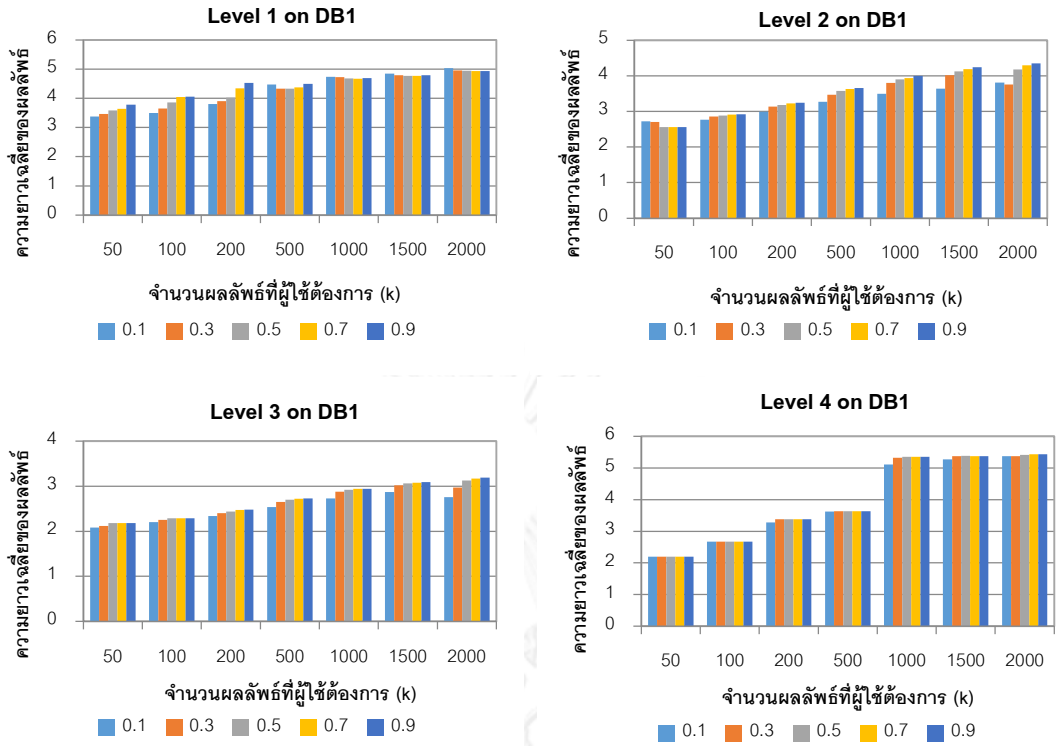
ภาพที่ 4-38 ค่าเฉลี่ยค่าความเชื่อมั่นของผลลัพธ์ทุกลำดับชั้นกับข้อมูล T20I6D100K (DB4 ถึง DB6)



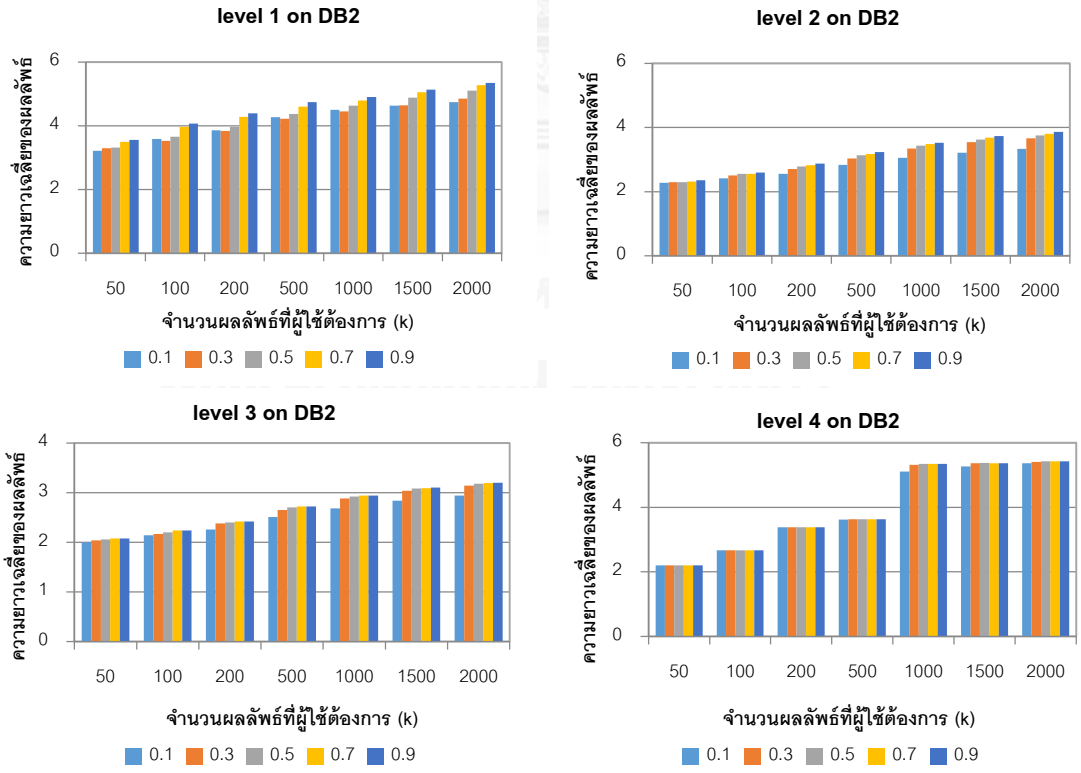
ภาพที่ 4-39 ค่าเฉลี่ยค่าความเชื่อมั่นของผลลัพธ์ทุกลำดับชั้นกับข้อมูล T40I10D100K (DB7 ถึง DB9)

4.3.5 ผลวิเคราะห์ความยาวเฉลี่ยของกฎความสัมพันธ์ที่น่าสนใจสุดจำแนกแต่ละลำดับชั้น

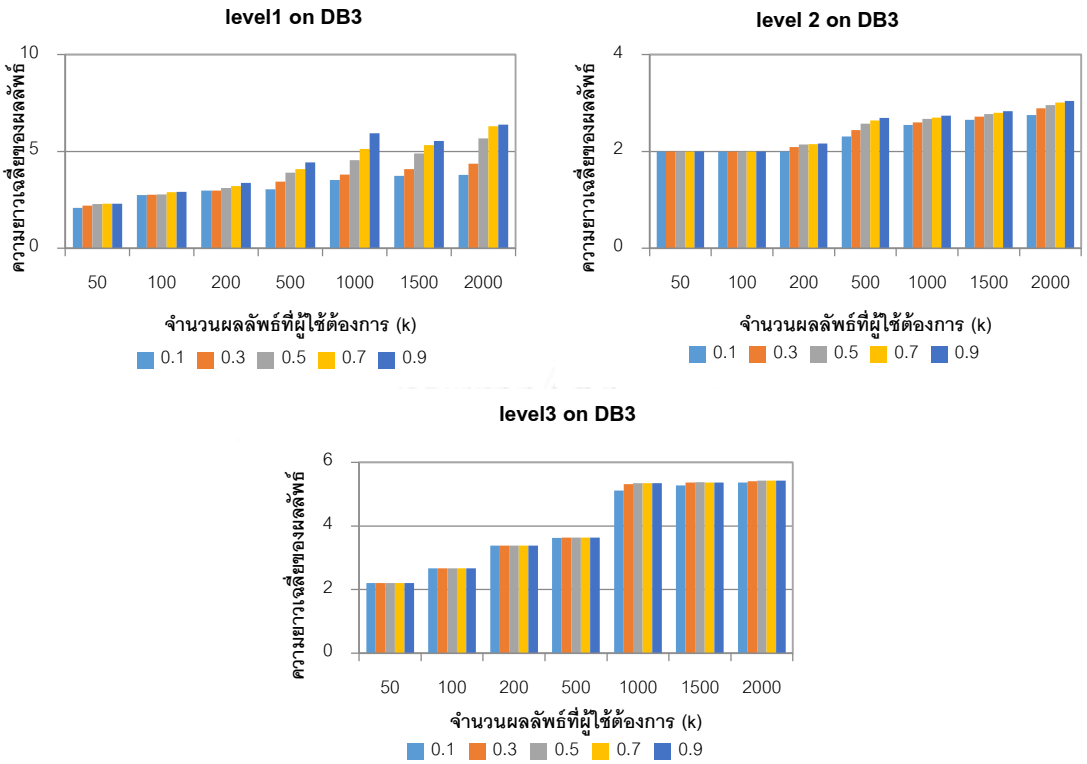
จากภาพที่ 4-40 ถึง 4-48 ที่วิเคราะห์หาความยาวเฉลี่ยของผลลัพธ์แต่ละลำดับชั้นโดยทดสอบกับทุกชุดข้อมูล พบว่า เมื่อต้องการจำนวนผลลัพธ์ที่มากขึ้น จะได้กฎความสัมพันธ์ที่น่าสนใจที่สุดในแต่ละลำดับชั้นที่มีความยาวของกฎเพิ่มมากขึ้น ซึ่งทำให้ค่าความยาวเฉลี่ยของผลลัพธ์ที่น่าสนใจมีค่ามากขึ้นด้วย และหากค่าถ่วงน้ำหนักความน่าสนใจของผลลัพธ์มีค่ามากขึ้น ก็จะได้ผลลัพธ์ที่น่าสนใจของกฎความสัมพันธ์มีความยาวของกฎมากขึ้น จึงทำให้ค่าเฉลี่ยค่าความยาวของผลลัพธ์ที่น่าสนใจก็มีค่ามากขึ้นเช่นเดียวกัน แสดงให้เห็นว่า ผลลัพธ์ที่ได้ในแต่ละค่าที่กำหนดขึ้นทั้งค่าถ่วงน้ำหนักและจำนวนผลลัพธ์ที่ต่างกัน ก็จะให้กฎความสัมพันธ์ที่มีลักษณะที่แตกต่างกัน ทั้งนี้ขึ้นกับค่าความสนใจของแต่ละกฎที่ได้จากการคำนวณ



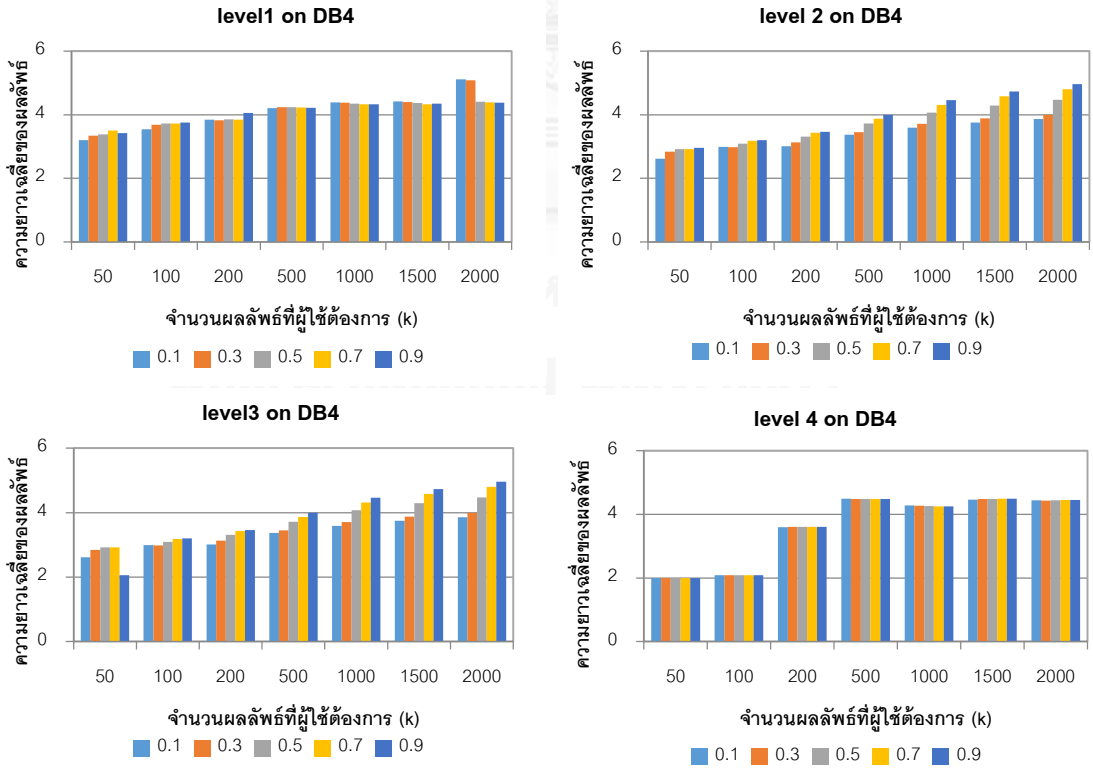
ภาพที่ 4-40 ความยาวเฉลี่ยของผลลัพธ์ที่น่าสนใจที่สุดแต่ละลำดับชั้นกับชุดข้อมูล DB1



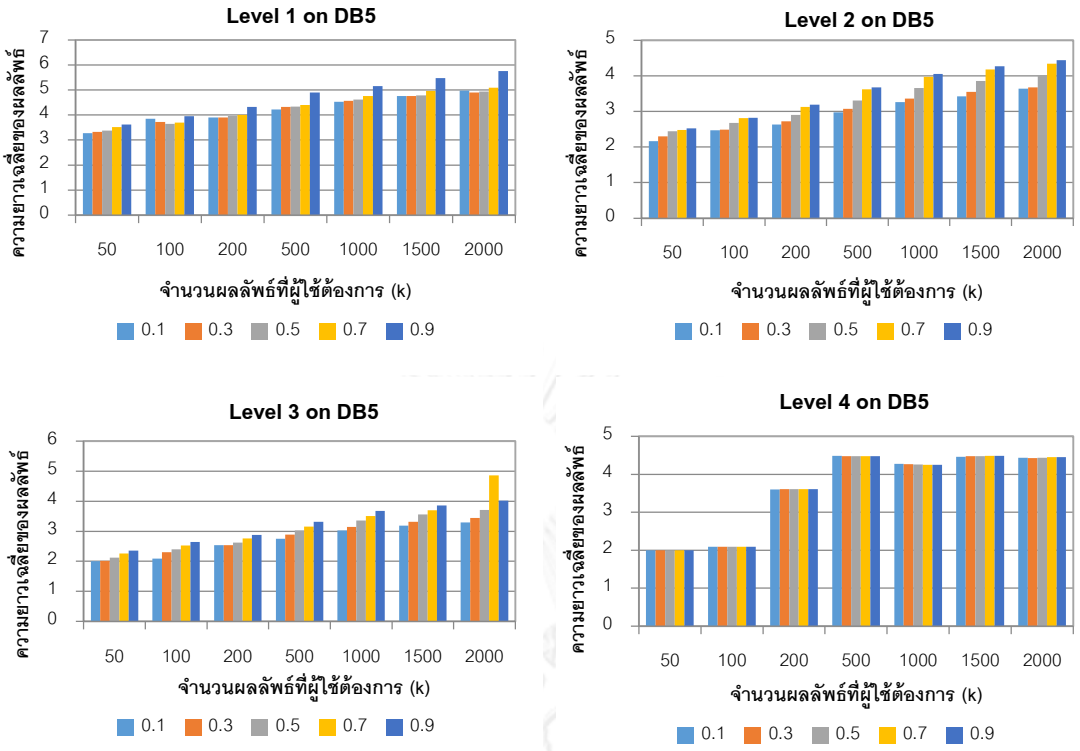
ภาพที่ 4-41 ความยาวเฉลี่ยของผลลัพธ์ที่น่าสนใจที่สุดแต่ละลำดับชั้นกับชุดข้อมูล DB2



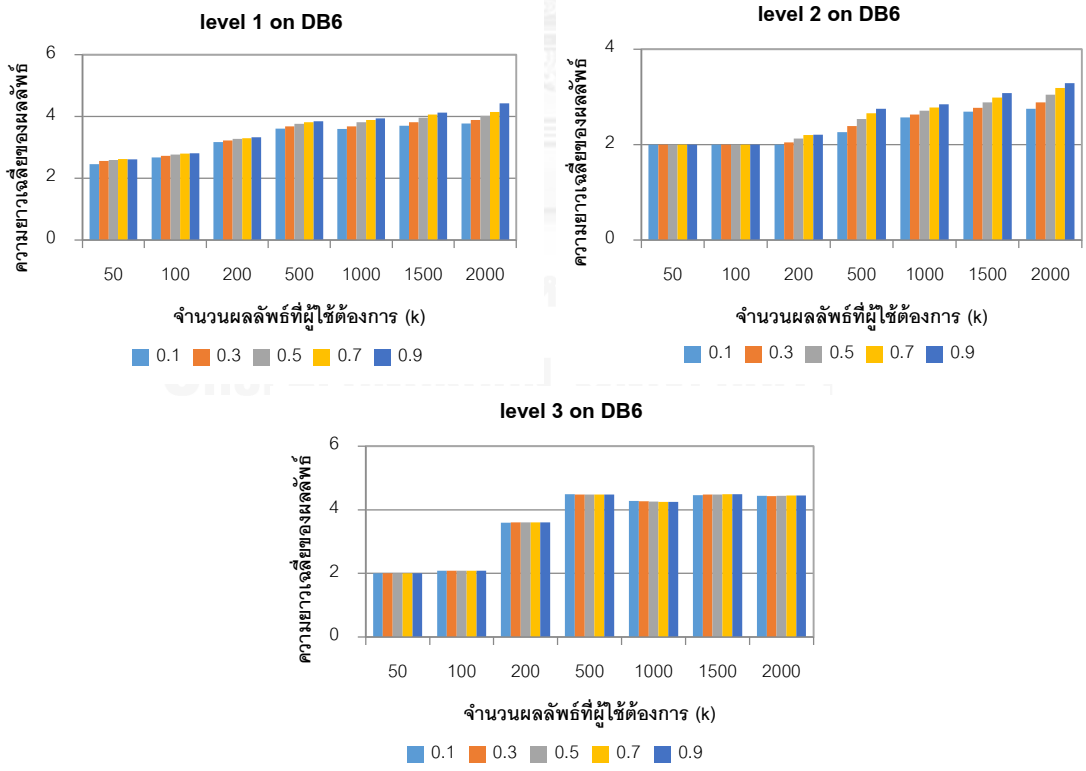
ภาพที่ 4-42 ความยาวเฉลี่ยของผลลัพธ์ที่น่าสนใจที่สุดแต่ละลำดับชั้นกับชุดข้อมูล DB3



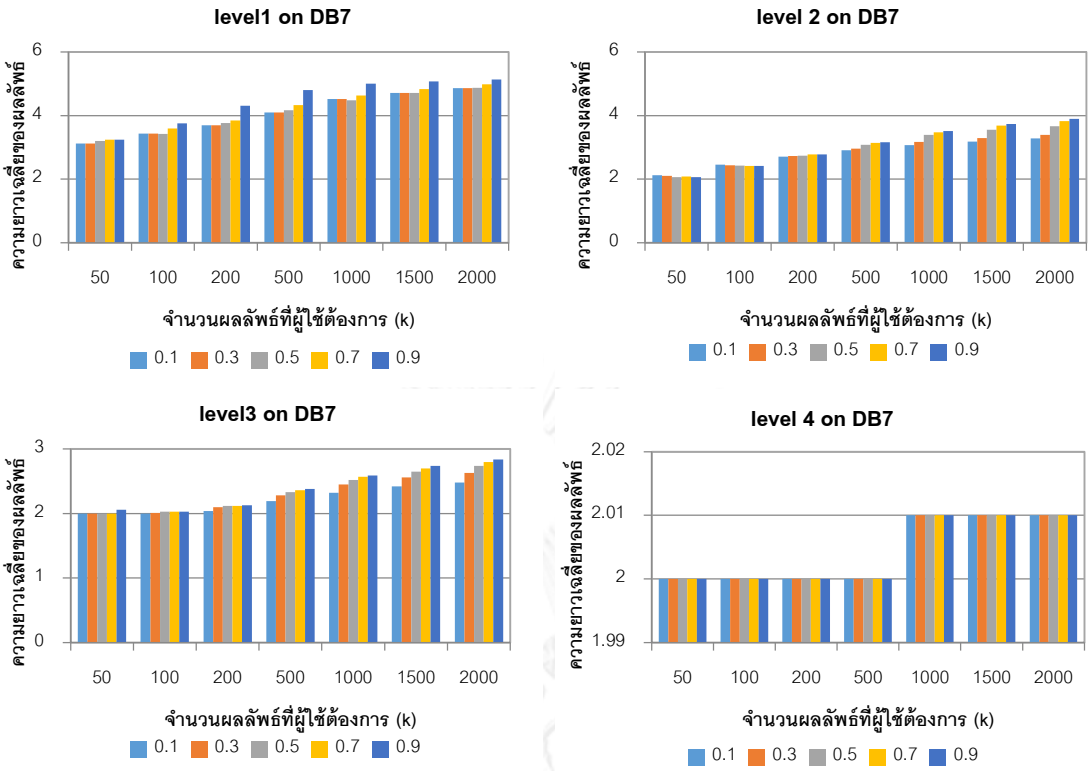
ภาพที่ 4-43 ความยาวเฉลี่ยของผลลัพธ์ที่น่าสนใจที่สุดแต่ละลำดับชั้นกับชุดข้อมูล DB4



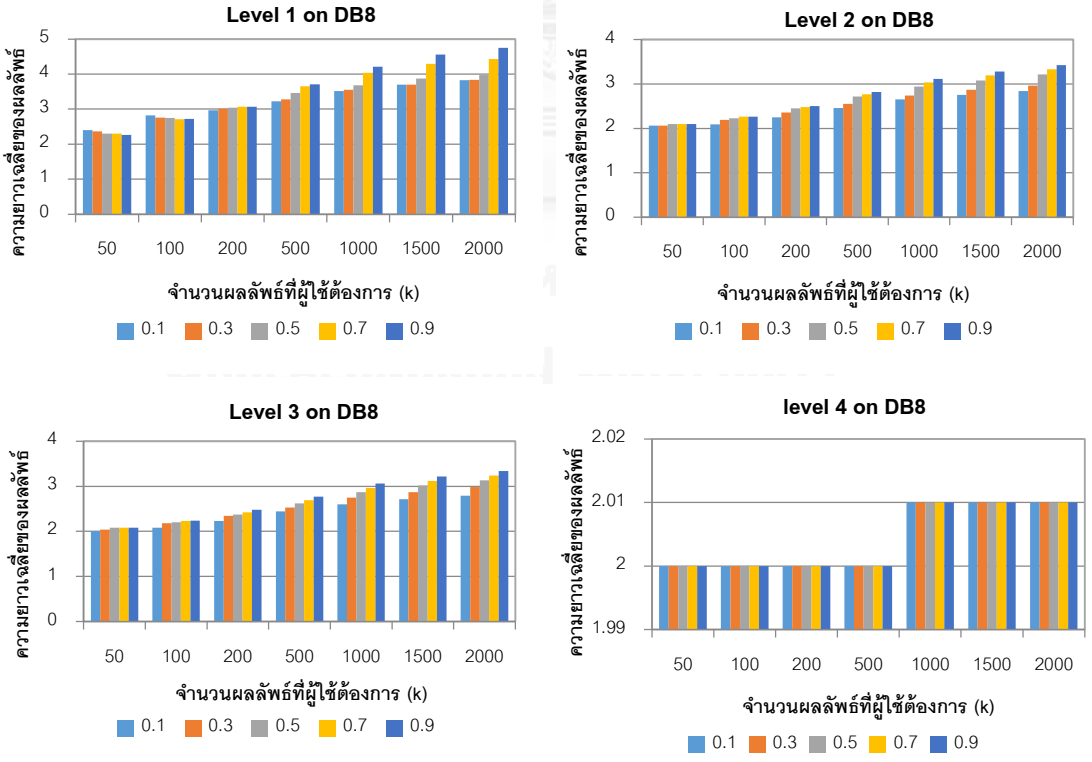
ภาพที่ 4-44 ความยาวเฉลี่ยของผลลัพธ์ที่น่าสนใจที่สุดแต่ละลำดับชั้นกับชุดข้อมูล DB5



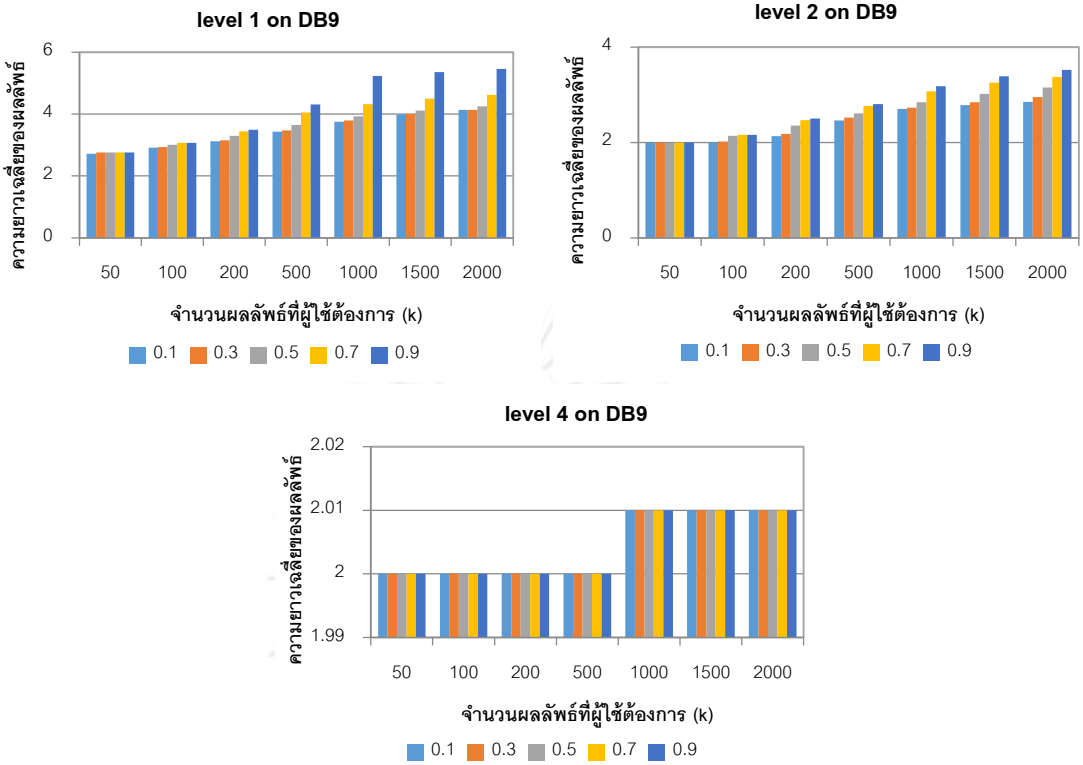
ภาพที่ 4-45 ความยาวเฉลี่ยของผลลัพธ์ที่น่าสนใจที่สุดแต่ละลำดับชั้นกับชุดข้อมูล DB6



ภาพที่ 4-46 ความยาวเฉลี่ยของผลลัพธ์ที่น่าสนใจที่สุดแต่ละลำดับชั้นกับชุดข้อมูล DB7



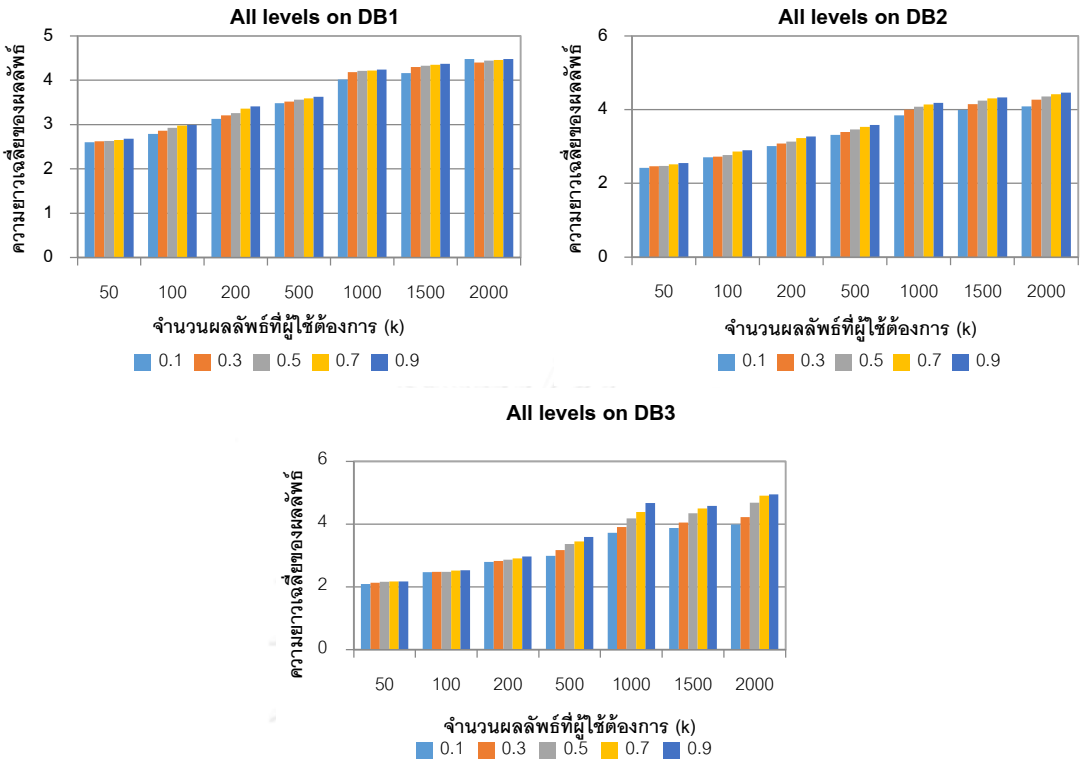
ภาพที่ 4-47 ความยาวเฉลี่ยของผลลัพธ์ที่น่าสนใจที่สุดแต่ละลำดับชั้นกับชุดข้อมูล DB8



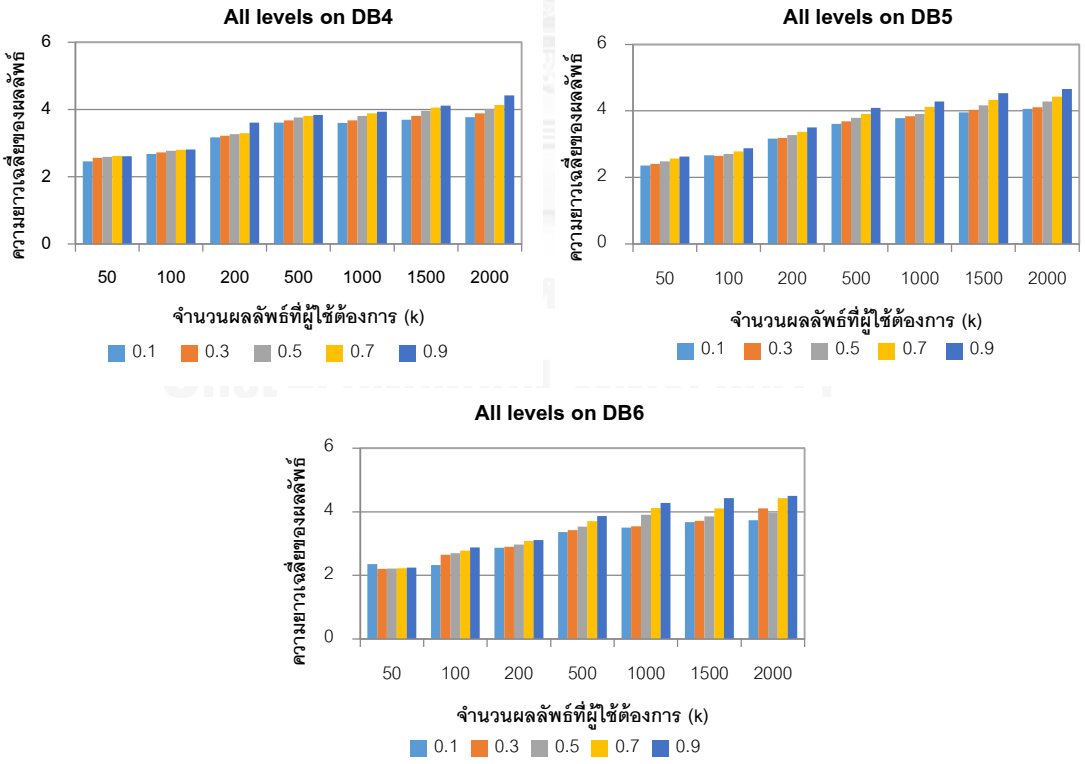
ภาพที่ 4-48 ความยาวเฉลี่ยของผลลัพธ์ที่น่าสนใจที่สุดแต่ละลำดับชั้นกับชุดข้อมูล DB9

4.3.6 ผลวิเคราะห์ความยาวเฉลี่ยของกฎความสัมพันธ์ที่น่าสนใจที่สุดทุกๆ ลำดับชั้น

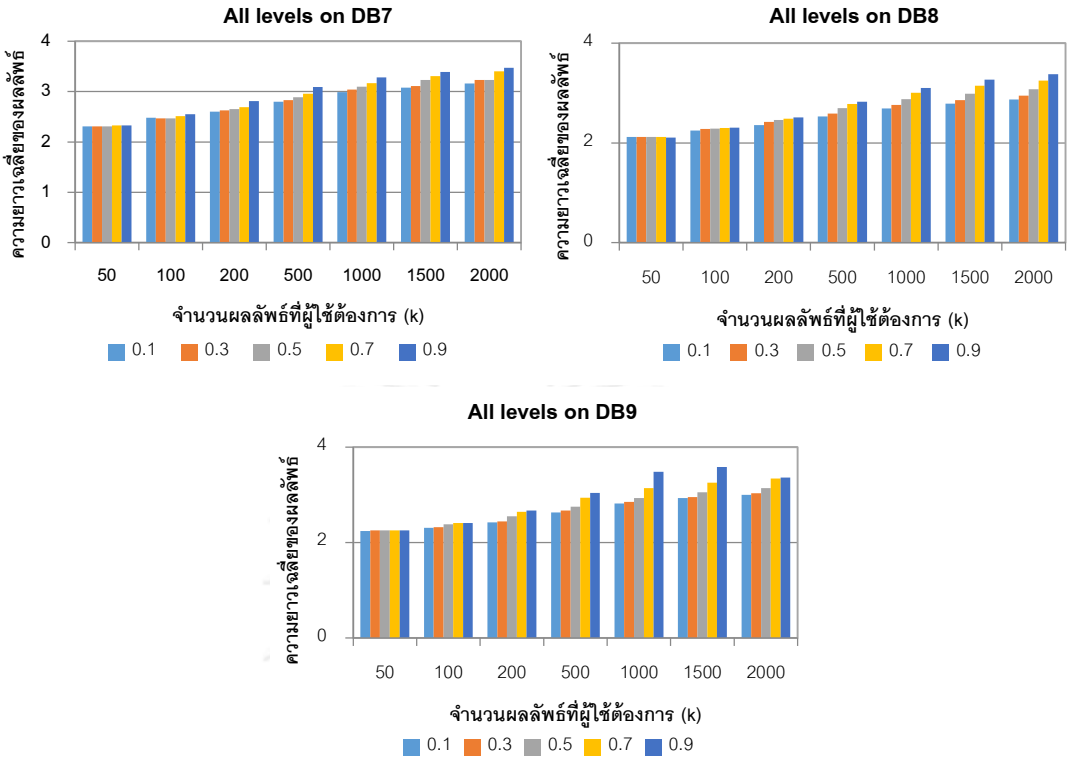
จากภาพที่ 4-49 ถึง 4-51 ที่วิเคราะห์หาความยาวเฉลี่ยของผลลัพธ์ทุกลำดับชั้นโดยทดสอบกับทุกชุดข้อมูล พบว่าหากต้องการจำนวนผลลัพธ์ที่มากขึ้น ความยาวเฉลี่ยของผลลัพธ์ที่น่าสนใจในทุกๆ ลำดับชั้นมีค่าเพิ่มขึ้นด้วย ซึ่งมีผลมาจากความยาวเฉลี่ยของกฎความสัมพันธ์ที่น่าสนใจที่สุดในแต่ละลำดับชั้นมีค่าเฉลี่ยที่สูงขึ้นด้วย นอกจากนี้เมื่อกำหนดค่าถ่วงน้ำหนักความน่าสนใจของผลลัพธ์เพิ่มขึ้น ความยาวเฉลี่ยของผลลัพธ์ที่น่าสนใจที่สุดที่ได้ในทุกลำดับชั้นก็มีค่าเฉลี่ยมากขึ้น สอดคล้องกับข้อ 4.3.5



ภาพที่ 4-49 ความยาวเฉลี่ยของผลลัพธ์ทุกลำดับชั้นกับชุดข้อมูล T10I4D100K (DB1 ถึง DB3)



ภาพที่ 4-50 ความยาวเฉลี่ยของผลลัพธ์ทุกลำดับชั้นกับชุดข้อมูล T20I6D100K (DB4 ถึง DB6)



ภาพที่ 4-51 ความยาวเฉลี่ยของผลลัพธ์ทุกลำดับชั้นกับชุดข้อมูล T40I10D100K (DB7 ถึง DB9)

บทที่ 5
สรุปผลการวิจัย

5.1 บทสรุป

จากปัญหาการกำหนดค่าขีดแบ่งสนับสนุนและ/หรือค่าขีดแบ่งความเชื่อมั่นเพื่อหากฎความสัมพันธ์แบบหลายลำดับชั้นที่เหมาะสม ตรงตามกับความต้องการของผู้ใช้งานนั้นกำหนดได้ยาก ซึ่งงานวิจัยนี้เป็นงานวิจัยทางเลือกใหม่สำหรับผู้ใช้ในการหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจที่สุด โดยที่ผู้ใช้นั้นสามารถกำหนดจำนวนผลลัพธ์ (ค่า K) และค่าถ่วงน้ำหนักค่าความน่าสนใจของผลลัพธ์เพื่อหากฎความสัมพันธ์ที่น่าสนใจที่สุดตามที่จำนวนผู้ใช้ต้องการ ซึ่งสามารถช่วยขจัดปัญหาของผู้ใช้ในการกำหนดค่าขีดแบ่งสนับสนุนและ/หรือค่าขีดแบ่งความเชื่อมั่นที่เหมาะสม และได้ผลลัพธ์ตรงตามที่ผู้ใช้ต้องการได้

งานวิจัยนี้ได้เสนอขั้นตอนวิธีการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจที่สุดตามจำนวนผลลัพธ์ที่ผู้ใช้ต้องการ ที่มีชื่อว่าเอ็นเอ็มแอลเอฟพี (NMLFP) โดยใช้โครงสร้างต้นไม้แสดงรูปแบบการเกิดขึ้นของข้อมูล เพื่อหารูปแบบเซตที่ปรากฏบ่อย และนำต้นไม้นั้นไปสร้างต้นไม้แสดงรูปแบบ ณ ลำดับชั้นข้อมูลที่สูงกว่าด้วย สำหรับการหารูปแบบเซตที่ปรากฏบ่อยในแต่ละลำดับชั้นนั้นจะเริ่มพิจารณาข้อมูลที่มีความถี่สูงสุด (ปรากฏบ่อยสุด) อันดับสองก่อน เพื่อการค้นหานั้นจะได้รูปแบบเซตที่มีค่าสนับสนุนสูงๆ ก่อน ซึ่งเป็นการลดเวลาในการค้นหาผลลัพธ์ และยังสามารถเพิ่มเกณฑ์ค่าสนับสนุนที่สามารถตัดรูปแบบเซตที่มีค่าสนับสนุนต่ำๆ ได้ด้วย สำหรับการสร้างต้นไม้แสดงรูปแบบการเกิดขึ้นของข้อมูล ณ ลำดับชั้นที่สูงกว่า ได้นับค่าสนับสนุนข้อมูลลำดับชั้นใหม่จากต้นไม้แสดงรูปแบบต้นเดิม จากนั้นทำการเรียงข้อมูลลำดับชั้นใหม่ตามค่าสนับสนุนจากมากไปน้อย และทำการสำรวจต้นไม้เดิมอีกครั้งเพื่อเรียงลำดับข้อมูลในแต่ละกิ่งให้สอดคล้องกับการเรียงข้อมูลตามค่าสนับสนุน และนำกิ่งเหล่านั้นไปสร้างเป็นต้นไม้แสดงรูปแบบการเกิดขึ้น ณ ลำดับชั้นใหม่ต่อไป สำหรับการสร้างกฎความสัมพันธ์ที่น่าสนใจที่สุดในแต่ละลำดับชั้น ได้ใช้ค่าถ่วงน้ำหนักความน่าสนใจของผลลัพธ์ เพื่อให้ผู้ใช้นั้นสามารถกำหนดคุณลักษณะความน่าสนใจของกฎความสัมพันธ์ได้ ซึ่งหมายถึงผู้ใช้อาจต้องการกฎความสัมพันธ์ที่สนใจค่าสนับสนุนสูงๆ หรือผู้ใช้อาจต้องการกฎความสัมพันธ์ที่สนใจค่าความเชื่อมั่นสูงๆ โดยการกำหนดค่าถ่วงน้ำหนักความน่าสนใจของค่าสนับสนุนและ/หรือค่าความเชื่อมั่นของผลลัพธ์นั้นจะกำหนดในช่วง 0 ถึง 1 เท่านั้น

จากผลการทดลองในบทที่ 4 เมื่อเทียบจำนวนผลลัพธ์ที่ต้องการและค่าถ่วงน้ำหนักความน่าสนใจของค่าความเชื่อมั่น (W_C) จากขั้นตอนวิธีเอ็นเอ็มแอลเอฟพีกับค่าขีดแบ่งสนับสนุนและค่าขีดแบ่งความเชื่อมั่นที่ใช้สำหรับการทดสอบจากขั้นตอนวิธีเอฟพีเอ็ม-ที พบว่า เวลาที่ใช้ในการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจสุด โดยส่วนใหญ่ขั้นตอนวิธีเอ็นเอ็มแอลเอฟพีจะใช้เวลาเร็วกว่าการค้นหาผลลัพธ์ด้วยขั้นตอนวิธีเอฟพีเอ็ม-ที ยกเว้นเมื่อกำหนดค่าถ่วงน้ำหนักที่ 0.9 ขั้นตอนวิธีเอ็นเอ็มแอลเอฟพีจะใช้เวลามากกว่าการค้นหาด้วยขั้นตอนวิธีเอฟพีเอ็ม-ที เนื่องจาก ค่าขีดแบ่งที่ใช้นั้นเป็นค่าขีดแบ่งที่เหมาะสมสำหรับการค้นหาผลลัพธ์ที่เหมาะสำหรับผู้ใช้ และได้จำนวนผลลัพธ์ที่ใกล้เคียงกับจำนวนที่ผู้ใช้ต้องการ ในส่วนด้านการใช้หน่วยความจำ การค้นหาจากขั้นตอนวิธีเอ็นเอ็ม

แอลเอฟพีใช้หน่วยความจำน้อยกว่าการค้นหาจากขั้นตอนวิธีเอฟพีเอ็ม-ที เพราะการค้นหาผลลัพธ์ที่กำหนดจำนวนผลลัพธ์สามารถที่จะประมาณการการใช้หน่วยความจำตามจำนวนผลลัพธ์ได้

นอกจากนี้ได้วิเคราะห์ผลลัพธ์ที่ได้จากการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจที่สุด โดยหากกำหนดจำนวนผลลัพธ์ที่มีค่าเพิ่มขึ้น มีผลทำให้ค่าเฉลี่ยสนับสนุนของผลลัพธ์ทั้งในแต่ละลำดับชั้นและรวมค่าเฉลี่ยทุกลำดับชั้นมีค่าน้อยลง แต่ทำให้ความยาวเฉลี่ยของผลลัพธ์ที่ได้แต่ละลำดับชั้นและรวมทุกลำดับชั้นมีค่ามากขึ้น มากกว่านั้นไม่ว่าจำนวนผลลัพธ์จะมีค่าเพิ่มขึ้นเพียงใด ค่าเฉลี่ยค่าความเชื่อมั่นที่ได้จะมีค่าที่ใกล้เคียงกัน ซึ่งแสดงให้เห็นว่าเมื่อกำหนดจำนวนผลลัพธ์ที่ต่างกัน กฎความสัมพันธ์ที่ได้ถึงแม้ว่าจะมีค่าสนับสนุนที่แตกต่างกัน แต่ก็ให้ค่าความเชื่อมั่นของกฎที่ใกล้เคียงกัน และถ้าหากค่าถ่วงน้ำหนักความน่าสนใจของความค่าเชื่อมั่นมีค่าเพิ่มขึ้น โดยพิจารณาจำนวนผลลัพธ์คงที่ จะพบว่าค่าเฉลี่ยสนับสนุนของผลลัพธ์แต่ละลำดับชั้นและทุกลำดับชั้นมีค่าน้อยลง แต่ทำให้ค่าเฉลี่ยความเชื่อมั่นของผลลัพธ์ที่ได้กับความยาวเฉลี่ยของผลลัพธ์แต่ละลำดับชั้นและทุกลำดับชั้นมีค่ามากขึ้น

5.2 ปัญหาและข้อจำกัดที่พบ

จากการทดสอบการค้นหากฎความสัมพันธ์ที่น่าสนใจสุดตามจำนวนที่ผู้ใช้ต้องการ เมื่อใช้ชุดข้อมูลทุกชุดข้อมูลที่มีการจัดแบ่งลำดับชั้นของข้อมูลแบบ 8-5-5-5 พบว่า หากกำหนดจำนวนผลลัพธ์สูงขึ้น จะทำให้รูปแบบเซตที่ปรากฏบ่อยสุดหรือกฎความสัมพันธ์ที่น่าสนใจสุด ณ ลำดับชั้นบนสุด (ลำดับชั้นที่หนึ่ง) ให้ผลลัพธ์ทุกๆ รูปแบบเซตที่เป็นไปได้ ซึ่งผลลัพธ์นั้นอาจเป็นผลลัพธ์ที่ไม่ก่อให้เกิดเป็นองค์ความรู้ที่ดีสำหรับผู้ใช้ได้

5.3 ข้อเสนอแนะ

เนื่องจากผลลัพธ์ที่ได้จากการค้นหากฎความสัมพันธ์แบบหลายลำดับชั้นที่น่าสนใจที่สุดตามจำนวนที่ผู้ใช้ต้องการนั้นได้ให้ผลลัพธ์เฉพาะกฎความสัมพันธ์แบบลำดับชั้นต่อลำดับชั้น อย่างไรก็ดี ผู้ใช้อาจต้องการกฎความสัมพันธ์แบบข้ามลำดับชั้นด้วย เช่น ถ้าซื้อนมช็อกโกแลตก็จะซื้อคุกกี้ด้วย (chocolate milk → cookies) ดังนั้นจึงมีข้อเสนอแนะที่จะคิดค้นหาขั้นตอนวิธีที่เหมาะสมสำหรับการหากฎความสัมพันธ์แบบข้ามลำดับชั้นตามจำนวนที่ผู้ใช้ต้องการ โดยอาจนำรูปแบบเซตที่ปรากฏบ่อยในแต่ละลำดับชั้นที่ได้จากการค้นหาในงานวิจัย มาทำการสร้างกฎความสัมพันธ์แบบข้ามลำดับชั้นที่น่าสนใจที่สุด โดยที่ค่าสนับสนุนของกฎความสัมพันธ์แบบข้ามลำดับชั้นที่ได้นั้นจะต้องมีค่าไม่น้อยกว่าค่าสนับสนุนของรูปแบบเซต ณ ลำดับชั้นล่างสุด ที่นำมาสร้างกฎความสัมพันธ์แบบข้ามลำดับชั้น นอกจากนั้น จำนวนกฎความสัมพันธ์ในแต่ละลำดับชั้นที่ได้จากงานวิจัยนี้มีจำนวนเท่ากัน ดังนั้นจึงข้อเสนอแนะในการปรับปรุงขั้นตอนวิธีของงานวิจัยนี้ ให้ผู้ใช้นั้นสามารถกำหนดจำนวนผลลัพธ์ที่ต้องการในแต่ละลำดับชั้นที่แตกต่างกันได้

รายการอ้างอิง

1. Agrawal, R., et al., *Mining association rules between sets of items in large databases*, in *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*. 1993, ACM: Washington, D.C., USA. p. 207-216.
2. Agrawal, R. and R. Srikant, *Fast Algorithms for Mining Association Rules in Large Databases*, in *Proceedings of the 20th International Conference on Very Large Data Bases*. 1994, Morgan Kaufmann Publishers Inc. p. 487-499.
3. Han, J. and Y. Fu, *Discovery of Multiple-Level Association Rules from Large Databases*, in *Proceedings of the 21th International Conference on Very Large Data Bases*. 1995, Morgan Kaufmann Publishers Inc. p. 420-431.
4. Fu, A.W.-c., R.W.-w. Kwong, and J. Tang, *Mining n-most interesting itemsets*, in *Foundations of Intelligent Systems*. 2000, Springer. p. 59-67.
5. Han, J., et al., *Mining Top.K Frequent Closed Patterns without Minimum Support*, in *Proceedings of the 2002 IEEE International Conference on Data Mining*. 2002, IEEE Computer Society. p. 211.
6. Songram, P. and V. Boonjing. *N-Most Interesting Closed Itemset Mining*. in *Convergence and Hybrid Information Technology, 2008. ICCIT '08. Third International Conference on*. 2008.
7. Wong, R.-W. and A.-C. Fu, *Mining top-K frequent itemsets from data streams.* Data Mining and Knowledge Discovery, 2006. **13**(2): p. 193-217.
8. Amphawan, K., *Mining top-k regular-frequent itemsets from transactional database*, in *Computer Engineering*. 2010, Chulalongkorn University.
9. Han, J., J. Pei, and Y. Yin, *Mining frequent patterns without candidate generation*, in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 2000, ACM: Dallas, Texas, USA. p. 1-12.
10. El-Hajj, M. and O.R. Zaïane. *COFI-tree mining: A new approach to pattern growth with reduced candidacy generation*. in *Workshop on Frequent Itemset Mining Implementations (FIMI'03) in conjunction with IEEE-ICDM*. 2003.
11. Sucahyo, Y.G. and R.P. Gopalan. *CT-PRO: A Bottom-Up Non Recursive Frequent Itemset Mining Algorithm Using Compressed FP-Tree Data Structure*. in *FIMI*. 2004.
12. Jian, P., et al. *H-mine: hyper-structure mining of frequent patterns in large databases*. in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. 2001.

13. Fortin, S. and L. Liu, *An object-oriented approach to multi-level association rule mining*, in *Proceedings of the fifth international conference on Information and knowledge management*. 1996, ACM: Rockville, Maryland, USA. p. 65-72.
14. Hong, T.-P., K.-Y. Lin, and S.-L. Wang, *Fuzzy data mining for interesting generalized association rules.* Fuzzy Sets Syst., 2003. **138**(2): p. 255-269.
15. Lee, Y.-C., T.-P. Hong, and T.-C. Wang, *Multi-level fuzzy mining with multiple minimum supports.* Expert Syst. Appl., 2008. **34**(1): p. 459-468.
16. Arya, S.a.R.A., *Mining Multiple Level Association Rules to Mining Multiple Level Correlations to discover Complex Pattern.* International Journal of Computer Applications, 2012. **58**.
17. Ong, K.-L., W.-K. Ng, and E.-P. Lim. *Mining multi-level rules with recurrent items using FP'-Tree*. in *Proceedings of the Third International Conference on Information, Communications and Signal Processing*. 2001. Citeseer.
18. Eavis, T. and X. Zheng, *Multi-level Frequent Pattern Mining*, in *Proceedings of the 14th International Conference on Database Systems for Advanced Applications*. 2009, Springer-Verlag: Brisbane, Australia. p. 369-383.
19. Kaya, M. and R. Alhajj. *Mining multi-cross-level fuzzy weighted association rules*. in *Intelligent Systems, 2004. Proceedings. 2004 2nd International IEEE Conference*. 2004.
20. Huang, Y.-M., J.-N. Chen, and S.-C. Cheng, *A Method of Cross-level Frequent Pattern Mining for Web-based Instruction.* Journal of Educational Technology & Society, 2007. **10**(3).
21. Ngan, S.-C., et al., *Mining N-most interesting itemsets without support threshold by the COFI-tree.* IJBIDM, 2005. **1**(1): p. 88-106.
22. Salam, A. and M.S. Khayal, *Mining top–k frequent patterns without minimum support threshold.* Knowledge and Information Systems, 2012. **30**(1): p. 57-86.
23. Fournier-Viger, P., C.-W. Wu, and V.S. Tseng, *Mining top-k association rules*, in *Proceedings of the 25th Canadian conference on Advances in Artificial Intelligence*. 2012, Springer-Verlag: Toronto, ON, Canada. p. 61-73.
24. Goethals, B. *Frequent Itemset Mining Dataset Repository*. 2003; Available from: http://fimi.ua.ac.be/data/.

ภาคผนวก

ประวัติผู้เขียนวิทยานิพนธ์

นายสรพล ชมไพศาล เกิดเมื่อวันที่ 24 กันยายน พ.ศ. 2530 ที่กรุงเทพมหานคร สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยหอการค้าไทย ในปีการศึกษา 2552 จากนั้นทำงานที่ศูนย์เทคโนโลยีสารสนเทศซาเลเซียน ในตำแหน่งผู้ดูแลและพัฒนาเว็บไซต์ของหน่วยงานต่างในคณะซาเลเซียน และปีพ.ศ. 2554 ได้เข้าศึกษาต่อในหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย