การเปรียบเทียบประสิทธิภาพของการประยุกต์เทคนิคการค้นหากฎความสัมพันธ์บนข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นและตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่

นายสัญชัย พิทักษ์ชลทรัพย์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาการพัฒนาซอฟต์แวร์ด้านธุรกิจ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2553

A COMPARISON OF THE EFFICIENCY OF APPLYING ASSOCIATION RULE DISCOVERY ON SOFTWARE ARCHIVE USING SUPPORT – CONFIDENCE MODEL AND SUPPORT – NEW CONFIDENCE MODEL

Mr. Sunchai Pitakchonlasup

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science Program in Business Software Development

Department of Statistics

Faculty of Commerce and Accountancy

Chulalongkorn University

Academic Year 2010

หัวข้อวิทยานิพนธ์	การเปรียบเทียบประสิทธิภาพของการประยุกต์เทคนิคการค้นหากฎความสัมพันธ์บนข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นและตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่
โดย	นาย สัญชัย พิทักษ์ชลทรัพย์
สาขาวิชา	การพัฒนาซอฟต์แวร์ด้านธุรกิจ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร. อัษฎาพร ทรัพย์สมบูรณ์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี
(รองศาสตราจารย์ ดร.อรรณพ ตันละมัย)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สมจารี ปรียานนท์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร. อัษฎาพร ทรัพย์สมบูรณ์)

..... กรรมการ
(อาจารย์ ดร.จันทร์เจ้า มงคลนาวิน)

..... กรรมการภายนอกมหาวิทยาลัย
(อาจารย์ ดร.รัชต พีชวณิชย์)

สัญชัย พิทักษ์ชลทรัพย์ : การเปรียบเทียบประสิทธิภาพของการประยุกต์เทคนิคการค้นหากฎความสัมพันธ์บนข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นและตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่. (A Comparison of the Efficiency of Applying Association Rule Discovery on Software Archive using Support – Confidence Model and Support – New Confidence Model)
อ. ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ.ดร. อัษฎาพร ทรัพย์สมบูรณ์, 261 หน้า.

ข้อมูลซอฟต์แวร์อาร์ไคฟ์เป็นข้อมูลที่รวบรวมกระบวนการพัฒนาระบบซอฟต์แวร์ในอดีตเอาไว้ทั้งหมด การประยุกต์เทคนิคการค้นหากฎความสัมพันธ์บนข้อมูลซอฟต์แวร์อาร์ไคฟ์สามารถสกัดรูปแบบของการพัฒนาที่เกิดขึ้นในอดีตออกมาได้ รูปแบบของการพัฒนาที่เกิดขึ้นในอดีตเหล่านี้มีประโยชน์มากต่อขั้นตอนการพัฒนาและการบำรุงรักษาซอฟต์แวร์ อาทิเช่น แนะนำนักพัฒนาว่าเมื่อแก้ไขซอฟต์แวร์ในส่วนนี้แล้ว นักพัฒนาควรจะต้องไปแก้ไขซอฟต์แวร์ที่ส่วนใดอีก งานวิจัยในอดีตที่ศึกษาการประยุกต์เทคนิคดังกล่าวมักจะใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นเป็นตัวแบบในการประเมินความน่าสนใจของกฎความสัมพันธ์ แต่ตัวแบบนี้มีข้อบกพร่องที่สำคัญคือสามารถให้ผลลัพธ์ที่เป็นผลบวกลวงจำนวนมาก ต่อมา Liu และคณะ (Liu et al., 2008) ได้เสนอตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่เพื่อปรับปรุงข้อบกพร่องดังกล่าวของตัวแบบเดิม

งานวิจัยนี้เป็นงานวิจัยเชิงทดลองเพื่อเปรียบเทียบประสิทธิภาพของการค้นหากฎความสัมพันธ์บนข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบเดิมและตัวแบบใหม่ ผู้วิจัยสร้างเครื่องมือขึ้นมาเพื่อใช้ในการทดสอบเปรียบเทียบกับข้อมูลซอฟต์แวร์อาร์ไคฟ์ของโครงการพัฒนาซอฟต์แวร์ชื่อเคมายมันนี่ (KMyMoney) ผลการประเมินแสดงให้เห็นว่าประสิทธิภาพของกฎความสัมพันธ์ที่ได้จากการใช้ตัวแบบใหม่มีค่าสูงกว่าการใช้ตัวแบบเดิมในสถานการณ์การนำทางอย่างมีนัยสำคัญ แต่มีค่าไม่ต่างกันหรือน้อยกว่าในสถานการณ์การป้องกันการเกิดข้อผิดพลาดและสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

ภาควิชา.....สถิติ.....ลายมือชื่อนิสิต.....สัญชัย พิทักษ์ชลทรัพย์.....
สาขาวิชา..การพัฒนาซอฟต์แวร์ด้านธุรกิจ..ลายมือชื่ออ.ที่ปรึกษาวิทยานิพนธ์หลัก..[signature]...
ปีการศึกษา.....2553.....

5181932726 : MAJOR BUSSINESS SOFTWARE DEVELOPMENT
KEYWORDS : ASSOCIATION RULE DISCOVERY / VERSION CONTROL SYSTEM / INTERESTINGNESS MEASURES / SOFTWARE ARCHIVES / CONFIDENCE

SUNCHAI PITAKCHONLASUP : A COMPARISON OF THE EFFICIENCY OF APPLYING ASSOCIATION RULE DISCOVERY ON SOFTWARE ARCHIVE USING SUPPORT – CONFIDENCE MODEL AND SUPPORT – NEW CONFIDENCE MODEL. THESIS ADVISOR : ASST.PROF. ASSADAPORN SAPSOMBOON, Ph.D., 261 pp.

Software archives contain historical information about the development process of a software system. Applying association rule discovery on these archives, the software development patterns can be extracted. This information is useful to support software modification activities, such as it indicate which modules are usually modified together during software maintenance. All previous works on association rule mining technique focused on classical interestingness measure model called support-confidence. This model has a defect which offers a number of false positives. The new model, named support-new confidence, was proposed by Liu et al (Liu et al., 2008), to improve the false positive issue in the classical model.

This research is an experimental research of the compare the efficiency of applying association rule discovery on software archive using classical model and Liu et al's model. The experiments were conducted on software archive of KMyMoney software project. The results show that the efficiency of the rules obtained in new model are significantly higher than the rules obtained in classical model in nevigation scenario but are not difference or lower than the rules obtained in classical model in error prevention and closure scenario.

Department : .....Statistics..... Student's Signature...SUNCHAI PITAKCHONLASUP
Field of Study :Business Software DevelopmentAdvisor's Signature...Assadaporn Sapsomboon
Academic Year : .....2010.....

กิตติกรรมประกาศ

วิทยานิพนธ์นี้จะสำเร็จลุล่วงและสมบูรณ์ไปได้ด้วยดีต้องขอกราบขอบพระคุณผู้ช่วยศาสตราจารย์ ดร. อัษฎาพร ทรัพย์สมบูรณ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์เป็นอย่างยิ่งที่ได้ให้คำแนะนำและข้อคิดเห็นต่างๆ ตรวจแก้วิทยานิพนธ์ฉบับนี้อย่างละเอียด ตลอดจนแนวทางในการวิจัยด้วยดีตลอดมา ขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.สมจารี ปรียานนท์ และอาจารย์ ดร. จันทร์เจ้า มงคลนาวิน กรรมการสอบวิทยานิพนธ์ที่กรุณาเสียสละเวลาให้คำปรึกษาและให้คำแนะนำ เพื่อแก้ไขรูปแบบและเนื้อหาวิทยานิพนธ์ฉบับนี้จนเสร็จสมบูรณ์ และขอบพระคุณอาจารย์ ดร. อรุณี กำลัง ที่ให้คำปรึกษาเกี่ยวกับการวิเคราะห์ข้อมูลในการทดลอง

ขอบคุณเพื่อนๆ ที่ให้คำปรึกษาและความช่วยเหลือในด้านต่างๆ ซึ่งทำให้งานวิจัยเป็นไปได้อย่างราบรื่นตลอดจนกำลังใจที่มอบให้เสมอมา

สุดท้ายนี้ ผู้วิจัยใคร่กราบขอบพระคุณบิดา มารดาและครอบครัวที่คอยช่วยเหลือให้การสนับสนุนและคอยเป็นแรงกระตุ้นให้แก่ผู้วิจัยเสมอจนสำเร็จการศึกษา

สารบัญ

หน้า

สารบัญตาราง

สารบัญภาพ

บทที่ 1
ที่มาและความสำคัญของปัญหา

1.1 บทนำ

ในโครงการพัฒนาซอฟต์แวร์ขนาดใหญ่ที่มีช่วงการพัฒนาและช่วงการบำรุงรักษาที่ยาวนาน นักพัฒนามักจะประสบปัญหาหลายประการในแง่ของการทำงานเป็นทีม เช่น ปัญหาการทำความเข้าใจพัฒนาการของซอฟต์แวร์ของนักพัฒนาที่เข้าร่วมทีมใหม่ ปัญหาการเปลี่ยนแปลงแก้ไขซอฟต์แวร์ในช่วงการบำรุงรักษาที่อาจเกิดกับทีมบำรุงรักษาที่ไม่ใช่ทีมเดียวกับทีมพัฒนา ปัญหาการเรียกใช้ซอฟต์แวร์ไลบรารี่ที่ผิดของนักพัฒนาที่เข้าร่วมทีมใหม่ที่อาจนำไปสู่การเกิดข้อผิดพลาด เป็นต้น ปัญหาต่างๆเหล่านี้สามารถตอบสนองได้โดยการประยุกต์ใช้เทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ (Association Rule on Software Archives) งานวิจัยในอดีตที่ศึกษาการประยุกต์เทคนิคดังกล่าวมักจะใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น (Support-Confidence Model) เป็นตัวแบบในการประเมินความน่าสนใจของกฎความสัมพันธ์ แต่ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นมีข้อบกพร่องที่สำคัญ คือ การให้ผลลัพธ์ที่เป็นผลบวกลวงจำนวนมาก ต่อมาในปีค.ศ. 2008 Liu และคณะ (Liu et al., 2008) ได้เสนอตัวแบบในการประเมินความน่าสนใจของกฎความสัมพันธ์ใหม่ขึ้นมาและให้ชื่อว่าค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ (Support-New confidence Model) เพื่อปรับปรุงข้อบกพร่องดังกล่าวของตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นโดยเฉพาะ ผู้วิจัยเห็นว่าการนำค่าสนับสนุน-ค่าความเชื่อมั่นใหม่มาประยุกต์ใช้กับการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์น่าจะสามารถเพิ่มประสิทธิภาพให้กับระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ได้ ส่งผลให้นักพัฒนาสามารถทำงานเป็นทีมได้อย่างมีประสิทธิภาพมากขึ้นด้วย

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาเปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008) ในสถานการณ์ของการให้คำแนะนำนักพัฒนาใน 3 สถานการณ์ได้แก่ 1) สถานการณ์การนำทาง (Navigation) 2) สถานการณ์การป้องกันการเกิดข้อผิดพลาด (Error Prevention) และ 3) สถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (Closure)

ในนี้ผู้วิจัยจะนำเสนอความสำคัญของปัญหาในเบื้องต้นที่ประกอบไปด้วย ความเป็นมาและความสำคัญของปัญหา วัตถุประสงค์ของการวิจัย ตัวแปรที่ศึกษา ประโยชน์ที่คาดว่าจะได้รับ ข้อจำกัดของงานวิจัยนี้ และนิยามของศัพท์สำคัญในงานวิจัยนี้

1.2 ความเป็นมาและความสำคัญของปัญหา

การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์เป็นเทคนิคหนึ่งที่ได้รับความนิยมและถูกนำไปประยุกต์ใช้กับข้อมูลหลายหลากแขนง หนึ่งในนั้นก็คือการประยุกต์ใช้กับข้อมูลซอฟต์แวร์อาร์ไคฟ์หรือข้อมูลการเปลี่ยนแปลงแก้ไขซอฟต์แวร์ในอดีตที่ได้มาจากระบบควบคุมการเปลี่ยนแปลงแก้ไข (Revision Control System, Version Control System) เพื่อประโยชน์ในการแก้ไขปัญหาต่างๆที่เกิดขึ้นกับนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ ความสำคัญของระบบควบคุมการเปลี่ยนแปลงแก้ไขและปัญหาของนักพัฒนาในโครงการพัฒนาซอฟต์แวร์ขนาดใหญ่ถูกรวบรวมเอาไว้หัวข้อ 1.2.1 แต่การตอบสนองปัญหาเหล่านั้นด้วยการประยุกต์ใช้เทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ในอดีต เลือกใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นเป็นตัวแบบในการประเมินความน่าสนใจของกฎความสัมพันธ์ทั้งหมด ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นเป็นตัวแบบที่ได้รับความนิยมมากแต่ในบางกรณีที่ประยุกต์ใช้กับข้อมูลบางประเภทก็สามารถทำให้เกิดผลลัพธ์ที่เป็นผลบวกลวงจำนวนมากได้ ปัญหาของค่าประเมินความน่าสนใจของกฎความสัมพันธ์ที่อาจมีผลกระทบมาถึงประสิทธิภาพของการนำไปประยุกต์ใช้รวบรวมเอาไว้ในหัวข้อ 1.2.2 ส่วนการประยุกต์ใช้เทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ของงานวิจัยในอดีตรวมถึงการประยุกต์ใช้กในระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์แสดงไว้ในหัวข้อ 1.2.3

1.2.1 ความสำคัญของระบบควบคุมการเปลี่ยนแปลงแก้ไขและปัญหาของนักพัฒนาในโครงการพัฒนาซอฟต์แวร์ขนาดใหญ่

เป็นที่ยอมรับกันว่าโลกธุรกิจทางด้านซอฟต์แวร์ในปัจจุบันมีการแข่งขันสูงมาก บริษัทที่จะอยู่ในตลาดได้จำเป็นอย่างยิ่งที่จะต้องมีเงินทุน บุคคลากรและกลยุทธ์ที่น่าสนใจ ฉะนั้นการสร้างความได้เปรียบทางการแข่งขันจำเป็นต้องมีกลยุทธ์ด้านกระบวนการที่มีประสิทธิภาพและมีมาตรฐานเป็นที่ยอมรับในระดับสากล เพื่อให้การผลิตซอฟต์แวร์มีคุณภาพ ตอบสนองความ

ต้องการและสร้างความพึงพอใจสูงสุดต่อลูกค้า มาตรฐานซีเอ็มเอ็มไอ (CMMI, Capability Maturity Model Integration) เป็นตัวแบบของการวัดระดับวุฒิภาวะ (Maturity) ความสามารถในการทำงานของบริษัท มาตรฐานซีเอ็มเอ็มไอที่ใช้ในปัจจุบันคือเวอร์ชั่น 2.1 ระดับวุฒิภาวะของมาตรฐานซีเอ็มเอ็มไอมีทั้งหมด 5 ระดับระดับวุฒิภาวะทั้งหมดประกอบด้วยกลุ่มกระบวน 22 กลุ่มกระบวนการ ในการบรรลุระดับวุฒิภาวะที่ 2 ของมาตรฐานซีเอ็มเอ็มไอ บริษัทจำเป็นต้องบรรลุเป้าหมายของกลุ่มกระบวนการทั้งหมด 7 กลุ่ม หนึ่งในนั้นคือกลุ่มกระบวนการการจัดการการตั้งค่าองค์ประกอบ (CM: Configuration Management) ซึ่งมีเป้าหมายเฉพาะเจาะจง (Specific Goal) ที่จำเป็นต้องบรรลุให้ได้ 3 เป้าหมาย และ 1 ใน 3 ของเป้าหมายเฉพาะเจาะจงนั้นคือ การติดตามและควบคุมการเปลี่ยนแปลงแก้ไข (Track and Control Changes) การบรรลุเป้าหมายข้อนี้นั้นจำเป็นต้องใช้เครื่องมือที่มีชื่อว่าระบบควบคุมการเปลี่ยนแปลงแก้ไข (Revision Control System, Version Control System) เข้ามาช่วย (Grune et al., 2006)

ระบบควบคุมการเปลี่ยนแปลงแก้ไข คือ ระบบที่ใช้ในจัดการการจัดเก็บ การค้นคืน การระบุและการผสานการเปลี่ยนแปลงแก้ไขแฟ้มข้อมูลซอร์สโค้ดของโปรแกรมประยุกต์ และสารสนเทศสำคัญอื่นๆที่พัฒนาขึ้นมาโดยทีมอย่างเป็นอัตโนมัติ ในซอฟต์แวร์ควบคุมการเปลี่ยนแปลงแก้ไขนั้นจะมีการบันทึกแฟ้มข้อมูลซอร์สโค้ดและแฟ้มข้อมูลบันทึก (Log Files) ที่บรรจุข้อมูลที่เกี่ยวข้องกับการเปลี่ยนแปลงแก้ไขอื่นๆ อาทิเช่น ซอร์สโค้ดส่วนใดที่ถูกแก้ไข นักพัฒนาแก้ไข วันเวลาบันทึกเวอร์ชั่นใหม่ของซอร์สโค้ด และหมายเหตุของการบันทึกเวอร์ชั่นใหม่เป็นต้น แฟ้มข้อมูลซอร์สโค้ดและแฟ้มข้อมูลบันทึกทั้งหมดจะถูกเรียกรวมกันว่า ซอฟต์แวร์อาร์ไคฟ์ (Software Archives) (Zimmermann et al., 2004; Zimmermann et al., 2005)

ระบบควบคุมการเปลี่ยนแปลงแก้ไขที่ได้รับความนิยมและถูกนำไปใช้อย่างแพร่หลายกว่า 2 ทศวรรษ คือ ระบบควบคุมการเปลี่ยนแปลงแก้ไขที่มีชื่อว่า ระบบคอนเคอเรนท์เวอร์ชั่น (Concurrent Versions System) (O'Sullivan et al., 2009) ระบบคอนเคอเรนท์เวอร์ชั่นถูกสร้างขึ้นมาให้บูรณาการรวมกับไอดีอี (IDE: Integrated Development Environment) ทำให้นักพัฒนาสามารถบรรลุเป้าหมายการติดตามและควบคุมการเปลี่ยนแปลงแก้ไขได้ ในขณะที่กำลังพัฒนาซอฟต์แวร์ในขั้นตอนการพัฒนาซอฟต์แวร์ (Development Phase) ของวงจรชีวิตการพัฒนาซอฟต์แวร์ (Software Development Life Cycle) ได้

ปัญหาที่มักจะเกิดขึ้นกับนักพัฒนาในโครงการพัฒนาซอฟต์แวร์ขนาดใหญ่ที่ต้องมีการทำงานเป็นทีมในแง่ของการควบคุมและติดตามการเปลี่ยนแปลงแก้ไขมีหลายประการ เช่น ปัญหา

การเกิดขึ้นของการเชื่อมโยงกัน (Evolution coupling) ระหว่างคลาสหรือระหว่างไฟล์ที่ไม่สามารถดักจับได้ในช่วงของการออกแบบ (Design phase) (Gall et al., 1998; Bieman et al., 2003; Burch et al., 2005) ปัญหาการทำความเข้าใจพัฒนาการของซอฟต์แวร์ (Software Evolution) (Ball et al., 1997) ที่อาจเกิดกับนักพัฒนาที่เข้าร่วมทีมใหม่ ปัญหาการเปลี่ยนแปลงแก้ไขซอฟต์แวร์ในช่วงการบำรุงรักษา (Maintenance phase) ที่อาจเกิดกับทีมบำรุงรักษาที่ไม่ใช่ทีมเดียวกับทีมพัฒนา ปัญหาการเรียกใช้ซอฟต์แวร์ไลบรารี่ที่ผิดและนำไปสู่การเกิดข้อผิดพลาด (Li et al., 2005; Livshits et al., 2005; Williams et al., 2005) นอกจากปัญหาต่างๆข้างต้นแล้ว ในระหว่างการพัฒนาซอฟต์แวร์นั้นอาจทำให้เกิดความต้องการบางอย่างเกิดขึ้นด้วย เช่น ความต้องการนำรูปแบบการเรียกใช้ซอฟต์แวร์ไลบรารี่ (Software Libraries) ที่ถูกต้องกลับมาใช้ใหม่ (Michail, 2000) ความต้องการระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ (Zimmermann et al., 2004; Zimmermann et al., 2005; Methanias et al., 2009) ปัญหาและความต้องการที่กล่าวมาข้างต้นนี้สามารถตอบสนองได้โดยการนำข้อมูลซอฟต์แวร์อาร์ไคฟร์ (Software Archives) ที่ได้มาจากระบบคอนเคอเรนท์เวอร์ชั่นมาวิเคราะห์และสร้างวิธีการในการแก้ปัญหาและตอบสนองความต้องการดังกล่าวได้ ซึ่งจะกล่าวถึงในหัวข้อที่ 1.2.3 ต่อไป

1.2.2 ความสำคัญและปัญหาของกฎความสัมพันธ์และค่าประเมินความน่าสนใจของกฎความสัมพันธ์

ทุกครั้งที่ผู้ใช้เข้าไปใช้บริการเลือกซื้อหนังสือหรือสินค้าต่างๆ ภายในเว็บไซต์อเมซอนดอทคอม (Amazon.com) ผู้ใช้จะสามารถมองเห็นส่วนหนึ่งของหน้าเว็บไซต์ปรากฏข้อความที่ว่า “ลูกค้าหลายๆคนที่ซื้อหนังสือเล่มนี้ (หรือสินค้าชิ้นนี้) มักจะซื้อหนังสือ (หรือสินค้า) …ด้วย” พร้อมกับแสดงรายการหนังสือ (หรือสินค้า) ที่มักจะถูกซื้อร่วมกันด้วย ข้อมูลสารสนเทศที่เว็บไซต์อเมซอนดอทคอมนำมาใช้เพื่อประโยชน์ในการเพิ่มยอดขายนี้เป็นข้อมูลสารสนเทศที่สร้างมาจากการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ (Association Rules Discovery) ทั้งสิ้น การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลการซื้อของลูกค้านั้นยังสามารถนำไปประยุกต์ใช้ในการออกแบบแค็ตตาล็อกสินค้า การขายสินค้าแขวน การออกแบบรายการส่งเสริมการขาย การจัดวางสินค้าภายในร้าน การแบ่งกลุ่มลูกค้าตามรูปแบบของพฤติกรรมการซื้อสินค้า เป็นต้น (Agrawal et al., 1994) นอกจากนั้นแล้วในตลอดช่วงทศวรรษที่ผ่านมาการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์นี้ยังมีบทบาทที่สำคัญในการค้นหารูปแบบความสัมพันธ์ที่มีคุณค่าในข้อมูลประเภทอื่นๆอีก เช่น ข้อมูลเครือข่ายโทรคมนาคม

ข้อมูลการจัดการความเสี่ยง ข้อมูลการควบคุมคลังสินค้า และข้อมูลทางพันธุกรรมของสิ่งมีชีวิต เป็นต้น (Kotsiantis et al., 2006)

แนวคิดของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์นี้ถูกนำเสนอขึ้นมาครั้งแรกในปีค.ศ. 1993 โดย Agrawal และคณะ (Agrawal et al., 1993) ต่อมาในปีค.ศ. 1994 Agrawal และคณะ (Agrawal et al., 1994) ได้นำเสนอขั้นตอนวิธีในการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ใหม่ขึ้นมาชื่อว่าขั้นตอนวิธีอพริโอริ (Apriori Algorithm) นอกจากนั้น Agrawal และคณะ (Agrawal et al., 1993) ยังได้นำเสนอตัวแบบของการประเมินระดับความตรงประเด็นหรือระดับความน่าสนใจของกฎความสัมพันธ์ตัวแบบแรกและดั้งเดิมคือตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น (Support-Confidence Model) ซึ่งใช้ค่า 2 ค่าในการประเมิน คือ ค่าสนับสนุน (Support) และค่าความเชื่อมั่น (Confidence) ของกฎความสัมพันธ์ ในตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นนี้ ค่าสนับสนุนถูกใช้ในการคัดกรองรายการที่มีความถี่สูงออกมา และค่าความเชื่อมั่นจะถูกใช้เป็นค่าที่วัดระดับความน่าสนใจของกฎความสัมพันธ์ หลังจากนั้นต่อมามีงานวิจัยหลายงานวิจัยออกมาเสนอค่าประเมินค่าอื่นๆ ที่ใช้ในการประเมินความน่าสนใจของกฎความสัมพันธ์แทนการใช้ค่าความเชื่อมั่น ผู้วิจัยรวบรวมงานวิจัยที่เสนอค่าประเมินความน่าสนใจของกฎความสัมพันธ์ที่ได้รับความนิยมและถูกนำไปประยุกต์กับต่างๆ อย่างละเอียดไว้ในบทที่ 2 ซึ่งสามารถสรุปได้ดังตารางต่อไปนี้

กำหนดให้ $P(X)$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

$P(\overline{X})$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่ไม่มีรายการ X ในฐานข้อมูล

$P(X \text{ and } Y)$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และ Y ในฐานข้อมูล

$P(\overline{X} \text{ and } \overline{Y})$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่ไม่มีทั้งรายการ X และ Y ในฐานข้อมูล

$P(X \text{ and } \overline{Y})$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และไม่รายการ Y ในฐานข้อมูล

ตารางที่ 1-1 แสดงตารางสรุปค่าประเมินความน่าสนใจของกฎความสัมพันธ์

ชื่อค่าประเมินฯ	สมการคำนวณ	อ้างอิง
ค่าสนับสนุน (Support)	Support($X \rightarrow Y$) = P(X and Y)	(Agrawal et al., 1993)
ค่าความเชื่อมั่น (Confidence)	Conf($X \rightarrow Y$) = $\frac{\mathrm{P(X\ and\ Y)}}{\mathrm{P(X)}}$	(Agrawal et al., 1993)
ค่าคอนวิคชั่น (Conviction)	Conviction($X \rightarrow Y$) = $\frac{\mathrm{P(X)P(\overline{Y})}}{\mathrm{P(X\ and\ \overline{Y})}}$	(Brin et al., 1997)
ค่าลิฟท์ (Lift)	Lift($X \rightarrow Y$) = $\frac{\mathrm{P(X\ and\ Y)}}{\mathrm{P(X)P(Y)}}$	(Brin et al., 1997)
ค่าเลฟเวอเรจ (Leverage)	Leverage($X \rightarrow Y$) = P(X and Y) - P(X)P(Y)	(Piatetsky-Shapiro et al., 1991)
ค่าคัฟเวอเรจ (Coverage)	Coverage($X \rightarrow Y$) = P(X)	(Michael., 2009)
ค่าสหสัมพันธ์ (Correlation)	Corr($X \rightarrow Y$) = $\frac{\mathrm{(P(X\ and\ Y) - P(X)P(Y))}}{\sqrt{P(X)P(Y)P(1-P(X))(1-P(Y))}}$	(Sheikh et al., 2004)
ค่าอัตราส่วนออดส์ (Odds Ratio)	Odds($X \rightarrow Y$) = $\frac{\mathrm{(P(X\ and\ Y)\ P(\overline{X}\ and\ \overline{Y}))}}{\mathrm{(P(X\ and\ \overline{Y})\ P(\overline{X}\ and\ Y))}}$	(Sheikh et al., 2004)

ต่อมาปีค.ศ. 2008 Liu และคณะ (Liu et al., 2008) ได้นำเสนอข้อบกพร่องประการหนึ่งของการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นและการใช้ค่าความเชื่อมั่นเป็นค่าประเมินความน่าสนใจของกฎความสัมพันธ์ โดยการยกตัวอย่างฐานข้อมูลการซื้อสินค้าของลูกค้าในกรณีที่ทำให้การใช้ค่าความเชื่อมั่นเป็นค่าประเมินความน่าสนใจของกฎความสัมพันธ์มีผลลัพธ์ที่ออกมาเป็นกฎความสัมพันธ์ที่มีเซตรายการที่มาก่อนมีความสัมพันธ์เชิงลบกับเซตรายการที่ตามมา กล่าวคือทรานแซคชั่นส่วนใหญ่ถ้ามีเซตรายการที่มาก่อนมักจะไม่มีเซตรายการที่ตามมาของกฎนั้นนั่นเอง หรือก็คือได้กฎความสัมพันธ์ที่เป็นผลบวกลวง (False Positive) นั่นเอง ด้วยสาเหตุนี้ Liu และคณะ (Liu et al., 2008) จึงได้นำเสนอค่าประเมินความน่าสนใจของกฎความสัมพันธ์ใหม่ขึ้นมาและให้ชื่อว่าค่าความเชื่อมั่นใหม่ (New Confidence) พร้อมกับพิสูจน์ว่าค่าความเชื่อมั่นใหม่นี้ไม่ขัดแย้งกับค่าสหสัมพันธ์และค่าความเชื่อมั่นเดิมซึ่งเป็นค่าสถิติ นอกจากนั้นยังได้แสดงตัวอย่างของฐานข้อมูลทรานแซคชั่นสมมุติชุดหนึ่งขึ้นมาเพื่อพิสูจน์ว่าค่าความเชื่อมั่นใหม่สามารถลดการเกิดกฎความสัมพันธ์ที่เป็นผลบวกลวงด้วย ค่าความเชื่อมั่นใหม่สามารถคำนวณได้จากสูตรดังต่อไปนี้

กำหนดให้ P(X) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

P($\overline{X}$) คือ ค่าความน่าจะเป็นในการไม่พบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

P(X and Y) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และรายการ Y

ในฐานข้อมูล

$P(X \text{ and } \bar{Y})$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และไม่รายการ Y ในฐานข้อมูล

$$NConf(X \rightarrow Y) = \frac{P(X \text{ and } Y)}{P(Y)} - \frac{P(X \text{ and } \bar{Y})}{P(\bar{Y})}$$

ในงานวิจัยของ Liu และคณะ (Liu et al., 2008) ที่ได้เสนอค่าความเชื่อมั่นใหม่ข้างต้นนั้น Liu และคณะได้ทำการเปรียบเทียบความสามารถของค่าความเชื่อมั่นใหม่กับค่าประเมินความน่าสนใจของกฎความสัมพันธ์อื่นๆ ทั้งหมด 8 ค่าคือ ค่าสนับสนุน (Support), ค่าความเชื่อมั่น (Confidence), ค่าคอนวิคชั่น (Conviction), ค่าลิฟท์ (Lift), ค่าเลฟเวอเรจ (Leverage), ค่าคัฟเวอเรจ (Coverage), ค่าสหสัมพันธ์ (Correlation) และ ค่าอัตราส่วนออดส์ (Odds Ratio) โดยใช้ฐานข้อมูลทรานแซคชั่นสมมุติขนาด 10 ทรานแซคชั่น ผลของการเปรียบเทียบคือ ค่าความเชื่อมั่นใหม่สามารถบ่งบอกทิศของความสัมพันธ์ได้อย่างถูกต้องและสอดคล้องกับค่าเลฟเวอเรจและค่าสหสัมพันธ์ แต่ค่าความเชื่อมั่นใหม่นั้นสามารถระบุความแตกต่างของความน่าสนใจของกฎความสัมพันธ์ 2 กฎความสัมพันธ์ใดๆที่ค่าเลฟเวอเรจและค่าสหสัมพันธ์ไม่สามารถระบุได้ (กล่าวคือกฎความสัมพันธ์ 2 กฎที่คำนวณค่าค่าเลฟเวอเรจหรือค่าสหสัมพันธ์ได้เท่ากันทั้ง 2 กฎ แต่ค่าความเชื่อมั่นใหม่ให้ค่าที่แตกต่างกันระหว่าง 2 กฎ) ส่วนค่าประเมินความน่าสนใจของกฎความสัมพันธ์อื่นๆให้ค่าที่ขัดแย้งกับค่าสหพันธ์ (กล่าวคือเซตรายการที่มาก่อนและเซตรายการที่ตามมาของกฎความสัมพันธ์นั้นมีความสัมพันธ์เชิงลบต่อกันแต่กลับให้ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ที่สูงออกมา)

จากงานวิจัยในอดีตที่ทำการเสนอค่าประเมินความน่าสนใจของกฎความสัมพันธ์ต่างๆที่กล่าวไปข้างต้น แต่ละค่านั้นก็แสดงคุณสมบัติเฉพาะตัวที่แตกต่างกัน ผลลัพธ์ของกฎความสัมพันธ์ที่ได้ออกมาก็แตกต่างกันออกไป ผู้วิจัยจึงสนใจที่จะทำการเปรียบเทียบความสามารถของแต่ละค่าประเมินความน่าสนใจของกฎความสัมพันธ์ ผู้วิจัยจึงทบทวนและรวบรวมงานวิจัยที่เสนอคุณสมบัติต่างๆที่ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ควรจะมี และทำการเปรียบเทียบความสามารถของค่าประเมินความน่าสนใจของกฎความสัมพันธ์เหล่านั้นด้วยคุณสมบัติที่ควรจะมีทั้งหมด คุณสมบัติของค่าประเมินความน่าสนใจของกฎความสัมพันธ์ทั้งหมด 16 คุณสมบัติอธิบายอย่างละเอียดไว้ในบทที่ 2 หัวข้อ 2.5 และสามารถสรุปได้ดังนี้

- คุณสมบัติ 3 ข้อของ Piatetsky-Shapiro และคณะ (Piatetsky-Shapiro, 1991)
- คุณสมบัติ 1 ข้อของ Major และ Mangano (Major and Mangano, 1995) เพิ่มเติมจากของ Piatetsky-Shapiro และคณะ
- คุณสมบัติ 5 ข้อของ Tan และคณะ (Tan et al, 2002)
- คุณสมบัติ 5 ข้อของ Lenca และคณะ (Lenca et al, 2004)
- คุณสมบัติ 2 ข้อของ Geng และ Hamilton (Geng and Hamilton, 2006)

คุณสมบัติที่ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ควรจะมีทั้งหมด 16 ข้อข้างต้น คุณสมบัติที่ได้รับการยอมรับและถูกอ้างอิงถึงโดยงานวิจัยต่างๆ (Freitas, 1999; Major and Mangano, 1995; Mcgarry, 2005; Geng and Hamilton, 2006; Liu et al., 2008; Heravi, 2009) มากที่สุดคือคุณสมบัติ P1 P2 และ P3 ของ Piatetsky-Shapiro และคณะ (Piatetsky-Shapiro, 1991) และคุณสมบัติ P4 ของ Major และ Mangano (Major and Mangano, 1995)

เนื่องจากงานวิจัยของ Liu และคณะในปี 2008 (Liu et al., 2008) ได้ทำการพิสูจน์คุณสมบัติของค่าความเชื่อมั่นใหม่ไว้ทั้งหมดเพียง 5 คุณสมบัติคือ คุณสมบัติ P1 P2 P3 O1 และ O2 เท่านั้น ดังนั้นผู้วิจัยจึงพิสูจน์คุณสมบัติ P4 O3 O4 O5 Q1 Q2 Q3 S1 และ S2 ของค่าความเชื่อมั่นใหม่อย่างละเอียดและแสดงไว้ในบทที่ 2 หัวข้อ 2.5 ผลของการพิสูจน์คุณสมบัติของค่าความเชื่อมั่นใหม่แสดงไว้ในตารางที่ 2-4 ข้างต้น

การเปรียบเทียบในตารางที่ 2-4 แสดงให้เห็นว่าค่าความเชื่อมั่นใหม่มีคุณสมบัติที่ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ควรจะมี 10 คุณสมบัติจากทั้งหมด 14 คุณสมบัติซึ่งมากกว่าค่าประเมินความน่าสนใจของกฎความสัมพันธ์อื่นๆ โดยเฉพาะอย่างยิ่งการมีคุณสมบัติ O3 ของค่าความเชื่อมั่นใหม่จะทำให้การนำค่าความเชื่อมั่นใหม่ไปใช้นั้นจะสามารถขจัดการเกิดกฎความสัมพันธ์ที่มีเซตรายการที่มาก่อนและเซตรายการที่ตามที่มีความสัมพันธ์เชิงลบออกไปได้ ผู้วิจัยจึงเชื่อว่าถ้านำค่าความเชื่อมั่นใหม่นี้ไปประยุกต์ใช้กับการค้นหากฎความสัมพันธ์กับข้อมูลประเภทต่างๆรวมถึงข้อมูลซอฟต์แวร์อาร์ไควฟ์ แล้วน่าจะทำให้กฎความสัมพันธ์ที่ได้มาเป็นกฎความสัมพันธ์ที่น่าสนใจและช่วยลดการเกิดกฎความสัมพันธ์ที่เป็นผลบวกลวง (False Positive) ได้

1.2.3 ความสำคัญและปัญหาของการประยุกต์ใช้เทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์

ในช่วงต้นของการคิดค้นและพัฒนาแนวคิดการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์นั้น การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ถูกพัฒนามาเพื่อการค้นหารูปแบบความสัมพันธ์ของพฤติกรรมการซื้อสินค้าของลูกค้าจากฐานข้อมูลรายการซื้อสินค้าขนาดใหญ่ ต่อจากนั้นมาไม่นานเริ่มมีนักวิจัยหลายคณะนำการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์มาประยุกต์ใช้กับข้อมูลประเภทต่างๆมากมาย หนึ่งในนั้นคือข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Software Archive) ที่ได้จากระบบคอนเคอเรนท์เวอร์ชั่นในขั้นตอนการพัฒนาซอฟต์แวร์ (Development Phase) ในวงจรชีวิตการพัฒนาซอฟต์แวร์ (Software Development Life Cycle) คณะนักวิจัยเหล่านั้นนำข้อมูลซอฟต์แวร์อาร์ไคฟว์มาวิเคราะห์ในรูปแบบต่างๆกันแบ่งตามการนำไปใช้ ดังต่อไปนี้

1) การวิเคราะห์เพื่อความเข้าใจพัฒนาการของการพัฒนาซอฟต์แวร์ (Ball et al., 1997)
2) การวิเคราะห์เพื่อดักจับพัฒนาการของการเชื่อมโยงกัน (Gall et al.,1998; Bieman et al., 2003; Zimmermann et al., 2004)
3) การวิเคราะห์เพื่อเปิดเผยรูปแบบการเรียกใช้งานซอฟต์แวร์ไลบรารี่ (Michail, 1999; Michail, 2000; Li et al., 2005; Livshits et al., 2005; Williams et al., 2005)
4) การวิเคราะห์เพื่อสร้างคำแนะนำในการเปลี่ยนแปลงแก้ไข (Zimmermann et al., 2005; Methanias et al., 2009)

รายละเอียดของแต่ละงานวิจัยแสดงในหัวข้อ 2.7 จากงานวิจัยที่นำข้อมูลซอฟต์แวร์อาร์ไคฟว์มาวิเคราะห์ทั้งหมด มีคณะวิจัยของ Li และคณะวิจัยของ Livshits (Li et al., 2005; Livshits et al., 2005) ได้แสดงให้เห็นว่าการประยุกต์เทคนิคการทำเหมืองข้อมูลด้วยกฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์เป็นเทคนิคที่ค่อนข้างมีประสิทธิภาพแต่ในบางกรณีก็สามารถทำให้เกิดผลลัพธ์ของการค้นหาที่เป็นผลบวกลวง (False Positive) เป็นจำนวนมากได้

ในปีค.ศ. 2005 Zimmermann และคณะ (Zimmermann et al., 2005) ได้เสนอการประยุกต์ใช้การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์อีกรูปแบบหนึ่ง คือการสร้างคำแนะนำในการเปลี่ยนแปลงแก้ไขให้กับนักพัฒนาในระหว่างขั้นตอนการพัฒนาซอฟต์แวร์ ในงานวิจัยนี้ Zimmermann และคณะเสนอว่าระบบสร้าง

คำแนะนำนักพัฒนานั้นควรจะสามารถให้คำแนะนำแก่นักพัฒนาทั้งหมด 3 สถานการณ์คือ 1) สถานการณ์การนำทาง (Navigation) คือ สถานการณ์ที่นักพัฒนามีการเปลี่ยนแปลงแก้ไขที่เอนทิตี้หนึ่งแล้ว ระบบจะให้คำแนะนำกับนักพัฒนาให้แก้ไขเอนทิตี้ใดต่อไป 2) สถานการณ์การป้องกันการเกิดข้อผิดพลาด (Error Prevention) คือ สถานการณ์ที่นักพัฒนามีการเปลี่ยนแปลงแก้ไขที่เอนทิตี้หลายๆเอนทิตี้ต่อเนื่องกันแต่ยังขาดการเปลี่ยนแปลงแก้ไขเอนทิตี้อีกหนึ่งเอนทิตี้จึงจะสมบูรณ์ ระบบจะให้คำแนะนำกับนักพัฒนาให้แก้ไขเอนทิตี้ที่เหลือนั่น และ 3) สถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (Closure) คือ สถานการณ์ที่นักพัฒนามีการเปลี่ยนแปลงแก้ไขที่เอนทิตี้หลายๆเอนทิตี้ต่อเนื่องกันจนสมบูรณ์แล้ว ระบบจะไม่ควรให้คำแนะนำที่เป็นผลบวกลวง (False Positive) ออกมาแก่นักพัฒนา งานวิจัยของ Zimmermann และคณะ (Zimmermann et al., 2005) ได้ทำการทดสอบประสิทธิภาพของการให้คำแนะนำในรูปแบบต่างๆหลายรูปแบบกับข้อมูลซอฟต์แวร์อาร์ไคฟ์วของโครงการพัฒนาซอฟต์แวร์แบบโอเพนซอร์ส (Open Source) และข้อสรุปของการทดสอบได้แนะนำสิ่งที่เป็นประโยชน์ต่อการต่อยอดการประยุกต์ใช้การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ได้เป็นอย่างดี เช่น การให้คำแนะนำในระดับเอนทิตี้ละเอียด (ตัวแปร เมธอดหรือฟังก์ชั่น) ให้ประสิทธิภาพไม่ต่างกับการให้คำแนะนำในแฟ้มข้อมูลหรือคลาส การให้คำแนะนำนักพัฒนามีประสิทธิภาพมากสำหรับการเปลี่ยนแปลงแก้ไขในช่วงการบำรุงรักษา (Maintenance Phase) (เน้นที่การแก้ไข (alter) มากกว่าการเพิ่ม (add to) กับการลบ (delete from)) เป็นต้น นอกจากนั้นผู้วิจัยสังเกตเห็นว่าผลการทดสอบประสิทธิภาพในงานวิจัยของ Zimmermann และคณะ (Zimmermann et al., 2005) ยังให้ประสิทธิภาพไม่สูงเท่าที่ควร

จากความสำคัญและปัญหาที่กล่าวไปทั้งหมดในหัวข้อ 1.2.1 ถึง 1.2.3 ข้างต้นผู้วิจัยเห็นว่าการเพิ่มประสิทธิภาพให้กับระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ของไอดีอีนั้นควรเพิ่มประสิทธิภาพโดยการปรับปรุงขั้นตอนวิธีในการค้นหากฎความสัมพันธ์ให้ดีขึ้นและลดจำนวนของการเกิดผลบวกลวงลง ผู้วิจัยจึงคาดว่าการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008) นั้นน่าจะมีประสิทธิภาพที่ดีกว่าการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความ

เชื่อมั่นเดิมจากการแสดงคุณสมบัติที่ค่าประเมินความน่าสนใจควรจะมีมากที่สุดและโดยเฉพาะอย่างยิ่งการแสดงคุณสมบัติ O3

โดยทั่วไปแล้วระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ของไอดีอีจะทำหน้าที่หลัก 2 ประการคือ 1) การชี้นำนักพัฒนาว่าควรเปลี่ยนแปลงแก้ไขส่วนใดต่อเมื่อมีการเปลี่ยนแปลงแก้ไขส่วนนี้แล้ว และ 2) การแจ้งเตือนนักพัฒนาก่อนการบันทึกว่ายังทำการแก้ไขเปลี่ยนแปลงไม่สมบูรณ์เพื่อป้องกันการเกิดข้อผิดพลาด (error) นอกจากหน้าที่ 2 ประการนี้แล้วสิ่งที่สำคัญอีกประการหนึ่งคือระบบให้คำแนะนำต้องไม่มีการให้คำแนะนำใดๆออกมาถ้าการเปลี่ยนแปลงแก้ไขทั้งหมดสมบูรณ์ดีแล้วด้วย (Zimmermann et al., 2005) ดังนั้นในการทดสอบประสิทธิภาพของระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ของไอดีอีควรทดสอบสถานการณ์ของการให้คำแนะนำนักพัฒนาต่างๆกัน 3 สถานการณ์ได้แก่ 1) สถานการณ์การนำทาง (Navigation) 2) สถานการณ์การป้องกันการเกิดข้อผิดพลาด (Error Prevention) 3) สถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (Closure)

ดังนั้นผู้วิจัยจึงต้องการศึกษาเปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น (Support-Confidence Model) ดั้งเดิมกับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008) ในสถานการณ์ของการให้คำแนะนำนักพัฒนาต่างๆกัน 3 สถานการณ์

1.3 วัตถุประสงค์ของงานวิจัย

1. เปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น (Support-Confidence Model) ดั้งเดิมกับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008) โดยที่ประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับ

ข้อมูลซอฟต์แวร์อาร์ไคฟ์นั้นสามารถแบ่งออกได้เป็นประสิทธิภาพใน 3 สถานการณ์ของการพัฒนาซอฟต์แวร์ดังนี้

- สถานการณ์การนำทาง (Navigation)
- สถานการณ์การป้องกันการเกิดข้อผิดพลาด (Error Prevention)
- สถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (Closure)

1.4 ขั้นตอนโดยสรุปของการทำวิจัย

1. ศึกษารายละเอียดเกี่ยวกับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ที่มีอยู่ในปัจจุบัน
2. ศึกษารายละเอียดเกี่ยวกับการประเมินระดับความน่าสนใจของกฎความสัมพันธ์ (Interestingness Measure of Association Rules) ในการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008) และตัวแบบต่างๆที่มีอยู่ในปัจจุบัน
3. ออกแบบเครื่องมือทดสอบต่างๆตามที่ได้ศึกษา
4. พัฒนาเครื่องมือทดสอบตามที่ได้ออกแบบไว้
5. ทดสอบการทำงานของเครื่องมือที่พัฒนา
6. ประเมินการทำงานของเครื่องมือ
7. วิเคราะห์ผลการทดลองและสำรวจข้อมูลเพิ่มจากผลการทดลอง
8. จัดทำเอกสารสรุปงานวิจัย และข้อเสนอแนะ

1.5 ตัวแปรที่ศึกษา

1. **ตัวแปรอิสระ** (Independent variables)

งานวิจัยนี้สนใจว่าการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008) สามารถเพิ่มประสิทธิภาพของระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์

ของไอดีอี (IDE: Integrated Development Environment) ได้หรือไม่ ดังนั้นตัวแปรต้นของการศึกษานี้ก็คือตัวแบบของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ ซึ่งงานวิจัยนี้จะศึกษาเปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ทั้งหมด 2 ตัวแบบ ดังนี้

1) การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น (Support-Confidence Model)
2) การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Support-New Confidence Model) (Liu et al., 2008)

โดยในงานวิจัยนี้จะเปรียบเทียบประสิทธิภาพของทั้ง 2 ตัวแบบข้างต้นในสถานการณ์ที่ต่างกัน 3 สถานการณ์ คือ สถานการณ์การนำทาง (Navigation) สถานการณ์การป้องกันการเกิดข้อผิดพลาด (Error Prevention) และสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (Closure) เช่นเดียวกับงานวิจัยของ Zimmermann และคณะในปี ค.ศ. 2005 (Zimmermann et al., 2005) และงานวิจัยของ Methanias และคณะในปี ค.ศ. 2009 (Methanias et al., 2009)

2. **ตัวแปรตาม** (Dependent variables)

ตัวแปรตามของงานวิจัยนี้ คือ ประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบต่างๆ การเปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์จะพิจารณาจากคำแนะนำสำหรับนักพัฒนาที่ถูกสร้างมาจากกฎความสัมพันธ์ที่ได้มานั้นมีความถูกต้องแม่นยำในการทำนายและให้คำแนะนำในระหว่างการพัฒนาซอฟต์แวร์มากน้อยเพียงใด โดยการวัดประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ในแต่ละสถานการณ์มีวิธีการในการประเมินที่แตกต่างกันออกไปดังนี้

- ในสถานการณ์ *การนำทาง (Navigation)* สามารถประเมินประสิทธิภาพจากค่าเอฟเมสเซอร์ (F-measure) ที่คำนวณมาจากค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) (Methanias et al., 2009) รายละเอียดและวิธีการคำนวณค่าเอฟเมสเซอร์ ค่าความถูกต้องและค่าเรียกคืนสำหรับการทำเหมืองข้อมูลด้วย

เทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์นั้นได้กล่าวเอาไว้ในบทที่ 2

- ในสถานการณ์ *การป้องกันการเกิดข้อผิดพลาด (Error Prevention)* สามารถประเมินประสิทธิภาพจากค่าเอฟเมสเซอร์ (F-measure) ที่คำนวณมาจากค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) (Methanias et al., 2009) รายละเอียดและวิธีการคำนวณค่าเอฟเมสเซอร์ ค่าความถูกต้องและค่าเรียกคืนสำหรับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์นั้นได้กล่าวเอาไว้ในบทที่ 2
- ในสถานการณ์ *การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (Closure)* สามารถประเมินประสิทธิภาพจากค่าผลสะท้อนกลับ (Feedback) (Zimmermann et al., 2005; Methanias et al., 2009) รายละเอียดและวิธีการคำนวณค่าผลสะท้อนกลับสำหรับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์นั้นได้กล่าวเอาไว้ในบทที่ 2

3. **ตัวแปรควบคุม**

ตัวแปรควบคุมกับการเปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ มีทั้งหมด 2 ตัวแปร ได้แก่

1) ข้อสอบถาม สำหรับงานวิจัยนี้ คือ เซตที่ประกอบไปด้วยเซตเหตุการณ์การเปลี่ยนแปลงแก้ไขและเซตผลลัพธ์ที่คาดไว้ โดยจะมีข้อสอบถามทั้งหมด 3 แบบสำหรับ 3 สถานการณ์ที่แตกต่างกันคือสถานการณ์การนำทาง สถานการณ์การป้องกันข้อผิดพลาด และสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (Zimmermann et al., 2005; Methanias et al., 2009)
2) เครื่องมือที่ใช้ในงานวิจัย ประกอบด้วยเครื่องมือทั้งหมด 5 เครื่องมือได้แก่ เครื่องมือจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ เครื่องมือสร้างข้อสอบถามสำหรับการทดสอบ 3 สถานการณ์ เครื่องมือการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบสำหรับ 3 สถานการณ์ เครื่องมือสร้างเซตของคำแนะนำสำหรับเหตุการณ์ และเครื่องมือประเมินผลการทดสอบ

1.6 ขอบเขตของการวิจัย

1. ข้อมูลซอฟต์แวร์อาร์ไคฟว์ที่นำมาใช้ในการศึกษาเป็นข้อมูลซอฟต์แวร์อาร์ไคฟว์ของโครงการพัฒนาซอฟต์แวร์ที่พัฒนาด้วยภาษาซีพลัสพลัส (C++) เท่านั้น
2. ข้อมูลซอฟต์แวร์อาร์ไคฟว์ที่นำมาใช้ในการศึกษาเป็นข้อมูลซอฟต์แวร์อาร์ไคฟว์ที่มาจากซอฟต์แวร์ควบคุมการแก้ไขปรับปรุง (Revision Control) ที่ชื่อว่าระบบคอนเคอเรนท์เวอร์ชั่น (Concurrent Version System) เท่านั้น
3. คำแนะนำสำหรับนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ในงานวิจัยนี้ หมายถึง คำแนะนำที่ได้มาจากการค้นหารูปแบบของการเปลี่ยนแปลงแก้ไขในการพัฒนาซอฟต์แวร์ที่เกิดขึ้นบ่อยในอดีตเท่านั้น ไม่รวมถึงรูปแบบของการเปลี่ยนแปลงแก้ไขที่เกิดขึ้นไม่บ่อยแต่มีความสำคัญมาก
4. คำแนะนำสำหรับนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ในงานวิจัยนี้ สนใจเพียงการเปลี่ยนแปลงแก้ไขที่ควรจะอยู่ในทรานแซคชั่นเดียวกันเท่านั้น ไม่สนใจลำดับของการเปลี่ยนแปลงแก้ไข
5. ทรานแซคชั่นของการเปลี่ยนแปลงแก้ไขที่นำมาทดสอบในงานวิจัยนี้ ไม่รวมทรานแซคชั่นที่ถูกระบุว่าเป็นสิ่งแปลกปลอม 2 ประเภทคือ ทรานแซคชั่นขนาดใหญ่และทรานแซคชั่นการผสานกิ่ง
6. ทรานแซคชั่นของการเปลี่ยนแปลงแก้ไขที่นำมาทดสอบในงานวิจัยนี้ คือ ทรานแซคชั่นของการเปลี่ยนแปลงแก้ไขในระดับของแฟ้มข้อมูลและคลาสเท่านั้น ไม่ได้พิจารณาถึงการเปลี่ยนแปลงแก้ไขในระดับเอนทิตี้ที่ละเอียดเช่น เมธอดหรือตัวแปร
7. การทดสอบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ 3 สถานการณ์แยกจากกัน แต่ใช้ทรานแซคชั่นชุดทดสอบชุดเดียวกันทั้ง 3 สถานการณ์
8. การทดสอบของงานวิจัยนี้เลือกตัวอย่างข้อมูลซอฟต์แวร์อาร์ไคฟว์เพียงตัวอย่างเดียว เนื่องจากงานวิจัยนี้ต้องการวิจัยเพื่อหาข้อมูลเบื้องต้น (Exploratory Research) เท่านั้น

1.7 ประโยชน์ที่คาดว่าจะได้รับ

1. ผู้ที่ต้องการพัฒนาระบบให้คำแนะนำสำหรับนักพัฒนาระหว่างการพัฒนาซอฟต์แวร์บนไอดีอี สามารถนำงานวิจัยนี้ไปเป็นแนวทางในการประยุกต์ใช้เข้ากับไอดีอีได้
2. ผู้ที่ต้องการพัฒนาระบบติดตามพัฒนาการการเกิดความเชื่อมโยงกัน (Evolution Coupling) สามารถนำงานวิจัยนี้ไปเป็นแนวทางในการประยุกต์ใช้ได้
3. ผู้ที่ต้องการพัฒนาระบบค้นหารูปแบบการเรียกใช้งานไลบรารี่ (Software Library Call Pattern) สามารถนำงานวิจัยนี้ไปเป็นแนวทางในการประยุกต์ใช้ได้
4. ผลการทดสอบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์วในงานวิจัยนี้ ทำให้ทราบตัวแบบที่เหมาะสำหรับการค้นหากฎความสัมพันธ์ในสถานการณ์การนำทาง (Navigation) สถานการณ์การป้องกันการเกิดข้อผิดพลาด (Error Prevention) และสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (Closure) ในการพัฒนาซอฟต์แวร์ได้
5. ผลการทดสอบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์วในงานวิจัยนี้เป็นประโยชน์สำหรับนักวิจัยที่ต้องการต่อยอดศึกษาการประยุกต์ใช้กฎความสัมพันธ์กับระบบให้คำแนะนำสำหรับนักพัฒนาระหว่างการพัฒนาซอฟต์แวร์ต่อไป

1.8 นิยามศัพท์

1. ซอฟต์แวร์อาร์ไคฟ์ว (Software Archives) คือ แฟ้มข้อมูลซอร์สโค้ด (Source Code Files) ทุกเวอร์ชั่น และแฟ้มข้อมูลบันทึกที่ได้จากระบบควบคุมการพัฒนาเวอร์ชั่น (CVS Log File)
2. ซอร์สโค้ด (Source Code) คือ รหัสคอมพิวเตอร์ซึ่งได้รับการเปลี่ยนเป็นภาษาทางเครื่องคอมพิวเตอร์ก่อนทำงานบนเครื่องคอมพิวเตอร์

3. เอนทิตี้ (Entity) คือ เอกลักษณ์หรือสิ่งที่ผู้วิจัยสนใจศึกษา ในที่นี้คำว่า เอนทิตี้ สามารถหมายถึง แฟ้มข้อมูลเอกสาร (File) คลาส (Class) เมธอดหรือฟังก์ชัน (Method or Function) และตัวแปร (Variable)
4. การเปลี่ยนแปลงแก้ไข (Changes) คือ การที่มีนักพัฒนาแก้ไขเอนทิตี้ใดๆ คำว่าการเปลี่ยนแปลงแก้ไขในที่นี้สามารถแสดงได้ 3 มิติ คือ 1) การแก้ไขเอนทิตี้ (alter) 2) การเพิ่มลงในเอนทิตี้ (add to) และ 3) การลบออกจากเอนทิตี้ (delete from)
5. ทรานแซคชั่น (Transaction) สำหรับงานวิจัยนี้ คือ เซตของการเปลี่ยนแปลงแก้ไขที่เกิดขึ้นพร้อมกันหรือในเวลาใกล้เคียงกันและถูกบันทึกเข้าสู่ระบบคอนเคอเรนท์เวอร์ชั่นโดยนักพัฒนาคนเดียวกัน
6. เหตุการณ์ (Situation) คือเซตของการเปลี่ยนแปลงแก้ไขใดๆ ที่เกิดขึ้นจากนักพัฒนา
7. กฎความสัมพันธ์ (Association Rules) สามารถนิยามได้ดังนี้ กำหนดให้ เซต $I = \{i_1, i_2, \ldots, i_m\}$ เป็นเซตของรายการข้อมูล (items) ที่มีอยู่ทั้งหมดและให้ เซต $T = \{t_1, t_2, \ldots, t_n\}$ เป็นเซตของทรานแซคชั่น โดยที่แต่ละทรานแซคชั่น t_n ประกอบด้วยเซตย่อย I_j (j = 1, 2, .. m) ที่เป็นเซตย่อยของเซตของรายการข้อมูล I เซตของรายการข้อมูล I_j นั้นถูกเรียกว่า เซตรายการ (Itemset) ดังนั้นกฎความสัมพันธ์ r ก็คือคู่ของเซตรายการ I_1 และเซตรายการ I_2 โดยที่เซตรายการ I_1 และเซตรายการ I_2 เป็นเซตย่อยของเซต I ที่ไม่มีสมาชิกที่ซ้อนทับกันและเซตรายการ I_2 ไม่เท่ากับเซตว่าง เซตรายการ I_1 ถูกเรียกว่า เซตรายการที่มาก่อน (Antecedent Itemset) และเซตรายการ I_2 ถูกเรียกว่า เซตรายการที่ตามมา (Consequent Itemset) และกำหนดสัญลักษณ์ $I_1 \rightarrow I_2$ แทนกฎความสัมพันธ์ R ที่มีเซตรายการ I_1 เป็นเซตรายการที่มาก่อน และเซตรายการ I_2 เป็นเซตรายการที่ตามมา โดยที่ I_2 ไม่ใช่เซตว่าง (Olivier et al., 2008) สำหรับงานวิจัยนี้ กฎความสัมพันธ์ คือ กฎความสัมพันธ์ที่ตอบข้อถามที่ว่า ถ้านักพัฒนาเปลี่ยนแปลงแก้ไข (เปลี่ยนแปลง เพิ่มลง หรือ ลบออก) เอนทิตี้ใดเอนทิตี้หนึ่งแล้วนักพัฒนาคนนั้นควรจะต้องเปลี่ยนแปลงแก้ไข (เปลี่ยนแปลง เพิ่มลง หรือ ลบออก) เอนทิตี้ใดด้วยต่อไป (Zimmermann et al., 2005; Methanias et al., 2009)

8. คำแนะนำสำหรับเหตุการณ์ Q (Suggestions for Situation Q) คือ เซตของการเปลี่ยนแปลงแก้ไขที่นักพัฒนาควรจะทำตามหลังจากที่นักพัฒนาได้เปลี่ยนแปลงแก้ไขตามเหตุการณ์ Q โดยอ้างอิงมาจากเซตของกฎความสัมพันธ์ที่มีเซตรายการที่มาก่อนเป็นเซตเหตุการณ์ Q (Zimmermann et al., 2005)
9. สถานการณ์การนำทาง (Navigation) คือ สถานการณ์ที่นักพัฒนามีการเปลี่ยนแปลงแก้ไขที่เอนทิตี้หนึ่งแล้ว ระบบจะต้องให้คำแนะนำกับนักพัฒนาให้แก้ไขเอนทิตี้ใดต่อไป (Zimmermann et al., 2005; Methanias et al., 2009)
10. สถานการณ์การป้องกันการเกิดข้อผิดพลาด (Error Prevention) คือ สถานการณ์ที่นักพัฒนามีการเปลี่ยนแปลงแก้ไขที่เอนทิตี้หลายๆเอนทิตี้ต่อเนื่องกันแต่ยังขาดการเปลี่ยนแปลงแก้ไขเอนทิตี้อีกหนึ่งเอนทิตี้จึงจะสมบูรณ์ ระบบจะต้องให้คำแนะนำกับนักพัฒนาให้แก้ไขเอนทิตี้ที่เหลือนั้นได้ (Zimmermann et al., 2005; Methanias et al., 2009)
11. สถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (Closure) คือ สถานการณ์ที่นักพัฒนามีการเปลี่ยนแปลงแก้ไขที่เอนทิตี้หลายๆเอนทิตี้ต่อเนื่องกันอย่างครบถ้วนแล้ว ระบบจะต้องไม่ให้คำแนะนำแก้ไขเอนทิตี้ใดๆกับนักพัฒนา (Zimmermann et al., 2005; Methanias et al., 2009)
12. ไอดีอี (IDE: Integrated Development Environment) คือ โปรแกรมประยุกต์ที่จัดเตรียมสิ่งแวดล้อมซึ่งอำนวยความสะดวกให้แก่นักพัฒนาซอฟต์แวร์ โดยปกติแล้วประกอบด้วย เครื่องมือพัฒนาซอร์สโค้ด (Source Code Editor) ตัวแปลภาษาคอมพิวเตอร์ (Compiler) หรือ ตัวแปลคำสั่งคอมพิวเตอร์ (interpreter) หรือทั้งสอง เครื่องมือสร้างระบบอัตโนมัติ (Build Automation Tools) และเครื่องมือตรวจแก้ข้อผิดพลาด (Debugger) เป็นพื้นฐาน

บทที่ 2
ทบทวนวรรณกรรมที่เกี่ยวข้อง.

2.1 บทนำ

วรรณกรรมที่เกี่ยวข้องกับงานวิจัยนี้ถูกเรียบเรียงไว้ตามลำดับ เริ่มต้นจากแนวคิดของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ (Association Rules Discovery) รวมถึงค่าที่ใช้ในการประเมินความน่าสนใจของกฎความสัมพันธ์ (Interestingness Measure of Association Rules) ที่มีงานวิจัยในอดีตเคยเสนอไว้ทั้งหมด 8 ค่า ต่อมาจะอธิบายถึงตัวแบบการประเมินความน่าสนใจของกฎความสัมพันธ์ใหม่ โดยเริ่มจากการอธิบายถึงข้อบกพร่องของตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นดั้งเดิม ตามด้วยการอธิบายตัวแบบการประเมินความน่าสนใจของกฎความสัมพันธ์แบบใหม่ และต่อด้วยการอธิบายคุณสมบัติที่ค่าประเมินความน่าสนใจควรจะมีทั้งหมด 16 คุณสมบัติ รวมถึงสรุปคุณสมบัติที่ค่าประเมินความน่าสนใจทั้ง 8 ค่าและค่าความเชื่อมั่นใหม่มี หัวข้อถัดมาเป็นการอธิบายแนวคิดของการควบคุมการเปลี่ยนแปลงแก้ไข (Concept of Revision Control, Version Control) ซอฟต์แวร์ควบคุมการเปลี่ยนแปลงแก้ไข (Revision Control Software, Version Control Software) รวมถึงระบบคอนเคอเรนท์เวอร์ชั่น (Concurrent Versions System, CVS) ซึ่งเป็นระบบที่ผู้วิจัยสนใจนำข้อมูลมาทำการทดสอบ หัวข้อถัดมาเป็นการรวบรวมวรรณกรรมที่เกี่ยวข้องกับการประยุกต์ใช้การทำเหมืองข้อมูลด้วยเทคนิคค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Applying Association Rule Discovery in Software Archive) ตั้งแต่อดีตจนถึงปัจจุบัน รวมถึงอธิบายถึงขั้นตอนวิธีการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Data Mining in Software Archives) ของงานวิจัยต่างๆในอดีตอย่างละเอียด ตั้งแต่การจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Preparing Data for Mining in Software Archives) จนถึงขั้นตอนการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Data Mining in Software Archives) และขั้นตอนการสร้างคำแนะนำจากกฎความสัมพันธ์ (Generating Suggestions for Situation) สุดท้ายจะกล่าวถึงการวัดประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์

2.2 การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ (Association Rules Discovery)

การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ (Association Rules discovery) คือการค้นหาความสัมพันธ์ที่มีความเกี่ยวข้องและเชื่อมโยงกันระหว่างรายการข้อมูล (Data Items) ภายในฐานข้อมูลขนาดใหญ่ที่มีอยู่เพื่อนำไปใช้ในการวิเคราะห์หรือทำนายปรากฏการณ์ต่างๆ (Agrawal et al., 1993) ตัวอย่างในการนำกฎความสัมพันธ์ไปประยุกต์ใช้ในทางปฏิบัติที่ได้รับความนิยมมากที่สุดคือ การวิเคราะห์พฤติกรรมการซื้อของลูกค้า (Market Basket Analysis)

นิยามทางคณิตศาสตร์ของการค้นหากฎความสัมพันธ์นั้นสามารถอธิบายได้ดังนี้ กำหนดให้ เซต $I = \{i_1, i_2, \ldots, i_m\}$ เป็นเซตของรายการข้อมูล (items) ที่มีอยู่ทั้งหมดและให้ เซต $T = \{t_1, t_2, \ldots, t_n\}$ เป็นเซตของทรานแซคชั่น โดยที่แต่ละทรานแซคชั่น t_n ประกอบด้วยเซตย่อย I_j (j = 1, 2, .. m) ที่เป็นเซตย่อยของเซตของรายการข้อมูล I เซตของรายการข้อมูล I_j นั้นถูกเรียกว่า เซตรายการ (Itemset) ดังนั้นกฎความสัมพันธ์ r ก็คือคู่ของเซตรายการ I_1 และเซตรายการ I_2 โดยที่เซตรายการ I_1 และเซตรายการ I_2 เป็นเซตย่อยของเซต I ที่ไม่มีสมาชิกที่ซ้อนทับกันและเซตรายการ I_2 ไม่เท่ากับเซตว่าง เซตรายการ I_1 ถูกเรียกว่า เซตรายการที่มาก่อน (Antecedent Itemset) และเซตรายการ I_2 ถูกเรียกว่า เซตรายการที่ตามมา (Consequent Itemset) และกำหนดสัญลักษณ์ $I_1 \rightarrow I_2$ แทนกฎความสัมพันธ์ที่มีเซตรายการ I_1 เป็นเซตรายการที่มาก่อน และเซตรายการ I_2 เป็นเซตรายการที่ตามมา โดยที่ I_2 ไม่ใช่เซตว่าง (Olivier et al., 2008)

วิธีการที่ดีวิธีหนึ่งที่จะระบุว่ากฎความสัมพันธ์นั้นๆมีความตรงประเด็นมากน้อยเพียงใด คือการกำหนดค่าที่ใช้ในการประเมินคุณภาพของกฎความสัมพันธ์หรือที่เรียกว่าค่าประเมินความน่าสนใจของกฎความสัมพันธ์ ตัวแบบของการประเมินความน่าสนใจของกฎความสัมพันธ์ที่เป็นตัวแบบแรกและดั้งเดิมคือตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น (Support-Confidence Model) ซึ่งใช้ค่า 2 ค่าในการประเมิน ค่าแรกถูกเรียกว่าค่าสนับสนุน (Support) ของกฎความสัมพันธ์ ใช้สัญลักษณ์ Support($I_1 \rightarrow I_2$) ค่าสนับสนุนนี้สามารถคำนวณได้จากจำนวนของทรานแซคชั่นที่มีทั้งส่วนเซตรายการที่มาก่อน (I_1) และเซตรายการที่ตามมา (I_2) ของกฎความสัมพันธ์หารด้วยจำนวนของทรานแซคชั่นทั้งหมด ซึ่งค่าสนับสนุนนี้ก็คือค่าความน่าจะเป็นที่จะมีทรานแซคชั่นที่มีทั้งรายการในเซตรายการที่มาก่อนและเซตรายการที่ตามมาในทรานแซคชั่นทั้งหมดนั่นเอง ในบางครั้งค่าสนับสนุนของกฎความสัมพันธ์นั้นอาจถูกนำไปใช้ในรูปร้อยละของจำนวนทรานแซคชั่น

ทั้งหมด ค่าสนับสนุนถูกใช้ในการประเมินความถี่ (Frequent) ในการปรากฏของรายการ (Item) หรือเซตรายการ หรือกฎความสัมพันธ์ นอกจากการพิจารณาค่าสนับสนุนของกฎความสัมพันธ์ก็คือการพิจารณาค่าความเชื่อมั่น (Confidence) ของกฎความสัมพันธ์ ใช้สัญลักษณ์ Conf($I_1 \rightarrow I_2$)) ค่าความเชื่อมั่นนี้ก็คืออัตราส่วนของค่าสนับสนุนของกฎความสัมพันธ์กับค่าสนับสนุนของเซตรายการที่มาก่อน หรืออาจเรียกได้ว่าค่าความเชื่อมั่นก็คือค่าความน่าจะเป็นที่จะมีรายการทุกรายการในเซตรายการที่มาก่อนแล้วจะมีรายการทุกรายการในเซตรายการที่ตามมาด้วย ค่าความเชื่อมั่นนั้นถูกนำไปใช้ในการประเมินระดับความน่าสนใจของกฎความสัมพันธ์

นอกจากตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น (Support-Confidence Model) แล้วการประเมินระดับความน่าสนใจของกฎความสัมพันธ์ยังสามารถใช้วิธีการอื่นๆได้ ซึ่งถูกรวบรวมเอาไว้และอธิบายอย่างละเอียดในหัวข้อถัดไป

2.3 **การประเมินความน่าสนใจของกฎความสัมพันธ์** (Interestingness Measure of Association Rules)

หัวข้อนี้ได้รวบรวมวรรณกรรมต่างๆที่เกี่ยวข้องกับการเสนอตัวแบบที่นำมาใช้ในการประเมินความน่าสนใจของกฎความสัมพันธ์ตั้งแต่ในอดีตจนถึงในปัจจุบันอันได้แก่ ค่าสนับสนุน (Support), ค่าความเชื่อมั่น (Confidence), ค่าคอนวิคชั่น (Conviction), ค่าลิฟท์ (Lift), ค่าเลฟเวอเรจ (Leverage), ค่าคัฟเวอเรจ (Coverage), ค่าสหสัมพันธ์ (Correlation) และ ค่าอัตราส่วนออดส์ (Odds Ratio) ตามลำดับ รวมถึงวรรณกรรมที่มีการสนับสนุนตัวแบบข้างต้นและวรรณกรรมที่พิสูจน์ข้อบกพร่องของตัวแบบบางตัวด้วย เพื่อให้ง่ายต่อการเปรียบเทียบกันของทุกตัวแบบที่นิยามขึ้นมาจากค่าทางสถิติจึงกำหนดนิยามค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูลดังนี้

กำหนดให้ P(X) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล
n(X) คือ จำนวนของทรานแซคชั่นที่มีรายการ X ปรากฏอยู่
N คือ จำนวนของทรานแซคชั่นทั้งหมดในฐานข้อมูล

$$P(X) = \frac{n(X)}{N}$$

2.3.1 **ค่าสนับสนุน** (Support)

ค่าสนับสนุนถูกนำเสนอขึ้นมาครั้งแรกโดย Agrawal และคณะในปีค.ศ. 1993 (Agrawal et al., 1993) ค่าสนับสนุนคือค่าความน่าจะเป็นของรายการนั้นนั่นเอง โดยปกติค่าสนับสนุนนั้นเป็นค่าที่ถูกใช้ในการแสดงความถี่ในการปรากฏของแต่ละรายการ แต่ค่าสนับสนุนนี้ก็สามารถนำมาใช้ในการประเมินความน่าสนใจของกฎความสัมพันธ์ได้ สมการในการคำนวณค่าสนับสนุนของรายการ X แสดงได้ดังนี้

กำหนดให้ Support(X) คือ ค่าสนับสนุนของรายการ X

P(X) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

$$Support(X) = P(X)$$

การหาค่าสนับสนุนของรายการหรือบางที่ก็ถูกเรียกว่าค่าความถี่ของรายการ เซตรายการที่มีค่าสนับสนุนมากกว่าค่าสนับสนุนขั้นต่ำที่กำหนดไว้จะเรียกเซตรายการนั้นว่าเซตรายการความถี่สูง (Frequent Itemset) หรือเซตรายการใหญ่ (large Itemset) หรือกล่าวคือค่าสนับสนุนนี้จะถูกนำมาใช้เพื่อเป็นการคัดกรองรายการทั้งหมดในฐานข้อมูลให้เหลือแต่เพียงรายการที่มีความถี่สูงตามที่กำหนดไว้เพื่อนำเซตรายการเหล่านั้นไปค้นหากฎความสัมพันธ์ในภายหลัง

ในกรณีที่จะใช้ค่าสนับสนุนมาประเมินระดับความน่าสนใจ ก็คือค่าความถี่ของกฎความสัมพันธ์นั้นนั่นเอง สมการในการคำนวณค่าสนับสนุนของกฎความสัมพันธ์ แสดงได้ดังนี้

กำหนดให้ Support(R) คือ ค่าสนับสนุนของกฎความสัมพันธ์ R

P(X) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

P(Y) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ Y ในฐานข้อมูล

P(X and Y) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และ Y ในฐานข้อมูล

$$\text{Support}(X \rightarrow Y) = P(X \text{ and } Y)$$

ข้อเสียของการใช้ค่าสนับสนุนก็คือการก่อให้เกิดปัญหาขาดแคลนรายการ (rare item problem) เนื่องจากบางรายการนั้นเป็นรายการที่ปรากฏอยู่ในฐานข้อมูลจำนวนน้อยจึงถูกคัดออกไป แต่ในความจริงนั้นรายการที่ปรากฏอยู่ในฐานข้อมูลจำนวนน้อยอาจสามารถนำไปสร้างเป็นกฎความสัมพันธ์ที่น่าสนใจหรือมีคุณค่าได้ ค่าที่เป็นไปได้ของค่าสนับสนุนนั้นอยู่ในพิสัย [0, 1] ถ้าเซตรายการที่มาก่อนและเซตรายการที่ตามมาของกฎความสัมพันธ์เป็นอิสระต่อกันแล้วค่าสนับสนุนจะเท่ากับ 0 ถ้าเซตรายการที่มาก่อนและเซตรายการที่ตามมาของกฎความสัมพันธ์มีการปรากฏขึ้นพร้อมกันเสมอค่าสนับสนุนจะเท่ากับ 1 (Agrawal et al., 1993; Sheikh et al., 2004)

2.3.2 **ค่าความเชื่อมั่น** (Confidence)

ค่าความเชื่อมั่นถูกนำเสนอขึ้นมาครั้งแรกโดย Agrawal และคณะในปีค.ศ. 1993 (Agrawal et al., 1993) เช่นเดียวกับค่าสนับสนุนนิยามของค่าความเชื่อมั่นคือความน่าจะเป็นที่จะพบกฎความสัมพันธ์ $X \rightarrow Y$ ในทรานแซคชั่นที่มีรายการ X ปรากฏอยู่ ดังนั้นการคำนวณค่าความเชื่อมั่นของกฎความสัมพันธ์ $X \rightarrow Y$ และกฎความสัมพันธ์ $Y \rightarrow X$ นั้นจะให้ค่าที่แตกต่างกัน สมการในการคำนวณค่าความเชื่อมั่นแสดงได้ดังนี้

กำหนดให้ Conf(R) คือ ค่าความเชื่อมั่นของกฎความสัมพันธ์ R

P(X) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

P(X and Y) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และ Y ในฐานข้อมูล

P(Y|X) คือ ค่าความน่าจะเป็นในการพบรายการ Y ในทรานแซคชั่นที่มีรายการ X อยู่แล้ว

$$\text{Conf}(X \rightarrow Y) = \frac{\mathbf{P(X \text{ and } Y)}}{\mathbf{P(X)}}$$

หรือ $$\text{Conf}(X \rightarrow Y) = P(Y|X)$$

เนื่องจากค่าความเชื่อมั่นและค่าสนับสนุนนั้นถูกพัฒนาขึ้นมาพร้อมกันโดย Agrawal และคณะ ค่าสนับสนุนจะถูกนำมาใช้เพื่อเป็นการคัดกรองรายการทั้งหมดในฐานข้อมูลให้เหลือแต่เพียงรายการที่มีความถี่สูงตามที่กำหนดไว้ ต่อจากนั้นค่าความเชื่อมั่นจะถูกนำมาใช้ในการสร้างกฎ

ความสัมพันธ์ขึ้นมาจากผลลัพธ์เซตรายการที่มีความถี่สูงที่ได้มาจากการหาค่าสนับสนุนโดยการคำนวณหาค่าความเชื่อมั่นของทุกกฎความสัมพันธ์ที่เป็นไปได้จากเซตรายการที่มีความถี่สูงทั้งหมด ค่าความเชื่อมั่นของกฎความสัมพันธ์ใดที่มากกว่าค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence) ที่กำหนดไว้จะถือว่ากฎความสัมพันธ์นั้นมีอยู่จริงและมีความน่าสนใจ

ปัญหาที่เกิดขึ้นจากการใช้ค่าความเชื่อมั่นคือค่าความเชื่อมั่นนั้นค่อนข้างจะอ่อนไหวง่ายต่อความถี่ของเซตรายการที่ตามมา กล่าวคือถ้ากฎความสัมพันธ์นั้นมีเซตรายการที่ตามมาที่มีความถี่สูงมาก (ในขณะที่เซตรายการที่มาก่อนมีความถี่ที่น้อยกว่ามาก) จะทำให้ค่าความเชื่อมั่นของกฎความสัมพันธ์มีค่าสูงส่งผลให้กฎความสัมพันธ์นั้นถูกพิจารณาว่ามีความน่าสนใจ ซึ่งในความจริงเซตรายการที่มาก่อนและเซตรายการที่ตามมาอาจมีความสัมพันธ์กันน้อยมากก็ได้ (Sheikh et al., 2004) ค่าที่เป็นไปได้ของค่าความเชื่อมั่นนั้นอยู่ในพิสัย [0, 1] ถ้าเซตรายการที่มาก่อนและเซตรายการที่ตามมาของกฎความสัมพันธ์เป็นอิสระต่อกันแล้วค่าความเชื่อมั่นจะเท่ากับ 0 ถ้าเซตรายการที่มาก่อนและเซตรายการที่ตามมาของกฎความสัมพันธ์มีการปรากฏขึ้นพร้อมกันเสมอค่าความเชื่อมั่นจะเท่ากับ 1 (Agrawal et al., 1993; Sheikh et al., 2004)

2.3.3 **ค่าคอนวิคชั่น** (Conviction)

ค่าคอนวิคชั่นถูกเสนอขึ้นมาครั้งแรกโดย Brin และคณะในปีค.ศ. 1997 (Brin et al., 1997) ค่าคอนวิคชั่นถูกพัฒนาขึ้นมาเพื่อแก้ไขข้อด้อยของค่าความเชื่อมั่นที่ไม่สามารถจะบ่งบอกทิศของกฎความสัมพันธ์ได้ภายในการคำนวณครั้งเดียว กล่าวคือการใช้ค่าความเชื่อมั่นจะต้องมีการคำนวณทั้งค่าความเชื่อมั่นของกฎความสัมพันธ์ทั้งไปและกลับเพื่อเลือกกฎความสัมพันธ์ที่มีค่าค่าความเชื่อมั่นสูงกว่ากัน ค่าคอนวิคชั่นของกฎความสัมพันธ์ $X \rightarrow Y$ นั้นจะเปรียบเทียบความน่าจะเป็นในการมีรายการ X ปรากฏโดยที่ไม่มีรายการ Y ปรากฏ สมการในการคำนวณค่าคอนวิคชั่นแสดงได้ดังนี้

กำหนดให้ Conviction(R) คือ ค่าคอนวิคชั่นของกฎความสัมพันธ์ R

P(X) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

$P(\overline{X})$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่ไม่มีรายการ X ในฐานข้อมูล

$P(X \text{ and } \overline{Y})$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X แต่ไม่มี

รายการ Y ในฐานข้อมูล

$$\text{Conviction}(X \rightarrow Y) = \frac{\mathbf{P(X)P(\overline{Y})}}{\mathbf{P(X \text{ and } \overline{Y})}}$$

โดยที่ $P(\overline{Y})$ คือความน่าจะเป็นที่ไม่มีรายการ Y ปรากฏในทรานแซคชั่น ค่าที่เป็นไปได้ของค่าคอนวิคชั่นนั้นอยู่ในพิสัย [0, $+\infty$] ถ้าเซตรายการที่มาก่อนและเซตรายการที่ตามมาของกฎความสัมพันธ์เป็นอิสระต่อกันแล้วค่าคอนวิคชั่นจะเท่ากับ 1 แต่ถ้าเซตรายการที่มาก่อนและเซตรายการที่ตามมาของกฎความสัมพันธ์มีการปรากฏขึ้นพร้อมกันเสมอค่าคอนวิคชั่นจะเท่ากับ $+\infty$ (Brin et al., 1997; Sheikh et al., 2004)

2.3.4 **ค่าลิฟท์** (Lift)

ค่าลิฟท์ถูกเสนอขึ้นมาครั้งแรกโดย Brin และคณะในปีค.ศ. 1997 (Brin et al., 1997) ค่าลิฟท์นั้นเป็นค่าที่ใช้ในการประเมินความถี่ในการปรากฏร่วมกันของเซตรายการที่มาก่อนและเซตรายการที่ตามมาของกฎความสัมพันธ์โดยที่เซตรายการที่มาก่อนและเซตรายการที่ตามมานั้นเป็นอิสระต่อกันทางสถิติ ข้อดีประการหนึ่งของการคำนวณค่าลิฟท์คือการใช้ค่าลิฟท์จะไม่พบกับปัญหาขาดแคลนรายการ สมการในการคำนวณค่าลิฟท์แสดงได้ดังนี้

กำหนดให้ Lift(R) คือ ค่าลิฟท์ของกฎความสัมพันธ์ R

P(X) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

P(X and Y) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และรายการ Y ในฐานข้อมูล

P(Y|X) คือ ค่าความน่าจะเป็นในการพบรายการ Y ในทรานแซคชั่นที่มีรายการ X อยู่แล้ว

Conf(R) คือ ค่าความเชื่อมั่นของกฎความสัมพันธ์ R

$$\text{Lift}(X \rightarrow Y) = \frac{\mathbf{P(X \text{ and } Y)}}{\mathbf{P(X)P(Y)}}$$

หรือ

$$\text{Lift}(X \rightarrow Y) = \frac{\mathbf{P(Y|X)}}{\mathbf{P(Y)}}$$

หรือ $$\text{Lift}(X \rightarrow Y) = \frac{\text{Conf}(X \rightarrow Y)}{\text{P(Y)}}$$

ค่าที่เป็นไปได้ของค่าลิฟท์นั้นอยู่ในพิสัย [0, $+\infty$] ถ้าเซตรายการที่มาก่อนและเซตรายการที่ตามมาของกฎความสัมพันธ์เป็นอิสระต่อกันแล้วค่าลิฟท์จะเท่ากับ 1 ถ้าค่าลิฟท์มีค่ามากกว่า 1 หมายถึงเซตรายการที่มาก่อนและเซตรายการที่ตามมามีความสัมพันธ์เชิงบวกต่อกัน ถ้าค่าลิฟท์มีค่าน้อยกว่า 1 หมายถึงเซตรายการที่มาก่อนและเซตรายการที่ตามมามีความสัมพันธ์เชิงลบต่อกัน (Brin et al., 1997; Sheikh et al., 2004)

2.3.5 **ค่าเลฟเวอเรจ** (Leverage)

ค่าเลฟเวอเรจถูกเสนอขึ้นมาครั้งแรกโดย Piatetsky-Shapiro และคณะในปี 1991 (Piatetsky-Shapiro et al., 1991) ค่าเลฟเวอเรจถูกพัฒนาขึ้นมาเพื่อประเมินความต่างของค่าความน่าจะเป็นของการปรากฏเซตรายการที่มาก่อนและเซตรายการที่ตามมาขึ้นพร้อมกันโดยที่เซตรายการทั้งสองนั้นมีการขึ้นต่อกันทางสถิติกับค่าความน่าจะเป็นของการปรากฏเซตรายการที่มาก่อนและเซตรายการที่ตามมาที่ปรากฏอย่างเป็นอิสระต่อกันในฐานข้อมูล สมการในการคำนวณค่าเลฟเวอเรจแสดงได้ดังนี้

กำหนดให้ Leverage(R) คือ ค่าเลฟเวอเรจของกฎความสัมพันธ์ R

P(X) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

P(X and Y) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และรายการ Y ในฐานข้อมูล

$$\text{Leverage}(X \rightarrow Y) = \text{P(X and Y)} - \text{P(X)P(Y)}$$

เหตุผลสำคัญที่มีการพัฒนาค่าเลฟเวอเรจขึ้นมานั้นมาจากความต้องการตั้งวิธีการขายสินค้า โดยทำการค้นหาว่าการขายสินค้าทั้งสองเซตรายการร่วมกันกับการขายที่เป็นอิสระต่อกันแบบไหนมีค่ามากกว่ากัน ข้อด้อยสำคัญของค่าเลฟเวอเรจก็คือการใช้ค่านี้อาจทำให้ประสบกับปัญหาขาดแคลนรายการได้ ค่าที่เป็นไปได้ของค่าเลฟเวอเรจนั้นอยู่ในพิสัย [$-\infty$, $+\infty$] (Piatetsky-Shapiro et al., 1991; Sheikh et al., 2004)

2.3.6 **ค่าคัฟเวอเรจ** (Coverage)

ค่าคัฟเวอเรจคือค่าสนับสนุนของเซตรายการที่มาก่อน หรือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการที่มาก่อนในฐานข้อมูลนั่นเอง แนวคิดของค่าคัฟเวอเรจคือการใช้ความถี่ของเซตรายการที่มาก่อนของกฎความสัมพันธ์มาใช้เป็นตัวระบุถึงความน่าสนใจของกฎความสัมพันธ์นั้น สมการในการคำนวณค่าคัฟเวอเรจแสดงได้ดังนี้

กำหนดให้ Coverage (R) คือ ค่าคัฟเวอเรจของกฎความสัมพันธ์ R

P(X) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

$$\text{Coverage}(X \rightarrow Y) = P(X)$$

ค่าที่เป็นไปได้ของค่าคัฟเวอเรจนั้นอยู่ในพิสัย [0, 1] (Michael., 2009)

2.3.7 **ค่าสหสัมพันธ์** (Correlation)

ค่าสหสัมพันธ์หรือค่าสหสัมพันธ์นี้เป็นเทคนิคทางสถิติที่สามารถแสดงได้ว่าเซตรายการใดมีความสัมพันธ์กันอย่างแข็งแรง สมการในการคำนวณค่าสหสัมพันธ์แสดงได้ดังนี้

กำหนดให้ Corr(R) คือ ค่าสหสัมพันธ์ของกฎความสัมพันธ์ R

P(X) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

P(X and Y) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และรายการ Y ในฐานข้อมูล

$$\text{Corr}(X \rightarrow Y) = \frac{(P(X \text{ and } Y) - P(X)P(Y))}{\sqrt{P(X)P(Y)P(1-P(X))(1-P(Y))}}$$

ค่าสหสัมพันธ์นี้สามารถนำมาใช้ประเมินลักษณะของความสัมพันธ์ระหว่างเซตรายการที่มาก่อนและเซตรายการที่ตามมาได้ ค่าที่เป็นไปได้ของค่าสหสัมพันธ์นั้นอยู่ในพิสัย [-1, 1] โดยที่ค่าสหสัมพันธ์ที่เป็นค่าลบ หมายถึง เซตรายการทั้งสองแปรผกผันกัน ถ้าค่าสหสัมพันธ์เท่า -1 จะหมายถึงเซตรายการทั้งสองแปรผกผันกันอย่างสมบูรณ์ ค่าสหสัมพันธ์ที่เป็นค่าบวก หมายถึง เซต

รายการทั้งสองแปรตามกัน ถ้าค่าสหสัมพันธ์เท่า 1 จะหมายถึงเซตรายการทั้งสองแปรผกตามกันอย่างสมบูรณ์ และถ้าค่าสหสัมพันธ์เป็น 0 หมายถึงเซตรายการทั้งสองไม่มีความสัมพันธ์ต่อกันเลย (Sheikh et al., 2004)

2.3.8 **ค่าอัตราส่วนออดส์** (Odds Ratio)

ค่าอัตราส่วนออดส์ คือค่าประเมินทางสถิติที่คำนวณหาความสัมพันธ์ระหว่างตัวแปร 2 ตัวที่แต่ละตัวเป็นตัวแปรจัดกลุ่มที่แบ่งออกเป็น 2 กลุ่ม แนวคิดของออดส์มาจากการเสี่ยงโชค (gambling) ตัวอย่างเช่น ออดส์ของม้าที่จะแข่งชนะเป็น 3 : 1 หมายถึงความน่าจะเป็นที่ม้าจะชนะ 3 ครั้งต่อการไม่ชนะ 1 ครั้ง เมื่อนำค่าทางสถิตินี้มาประยุกต์ใช้กับการประเมินกฎความสัมพันธ์จะได้ว่า ค่าอัตราส่วนออดส์ คืออัตราส่วนระหว่างออดส์ของเหตุการณ์ที่มีเซตรายการที่มาก่อนและเซตรายการที่ตามมาปรากฏหรือไม่ปรากฏทั้งคู่กับออดส์ของเหตุการณ์ที่มีเซตรายการที่มาก่อนปรากฏแต่เซตรายการที่ตามมาไม่ปรากฏหรือเหตุการณ์ที่ไม่มีเซตรายการที่มาก่อนปรากฏแต่มีเซตรายการที่ตามมาปรากฏ โดยค่าอัตราส่วนออดส์ มีสูตรคำนวณดังนี้ สมการในการคำนวณค่าคอนวิคชั่นแสดงได้ดังนี้

กำหนดให้ Odds(R) คือ ค่าอัตราส่วนออดส์ของกฎความสัมพันธ์ R

P(X) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

$P(\overline{X})$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่ไม่มีรายการ X ในฐานข้อมูล

P(X and Y) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และรายการ Y ในฐานข้อมูล

$P(\overline{X} \text{ and } \overline{Y})$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่ไม่มีรายการ X และรายการ Y ในฐานข้อมูล

$P(X \text{ and } \overline{Y})$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และไม่รายการ Y ในฐานข้อมูล

$$\text{Odds}(X \rightarrow Y) = \frac{(P(X \text{ and } Y)\, P(\overline{X} \text{ and } \overline{Y}))}{(P(X \text{ and } \overline{Y})\, P(\overline{X} \text{ and } Y))}$$

ค่าที่เป็นไปได้ของค่าอัตราส่วนออดส์นั้นอยู่ในพิสัย [0, +∞] ถ้าเซตรายการที่มาก่อนและเซตรายการที่ตามมาของกฎความสัมพันธ์เป็นอิสระต่อกันแล้วค่าอัตราส่วนออดส์จะเท่ากับ 0 กฎความสัมพันธ์ที่มีความน่าสนใจมากจะมีค่าอัตราส่วนออดส์เท่ากับ +∞ (Sheikh et al., 2004)

2.4 ตัวแบบการประเมินความน่าสนใจของกฎความสัมพันธ์ใหม่

การค้นหากฎความสัมพันธ์นั้นจัดเป็นหนึ่งในงานที่สำคัญและได้รับความนิยมมากในการทำเหมืองข้อมูล ตัวแบบดั้งเดิมที่ใช้ในประเมินในการค้นหากฎความสัมพันธ์นั้นก็คือตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น ที่ได้กล่าวไปแล้วในข้างต้น ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นนี้จะใช้การประเมินค่าความเชื่อมั่นเป็นตัวชี้วัดระดับความน่าสนใจของกฎความสัมพันธ์ การประเมินความน่าสนใจของกฎความสัมพันธ์ด้วยวิธีแบบดั้งเดิมนี้มีข้อบกพร่องอยู่หลายประการซึ่งจะอธิบายข้อบกพร่องเหล่านี้อย่างละเอียดในหัวข้อ 2.4.1 ต่อจากนั้นจะกล่าวถึงตัวแบบการประเมินความน่าสนใจของกฎความสัมพันธ์ที่เสนอโดย Liu และคณะในปี 2008 (Liu et al., 2008) ในงานวิจัยนี้ผู้วิจัยจะเรียกตัวแบบใหม่ดังกล่าวนี้ว่า ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ (Support-New Confidence Model) ของ Liu และคณะ (Liu et al., 2008)

2.4.1 ข้อบกพร่องของตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น

การทำเหมืองข้อมูลของกฎความสัมพันธ์โดยใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น (Support-Confidence Model) นั้นค่าที่ใช้ในการประเมินความน่าสนใจของกฎความสัมพันธ์ $X \rightarrow Y$ ก็คือค่าความเชื่อมั่นของกฎความสัมพันธ์ (Conf($X \rightarrow Y$)) แต่ในบางครั้งการทำเหมืองข้อมูลของกฎความสัมพันธ์โดยตัวแบบนี้อาจจะทำให้ได้รับกฎความสัมพันธ์ที่ไม่มีความเกี่ยวข้องกันจริงและอาจนำไปสู่การกำหนดกฎความสัมพันธ์ที่ผิดหรือเป็นผลบวกลวง (False Positive) ได้ การนำตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นไปใช้งานจึงมีข้อจำกัดดังแสดงในตัวอย่างต่อไปนี้

<u>ตัวอย่างที่ 1</u> กำหนดให้ตารางด้านล่างนี้เป็นตารางแสดงทรานแซคชั่นอย่างง่าย แต่ละแถวแทนรายการต่างๆในทรานแซคชั่น แต่ละหลักแทนทรานแซคชั่น 1 ทรานแซคชั่น จำนวนทรานแซคชั่นทั้งหมดที่แสดงในตารางคือ 8 ทรานแซคชั่นคือทรานแซคชั่น T1-T8 มีรายการทั้งหมด 3 รายการคือ รายการ X รายการ Y และรายการ Z หมายเลข 1 หมายถึงมีรายการของแถวนั้นปรากฏอยู่บนทรานแซคชั่นของหลักนั้น และหมายเลข 0 หมายถึงไม่มีรายการของแถวนั้นปรากฏอยู่บนทรานแซคชั่นของหลักนั้น (Tan et al., 2002; Liu et al., 2008)

ตารางที่ 2-1 แสดงตัวอย่างทรานแซคชั่นที่ประกอบด้วยรายการ X Y และ Z

	T1	T2	T3	T4	T5	T6	T7	T8
X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	0	1	1	1	1	1	1	1

พิจารณาตารางข้างต้นจะเห็นได้ว่ารายการ X และรายการ Y นั้นมีความสัมพันธ์ที่แปรผันตรงกันหรือมีสหสัมพันธ์เชิงบวกต่อกัน (Positively Correlated) เป็นส่วนใหญ่ กล่าวคือถ้ามีรายการ X ปรากฏอยู่บนทรานแซคชั่นก็จะมีรายการ Y ปรากฏบนทรานแซคชั่นเป็นส่วนใหญ่และในทางกลับกันถ้าไม่มีรายการ X ปรากฏอยู่บนทรานแซคชั่นก็จะไม่มีรายการ Y ปรากฏบนทรานแซคชั่นเป็นส่วนใหญ่เช่นกัน นอกจากนั้นยังสามารถสังเกตได้ว่ารายการ X และรายการ Z นั้นมีความสัมพันธ์ที่แปรผกผันกันหรือมีสหสัมพันธ์เชิงลบต่อกัน (Negatively Correlated) เป็นส่วนใหญ่ (ร้อยละ 62.5) ด้วย กล่าวคือถ้ามีรายการ X ปรากฏอยู่บนทรานแซคชั่นก็จะไม่มีรายการ Z ปรากฏบนทรานแซคชั่นเป็นส่วนใหญ่และในทางกลับกันถ้าไม่มีรายการ X ปรากฏอยู่บนทรานแซคชั่นก็จะมีรายการ Z ปรากฏบนทรานแซคชั่นเป็นส่วนใหญ่เช่นกัน แต่อย่างไรก็ตามเมื่อได้คำนวณหาค่าสนับสนุนและค่าความเชื่อมั่นของกฎความสัมพันธ์ $X \rightarrow Y$ แล้วได้ค่า 25% และ 50% ตามลำดับ และเมื่อคำนวณหาค่าสนับสนุนและค่าความเชื่อมั่นของกฎความสัมพันธ์ $X \rightarrow Z$ แล้วได้ค่า 37.5% และ 75% ตามลำดับ จากค่าสนับสนุนและค่าความเชื่อมั่นของกฎความสัมพันธ์ทั้งสองบ่งชี้ว่ากฎความสัมพันธ์ $X \rightarrow Z$ มีความน่าสนใจมากกว่ากฎความสัมพันธ์ $X \rightarrow Y$ ตัวอย่างนี้ได้แสดงให้เห็นจุดบกพร่องของการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น ในการค้นหากฎความสัมพันธ์คือ กฎความสัมพันธ์ที่มีค่าความเชื่อมั่นที่มีค่าสูงในตัวอย่างกลับเป็นกฎความสัมพันธ์ที่มีความสัมพันธ์แปรผกผันกันหรือมีความสัมพันธ์เชิงลบต่อกันสูง

<u>ตัวอย่างที่ 2</u> กำหนดให้ตารางด้านล่างนี้เป็นตารางแสดงการปรากฏของรายการชาและรายการกาแฟในทรานแซคชั่นการซื้อสินค้าในร้านขายของชำแห่งหนึ่ง แถว t และ t' คือร้อยละของทรานแซคชั่นที่มีรายการชาปรากฏอยู่และไม่มีรายการชาปรากฏอยู่ตามลำดับ หลัก c และ c' คือร้อยละของทรานแซคชั่นที่มีรายการกาแฟปรากฏอยู่และไม่มีรายการกาแฟปรากฏอยู่ตามลำดับ (Brin et al., 1997; Liu et al., 2008)

ตารางที่ 2-2 แสดงตัวอย่างร้อยละของการปรากฏของรายการบนทรานแซคชั่นของร้านขายของชำแห่งหนึ่ง

	c	c'	รวม
t	20	5	25
t'	70	5	75
รวม	90	10	100

จากข้อมูลในตารางข้างต้น เมื่อนำไปใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น เพื่อค้นหากฎความสัมพันธ์จะได้ผลลัพธ์คือได้กฎความสัมพันธ์ $t \rightarrow c$ มีค่าสนับสนุนเท่ากับ 20% ซึ่งในทางปฏิบัติถือว่าเป็นค่าที่ค่อนข้างสูงมาก ค่าความเชื่อมั่นของกฎความสัมพันธ์ $t \rightarrow c$ นี้จะเป็นตัวบ่งชี้ถึงความน่าจะเป็นที่ลูกค้าจะซื้อกาแฟเมื่อลูกค้านั้นซื้อชา ซึ่งมีค่าเท่ากับความน่าจะเป็นที่จะซื้อชาและกาแฟหารด้วยความน่าจะเป็นที่จะซื้อชา นั่นคือ 20/25 = 0.8 หรือ 80% ซึ่งจัดได้ว่าเป็นค่าความเชื่อมั่นที่สูงมาก จากค่าสนับสนุนและค่าความเชื่อมั่นที่คำนวณไว้สามารถสรุปได้ว่ากฎความสัมพันธ์ $t \rightarrow c$ นี้เป็นกฎความสัมพันธ์ที่มีอยู่จริงอย่างสมเหตุสมผล

แต่ในความจริงแล้วข้อมูลที่แสดงในตารางไม่ใช่ข้อมูลทั้งหมดในฐานข้อมูลของร้านขายของชำแห่งนี้ ซึ่งอาจเป็นไปได้ว่าความสัมพันธ์ระหว่างการซื้อชาและการซื้อกาแฟในทรานแซคชั่นใดๆเป็นความสัมพันธ์แปรผกผันกัน จากตารางจะสามารถเห็นได้เพียงค่าร้อยละของการซื้อชาและซื้อกาแฟด้วยคือร้อยละ 20 ค่าที่กำหนดให้ในตารางไม่เพียงพอจะนำมาคำนวณค่าสหสัมพันธ์ได้โดยตรงแต่ว่าสามารถทราบทิศของความสัมพันธ์ระหว่างการซื้อชาและการซื้อกาแฟได้โดยการใช้การคำนวณค่าลิฟท์ดังสูตร P(tUc)/(P(t)xP(c)) = 0.2/(0.25x0.9) = 0.89 ค่าลิฟท์ที่คำนวณมาได้นั้นมีค่าน้อยกว่า 1 ซึ่งบ่งชี้ว่าระหว่างการซื้อชาและการซื้อกาแฟมีความสัมพันธ์เชิงลบต่อกันหรือการซื้อชาและการซื้อกาแฟแปรผกผันกันนั่นเอง จากตัวอย่างนี้แสดงเห็นถึงข้อบกพร่องของการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น ที่ทำให้ได้กฎความสัมพันธ์ที่ขัดแย้งต่อทิศของสหสัมพันธ์

2.4.2 ตัวแบบการประเมินความน่าสนใจของกฎความสัมพันธ์ใหม่

การประเมินความน่าสนใจของกฎความสัมพันธ์ด้วยวิธีการต่างๆ ที่กล่าวไปในข้างต้นนั้นแสดงให้เห็นอย่างชัดเจนว่าการใช้ตัวแบบประเมินที่แตกต่างกันส่งผลให้ได้กฎความสัมพันธ์ที่น่าสนใจต่างกันออกไป (Sheikh et al., 2004) เหตุผลสำคัญที่ทำให้เป็นเช่นนั้นก็เพราะความไม่สอดคล้องกันของความน่าจะเป็นของการมีกฎความสัมพันธ์ปรากฏกับค่าสหสัมพันธ์ของกฎ

ความสัมพันธ์ ในปี 2008 Liu และคณะ (Liu et al., 2008) ได้นำเสนอตัวแบบการประเมินความน่าสนใจของกฎความสัมพันธ์ขึ้นมาใหม่ โดยการพิสูจน์ทฤษฎีบทที่เกี่ยวข้องกับตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น กับทฤษฎีบทค่าสหสัมพันธ์ และตั้งชื่อการประเมินความน่าสนใจของกฎความสัมพันธ์แบบใหม่นี้ว่า ค่าความเชื่อมั่นใหม่ (New Confidence) โดยใช้สัญลักษณ์ NConf($X \rightarrow Y$) แทนค่าความเชื่อมั่นใหม่นี้ ซึ่งสามารถคำนวณได้จากสูตรดังต่อไปนี้

กำหนดให้ NConf(R) คือ ค่าความเชื่อมั่นใหม่ของกฎความสัมพันธ์ R

P(X) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

$P(\overline{X})$ คือ ค่าความน่าจะเป็นในการไม่พบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

P(X and Y) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และรายการ Y ในฐานข้อมูล

$P(X \text{ and } \overline{Y})$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และไม่รายการ Y ในฐานข้อมูล

P(X|Y) คือ ค่าความน่าจะเป็นในการพบรายการ X ในทรานแซคชั่นที่มีรายการ Y อยู่แล้ว

$$\text{NConf}(X \rightarrow Y) = \frac{P(X \text{ and } Y)}{P(Y)} - \frac{P(X \text{ and } \overline{Y})}{P(\overline{Y})}$$

หรือ $\text{NConf}(X \rightarrow Y) = P(X|Y) - P(X|\overline{Y})$

ค่าที่เป็นไปได้ของค่าความเชื่อมั่นใหม่นั้นอยู่ในพิสัย [-1, 1] ถ้าค่าความเชื่อมั่นใหม่มีค่าเท่ากับ 0 หมายความว่าเซตรายการที่มาก่อนและเซตรายการที่ตามมาของกฎความสัมพันธ์เป็นอิสระต่อกัน ถ้าค่าความเชื่อมั่นใหม่น้อยกว่า 0 แสดงว่าทั้งเซตรายการที่มาก่อนและเซตรายการที่ตามมาแปรผกผันกันหรือสหสัมพันธ์เชิงลบ และถ้าค่าความเชื่อมั่นใหม่มากกว่า 0 แสดงว่าทั้งเซตรายการที่มาก่อนและเซตรายการที่ตามมาแปรผันตามกันหรือสหสัมพันธ์เชิงบวก (Liu et al., 2008)

ในงานวิจัยของ Liu และคณะ (Liu et al., 2008) ที่ได้เสนอค่าความเชื่อมั่นใหม่ข้างต้นนั้น Liu และคณะได้ทำการเปรียบเทียบความสามารถของค่าความเชื่อมั่นใหม่กับค่าประเมินความน่าสนใจของกฎความสัมพันธ์อื่นๆ ทั้งหมด 8 ค่าคือ ค่าสนับสนุน (Support), ค่าความเชื่อมั่น (Confidence), ค่าคอนวิคชั่น (Conviction), ค่าลิฟท์ (Lift), ค่าเลฟเวอเรจ (Leverage), ค่าคัฟเวอเรจ (Coverage), ค่าสหสัมพันธ์ (Correlation) และ ค่าอัตราส่วนออดส์ (Odds Ratio) โดยใช้ฐานข้อมูลทรานแซคชั่นสมมุติขนาด 10 ทรานแซคชั่น ผลของการเปรียบเทียบคือ ค่าความเชื่อมั่นใหม่สามารถบ่งบอกทิศของความสัมพันธ์ได้อย่างถูกต้องและสอดคล้องกับค่าเลฟเวอเรจและค่าสหสัมพันธ์ แต่ค่าความเชื่อมั่นใหม่นั้นสามารถระบุความแตกต่างของความน่าสนใจของกฎความสัมพันธ์ 2 กฎความสัมพันธ์ใดๆที่ค่าเลฟเวอเรจและค่าสหสัมพันธ์ไม่สามารถระบุได้ (กล่าวคือกฎความสัมพันธ์ 2 กฎที่คำนวณค่าค่าเลฟเวอเรจหรือค่าสหสัมพันธ์ได้เท่ากันทั้ง 2 กฎ แต่ค่าความเชื่อมั่นใหม่ให้ค่าที่แตกต่างกันระหว่าง 2 กฎ) ส่วนค่าประเมินความน่าสนใจของกฎความสัมพันธ์อื่นๆให้ค่าที่ขัดแย้งกับค่าสหพันธ์ (กล่าวคือเซตรายการที่มาก่อนและเซตรายการที่ตามมาของกฎความสัมพันธ์นั้นมีความสัมพันธ์เชิงลบต่อกันแต่กลับให้ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ที่สูงออกมา)

เนื่องจากการเปรียบเทียบค่าความเชื่อมั่นใหม่กับค่าอื่นๆในงานวิจัยของ Liu และคณะ (Liu et al., 2008) เป็นการทดสอบกับฐานข้อมูลทรานแซคชั่นสมมุติขนาด 10 ทรานแซคชั่นเท่านั้น จึงค่อนข้างขาดความน่าเชื่อถือว่าค่าความเชื่อมั่นใหม่จะให้ประสิทธิภาพในการค้นหากฎความสัมพันธ์ที่ดีกว่าค่าอื่นๆจริง ผู้วิจัยจึงรวบรวมวรรณกรรมที่เกี่ยวข้องกับคุณสมบัติที่ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ควรจะมีในหัวข้อ 2.5 ทั้งหมด 16 คุณสมบัติ และทำการเปรียบเทียบคุณสมบัติที่ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ทั้ง 8 ค่ารวมถึงค่าความเชื่อมั่นใหม่มี

2.5 คุณสมบัติที่ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ควรจะมี

ในปีค.ศ. 1991 Piatetsky-Shapiro และคณะ (Piatetsky-Shapiro, 1991) ได้เสนอคุณสมบัติ 3 ประการสำคัญที่ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะมี ดังนี้

กำหนดให้ M คือ ค่าประเมินความน่าสนใจของกฎความสัมพันธ์

$X \rightarrow Y$ คือ กฎความสัมพันธ์ที่มีเซตรายการที่มาก่อนคือเซต X และเซตรายการที่ตามมาคือเซต Y

P(X) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

P(X and Y) คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และ Y ในฐานข้อมูล

1. ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ต้องเท่ากับ 0 เมื่อเซตรายการที่มาก่อนและเซตรายการที่ตามมาของกฎความสัมพันธ์เป็นอิสระต่อกันทางสถิติ (M = 0 ถ้า P(X and Y) = P(X)P(Y))
2. ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรต้องเพิ่มขึ้นเมื่อ P(X and Y) เพิ่มขึ้นในขณะที่ P(X) และ P(Y) คงที่
3. ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรต้องเพิ่มขึ้นเมื่อ P(X) (หรือ P(Y)) ลดลงในขณะที่ P(X and Y) และ P(Y) (หรือ P(X)) คงที่

Piatetsky-Shapiro และคณะ (Piatetsky-Shapiro, 1991) เสนอคุณสมบัติของค่าประเมินความน่าสนใจของกฎความสัมพันธ์ทั้ง 3 คุณสมบัติขึ้นมาให้มีความเป็นทั่วไป (General) มากที่สุด โดยในทั้ง 3 คุณสมบัติจะอ้างอิงถึงพารามิเตอร์เพียง 3 ตัวคือ P(X) P(Y) และ P(X and Y) ซึ่งเป็นค่าทางสถิติที่เกี่ยวข้องกับกฎความสัมพันธ์เท่านั้น (Freitas, 1999) ในงานวิจัยนี้ผู้วิจัยจะใช้สัญลักษณ์ P1 P2 และ P3 แทนการอ้างถึงคุณสมบัติที่ 1 2 และ 3 ของ Piatetsky-Shapiro และคณะ (Piatetsky-Shapiro, 1991)

คุณสมบัติ P1 นั้นเป็นคุณสมบัติที่อธิบายว่าถ้าเซตรายการที่มาก่อนและเซตรายการที่ตามมานั้นเป็นอิสระต่อกันแล้วกฎความสัมพันธ์นั้นก็ควรจะต้องไม่มีความน่าสนใจเลยหรือมีค่าประเมินความน่าสนใจของกฎความสัมพันธ์นั้นเท่ากับ 0 คณะวิจัยหลายคณะวิจารณ์ว่าคุณสมบัติข้อนี้นั้นไม่ค่อยยืดหยุ่นและจำกัดมากเกินไป (Tan et al, 2002) ในปี ค.ศ. 2002 Tan และคณะ (Tan et al, 2002) จึงเสนอให้ผ่อนคลายคุณสมบัติข้อนี้ใหม่เป็น ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ต้องเท่ากับ k เมื่อเซตรายการที่มาก่อนและเซตรายการที่ตามมาของกฎความสัมพันธ์เป็นอิสระต่อกันทางสถิติ (M = k ถ้า P(X and Y) = P(X)P(Y)) โดยที่ k คือค่าคงที่ แต่อย่างไรก็ตามงานวิจัยต่างๆ (Freitas , 1999; Major and Mangano, 1995; Mcgarry, 2005; Geng and Hamilton, 2006; Liu et al., 2008; Heravi, 2009) ที่นำคุณสมบัติทั้ง 3 ข้อของ Piatetsky-Shapiro และคณะไปใช้หรืออ้างอิงถึงก็ยังคงอ้างอิงตามต้นฉบับคือค่าประเมินความน่าสนใจของกฎความสัมพันธ์ควรจะเท่ากับ 0 เมื่อกฎนั้นไม่มีความน่าสนใจเลย คุณสมบัติ P2 นั้น

เป็นคุณสมบัติที่อธิบายว่าในขณะที่ค่าสนับสนุนของเซตรายการที่มาก่อนและค่าสนับสนุนของเซตรายการที่ตามมาคงที่ ถ้าค่าสนับสนุนของทั้งกฎความสัมพันธ์เพิ่มขึ้น ความน่าสนใจของกฎความสัมพันธ์นั้นก็ควรจะเพิ่มขึ้นตามด้วย กล่าวคือถ้ากฎความสัมพันธ์มีความสัมพันธ์เชิงบวกกันมากขึ้น กฎความสัมพันธ์นั้นก็ควรจะน่าสนใจเพิ่มขึ้น และคุณสมบัติ P3 นั้นเป็นคุณสมบัติที่อธิบายว่าในขณะที่ค่าสนับสนุนของกฎความสัมพันธ์และค่าสนับสนุนของเซตรายการที่มาก่อน (หรือ ค่าสนับสนุนของเซตรายการที่ตามมา) คงที่ ถ้าค่าสนับสนุนของเซตรายการที่ตามมา (หรือค่าสนับสนุนของเซตรายการที่มาก่อน) ลดลงแล้ว ความน่าสนใจของกฎความสัมพันธ์นั้นควรจะเพิ่มขึ้น (Piatetsky-Shapiro, 1991; Geng and Hamilton, 2006)

ในปีค.ศ. 1995 Major และ Mangano (Major and Mangano, 1995) เสนอให้นำคุณสมบัติทั้ง 3 ข้อของ Piatetsky-Shapiro และคณะ (Piatetsky-Shapiro, 1991) พร้อมกับแนะนำให้เพิ่มอีก 1 คุณสมบัติเข้าไปดังนี้

กำหนดให้ M คือ ค่าประเมินความน่าสนใจของกฎความสัมพันธ์

$X \rightarrow Y$ คือ กฎความสัมพันธ์ที่มีเซตรายการที่มาก่อนคือเซต X และเซตรายการที่ตามมาคือเซต Y

$P(X)$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

$P(\overline{X})$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่ไม่มีรายการ X ในฐานข้อมูล

$Conf(X \rightarrow Y)$ คือ ค่าความเชื่อมั่นของกฎความสัมพันธ์ $X \rightarrow Y$

1. ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรต้องเพิ่มขึ้นเมื่อ P(X) เพิ่มขึ้น ในขณะที่ P(Y) $P(\overline{Y})$ และ $Conf(X \rightarrow Y)$ คงที่

คุณสมบัติของนี้มักจะถูกอ้างถึงและนำไปใช้พร้อมๆกับคุณสมบัติทั้ง 3 ข้อของ Piatetsky-Shapiro และคณะ (Piatetsky-Shapiro, 1991) ดังนั้นคุณสมบัตินี้มักจะถูกเรียกว่าคุณสมบัติของที่ 4 ที่ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ควรจะมี ในงานวิจัยนี้ผู้วิจัยจะใช้สัญลักษณ์ P4 แทนการอ้างถึงคุณสมบัติข้อนี้ของ Major และ Mangano (Major and Mangano, 1995)

คุณสมบัติ P4 นั้นเป็นคุณสมบัติที่อธิบายว่าในขณะที่ค่าความเชื่อมั่นของกฎความสัมพันธ์มีค่าคงที่ เมื่อค่าสนับสนุนของของเซตรายการที่มาก่อนเพิ่มขึ้นแล้ว ค่าประเมินความ

น่าสนใจของกฎความสัมพันธ์ M ก็ควรจะเพิ่มขึ้นตามด้วย ในขณะที่ P(Y) P($\overline{Y}$) และ P(Y|X) (หรือ Conf(X→Y) นั่นเอง) คงที่

ในปีค.ศ. 2002 Tan และคณะ (Tan et al, 2002) เสนอคุณสมบัติอีก 5 ประการที่ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะมี เพิ่มเติมจาก 4 ข้อของคณะวิจัย Piatetsky-Shapiro (Piatetsky-Shapiro, 1991) และ Major กับ Mangano (Major and Mangano, 1995) โดยที่คุณสมบัติ 5 ประการของ Tan และคณะนั้นเป็นคุณสมบัติที่อยู่บนฐานของการดำเนินการบนตารางหลายตัวแปร (Contingency Table) ขนาด 2 x 2 ดังนี้

ตารางที่ 2-3 แสดงตารางหลายตัวแปร (Contingency Table) ขนาด 2 x 2 ของกฎความสัมพันธ์ X→Y

	Y	$\overline{Y}$	
X	n(X and Y)	n(X and $\overline{Y}$)	n(X)
$\overline{X}$	n($\overline{X}$ and Y)	n($\overline{X}$ and $\overline{Y}$)	n($\overline{X}$)
	n(Y)	n($\overline{Y}$)	N

กำหนดให้ M คือ ค่าประเมินความน่าสนใจของกฎความสัมพันธ์

X→Y คือ กฎความสัมพันธ์ที่มีเซตรายการที่มาก่อนคือเซต X และเซตรายการที่ตามมาคือเซต Y

N คือ จำนวนของทรานแซคชั่นทั้งหมดในฐานข้อมูล

n(X) คือ จำนวนของทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

n($\overline{X}$) คือ จำนวนของทรานแซคชั่นที่ไม่มีรายการ X ในฐานข้อมูล

n(X and Y) คือ จำนวนของทรานแซคชั่นที่มีรายการ X และ Y ในฐานข้อมูล

1. ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรมีคุณสมบัติสมมาตรภายในการเปลี่ยนแปลงลำดับ
2. ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรคงที่เมื่อมีการขยายตามแถวหรือขยายตามหลัก

3. ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรสามารถบ่งชี้ทิศทางของความสัมพันธ์ได้
4. ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรคงที่เมื่อมีการดำเนินการทั้งตามแถวและตามหลักพร้อมกัน
5. ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M จะต้องไม่มีความสัมพันธ์กับจำนวนของทรานแซคชั่นที่มีเซตรายการ X เซตรายการ Y หรือทั้งคู่

ในงานวิจัยนี้ผู้วิจัยจะใช้สัญลักษณ์ O1 O2 O3 O4 และ O5 แทนการอ้างถึงคุณสมบัติที่ 1 2 3 4 และ 5 ของ Tan และคณะ (Tan et al, 2002)

คุณสมบัติ O1 นั้นเป็นคุณสมบัติที่อธิบายว่ากฎความสัมพันธ์ $X \rightarrow Y$ และกฎความสัมพันธ์ $Y \rightarrow X$ ควรมีค่าประเมินความน่าสนใจของกฎความสัมพันธ์ที่เท่ากัน Tan และคณะ (Tan et al, 2002) ผู้เสนอคุณสมบัติข้อนี้กล่าวว่าค่าประเมินความน่าสนใจของกฎความสัมพันธ์ที่ไม่แสดงคุณสมบัติสมมาตรสามารถแก้ไขให้มีคุณสมบัติสมมาตรได้โดยการกำหนดให้ค่า M ของกฎความสัมพันธ์ $X \rightarrow Y$ และค่า M ของกฎความสัมพันธ์ $Y \rightarrow X$ มีค่าเท่ากับค่าใดค่าหนึ่งที่มากกว่าหรือเท่ากับ $\max(M(X \rightarrow Y), M(Y \rightarrow X))$ นั่นเอง คุณสมบัติข้อนี้ถูกปฏิเสธว่าไม่เป็นจริงในหลายงานวิจัย (Geng and Hamilton, 2006) คุณสมบัติ O2 นั้นเป็นคุณสมบัติที่อธิบายว่าค่าการประเมินความน่าสนใจของกฎความสัมพันธ์ $X \rightarrow Y$ เมื่อตอนที่ n(X and Y) มีค่าเท่ากับ Z ควรจะเท่ากับค่าการประเมินความน่าสนใจของกฎความสัมพันธ์ $X \rightarrow Y$ เมื่อตอนที่ n(X and Y) มีค่าเท่ากับ k_1k_2Z โดยที่ k_1, k_2 เป็นค่าคงที่บวกที่มาขยายตามแถวและตามหลักตามลำดับ เป็นต้น คุณสมบัติ O3 นั้นเป็นคุณสมบัติที่อธิบายว่าค่าการประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะต้องบ่งบอกได้ว่าเซตรายการที่มาก่อนและเซตรายการที่ตามมามีความสัมพันธ์เชิงบวกหรือความสัมพันธ์เชิงลบต่อกันได้ คุณสมบัติ O4 นั้นเป็นคุณสมบัติที่อธิบายว่าค่าการประเมินความน่าสนใจของกฎความสัมพันธ์ $X \rightarrow Y$ ควรเท่ากับค่าการประเมินความน่าสนใจของกฎความสัมพันธ์ $\bar{X} \rightarrow \bar{Y}$ ด้วย และคุณสมบัติ O5 นั้นเป็นคุณสมบัติที่อธิบายว่าค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M จะต้องไม่เปลี่ยนแปลงเมื่อ n($\bar{X}$ and $\bar{Y}$) เปลี่ยนแปลง กล่าวคือค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M จะต้องไม่มีความสัมพันธ์กับจำนวนของทรานแซคชั่นที่ไม่มีทั้งเซตรายการที่มาก่อนและเซตรายการที่ตาม (Tan et al, 2002)

ในปีค.ศ. 2004 Lenca และคณะ (Lenca et al, 2004) ก็เสนอคุณสมบัติ 5 ข้อที่ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะมีเช่นกัน หลังจากนั้นในปี ค.ศ. 2007 Lenca และคณะได้ปรับปรุงคุณสมบัติทั้ง 5 ข้อนั้นเล็กน้อยเพื่อให้มีความยืดหยุ่นมากขึ้น (Lenca et al, 2007) ดังนี้

กำหนดให้ M คือ ค่าประเมินความน่าสนใจของกฎความสัมพันธ์

$X \rightarrow Y$ คือ กฎความสัมพันธ์ที่มีเซตรายการที่มาก่อนคือเซต X และเซตรายการที่ตามมาคือเซต Y

$P(X)$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

$P(\overline{X})$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่ไม่มีรายการ X ในฐานข้อมูล

$P(X \text{ and } Y)$ คือ ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และ Y ในฐานข้อมูล

N คือ จำนวนของทรานแซคชั่นทั้งหมดในฐานข้อมูล

1. ถ้า $P(X \text{ and } \overline{Y}) = 0$ แล้ว ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะเป็นค่าคงที่หรือเป็นอนันต์ (infinity)
2. ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะลดลงแบบเส้นตรง แบบพาราโบล่าหงาย หรือแบบพาราโบล่าคว่ำ เมื่อ $P(X \text{ and } \overline{Y})$ มีค่าเพิ่มขึ้น
3. ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะเพิ่มขึ้นเมื่อ N เพิ่มขึ้นในขณะที่ P(X and Y) P(X) และ P(Y) คงที่
4. ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ที่ดีควรจะต้องสามารถหาค่าเธรสโฮลด์ (threshold) ที่แบ่งแยกระหว่างกฎความสัมพันธ์ที่น่าสนใจออกจากกฎความสัมพันธ์ที่ไม่น่าสนใจได้ง่าย
5. ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ที่ดีควรจะต้องมีอรรถศาสตร์ (semantics) ที่ผู้ใช้สามารถเข้าใจได้ง่าย

ในงานวิจัยนี้ผู้วิจัยจะใช้สัญลักษณ์ Q1 Q2 Q3 Q4 และ Q5 แทนการอ้างถึงคุณสมบัติที่ 1 2 3 4 และ 5 ของ Lenca และคณะ (Lenca et al, 2004; Lenca et al, 2007)

คุณสมบัติ Q1 นั้นเป็นคุณสมบัติที่อธิบายว่าเมื่อกฎความสัมพันธ์นั้นมีความน่าสนใจสูงที่สุด หรือค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X แต่ไม่มีรายการ Y ในฐานข้อมูลเท่ากับ 0 (หรือ ค่าความเชื่อมั่นเท่ากับ 1 นั่นเอง) แล้วค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะเป็นค่าคงที่ค่าใดค่าหนึ่งหรือเป็นค่าอนันต์เพื่อสื่อความหมายอย่างชัดเจนว่ากฎความสัมพันธ์นั้นน่าสนใจสูงที่สุด คุณสมบัติ Q2 นั้นเป็นคุณสมบัติที่อธิบายว่าค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะต้องมีการลงลดเมื่อมีทรานแซคชั่นที่มีรายการ X แต่ไม่มีรายการ Y ถูกเพิ่มเข้ามาในฐานข้อมูล ลักษณะของการลดลงจะเป็นอย่างไรนั้นขึ้นอยู่กับวัตถุประสงค์ของงานที่นำไปประยุกต์ใช้ ตัวอย่างเช่น ถ้าต้องการให้ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M มีความคงทน (Tolerated) ต่อการเพิ่มขึ้นของทรานแซคชั่นที่มีรายการ X แต่ไม่มีรายการ Y (กล่าวคือเมื่อมีทรานแซคชั่นที่มีรายการ X แต่ไม่มีรายการ Y เพิ่มขึ้น ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรลดลงทีละเพียงเล็กน้อยเท่านั้น) ก็ควรกำหนดให้ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ลดลงแบบพาราโบล่าหงาย (Concave up) เมื่อมีทรานแซคชั่นที่มีรายการ X แต่ไม่มีรายการ Y เพิ่มขึ้น แต่ถ้าต้องการให้ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M มีความอ่อนไหวมาก (กล่าวคือถ้ามีทรานแซคชั่นที่มีรายการ X แต่ไม่มีรายการ Y เพิ่มขึ้นมาเพียงเล็กน้อยก็จะทำให้ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ลดลงอย่างรวดเร็ว) ก็ควรกำหนดให้ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ลดลงแบบพาราโบล่าคว่ำ (Concave down, Convex) เมื่อมีทรานแซคชั่นที่มีรายการ X แต่ไม่มีรายการ Y เพิ่มขึ้น เป็นต้น คุณสมบัติ Q3 นั้นเป็นคุณสมบัติที่อธิบายว่าค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะเพิ่มขึ้นเมื่อจำนวนของทรานแซคชั่นทั้งหมดเพิ่มขึ้นในขณะที่ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และ Y ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ Y คงที่ คุณสมบัติ Q4 นั้นเป็นคุณสมบัติที่อธิบายว่าค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะต้องสามารถหาค่าเธรสโชลด์ (Threshold) ที่เป็นจุดแบ่งแยกระหว่างกฎความสัมพันธ์ที่น่าสนใจกับกฎความสัมพันธ์ที่ไม่น่าสนใจได้ง่าย และค่าเธรสโชลด์นั้นจะต้องสื่อความหมายที่ผู้ใช้สามารถเข้าใจได้ง่ายด้วย กล่าวคือค่าเธรสโชลด์นั้นควรเป็นค่ากลางระหว่างค่าสูงสุดกับค่าต่ำสุดนั่นเอง คุณสมบัติ Q5 นั้นเป็นคุณสมบัติที่อธิบายว่าค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ที่ดีควรจะต้องสื่อความหมายที่ผู้ใช้สามารถเข้าใจได้ง่าย กล่าวคือเมื่อผู้ใช้อ่านค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M แล้วจะต้องสามารถเข้าใจได้โดยง่ายว่าค่านี้หมายถึงอะไร โดยไม่ต้องอธิบายใดๆเพิ่มเติม (Lenca et al, 2004; Lenca et al, 2007)

ในปีค.ศ. 2006 Geng และ Hamilton (Geng and Hamilton, 2006) ก็เสนอคุณสมบัติ 2 ข้อที่ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะมีเช่นกัน แต่คุณสมบัติทั้ง 2 ข้อของ Geng และ Hamilton (Geng and Hamilton, 2006) นี้จะเน้นประเมินที่ความสัมพันธ์ระหว่างค่าประเมินความน่าสนใจของกฎความสัมพันธ์ ค่าสนับสนุน และค่าความเชื่อมั่น ดังนี้

กำหนดให้ M คือ ค่าประเมินความน่าสนใจของกฎความสัมพันธ์

X→Y คือ กฎความสัมพันธ์ที่มีเซตรายการที่มาก่อนคือเซต X และเซตรายการที่ตามมาคือเซต Y

Sup(X→Y) คือ ค่าสนับสนุนของกฎความสัมพันธ์ X→Y

Conf(X→Y) คือ ค่าความเชื่อมั่นของกฎความสัมพันธ์ X→Y

N คือ จำนวนของทรานแซคชั่นทั้งหมดในฐานข้อมูล

n(X) คือ จำนวนของทรานแซคชั่นที่มีรายการ X ในฐานข้อมูล

n($\overline{X}$) คือ จำนวนของทรานแซคชั่นที่ไม่มีรายการ X ในฐานข้อมูล

n(X and Y) คือ จำนวนของทรานแซคชั่นที่มีรายการ X และ Y ในฐานข้อมูล

1. ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรอยู่ในรูปของฟังก์ชั่นที่ขึ้นกับค่าสนับสนุน (*f*(Sup(X→Y))) โดยที่ค่าของฟังก์ชั่นนี้ควรเพิ่มขึ้นเมื่อค่าสนับสนุนเพิ่มขึ้น ในขณะที่ขอบ (Margins) ของตารางหลายตัวแปร (ตารางที่ 2-3) คงที่
2. ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรอยู่ในรูปของฟังก์ชั่นที่ขึ้นกับค่าความเชื่อมั่น (*f*(Conf(X→Y))) โดยที่ค่าของฟังก์ชั่นนี้ควรเพิ่มขึ้นเมื่อค่าความเชื่อมั่นเพิ่มขึ้น ในขณะที่ขอบ (Margins) ของตารางหลายตัวแปร (ตารางที่ 2-3) คงที่

ในงานวิจัยนี้ผู้วิจัยจะใช้สัญลักษณ์ S1 และ S2 แทนการอ้างถึงคุณสมบัติที่ 1 และ 2 ของ Geng และ Hamilton (Geng and Hamilton, 2006)

สมมุติให้ n(X) = a, n($\overline{X}$) = N – a, n(Y) = b, n($\overline{Y}$) = N – b, Sup(X→Y) = x และ Conf(X→Y) = y คุณสมบัติ S1 นั้นเป็นคุณสมบัติที่อธิบายว่า เมื่อค่า n(X) n($\overline{X}$) n(Y) และ n($\overline{Y}$) คงที่จะได้ว่า P(X and Y) = x, P($\overline{X}$ and Y) = ($\frac{b}{N}$) – x, P(X and $\overline{Y}$) = ($\frac{a}{N}$)) – x และ P($\overline{X}$ and $\overline{Y}$) = 1 – $\frac{a+b}{N}$) + x แล้วค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ที่เขียนให้อยู่ในรูปของฟังก์ชั่นที่ขึ้นกับ x นั้นควรจะเพิ่มขึ้นเมื่อ x เพิ่มขึ้น และคุณสมบัติ S2 ก็เช่นเดียวกับคุณสมบัติ S1

คือเป็นคุณสมบัติที่อธิบายว่า เมื่อค่า n(X) n($\overline{X}$) n(Y) และ n($\overline{Y}$) คงที่จะได้ว่า P(X and Y) = $\frac{ay}{N}$, P($\overline{X}$ and Y) = $\frac{(b-ay)}{N}$, P(X and $\overline{Y}$) = $\frac{a(1-y)}{N}$ และ P($\overline{X}$ and $\overline{Y}$) = 1 – $\frac{(a+b)}{N}$ + $\frac{ay}{N}$ แล้วค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ที่เขียนให้อยู่ในรูปของฟังก์ชั่นที่ขึ้นกับ y นั้นควรจะเพิ่มขึ้นเมื่อ y เพิ่มขึ้น (Geng and Hamilton, 2006)

คุณสมบัติ S1 สอดคล้องโดยตรงกับคุณสมบัติ P2 (Heravi, 2009) และคุณสมบัติ S2 สอดคล้องโดยตรงกับคุณสมบัติ Q2 (Geng and Hamilton, 2006)

คุณสมบัติที่ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ควรจะมีทั้งหมด 16 ข้อข้างต้น คุณสมบัติที่ได้รับการยอมรับและถูกอ้างอิงถึงโดยงานวิจัยต่างๆ (Freitas , 1999; Major and Mangano, 1995; Mcgarry, 2005; Geng and Hamilton, 2006; Liu et al., 2008; Heravi, 2009) มากที่สุดคือคุณสมบัติ P1 P2 และ P3 ของ Piatetsky-Shapiro และคณะ (Piatetsky-Shapiro, 1991) และคุณสมบัติ P4 ของ Mango และ Mangano (Major and Mangano, 1995)

จากหัวข้อที่ 2.3 ผู้วิจัยได้ทบทวนวรรณกรรมของค่าประเมินความน่าสนใจของกฎความสัมพันธ์ทั้งหมด 8 ค่าคือ ค่าสนับสนุน (Support), ค่าความเชื่อมั่น (Confidence), ค่าคอนวิคชั่น (Conviction), ค่าลิฟท์ (Lift), ค่าเลฟเวอเรจ (Leverage), ค่าคัฟเวอเรจ (Coverage), ค่าสหสัมพันธ์ (Correlation) และ ค่าอัตราส่วนออดส์ (Odds Ratio) และหัวข้อ 2.4 ที่ผู้วิจัยทบทวนวรรณกรรมของค่าประเมินความน่าสนใจของกฎความสัมพันธ์ใหม่ที่ชื่อว่า ค่าความเชื่อมั่นใหม่ (New Confidence) ที่ถูกเสนอขึ้นโดย Liu และคณะ (Liu et al., 2008) ผู้วิจัยจึงรวบรวมงานวิจัยที่ทำการทดสอบคุณสมบัติทั้ง 14 คุณสมบัติกับค่าประเมินความน่าสนใจของกฎความสัมพันธ์ทั้งหมด 8 ค่าและค่าความเชื่อมั่นใหม่ การเปรียบเทียบคุณสมบัติของค่าประเมินความน่าสนใจของกฎความสัมพันธ์ 8 ค่าที่รวบรวมโดยคณะวิจัยของ Geng กับ Hamilton ในปี 2006 และคณะวิจัยของ Heravi ในปี 2009 (Geng and Hamilton, 2006; Heravi, 2009) และค่าความเชื่อมั่นใหม่ที่รวบรวมโดยผู้วิจัยแสดงดังตารางที่ 2-4 ในบทที่ 1

เนื่องจากงานวิจัยของ Liu และคณะในปี 2008 (Liu et al., 2008) ได้ทำการพิสูจน์คุณสมบัติค่าความเชื่อมั่นใหม่ไว้ทั้งหมดเพียง 5 คุณสมบัติคือ คุณสมบัติ P1 P2 P3 O1 และ O2 เท่านั้น ดังนั้นผู้วิจัยจึงพิสูจน์คุณสมบัติ P4 O3 O4 O5 Q1 Q2 Q3 S1 และ S2 ของค่าความเชื่อมั่นใหม่ดังนี้

<u>พิสูจน์คุณสมบัติ P4</u>

คุณสมบัติ P4 นั้นเป็นคุณสมบัติที่อธิบายว่าในขณะที่ค่าความเชื่อมั่นของกฎความสัมพันธ์มีค่าคงที่ เมื่อค่าสนับสนุนของของเซตรายการที่มาก่อนเพิ่มขึ้นแล้ว ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ก็ควรจะเพิ่มขึ้นตามด้วย ในขณะที่ P(Y) P($\overline{Y}$) และ P(Y|X) (หรือ Conf(X→Y) นั่นเอง) คงที่

จาก $NConf(X \rightarrow Y) = P(X|Y) - P(X|\overline{Y})$

$$NConf(X \rightarrow Y) = \frac{P(X \text{ and } Y)}{P(Y)} - \frac{P(X \text{ and } \overline{Y})}{P(\overline{Y})}$$

$$NConf(X \rightarrow Y) = \frac{P(X \text{ and } Y)}{P(Y)} - \frac{P(\overline{Y}|X)P(X)}{P(\overline{Y})}$$

(เนื่องจาก $P(X \text{ and } \overline{Y}) = P(\overline{Y}|X)P(X)$)

$$NConf(X \rightarrow Y) = \frac{P(Y|X)P(X)}{P(Y)} - \frac{P(\overline{Y}|X)P(X)}{P(\overline{Y})}$$

(เนื่องจาก $P(X \text{ and } Y) = P(Y|X)P(X)$)

$$NConf(X \rightarrow Y) = P(X)\left(\frac{P(Y|X)}{P(Y)} - \frac{P(\overline{Y}|X)}{P(\overline{Y})}\right)$$

จากข้อกำหนด P(Y) P($\overline{Y}$) และ P(Y|X) คงที่

กรณีที่พจน์ $\left(\frac{P(Y|X)}{P(Y)} - \frac{P(\overline{Y}|X)}{P(\overline{Y})}\right)$ มีค่าเป็นบวก จะทำให้ $NConf(X \rightarrow Y)$ มีค่าเพิ่มขึ้นเมื่อ P(X) เพิ่มขึ้น

กรณีที่พจน์ $\left(\frac{P(Y|X)}{P(Y)} - \frac{P(\overline{Y}|X)}{P(\overline{Y})}\right)$ มีค่าเป็นลบ จะทำให้ $NConf(X \rightarrow Y)$ มีค่าลดลงเมื่อ P(X) ลดลง

ดังนั้น ค่าความเชื่อมั่นใหม่ไม่มีคุณสมบัติ P4

<u>พิสูจน์คุณสมบัติ O3</u>

คุณสมบัติ O3 นั้นเป็นคุณสมบัติที่อธิบายว่าค่าการประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะต้องบ่งบอกได้ว่าเซตรายการที่มาก่อนและเซตรายการที่ตามมามีความสัมพันธ์เชิงบวกหรือความสัมพันธ์เชิงลบต่อกันได้

เนื่องจาก ค่าที่เป็นไปได้ของค่าความเชื่อมั่นใหม่นั้นอยู่ในพิสัย [-1, 1] ถ้าค่าความเชื่อมั่นใหม่มีค่าเท่ากับ 0 หมายความว่าเซตรายการที่มาก่อนและเซตรายการที่ตามมาของกฎความสัมพันธ์เป็นอิสระต่อกัน ถ้าค่าความเชื่อมั่นใหม่น้อยกว่า 0 แสดงว่าทั้งเซตรายการที่มาก่อนและเซตรายการที่ตามมาแปรผกผันกันหรือสหสัมพันธ์เชิงลบ และถ้าค่าความเชื่อมั่นใหม่มากกว่า 0 แสดงว่าทั้งเซตรายการที่มาก่อนและเซตรายการที่ตามมาแปรผันตามกันหรือสหสัมพันธ์เชิงบวก (Liu et al., 2008)

ดังนั้นค่าความเชื่อมั่นใหม่มีคุณสมบัติ O3

<u>พิสูจน์คุณสมบัติ O4</u>

คุณสมบัติ O4 นั้นเป็นคุณสมบัติที่อธิบายว่าค่าการประเมินความน่าสนใจของกฎความสัมพันธ์ $X \rightarrow Y$ ควรเท่ากับค่าการประเมินความน่าสนใจของกฎความสัมพันธ์ $\overline{X} \rightarrow \overline{Y}$

จาก $\text{NConf}(X \rightarrow Y) = P(X|Y) - P(X|\overline{Y})$

$$\text{NConf}(X \rightarrow Y) = \frac{P(X \text{ and } Y)}{P(Y)} - \frac{P(X \text{ and } \overline{Y})}{P(\overline{Y})}$$

และ $\text{NConf}(\overline{X} \rightarrow \overline{Y}) = P(\overline{X}|\overline{Y}) - P(\overline{X}|Y)$

$$\text{NConf}(\overline{X} \rightarrow \overline{Y}) = \frac{P(\overline{X} \text{ and } \overline{Y})}{P(\overline{Y})} - \frac{P(\overline{X} \text{ and } Y)}{P(Y)}$$

ต้องการพิสูจน์ว่า $\text{NConf}(X \rightarrow Y) = \text{NConf}(\overline{X} \rightarrow \overline{Y})$

$$\frac{P(X \text{ and } Y)}{P(Y)} - \frac{P(X \text{ and } \overline{Y})}{P(\overline{Y})} = \frac{P(\overline{X} \text{ and } \overline{Y})}{P(\overline{Y})} - \frac{P(\overline{X} \text{ and } Y)}{P(Y)}$$

คูณ $P(Y)P(\overline{Y})$ ทั้ง 2 ข้าง

$$P(\overline{Y})P(X \text{ and } Y) - P(Y)P(X \text{ and } \overline{Y}) = P(Y)P(\overline{X} \text{ and } \overline{Y}) - P(\overline{Y})P(\overline{X} \text{ and } Y)$$

เนื่องจาก $P(X \text{ and } \overline{Y}) = P(X) - P(X \text{ and } Y)$

$$P(\overline{Y})P(X \text{ and } Y) - P(Y)(P(X) - P(X \text{ and } Y)) = P(Y)P(\overline{X} \text{ and } \overline{Y}) - P(\overline{Y})P(\overline{X} \text{ and } Y)$$

เนื่องจาก $P(X \text{ and } \overline{Y}) = P(Y) - P(X \text{ and } Y)$

$$P(\overline{Y})P(X \text{ and } Y) - P(Y)(P(X) - P(X \text{ and } Y)) = P(Y)P(\overline{X} \text{ and } \overline{Y}) - P(\overline{Y})(P(Y) - P(X \text{ and } Y))$$

เนื่องจาก $P(\overline{X} \text{ and } \overline{Y}) = 1 - P(X) - P(Y) + P(X \text{ and } Y)$

$$P(\overline{Y})P(X \text{ and } Y) - P(Y)(P(X) - P(X \text{ and } Y)) = P(Y)(1 - P(X) - P(Y) + P(X \text{ and } Y)) - P(\overline{Y})(P(Y) - P(X \text{ and } Y))$$

$$P(\overline{Y})P(X \text{ and } Y) - P(X)P(Y) + P(Y)P(X \text{ and } Y) = P(Y) - P(X)P(Y) - P(Y)P(Y) + P(Y)P(X \text{ and } Y) - P(\overline{Y})(P(Y) - P(X \text{ and } Y))$$

เนื่องจาก $P(\overline{Y}) = 1 - P(Y)$

$$(1 - P(Y))P(X \text{ and } Y) - P(X)P(Y) + P(Y)P(X \text{ and } Y) = P(Y) - P(X)P(Y) - P(Y)P(Y) + P(Y)P(X \text{ and } Y) - (1 - P(Y))(P(Y) - P(X \text{ and } Y))$$

$$P(X \text{ and } Y) - P(X)P(Y) = P(Y) - P(X)P(Y) - P(Y) + P(X \text{ and } Y)$$

$$1 = 1 \quad \text{เป็นจริง}$$

ดังนั้นค่าความเชื่อมั่นใหม่มีคุณสมบัติ O4

<u>พิสูจน์คุณสมบัติ O5</u>

คุณสมบัติ O5 นั้นเป็นคุณสมบัติที่อธิบายว่าค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M จะต้องไม่เปลี่ยนแปลงเมื่อ $n(\overline{X} \text{ and } \overline{Y})$ เปลี่ยนแปลง กล่าวคือค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะคำนวณมาจาก n(X) n(Y) หรือ/และ n(X and Y) เท่านั้น (ไม่มี $n(\overline{X} \text{ and } \overline{Y})$ หรือ N มาเกี่ยวข้อง)

จาก $\text{NConf}(X \rightarrow Y) = P(X|Y) - P(X|\overline{Y})$

$$\text{NConf}(X \rightarrow Y) = \frac{P(X \text{ and } Y)}{P(Y)} - \frac{P(X \text{ and } \overline{Y})}{P(\overline{Y})}$$

$$\text{NConf}(X \rightarrow Y) = \frac{n(\text{X and Y})\text{N}}{n(\text{Y})\text{N}} - \frac{n(\text{X and } \bar{\text{Y}})\text{N}}{n(\bar{\text{Y}})\text{N}}$$

$$\text{NConf}(X \rightarrow Y) = \frac{n(\text{X and Y})}{n(\text{Y})} - \frac{n(\text{X and } \bar{\text{Y}})}{n(\bar{\text{Y}})}$$

$$\text{NConf}(X \rightarrow Y) = \frac{n(\text{X and Y})}{n(\text{Y})} - \frac{(n(\text{X}) - n(\text{X and Y}))}{(\text{N} - n(\text{Y}))}$$

มีพจน์ที่เกี่ยวข้องกับจำนวนทรานแซคชั่นทั้งหมด (N) คือพจน์ $(\text{N} - n(\text{Y}))$

ดังนั้นค่าความเชื่อมั่นใหม่ไม่มีคุณสมบัติ O5

พิสูจน์คุณสมบัติ Q1

คุณสมบัติ Q1 นั้นเป็นคุณสมบัติที่อธิบายว่าเมื่อกฎความสัมพันธ์นั้นมีความน่าสนใจสูงที่สุด หรือค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X แต่ไม่มีรายการ Y ในฐานข้อมูลเท่ากับ 0 (หรือ ค่าความเชื่อมั่นเท่ากับ 1 นั่นเอง) แล้วค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะเป็นค่าคงที่ค่าใดค่าหนึ่งหรือเป็นค่าอนันต์

จาก $\text{NConf}(X \rightarrow Y) = \text{P(X|Y)} - \text{P(X|}\bar{\text{Y}}\text{)}$

$$\text{NConf}(X \rightarrow Y) = \frac{\text{P(X and Y)}}{\text{P(Y)}} - \frac{\text{P(X and } \bar{\text{Y}})}{\text{P}(\bar{\text{Y}})}$$

เมื่อกฎความสัมพันธ์นั้นมีความน่าสนใจสูงที่สุด นั่นคือ $\text{P(X and } \bar{\text{Y}}\text{)} = 0$

จะได้ $$\text{NConf}(X \rightarrow Y) = \frac{\text{P(X and Y)}}{\text{P(Y)}}$$

$$\text{NConf}(X \rightarrow Y) = \frac{n(\text{X and Y})\text{N}}{n(\text{Y})\text{N}}$$

$$\text{NConf}(X \rightarrow Y) = \frac{n(\text{X and Y})}{n(\text{Y})}$$

เมื่อกฎความสัมพันธ์นั้นมีความน่าสนใจสูงที่สุด พจน์ n(X and Y) จะเป็นจำนวนเต็มบวกเท่านั้น (ไม่เป็น 0) และพจน์ n(Y) คือจำนวนเต็มบวก (ไม่เป็น 0)

จะได้ เมื่อกฎความสัมพันธ์นั้นมีความน่าสนใจสูงที่สุด ค่าความเชื่อมั่นใหม่จะมีค่าเป็น 1 ซึ่งเป็นค่าคงที่บวกเสมอ

ดังนั้นค่าความเชื่อมั่นใหม่มีคุณสมบัติ Q1

พิสูจน์คุณสมบัติ Q2

คุณสมบัติ Q2 นั้นเป็นคุณสมบัติที่อธิบายว่าค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะต้องมีการลงลดเมื่อมีทรานแซคชั่นที่มีรายการ X แต่ไม่มีรายการ Y ถูกเพิ่มเข้ามาในฐานข้อมูล โดยที่ลักษณะของการลดลงนั้นสามารถเป็นได้ 3 แบบคือ ลดลงเป็นเส้นตรง ลดลงเป็นพาราโบล่าคว่ำ หรือลดลงเป็นพาราโบล่าหงาย อย่างใดอย่างหนึ่ง

จาก $\text{NConf}(X \rightarrow Y) = P(X|Y) - P(X|\bar{Y})$

$$\text{NConf}(X \rightarrow Y) = \frac{P(X \text{ and } Y)}{P(Y)} - \frac{P(X \text{ and } \bar{Y})}{P(\bar{Y})}$$

$$\text{NConf}(X \rightarrow Y) = \frac{n(X \text{ and } Y)N}{n(Y)N} - \frac{n(X \text{ and } \bar{Y})N}{n(\bar{Y})N}$$

$$\text{NConf}(X \rightarrow Y) = \frac{n(X \text{ and } Y)}{n(Y)} - \frac{n(X \text{ and } \bar{Y})}{n(\bar{Y})}$$

จะเห็นว่า เมื่อไม่มีทรานแซคชั่นที่มีรายการ X แต่ไม่มีรายการ Y นั่นคือ $n(X \text{ and } \bar{Y}) = 0$

จะได้ $$\text{NConf}(X \rightarrow Y) = \frac{n(X \text{ and } Y)}{n(Y)}$$

เมื่อเริ่มมีทรานแซคชั่นที่มีรายการ X แต่ไม่มีรายการ Y เพิ่มขึ้น นั่นคือ $n(X \text{ and } \bar{Y}) > 0$

จะได้ $$\text{NConf}(X \rightarrow Y) = \frac{n(X \text{ and } Y)}{n(Y)} - \frac{n(X \text{ and } \bar{Y})}{n(\bar{Y})}$$

จากสมการเส้นตรง $y = mx+c$

จะได้ว่าเมื่อเริ่มมีทรานแซคชั่นที่มีรายการ X แต่ไม่มีรายการ Y เพิ่มขึ้น สมการนี้เป็นสมการเส้นตรง

โดยที่ y คือ $\text{NConf}(X \rightarrow Y)$

x คือ $n(X \text{ and } \bar{Y})$

m คือ $-\frac{1}{n(\bar{Y})}$

c คือ $\frac{n(X \text{ and } Y)}{n(Y)}$

จะได้สมการเส้นตรงที่มีความชันเป็นลบ นั่นคือ $\text{NConf}(X \rightarrow Y)$ ลดลง เมื่อมี $n(X \text{ and } \bar{Y})$ เพิ่มขึ้น

ดังนั้นค่าความเชื่อมั่นใหม่มีคุณสมบัติ Q2 และมีลักษณะของการลดลงแบบเส้นตรง (ใส่หมายเลข 1 ในตารางที่ 2-4)

พิสูจน์คุณสมบัติ Q3

คุณสมบัติ Q3 นั้นเป็นคุณสมบัติที่อธิบายว่าค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ควรจะเพิ่มขึ้นเมื่อจำนวนของทรานแซคชั่นทั้งหมด (N) เพิ่มขึ้นในขณะที่ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และ Y ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ X และค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีรายการ Y คงที่

จาก $NConf(X \rightarrow Y) = P(X|Y) - P(X|\overline{Y})$

$$NConf(X \rightarrow Y) = \frac{P(X \text{ and } Y)}{P(Y)} - \frac{P(X \text{ and } \overline{Y})}{P(\overline{Y})}$$

เนื่องจาก $P(X \text{ and } \overline{Y}) = P(X) - P(X \text{ and } Y)$

และ $P(\overline{Y}) = 1 - P(Y)$

จะได้ $$NConf(X \rightarrow Y) = \frac{P(X \text{ and } Y)}{P(Y)} - \frac{(P(X) - P(X \text{ and } Y))}{(1 - P(Y))}$$

จะเห็นว่าเมื่อ N เพิ่มขึ้นในขณะที่ P(X and Y) P(X) และ P(Y) คงที่ ค่า $NConf(X \rightarrow Y)$ ก็คงที่

ดังนั้นค่าความเชื่อมั่นใหม่มีคุณสมบัติ Q3

พิสูจน์คุณสมบัติ S1

คุณสมบัติ S1 นั้นเป็นคุณสมบัติที่อธิบายว่าค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ที่เขียนให้อยู่ในรูปของฟังก์ชั่นที่ขึ้นกับค่าสนับสนุนนั้นควรจะเพิ่มขึ้นเมื่อค่าสนับสนุนเพิ่มขึ้น เมื่อค่า $n(X)$ $n(\overline{X})$ $n(Y)$ และ $n(\overline{Y})$ คงที่

จาก $NConf(X \rightarrow Y) = P(X|Y) - P(X|\overline{Y})$

$$NConf(X \rightarrow Y) = \frac{P(X \text{ and } Y)}{P(Y)} - \frac{P(X \text{ and } \overline{Y})}{P(\overline{Y})}$$

สมมุติให้ $n(X) = a$, $n(\overline{X}) = N - a$, $n(Y) = b$, $n(\overline{Y}) = N - b$, $Sup(X \rightarrow Y) = x$

จะได้ว่า P(X and Y) = x, P($\overline{X}$ and Y) = ($\frac{b}{N}$) – x, P(X and $\overline{Y}$) = ($\frac{a}{N}$)) – x และ P($\overline{X}$ and $\overline{Y}$) = 1 – $\frac{a+b}{N}$) + x

ดังนั้น $$NConf(X \rightarrow Y) = \frac{xN}{b} - \frac{a - xN}{N - b}$$

$$NConf(X \rightarrow Y) = \frac{xN}{b} - \frac{a}{N - b} + \frac{xN}{N - b}$$

พจน์ $\frac{a}{N - b}$ คงที่ เนื่องจาก n(X) และ n($\overline{Y}$) คงที่

และเมื่อ x เพิ่ม พจน์ $\frac{xN}{b}$ และพจน์ $\frac{xN}{N - b}$ จะเพิ่มขึ้นด้วย เนื่องจาก n(Y) และ n($\overline{Y}$) คงที่

ดังนั้นเมื่อ x เพิ่ม ค่า NConf($X \rightarrow Y$) ก็จะเพิ่มด้วย

นั่นคือ Sup(X→Y) เพิ่ม ค่า NConf($X \rightarrow Y$) ก็จะเพิ่มด้วย

ดังนั้นค่าความเชื่อมั่นใหม่มีคุณสมบัติ S1 (ใส่หมายเลข 0 ในตารางที่ 2-4)

<u>พิสูจน์คุณสมบัติ S2</u>

คุณสมบัติ S1 นั้นเป็นคุณสมบัติที่อธิบายว่าค่าประเมินความน่าสนใจของกฎความสัมพันธ์ M ที่เขียนให้อยู่ในรูปของฟังก์ชั่นที่ขึ้นกับค่าความเชื่อมั่นนั้นควรจะเพิ่มขึ้นเมื่อค่าความเชื่อมั่นเพิ่มขึ้น เมื่อค่า n(X) n($\overline{X}$) n(Y) และ n($\overline{Y}$) คงที่

จาก $$NConf(X \rightarrow Y) = P(X|Y) - P(X|\overline{Y})$$

$$NConf(X \rightarrow Y) = \frac{P(X \text{ and } Y)}{P(Y)} - \frac{P(X \text{ and } \overline{Y})}{P(\overline{Y})}$$

สมมุติให้ n(X) = a, n($\overline{X}$) = N – a, n(Y) = b, n($\overline{Y}$) = N – b, Conf(X→Y) = y

จะได้ว่า P(X and Y) = $\frac{ay}{N}$, P($\overline{X}$ and Y) = $\frac{(b-ay)}{N}$, P(X and $\overline{Y}$) = $\frac{a(1-y)}{N}$ และ P($\overline{X}$ and $\overline{Y}$) = $1 - \frac{(a+b)}{N} + \frac{ay}{N}$

ดังนั้น $$NConf(X \rightarrow Y) = \frac{\frac{ay}{N}}{\frac{b}{N}} - \frac{\frac{a(1-y)}{N}}{\frac{N-b}{N}}$$

$$\text{NConf}(X \rightarrow Y) = \frac{ay}{b} - \frac{a(1-y)}{N-b}$$

$$\text{NConf}(X \rightarrow Y) = \frac{ay}{b} - \frac{a}{N-b} + \frac{ay}{N-b}$$

พจน์ $\frac{a}{N-b}$ คงที่ เนื่องจาก n(X) และ n($\bar{Y}$) คงที่

และเมื่อ y เพิ่ม พจน์ $\frac{ay}{b}$ และพจน์ $\frac{ay}{N-b}$ จะเพิ่มขึ้นด้วย เนื่องจาก n(Y) และ n($\bar{Y}$) คงที่

ดังนั้นเมื่อ y เพิ่ม ค่า NConf($X \rightarrow Y$) ก็จะเพิ่มด้วย

นั่นคือ Conf(X→Y) เพิ่ม ค่า NConf($X \rightarrow Y$) ก็จะเพิ่มด้วย

ดังนั้นค่าความเชื่อมั่นใหม่มีคุณสมบัติ S2 (ใส่หมายเลข 0 ในตารางที่ 2-4)

การทดสอบคุณสมบัติข้างต้นไม่รวมคุณสมบัติ Q4 และ Q5 เนื่องจากคุณสมบัติทั้ง 2 นั้นเป็นคุณสมบัติเชิงอัตวิสัย (Subjective Properties) หรือคุณสมบัติที่ขึ้นอยู่กับผู้ใช้และโดเมนที่นำไปใช้ ไม่สามารถตัดสินได้ว่ามีคุณสมบัติทั้ง 2 ข้อหรือไม่ถ้าไม่ได้อ้างอิงถึงโดเมนและวัตถุประสงค์ที่นำไปประยุกต์ใช้ (Geng and Hamilton, 2006; Heravi, 2009) กับค่าประเมินความน่าสนใจของกฎความสัมพันธ์ทั้งหมด 8 ค่าและค่าความเชื่อมั่นใหม่ การเปรียบเทียบคุณสมบัติของค่าประเมินความน่าสนใจของกฎความสัมพันธ์ 8 ค่าที่รวบรวมโดยคณะวิจัยของ Geng กับ Hamilton ในปี 2006 และคณะวิจัยของ Heravi ในปี 2009 (Geng and Hamilton, 2006; Heravi, 2009) และค่าความเชื่อมั่นใหม่ที่รวบรวมโดยผู้วิจัย แสดงดังตารางต่อไปนี้

ตารางที่ 2-4 แสดงการเปรียบเทียบคุณสมบัติของค่าประเมินความน่าสนใจของกฎความสัมพันธ์ทั้งหมด

	P1	P2	P3	P4	O1	O2	O3	O4	O5	Q1	Q2	Q3	S1	S2	รวม
ค่าสนับสนุน	✗	✓	✗	✓	✓	✗	✗	✗	✗	✗	1	✗	0	0	6
ค่าความเชื่อมั่น	✗	✓	1	✗	✗	✗	✗	✗	✓	✓	1	✗	0	0	7
ค่าคอนวิคชั่น	✗	✓	2	✗	✗	✗	✗	✓	✗	✓	0	✗	0	0	7
ค่าลิฟท์	✗	✓	2	✗	✓	✗	✗	✗	✗	✗	2	✗	0	0	6
ค่าเลฟเวอเรจ	✗	✓	2	✗	✗	✗	✗	✗	✗	✗	1	✗	0	0	5
ค่าคัฟเวอเรจ	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	3	✗	1	0	3
ค่าสหสัมพันธ์	✓	✓	2	✗	✓	✗	✓	✓	✗	✗	0	✗	0	0	9
ค่าอัตราส่วนออดส์	✗	✓	2	✗	✓	✓	✗	✓	✗	✓	0	✗	4	0	9
ค่าความเชื่อมั่นใหม่	✓	✓	1	✗	✗	✓	✓	✓	✗	✓	1	✗	0	0	10

จากตารางข้างต้นเครื่องหมาย ✓ ในตารางหมายถึงค่าประเมินความน่าสนใจของกฎความสัมพันธ์ที่อยู่ในแถวนั้นมีคุณสมบัติของหลักนั้น เครื่องหมาย ✗ ในตารางหมายถึงค่าประเมินความน่าสนใจของกฎความสัมพันธ์ที่อยู่ในแถวนั้นไม่มีคุณสมบัติของหลักนั้น ในหลักของคุณสมบัติ P3 หมายเลข 0 หมายถึง ค่าประเมินความน่าสนใจของกฎความสัมพันธ์เพิ่มขึ้นเมื่อค่าความน่าจะเป็นของเซตรายการที่มาก่อนลดลง หมายเลข 1 หมายถึง ค่าประเมินความน่าสนใจของกฎความสัมพันธ์เพิ่มขึ้นเมื่อค่าความน่าจะเป็นของเซตรายการที่ตามมาลดลง หมายเลข 2 หมายถึง ค่าประเมินความน่าสนใจของกฎความสัมพันธ์เพิ่มขึ้นเมื่อทั้งค่าความน่าจะเป็นของเซตรายการที่มาก่อนและค่าความน่าจะเป็นของเซตรายการที่ตามมาลดลงเท่านั้น ในหลักของคุณสมบัติ Q2 หมายเลข 0 หมายถึง ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ลดลงแบบพาราโบล่าคว่ำ หมายเลข 1 หมายถึง ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ลดลงแบบเส้นตรง หมายเลข 2 หมายถึง ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ลดลงแบบพาราโบล่าหงาย หมายเลข 3 หมายถึง ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ลดลงแต่ขึ้นอยู่กับพารามิเตอร์ หมายเลข 4 หมายถึง ค่าประเมินความน่าสนใจของกฎความสัมพันธ์คงที่ หมายเลข 5 หมายถึง ค่าประเมินความน่าสนใจของกฎความสัมพันธ์เพิ่มขึ้น หมายเลข 6 หมายถึง ค่าประเมินความน่าสนใจของกฎความสัมพันธ์เพิ่มหรือลดไม่แน่นอนขึ้นอยู่กับพารามิเตอร์ ในหลัก

ของคุณสมบัติ S1 และ S2 หมายเลข 0 หมายถึงค่าประเมินความน่าสนใจของกฎความสัมพันธ์เพิ่มขึ้นเมื่อค่าสนับสนุนเพิ่มขึ้น หมายเลข 1 หมายถึงค่าประเมินความน่าสนใจของกฎความสัมพันธ์คงที่เมื่อค่าสนับสนุนเพิ่มขึ้น หมายเลข 2 หมายถึงค่าประเมินความน่าสนใจของกฎความสัมพันธ์ลดลงเมื่อค่าสนับสนุนเพิ่มขึ้น หมายเลข 3 หมายถึงไม่สามารถประเมินได้ (not applicable) และหมายเลข 4 หมายถึงค่าประเมินความน่าสนใจของกฎความสัมพันธ์เพิ่มหรือลดขึ้นอยู่กับพารามิเตอร์ (Geng and Hamilton, 2006; Heravi, 2009) หลักสุดท้ายของตารางคือหลักที่แสดงการรวมจำนวนคุณสมบัติทั้งหมดที่ค่าประเมินความน่าสนใจของกฎความสัมพันธ์นั้นๆ มี

จากผลการพิสูจน์คุณสมบัติของค่าประเมินความน่าสนใจของกฎความสัมพันธ์ข้างต้น และตารางที่ 2-4 แสดงให้เห็นว่าค่าความเชื่อมั่นใหม่มีคุณสมบัติที่ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ควรจะมี 10 คุณสมบัติจากทั้งหมด 14 คุณสมบัติ ซึ่งมากที่สุดเมื่อเทียบกับค่าประเมินความน่าสนใจของกฎความสัมพันธ์อื่นๆ โดยเฉพาะอย่างยิ่งการมีคุณสมบัติ O3 ของค่าความเชื่อมั่นใหม่จะทำให้การนำค่าความเชื่อมั่นใหม่ไปใช้นั้นจะสามารถขจัดการเกิดกฎความสัมพันธ์ที่มีเซตรายการที่มาก่อนและเซตรายการที่ตามที่มีความสัมพันธ์เชิงลบออกไปได้ ผู้วิจัยจึงเชื่อว่าถ้านำค่าความเชื่อมั่นใหม่นี้ไปประยุกต์ใช้กับการค้นหากฎความสัมพันธ์กับข้อมูลประเภทต่างๆรวมถึงข้อมูลซอฟต์แวร์อาร์ไควฟ์ แล้วน่าจะทำให้กฎความสัมพันธ์ที่ได้มาเป็นกฎความสัมพันธ์ที่น่าสนใจและช่วยลดการเกิดกฎความสัมพันธ์ที่เป็นผลบวกลวง (False Positive) ได้

2.6 การควบคุมการเปลี่ยนแปลงแก้ไข (Revision Control, Version Control)

ซอฟต์แวร์ที่ถูกพัฒนาขึ้นในปัจจุบันนั้นมีความซับซ้อนและขนาดที่ใหญ่กว่าซอฟต์แวร์ที่ถูกพัฒนาขึ้นมาในอดีตเป็นอย่างมาก ความยากและความซับซ้อนเหล่านั้นถูกสะท้อนออกมาในการบริหารจัดการการพัฒนาและการบำรุงรักษาของซอฟต์แวร์นั้น แม้ว่าหลายองค์กรในปัจจุบันจะใช้ซอฟต์แวร์ควบคุมการเปลี่ยนแปลงแก้ไข (Revision Control Software, Version Control Software) คอยติดตามและจัดการกับพัฒนาการของความซับซ้อนของโครงการพัฒนาซอฟต์แวร์กันอย่างมาก แต่ทว่าแนวคิดของการควบคุมการเปลี่ยนแปลงแก้ไข (Concept of Revision Control, Version Control) นั้นกลับเป็นศาสตร์ที่ได้ไม่ค่อยถูกพูดถึงและไม่ค่อยมีวัฒนาการมาก

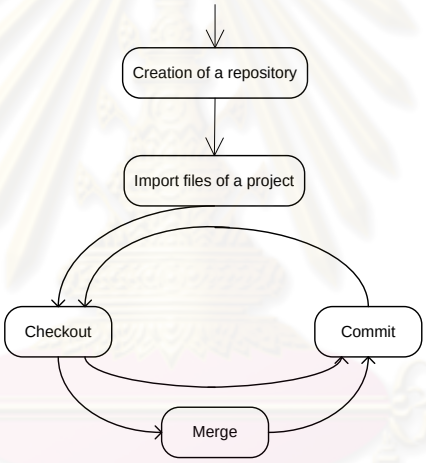
เท่าไรในตลอดทศวรรษที่ผ่านมา (Löh et al., 2007) ในหัวข้อนี้ได้เรียบเรียงวรรณกรรมที่เกี่ยวข้องกับการควบคุมการเปลี่ยนแปลงแก้ไข เริ่มตั้งแต่แนวคิดของการควบคุมการเปลี่ยนแปลงแก้ไข ซอฟต์แวร์ที่ประยุกต์แนวคิดของการควบคุมการเปลี่ยนแปลงแก้ไข และสุดท้ายจะกล่าวถึงระบบคอนเคอเรนท์เวอร์ชั่น (Concurrent Version System, CVS) ซึ่งเป็นซอฟต์แวร์โอเพนซอร์สที่ประยุกต์แนวคิดของการควบคุมการเปลี่ยนแปลงแก้ไขที่ได้รับความนิยมสูงมากนานกว่าทศวรรษ (O'Sullivan et al., 2009)

2.6.1 **แนวคิดของการควบคุมการเปลี่ยนแปลงแก้ไข** (Concept of Revision Control, Version Control)

การควบคุมการเปลี่ยนแปลงแก้ไข (Revision Control, Version Control) เป็นลักษณะอย่างหนึ่งของการควบคุมเอกสาร (Documentation Control) แนวคิดของการควบคุมการเปลี่ยนแปลงแก้ไข คือ การควบคุมพัฒนาการของเนื้อหาภายในเอกสารอิเล็กทรอนิกส์ตลอดช่วงอายุของเอกสาร รวมถึงการเรียกเนื้อหาในอดีตกลับคืน (Recovery) การบ่งชี้ข้อแตกต่างระหว่างเวอร์ชั่นของเนื้อหา และการให้รายละเอียดของพัฒนาการแต่ละครั้งด้วย (Tichy, 1982; Junqueira et al., 2008) การควบคุมการเปลี่ยนแปลงแก้ไขนั้นถูกนำไปประยุกต์ใช้ในทางวิศวกรรมแขนงต่างๆเพื่อการจัดการการพัฒนาที่ต่อเนื่องไปของเอกสารอิเล็กทรอนิกส์ เช่น ซอร์สโค้ดของโปรแกรมประยุกต์ พิมพ์เขียว แบบจำลองอิเล็กทรอนิกส์ และสารสนเทศสำคัญอื่นๆ ซึ่งพัฒนาโดยทีม การเปลี่ยนแปลงเอกสารเหล่านี้ในแต่ละครั้งจะถูกระบุโดยใช้การเพิ่มหมายเลขการเปลี่ยนแปลงแก้ไข (Revision Number) และมีการเชื่อมโยงกับผู้กระทำการเปลี่ยนแปลงแก้ไขด้วย

การไหลของกิจกรรมการควบคุมการเปลี่ยนแปลงแก้ไขเริ่มต้นจากการสร้างรีพอสิโทริ (Creation of a Pepository) สำหรับโครงการขึ้นมา จากนั้นนำเข้าแฟ้มข้อมูลทั้งหมดของโครงการ (Import Files of Project) ลงสู่รีพอสิโทริที่สร้างไว้ แฟ้มข้อมูลที่นำเข้ามาครั้งแรกจะถูกกำหนดค่าเริ่มต้น (Initial Set) ให้มีหมายเลขการแก้ไขหรือหมายเลขเวอร์ชั่นเป็น 1 เมื่อมีความต้องการเปลี่ยนแปลงแก้ไขแฟ้มข้อมูลผู้ใช้ก็จะต้องทำการลงทะเบียนออก (Check out) แฟ้มข้อมูลนั้นออกมาจากรีพอสิโทริ การลงทะเบียนออกของแฟ้มข้อมูลก็คือการสำเนาแฟ้มข้อมูลเวอร์ชั่นล่าสุดจากรีพอสิโทริมาเป็นแฟ้มข้อมูลบนเครื่องของผู้ใช้ (Local File) เมื่อผู้ใช้เปลี่ยนแปลงแก้ไขแฟ้มข้อมูลเรียบร้อยแล้วจะต้องการลงทะเบียนเข้า (Check in) หรือคอมมิท (Commit) แฟ้มข้อมูลนั้นกลับเข้าสู่รีพอสิโทริ การคอมมิทแต่ละครั้งจะมีผลให้หมายเลขการแก้ไขหรือหมายเลขเวอร์ชั่นมีค่าเพิ่มขึ้นไปเรื่อยๆ การไหลของกิจกรรมจะวนลูปเช่นนี้ไปเรื่อยๆตลอดช่วงอายุของโครงการ ใน

กรณีที่มีผู้ใช้หลายคนทำการเปลี่ยนแปลงแก้ไขแฟ้มข้อมูลเดียวกันในเวลาเดียวกัน (มีการลงทะเบียนออกของแฟ้มข้อมูลโดยผู้ใช้มากกว่า 1 คนในช่วงเวลาเดียวกัน) จะทำให้เกิดกิจกรรมที่สำคัญขึ้นอีกกิจกรรมคือการผสานเนื้อหา (Merge) การผสานเนื้อหาคือการปรับปรุง (Update) เนื้อหาภายในแฟ้มข้อมูลบนเครื่องผู้ใช้ที่แก้ไขไปแล้วกับเนื้อหาล่าสุดของแฟ้มข้อมูลนั้นในรีพอสิโทริซึ่งอาจถูกคอมมิทเวอร์ชั่นใหม่จากผู้ใช้อื่นในระหว่างที่ผู้ใช้กำลังแก้ไขอยู่ กล่าวคือเป็นการปรับปรุงเนื้อหาของแฟ้มข้อมูลนั้นบนเครื่องของผู้ใช้ (Local File) ให้สอดคล้องกับเวอร์ชั่นบนรีพอสิโทรินั่นเอง ผู้ใช้สามารถทำการผสานเนื้อหาได้โดยการเรียกใช้คำสั่งที่ชื่อว่า update (Tichy, 1982; Ambriola et al., 1990; Junqueira et al., 2008) การไหลของกิจกรรมการควบคุมการเปลี่ยนแปลงแก้ไขสามารถแสดงได้ดังรูปต่อไปนี้ (Junqueira et al., 2008)



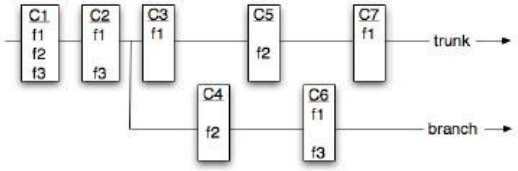
รูปที่ 2-1 แสดงการไหลของกิจกรรมการควบคุมการเปลี่ยนแปลงแก้ไข

แนวคิดของการควบคุมการเปลี่ยนแปลงแก้ไขที่กล่าวไปข้างต้นนั้นเป็นแนวคิดของการควบคุมการเปลี่ยนแปลงแก้ไขสำหรับเอกสารอิเล็กทรอนิกส์ทั่วไป สำหรับแนวคิดของการควบคุมการเปลี่ยนแปลงแก้ไขในวิศวกรรมซอฟต์แวร์จะหมายถึงเทคนิคและเครื่องมือที่นำมาใช้ในการควบคุมพัฒนาการของแฟ้มข้อมูลซอร์สโค้ดและแฟ้มข้อมูลอื่นๆภายในโครงการพัฒนาซอฟต์แวร์ (Junqueira et al., 2008) แฟ้มข้อมูลซอร์สโค้ดนั้นเป็นแฟ้มข้อมูลที่มีคุณลักษณะเฉพาะตัวมากกว่าแฟ้มข้อมูลของเอกสารอิเล็กทรอนิกส์อื่นๆ ทำให้เทคนิคที่ใช้ในการควบคุมการเปลี่ยนแปลงแก้ไขแฟ้มข้อมูลซอร์ส โค้ดมีความละเอียดและเฉพาะตัวมากกว่าการควบคุมการเปลี่ยนแปลงแก้ไขของเอกสารอิเล็กทรอนิกส์อื่นๆ คุณลักษณะเฉพาะนั้นก็คือเนื้อหา (Contents) ภายในแฟ้มข้อมูลซอร์สโค้ดนั้นเป็นเนื้อหาที่มีโครงสร้างและโครงสร้างนั้นประกอบด้วยหน่วยย่อยที่

มีความสัมพันธ์กันอยู่ภายใน ด้วยเหตุนี้การควบคุมการเปลี่ยนแปลงแก้ไขแฟ้มข้อมูลซอร์สโค้ดจึงต้องมีการดำเนินการขั้นสูง (Advanced Operations) 2 อย่างคือการต่อกิ่ง (Branches) และการผสานกิ่ง (Merge) (Chadd et al., 2008)

คำว่า กิ่ง (Branch) ในที่นี้หมายถึง กลุ่มของเวอร์ชั่นต่างๆที่ถูกสร้างขึ้นมาด้วยเหตุผลในการทดสอบอะไรบางอย่างที่ยังไม่มั่นใจเพียงพอที่จะนำเวอร์ชั่นเหล่านี้ไปรวมกับเวอร์ชั่นหลักที่เสถียรแล้ว การต่อกิ่งเป็นการเปรียบเทียบการพัฒนาซอร์สโค้ดในลักษณะของต้นไม้ (Tree) คือการกำหนดให้ซอร์สโค้ดหลัก (เวอร์ชั่นของซอร์สโค้ดที่มั่นใจว่ามีความเสถียรสูงแล้ว) เป็นลำต้น (Trunk) ของต้นไม้ โดยปกติแล้วเมื่อนักพัฒนาแก้ไขหรือพัฒนาซอร์สโค้ดเพิ่มเติมเข้าไปจะทำให้เกิดเวอร์ชั่นใหม่ขึ้นมา เวอร์ชั่นใหม่ที่เสถียรเหล่านี้จะเรียงต่อกันไปเป็นลำดับ แต่ในกรณีที่นักพัฒนามีการแก้ไขหรือพัฒนาซอร์สโค้ดส่วนใดส่วนหนึ่ง ซึ่งเป็นการแก้ไขเพื่อทดสอบอะไรบางอย่างและยังไม่มั่นใจในความเสถียรของการแก้ไขครั้งนั้น เวอร์ชั่นใหม่ที่ถูกสร้างขึ้นมานี้จะถูกกำหนดให้เป็นกิ่ง (Branches) ต่อออกมาจากส่วนลำต้นหลักของต้นไม้ ทิศทางการพัฒนาจะขยายออกจากส่วนลำต้นของต้นไม้ การต่อกิ่งนี้เป็นเทคนิคที่นิยมใช้มากในการควบคุมการเปลี่ยนแปลงแก้ไขของโครงการพัฒนาซอฟต์แวร์ที่มีขนาดใหญ่ เทคนิควิธีการต่อกิ่งสามารถใช้เพื่อการทดสอบซอร์สโค้ดและเพิ่มเติมซอร์สโค้ดเมื่อเราต้องการที่จะขยายระบบหรือซอฟต์แวร์เพิ่มเติม ในขณะที่มีการพัฒนาส่วนกิ่งอยู่นั้นส่วนลำต้นก็ยังคงสามารถพัฒนาต่อออกไปได้เช่นกัน เมื่อพัฒนาส่วนที่เป็นกิ่งเสร็จแล้วและมั่นใจความเสถียรของกิ่งนั้นแล้ว นักพัฒนาสามารถที่จะนำส่วนกิ่งนั้นมารวมเข้ากับส่วนลำต้นหลักได้ เรียกว่าการผสานกิ่ง (Merge)

ถ้ากำหนดให้การเปลี่ยนแปลงแก้ไข (Change, Revision) คือการที่นักพัฒนาแก้ไข (alter) เพิ่ม (add) หรือ ลบ (delete) อย่างใดอย่างหนึ่งบนแฟ้มข้อมูลซอร์สโค้ด การคอมมิท (Commit) ก็คือเซตของการเปลี่ยนแปลงแก้ไขที่กระทำโดยนักพัฒนาคนเดียวกันในเวลาเดียวกันหรือใกล้เคียงกันโดยที่ไม่ใช่เซตว่างนั่นเอง รูปด้านล่างนี้แสดงตัวอย่างของการคอมมิท ลงบนลำต้นและบนกิ่งโดยกำหนดให้ สัญลักษณ์ C# คือการคอมมิทครั้งที่ # และสัญลักษณ์ f# คือการเปลี่ยนแปลงแก้ไขบนแฟ้มข้อมูลที่ # (Junqueira et al., 2008)



รูปที่ 2-2 แสดงตัวอย่างการคอมมิทลงบนลำต้นและบนกิ่ง

จากรูปที่ 2-2 ข้างต้นนี้ถ้ามีการผสานกิ่งเกิดขึ้นหลังจากการคอมมิทครั้งที่ 7 การดำเนินการของรีพอสิโทรีที่จะเกิดขึ้นก็คือ พิจารณาว่าตั้งแต่เกิดการต่อกิ่งนี้ขึ้นมาจนถึงการผสานกิ่งมีการเปลี่ยนแปลงแก้ไขกับแฟ้มข้อมูลใดบ้าง ถ้าการเปลี่ยนแปลงแก้ไขแฟ้มข้อมูลใดเกิดขึ้นทั้งบนกิ่งและบนลำต้น การเปลี่ยนแปลงแก้ไขแฟ้มข้อมูลนั้นจะต้องถูกจับคู่กันแล้วนำไปวิเคราะห์เชิงลึกว่ามีการเปลี่ยนแปลงแก้ไขที่เกิดความขัดแย้ง (Conflict) ขึ้นหรือไม่ ถ้าไม่เกิดความขัดแย้งขึ้นเปลี่ยนแปลงแก้ไขแฟ้มข้อมูลนั้นทั้งบนกิ่งและบนลำต้นก็จะสามารถนำมาผสานกันได้ทันที แต่ถ้าเกิดความขัดแย้งขึ้นจำเป็นจะต้องให้ผู้ที่ทำการผสานกิ่งเป็นผู้ตัดสินว่าจะแก้ไขความขัดแย้งนี้อย่างไร ตัวอย่างเช่น จากรูปข้างต้น ตั้งแต่จุดเริ่มต้นการต่อกิ่งจนถึงจุดการผสานกิ่งนั้นมีการคอมมิทบนลำต้นทั้งหมด 3 ครั้งคือ การคอมมิทครั้งที่ 3 การคอมมิทครั้งที่ 5 และการคอมมิทครั้งที่ 7 และมีการคอมมิทบนกิ่งทั้งหมด 2 ครั้งคือ การคอมมิทครั้งที่ 4 และการคอมมิทครั้งที่ 6 การเปลี่ยนแปลงแก้ไขบนแฟ้มข้อมูลที่ 2 เกิดขึ้นทั้งการคอมมิทบนกิ่ง (การคอมมิทครั้งที่ 4) และการคอมมิทบนลำต้น (การคอมมิทครั้งที่ 5) ดังนั้นการเปลี่ยนแปลงแก้ไขบนแฟ้มข้อมูลที่ 2 ทั้งสองอันนี้จะต้องถูกนำไปวิเคราะห์เชิงลึกต่อไป ถ้าพบว่าเกิดความขัดแย้งขึ้นจะต้องมีการตัดสินความขัดแย้งนั้น แต่ถ้าไม่มีการเปลี่ยนแปลงแก้ไขบนแฟ้มข้อมูลที่ 2 ทั้งสองอันจะถูกผสานเข้าด้วยกัน (Junqueira et al., 2008)

เทคนิคและขั้นตอนวิธีในการต่อกิ่ง การผสานกิ่ง และการเปรียบเทียบความแตกต่างระดับบรรทัดของแต่ละซอฟต์แวร์ควบคุมการเปลี่ยนแปลงแก้ไขนั้น มีเทคนิคและขั้นตอนวิธีที่แตกต่างกันออกไปตามซอฟต์แวร์ควบคุมการเปลี่ยนแปลงแก้ไข ซึ่งจะไม่กล่าวถึงในที่นี้

2.6.2 **ซอฟต์แวร์ควบคุมการเปลี่ยนแปลงแก้ไข** (Revision Control Software, Version Control Software)

ซอฟต์แวร์ควบคุมการเปลี่ยนแปลงแก้ไข (Revision Control Software, Version Control Software) คือ ซอฟต์แวร์ที่นำแนวคิดของการควบคุมการเปลี่ยนแปลงแก้ไขมาประยุกต์ใช้จริงในอุตสาหกรรมการพัฒนาซอฟต์แวร์ ซอฟต์แวร์ควบคุมการเปลี่ยนแปลงแก้ไขคือซอฟต์แวร์ที่ใช้ใน

จัดการการจัดเก็บ การค้นคืน การระบุและการผสานการเปลี่ยนแปลงแก้ไขแฟ้มข้อมูลซอร์สโค้ดของโปรแกรมประยุกต์ และสารสนเทศสำคัญอื่นๆที่พัฒนาขึ้นมาโดยทีมอย่างเป็นอัตโนมัติ ในซอฟต์แวร์ควบคุมการเปลี่ยนแปลงแก้ไขนั้นจะมีการบันทึกแฟ้มข้อมูลซอร์สโค้ดทั้งโครงการเอาไว้ นอกจากนั้นข้อมูลที่เกี่ยวข้องกับการเปลี่ยนแปลงแก้ไขอื่นๆ อาทิเช่น ซอร์สโค้ดส่วนใดที่ถูกแก้ไข นักพัฒนาผู้บันทึกเวอร์ชั่นใหม่ของซอร์สโค้ด วันเวลาบันทึกเวอร์ชั่นใหม่ของซอร์สโค้ด และหมายเหตุของการบันทึกเวอร์ชั่นใหม่ของซอร์สโค้ดนั้นจะถูกบันทึกอยู่ในรูปของแฟ้มข้อมูลบันทึก (Log Files) แฟ้มข้อมูลบันทึกเหล่านี้ถูกแก้ไขใหม่ทุกครั้งที่มีนักพัฒนาทำการคอมมิท แฟ้มข้อมูลซอร์สโค้ดแต่ละเวอร์ชั่นในอดีตและแฟ้มข้อมูลบันทึกทั้งหมดจะถูกเรียกรวมกันว่า ซอฟต์แวร์อาร์ไคฟว์ (Software Archives) (Zimmermann et al., 2004; Zimmermann et al., 2005)

ซอฟต์แวร์แรกที่ได้ประยุกต์แนวคิดของการควบคุมการเปลี่ยนแปลงแก้ไขขึ้นมาจริง คือ ระบบควบคุมซอร์สโค้ด (SCCS: Source Code Control System) ซึ่งถูกแนะนำขึ้นมาและออกจำหน่ายครั้งแรกในปี ค.ศ. 1972 โดยแบลแลบส์ (Bell Labs) (Rochkind et al., 1975) ในช่วงแรกนั้นระบบควบคุมซอร์สโค้ดถูกนำไปใช้อยู่เพียงแค่ในวงจำกัดและไม่ค่อยได้รับความนิยมมากเท่าไร (Baudis, 2009) หลังจากนั้นในปี ค.ศ. 1985 เป็นปีแรกที่ระบบควบคุมการเปลี่ยนแปลงแก้ไข (RCS: Revision Control System) ได้ถูกเผยแพร่ออกมาและสามารถนำไปใช้ได้โดยไม่เสียค่าใช้จ่าย (Tichy, 1985) หลังจากที่ถูกเผยแพร่ได้ไม่นานระบบควบคุมการเปลี่ยนแปลงแก้ไขได้ประสบความสำเร็จอย่างมากในอุตสาหกรรมพัฒนาซอฟต์แวร์ (Baudis, 2009) ถึงแม้ว่าในปัจจุบันระบบควบคุมการเปลี่ยนแปลงแก้ไขจะไม่ได้มีวิวัฒนาการใดๆเพิ่มขึ้นมาและไม่ได้ถูกนำไปใช้แล้วแต่ระบบควบคุมการเปลี่ยนแปลงแก้ไขยังคงได้รับการกล่าวถึงอยู่เรื่อยๆ ในฐานะเป็นต้นแบบของแนวคิดและรูปแบบแฟ้มข้อมูล (File Format) ของซอฟต์แวร์ควบคุมการเปลี่ยนแปลงแก้ไขที่ดี (Baudis, 2009) ต่อมาในปี ค.ศ. 1990 ระบบคอนเคอเรนท์เวอร์ชั่น (CVS: Concurrent Versions System) ที่ถูกพัฒนาขึ้นมาโดยมีระบบควบคุมการเปลี่ยนแปลงแก้ไข (RCS, Revision Control System) เป็นต้นแบบได้ถูกเผยแพร่ออกมาและได้รับความนิยมอย่างมากและรวดเร็ว นอกจากซอฟต์แวร์ทั้ง 3 ซอฟต์แวร์ในข้างต้นแล้วหลังจากนั้นในช่วงทศวรรษที่ 1990 ก็มีซอฟต์แวร์ที่ประยุกต์แนวคิดของการควบคุมการเปลี่ยนแปลงแก้ไขเกิดขึ้นมาอีกอย่างมากมาย ทั้งที่เป็นซอฟต์แวร์เชิงพาณิชย์ (Commercial Software) และซอฟต์แวร์เสรี (Free/Open Source Software) ตัวอย่างซอฟต์แวร์ควบคุมการเปลี่ยนแปลงแก้ไขเชิงพาณิชย์เช่น ClearCase (ปี ค.ศ. 1992), Visual SourceSafe (ปี ค.ศ. 1994), Perforce (ปี ค.ศ. 1995) และ Code Co-op

(ปี ค.ศ. 1997) ตัวอย่างซอฟต์แวร์ควบคุมการเปลี่ยนแปลงแก้ไขที่เป็นซอฟต์แวร์เสรีเช่น CVSNT (ปี ค.ศ. 1998) และ Subversion (ปี ค.ศ. 2000)

เนื่องจากงานวิจัยนี้สนใจศึกษาข้อมูลซอฟต์แวร์อาร์ไคฟ์ที่ได้จากระบบคอนเคอเรนท์เวอร์ชั่นดังนั้นผู้วิจัยจึงเรียบเรียงรายละเอียดเกี่ยวกับระบบคอนเคอเรนท์เวอร์ชั่นรวมถึงข้อมูลซอฟต์แวร์อาร์ไคฟ์ของระบบคอนเคอเรนท์เวอร์ชั่นไว้ในหัวข้อถัดไป

2.6.3 **ระบบคอนเคอเรนท์เวอร์ชั่น** (Concurrent Versions System, CVS)

ระบบคอนเคอเรนท์เวอร์ชั่น (Concurrent Versions System, CVS) คือซอฟต์แวร์ควบคุมการเปลี่ยนแปลงแก้ไข (Revision Control Software, Version Control Software) ซอฟต์แวร์หนึ่งที่ได้รับความนิยมสูงสุดในปัจจุบัน (O'Sullivan et al., 2009) ระบบคอนเคอเรนท์เวอร์ชั่นนั้นถูกพัฒนาขึ้นมาโดยมีระบบควบคุมการเปลี่ยนแปลงแก้ไข (Revision Control System, RCS) เป็นต้นแบบ และพัฒนาขึ้นมาอย่างโอเพนซอร์ส ดังนั้นระบบคอนเคอเรนท์เวอร์ชั่นนี้จึงสามารถบรรจุลง (download) มาใช้ได้โดยไม่เสียค่าใช้จ่าย ความสำคัญของระบบคอนเคอเรนท์เวอร์ชั่นคือระบบคอนเคอเรนท์เวอร์ชั่นนี้ถูกสร้างขึ้นมาให้เป็นส่วนประกอบ (component) ที่สำคัญส่วนประกอบหนึ่งที่ช่วยให้บรรลุถึงมาตรฐานของการจัดการค่าองค์ประกอบของซอฟต์แวร์ (SCM, Software Configuration Management) ซึ่งเป็น 1 ใน 6 กลุ่มกระบวนการหลัก (Process Area) ในระดับที่ 2 ของมาตรฐาน CMM (Capability Maturity Model) ได้ (Grune et al., 2006)

วิวัฒนาการของระบบคอนเคอเรนท์เวอร์ชั่นนั้นเริ่มต้นขึ้นในเดือนกรกฎาคมปี ค.ศ. 1986 Grune และคณะได้เริ่มต้นการเผยแพร่บางส่วนระบบคอนเคอเรนท์เวอร์ชั่น (Concurrent Versions System, CVS) ในรูปแบบของเชลล์สคริพท์ (Shell Script) ลงสู่กลุ่มข่าว (News-Group) ที่ชื่อว่า comp.sources.unix ต่อมาในเดือนเมษายนปี ค.ศ. 1989 Berliner และ Polk (Berliner et al., 1989) ได้ออกแบบและพัฒนาระบบคอนเคอเรนท์เวอร์ชั่นเพิ่มเติมซึ่งเวอร์ชั่นของ Berliner นี้ก็คือเวอร์ชั่นที่ใช้กันอยู่ในปัจจุบัน (Cederqvist, 2006; Grune et al., 2006) จนกระทั่งวันที่ 19 พฤศจิกายน ปี ค.ศ. 1990 ระบบคอนเคอเรนท์เวอร์ชั่นที่สมบูรณ์พร้อมใช้งานเวอร์ชั่นที่ 1 ได้ถูกเสนอเข้าสู่มูลนิธิซอฟต์แวร์เสรี (Free Software Foundation) เพื่อการพัฒนาและการเผยแพร่ต่อไป (Grune et al., 2006)

ระบบคอนเคอเรนท์เวอร์ชั่นเป็นซอฟต์แวร์ควบคุมการเปลี่ยนแปลงแก้ไขที่ใช้สถาปัตยกรรมลูกข่าย-แม่ข่าย (Client-Server) การทำงานหลักของระบบคอนเคอเรนท์เวอร์ชั่น

คือ การเก็บประวัติการทำงานของแฟ้มข้อมูลต่างๆในโครงการพัฒนาซอฟต์แวร์ เพื่อช่วยแก้ปัญหาในการค้นหาความแตกต่างของซอร์สโค้ดพื้นฐานเดิมกับซอร์สโค้ดที่ทำการแก้ไขใหม่ นักพัฒนาสามารถที่จะทำการเปลี่ยนแปลงซอร์สโค้ด แก้ไขข้อผิดพลาดและรวม (Merge) ซอร์สโค้ดที่ตนพัฒนาอยู่กับซอร์สโค้ดล่าสุดของนักพัฒนาคนอื่นๆได้ ระบบคอนเคอเรนท์เวอร์ชั่นนั้นมีรูปแบบการทำงานแบบ 1 รีพอสิโทริ (Repository) แต่หลายๆ พื้นที่ทำงาน (Workspace) ภายใต้รูปแบบการทำงานนี้นักพัฒนาแต่ละคนสามารถนำซอร์สโค้ดมาพัฒนาเพียงงานเดียวหรือหลายๆงานก็ได้ เมื่อแก้ไขเรียบร้อยแล้วก็ทำการรวม (Merge) งานที่ทำเข้าด้วยกัน และสามารถช่วยให้นักพัฒนาค้นหาได้ว่าข้อผิดพลาดเกิดขึ้นนั้น เกิดขึ้นที่ไหนและกระทำโดยใคร โดยปกติแล้วการบันทึกเวอร์ชั่นทุกๆเวอร์ชั่นของแต่ละเอกสารเอาไว้นั้นเป็นการกระทำที่ค่อนข้างสิ้นเปลืองเนื้อที่บันทึก (Disk) เป็นอย่างมาก แต่ระบบคอนเคอเรนท์เวอร์ชั่นนั้นมีวิธีการบันทึกแต่ละเวอร์ชั่นของเอกสารภายในแฟ้มข้อมูลแฟ้มเดียวด้วยวิธีที่มีประสิทธิภาพ นั่นคือระบบคอนเคอเรนท์เวอร์ชั่นจะบันทึกเฉพาะส่วนที่แตกต่างกันของแต่ละเวอร์ชั่นเอาไว้เท่านั้น (Cederqvist, 2006)

เนื่องจากระบบคอนเคอเรนท์เวอร์ชั่น (CVS) เป็นระบบที่ถูกพัฒนาขึ้นมาโดยมีระบบควบคุมการเปลี่ยนแปลงแก้ไข (Revision Control System, RCS) เป็นต้นแบบ รูปแบบแฟ้มข้อมูล (File Format) ของระบบคอนเคอเรนท์เวอร์ชั่นจึงใช้รูปแบบเดียวกับระบบควบคุมการเปลี่ยนแปลงแก้ไขทั้งหมดรวมถึงรูปแบบของแฟ้มข้อมูลบันทึก (CVS Log File) ด้วย การดึงแฟ้มข้อมูลบันทึกของโครงการที่เก็บอยู่บนรีพอสิโทริสามารถทำได้โดยการใช้คำสั่ง cvs log และแฟ้มข้อมูลบันทึกจะถูกทำสำเนาและส่งกลับมายังเครื่องที่ใช้คำสั่งนี้ (Fischer et al., 2003) ตัวอย่างบางส่วนของแฟ้มข้อมูลบันทึกของคอนเคอเรนท์เวอร์ชั่นแสดงดังรูปด้านล่างนี้ (Fischer et al., 2003)

```
RCS file: /cvsroot/mozilla/layout/html/style/src/nsCSSFrameConstructor.cpp,v
Working file: nsCSSFrameConstructor.cpp
head: 1.804
branch:
locks: strict
access list:
symbolic names:
        MOZILLA_1_3a_RELEASE: 1.800
        NETSCAPE_7_01_RTM_RELEASE: 1.727.2.17
        PHOENIX_0_5_RELEASE: 1.800
        ...
        RDF_19990305_BASE: 1.46
        RDF_19990305_BRANCH: 1.46.0.2
keyword substitution: kv
total revisions: 976;   selected revisions: 976
description:
----------------------------
revision 1.804
date: 2002/12/13 20:13:16;  author: doe@netscape.com;  state: Exp;  lines: +15 -47
Don't set NS_BLOCK_SPACE_MGR and NS_BLOCK_WRAP_SIZE on ...
----------------------------
...
----------------------------
revision 1.638
date: 2001/09/29 02:20:52;  author: doe@netscape.com;  state: Exp;  lines: +14 -4
branches:  1.638.4;
bug 94341 keep a separate pseudo frame list for a new pseudo block or inline frame ...
----------------------------
....
=============================================================================
```

รูปที่ 2-3 แสดงตัวอย่างแฟ้มข้อมูลบันทึกของคอนเคอเรนท์เวอร์ชั่น

แฟ้มข้อมูลบันทึกของระบบคอนเคอเรนท์เวอร์ชั่นประกอบด้วยหลายส่วน (Sections) โดยที่แต่ละส่วนจะอธิบายถึงเวอร์ชั่นในอดีตของแต่ละแฟ้มข้อมูลที่อยู่ในโครงการ (1 ส่วนต่อ 1 แฟ้มข้อมูล) ส่วนแต่ละส่วนจะถูกแบ่งออกจากกันด้วยบรรทัดของเครื่องหมายเท่ากับ (=) ภายในแต่ละส่วนจะประกอบด้วยแอททริบิวต์ (Attributes) และค่าของแอททริบิวต์นั้นๆ แอททริบิวต์ที่สำคัญและถูกนำไปใช้ในการวิเคราะห์ต่างๆ มีรายละเอียดดังนี้ (Fischer et al., 2003)

- แอททริบิวต์ RCS file คือ พาธ (Path) ที่ระบุตำแหน่งของแฟ้มข้อมูลที่สัมพันธ์กับส่วนนี้ที่บันทึกอยู่บนรีพอสิโทริของระบบคอนเคอเรนท์เวอร์ชั่น โดยอ้างอิงจากตำแหน่งราก (Root) ของรีพอสิโทริ จากรูปที่ 2-3 ระบุว่าส่วนนี้เป็นส่วนของแฟ้มข้อมูลชื่อ nsCSSFrameConstructor.cpp ซึ่งมีพาธอยู่ที่ /cvsroot/mozilla/layout/html/style/src/nsCSSFrameConstructor.cpp
- แอททริบิวต์ symbolic names คือ รายการของคู่ชื่อแท็ก (Tag Name) กับหมายเลขการเปลี่ยนแปลงแก้ไข ที่นักพัฒนาตั้งชื่อไว้เพื่อสะดวกในการอ้างอิงถึง จากรูปที่ 2-3 ระบุว่าแฟ้มข้อมูลชื่อ nsCSSFrameConstructor.cpp ถูกตั้งชื่อแท็กไว้หลายชื่อ ตัวอย่างเช่นแท็กชื่อ MOZILLA_1_3a_RELEASE คู่กับหมายเลขการเปลี่ยนแปลงแก้ไขที่ 1.800

- แอททริบิวต์ description คือ รายการของการเปลี่ยนแปลงแก้ไขที่เกิดขึ้นกับแฟ้มข้อมูลนี้ เริ่มตั้งแต่ที่แฟ้มข้อมูลนี้ได้ถูกลงคอมมิทเข้าสู่รีพอสิโทริจนถึงเวอร์ชั่นล่าสุดของแฟ้มข้อมูลนี้ นอกจากการเปลี่ยนแปลงแก้ไขที่ถูกบันทึกลงบนลำต้นหลัก (Main Trunk) แล้วทุกๆการเปลี่ยนแปลงแก้ไขที่ถูกบันทึกลงบนกิ่ง (Branches) ก็ถูกบันทึกไว้ในส่วนนี้ด้วยเช่นกัน การเปลี่ยนแปลงแก้ไขแต่ละการเปลี่ยนแปลงแก้ไขจะถูกแบ่งออกจากกันด้วยบรรทัดเครื่องหมายยัติภังค์ (-) ภายในแต่ละส่วนย่อยนี้อธิบายการเปลี่ยนแปลงแก้ไขทั้งหมดที่เคยเกิดขึ้นกับแฟ้มข้อมูล ประกอบด้วย 1) หมายเลขการเปลี่ยนแปลงแก้ไข (Revision Number) 2) วันและเวลา (Date) ที่คอมมิทการเปลี่ยนแปลงแก้ไขนี้เข้ามา 3) นักพัฒนา (Author) ที่เป็นผู้คอมมิท 4) สเตท (Atate) ระบุถึงสถานภาพในขณะนั้นของแฟ้มข้อมูล มีค่าที่เป็นไปได้ 2 ค่า คือ Exp หมายถึง สถานภาพทดสอบ (Experimental State) และ Dead หมายถึง แฟ้มข้อมูลถูกลบไปแล้ว 4) บรรทัด (Lines) ระบุจำนวนบรรทัดที่ถูกเพิ่มเข้าไป (นำหน้าด้วยเครื่องหมายบวก) และจำนวนบรรทัดที่ถูกลบออกไป (นำหน้าด้วยเครื่องหมายลบ) ของแฟ้มข้อมูลสำหรับการคอมมิทครั้งนี้ 5) รายการหมายเลขกิ่ง (Branches) คือรายการของหมายเลขการเปลี่ยนแปลงแก้ไขบนกิ่งที่มีการเปลี่ยนแปลงแก้ไขนี้เป็นจุดต่อกิ่ง และ 6) บรรทัดสุดท้ายของส่วนการเปลี่ยนแปลงแก้ไขระบุข้อความหมายเหตุ (Comment) ที่ผู้พัฒนาเขียนระบุไว้ในการคอมมิท ตัวอย่างรายการการเปลี่ยนแปลงแก้ไขหนึ่งจากรูปที่ 2-3 คือรายการการเปลี่ยนแปลงแก้ไขที่มีหมายเลขการเปลี่ยนแปลงแก้ไขที่ 1.804 บันทึกวันที่ 13 ธันวาคม ค.ศ. 2002 เวลา 20 นาฬิการ 13 นาที 16 วินาที บันทึกโดย doe@netscape.com อยู่ในสถานภาพทดสอบ การเปลี่ยนแปลงแก้ไขครั้งนี้มีบรรทัดที่ถูกเพิ่มเข้าไปจำนวน 15 บรรทัด บรรทัดที่ถูกลบออกไป 47 บรรทัด และมีข้อความหมายเหตุจากผู้บันทึกว่า Don't set NS_BLOCK_SPACE_MGR and NS_BLOCK_WARP_SIZE on ...

2.7 การประยุกต์ใช้การทำเหมืองข้อมูลด้วยเทคนิคค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Applying Association Rule Discovery in Software Archive)

ในช่วงต้นของการคิดค้นและพัฒนาแนวคิดการทำเหมืองข้อมูลด้วยด้วยเทคนิคการค้นหากฎความสัมพันธ์ (Association Rule Mining) นั้น การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ถูกพัฒนามาเพื่อการค้นหารูปแบบความสัมพันธ์ของพฤติกรรมการซื้อสินค้าของลูกค้า ตัวอย่างเช่นการค้นหาว่าลูกค้าที่ซื้อหนังสือ ก. มักจะซื้อหนังสืออะไรด้วย จากฐานข้อมูล

รายการซื้อหนังสือขนาดใหญ่ นอกจากความสามารถในการค้นหารูปแบบความสัมพันธ์ของพฤติกรรมการซื้อสินค้าของลูกค้าแล้ว ต่อจากนั้นมาไม่นานนักเริ่มมีนักวิจัยหลายคณะนำการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์มาประยุกต์ใช้กับข้อมูลในแขนงต่างๆ เช่น ข้อมูลเครือข่ายโทรคมนาคม ข้อมูลการจัดการความเสี่ยง ข้อมูลการควบคุมคลังสินค้า และข้อมูลทางพันธุกรรมของสิ่งมีชีวิต เป็นต้น (Kotsiantis et al., 2006) นอกจากนั้นการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ยังสามารถประยุกต์ใช้กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Software Archive) ของขั้นตอนการพัฒนาซอฟต์แวร์ในวงจรชีวิตการพัฒนาซอฟต์แวร์ได้อีกด้วย คณะนักวิจัยต่างๆนำข้อมูลซอฟต์แวร์อาร์ไคฟว์มาวิเคราะห์ในหลากหลายรูปแบบต่างๆกัน เช่น การวิเคราะห์เพื่อทำความเข้าใจพัฒนาการของการพัฒนาซอฟต์แวร์ (Software Evolution) (Ball et al., 1997) การวิเคราะห์เพื่อดักจับพัฒนาการของการเชื่อมโยงกัน (Evolution Coupling) ระหว่างคลาส (Bieman et al., 2003) หรือระหว่างไฟล์ (Gall et al.,1998; Burch et al., 2005) ต่างๆในระหว่างการพัฒนาโปรแกรม การวิเคราะห์เพื่อดักจับพัฒนาการของการเกิดขึ้นต่อกันระหว่างไฟล์พร้อมระบุลำดับการเปลี่ยนแปลง (Burch et al., 2005) การวิเคราะห์เพื่อดักจับพัฒนาการของการเกิดขึ้นต่อกันอย่างละเอียดระหว่างส่วนย่อยภายในโปรแกรม (Program Entities) อย่างเช่น ระหว่างฟังก์ชั่นหรือระหว่างตัวแปร (ซึ่งจัดว่าละเอียดกว่าการพิจารณาเพียงแค่ระดับระหว่างคลาสหรือระหว่างไฟล์) (Zimmermann et al., 2004) การวิเคราะห์เพื่อค้นหารูปแบบการเรียกใช้ซอฟต์แวร์ไลบรารี่ (Software Libraries) ที่ถูกต้องเพื่อการนำรูปแบบเหล่านั้นกลับมาใช้ใหม่ (Michail, 2000) และรูปแบบการเรียกใช้ซอฟต์แวร์ไลบรารี่ที่ผิดและนำไปสู่การเกิดข้อผิดพลาดได้ (Li et al., 2005; Livshits et al., 2005; Williams et al., 2005) งานวิจัยในอดีตเหล่านี้แสดงให้เห็นถึงประโยชน์ของการทำเหมืองข้อมูลด้วยกฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ ซึ่งมีส่วนช่วยในการทำงานร่วมกันของนักพัฒนาซอฟต์แวร์ระหว่างขั้นตอนการพัฒนาซอฟต์แวร์ในวงจรชีวิตการพัฒนาซอฟต์แวร์เป็นอย่างมาก ประโยชน์ของการทำเหมืองข้อมูลดังกล่าวได้แก่ 1) แนะนำและคาดการณ์ล่วงหน้าว่านักพัฒนาซอฟต์แวร์ควรจะแก้ไขคลาส ไฟล์ หรือฟังก์ชั่นใดต่อไปหลังจากที่ได้แก้ไขคลาส ไฟล์ หรือฟังก์ชั่นหนึ่งไปแล้ว 2) ป้องกันการเกิดข้อผิดพลาดจากการแก้ไขคลาส ไฟล์ หรือฟังก์ชั่นใดอย่างไม่สมบูรณ์ 3) สามารถดักจับการเกิดของการขึ้นต่อกันระหว่างคลาส ไฟล์ หรือฟังก์ชั่นได้ โดยเฉพาะอย่างยิ่งการขึ้นต่อกันที่ไม่สามารถดักจับได้ในระหว่างขั้นตอนการวิเคราะห์และออกแบบซอฟต์แวร์ (Zimmermann et al., 2004)

ในงานวิจัยนี้สนใจเฉพาะการประยุกต์ใช้การทำเหมืองข้อมูลด้วยกฎความสัมพันธ์จากข้อมูลซอฟต์แวร์อาร์ไคฟว์ในขั้นตอนการพัฒนาซอฟต์แวร์ โดยจะเน้นเฉพาะความสัมพันธ์ในระดับ

ที่ละเอียดที่สุด นั่นก็คือในระดับความสัมพันธ์ระหว่างฟังก์ชั่นหรือเมธอด รูปแบบการเรียกใช้ฟังก์ชั่นหรือเมธอดในคลาส (Function-Call-Usage Pattern) คือเซตหรือบัญชีรายการของการเรียกใช้ฟังก์ชั่นหรือเมธอดในคลาสที่พบอยู่ในซอร์สโค้ดของซอฟต์แวร์ รูปแบบการเรียกใช้ฟังก์ชั่นหรือเมธอดในคลาสเหล่านี้มักจะเกิดขึ้นมาจากสัญชาตญาณและองค์ความรู้สามัญของนักพัฒนาซอฟต์แวร์ รูปแบบการเรียกใช้ฟังก์ชั่นหรือเมธอดในคลาสบางรูปแบบก็เกิดขึ้นมาโดยที่นักพัฒนาซอฟต์แวร์ไม่รู้ตัว ซึ่งทั้งหมดนี้มักจะไม่ได้ถูกนำมาจัดทำเป็นเอกสารอย่างเป็นทางการและอยู่นอกเหนือความสามารถของการตรวจหาจุดบกพร่องของระบบตรวจจุดบกพร่องภายในซอฟต์แวร์ด้วย

ในช่วงปี 2005 คณะวิจัยของ Li และคณะวิจัยของ Livshits (Li et al., 2005; Livshits et al., 2005) ได้ทำการประยุกต์เทคนิคการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ในการค้นหารูปแบบการเรียกใช้ฟังก์ชั่นหรือเมธอดในคลาสของระบบซอฟต์แวร์ขนาดใหญ่เพื่อที่จะแก้ไขปัญหาดังที่กล่าวไว้ งานวิจัยของคณะวิจัยทั้ง 2 ได้แสดงให้เห็นว่าการประยุกต์เทคนิคการทำเหมืองข้อมูลด้วยกฎความสัมพันธ์ในการค้นหารูปแบบการเรียกใช้ฟังก์ชั่นหรือเมธอดในคลาสนั้นเป็นเทคนิคที่ค่อนข้างมีประสิทธิภาพมากแต่ในบางกรณีนั้นสามารถทำให้เกิดผลลัพธ์ของการค้นหาที่เป็นผลบวกลวง (False Positive) เป็นจำนวนมาก กล่าวคือการใช้เทคนิคดังกล่าวอาจก่อให้เกิดกฎความสัมพันธ์ที่เป็นไปได้ออกมาโดยความจริงแล้วไม่ได้มีกฎความสัมพันธ์นั้นอยู่จริงๆ

นอกจากการนั้นการใช้เทคนิคการค้นหากฎความสัมพันธ์ก็ยังคงมีข้อด้อยที่สำคัญในการค้นหารูปแบบที่เรียกใช้ฟังก์ชั่นหรือเมธอดในคลาสก็คือ ธรรมชาติการเรียกใช้ฟังก์ชั่นหรือเมธอดในคลาสนั้นลำดับของการเรียกใช้ฟังก์ชั่นหรือเมธอดในคลาสย่อมมีความสำคัญ เช่น เมธอด open() ย่อมถูกเรียกใช้งานก่อนเมธอด close() แต่การใช้เทคนิคการค้นหากฎความสัมพันธ์ไม่สามารถบ่งชี้ถึงลำดับได้ (Huzefa Kagdi et al., 2007)

ต่อมาไม่นานงานวิจัยของ Burch และคณะ (Burch et al., 2005) ได้ทำการประยุกต์การทำเหมืองข้อมูลด้วยเทคนิคการค้นหารูปแบบลำดับ (Sequential pattern Mining) กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ เพื่อแก้ไขข้อบกพร่องการใช้เทคนิคค้นหากฎความสัมพันธ์ที่ไม่ได้ให้ความสำคัญกับลำดับของการเรียกฟังก์ชั่นหรือเมธอดในคลาส ผลลัพธ์ที่ได้จากการใช้เทคนิคการค้นหารูปแบบลำดับนั้นมีความแม่นยำสูงมากและให้ผลลัพธ์ที่เป็นผลบวกลวงน้อยกว่าเทคนิคอื่นๆ แต่อย่างไรก็ดีการทำเหมืองข้อมูลด้วยเทคนิคการค้นหารูปแบบลำดับนั้นมาพร้อมกับค่าใช้จ่ายในการประมวลผลที่สูงมาก ด้วยเหตุนี้ การทำเหมืองข้อมูลเพื่อค้นหารูปแบบที่มีความสัมพันธ์กันใน

ข้อมูลซอฟต์แวร์อาร์ไคฟว์ในทางปฏิบัตินั้นเหมาะกับการใช้เทคนิคการค้นหากฎความสัมพันธ์มากกว่าการใช้เทคนิคการค้นหารูปแบบลำดับถึงแม้ว่าการใช้เทคนิคการค้นหารูปแบบลำดับจะให้ผลลัพธ์ที่แม่นยำกว่าก็ตาม (Burch et al., 2005)

2.8 ขั้นตอนวิธีการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Data Mining in Software Archives)

วัตถุประสงค์ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Data Mining in Software Archives) นั้นก็คือ การสร้างเซตของคำแนะนำในการเปลี่ยนแปลงแก้ไขให้กับนักพัฒนาในระหว่างขั้นตอนการพัฒนาซอฟต์แวร์เมื่อนักพัฒนาได้ทำให้เกิดเหตุการณ์ใดเหตุการณ์หนึ่งเกิดขึ้น การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์มีขั้นตอนหลักทั้งหมด 2 ขั้นตอน คือ 1) การจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Preparing Data for Mining in Software Archives) และ 2) การทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Data Mining in Software Archives) รายละเอียดของแต่ละขั้นตอนอธิบายได้ดังต่อไปนี้

2.8.1 การจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Preparing Data for Mining in Software Archives)

ในปี ค.ศ. 1997 คณะวิจัยของ Ball เป็นคณะวิจัยแรกที่ได้เริ่มทำงานวิจัยเกี่ยวกับข้อมูลซอฟต์แวร์อาร์ไคฟว์และได้พบว่าข้อมูลซอฟต์แวร์อาร์ไคฟว์ซึ่งประกอบด้วยแฟ้มข้อมูลซอร์สโค้ดและแฟ้มข้อมูลบันทึกของซอฟต์แวร์ควบคุมการเปลี่ยนแปลงแก้ไขนั้นสามารถแสดงได้ถึงพัฒนาการของการพัฒนาซอฟต์แวร์หรือระบบนั้นๆ และยังสามารถแสดงถึงการเชื่อมโยงกันหรือความสัมพันธ์กันระหว่าง 2 คลาส (Class) ภายในซอฟต์แวร์หรือระบบได้ด้วย นอกจากนั้นยังแสดงให้เห็นถึงความสัมพันธ์ของสิ่งต่างๆ (aspects) ในขั้นตอนการพัฒนาซอฟต์แวร์ที่มีผลต่อการเปลี่ยนแปลงข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Ball et al., 1997)

งานวิจัยของ Ball และคณะในปี ค.ศ. 1997 นี้ถือเป็นจุดเริ่มต้นแรกให้เกิดงานวิจัยเกี่ยวกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ขึ้นมาอีกมากมายในเวลาต่อมา ซึ่งสามารถยกตัวอย่างได้ดังต่อไปนี้

1) การประยุกต์ใช้การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ เพื่อช่วยเหลือการทำงานของนักพัฒนาซอฟต์แวร์อาทิเช่น คาดการณ์และแนะนำนักพัฒนาซอฟต์แวร์ว่าควรจะต้องแก้ไขซอร์สโค้ดส่วนไหนต่อไป แสดงการเชื่อมโยงกันของคลาสที่เกิดขึ้นในระหว่างการพัฒนาซึ่งไม่สามารถดักจับพบได้ในช่วงของการออกแบบซอฟต์แวร์ และป้องกันไม่ให้เกิดข้อผิดพลาดที่เกิดขึ้นจากการแก้ไขซอร์สโค้ดไม่สมบูรณ์ (Zimmermann et al., 2004; Ying et al., 2004; Livshits et al., 2005; Weißgerber et al., 2005; Yu et al., 2007)
2) การประยุกต์ใช้การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ เพื่อค้นหารูปแบบการเรียกใช้ซอฟต์แวร์ไลบรารี่ (Software Libraries) ที่ถูกต้องเพื่อการนำรูปแบบเหล่านั้นกลับมาใช้ใหม่ (Michail, 2000) และค้นหารูปแบบการเรียกใช้ซอฟต์แวร์ไลบรารี่ที่ผิดและนำไปสู่การเกิดข้อผิดพลาดได้ (Li et al., 2005; Livshits et al., 2005; Williams et al., 2005)
3) การจินตทัศน์ผลของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ เพื่อแสดงให้นักพัฒนาซอฟต์แวร์เห็นถึงพัฒนาการของการเชื่อมโยงกัน (Evoluation Coupling) ของคลาสที่เกิดขึ้นในระหว่างการพัฒนาซอฟต์แวร์ (Burch et al., 2005; Voinea et al., 2005; Voinea et al., 2006; Weissgerber et al., 2007)

การทำเหมืองข้อมูลกับข้อมูลประเภทต่างๆนั้นจะต้องมีขั้นตอนสามัญขั้นตอนหนึ่งที่ต้องทำก่อนเป็นอันดับแรกเสมอนั้นก็คือขั้นตอนการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูล (Preparing Data for Mining) หรือที่เรียกว่า ขั้นตอนก่อนกระบวนการทำเหมืองข้อมูล (Preprocessing) งานวิจัยทุกงานวิจัยเกี่ยวกับการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ที่กล่าวไปในข้างต้นนั้นก็ต้องผ่านการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Preparing Data for Mining in Software Archives) ด้วยเช่นกัน ขั้นตอนการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลนี้ถือเป็นขั้นตอนที่มีความสำคัญมากสำหรับการทำเหมืองข้อมูลในทุกๆประเภท เพราะขั้นตอนการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลนั้นจะส่งผลกระทบโดยตรงกับคุณภาพของการวิเคราะห์ผลลัพธ์ที่จะได้มาจากการทำเหมืองข้อมูล

ในหัวข้อนี้จะกล่าวถึงขั้นตอนสำคัญ 4 ขั้นตอนที่จะเกิดขึ้นในการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ ที่ถูกเสนอขึ้นมาโดย Zimmermann และคณะ ในปี

2004 (Zimmermann et al., 2004) และเป็นขั้นตอนวิธีที่มีงานวิจัยที่เกี่ยวข้องกับการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์หลายงายวิจัย (Zimmermann et al., 2005; Livshits et al., 2005; Williams et al., 2005; Breu et al., 2006; Weißgerber et al, 2006) นำไปใช้ ขั้นตอนทั้ง 4 ขั้นตอนได้แก่ 1) การสกัดข้อมูล (Data Extraction) 2) การซ่อมแซมทรานแซคชั่น (Restoring Transactions) 3) การระบุการเปลี่ยนแปลงแก้ไขในระดับเอนทิตี้ (Mapping Changes to Entities) และ 4) การกำจัดสิ่งแปลกปลอม (Data Cleaning) แต่ก่อนที่จะเริ่มการอธิบายรายละเอียดของการทำงานในแต่ละขั้นตอนข้างต้น ผู้วิจัยจำเป็นต้องนิยามคำศัพท์และสัญลักษณ์ที่เกี่ยวข้องกับการอธิบายขั้นตอนการทำงานดังต่อไปนี้

- เอนทิตี้ (Entity) คือ เอกลักษณ์หรือสิ่งที่ผู้วิจัยสนใจศึกษา ในที่นี้คำว่า เอนทิตี้ สามารถหมายถึง แฟ้มข้อมูลเอกสาร คลาส เมธอดหรือฟังก์ชัน และตัวแปร นิยามให้เอนทิตี้ e ถูกเขียนอยู่ในรูปแบบ (c, i, p) โดยที่ c คือหมวดหรือประเภทของเอนทิตี้นั้น (syntactic category) i คือชื่อของเอนทิตี้นั้น (identifier) และ p คือเอนทิตี้ที่เป็นเอนทิตี้แม่ของเอนทิตี้นั้นหรือใช้สัญลักษณ์ … แทนในกรณีที่เอนทิตี้นั้นคือเอนทิตี้รากหรือในกรณีที่ต้องการละไว้ในฐานที่เข้าใจ ตัวอย่างของเอนทิตี้เช่น (method, initDefaults(), (Class, Comp, (file, Comp.java,…))) แสดงถึงเมธอดชื่อ initDefaults() ของคลาส Comp ในแฟ้มข้อมูล Comp.java
- การเปลี่ยนแปลงแก้ไข (Changes, Revisions) คือ เหตุการณ์ที่มีนักพัฒนาแก้ไขเอนทิตี้ใดๆ คำว่า เปลี่ยนแปลงแก้ไข ในที่นี้สามารถแสดงได้ 3 มิติ คือ 1) การเปลี่ยนแปลงเอนทิตี้ (alter) ใช้สัญลักษณ์ alter(e) แทนการเปลี่ยนแปลงอะไรบางอย่างภายในเอนทิตี้ e 2) การเพิ่มลงในเอนทิตี้ (add to) ใช้สัญลักษณ์ add_to(e) แทนการเพิ่มเอนทิตี้ใหม่เข้าไปในเอนทิตี้ e 3) การลบออกจากเอนทิตี้ (delete from) ใช้สัญลักษณ์ del_from(e) แทนการลบเอนทิตี้ใดๆออกจากเอนทิตี้ e
- ทรานแซคชั่น (Transaction) คือ เซตของการเปลี่ยนแปลงแก้ไขที่เกิดขึ้นพร้อมกันและถูกคอมมิทเข้าสู่ระบบ โดยหนึ่งทรานแซคชั่นใดๆจะเป็นของนักพัฒนาเพียงคนเดียวเท่านั้น ตัวอย่างของทรานแซคชั่นเช่น T = {alter(method, initDefaults(), …), alter(field, fKeys[], …), add_to(file, Comp.java, …)} การ

เปลี่ยนแปลงแก้ไขที่อยู่ภายในทรานแซคชั่นบางครั้งอาจถูกเรียกว่า รายการ (item)

- เหตุการณ์ (Situation) คือเซตของการเปลี่ยนแปลงแก้ไขใดๆ ใช้สัญลักษณ์ Q แทนเหตุการณ์ ตัวอย่างของเหตุการณ์เช่น Q = {alter(method, initDefaults(), …)}
- กฎความสัมพันธ์ (Association Rules) ที่ได้จากการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ คือกฎความสัมพันธ์ที่ตอบข้อถามที่ว่า *ถ้านักพัฒนาเปลี่ยนแปลงแก้ไข (เปลี่ยนแปลง เพิ่มลง หรือ ลบออก) เอนทิตี้ใดเอนทิตี้หนึ่งแล้วนักพัฒนาคนนั้นควรจะต้องเปลี่ยนแปลงแก้ไขอะไรด้วย* ตัวอย่างของกฎความสัมพันธ์สามารถเขียนได้ดังนี้ {alter(field, fKeys[], …)} → {alter(method, initDefaults(), …), alter(file, plug.properties, …)} แทนกฎความสัมพันธ์ที่ว่าเมื่อมีการเปลี่ยนแปลงตัวแปรชื่อ fKey[] แล้วจะมีการเปลี่ยนแปลงเมธอดชื่อ initDefaults() และการเปลี่ยนแปลงแฟ้มข้อมูลชื่อ plug.properties ด้วย
- เซตของคำแนะนำสำหรับเหตุการณ์ Q (The Set of Suggestions for Situation Q) คือ เซตของการเปลี่ยนแปลงแก้ไขที่นักพัฒนาควรจะทำตามหลังจากที่นักพัฒนาได้เปลี่ยนแปลงแก้ไขตามเหตุการณ์ Q โดยอ้างอิงมาจากเซตของกฎความสัมพันธ์ R เซตของคำแนะนำสำหรับเหตุการณ์ Q สามารถเขียนเป็นสัญลักษณ์ได้คือ $apply_R(\mathrm{Q}) = \mathrm{U}_{(\mathrm{Q}\rightarrow x_2)\in R}\, x_2$ ตัวอย่างเซตของคำแนะนำสำหรับเหตุการณ์ Q เช่น $apply_R(\mathrm{Q})$ = {alter(method, initDefaults(), …), add_to(file, Comp.java, …)}

เมื่อได้ทำความเข้าใจกับนิยามและสัญลักษณ์ของคำศัพท์แล้ว ส่วนต่อไปนี้คือการอธิบายรายละเอียดของการทำงานในแต่ละขั้นตอนของการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์

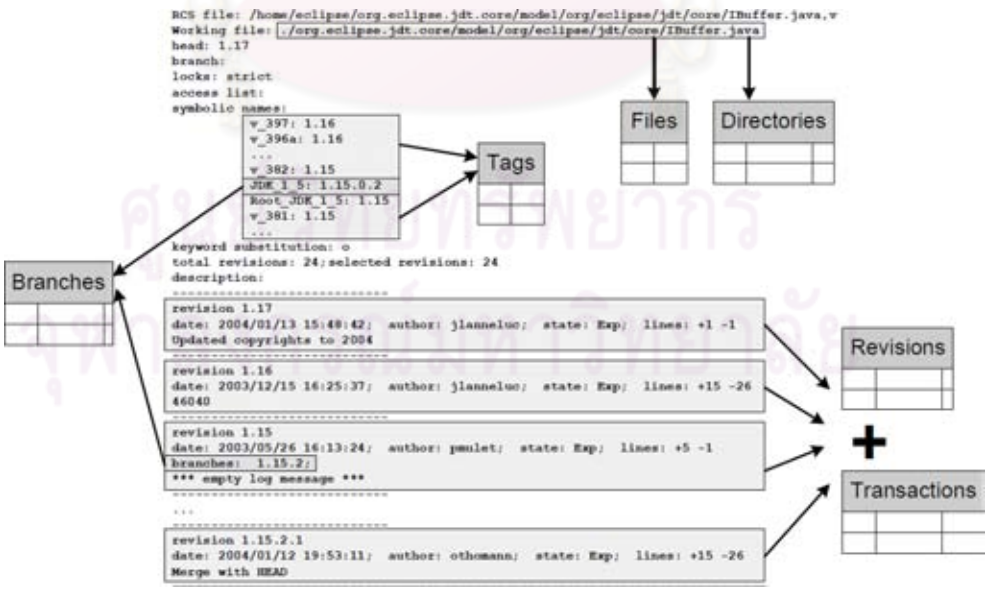
2.8.1.1 การสกัดข้อมูล (Data Extraction)

จุดประสงค์หนึ่งของการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ คือ เพื่อให้สามารถเข้าถึงข้อมูลภายในระบบคอนเคอเรนท์เวอร์ชั่นได้อย่างรวดเร็วในขณะที่

กำลังทำเหมืองข้อมูล วิธีการที่ดีที่สุดที่จะบรรลุจุดประสงค์นั้นได้ก็คือการสำเนาแฟ้มข้อมูลบันทึกจากรีพอสิโทริของระบบคอนเคอเรนท์เวอร์ชั่น ทำการสกัดข้อมูลจากแฟ้มข้อมูลบันทึกนั้น และนำมาบันทึกไว้บนฐานข้อมูลของผู้วิจัยเอง (Zimmermann et al., 2004)

โดยทั่วไปแล้ว ข้อมูลที่จะถูกสกัดออกมาจะออกมาเป็นอย่างไรนั้นจะขึ้นอยู่กับว่ามีความต้องการวิเคราะห์ข้อมูลอะไร ตัวอย่างเช่น ถ้ามีความต้องการที่จะวิเคราะห์พัฒนาการของซอฟต์แวร์ (Software Evolution) ก็จะให้ความสนใจกับข้อมูลทุกอย่างที่บันทึกเอาไว้รวมถึงแฟ้มข้อมูลที่ถูกลบไปแล้วด้วย (Zimmermann et al., 2004) แต่ถ้ามีความต้องการที่จะวิเคราะห์เพื่อให้คำแนะนำกับนักพัฒนาในการแก้ไขแฟ้มข้อมูลที่เกี่ยวข้องกัน ก็จะให้ความสนใจกับข้อมูลที่บันทึกเอาไว้ในปัจจุบันเท่านั้น (Zimmermann et al., 2004)

การสกัดข้อมูลของระบบคอนเคอเรนท์เวอร์ชั่นเริ่มต้นจากการเรียกใช้คำสั่ง CVS log จากไดเรคทอรี่ราก (Root Directory) ของโครงการพัฒนาซอฟต์แวร์ที่ต้องการ ผลลัพธ์ที่ส่งกลับคืนมาก็คือแฟ้มข้อมูลซอร์สโค้ด (Source Code Files) และแฟ้มข้อมูลบันทึก (Log File) ของระบบคอนเคอเรนท์เวอร์ชั่นที่บันทึกอยู่บนรีพอสิโทริ แฟ้มข้อมูลบันทึกที่ได้มาจะถูกนำมาวิเคราะห์รูปแบบไวยากรณ์ (Parse) และถูกนำไปบันทึกลงฐานข้อมูลตามแผนภาพข้างล่างนี้ (Zimmermann et al., 2004)



รูปที่ 2-4 แสดงการทำงานของขั้นตอนการสกัดข้อมูล (Data Extraction)

ข้อมูลที่ได้จากการสกัดข้อมูลจากแฟ้มข้อมูลบันทึกก็คือ 1) แฟ้มข้อมูล (Files) ทั้งหมดที่อยู่ในโครงการนี้ ทั้งที่เป็นแฟ้มข้อมูลซอร์สโค้ดและแฟ้มข้อมูลอื่นๆ ด้วย 2) ไดเรคทอรี่ (Directories) ทั้งหมดที่อยู่ในโครงการนี้ 3) การเปลี่ยนแปลงแก้ไขแฟ้มข้อมูล (Revisions) 4) ทรานแซคชั่นของการเปลี่ยนแปลงแก้ไข (Transactions) 5) การตั้งชื่อแท็ก (Tags) และ 6) การต่อกิ่ง (Branches) รายละเอียดของขั้นตอนการสกัดข้อมูลสามารถอธิบายได้ดังต่อไปนี้ (Zimmermann et al., 2004)

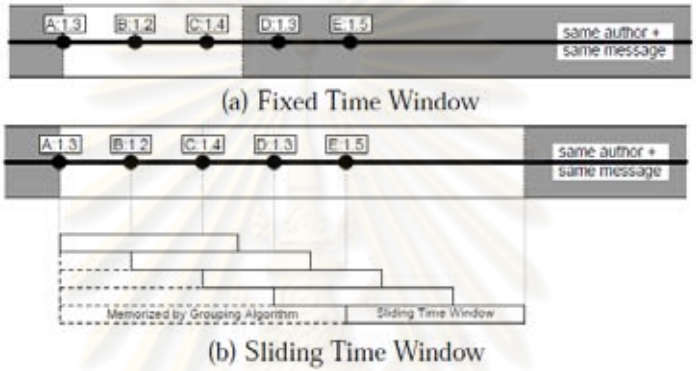
- แอททริบิวต์ RCS file ของแต่ละส่วน (Sections) ในแฟ้มข้อมูลบันทึกสามารถสกัดข้อมูลออกมาเป็นรายชื่อและรายละเอียดของแฟ้มข้อมูล (Files) และไดเรคทอรี่ (Directories) ทั้งหมดของโครงการ ตัวอย่างจากรูปข้างต้นจะได้ แฟ้มข้อมูลชื่อ IBuffer.java และไดเรคทอรี่ ./org.eclipse.jdt.core/Model/org/eclipse/jdt/core/
- แอททริบิวต์ description ประกอบด้วยส่วนย่อยหลายส่วนแต่ละส่วนแสดงถึงการเปลี่ยนแปลงแก้ไขแต่ละครั้งที่เกิดขึ้น ข้อมูลในแต่ละส่วนย่อยนี้สามารถสกัดข้อมูลออกมาเป็นรายการการเปลี่ยนแปลงแก้ไขแฟ้มข้อมูล (Revisions) ได้ ตัวอย่างจากรูปข้างต้นจะได้ รายการการเปลี่ยนแปลงแก้ไขแฟ้มข้อมูลทั้งหมด 4 รายการ คือ รายการการเปลี่ยนแปลงแก้ไขหมายเลข 1.15.2.1 1.15 1.16 และ 1.17
- รายการการเปลี่ยนแปลงแก้ไขแฟ้มข้อมูลที่ได้มาข้างต้นจะถูกนำมาพิจารณาว่ารายการใดบางที่เกิดขึ้นในเวลาเดียวกันและเกิดขึ้นโดยนักพัฒนาคนเดียวกันจะถูกรวมกันไว้เป็นทรานแซคชั่นของการเปลี่ยนแปลงแก้ไข (Transactions) เดียวกัน ตัวอย่างจากรูปข้างต้นจะได้ว่ามีทรานแซคชั่นทั้ง 4 ทรานแซคชั่นและแต่ละทรานแซคชั่นประกอบด้วยรายการการเปลี่ยนแปลงแก้ไขเพียง 1 รายการ
- แอททริบิวต์ symbolic name ในแต่ละส่วนย่อยของแอททริบิวต์ description ในแฟ้มข้อมูลบันทึกสามารถสกัดข้อมูลออกมาเป็นรายชื่อของแท็ก (Tags) ที่นักพัฒนาตั้งไว้ให้กับการเปลี่ยนแปลงแก้ไขนั้นๆได้ ตัวอย่างจากรูปข้างต้นจะได้ว่าสำหรับแฟ้มข้อมูล IBuffer.java มีการตั้งชื่อแท็กทั้งหมดหลายชื่อและมีชื่อหนึ่งคือ V_397 ที่ตั้งไว้ให้กับรายการการเปลี่ยนแปลงแก้ไขหมายเลข 1.16
- การต่อกิ่ง (Branches) สามารถสกัดมาจากรายการการเปลี่ยนแปลงแก้ไขและชื่อแท็ก ตัวอย่างจากรูปข้างต้น เช่น ชื่อแท็ก JDK_1_5 ตั้งให้รายการการเปลี่ยนแปลงแก้ไขหมายเลข 1.15.0.2 จะได้ชื่อของการต่อกิ่งนี้เป็น 1.15.2 ส่วนจุดต่อกิ่งนั้นได้มาจากตารางแฮช (Hash Map) ที่ใช้ชื่อของการต่อกิ่งเป็นคีย์

2.8.1.2 การซ่อมแซมทรานแซคชั่น (Restoring Transactions)

ระบบคอนเคอเรนท์เวอร์ชั่นโดยทั่วไปนั้นจะไม่มีการบันทึกเอาไว้ว่าแฟ้มข้อมูลใดบ้างที่ถูกเปลี่ยนแปลงแก้ไขร่วมกันในการคอมมิทแต่ละครั้ง แต่ว่าข้อมูลการเปลี่ยนแปลงแก้ไขดังกล่าวมีความจำเป็นต่อการนำมาวิเคราะห์ข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Zimmermann et al., 2004; Gall et al., 2003) วิธีการให้ได้มาซึ่งข้อมูลการเปลี่ยนแปลงแก้ไขที่เกิดร่วมกันในแต่ละการคอมมิทก็คือการเข้าไปอ่านข้อมูลที่ถูกบันทึกเอาไว้ในแฟ้มข้อมูลบันทึก (Log File) บนเครื่องแม่ข่ายของระบบคอนเคอเรนท์เวอร์ชั่นว่ามีข้อความบันทึก (Log Message) ใดบ้างที่ระบุการเปลี่ยนแปลงแก้ไขทั้งหมดที่เกิดจากนักพัฒนาคนเดียวกันและเกิดขึ้นในเวลาเดียวกัน ข้อมูลการเปลี่ยนแปลงแก้ไขที่เกิดจากนักพัฒนาคนเดียวกันและเกิดขึ้นในเวลาเดียวกันจะถูกนำมาเปลี่ยนเป็นทรานแซคชั่น 1 ทรานแซคชั่นแล้วบันทึกลงในตาราง Transactions บนฐานข้อมูล คำว่า ในเวลาเดียวกัน ในที่นี้นั้นหมายรวมถึงในเวลาที่ใกล้เคียงกันด้วย เนื่องจากการคอมมิทในแต่ละครั้งอาจใช้เวลาในการดำเนินการหลายวินาทีหรือหลายนาที โดยเฉพาะอย่างยิ่งการคอมมิทที่ละหลายๆแฟ้มข้อมูล (Zimmermann et al., 2004) ดังนั้นในแต่ทางปฏิบัติแล้วนอกจากการพิจารณาที่การคอมมิทในเวลาเดียวกันแล้วยังต้องมีวิธีการพิจารณาที่การคอมมิทระหว่างช่วงของเวลา (Time Interval) เดียวกันด้วย วิธีการพิจารณาการเปลี่ยนแปลงแก้ไขระหว่างช่วงของเวลานั้นมี 2 วิธีคือ

1) วิธีกำหนดกรอบเวลาที่แน่นอน (Fixed Time Windows) คือ การกำหนดกรอบของช่วงเวลามากสุดที่จะพิจารณาว่าการเปลี่ยนแปลงแก้ไขดังกล่าวนั้นอยู่ภายในทรานแซคชั่นเดียวกันหรือไม่ โดยกำหนดให้กรอบของช่วงเวลาจะเริ่มต้นขึ้นใหม่เมื่อมีการเปลี่ยนแปลงแก้ไขใหม่เกิดขึ้นครั้งแรกและกรอบจะคงที่เช่นนี้จนจบการพิจารณา การเปลี่ยนแปลงแก้ไขใดที่อยู่ในกรอบนี้ทั้งหมดจะถูกพิจารณาว่าอยู่ในทรานแซคชั่นเดียวกัน วิธีนี้เป็นวิธีที่ง่ายและสะดวกในการนำไปประยุกต์ใช้ วิธีการพิจารณาที่การเปลี่ยนแปลงแก้ไขระหว่างช่วงของเวลาเดียวโดยกำหนดกรอบเวลาที่แน่นอนถูกนำไปใช้ในงานวิจัยของ Ubranic´ และคณะ (Ubranic´ et al., 2003) และในงานวิจัยของ Gall และคณะในปี ค.ศ. 2003 (Gall et al., 2003)
2) วิธีเลื่อนกรอบเวลา (Sliding Time Windows) คือ การกำหนดช่องว่างระหว่างการเปลี่ยนแปลงแก้ไข 2 ครั้งที่มากที่สุด จุดเริ่มต้นของกรอบของช่วงเวลาจะถูกเลื่อนไปที่การเปลี่ยนแปลงแก้ไขครั้งต่อไปเสมอตราบใดที่การเปลี่ยนแปลงแก้ไขครั้งต่อไปนั้นมี

จุดเริ่มต้นอยู่ภายในกรอบเวลาของการเปลี่ยนแปลงแก้ไขครั้งก่อนหน้า ดังนั้นการพิจารณาด้วยวิธีเลื่อนกรอบเวลานี้จะสามารถรู้จำ (Recognize) การคอมมิทที่ใช้ระยะเวลาในการดำเนินการนานกว่าจะสมบูรณ์ได้ดีกว่าการพิจารณาด้วยวิธีกำหนดกรอบเวลาที่แน่นอน วิธีเลื่อนกรอบเวลานี้มีต้นกำเนิดมาจากโปรแกรมประยุกต์ที่มีชื่อว่า cvs2cl (http://www.red-bean.com/cvs2cl) และ CVSps (http://www.cobite.com/cvsps) (Zimmermann et al., 2004)



รูปที่ 2-5 แสดงการพิจารณาช่วงเวลาของการคอมมิทด้วยวิธีกำหนดกรอบเวลาที่แน่นอนและวิธีเลื่อนกรอบเวลา

รูปที่ 2-5 แสดงการพิจารณาช่วงเวลาของการคอมมิทด้วยวิธีกำหนดกรอบเวลาที่แน่นอนและวิธีเลื่อนกรอบเวลา (Zimmermann et al., 2004) ส่วน (a) แสดงการพิจารณาด้วยวิธีกำหนดกรอบเวลาที่แน่นอน บริเวณสีขาวเป็นบริเวณที่อยู่ภายในกรอบเวลาที่กำหนดเอาไว้แน่นอน ดังนั้นการแก้ไขเปลี่ยนแปลงที่แฟ้มข้อมูล A เวอร์ชั่นที่ 1.3 แฟ้มข้อมูล B เวอร์ชั่นที่ 1.2 และแฟ้มข้อมูล C เวอร์ชั่นที่ 1.4 จะถูกพิจารณาว่าเกิดขึ้นพร้อมกันและอยู่ภายในทรานแซคชั่นเดียวกัน รูปที่ 2-5 ส่วน (b) แสดงการพิจารณาด้วยวิธีเลื่อนกรอบเวลา โดยเริ่มต้นกรอบเวลาที่มีช่วงแน่นอนเริ่มต้นที่จุดการแก้ไขเปลี่ยนแปลงที่แฟ้มข้อมูล A เวอร์ชั่นที่ 1.3 และกรอบเวลาถูกเลื่อนไปเรื่อยๆ จนสิ้นสุดดังรูป ดังนั้นการแก้ไขเปลี่ยนแปลงที่แฟ้มข้อมูล A เวอร์ชั่นที่ 1.3 แฟ้มข้อมูล B เวอร์ชั่นที่ 1.2 แฟ้มข้อมูล C เวอร์ชั่นที่ 1.4 แฟ้มข้อมูล D เวอร์ชั่นที่ 1.3 และ แฟ้มข้อมูล E เวอร์ชั่นที่ 1.5 จะถูกพิจารณาว่าเกิดขึ้นพร้อมกันและอยู่ภายในทรานแซคชั่นเดียวกัน

โดยปกติ การใช้วิธีเลื่อนกรอบเวลาสำหรับทุกการเปลี่ยนแปลงแก้ไข α_1, α_2, …, α_k (เรียงตามลำดับเวลาที่บันทึก (time(α_i))) ที่เป็นส่วนหนึ่งของทรานแซคชั่น T เดียวกันนั้น จะต้องอยู่ภายใต้เงื่อนไข

$$\forall \alpha_i \in T : author(\alpha_i) = author(\alpha_1)$$

$$\forall \alpha_i \in T : \log\_message(\alpha_i) = \log\_message(\alpha_1)$$

$$\forall i \in \{2, \dots, k\} : |time(\alpha_i) - time(\alpha_{i-1})| \leq 200sec$$

นอกจากนั้นการเปลี่ยนแปลงแก้ไขเวอร์ชั่นของแต่ละแฟ้มข้อมูลจะปรากฏอยู่บน 1 ทรานแซคชั่นได้เพียงครั้งเดียว เนื่องจากระบบคอนเคอเรนท์เวอร์ชั่นไม่อนุญาตให้มีการคอมมิทการเปลี่ยนแปลงเวอร์ชั่นของแฟ้มข้อมูลเดียวกัน 2 ครั้งในเวลาเดียวกันได้ ดังนั้นจึงมีเงื่อนไขเพิ่มอีก 1 ข้อดังนี้

$$\forall \alpha_a, \alpha_b \in T : \alpha_a \neq \alpha_b \rightarrow file(\alpha_a) \neq file(\alpha_b)$$

ขั้นตอนวิธีในการรวมกลุ่ม (grouping) การเปลี่ยนแปลงแก้ไขเวอร์ชั่นของแต่ละแฟ้มข้อมูลให้มาเป็นทรานแซคชั่นนั้น เป็นขั้นตอนวิธีที่เรียบง่ายและตรงไปตรงมา ดังนี้ (Zimmermann et al., 2004)

1) เรียงลำดับการเปลี่ยนแปลงแก้ไขเวอร์ชั่นของแต่ละแฟ้มข้อมูลตาม เวลา ชื่อของนักพัฒนาที่ทำการเปลี่ยนแปลง และข้อความในแฟ้มข้อมูลบันทึก (log messager)
2) เริ่มต้นทรานแซคชั่นที่ i โดยวนซ้ำพิจารณาแต่ละ การเปลี่ยนแปลงแก้ไขเวอร์ชั่นของแต่ละแฟ้มข้อมูล ถ้ากรอบเวลาของทรานแซคชั่นปัจจุบันจบลง หรือ ชื่อนักพัฒนา เวลา และ/หรือ ข้อความในแฟ้มข้อมูลบันทึกของการเปลี่ยนแปลงแก้ไขเวอร์ชั่นของแฟ้มข้อมูลที่ i แตกต่างจากชื่อนักพัฒนา เวลา และ/หรือ ข้อความในแฟ้มข้อมูลบันทึกของการเปลี่ยนแปลงแก้ไขเวอร์ชั่นของแฟ้มข้อมูลที่ i-1 แล้ว ถือเป็นการสิ้นสุดทรานแซคชั่นที่ i
3) วนซ้ำข้อ 2 ไปจนกว่าจะหมดข้อมูลเปลี่ยนแปลงแก้ไขเวอร์ชั่นของแต่ละแฟ้มข้อมูล

งานวิจัยของ Zimmermann และคณะในปี ค.ศ. 2004 (Zimmermann et al., 2004) งานวิจัยของ Livshits และคณะในปี ค.ศ. 2005 (Livshits et al., 2005) งานวิจัยของ Weißgerber และคณะในปี ค.ศ. 2005 (Weißgerber et al., 2005) และปี ค.ศ. 2007 (Weißgerber et al., 2007) เลือกพิจารณาการเปลี่ยนแปลงแก้ไขเวอร์ชั่นระหว่างช่วงของเวลาด้วยวิธีเลื่อนกรอบเวลาและกำหนดช่วงของกรอบเวลาอยู่ที่ 200 วินาที

2.8.1.3 การระบุการเปลี่ยนแปลงแก้ไขในระดับเอนทิตี้ (Mapping Changes to Entities)

ข้อมูลที่ถูกจัดเก็บไว้ในรีพอสิโทริของระบบคอนเคอเรนท์เวอร์ชั่นนั้นมีเพียงข้อมูลแฟ้มข้อมูลทุกแฟ้มข้อมูลในโครงการและข้อมูลการเปลี่ยนแปลงแก้ไขในระดับแฟ้มข้อมูล (File) หรือคลาส (Class) ที่เก็บอยู่ในรูปของแฟ้มข้อมูลบันทึก (Log File) เท่านั้น แต่ไม่มีบันทึกว่าการเปลี่ยนแปลงแก้ไขที่เกิดขึ้นนั้นเกิดขึ้นกับฟังก์ชั่น (Function) หรือ เมธอด (Method) ใดบ้าง มีตัวแปร (Variable) ใดถูกเพิ่มเข้ามา แก้ไข หรือถูกลบออกไปบ้าง (Zimmermann et al., 2004)

การเปรียบเทียบความแตกต่างอย่างละเอียดถึงในระดับตัวแปร และ ฟังก์ชั่น หรือ เมธอด นี้มีวิธีการหลายวิธี วิธีการอย่างง่ายๆก็คือการเปรียบเทียบในระดับแฟ้มข้อมูล โดยการประยุกต์ใช้ฟังก์ชั่นดิฟฟ์ (diff) (Miller et al., 1985) กับแต่ละบรรทัดของซอร์สโค้ดระหว่างแฟ้มข้อมูลของเวอร์ชั่นเก่ากับแฟ้มข้อมูลเวอร์ชั่นใหม่ ผลลัพธ์ที่ได้คือส่วนของโค้ดที่มีการเปลี่ยนแปลง ส่วนที่มีการเพิ่มเข้าใหม่และส่วนที่ถูกลบออกไป วิธีการนี้มีข้อเสียที่สำคัญอยู่ 2 ประการคือ 1) คุณภาพของผลลัพธ์ที่ได้จะขึ้นอยู่กับคุณภาพของฟังก์ชั่นดิฟฟ์ (diff) ที่นำมาใช้ 2) วิธีการนี้ระบุความแตกต่างได้แค่ในระดับบรรทัดของซอร์สโค้ด ไม่สามารถระบุได้ว่าเป็นตรงไหนของบรรทัดนั้น (Zimmermann et al., 2004)

วิธีการที่มีความแม่นยำสูงกว่าแต่ก็มีค่าใช้จ่ายในการคำนวณสูงวิธีหนึ่ง คือการกำหนดเอนทิตี้ (ตัวแปร ฟังก์ชั่นหรือเมธอด คลาส แฟ้มข้อมูล) ทั้งหมดภายในแฟ้มข้อมูลทั้ง 2 เวอร์ชั่นโดยการนำไปผ่านตัววิเคราะห์ไวยากรณ์ (Parser) จากนั้นก็ทำการเปรียบเทียบซอร์สโค้ดของเอนทิตี้เดียวกันใน 2 เวอร์ชั่น หรือกล่าวคือเป็นการประยุกต์ใช้ฟังก์ชั่นดิฟฟ์ (diff) ในระดับของเอนทิตี้นั่นเอง วิธีการนี้สามารถดำเนินการได้ดังต่อไปนี้ (Zimmermann et al., 2004)

1) กำหนดเซต E_1 คือเซตของเอนทิตี้ที่มีอยู่ทั้งหมดในเวอร์ชั่น r_1 ของแฟ้มข้อมูล และกำหนดเซต E_2 คือเซตของเอนทิตี้ที่มีอยู่ทั้งหมดในเวอร์ชั่น r_2 ของแฟ้มข้อมูลเดียวกัน
2) เอนทิตี้ที่ถูกเพิ่มเข้ามาใหม่สามารถหาได้จาก $E_2 - E_1$
3) เอนทิตี้ที่ถูกลบออกไปสามารถหาได้จาก $E_1 - E_2$
4) ทุกๆเอนทิตี้ที่อยู่ในเซต $E_1 \cap E_2$ อาจจะเป็นเอนทิตี้ที่มีการเปลี่ยนแปลงภายใน การตัดสินว่าเอนทิตี้ใดบ้างที่มีการเปลี่ยนแปลงแก้ไขสามารถทำได้โดยการประยุกต์ใช้ฟังก์ชั่นดิฟฟ์ (diff) กับซอร์สโค้ดของเอนทิตี้นั้นๆของทั้ง 2 เวอร์ชั่น

ภายในแพลตฟอร์ม (Platform) ของอิคลิพซ์ (Eclipse) นั้นมีการจัดเตรียมโครงร่าง (Framework) สำหรับการเปรียบเทียบความแตกต่างระหว่าง 2 ซอร์สโค้ดใดๆที่มีประสิทธิภาพสูง และสามารถนำไปประยุกต์เพิ่มเติมได้ วิธีการทั้ง 2 วิธีการที่ได้กล่าวไปข้างต้นนั้นสามารถนำมาประยุกต์ใช้จริงได้โดยเรียกใช้ 2 โครงร่างดังต่อไปนี้ (Zimmermann et al., 2004)

1) โครงร่างเรนจ์ดิฟเฟอร์เรนเซอร์ (Range Differencer) คลาสเรนจ์ดิฟเฟอร์เรนเซอร์ (RangeDifferencer Class) ทำการเปรียบเทียบความแตกต่างของแฟ้มข้อมูล 2 เวอร์ชั่น โดยใช้โทเคน (Token) เป็นฐาน วิธีการนี้อ้างอิงวิธีการในการเปรียบเทียบมาจากขั้นตอนวิธีดิฟฟ์ (diff Algorithm) ดังเดิมของ Miller และ Myers (Miller et al., 1985) โทเคนแต่ละโทเคนจะถูกสร้างขึ้นมาโดยคลาสที่มีการอิมพลีเมนต์ (Implementing) อินเตอร์เฟส (Interface) ชื่อว่า ไอโทเคนคอมพาราเตอร์ (ITokenComparator interface) ตัวอย่างคลาสที่มีการอิมพลีเมนต์โทเคนคอมพาราเตอร์ คือ คลาสดอคไลน์คอมพาราเตอร์ (DocLineComparator Class) ที่สามารถคำนวณความแตกต่างของแต่ละบรรทัดในคลาสได้และให้ผลลัพธ์ออกมาเป็นรายการ (List object) ของบรรทัดที่แตกต่างกันระหว่าง 2 เวอร์ชั่นของแฟ้มข้อมูลที่นำมาเปรียบเทียบกัน
2) โครงร่างสตรัคเชอร์เมิร์จวิวเวอร์ (Structure Merge Viewer) คลาสดิฟเฟอร์เรนเซอร์ (Differrencer Class) ทำการเปรียบเทียบความแตกต่างของแฟ้มข้อมูล 2 เวอร์ชั่นโดยใช้โครงสร้างลำดับขั้น (hierarchical Structure) เป็นฐาน ผลลัพธ์ที่ได้จากการเปรียบเทียบคือต้นไม้ที่อธิบายการความแตกต่างระหว่าง 2 เวอร์ชั่นของแฟ้มข้อมูลอย่างละเอียด โครงสร้างลำดับขั้นแต่ละโครงสร้างที่เป็นตัวแทนของแฟ้มข้อมูลนั้นๆจะถูกสร้างขึ้นมาโดยคลาสที่มีการอิมพลีเมนต์อินเตอร์เฟสชื่อว่าไอสตรัคเชอร์ครีเอเตอร์ (IStructureCreator Interface) แต่ถ้าไม่ต้องการสร้างคลาสที่มีการอิมพลีเมนต์อินเตอร์เฟสไอสตรัคเชอร์ครีเอเตอร์ มาใช้เองก็สามารถใช้คลาสจาวาสตรัคเชอร์ครีเอเตอร์ (JavaStructureCreator Class) ที่แพลตฟอร์มอิคลิพซ์จัดเตรียมไว้ให้แล้วสำหรับการสร้างโครงสร้างลำดับขั้นของคลาสแฟ้มข้อมูลภาษาจาวา (Java)

2.8.1.4 การกำจัดสิ่งแปลกปลอม (Data Cleaning)

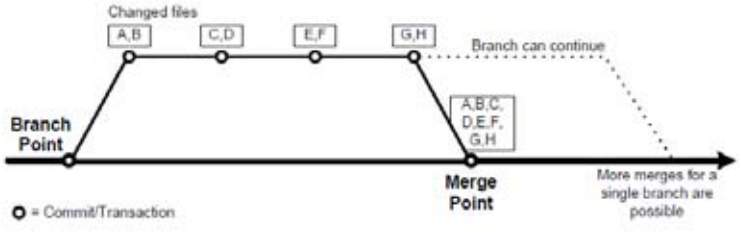
ในหัวข้อก่อนหน้านี้อธิบายถึงการสกัดข้อมูล การซ่อมแซมทรานแซคชั่น และการแปลงการเปลี่ยนแปลงแก้ไขไปสู่ระดับเอนทิตี้ ซึ่งผลลัพธ์ที่ได้มานั้นยังมีสิ่งแปลกปลอม (Noise) ปะปนมาด้วย ขั้นตอนการกำจัดสิ่งแปลกปลอม (Data Cleaning) เป็นขั้นตอนที่เข้าไปตรวจสอบข้อมูลทั้งหมดเพื่อค้นหาสิ่งแปลกปลอมและกำจัดสิ่งแปลกปลอมเหล่านั้นออกไป ลักษณะของข้อมูลทรานแซคชั่นที่จะถูกระบุว่าเป็นสิ่งแปลกปลอมมีอยู่ 2 ลักษณะคือ 1) ทรานแซคชั่นขนาดใหญ่ (Large Transactions) และ 2) ทรานแซคชั่นการผสานกิ่ง (Merge Transactions) รายละเอียดของสิ่งแปลกปลอมและวิธีการในการกำจัดสิ่งแปลกปลอมทั้ง 2 ลักษณะนี้สามารถอธิบายได้ดังต่อไปนี้ (Zimmermann et al., 2004)

- ทรานแซคชั่นขนาดใหญ่ (Large Transactions)

ทรานแซคชั่นขนาดใหญ่เป็นเหตุการณ์ปกติที่สามารถเกิดขึ้นได้จริงกับข้อมูลจริงในทุกประเภท ข้อมูลที่ได้จากระบบคอนเคอเรนท์เวอร์ชั่นนั้นอาจมีข้อมูลที่ไม่เกี่ยวกับข้องกับการนำไปวิเคราะห์มาปะปนอยู่ด้วย ข้อมูลที่ไม่เกี่ยวข้องกับการวิเคราะห์อาทิ ข้อความที่ระบุการเปลี่ยนแปลงแก้ไขแฟ้มข้อมูลที่เป็นการแก้ไขโครงสร้างภายในของแต่ละแฟ้มข้อมูลเองที่เกิดขึ้นต่อเนื่องกันแต่ไม่ได้มีความสัมพันธ์เชิงตรรกะระหว่างแฟ้มข้อมูลเหล่านั้น เช่น ข้อความ “Chage #include filenames from <foo.h> [sign] to <openssl.h>.” ที่ปรากฏต่อเนื่องกัน 552 ครั้งสำหรับการแก้ไขแฟ้มข้อมูล 552 แฟ้มข้อมูล ข้อความลักษณะดังกล่าวจะถูกระบุว่าเป็นสิ่งแปลกปลอมที่อาจส่งผลให้ผลลัพธ์ของการวิเคราะห์ผิดพลาดได้ วิธีกำจัดสิ่งแปลกปลอมลักษณะนี้ทำได้โดยการกรอง (filter out) ทรานแซคชั่นที่มีขนาดใหญ่เกินกว่าค่า N ที่กำหนดไว้ออก

- ทรานแซคชั่นการผสานกิ่ง (Merge Transactions)

ในระหว่างที่นักพัฒนากำลังพัฒนาซอฟต์แวร์ด้วยระบบคอนเคอเรนท์เวอร์ชั่นอยู่นั้นอาจมีการสร้างกิ่ง (Branches) ของเวอร์ชั่นขึ้นมาหลากหลายกิ่ง กิ่งเหล่านั้นบางกิ่งอาจจะถูกผสานกิ่ง (Merge) เข้ากับลำต้นเวอร์ชั่นหลักในอนาคต แต่บางกิ่งของเวอร์ชั่นก็ถูกปล่อยทิ้งเอาไว้



รูปที่ 2-6 แสดงตัวอย่างการต่อกิ่งและผสานกิ่ง

รูปที่ 2-6 แสดงตัวอย่างการต่อกิ่งและการผสานกิ่งอย่างง่ายในระบบคอนเคอเรนท์เวอร์ชั่น กิ่งที่ต่อออกมานั้นประกอบด้วยการคอมมิททั้งหมด 4 ทรานแซคชั่น ได้แก่ ทรานแซคชั่น {A, B} ทรานแซคชั่น {C, D} ทรานแซคชั่น {E, F} และทรานแซคชั่น {G, H} แฟ้มข้อมูล A B C D E F G และ H เหล่านี้จะถูกเปลี่ยนแปลงแก้ไขอีกครั้งตอนที่มีการผสานกิ่งเข้ากับลำต้นเกิดเป็นทรานแซคชั่น {A, B, C, D, E, F, G, H} ที่จุดผสาน (Merge Point) และเรียกทรานแซคชั่น {A, B, C, D, E, F, G, H} ว่า ทรานแซคชั่นการผสานกิ่ง

ทรานแซคชั่นการผสานกิ่งถูกระบุว่าเป็นสิ่งแปลกปลอมด้วยสาเหตุสำคัญ 2 สาเหตุคือ 1) ทรานแซคชั่นการผสานกิ่งเป็นทรานแซคชั่นที่ประกอบด้วยการเปลี่ยนแปลงแก้ไขแฟ้มข้อมูลที่ไม่มีความสัมพันธ์เกี่ยวข้องกันอย่างแท้จริง ตัวอย่างเช่น ทรานแซคชั่นการผสานกิ่งข้างต้นมีการเปลี่ยนแปลงแก้ไขแฟ้มข้อมูล B และการเปลี่ยนแปลงแก้ไขแฟ้มข้อมูล C อยู่ภายใน ซึ่งในความเป็นจริงทั้ง 2 แฟ้มข้อมูลไม่มีความสัมพันธ์เกี่ยวข้องกัน 2) ทรานแซคชั่นการผสานกิ่งถูกสร้างมาจากการนำการเปลี่ยนแปลงแก้ไขที่มีทั้งหมดบนกิ่งนั้นมาเรียงต่อกัน ดังนั้นทรานแซคชั่นการผสานกิ่งจึงเป็นทรานแซคชั่นที่มีความซ้ำซ้อนกับทรานแซคชั่นอื่นที่มีอยู่บนกิ่งนั้นอยู่แล้ว

จากเหตุผลทั้ง 2 ข้อข้างต้น การนำทรานแซคชั่นการผสานกิ่งไปใช้ในการทำเหมืองข้อมูลด้วย อาจทำให้ผลลัพธ์ที่ได้จากการทำเหมืองข้อมูลผิดพลาดได้ ดังนั้นทรานแซคชั่นการผสานกิ่งที่จะเกิดขึ้นทุกครั้งที่มีการผสานกิ่งเกิดขึ้นจะต้องถูกค้นหาให้พบและกำจัดออก

2.8.2 **การทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์** (Data Mining in Software Archives)

หลังจากที่ข้อมูลการเปลี่ยนแปลงแก้ไขภายในรีพอสิโทรีของระบบคอนเคอเรนท์เวอร์ชั่นได้ถูกนำมาดำเนินการตามขั้นตอนการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์

อาร์ไคฟ์ที่กล่าวไปข้างต้นและได้ผลลัพธ์ของการดำเนินการออกมาเป็นข้อมูลรายการทรานแซคชั่นแล้ว ข้อมูลรายการทรานแซคชั่นเหล่านั้นจะถูกนำมาเป็นข้อมูลเข้าสำหรับการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟ์ (Data Mining in Software Archives) ผลลัพธ์หรือข้อมูลออกที่ได้จากการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟ์ก็คือ กฎความสัมพันธ์ที่สามารถตอบข้อถามที่ว่า *ถ้านักพัฒนาเปลี่ยนแปลงแก้ไข (เปลี่ยนแปลง เพิ่มลง หรือ ลบออก) เอนทิตี้ใดเอนทิตี้หนึ่งแล้วนักพัฒนาคนนั้นควรจะต้องเปลี่ยนแปลงแก้ไขอะไรด้วย* ตัวอย่างเช่น

{alter(field, fKeys[], …)} → {alter(method, initDefaults(), …), alter(file, plug.properties, …)}

กฎความสัมพันธ์ในข้างต้นนั้นหมายความว่า เมื่อไรก็ตามที่นักพัฒนามีการเปลี่ยนแปลงอะไรบางอย่างกับตัวแปรชื่อ fkey[] แล้วนักพัฒนาควรจะต้องไปเปลี่ยนแปลงแก้ไขอะไรบางอย่างในเมธอดชื่อ initDefaults() และเปลี่ยนแปลงแก้ไขอะไรบางอย่างในแฟ้มข้อมูลชื่อ plug.properties

โดยทั่วไปแล้ว กฎความสัมพันธ์ r จะอยู่ในรูปของคู่ (x_1, x_2) โดยที่เซตรายการ x_1 และ x_2 ไม่มีส่วนซ้อนทับกัน (disjoint) หรืออยู่ในรูปของ $x_1 \rightarrow x_2$ ซึ่ง x_1 จะถูกเรียกว่าเซตรายการที่มาก่อน (Antecedent Itemset) และ x_2 จะถูกเรียกว่าเซตรายการที่ตามมา (Consequent Itemset) กฎความสัมพันธ์แต่ละกฎนั้นถูกสร้าง (derived) ขึ้นมาจากการพิจารณาความน่าจะเป็นจากทรานแซคชั่นที่เกิดขึ้นจริงในอดีต วิธีที่นิยมนำมาใช้ในการประเมินความน่าสนใจของแต่ละกฎความสัมพันธ์คือการพิจารณาค่าสนับสนุน (Support) และค่าความเชื่อมั่น (Confidence) ของกฎความสัมพันธ์ แต่สำหรับการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟ์นั้นจะใช้วิธีการประเมินความน่าสนใจของแต่ละกฎความสัมพันธ์โดยการพิจารณาที่ค่าสนับสนุนนับ (Support Count) และค่าความเชื่อมั่น (Confidence) ของกฎความสัมพันธ์ (Zimmermann et al., 2005) นิยามของค่าสนับสนุนนับ และค่าความเชื่อมั่น ของกฎความสัมพันธ์สำหรับการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟ์นั้น สามารถอธิบายได้ดังนี้

- ค่าสนับสนุนนับ (Support Count) คือ จำนวนของทรานแซคชั่นที่กฎความสัมพันธ์นั้นปรากฏอยู่ ตัวอย่างเช่น สมมุติให้ตัวแปร fKey[] เคยปรากฏว่าถูกเปลี่ยนแปลงแก้ไขไปทั้งหมด 11 ทรานแซคชั่น ภายใน 11 ทรานแซคชั่นนั้นมี 10 ทรานแซคชั่นที่มีการเปลี่ยนแปลงแก้ไขเมธอด initDefaults() และแฟ้มข้อมูล plug.properties ด้วย ดังนั้นค่าสนับสนุนนับของกฎความสัมพันธ์ {alter(field, fKeys[], …)} → {alter(method, initDefaults(), …), alter(file, plug.properties, …)} เท่ากับ 10 การทำเหมืองข้อมูลกับ

ข้อมูลซอฟต์แวร์อาร์ไคฟว์นั้นใช้ค่าสนับสนุนนับ แทนที่การใช้ค่าสนับสนุนอย่างในการทำเหมืองข้อมูลทั่วๆไปเพราะสาเหตุสำคัญ 2 ประการคือ 1) ค่าสนับสนุนนับ สามารถสื่อสารให้นักพัฒนาเข้าใจได้ดีกว่าค่าสนับสนุน กล่าวคือ ค่าสนับสนุนนับ เท่ากับ 10 หมายถึงที่ผ่านมาในอดีตเคยมีเหตุการณ์ตามกฎความสัมพันธ์เกิดขึ้นมาแล้วทั้งหมด 10 ครั้ง แต่ค่าสนับสนุน เท่ากับ 0.000145 นั้นไม่สามารถสื่อสารอะไรได้เลยถ้าไม่ได้บอกข้อมูลเพิ่มเติมว่าจำนวนทรานแซคชั่นที่มีอยู่ทั้งหมดนั้นเท่ากับเท่าไร 2) ค่าสนับสนุนนับ สามารถถูกนำไปใช้ในการวิเคราะห์อื่นๆได้ ตัวอย่างเช่น นอกจากการใช้กฎความสัมพันธ์ในการให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาแล้ว กฎความสัมพันธ์สามารถบอกได้ถึงพัฒนาการของการเชื่อมโยงกันระหว่างแฟ้มข้อมูล (Evoluation Coupling) ได้ ในการระบุพัฒนาการของการเชื่อมโยงกันระหว่างแฟ้มข้อมูลนั้นใช้แค่เพียงค่าสนับสนุนนับก็สามารถระบุได้ เพราะจำนวนทรานแซคชั่นทั้งหมดมีผลต่อการระบุพัฒนาการของการเชื่อมโยงกันระหว่างแฟ้มข้อมูลน้อยมาก (Zimmermann et al., 2005)

- ค่าความเชื่อมั่น (Confidence) คือ ค่าความน่าจะเป็นที่จะพบกฎความสัมพันธ์ในทรานแซคชั่นที่มีเซตรายการที่มาก่อนอยู่ จากตัวอย่างข้างต้นค่าความเชื่อมั่นของกฎความสัมพันธ์ {alter(field, fKeys[], …)} → {alter(method, initDefaults(), …), alter(file, plug.properties, …)} เท่ากับ 10/11 = 0.909

นิยามการเขียนกฎความสัมพันธ์พร้อมระบุค่าสนับสนุนนับ และค่าความเชื่อมั่น ในรูปสัญลักษณ์ดังนี้ r [s;c] โดยที่ r แทนกฎความสัมพันธ์ s แทนค่าสนับสนุนนับของกฎความสัมพันธ์นั้น และ c แทนค่าความเชื่อมั่นของกฎความสัมพันธ์นั้น ตัวอย่างเช่น {alter(field, fKeys[], …)} → {alter(method, initDefaults(), …)} [4;0.57]

สมมุติให้นักพัฒนามีการเปลี่ยนแปลงแก้ไขบางอย่างลงไป ทำให้เกิดเซตรายการของการเปลี่ยนแปลงแก้ไขขึ้นมา นิยามเซตรายการของการเปลี่ยนแปลงแก้ไขดังกล่าวว่าเป็นเซตเหตุการณ์ (Situation) และใช้สัญลักษณ์ Q แทนเหตุการณ์ เซตเหตุการณ์จะถูกปรับปรุง (Update) ทุกครั้งที่นักพัฒนาทำการบันทึกการเปลี่ยนแปลงแก้ไขลงสู่เครื่องลูกข่ายของนักพัฒนาเอง เมื่อนักพัฒนามั่นใจในเวอร์ชั่นใหม่ของแต่ละแฟ้มข้อมูลที่นักพัฒนาแก้ไขไปแล้วทำการคอมมิทเพื่อบันทึกการเปลี่ยนแปลงแก้ไขทั้งหมดลงสู่รีพอสิโทริของระบบคอนเคอเรนท์เวอร์ชั่น เซตเหตุการณ์สุดท้ายก็จะถูกเพิ่มเข้าไปฐานข้อมูลทรานแซคชั่น ตัวอย่างของเซตเหตุการณ์เช่น

Q = {alter(field, fKeys[], …)}

สมการข้างต้นแสดงเซตเหตุการณ์ที่ประกอบด้วย 1 รายการการเปลี่ยนแปลงแก้ไข ระบุว่านักพัฒนาได้ทำการเปลี่ยนแปลงเอนทิตี้ระดับตัวแปรชื่อ fKey[] ในแฟ้มข้อมูล ComparePerferencePage.java (ในสมการข้างต้นละการเขียนถึงเอนทิตี้แฟ้มข้อมูล ComparePerferencePage.java เอาไว้โดยแทนที่ด้วย ... เพื่อความสะดวกในการเขียน)

จุดมุ่งหมายในการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ของงานวิจัยนี้คือ การให้คำแนะนำนักพัฒนาว่าควรจะเปลี่ยนแปลงแก้ไขที่ใดต่อไปเมื่อนักพัฒนาทำให้เกิดเหตุการณ์นั้นๆ ขึ้น ในหัวข้อต่อไปนี้จะอธิบายถึงขั้นตอนวิธีในการสร้างกฎความสัมพันธ์จากการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ และอธิบายการสร้างคำแนะนำจากกฎความสัมพันธ์ที่ได้มา

2.8.2.1 การสร้างกฎความสัมพันธ์จากการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์

ขั้นตอนวิธีในการสร้างกฎความสัมพันธ์ที่ได้รับความนิยมมากที่สุดก็คือ ขั้นตอนวิธีอพริโอริ (Apriori algorithm) ที่ถูกนำเสนอโดย Agrawal และ Srikant ในปี ค.ศ. 1994 (Agrawal and Srikant, 1994) ขั้นตอนวิธีอพริโอรินั้นจำเป็นต้องระบุค่าสนับสนุนขั้นต่ำ (Minimum Support) (สำหรับงานวิจัยนี้เป็นค่าสนับสนุนนับขั้นต่ำ (Minimum Support Count) แทน) และค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence) เพื่อใช้ในการคำนวณผลลัพธ์ของขั้นตอนวิธีอพริโอริคือเซตของกฎความสัมพันธ์ทั้งหมดที่มีค่าสนับสนุนนับมากกว่าค่าสนับสนุนนับขั้นต่ำและมีค่าความเชื่อมั่นมากกว่าค่าความเชื่อมั่นขั้นต่ำ

การนำขั้นตอนวิธีอพริโอริมาประยุกต์ใช้กับการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์โดยคำนวณหากฎความสัมพันธ์ทั้งหมดเอาไว้ก่อนล่วงหน้าและสร้างคำแนะนำที่เหมาะสมกับเหตุการณ์ที่นักพัฒนาสร้างขึ้น แต่วิธีการดังกล่าวสามารถทำให้เกิดปัญหาตามมาได้ เนื่องจากเหตุการณ์ใหม่ๆที่เกิดขึ้นหลังจากการสร้างกฎความสัมพันธ์เก็บไว้แล้วจะไม่ได้ถูกนำไปสร้างเป็นทรานแซคชั่นและนำไปเป็นส่วนหนึ่งของข้อมูลที่ใช้ในการทำเหมืองข้อมูล ดังนั้นเมื่อเวลาผ่านไปเซตของกฎความสัมพันธ์ที่สร้างเอาไว้ล่วงหน้าจะได้ถูกต้องตรงกับความเป็นจริง วิธีที่ถูกต้องก็คือกฎความสัมพันธ์ทั้งหมดจะต้องถูกคำนวณใหม่อยู่เสมอทุกครั้งที่นักพัฒนาทำให้เกิดเหตุการณ์ใหม่ๆ แต่ในความเป็นจริงการคำนวณโดยใช้ขั้นตอนวิธีอพริโอริกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ใช้เวลานานมาก (ในกรณีโครงการใหญ่ๆและผ่านระยะเวลาการพัฒนามานานๆ ต้องใช้เวลาในการคำนวณหลายวัน) ปี ค.ศ. 2005 Zimmermann และคณะได้เสนอวิธีการในปรับปรุงขั้นตอนวิธีอพริ

โอริมาเพื่อให้เหมาะสม (Optimization) กับการทำเหมืองข้อมูลของข้อมูลซอฟต์แวร์อาร์ไคฟว์ โดยการเพิ่มข้อกำหนด 2 ข้อให้กับขั้นตอนวิธีอพริโอริ (Zimmermann et al., 2005) ดังนี้

- การค้นหากฎความสัมพันธ์เฉพาะกฎที่มีเซตรายการที่มาก่อนที่ต้องการเท่านั้น กล่าวคือ การทำเหมืองข้อมูลจะเกิดขึ้นในขณะที่ผู้พัฒนากำลังทำให้เกิดเหตุการณ์ขึ้นและค้นหาเฉพาะกฎความสัมพันธ์ที่มีเซตที่เหตุการณ์ Q ที่นักพัฒนาพึ่งจะกระทำเป็นเซตรายการที่มาก่อนของกฎ (Srikant et al., 1997) การกระทำเช่นนี้ทำให้การคำนวณทั้งหมดใช้เวลาเพียงเล็กน้อยเท่านั้น ข้อดีอีกประการหนึ่งของการใช้การค้นหากฎความสัมพันธ์เฉพาะกฎที่มีเซตรายการที่มาก่อนคือ วิธีการนี้สามารถกระทำได้แบบเพิ่มขึ้นต่อเนื่อง (Incrementally) ได้ จึงทำให้เหตุการณ์ใหม่ๆที่พึ่งการจะถูกนำไปสร้างเป็นทรานแซคชั่นและมีผลต่อการคำนวณหากฎความสัมพันธ์ในครั้งถัดไป
- การกำหนดให้ทุกกฎความสัมพันธ์ที่ค้นหามีเซตรายการที่ตามมาเพียง 1 รายการเท่านั้น การกระทำนี้ทำให้การทำเหมืองข้อมูลเร็วขึ้นกว่าเดิมมาก ดังนั้นกฎความสัมพันธ์ที่ได้หลังจากที่นักพัฒนากระทำให้เกิดเหตุการณ์ Q ก็คือ $Q \rightarrow \{e\}$ โดยที่ e คือสมาชิกของเซตรายการที่ตามมา การกำหนดให้ทุกกฎความสัมพันธ์ที่ค้นหามีเซตรายการที่ตามมาเพียงรายการเดียวนี้ไม่ได้ส่งผลกระทบให้ผลลัพธ์ของคำแนะนำที่จะได้นั้นผิดเพลี้ยนไปเพราะถึงแม้ว่าเซตของคำแนะนำที่ได้จะมีสมาชิกมากกว่า 1 รายการแต่นักพัฒนาก็สามารถกระทำตามคำแนะนำได้เพียงครั้งละ 1 รายการอยู่ดี หรือการพิสูจน์ทางคณิตศาสตร์ดังนี้ กำหนดให้ รายการ $e \in x_2$ ของกฎความสัมพันธ์ $Q \rightarrow x_2[s;c]$ ดังนั้นย่อมมีกฎความสัมพันธ์ r ที่มีเซตรายการที่ตามมารายการเดียวคือ กฎความสัมพันธ์ $Q \rightarrow \{e\}[s_r;c_r]$ ซึ่ง $s_r \geq s$ และ $c_r \geq c$ เนื่องจาก $Q \cup \{e\} \subseteq Q \cup x_2$ ทำให้ จำนวนรายการของ $Q \cup \{e\} \geq Q \cup x_2$

2.8.2.2 การสร้างคำแนะนำจากกฎความสัมพันธ์ (Generating Suggestions for Situation)

การสร้างเซตของคำแนะนำ (Suggestions) จากเซตของกฎความสัมพันธ์ R สำหรับเหตุการณ์ Q ที่นักพัฒนากระทำออกมาสามารถนิยามให้อยู่ในรูปของการยูเนียน (Union) ของเซตรายการที่ตามมาของกฎความสัมพันธ์ R ที่มีเซตรายการที่มาก่อนตรงกับเซตเหตุการณ์ Q ดังนี้

$$apply_R(\mathrm{Q}) = \mathrm{U}_{(\mathrm{Q}\rightarrow\{x_2\})\in R}\, x_2$$

ตัวอย่างการสร้างคำแนะนำจากกฎความสัมพันธ์ {alter(field, fKeys[], …)} → {alter(method, initDefaults(), …), alter(file, plug.properties, …)} สำหรับเหตุการณ์ Q จะทำให้ได้ผลลัพธ์คือเซตของคำแนะนำคือ {alter(method, initDefaults(), …), alter(file, plug.properties, …)} โดยที่คำแนะนำการเปลี่ยนแปลงแก้ไขที่อยู่ภายในเซตของคำแนะนำนั้นจะถูกเรียงลำดับตามค่าสนับสนุนนับ และค่าความเชื่อมั่น จำนวนของคำแนะนำภายในเซตของคำแนะนำนี้จะขึ้นอยู่กับการกำหนดค่าสนับสนุนนับ และค่าความเชื่อมั่นขั้นต่ำ โดยปกติแล้วมักจะเริ่มต้นกำหนดค่าสนับสนุนนับขั้นต่ำเป็น 1 และค่าความเชื่อมั่นขั้นต่ำเป็น 0.1

ในงานวิจัยของ Zimmermann และคณะ (Zimmermann et al., 2005) ได้ตั้งข้อสันนิษฐานไว้ว่า การให้คำแนะนำในการเปลี่ยนแปลงแก้ไขกับนักพัฒนานั้น คำแนะนำที่จะได้รับความสนใจจากนักพัฒนาก็คือคำแนะนำที่อยู่ใน 10 อันดับแรกเท่านั้น ดังนั้นในการสร้างเซตของคำแนะนำจึงควรให้ความสนใจกฎความสัมพันธ์ที่อยู่ใน 10 อันดับแรกโดยเรียงจากค่าสนับสนุนนับและค่าความเชื่อมั่นเท่านั้น ดังนั้นผู้วิจัยจึงกำหนดสมการในการสร้างเซตของคำแนะนำดังแสดงในสมการต่อไปนี้

กำหนดให้ q คือ ข้อสอบถาม (Query) ที่ประกอบด้วยเซตเหตุการณ์ (Situation) Q และเซตผลลัพธ์ที่คาดไว้ (Expected Result) E และเขียนให้อยู่ในรูป q = (Q, E)

R คือ เซตของกฎความสัมพันธ์ที่อยู่ในรูปแบบ Q→{x} โดยที่ x คือรายการการเปลี่ยนแปลงแก้ไข และกำหนดให้ R_{10} คือเซตของกฎความสัมพันธ์ที่มีระดับความน่าสนใจสูงสุด 10 กฎแรกซึ่งเรียงลำดับด้วยค่าความเชื่อมั่น โดยที่ $R_{10} \subset R$

A_q คือ เซตของรายการการเปลี่ยนแปลงแก้ไข x ที่ได้จากกฎความสัมพันธ์ใน

เซต R_{10} ที่สอดคล้องกับเซตเหตุการณ์ Q ของข้อสอบถาม q ซึ่งสามารถเขียนในรูป $A_q = apply_{R_{10}}(Q)$ ดังนั้นขนาดของเซต A_q จะน้อยกว่าหรือเท่ากับ 10 เสมอ

$$A_q = apply_{R_{10}}(Q)$$

สมมุติว่าเมื่อนักพัฒนาเห็นเซตของคำแนะนำแล้ว นักพัฒนาจะตัดสินใจดำเนินการเปลี่ยนแปลงแก้ไขตามคำแนะนำแรกเสมอ (คำแนะนำที่มีค่าความเชื่อมั่นสูงที่สุด) เมื่อนักพัฒนากระทำตามคำแนะนำนั้นก็จะทำให้เกิดเหตุการณ์ใหม่ขึ้นและเหตุการณ์ใหม่นั้นก็จะถูกนำไปสร้างเซตของคำแนะนำใหม่จากกฎความสัมพันธ์ที่ได้จากการทำเหมืองข้อมูลกับทรานแซคชั่นทั้งหมดที่รวมเอาเหตุการณ์ครั้งก่อนหน้าเอาไว้ด้วย

ขั้นตอนวิธีที่กล่าวมาทั้งหมดในหัวข้อ 2.8.1 และ 2.8.2 นี้ ได้ถูกนำไปสร้างเป็นระบบให้คำแนะนำสำหรับนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ชื่อโปรแกรมประยุกต์อีโรส (eROSE) โดย Zimmermann และคณะ (Zimmermann et al., 2005) โปรแกรมประยุกต์อีโรสประกอบด้วย 3 ส่วน คือ 1) ส่วนการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ 2) ส่วนการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ และ 3) ส่วนการให้คำแนะนำนักพัฒนาระหว่างการพัฒนาซอฟต์แวร์

2.9 การวัดประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์

การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์นั้นสามารถกระทำได้ในสถานการณ์ที่แตกต่างกัน 3 สถานการณ์ ได้แก่ 1) สถานการณ์การนำทาง (Navigation) คือ สถานการณ์ที่นักพัฒนามีการเปลี่ยนแปลงแก้ไขที่เอนทิตี้หนึ่งแล้ว ระบบจะให้คำแนะนำกับนักพัฒนาให้แก้ไขเอนทิตี้ใดต่อไปได้ถูกต้องหรือไม่ 2) สถานการณ์การป้องกันการเกิดข้อผิดพลาด (Error Prevention) คือ สถานการณ์ที่นักพัฒนามีการเปลี่ยนแปลงแก้ไขที่เอนทิตี้หลายๆเอนทิตี้ต่อเนื่องกันแต่ยังขาดการเปลี่ยนแปลงแก้ไขเอนทิตี้อีกหนึ่งเอนทิตี้จึงจะสมบูรณ์ ระบบจะให้คำแนะนำกับนักพัฒนาให้แก้ไขเอนทิตี้ที่เหลือนั้นได้ถูกต้องหรือไม่ 3) สถานการณ์การ

เปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (Closure) คือ สถานการณ์ที่นักพัฒนามีการเปลี่ยนแปลงแก้ไขที่เอนทิตี้หลายๆเอนทิตี้ต่อเนื่องกันจนสมบูรณ์แล้ว ระบบจะให้คำแนะนำที่เป็นผลบวกลวง (False Positive) เป็นผลลัพธ์แก่นักพัฒนาหรือไม่

การวัดประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์วใช้วิธีการวัดประสิทธิภาพเหมือนกับการวัดประสิทธิภาพของการค้นคืนข้อมูล (Information Retrieval) (Zimmermann et al., 2005) ก็คือการคำนวณหาค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของแต่ละหน่วยทดลอง จากนั้นนำค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ที่ได้มาคำนวณหาค่าเอฟเมสเซอร์ (F-measure) ซึ่งค่าเอฟเมสเซอร์นี้ก็คือค่าประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์วในสถานการณ์การนำทาง (Navigation) และสถานการณ์การป้องกันการเกิดข้อผิดพลาด (Error Prevention) นั่นเอง นอกนั้นผู้วิจัยยังสามารถคำนวณค่าผลสะท้อนกลับ (Feedback) หรือค่าร้อยละของข้อสอบถามที่ไม่ได้ให้เซตรายการการเปลี่ยนแปลงแก้ไขที่ถูกดึงขึ้นมาเป็นเซตว่างกับข้อสอบถามทั้งหมด ซึ่งค่าผลสะท้อนกลับนี้ก็คือค่าประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์วในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (Closure)

การใช้ค่าเอฟเมสเซอร์เป็นค่าที่ใช้การระบุระดับประสิทธิภาพของการค้นคืนข้อมูลเป็นที่นิยมในงานวิจัยที่เกี่ยวข้องกับการค้นคืนข้อมูล เช่น งานวิจัยการขยายคำในข้อสอบถามโดยรวมกฎความสัมพันธ์กับสิ่งที่ศึกษาร่วมกับเทคนิคการค้นคืนสารสนเทศ (Song et al., 2005) และงานวิจัยการจัดกลุ่มเอกสารทางเว็บโดยใช้เซตรายการที่มากที่สุด (Zhuang and Dai, 2004) เป็นต้น นอกจากนั้นงานวิจัยของ Methanias และคณะ (Methanias et al., 2009) ที่ทำการเปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลบนข้อมูลซอฟต์แวร์อาร์ไคฟ์วของโครงการซอฟต์แวร์สิ่งแวดล้อมอุตสาหกรรม (Industrial Environment) ก็ใช้ค่าเอฟเมสเซอร์เป็นค่าที่ใช้การระบุระดับประสิทธิภาพเช่นกัน ค่าความถูกต้อง (Precision) ค่าเรียกคืน (Recall) และค่าเอฟเมสเซอร์ (F-measure) นั้นสามารถคำนวณได้ดังนี้

กำหนดให้ q คือ ข้อสอบถาม (Query) ที่ประกอบด้วยเซตเหตุการณ์ (Situation) Q และเซตผลลัพธ์ที่คาดไว้ (Expected Result) E และเขียนให้อยู่ในรูป q = (Q, E)

R คือ เซตของกฎความสัมพันธ์ที่อยู่ในรูปแบบ $Q \rightarrow \{x\}$ โดยที่ x คือรายการ

การเปลี่ยนแปลงแก้ไข และกำหนดให้ R_{10} คือเซตของกฎความสัมพันธ์ที่มีระดับความน่าสนใจสูงสุด 10 กฎแรกซึ่งเรียงลำดับด้วยค่าความเชื่อมั่น โดยที่ $R_{10} \subset R$

A_q คือ เซตของรายการการเปลี่ยนแปลงแก้ไข x ที่ได้จากกฎความสัมพันธ์ในเซต R_{10} ที่สอดคล้องกับเซตเหตุการณ์ Q ของข้อสอบถาม q ซึ่งสามารถเขียนในรูป $A_q = apply_{R_{10}}(Q)$ ดังนั้นขนาดของเซต A_q จะน้อยกว่าหรือเท่ากับ 10 เสมอ

2.9.1 **ค่าความถูกต้อง** (Precision)

ค่าความถูกต้อง (Precision) เป็นสัดส่วนรายการการเปลี่ยนแปลงแก้ไขที่ถูกดึงขึ้นมาแล้วตรงกับรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตผลลัพธ์ที่คาดไว้ เทียบกับรายการการเปลี่ยนแปลงแก้ไขทั้งหมดที่ถูกดึงขึ้น ดังสมการต่อไปนี้ (Zimmermann et al., 2005)

กำหนดให้ $|A_q \cap E|$ คือ จำนวนรายการการเปลี่ยนแปลงแก้ไขที่ถูกดึงขึ้นมาแล้วตรงกับรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตผลลัพธ์ที่คาดไว้

$|A_q|$ คือ จำนวนรายการการเปลี่ยนแปลงแก้ไขที่ถูกดึงขึ้นมา

$$precision = \frac{|A_q \cap E|}{|A_q|}$$

ในกรณีที่เซตรายการการเปลี่ยนแปลงแก้ไขที่ถูกดึงขึ้นมาเป็นเซตว่าง ($|A_q| = 0$) จะกำหนดให้ค่าความถูกต้องมีค่าเท่ากับ 1

2.9.2 **ค่าเรียกคืน** (Recall)

ค่าเรียกคืน (Recall) เป็นสัดส่วนรายการการเปลี่ยนแปลงแก้ไขที่ถูกดึงขึ้นมาแล้วตรงกับรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตผลลัพธ์ที่คาดไว้ เทียบกับรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตผลลัพธ์ที่คาดไว้ ดังสมการต่อไปนี้ (Zimmermann et al., 2005)

กำหนดให้ $|A_q \cap E|$ คือ จำนวนรายการการเปลี่ยนแปลงแก้ไขที่ถูกดึงขึ้นมาแล้วตรงกับรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตผลลัพธ์ที่คาดไว้

|E| คือ จำนวนรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตผลลัพธ์ที่คาดไว้

$$recall = \frac{|A_q \cap E|}{|E|}$$

ในกรณีที่เซตผลลัพธ์ที่คาดไว้เป็นเซตว่าง (|E| = 0) จะกำหนดให้ค่าความถูกต้องมีค่าเท่ากับ 1

2.9.3 ค่าเอฟเมสเซอร์ (F-measure)

ค่าเอฟเมสเซอร์ (F-measure) หรือค่าเอฟเมสเซอร์แบบถ่วงน้ำหนัก (Weighted Harmonic mean of Precision and Recall) ถูกเสนอขึ้นโดย Rijsbergen ในปี ค.ศ. 1979 (Rijsbergen, 1979) คือ ค่าที่ใช้ในการประเมินความถูกต้องแม่นยำ (accuracy) ของการทดสอบ ซึ่งพิจารณาจากทั้งค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการทดสอบโดยมีการถ่วงน้ำหนักให้กับค่าทั้งสองด้วย ค่าเอฟเมสเซอร์มีพิสัยอยู่ระหว่าง 0 กับ 1 ค่าเอฟเมสเซอร์ที่เท่ากับ 1 แสดงว่ามีความถูกต้องแม่นยำมากที่สุด ส่วนค่าเอฟเมสเซอร์ที่เท่ากับ 0 แสดงว่ามีความถูกต้องแม่นยำน้อยที่สุด สมการรูปทั่วไปของค่าเอฟเมสเซอร์แสดงได้ดังต่อไปนี้ (Tsunenori, 2003)

กำหนดให้ F คือ ค่าเอฟเมสเซอร์ (F-measure)

Precision คือ ค่าความถูกต้อง

Recall คือ ค่าเรียกคืน

β คือ ค่าอัตราส่วนน้ำหนักของค่าความถูกต้องต่อค่าเรียกคืน

$$F_\beta = \frac{(1+\beta^2) * precision * recall}{\beta^2 * precision + recall}$$

โดยที่ ค่า β หรือค่าอัตราส่วนน้ำหนักของค่าความถูกต้องต่อค่าเรียกคืนเป็นจำนวนจริงบวก ตัวอย่างของการกำหนดค่า เช่น ค่า β = 1 หมายความว่าให้น้ำหนักของค่าความถูกต้องกับค่าเรียกคืนมีน้ำหนักความสำคัญเท่ากัน ค่า β = 2 หมายถึงกำหนดให้ค่าความถูกต้องมีน้ำหนักความสำคัญเป็นสองเท่าเมื่อเทียบกับค่าเรียกคืน และค่า β = 0.5 หมายถึงกำหนดให้ค่าเรียกคืนมีน้ำหนักความสำคัญเป็นสองเท่าเมื่อเทียบกับค่าความถูกต้อง เป็นต้น สำหรับที่กำหนดให้ β = 1 นั้นค่าเอฟเมสเซอร์จะถูกเรียกว่า ค่าบาลานซ์เอฟเมสเซอร์ (balanced F-score) หรือค่าเอฟหนึ่ง

สกอร์ (F_1 score) หรือก็คือค่าเอฟเมสเซอร์ที่ถ่วงน้ำหนักอย่างสมดุลนั่นเอง สมการของค่าเอฟหนึ่งสกอร์แสดงได้ดังนี้ (Tsunenori, 2003)

กำหนดให้ F_1 คือ ค่าเอฟหนึ่งสกอร์ (F_1 score)
Precision คือ ค่าความถูกต้อง
Recall คือ ค่าเรียกคืน

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

ค่าเอฟเมสเซอร์เป็นค่าที่นิยมใช้ในการวัดประสิทธิภาพของการค้นคืนข้อมูล (Information Retrieval) อย่างแพร่หลาย (Tsunenori, 2003) งานวิจัยที่เกี่ยวข้องกับการค้นหากฎความสำคัญหลายงานวิจัย (Beil et al., 2002; Geyer-Schulz and Hahsler, 2002) แนะนำให้ใช้ค่าเอฟหนึ่งสกอร์หรือค่าเอฟเมสเซอร์ที่ให้น้ำหนักของค่าความถูกต้องและค่าเรียกคืนอย่างสมดุล

ในปี 2005 งานวิจัยของ Zimmermann และคณะ (Zimmermann et al., 2005)) ทำการเปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลบนข้อมูลซอฟต์แวร์อาร์ไคฟ์วกับโครงการพัฒนาระบบปฏิบัติการคอมพิวเตอร์ต่างๆ (Operating system) และกล่าวว่า จุดมุ่งหมายของการทดสอบระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์คือค่าความถูกต้องที่สูง (ค่าใกล้เคียง 1) และค่าเรียกคืนที่สูง (ค่าใกล้เคียง 1) นั่นคือต้องการให้ระบบสามารถแนะนำคำแนะนำทั้งหมด (ค่าเรียกคืนเท่ากับ 1) และแต่ละคำแนะนำนั้นถูกต้องหรือตรงกับเซตผลลัพธ์ที่คาดไว้ทั้งหมด (ค่าความถูกต้องเท่ากับ 1) ดังนั้นงานวิจัยของ Zimmermann และคณะจึงใช้ค่าเฉลี่ยฮาร์โมนิกของค่าความถูกต้องและค่าเรียกคืน (Harmonic mean of Precision and Recall) หรือก็คือค่าเอฟเมสเซอร์ที่ให้น้ำหนักของค่าความถูกต้องและค่าเรียกคืนอย่างสมดุล เป็นค่าประเมินประสิทธิภาพของการทำเหมืองข้อมูล (Zimmermann et al., 2005)

ต่อมาในปี 2009 งานวิจัยของ Methanias และคณะ (Methanias et al., 2009) ทำการเปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลบนข้อมูลซอฟต์แวร์อาร์ไคฟ์วของโครงการซอฟต์แวร์สิ่งแวดล้อมอุตสาหกรรม (Industrial Environment) ใน 3 สถานการณ์เช่นเดียวกัน และแนะนำให้ใช้ค่าเอฟเมสเซอร์ที่ให้น้ำหนักของค่าความถูกต้องและค่าเรียกคืนอย่างสมดุลสำหรับการประเมินประสิทธิภาพในสถานการณ์การนำทางและสถานการณ์การป้องกันข้อผิดพลาด (Methanias et al., 2009)

2.9.4 **ค่าผลสะท้อนกลับ** (Feedback)

ค่าผลสะท้อนกลับ (Feedback) คือ ค่าร้อยละของเซตรายการการเปลี่ยนแปลงแก้ไขที่ถูกดึงขึ้นมาที่ไม่ใช่เซตว่างกับเซตรายการการเปลี่ยนแปลงแก้ไขที่ถูกดึงขึ้นมาทั้งหมด ค่าผลสะท้อนกลับแสดงได้ดังสมการต่อไปนี้ (Zimmermann et al., 2005)

กำหนดให้ $feedback$ คือ ค่าผลสะท้อนกลับ

$|Z^*|$ คือ จำนวนข้อสอบถามที่อยู่ในเซตของข้อสอบถามที่มีเซตของคำแนะนำที่ไม่เป็นเซตว่าง โดยที่ $Z^* = \{q \mid q = (Q, E) \in Z, apply_{R_{10}}(Q) \neq \emptyset\}$

$|Z|$ คือ จำนวนข้อสอบถามทั้งหมด โดยที่ $Z = \{q \mid q = (Q, E)\}$

$$feedback = \frac{|Z^*|}{|Z|}$$

เมื่อนำค่าผลสะท้อนกลับมาใช้ประเมินในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วจะสามารถแสดงให้เห็นถึงร้อยละของการเกิดการแจ้งเตือนที่ผิด (False Alarm) หรือการให้คำแนะนำที่เป็นผลบวกลวงนั่นเอง เนื่องจากค่าผลสะท้อนกลับมีพิสัยอยู่ระหว่าง 0 กับ 1 ค่าผลสะท้อนกลับที่มีค่าเท่ากับ 0 หมายถึงไม่มีข้อสอบถามใดเลยที่ให้คำแนะนำที่เป็นผลบวกลวงออกมาในสถานการณ์นี้นั่นคือมีประสิทธิภาพดีที่สุด และค่าผลสะท้อนกลับที่มีค่าเท่ากับ 1 หมายถึงข้อสอบถามทั้งหมดให้คำแนะนำที่เป็นผลบวกลวงออกมานั่นคือมีประสิทธิภาพไม่ดีที่สุด ดังนั้นการวัดประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์วในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วนั้นจะต้องเปรียบเทียบค่าผลสะท้อนกลับและค่าผลสะท้อนกลับที่น้อยกว่าจะมีความหมายว่ามีประสิทธิภาพดีกว่า (Zimmermann et al., 2005)

บทที่ 3
ระเบียบวิธีวิจัย

3.1 บทนำ

ในบทนี้จะกล่าวถึงระเบียบวิธีวิจัยซึ่งประกอบด้วยการอธิบายแผนแบบการทดลอง (Experimental Design) การทดสอบสมมติฐาน การทำงานของเครื่องมือทดสอบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบต่างๆ ที่งานวิจัยกำหนด รวมทั้งวิธีการพัฒนาเครื่องมือทดสอบประสิทธิภาพของ การทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ ประเด็นความน่าเชื่อถือได้ (Reliability) ความถูกต้อง (Validity) และกรอบการวิเคราะห์ข้อมูล (Data Analysis Framework) ดังรายละเอียดต่อไปนี้

3.2 แผนแบบการทดลอง

งานวิจัยนี้มีวัตถุประสงค์ในการทดลองเพื่อศึกษาเปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น (Support-Confidence Model) ดั้งเดิมกับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008) ในสถานการณ์ของการให้คำแนะนำนักพัฒนาต่างๆกัน 3 สถานการณ์ได้แก่ 1) สถานการณ์การนำทาง (Navigation) คือ สถานการณ์ที่นักพัฒนามีการเปลี่ยนแปลงแก้ไขที่เอนทิตี้หนึ่งแล้ว ระบบจะให้คำแนะนำกับนักพัฒนาให้แก้ไขเอนทิตี้ใดต่อไปได้ถูกต้องหรือไม่ 2) สถานการณ์การป้องกันการเกิดข้อผิดพลาด (Error Prevention) คือ สถานการณ์ที่นักพัฒนามีการเปลี่ยนแปลงแก้ไขที่เอนทิตี้หลายๆเอนทิตี้ต่อเนื่องกันแต่ยังขาดการเปลี่ยนแปลงแก้ไขเอนทิตี้อีกหนึ่งเอนทิตี้จึงจะสมบูรณ์ ระบบจะให้คำแนะนำกับนักพัฒนาให้แก้ไขเอนทิตี้ที่เหลือนั้นได้ถูกต้องหรือไม่ 3) สถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (Closure) คือ สถานการณ์ที่นักพัฒนามีการเปลี่ยนแปลงแก้ไขที่เอนทิตี้หลายๆเอนทิตี้ต่อเนื่องกันจนสมบูรณ์แล้ว ระบบจะให้คำแนะนำที่เป็นผลบวกลวง (False Positive) ออกมาแก่นักพัฒนาหรือไม่

จากวัตถุประสงค์งานวิจัยที่กล่าวข้างต้น ผู้วิจัยจึงเลือกใช้แผนแบบการทดลอง แบบการเปรียบเทียบกลุ่มสถิต (Static Group Comparison) ซึ่งเป็นแผนแบบการทดลองที่เหมาะสมกับการทดลองที่ต้องการวัดค่าตัวแปรตามของกลุ่มควบคุมและกลุ่มทดลองหลังจากทำการทดสอบมีค่าแตกต่างกันอย่างไร นั่นคือ การวัดค่าประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นดั้งเดิม (กลุ่มควบคุม) กับค่าประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008) (กลุ่มทดสอบ) หลังจากการทดสอบว่ามีค่าแตกต่างกันอย่างไร โดยกำหนดให้มีตัวแปรในการทดลองเปรียบเทียบการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบที่ต้องการทดสอบ ดังต่อไปนี้

3.2.1 ตัวแปรต้น

งานวิจัยนี้สนใจว่าการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008) สามารถเพิ่มประสิทธิภาพของระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ของไอดีอี (IDE: Integrated Development Environment) ได้หรือไม่ ดังนั้นตัวแปรต้นของการศึกษาครั้งนี้คือตัวแบบของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ทั้งหมด 2 ตัวแบบ ดังนี้

1) การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น (Support-Confidence Model)
2) การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Support-New Confidence Model) (Liu et al., 2008)

จากการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ทั้ง 2 ตัวแบบข้างต้น ผู้วิจัยจะเรียกการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น ด้วยคำว่า “การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1” ส่วนการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎ

ความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008) จะเรียกว่า “การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2”

3.2.2 ตัวแปรตาม

เนื่องจากงานวิจัยนี้สนใจเปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ 2 ตัวแบบดังกล่าวข้างต้น ดังนั้นการเปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์จะพิจารณาจากความถูกต้องแม่นยำในการทำนายและให้คำแนะนำกับนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์มากน้อยเพียงใด โดยการวัดประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ในแต่ละสถานการณ์มีวิธีการในการประเมินที่แตกต่างกันออกไปดังนี้

- ในสถานการณ์ *การนำทาง (Navigation)* สามารถประเมินประสิทธิภาพจากค่าเอฟเมสเซอร์ (F-measure) ที่คำนวณมาจากค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ซึ่งรายละเอียดและวิธีการคำนวณค่าเอฟเมสเซอร์ ค่าความถูกต้องและค่าเรียกคืนสำหรับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์นั้นได้กล่าวเอาไว้ในบทที่ 2
- ในสถานการณ์ *การป้องกันการเกิดข้อผิดพลาด (Error Prevention)* สามารถประเมินประสิทธิภาพจากค่าเอฟเมสเซอร์ (F-measure) ที่คำนวณมาจากค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ซึ่งรายละเอียดและวิธีการคำนวณค่าเอฟเมสเซอร์ ค่าความถูกต้องและค่าเรียกคืนสำหรับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์นั้นได้กล่าวเอาไว้ในบทที่ 2
- ในสถานการณ์ *การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (Closure)* สามารถประเมินประสิทธิภาพจากค่าผลสะท้อนกลับ (Feedback) ซึ่งรายละเอียดและวิธีการคำนวณค่าผลสะท้อนกลับสำหรับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์นั้นได้กล่าวเอาไว้ในบทที่ 2

3.3 สมมติฐานงานวิจัย

จากวัตถุประสงค์ของงานวิจัย ผู้วิจัยต้องการทดสอบว่าการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008) สามารถเพิ่มประสิทธิภาพของระบบให้คำแนะนำนักพัฒนาในสถานการณ์ต่างๆกัน 3 สถานการณ์คือ สถานการณ์การนำทาง (Navigation) สถานการณ์การป้องกันการเกิดข้อผิดพลาด (Error Prevention) และสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (Closure) ได้หรือไม่ ดังนั้นงานวิจัยนี้จึงต้องการศึกษาประสิทธิภาพของการให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ที่ได้มาจากการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ทั้ง 2 ตัวแบบตามที่กำหนดไว้ในหัวข้อตัวแปรต้น โดยผู้วิจัยจะตั้งสมมติฐานในแต่ละสถานการณ์จะที่ทำการทดสอบไว้ดังนี้

1) วิเคราะห์เปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ทั้ง 2 ตัวแบบในสถานการณ์การนำทาง ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ μ_1คือ ค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การนำทาง

μ_2คือ ค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การนำทาง

ผู้วิจัยต้องการทราบว่าค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบใดที่มีค่ามากกว่ากัน ดังนั้นผู้วิจัยจึงเปรียบเทียบค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 กับค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ผู้วิจัยเห็นว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 นั้นสามารถให้เกิดผลลัพธ์ของการทำเหมืองข้อมูลที่เป็นผลบวกลวง (False Positive) เป็นจำนวนมาก โดยเฉพาะอย่างยิ่งกับข้อมูลซอฟต์แวร์อาร์ไคฟ์ (Li et al., 2005) ซึ่งน่าจะเป็นผลมาจากการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 นั้นจะให้ผลลัพธ์ที่ไม่สอดคล้องกับสหสัมพันธ์ (Correlation) แต่การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 นั้นให้ผลลัพธ์ของการทำเหมืองข้อมูลที่สอดคล้องกับสหสัมพันธ์ด้วย (Liu et al., 2008) ดังนั้นผู้วิจัยจึงคาดว่าการ

ค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 นั้นมีประสิทธิภาพที่ดีกว่าการการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การนำทาง จึงตั้งสมมติฐานไว้ ดังนี้

$H_0 : \mu_2 \leq \mu_1$

$H_1 : \mu_2 > \mu_1$

การยอมรับ H_0 หมายถึง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 นั้นมีประสิทธิภาพที่ดีกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การนำทาง

การปฏิเสธ H_0 หมายถึง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 นั้นมีประสิทธิภาพที่ดีกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การนำทาง

2) วิเคราะห์เปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ μ_1 คือ ค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

μ_2 คือ ค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

ผู้วิจัยต้องการทราบว่าค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบใดที่มีค่ามากกว่ากัน ดังนั้นผู้วิจัยจึงเปรียบเทียบค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 กับค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ผู้วิจัยเห็นว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 นั้นสามารถให้เกิดผลลัพธ์ของการทำเหมืองข้อมูลที่เป็นผลบวกลวง (False Positive) เป็นจำนวนมาก โดยเฉพาะอย่างยิ่งกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Li et al., 2005) ซึ่งน่าจะเป็นผลมาจากการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 นั้นจะให้ผลลัพธ์ที่ไม่สอดคล้องกับสหสัมพันธ์ (Correlation) แต่การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 นั้นให้ผลลัพธ์ของการทำ

เหมืองข้อมูลที่สอดคล้องกับสหสัมพันธ์ด้วย (Liu et al., 2008) ดังนั้นผู้วิจัยจึงคาดว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 นั้นมีประสิทธิภาพที่ดีกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด จึงตั้งสมมติฐานไว้ ดังนี้

$H_0 : \mu_2 \leq \mu_1$

$H_1 : \mu_2 > \mu_1$

การยอมรับ H_0 หมายถึง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 นั้นมีประสิทธิภาพที่ดีกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

การปฏิเสธ H_0 หมายถึง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 นั้นมีประสิทธิภาพที่ดีกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

3) วิเคราะห์เปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ μ_1 คือ ค่าผลสะท้อนกลับของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

μ_2 คือ ค่าผลสะท้อนกลับของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

ผู้วิจัยต้องการทราบว่าค่าผลสะท้อนกลับของการค้นหากฎความสัมพันธ์ด้วยตัวแบบใดที่มีค่ามากกว่ากัน ดังนั้นผู้วิจัยจึงเปรียบเทียบค่าผลสะท้อนกลับของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 กับค่าผลสะท้อนกลับของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ผู้วิจัยเห็นว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 นั้นสามารถให้เกิดผลลัพธ์ของการทำเหมืองข้อมูลที่เป็นผลบวกลวง (False Positive) เป็นจำนวนมาก โดยเฉพาะอย่างยิ่งกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Li et al., 2005) ซึ่งน่าจะเป็นผลมาจาก

การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 นั้นจะให้ผลลัพธ์ที่ไม่สอดคล้องกับสหสัมพันธ์ (Correlation) แต่การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 นั้นให้ผลลัพธ์ของการทำเหมืองข้อมูลที่สอดคล้องกับสหสัมพันธ์ด้วย (Liu et al., 2008) ดังนั้นผู้วิจัยจึงคาดว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 นั้นจะให้ค่าผลสะท้อนกลับที่น้อยกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว จึงตั้งสมมติฐานไว้ ดังนี้

$H_0 : \mu_2 \geq \mu_1$

$H_1 : \mu_2 < \mu_1$

การยอมรับ H_0 หมายถึง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 นั้นมีประสิทธิภาพที่ดีกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

การปฏิเสธ H_0 หมายถึง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 นั้นมีประสิทธิภาพที่ดีกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

3.4 ประชากรและหน่วยตัวอย่าง

สำหรับงานวิจัยนี้ข้อมูลซอฟต์แวร์อาร์ไคฟ์ที่ได้จากระบบคอนเคอเรนท์เวอร์ชั่นนั้นเป็นประชากรของการทดลองในงานวิจัยนี้ ซึ่งผู้วิจัยหวังว่าจะทดสอบระบบกับข้อมูลซอฟต์แวร์อาร์ไคฟ์ของโครงการพัฒนาซอฟต์แวร์ทั้งหมด แต่ในทางปฏิบัตินั้นผู้วิจัยไม่สามารถนำข้อมูลซอฟต์แวร์อาร์ไคฟ์ของโครงการพัฒนาซอฟต์แวร์ที่พัฒนาขึ้นในเชิงพาณิชย์ได้ ดังนั้นผู้วิจัยจึงเลือกใช้ข้อมูลซอฟต์แวร์อาร์ไคฟ์จากระบบคอนเคอเรนท์เวอร์ชั่นในโครงการพัฒนาซอฟต์แวร์ที่เป็นโครงการโอเพนซอร์ส (Open Source Project) จำนวน 1 โครงการคือโครงการพัฒนาซอฟต์แวร์ทางการบัญชีชื่อเคมายมันนี่ (KMyMoney) มาเพื่อเป็นหน่วยตัวอย่างของงานวิจัยนี้ รายละเอียดของซอฟต์แวร์นี้แสดงดังตารางต่อไปนี้

ตารางที่ 3-1 แสดงรายละเอียดของซอฟต์แวร์ทางการบัญชีชื่อเคมายมันนี่ (KMyMoney) เก็บข้อมูลในวันที่ 10 มกราคม พ.ศ. 2553

ข้อมูลทั่วไป	
ประเภทซอฟต์แวร์	การบัญชี การจัดทำและการพยากรณ์งบประมาณ
ผู้บำรุงรักษา	โทมัส โบมการ์ท และ มิเชล เอ็ดเวิร์ด
ประเภทลิขสิทธิ์	จีพีแอล (GPL)
ผู้ใช้เป้าหมาย	ผู้ใช้ทั่วไป องค์กรไม่แสวงหาผลประโยชน์
ขนาดซอฟต์แวร์	14.2 เมกะไบต์
จำนวนการถูกบรรจุลง (ครั้ง)	223,889
ข้อมูลด้านเทคนิค	
ภาษาที่ใช้ในการพัฒนา	ซีพลัสพลัส (C++)
แพลตฟอร์มที่รองรับ	ลินุกซ์ (Linux) ยูนิกซ์ (Unix) และฟรีบีเอสดี (FreeBSD)
ส่วนติดต่อผู้ใช้	เคดีอี (KDE), คิวท์ (Qt)
ฐานข้อมูลที่สนับสนุน	มายเอสคิวแอล (MySQL) โพสต์เกรสเอสคิวแอล (PostgreSQL) และ เอสคิวแอลไลท์ (SQLite)
ข้อมูลจากระบบคอนเคอเรนท์เวอร์ชั่น (CVS)	
วันที่เริ่มต้นโครงการ	16/04/2000
วันที่ปรับปรุงล่าสุด	28/12/2009
ระยะเวลา (ปี)	9.7
จำนวนนักพัฒนา (คน)	10
จำนวนทรานแซคชั่น (ทรานแซคชั่น)	28261
จำนวนแท็ก (แท็ก)	30777
จำนวนแฟ้มข้อมูล (แฟ้ม)	2583
จำนวนไดเรกทอรี่	118
จำนวนเอนทิตี้ (เอนทิตี้)	21660
จำนวนเอนทิตี้ต่อแฟ้มข้อมูลโดยเฉลี่ย (เอนทิตี้/แฟ้มข้อมูล)	8.4

จำนวนกิ่งก้าน (กิ่ง)	636
ข้อมูลโครงสร้างของซอฟต์แวร์	
จำนวนเนมสเปส (เนมสเปส)	7
จำนวนคลาส อินเตอร์เฟส สตรัคท์ และยูเนียน (คอมโพเนนท์)	543
จำนวนฟังก์ชั่น (ฟังก์ชั่น)	4472
จำนวนตัวแปร (ตัวแปร)	2896
จำนวนไทพ์ดีฟ (ไทพ์ดีฟ)	27
จำนวนอีนัมมูเรชั่น (อีนัมมูเรชั่น)	112
จำนวนอีนัมมูเรเตอร์ (อีนัมมูเรเตอร์)	962
จำนวนพรอพเพอร์ตี้ (พรอพเพอร์ตี้)	10

ซอฟต์แวร์เคมายมันนี่ (KMyMoney) คือ ซอฟต์แวร์จัดการการเงิน (Finance Management Software) ที่สามารถรองรับการใช้งานในระดับบุคคลและระดับองค์กรขนาดเล็กได้ ถูกพัฒนาขึ้นมาด้วยภาษาซีพลัสพลัส (C++) บนเครื่องมือที่ชื่อว่าคิวท์ (Qt) ทำให้เป็นโปรแกรมประยุกต์ที่พัฒนาขึ้นครั้งเดียวแต่ทำงานได้บนระบบปฏิบัติการตระกูลยูนิกซ์เกือบทุกรุ่น ซอฟต์แวร์เคมายมันนี่แบ่งกลุ่มของการจัดการข้อมูลออกเป็น 10 ส่วนดังนี้

1. สถาบันทางการเงิน (Institutions) คือ ส่วนที่ผู้ใช้สามารถบริหารจัดการบัญชีของผู้ใช้โดยแบ่งแยกตามสถาบันทางการเงินหรือธนาคารของแต่และบัญชีที่มีอยู่ ในส่วนนี้ผู้ใช้สามารถสร้าง แก้ไข ลบ หรือเปิดดูข้อมูลการเงินของแต่ละบัญชีในสถาบันทางการเงินหรือธนาคารได้ตามต้องการ
2. บัญชี (Accounts) คือ ส่วนที่ผู้ใช้สามารถบริหารจัดการบัญชีของผู้ใช้โดยแบ่งแยกตามประเภทของแต่และบัญชีที่มีอยู่ เช่น บัญชีทรัพย์สิน บัญชีหนี้สิน เป็นต้น ในส่วนนี้ผู้ใช้สามารถสร้าง แก้ไข ลบ หรือเปิดดูข้อมูลการเงินของแต่ละบัญชีได้ตามต้องการ
3. ตารางเวลา (Schedules) คือ ส่วนที่ผู้ใช้สามารถสร้างและจัดการตารางเวลาการทำธุรกรรมทางการเงินได้ ผู้ใช้สามารถสร้างตารางเวลาล่วงหน้า ตั้งตารางเวลา

แบบซ้ำเป็นรอบวัน รอบสัปดาห์ หรือรอบเดือน ช่วยให้ผู้ใช้ไม่ลืมและสามารถทำธุรกรรมได้ตรงตามเวลา นอกจากนั้นผู้ใช้สามารถกำหนดให้ธุรกรรมที่อยู่บนตารางเวลานั้นไปปรากฏบนบัญชีแยกประเภท (Ledgers) ได้ด้วย

4. หมวดการทำธุรกรรม (Categories) คือ ส่วนที่ผู้ใช้สามารถ แก้ไข หรือลบหมวดของการทำธุรกรรม เช่น หมวดธุรกรรมด้านศึกษา หมวดธุรกรรมทั่วไป เป็นต้น ผู้ใช้สามารถดึงข้อมูลธุรกรรมที่จัดทำไว้ในส่วนอื่นๆมาจัดหมวดหมู่ในส่วนนี้
5. ผู้ร่วมทำธุรกรรม (Payees) คือ ส่วนที่ผู้ใช้สามารถจัดการข้อมูลของบุคคล กลุ่มบุคคล หรือองค์กร ที่ผู้ใช้ไปทำธุรกรรมด้วย ในส่วนนี้ผู้ใช้สามารถเรียกดูรายการทำธุรกรรมทั้งหมดโดยแบ่งแยกตามบุคคล กลุ่มบุคคล หรือองค์กร ที่ผู้ใช้ไปทำธุรกรรมด้วยได้
6. บัญชีแยกประเภท (Ledgers) คือ ส่วนที่ผู้ใช้สามารถจัดการรายการธุรกรรมของผู้ใช้ ลักษณะเดียวกับโปรแกรมไมโครซอฟต์มันนี่ (Microsoft Money) นอกจากนั้นยังมีความสามารถพิเศษที่ชื่อว่าเลคเจอะเลนส์ (ledger lens) ใช้สำหรับเลือกรายการธุรกรรมตั้งแต่หนึ่งถึงสามรายการเพื่อขยายดูรายละเอียดภายในของรายการธุรกรรมนั้นๆ ข้อมูลรายการธุรกรรมทั้งหมดสามารถเลือกให้เรียงลำดับรายการตามคอลัมภ์ที่ต้องการได้
7. การลงทุน (Investments) คือ ส่วนที่ผู้ใช้สามารถติดตามการลงทุนพื้นฐานต่างๆ ได้ เช่นการติดตามราคาหุ้น ราคาทองคำ อัตราแลกเปลี่ยนเงินตราต่างประเทศ และกองทุนต่างๆ ผู้ใช้สามารถเลือกเพิ่ม แก้ไขหรือลบรายการการลงทุนที่ต้องการติดตามได้
8. รายงานทางการเงิน (Reports) คือ ส่วนที่ผู้ใช้สามารถจัดทำเอกสารรายงานทางการเงินต่างๆได้ โดยที่ผู้ใช้สามารถเลือกกำหนดค่าองค์ประกอบต่างๆได้อย่างอิสระ ในเวอร์ชั่นปัจจุบันผู้ใช้สามารถสร้างกราฟ แผนภาพแบบต่างๆในเอกสารรายงานได้ด้วย
9. งบประมาณ (Budgets) คือ ส่วนที่ผู้ใช้สามารถจัดการหมวดหมู่ของรายการรายได้และค่าใช้จ่ายที่คาดหวังไว้ในภายในช่วงของเวลาที่กำหนด ผู้ใช้สามารถกำหนดช่วงของเวลาได้ 3 แบบคือ รายปี รายเดือน หรือระยะเวลาเฉพาะที่กำหนดเอง นอกจากนั้นระบบยังสามารถสร้างรายงานแสดงการเปรียบเทียบรายได้และรายจ่ายจริงกับและรายได้และรายจ่ายที่คาดหวังไว้ได้

10. การพยากรณ์ทางการเงิน (Forecast) คือ ส่วนที่ผู้ใช้สามารถเรียกดูการพยากรณ์ทางการเงินของแต่ละบัญชีที่ผู้ใช้มีอยู่ได้ การพยากรณ์ทางการเงินในที่นี้หมายถึงการคาดการณ์ยอดเงินคงเหลือของแต่ละบัญชีในจุดเวลาอนาคต การพยากรณ์ของระบบเกิดขึ้นมาจากข้อมูลที่บันทึกในส่วนตารางเวลาและส่วนบัญชีแยกประเภทปัจจุบัน รวมถึงการรวบรวมธุรกรรมที่เคยเกิดขึ้นในอดีตมาพิจารณาด้วย

จากตารางแสดงรายละเอียดด้านต่างๆและหน้าที่การทำงานที่อธิบายไปในข้างต้นของซอฟต์แวร์ชื่อเคมายมันนี่ ผู้วิจัยเห็นว่าซอฟต์แวร์ชื่อเคมายมันนี่มีขนาดใหญ่ มีความซับซ้อนมากในระดับหนึ่ง มีนักพัฒนาผู้ร่วมโครงการเป็นจำนวนมาก และมีโอกาสที่จะมีนักพัฒนาใหม่เข้าร่วมโครงการได้เสมอ นอกจากนั้นซอฟต์แวร์ชื่อเคมายมันนี่นั้นพัฒนาขึ้นมาด้วยภาษาซี พลัสพลัส (C++) ซึ่งเป็นภาษาเชิงวัตถุ (Object Oriented Programming Language) ผู้วิจัยจึงหวังว่าซอร์สโค้ดของโครงการพัฒนาซอฟต์แวร์ชื่อเคมายมันนี่ (KMyMoney) จะสามารถเป็นตัวแทนที่ดีของซอฟต์แวร์ขนาดใหญ่ที่ถูกพัฒนาด้วยภาษาเชิงวัตถุ และมีนักพัฒนาผู้ร่วมโครงการเป็นจำนวนมากได้ ดังนั้นผู้วิจัยจึงเลือกข้อมูลซอฟต์แวร์อาร์ไคฟ์ของโครงการพัฒนาซอฟต์แวร์ชื่อเคมายมันนี่นี้เป็นข้อมูลที่ใช้ในการทดสอบของงานวิจัยนี้

สาเหตุที่งานวิจัยนี้เลือกตัวอย่างข้อมูลซอฟต์แวร์อาร์ไคฟ์เพียงตัวอย่างเดียว เนื่องจากงานวิจัยนี้ต้องการวิจัยเพื่อหาข้อมูลเบื้องต้น (Exploratory Research) เพื่อประโยชน์ในการวิจัยในอนาคตเท่านั้น ไม่ได้ต้องการสรุปผลให้ครอบคลุมการนำไปประยุกต์กับกรณีทั่วไป

3.5 แนวทางการทำวิจัย

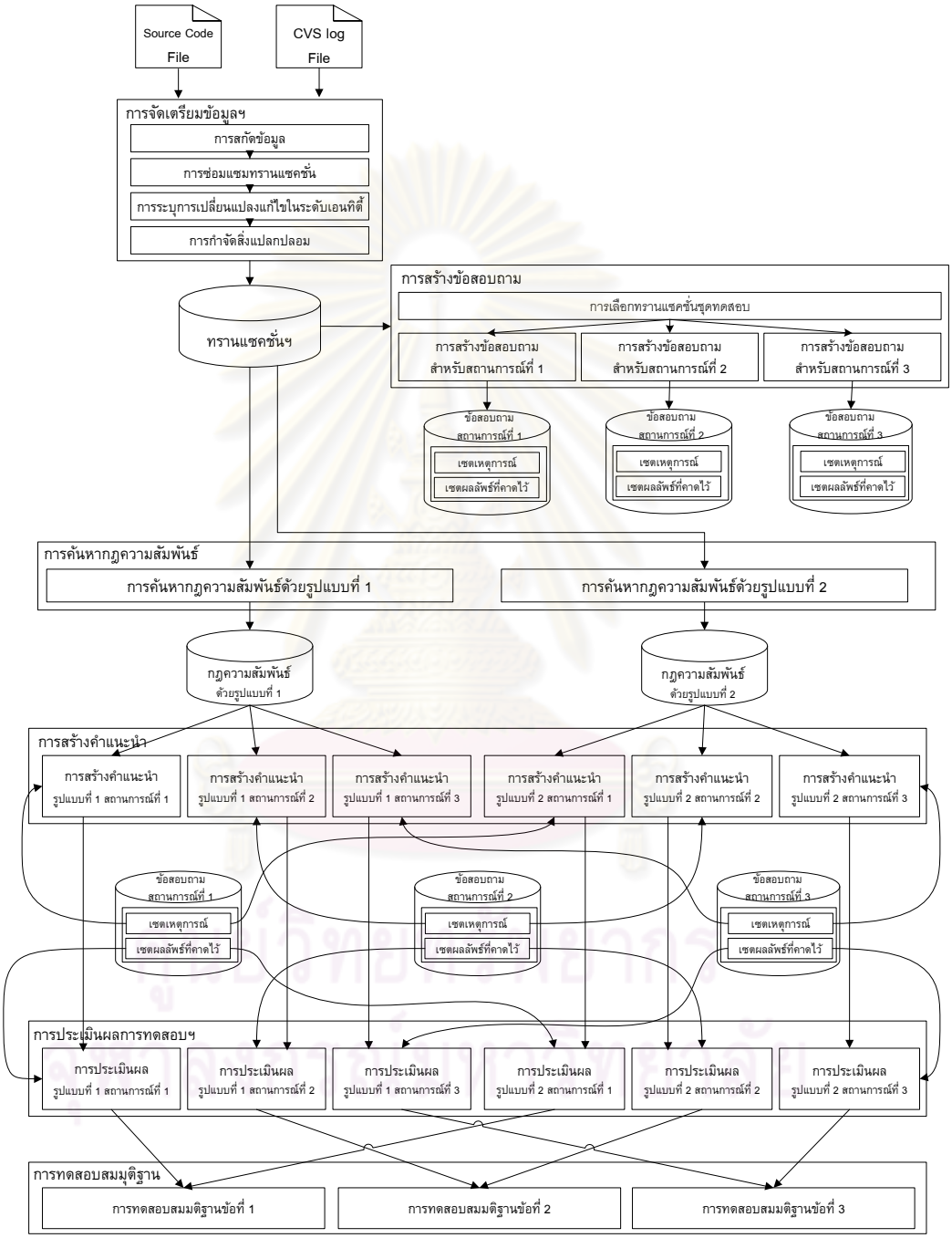
งานวิจัยนี้เป็นงานวิจัยเชิงทดลอง (Experimental Research) เนื่องจากเป็นการทดลองเพื่อเปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบ 2 ตัวแบบใน 3 สถานการณ์ สำหรับในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้นตัวแบบที่มีประสิทธิภาพดีกว่าคือตัวแบบที่ให้ค่าเอฟเมสเซอร์ที่สูงกว่า และสำหรับในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วนั้นตัวแบบที่มีประสิทธิภาพดีกว่าคือตัวแบบที่ให้ค่าผลสะท้อนกลับที่น้อยกว่า โดยในงานวิจัยนี้สนใจว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 หรือการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่

ของ Liu และคณะ (Liu et al., 2008) ในการทำเหมืองข้อมูลนั้นจะสามารถช่วยเพิ่มประสิทธิภาพให้กับระบบให้คำแนะนำนักพัฒนาระหว่างการพัฒนาซอฟต์แวร์ที่ได้มาจากผลของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ได้ ในงานวิจัยเชิงทดลองนี้จะควบคุมตัวแปรอื่นๆ ให้เหมือนกันหมดนั่นคือ ข้อมูลซอฟต์แวร์อาร์ไคฟว์ ข้อสอบถามความถูกต้องระหว่างข้อมูลซอฟต์แวร์อาร์ไคฟว์และข้อสอบถาม และเครื่องมือที่ใช้ในการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูล (Data preprocessing tool) และเครื่องมือในการทำเหมืองข้อมูล (Data Mining tool) เพื่อให้ตัวแปรควบคุมที่กำหนดนั้นมีผลกระทบกับตัวแปรตามน้อยที่สุดและผลของงานวิจัยจะได้เป็นผลที่เกิดจากการเปลี่ยนแปลงตัวแปรต้นอย่างแท้จริง นั่นคืองานวิจัยจะทดลองว่าผลลัพธ์ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์จะมีประสิทธิภาพเปลี่ยนแปลงไปอย่างไรเมื่อใช้ตัวแบบในการทำเหมืองข้อมูลแตกต่างกันสำหรับสถานการณ์ต่างๆ โดยได้สร้างเครื่องมือเพื่อทดสอบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ ดังนี้

1) การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น
2) การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008)

งานวิจัยนี้ได้พัฒนาระบบการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ออกเป็น 2 ตัวแบบดังกล่าว เนื่องจากผู้วิจัยสนใจว่าการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะนั้นสามารถเพิ่มประสิทธิภาพของระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ของไอดีอีได้หรือไม่ ดังนั้นในการทดลองจึงต้องพัฒนาระบบการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นดั้งเดิมไว้เป็นกลุ่มควบคุม เพื่อเป็นกลุ่มเปรียบเทียบกับระบบการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะซึ่งเป็นกลุ่มทดสอบ สำหรับสถานการณ์ทั้ง 3 สถานการณ์

3.6 ขั้นตอนทดสอบการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์



รูปที่ 3-1 แสดงขั้นตอนการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์

การออกแบบการทดสอบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ของงานวิจัยนี้มีขั้นตอนการทำงานทั้งหมดแสดงดังรูปที่ 3-1 ซึ่งการทดสอบในครั้งนี้จะแบ่งขั้นตอนในการทดสอบออกเป็น 6 ขั้นตอน คือ

1) การจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์
2) การสร้างข้อสอบถามสำหรับการทดสอบ 3 สถานการณ์
3) การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบสำหรับ 3 สถานการณ์
4) การสร้างเซตของคำแนะนำสำหรับเหตุการณ์
5) การประเมินผลการทดสอบ
6) การทดสอบสมมติฐาน

รายละเอียดข้อมูลเข้า กระบวนการทำงานและข้อมูลออกของแต่ละขั้นตอนการทดสอบอธิบายดังต่อไปนี้

3.6.1 การจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์

การจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ คือ ส่วนที่ทำหน้าที่ในการดึงแฟ้มข้อมูลซอร์สโค้ดและแฟ้มข้อมูลบันทึกจากรีพอสิโทริของระบบคอนเคอเรนท์เวอร์ชั่นแล้วนำแฟ้มข้อมูลเหล่านั้นมาผ่านกระบวนการ 4 กระบวนการเพื่อให้ได้ข้อมูลทรานแซคชั่นของการเปลี่ยนแปลงแก้ไข ดังนี้

1) การสกัดข้อมูล (Data Extraction)

2) การซ่อมแซมทรานแซคชั่น (Restoring Transactions)

3) การระบุการเปลี่ยนแปลงแก้ไขในระดับเอนทิตี้ (Mapping Changes to Entities)

4) การกำจัดสิ่งแปลกปลอม (Data Cleaning)

ขั้นตอนการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ 4 กระบวนการข้างต้นถูกเสนอขึ้นมาโดย Zimmermann และคณะ ในปี 2004 (Zimmermann et al., 2004) และเป็นขั้นตอนวิธีที่มีงานวิจัยที่เกี่ยวข้องกับการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์

ไคฟว์หลายงายวิจัย (Zimmermann et al., 2005; Livshits et al., 2005; Williams et al., 2005; Breu et al., 2006; Weißgerber et al, 2006) นำไปใช้ งานวิจัยนี้นำขั้นตอนการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 4 กระบวนการมาใช้เช่นกัน และกำหนดรายละเอียดต่างๆของแต่ละกระบวนการเหมือนกับที่ Zimmermann และคณะได้เสนอไว้ (Zimmermann et al., 2004) ผู้วิจัยจะอธิบายกระบวนการทำงานพร้อมกับยกตัวอย่างบางส่วนประกอบการอธิบาย เพื่อเพิ่มความเข้าใจในการทำงานของแต่ละกระบวนการมากขึ้นด้วย

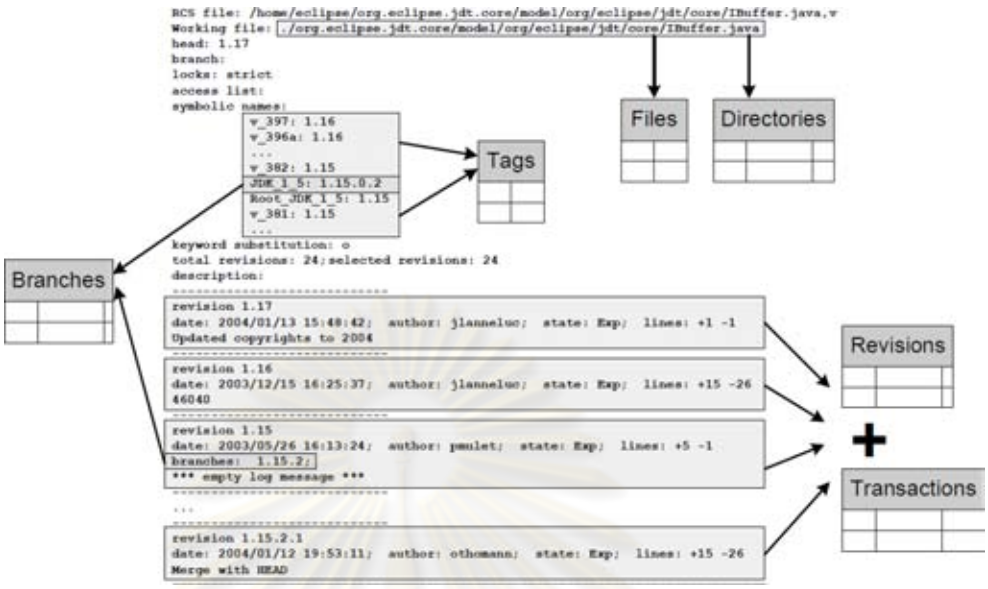
- **การสกัดข้อมูล** (Data Extraction)

การสกัดข้อมูลจากแฟ้มข้อมูลบันทึกของระบบคอนเคอเรนท์เวอร์ชั่นเริ่มต้นจากการเรียกใช้คำสั่ง CVS log จากไดเรคทอรี่ราก (root directory) ของโครงการพัฒนาซอฟต์แวร์ที่ต้องการ ผลลัพธ์ที่ส่งกลับคืนมาก็คือข้อมูลแฟ้มข้อมูลซอร์สโค้ดและแฟ้มข้อมูลบันทึก (Log File) ที่บันทึกอยู่บนรีพอสิโทริของระบบคอนเคอเรนท์เวอร์ชั่นนั้น แฟ้มข้อมูลบันทึกที่ได้มาจะถูกนำมาวิเคราะห์รูปแบบไวยากรณ์ (Parse) และถูกนำไปบันทึกลงฐานข้อมูล (Zimmermann et al., 2004)

ข้อมูลเข้า คือ แฟ้มข้อมูลบันทึกของระบบคอนเคอเรนท์เวอร์ชั่น (CVS Log File)

ข้อมูลออก คือ ทรานแซคชั่นของการเปลี่ยนแปลงแก้ไข (ฐานข้อมูลที่ประกอบด้วยตาราง 6 ตาราง คือตารางชื่อ Files ตารางชื่อ Directories ตารางชื่อ Tags ตารางชื่อ Branches ตารางชื่อ Revisions และตารางชื่อ Transactions)

กระบวนการทำงาน สามารถอธิบายได้โดยแผนภาพต่อไปนี้ (Zimmermann et al., 2004)



รูปที่ 3-2 แสดงตัวอย่างแฟ้มข้อมูลบันทึกของระบบคอนเคอเรนท์เวอร์ชั่น

รูปแบบไวยากรณ์ แอททริบิวต์ และความหมายของแต่ละแอททริบิวต์ อธิบายไว้ในบทที่ 2 รายละเอียดของข้อมูลทั้งหมดที่จะถูกบันทึกเอาไว้จากขั้นตอนการสกัดข้อมูลสามารถอธิบายได้ดังต่อไปนี้

- แอททริบิวต์ RCS file ของแต่ละส่วน (Sections) ในแฟ้มข้อมูลบันทึกสามารถสกัดข้อมูลออกมาเป็นรายชื่อและรายละเอียดของแฟ้มข้อมูล (Files) และไดเรคทอรี่ (Directories) ทั้งหมดของโครงการ จากตัวอย่างแฟ้มข้อมูลบันทึกข้างต้นจะทำให้เกิดระเบียน (Record) ใหม่ขึ้นมาในตารางชื่อ Files และ Directories ดังนี้

ตัวอย่างระเบียนของตารางชื่อ Directories

DirectoryID	DirectoryName	Depth
1	/org.eclipse.jdt.core/Model/org/elipse/jdt/core/	6

ตัวอย่างระเบียนของตารางชื่อ Files

FileID	FileName	DirectoryID	FileExtension	Depth	NumberOfRevisions
1	IBuffer.java	1	.java	7	0

- แอททริบิวต์ description ประกอบด้วยส่วนย่อยหลายส่วนแต่ละส่วนแสดงถึงการเปลี่ยนแปลงแก้ไขแต่ละครั้งที่เกิดขึ้น ข้อมูลในแต่ละส่วนย่อยนี้สามารถสกัดข้อมูลออกมาเป็นรายการการเปลี่ยนแปลงแก้ไขแฟ้มข้อมูล (Revisions) ได้ จากตัวอย่างแฟ้มข้อมูลบันทึกข้างต้นจะทำให้เกิดระเบียนใหม่ขึ้นมาในตารางชื่อ Revisions ดังนี้

ตัวอย่างระเบียนของตารางชื่อ Revisions

FileID	RevisionID	TransactionID	CheckinTime	Plus	Minus	State	BranchPrefix
1	1.17	1	2004-01-13 15:48:42	1	1	Exp	NULL

- รายการการเปลี่ยนแปลงแก้ไขแฟ้มข้อมูลที่ได้มาข้างต้นจะถูกนำมาพิจารณาว่ารายการใดบางที่เกิดขึ้นในเวลาเดียวกันและเกิดขึ้นโดยนักพัฒนาคนเดียวกันจะถูกรวมกันไว้เป็นทรานแซคชั่นของการเปลี่ยนแปลงแก้ไข (Transactions) เดียวกัน จากตัวอย่างแฟ้มข้อมูลบันทึกข้างต้นจะทำให้เกิดระเบียนใหม่ขึ้นมาในตารางชื่อ Transactions ดังนี้

ตัวอย่างระเบียนของตารางชื่อ Transactions

TransactionID	Author	Message	MessageMD5	BeginTime	EndTime	IsNoise
1	jlanneluc	Updated copyrights to 2004	817397A1A 8F94C3C8 1AF1C5DB E9F37F7	2004-01-13 15:48:42	2004-01-13 15:48:42	N

- แอททริบิวต์ symbolic name แต่ละส่วนย่อยของแอททริบิวต์ description ในแฟ้มข้อมูลบันทึกสามารถสกัดข้อมูลออกมาเป็นรายชื่อของแท็ก (Tags) ที่นักพัฒนาตั้งไว้ให้กับการเปลี่ยนแปลงแก้ไขนั้นๆได้ จากตัวอย่างแฟ้มข้อมูลบันทึกข้างต้นจะทำให้เกิดระเบียนใหม่ขึ้นมาในตารางชื่อ Tags ดังนี้

ตัวอย่างระเบียนของตารางชื่อ Tags

FileID	TagName	RevisionID
1	1.16	v_396a

- ตารางชื่อ Branches ในฐานข้อมูลนั้นจะบันทึกจุดต่อกิ่ง (Branch Points) และชื่อของการต่อกิ่ง (Branch Names) ข้อมูลทั้ง 2 ข้อมูลนี้ถูกเก็บมาจาก 2 ส่วนของข้อมูลที่ได้จากการเรียกคำสั่ง CVS log โดยที่ชื่อของการต่อกิ่งก็คือชื่อที่เป็นสัญลักษณ์ที่มีหมายเลขปรากฏอยู่ด้วย ตัวอย่างเช่น JDK_1_5 มีหมายเลขเวอร์ชั่นเป็น 1.15.0.2 จะได้ชื่อของการต่อกิ่งนี้เป็น 1.15.2 ส่วนจุดต่อกิ่งนั้นได้มาจากตารางแฮช (Hash Map) ที่ใช้ชื่อของการต่อกิ่งเป็นคีย์

ตัวอย่างระเบียนของตารางชื่อ Branches

FileID	BranchPrefix	OriginRevision	BranchName	InternalRevision
1	1.15.2	1.15	JDK_1_5	1.15.0.2

- **การซ่อมแซมทรานแซคชั่น** (Restoring Transactions)

การเข้าไปอ่านข้อมูลที่ถูกบันทึกเอาไว้ในแฟ้มข้อมูลบันทึก (Log File) ต่างๆบนเครื่องแม่ข่ายของระบบคอนเคอเรนท์เวอร์ชั่นว่ามีข้อความบันทึก (Log Message) ใดบ้างที่ระบุการเปลี่ยนแปลงแก้ไขทั้งหมดที่เกิดจากนักพัฒนาคนเดียวกันและเกิดขึ้นในเวลาเดียวกัน ข้อมูลการเปลี่ยนแปลงแก้ไขที่เกิดจากนักพัฒนาคนเดียวกันและเกิดขึ้นในเวลาเดียวกันจะถูกนำมาแปลงเป็นทรานแซคชั่น 1 ทรานแซคชั่นแล้วบันทึกลงในตาราง Transactions บนฐานข้อมูล คำว่า ในเวลาเดียวกัน ในที่นี้นั้นหมายรวมถึงในเวลาที่ใกล้เคียงกันด้วย เนื่องจากการคอมมิทในแต่ละครั้งอาจใช้เวลาในการดำเนินการหลายวินาทีหรือหลายนาที โดยเฉพาะอย่างยิ่งการคอมมิททีละหลายๆแฟ้มข้อมูล (Zimmermann et al., 2004) ดังนั้นในทางปฏิบัติแล้วนอกจากการพิจารณาที่การคอมมิทในเวลาเดียวกันแล้วยังต้องมีวิธีการพิจารณาที่การคอมมิทระหว่างช่วงของเวลา (Time Interval) เดียวกันด้วย วิธีการพิจารณาการเปลี่ยนแปลงแก้ไขระหว่างช่วงของเวลานั้นมี 2 วิธีคือวิธี

กำหนดกรอบเวลาที่แน่นอน (Fixed Time Windows) และวิธีเลื่อนกรอบเวลา (Sliding Time Windows) ตามที่ได้อธิบายอย่างละเอียดในบทที่ 2 (Zimmermann et al., 2004)

ข้อมูลเข้า คือ ทรานแซคชั่นของการเปลี่ยนแปลงแก้ไขและข้อมูลที่เกี่ยวข้องอื่นๆ (ฐานข้อมูลทั้ง 6 ตาราง คือตารางชื่อ Files ตารางชื่อ Directories ตารางชื่อ Tags ตารางชื่อ Branches ตารางชื่อ Revisions และตารางชื่อ Transactions)

ข้อมูลออก คือ ทรานแซคชั่นของการเปลี่ยนแปลงแก้ไข (เฉพาะตารางชื่อ Transactions ที่ถูกปรับปรุงข้อมูลใหม่)

กระบวนการทำงาน สำหรับในการทดสอบครั้งนี้ผู้วิจัยเลือกใช้วิธีเลื่อนกรอบเวลา (Sliding Time Windows) แบบเดียวกับที่งานวิจัยในอดีตหลายๆงานวิจัยเลือกใช้ (Zimmermann et al., 2004; Ying et al., 2004; Livshits et al., 2005; Weißgerber et al., 2005; Zimmermann et al., 2005; Kim et al., 2005; Breu et al 2006) หลักการของวิธีการนี้ คือ การกำหนดช่องว่างระหว่างการเปลี่ยนแปลงแก้ไข (Revision, Change) 2 ครั้งที่มากที่สุด จุดเริ่มต้นของกรอบของช่วงเวลาจะถูกเลื่อนไปที่การเปลี่ยนแปลงแก้ไขครั้งต่อไปเสมอตราบใดที่การเปลี่ยนแปลงแก้ไขครั้งต่อไปนั้นมีจุดเริ่มต้นอยู่ภายในกรอบเวลาของการเปลี่ยนแปลงแก้ไขครั้งก่อนหน้า

จากข้อมูลของการเปลี่ยนแปลงแก้ไขที่บันทึกไว้ในตารางชื่อ Revisions และข้อมูลทรานแซคชั่นที่อยู่ในตารางชื่อ Transactions จะต้องถูกดึงขึ้นมาพิจารณาระบุทรานแซคชั่นที่ถูกต้องใหม่ทั้งหมดตามวิธีเลื่อนกรอบเวลาข้างต้นโดยที่ในการทดสอบครั้งนี้ผู้วิจัยกำหนดให้ความกว้างของกรอบเวลาที่พิจารณาอยู่ที่ 200 วินาที เช่นเดียวกับงานวิจัยอื่นๆในอดีต (Zimmermann et al., 2004; Ying et al., 2004; Livshits et al., 2005; Weißgerber et al., 2005; Zimmermann et al., 2005; Kim et al., 2005; Breu et al 2006) สำหรับทุกการเปลี่ยนแปลงแก้ไข α_1, α_2, …, α_k (เรียงตามลำดับเวลาที่บันทึก (time(α_i))) ที่เป็นส่วนหนึ่งของทรานแซคชั่น T เดียวกันนั้น จะต้องอยู่ภายใต้เงื่อนไข

$$\forall \alpha_i \in T : author(\alpha_i) = author(\alpha_1)$$

$$\forall \alpha_i \in T : \log\_message(\alpha_i) = \log\_message(\alpha_1)$$

$$\forall i \in \{2, \dots, k\} : |time(\alpha_i) - time(\alpha_{i-1})| \leq 200sec$$

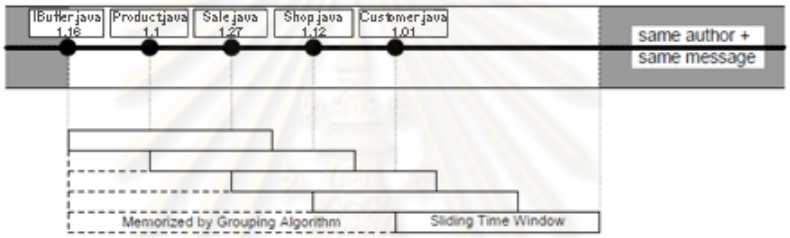
นอกจากนั้นการเปลี่ยนแปลงแก้ไขของแต่ละแฟ้มข้อมูลจะปรากฏอยู่บน 1 ทรานแซคชั่นได้เพียงครั้งเดียว เนื่องจากระบบคอนเคอเรนท์เวอร์ชั่นไม่อนุญาตให้มีการคอมมิทการ

เปลี่ยนแปลงเวอร์ชั่นของแฟ้มข้อมูลเดียวกัน 2 ครั้งในเวลาเดียวกันได้ ดังนั้นจึงมีเงื่อนไขเพิ่มมาอีก 1 ข้อดังนี้

$$\forall \alpha_a, \alpha_b \in T : \alpha_a \neq \alpha_b \rightarrow file(\alpha_a) \neq file(\alpha_b)$$

เมื่อพิจารณาทรานแซคชั่นทั้งหมดเรียบร้อยแล้วการทำการแก้ไขระเบียนภายในตารางชื่อ Transactions ใหม่ให้ถูกต้อง

ตัวอย่างการทำงานของวิธีเลื่อนกรอบเวลา แสดงได้ดังต่อไปนี้ (Zimmermann et al., 2004)



รูปที่ 3-3 แสดงตัวอย่างการพิจารณาการเปลี่ยนแปลงแก้ไขเวอร์ชั่นระหว่างช่วงของเวลาด้วยวิธีเลื่อนกรอบเวลา

รูปที่ 3-3 แสดงการพิจารณาด้วยวิธีเลื่อนกรอบเวลา โดยเริ่มต้นกรอบเวลาที่มีช่วงแน่นอนเริ่มต้นที่จุดของการเปลี่ยนแปลงแก้ไขที่แฟ้มข้อมูล IBuffer.java เวอร์ชั่นที่ 1.16 และกรอบเวลาถูกเลื่อนไปเรื่อยๆ จนสิ้นสุดดังรูป ดังนั้นการเปลี่ยนแปลงแก้ไขที่แฟ้มข้อมูล IBuffer.java เวอร์ชั่นที่ 1.16 แฟ้มข้อมูล Product.java เวอร์ชั่นที่ 1.1 แฟ้มข้อมูล Sale.java เวอร์ชั่นที่ 1.27 แฟ้มข้อมูล Shop.java เวอร์ชั่นที่ 1.12 และ แฟ้มข้อมูล Customer.java เวอร์ชั่นที่ 1.01 จะถูกพิจารณาว่าเกิดขึ้นพร้อมกันและอยู่ภายในทรานแซคชั่นเดียวกัน

- **การระบุการเปลี่ยนแปลงแก้ไขในระดับเอนทิตี้** (Mapping Changes to Entities)

ข้อมูลที่ถูกจัดเก็บไว้ในรีพอสิโทริของระบบคอนเคอเรนท์เวอร์ชั่นนั้นมีเพียงข้อมูลแฟ้มข้อมูลทุกแฟ้มข้อมูลในโครงการและข้อมูลการเปลี่ยนแปลงแก้ไขในระดับแฟ้มข้อมูล (File) หรือคลาส (Class) ที่เก็บอยู่ในรูปของแฟ้มข้อมูลบันทึก (Log File) เท่านั้น แต่ไม่มีบันทึกว่าการ

เปลี่ยนแปลงแก้ไขที่เกิดขึ้นนั้นเกิดขึ้นกับฟังก์ชั่น (Function) หรือ เมธอด (Method) ใดบ้าง มีตัวแปร (Variable) ใดถูกเพิ่มเข้ามา แก้ไข หรือถูกลบออกไปบ้าง วิธีการที่มีความแม่นยำสูงกว่าแต่ก็มีค่าใช้จ่ายในการคำนวณสูงวิธีหนึ่ง คือการกำหนดเอนทิตี้ (ตัวแปร ฟังก์ชั่นหรือเมธอด) ทั้งหมดภายในแฟ้มข้อมูลทั้ง 2 เวอร์ชั่นโดยการนำไปผ่านตัววิเคราะห์ไวยากรณ์ (Parser) จากนั้นก็ทำการเปรียบเทียบซอร์สโค้ดของเอนทิตี้เดียวกันใน 2 เวอร์ชั่น หรือกล่าวคือเป็นการประยุกต์ใช้ฟังก์ชั่นดิฟฟ์ (diff) (Miller et al., 1985) ในระดับของเอนทิตี้นั่นเอง (Zimmermann et al., 2004)

ข้อมูลเข้า คือ แฟ้มข้อมูลบันทึกของระบบคอนเคอเรนท์เวอร์ชั่น (CVS Log File) แฟ้มข้อมูลซอร์สโค้ดทุกเวอร์ชั่นของโครงการ และฐานข้อมูลตารางชื่อ Files ตารางชื่อ Directories ตารางชื่อ Tags ตารางชื่อ Branches ตารางชื่อ Revisions และตารางชื่อ Transactions

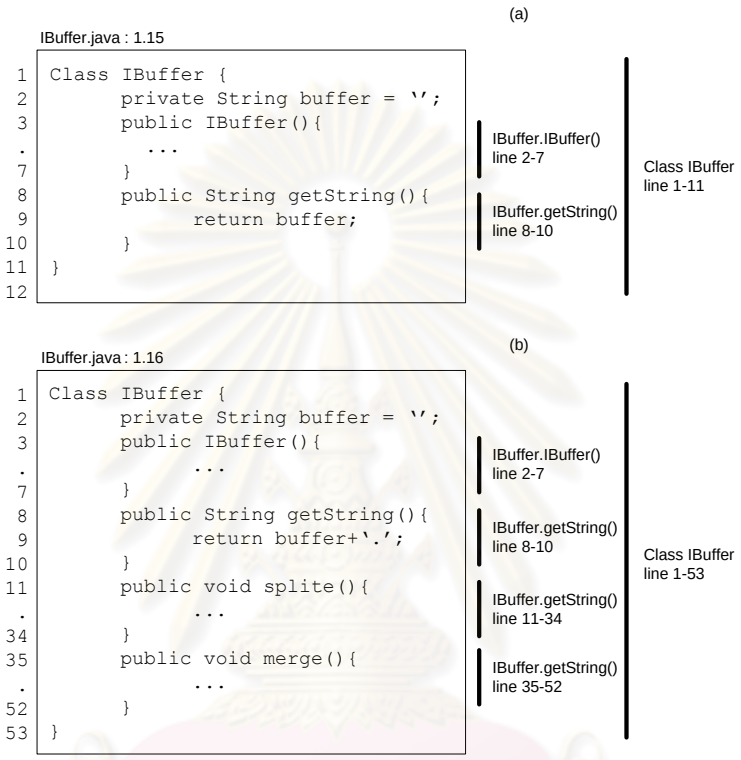
ข้อมูลออก คือ ทรานแซคชั่นของการเปลี่ยนแปลงแก้ไข (เฉพาะตารางชื่อ Transactions และตารางชื่อ Revisions ที่ถูกเพิ่มระเบียบใหม่เข้าไป หรือถูกปรับปรุงข้อมูลใหม่)

กระบวนการทำงาน สามารถดำเนินการได้ดังต่อไปนี้ (Zimmermann et al., 2004)

1) กำหนดเซต E_1 คือเซตของเอนทิตี้ที่มีอยู่ทั้งหมดในเวอร์ชั่น r_1 ของแฟ้มข้อมูล และกำหนดเซต E_2 คือเซตของเอนทิตี้ที่มีอยู่ทั้งหมดในเวอร์ชั่น r_2 ของแฟ้มข้อมูลเดียวกัน
2) เอนทิตี้ที่ถูกเพิ่มเข้ามาใหม่สามารถหาได้จาก $E_2 - E_1$
3) เอนทิตี้ที่ถูกลบออกไปสามารถหาได้จาก $E_1 - E_2$
4) ทุกๆ เอนทิตี้ที่อยู่ในเซต $E_1 \cap E_2$ อาจจะเป็นเอนทิตี้ที่มีการเปลี่ยนแปลงภายใน การตัดสินว่าเอนทิตี้ใดบ้างที่มีการเปลี่ยนแปลงแก้ไขสามารถทำได้โดยการประยุกต์ใช้ฟังก์ชั่นดิฟฟ์ (diff) กับซอร์สโค้ดของเอนทิตี้นั้นๆของทั้ง 2 เวอร์ชั่น

ภายในแพลตฟอร์ม (Platform) ของอิคลิพซ์ (Eclipse) นั้นมีการจัดเตรียมโครงร่าง (Framework) สำหรับการเปรียบเทียบความแตกต่างระหว่าง 2 ซอร์สโค้ดใดๆที่มีประสิทธิภาพสูง และสามารถนำไปประยุกต์เพิ่มเติมได้ทั้งหมด 2 โครงร่างได้แก่โครงร่างเรนจ์ดิฟเฟอร์เรนเซอร์ (Range Differencer) และโครงร่างสตรัคเชอร์เมิร์จวิวเวอร์ (Structure Merge Viewer) ซึ่งได้กล่าวไปแล้วในบทที่ 2 วิธีการเปรียบเทียบซอร์สโค้ดของเอนทิตี้เดียวกันใน 2 เวอร์ชั่นของงานวิจัยชิ้นนี้ผู้วิจัยเลือกใช้โครงร่างเรนจ์ดิฟเฟอร์เรนเซอร์ (Zimmermann et al., 2004)

ตัวอย่างของการระบุการเปลี่ยนแปลงแก้ไขในระดับเอนทิตี้ตั้งแต่ขั้นตอนการระบุเอนทิตี้ของทั้ง 2 เวอร์ชั่นของแฟ้มข้อมูลจนถึงขั้นตอนวิธีการเปรียบเทียบความแตกต่างของซอร์สโค้ดของเอนทิตี้เดียวกันใน 2 เวอร์ชั่น ดังต่อไปนี้



รูปที่ 3-4 แสดงตัวอย่างการระบุเอนทิตี้ภายในซอร์สโค้ด 2 เวอร์ชั่นของแฟ้มข้อมูล IBuffer.java

จากรูปที่ 3-4 (a) แสดงซอร์สโค้ดของแฟ้มข้อมูล IBuffer.java ในเวอร์ชั่นที่ 1.15 และรูปที่ 3-4 (b) แสดงซอร์สโค้ดของแฟ้มข้อมูล IBuffer.java ในเวอร์ชั่นที่ 1.16 เมื่อนำซอร์สโค้ดทั้ง 2 เวอร์ชั่นข้างต้นไปผ่านการวิเคราะห์ไวยากรณ์ (Parse) โดยตัววิเคราะห์ไวยากรณ์ (Parser) ทำให้ได้เซต E_1 คือเซตของเอนทิตี้ที่มีอยู่ทั้งหมดในเวอร์ชั่น 1.15 ของแฟ้มข้อมูล IBuffer.java และได้เซต E_2 คือเซตของเอนทิตี้ที่มีอยู่ทั้งหมดในเวอร์ชั่น 1.16 ของแฟ้มข้อมูลเดียวกัน ดังนี้

E_1 = { (variable, buffer, (Class, IBuffer, (file, IBuffer.java, …))),

(method, IBuffer(), (Class, IBuffer, (file, IBuffer.java, …))),

(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))),

(Class, IBuffer, (file, IBuffer.java, …)),

(file, IBuffer.java, …)

}

E_2 = { (variable, buffer, (Class, IBuffer, (file, IBuffer.java, …))),

(method, IBuffer(), (Class, IBuffer, (file, IBuffer.java, …))),

(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))),

(method, splite(), (Class, IBuffer, (file, IBuffer.java, …))),

(method, Merge(), (Class, IBuffer, (file, IBuffer.java, …))),

(Class, IBuffer, (file, IBuffer.java, …)),

(file, IBuffer.java, …)

}

เอนทิตี้ที่ถูกเพิ่มเข้ามาใหม่สามารถหาได้จาก E_2 - E_1 ดังนี้

E_2 - E_1 = { (method, splite(), (Class, IBuffer, (file, IBuffer.java, …))),

(method, Merge(), (Class, IBuffer, (file, IBuffer.java, …)))

}

จากเซต E_2 - E_1 ข้างต้นทำให้ทราบว่ามีการเพิ่มเอนทิตี้ประเภทเมธอดชื่อ splite() และ Merge() เข้าสู่เอนทิตี้ประเภทคลาสชื่อ IBuffer นั่นคือเกิดเซตของการเปลี่ยนแปลงแก้ไขดังนี้

{ add_to(Class, IBuffer, (file, IBuffer.java, …)),

add_to(Class, IBuffer, (file, IBuffer.java, …))

}

เอนทิตี้ที่ถูกลบออกไปสามารถหาได้จาก E_1 - E_2 ดังนี้

$E_1 - E_2 = \{ \}$

จากเซต $E_1 - E_2$ ข้างต้นเป็นเซตว่างดังนั้นทำให้เกิดเซตของการเปลี่ยนแปลงแก้ไขที่เป็นเซตว่างเช่นกัน

เอนทิตี้ที่อาจจะถูกเปลี่ยนแปลงแก้ไขภายในสามารถหาได้จาก $E_1 \cap E_2$ ดังนี้

$E_1 \cap E_2$ = { (variable, buffer, (Class, IBuffer, (file, IBuffer.java, …))),

(method, IBuffer(), (Class, IBuffer, (file, IBuffer.java, …))),

(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))),

(Class, IBuffer, (file, IBuffer.java, …)),

(file, IBuffer.java, …)

}

การตัดสินว่าเอนทิตี้ใดบ้างที่มีการเปลี่ยนแปลงแก้ไขภายในจริงๆสามารถทำได้โดยการประยุกต์ใช้โครงร่างเรนจ์ดิฟเฟอร์เรนเซอร์ (Range Differencer) สำหรับการเปรียบเทียบความแตกต่างระหว่าง 2 ซอร์สโค้ดใดๆ ผลของการเปรียบเทียบคือมีการเปลี่ยนแปลงภายในเพียงเอนทิตี้เดียวคือเอนทิตี้ประเภทเมธอดชื่อ getString() จาก return buffer; ในเวอร์ชั่น 1.15 เป็น return buffer+'.'; ในเวอร์ชั่น 1.16 นั่นคือมีเซตของการเปลี่ยนแปลงแก้ไขดังนี้

{ alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))) }

ดังนั้นเซตของการเปลี่ยนแปลงแก้ไขทั้งหมดที่เกิดขึ้นจากเวอร์ชั่น 1.15 ไปเป็นเวอร์ชั่น 1.16 ของแฟ้มข้อมูล IBuffer.java คือ

{ add_to(Class, IBuffer, (file, IBuffer.java, …)),

add_to(Class, IBuffer, (file, IBuffer.java, …)),

alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …)))

}

ดำเนินการขั้นตอนทั้งหมดนี้ซ้ำกับแฟ้มข้อมูล Product.java จากเวอร์ชั่น 1.07 ไปเป็นเวอร์ชั่นที่ 1.1 แฟ้มข้อมูล Sale.java จากเวอร์ชั่น 1.26 ไปเป็นเวอร์ชั่นที่ 1.27 แฟ้มข้อมูล Shop.java จากเวอร์ชั่น 1.11 ไปเป็นเวอร์ชั่นที่ 1.12 และ แฟ้มข้อมูล Customer.java จากเวอร์ชั่น 1.0 ไปเป็นเวอร์ชั่นที่ 1.01 เมื่อได้เซตของการเปลี่ยนแปลงแก้ไขทั้งหมดมาแล้วให้นำการเปลี่ยนแปลงแก้ไขเหล่านั้นมารวมกันเป็นเซตเดียวแล้วเรียกเซตนั้นว่าเซตทรานแซคชั่น 1 ทรานแซคชั่น เพื่อความสะดวกผู้วิจัยจึงสมมุติเซตของการเปลี่ยนแปลงแก้ไขของแฟ้มข้อมูลทั้งหมดที่ได้หลังจากทำขั้นตอนข้างต้นเรียบร้อยแล้ว ดังนี้

เซตของการเปลี่ยนแปลงแก้ไขทั้งหมดที่เกิดขึ้นจากเวอร์ชั่น 1.07 ไปเป็นเวอร์ชั่น 1.1 ของแฟ้มข้อมูล Product.java คือ

{ alter(method, getPrice(), (Class, Product, (file, Product.java, …))),

alter(method, getDetail(), (Class, Product, (file, Product.java, …))),

del_from(Class, Product, (file, Product.java, …))

}

เซตของการเปลี่ยนแปลงแก้ไขทั้งหมดที่เกิดขึ้นจากเวอร์ชั่น 1.26 ไปเป็นเวอร์ชั่น 1.27 ของแฟ้มข้อมูล Sale.java คือ

{ alter(method, add(), (Class, Sale, (file, Sale.java, …))),

alter(method, remove(), (Class, Sale, (file, Sale.java, …))),

alter(Class, Sale, (file, Sale.java, …)),

alter(Class, Sale, (file, Sale.java, …)))

}

เซตของการเปลี่ยนแปลงแก้ไขทั้งหมดที่เกิดขึ้นจากเวอร์ชั่น 1.11 ไปเป็นเวอร์ชั่น 1.12 ของแฟ้มข้อมูล Shop.java คือ

{ alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))),

alter(Class, Shop, (file, Shop.java, …)))

}

เซตของการเปลี่ยนแปลงแก้ไขทั้งหมดที่เกิดขึ้นจากเวอร์ชั่น 1.0 ไปเป็นเวอร์ชั่น 1.01 ของแฟ้มข้อมูล Customer.java คือ

{ alter(method, add(), (Class, Customer, (file, Customer.java, …))),

alter(method, getID(), (Class, Customer, (file, Customer.java, …))),

alter(Class, Customer, (file, Customer.java, …)),

del_from(Class, Customer, (file, Customer.java, …)))

}

สำหรับงานวิจัยนี้ผู้วิจัยสนใจทรานแซคชั่นของการเปลี่ยนแปลงแก้ไขเฉพาะในระดับของแฟ้มข้อมูล (file) และคลาส (Class) เท่านั้น แต่อย่างไรก็ตามขั้นตอนการระบุการเปลี่ยนแปลงแก้ไขในระดับเอนทิตี้นี้ก็ยังคงจำเป็นสำหรับงานวิจัยนี้เนื่องจากในภาษาซีพลัสพลัส (C++) นั้นอนุญาตให้ใน 1 แฟ้มข้อมูลสามารถมีคลาสได้มากกว่า 1 คลาส ดังนั้นขั้นตอนการระบุการเปลี่ยนแปลงแก้ไขในระดับเอนทิตี้นี้จะช่วยในการระบุเอนทิตี้ระดับคลาสของกรณีดังกล่าวได้

- **การกำจัดสิ่งแปลกปลอม (Data Cleaning)**

ขั้นตอนการกำจัดสิ่งแปลกปลอม (Data Cleaning) เป็นขั้นตอนที่เข้าไปตรวจสอบข้อมูลทั้งหมดเพื่อค้นหาสิ่งแปลกปลอมและกำจัดสิ่งแปลกปลอมเหล่านั้นออกไป ลักษณะของข้อมูลทรานแซคชั่นที่จะถูกระบุว่าเป็นสิ่งแปลกปลอมมีอยู่ 2 ลักษณะคือ 1) ทรานแซคชั่นขนาดใหญ่ (Large Transactions) และ 2) ทรานแซคชั่นการผสานกิ่ง (Merge Transactions) ตามที่ได้อธิบายรายละเอียดไว้ในบทที่ 2

ข้อมูลเข้า คือ ทรานแซคชั่นของการเปลี่ยนแปลงแก้ไข (เฉพาะตารางชื่อ Transactions และตารางชื่อ Revisions)

ข้อมูลออก คือ ทรานแซคชั่นของการเปลี่ยนแปลงแก้ไข (ฐานข้อมูลเฉพาะตารางชื่อ Transactions และตารางชื่อ Revisions ที่ถูกปรับปรุงข้อมูลใหม่)

กระบวนการทำงาน คือ พิจารณาทรานแซคชั่นที่ได้มาจากข้อมูลออกของขั้นตอนการระบุการเปลี่ยนแปลงแก้ไขในระดับเอนทิตี้ข้างต้นว่ามีทรานแซคชั่นใดมีลักษณะเข้าข่ายที่จะเป็นสิ่งแปลกปลอมทั้ง 2 ลักษณะหรือไม่ ถ้าพบทรานแซคชั่นที่เข้าข่ายดังกล่าวจะทำการลบทรานแซคชั่นเหล่านั้นออกไป

จากตัวอย่างของเซตรายการการเปลี่ยนแปลงแก้ไขที่ได้มาจากขั้นตอนการระบุการเปลี่ยนแปลงแก้ไขในระดับเอนทิตี้ในข้างนั้นมีทรานแซคชั่นที่มีลักษณะเป็นทรานแซคชั่นการผสานกิ่งตามเงื่อนไขที่ได้กล่าวไปในบทที่ 2 จึงทำให้ผลลัพธ์หลังจากผ่านขั้นตอนการกำจัดสิ่งแปลกปลอมเหลือเซตรายการการเปลี่ยนแปลงแก้ไข 3 เซต ดังนี้

```
{ add_to(Class, IBuffer, (file, IBuffer.java, …)),
  add_to(Class, IBuffer, (file, IBuffer.java, …)),
  alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …)))
}
{  alter(method, add(), (Class, Sale, (file, Sale.java, …))),
   alter(method, remove(), (Class, Sale, (file, Sale.java, …))),
   alter(Class, Sale, (file, Sale.java, …)),
   alter(Class, Sale, (file, Sale.java, …))),
}
{  alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))),
   alter(Class, Shop, (file, Shop.java, …)))
}
```

เซตของการเปลี่ยนแปลงแก้ไขเหล่านี้จะถูกนำมารวมกันและตัดรายการการเปลี่ยนแปลงแก้ไขที่ซ้ำซ้อนกันออกแล้วเรียกเซตนี้ว่าทรานแซคชั่น กล่าวคือจากตัวอย่างที่ยกมาข้างต้นนั้นเมื่อนำมาผ่านส่วนการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 4 ขั้นตอนแล้วทำให้เกิดทรานแซคชั่นขึ้น 1 ทรานแซคชั่นที่มีรายการของการเปลี่ยนแปลงแก้ไข 7 รายการดังนี้

T = { add_to(Class, IBuffer, (file, IBuffer.java, …)),

alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))),

alter(method, add(), (Class, Sale, (file, Sale.java, …))),

alter(method, remove(), (Class, Sale, (file, Sale.java, …))),

alter(Class, Sale, (file, Sale.java, …)),

alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))),

alter(Class, Shop, (file, Shop.java, …)))

}

ข้อมูลทรานแซคชั่นทั้งหมดที่ได้จากส่วนนี้คือข้อมูลทรานแซคชั่นของการเปลี่ยนแปลงแก้ไขที่จะนำไปใช้ในการสร้างกฎความสัมพันธ์ในส่วนที่ 3 นอกจากนั้นทรานแซคชั่นเหล่านี้จะถูกนำไปวิเคราะห์เพื่อคัดเลือกขึ้นมาสร้างเป็นข้อสอบถามสำหรับการทดสอบทั้ง 3 สถานการณ์ในส่วนที่ 2 ด้วย

3.6.2 การสร้างข้อสอบถาม

การสร้างข้อสอบถาม คือ ส่วนที่นำทรานแซคชั่นของการเปลี่ยนแปลงแก้ไขทั้งหมดที่มาจากส่วนการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์มาคัดเลือกทรานแซคชั่นจำนวนหนึ่งที่มีความหมายทางสถิติ เรียกว่าทรานแซคชั่นชุดทดสอบ (Test set) เพื่อ

นำทรานแซคชั่นเหล่านั้นมาใช้ในการทดสอบ หลังจากนั้นจึงนำทรานแซคชั่นที่คัดเลือกไว้ไปสร้างเป็นข้อสอบถาม (Query) สำหรับการทดสอบสถานการณ์ 3 สถานการณ์ คือ 1) ข้อสอบถามสำหรับการทดสอบสถานการณ์การนำทาง 2) ข้อสอบถามสำหรับการทดสอบสถานการณ์การป้องกันการเกิดข้อผิดพลาด และ 3) ข้อสอบถามสำหรับการทดสอบสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว ตามข้อกำหนดที่กำหนดไว้ในหัวข้อตัวแปรควบคุม

ข้อมูลเข้า คือ ทรานแซคชั่นของการเปลี่ยนแปลงแก้ไข (เฉพาะตารางชื่อ Revisions และตารางชื่อ Transactions)

ข้อมูลออก คือ ข้อสอบถามสำหรับการทดสอบสถานการณ์การนำทาง ข้อสอบถามสำหรับการทดสอบสถานการณ์การป้องกันการเกิดข้อผิดพลาด และข้อสอบถามสำหรับการทดสอบสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (ตารางชื่อ Queries)

กระบวนการทำงานของการสร้างข้อสอบถามสำหรับแต่ละสถานการณ์มีข้อกำหนดในการสร้างที่แตกต่างกันออกไป ข้อกำหนดในการสร้างข้อสอบสำหรับทั้ง 3 สถานการณ์สามารถแบ่งออกเป็น 2 ขั้นตอนย่อยดังนี้

1) การเลือกทรานแซคชั่นชุดทดสอบ
2) การสร้างข้อสอบถามแต่ละสถานการณ์

- **การเลือกทรานแซคชั่นชุดทดสอบ**

ข้อมูลซอฟต์แวร์อาร์ไคฟว์ที่นำมาใช้ในการวิจัยนี้คือข้อมูลซอฟต์แวร์อาร์ไคฟว์จากโครงการพัฒนาซอฟต์แวร์ทางการบัญชีชื่อเคมายมันนี่ (KMyMoney) ที่มี Thomas Baumgart และ Michael Edwardes เป็นผู้ก่อตั้งโครงการและปัจจุบันมีนักพัฒนาในโครงการนี้ทั้งหมด 10 คน ช่องทางการติดต่อสื่อสารระหว่างนักพัฒนาภายในโครงการคือการใช้ระบบไออาร์ซี (IRC Channel) ที่ต้องเข้าสู่ระบบโดยใช้รหัสผู้ใช้ (Username) และรหัสผ่าน (Password) ของนักพัฒนาซึ่งต้องร้องขอและได้รับอนุมัติในการร้องขอเพื่อเข้าร่วมเป็นนักพัฒนาของโครงการ นอกจากนั้นในกลุ่มนักพัฒนาทั้งหมดมีเพียง Thomas Baumgart และ Martin Preuss ที่เปิดเผยข้อมูลส่วนตัว ด้วยเหตุนี้ทำให้ผู้วิจัยมีข้อจำกัดในการสร้างข้อสอบถาม (Query) ที่ได้รับการประเมินชุดทดสอบจากผู้เชี่ยวชาญ (ในกรณีนี้ ผู้เชี่ยวชาญคือนักพัฒนาที่อยู่ในโครงการนี้) ผู้วิจัยจึงใช้วิธีการเลือก

ทรานแซคชั่นตัวแทนที่มีความหมายทางสถิติมาจำนวนหนึ่งเพื่อนำทรานแซคชั่นเหล่านั้นมาสร้างเป็นข้อสอบถามที่ใช้ในการวิจัยครั้งนี้

ทรานแซคชั่นชุดทดสอบที่ผู้วิจัยเลือกมาเป็นตัวแทนเพื่อสร้างข้อสอบถามสำหรับการทดสอบนี้หรือที่เรียกว่า ทรานแซคชั่นชุดทดสอบ (Test set) ผู้วิจัยกำหนดให้มีทรานแซคชั่นชุดทดสอบทั้งหมด 60 ทรานแซคชั่น ซึ่งถือว่าเพียงพอสำหรับข้อสอบถามในงานวิจัยทางด้านการค้นคืนข้อมูล (Information Retrieval) เนื่องจากข้อสอบถามในงานวิจัยทางด้านการค้นคืนข้อมูลควรมีอย่างน้อย 30 หน่วยทดสอบ (Baeza-Yates and Riberio-Neto, 1999) จากข้อจำกัดที่กล่าวไปข้างต้นการสร้างข้อสอบถามจึงต้องเลือกจากทรานแซคชั่นในฐานข้อมูลที่มีความหมายทางสถิติ โดยเลือกทรานแซคชั่นชุดดทดสอบ 60 ทรานแซคชั่นเป็นตัวแทนของทรานแซคชั่นที่มีขนาดสั้น กลาง ยาวและแบ่งเป็นทรานแซคชั่นที่พบบ่อยและพบไม่บ่อยด้วย ดังนั้นทรานแซคชั่นชุดทดสอบที่จะเลือกมาจะแบ่งได้เป็น 6 กลุ่มคือ 1) ทรานแซคชั่นขนาดสั้นและพบบ่อย 2) ทรานแซคชั่นขนาดสั้นและพบไม่บ่อย 3) ทรานแซคชั่นขนาดกลางและพบบ่อย 4) ทรานแซคชั่นขนาดกลางและพบไม่บ่อย 5) ทรานแซคชั่นขนาดยาวและพบบ่อย และ 6) ทรานแซคชั่นขนาดยาวและพบไม่บ่อย ผู้วิจัยเลือกกำหนดให้แต่ละกลุ่มมีจำนวนที่เท่ากันคือกลุ่มละ 10 ทรานแซคชั่น หรือสามารถแสดงได้ดังตารางต่อไปนี้

ตารางที่ 3-2 แสดงจำนวนของทรานแซคชั่นในแต่ละกลุ่มที่จะเลือกขึ้นมาสร้างเป็นข้อสอบถาม

การปรากฏ \ ขนาด	สั้น	กลาง	ยาว
พบบ่อย	10	10	10
พบไม่บ่อย	10	10	10

ขนาดของทรานแซคชั่นนั้นนับจากจำนวนของการเปลี่ยนแปลงแก้ไขทั้งหมดในทรานแซคชั่น ส่วนวิธีการนับจำนวนการปรากฏของรูปแบบทรานแซคชั่นจะนับจากทรานแซคชั่นที่เป็นซูเปอร์เซตของรูปแบบทรานแซคชั่นนั้นได้

รายละเอียดของการเลือกทรานแซคชั่นชุดทดสอบที่เฉพาะเจาะจงกับโครงการพัฒนาซอฟต์แวร์เคมายมันนี่ (KMyMoney) นั้นอธิบายไว้ในภาคผนวก ก และนำทรานแซคชั่นที่ได้มาทำการสร้างข้อสอบถามของแต่ละสถานการณ์ด้วยวิธีการสร้างข้อสอบถามที่อธิบายในหัวข้อถัดไป

- **การสร้างข้อสอบถามแต่ละสถานการณ์**

ข้อสอบถาม (Query) สำหรับงานวิจัยนี้ คือ เซตที่ประกอบไปด้วยเซตเหตุการณ์การเปลี่ยนแปลงแก้ไขและเซตผลลัพธ์ที่คาดไว้ โดยจะมีข้อสอบถามทั้งหมด 3 แบบสำหรับ 3 สถานการณ์ที่แตกต่างกันคือสถานการณ์การนำทาง สถานการณ์การป้องกันข้อผิดพลาด และสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว และถูกเขียนในรูปแบบ q = (Q, E) เช่นเดียวกับงานวิจัยของ Zimmermann และคณะในปีค.ศ. 2005 (Zimmermann et al., 2005) และงานวิจัยของ Methanias และคณะในปีค.ศ. 2009 (Methanias et al., 2009) ที่ทำการทดสอบประสิทธิภาพของการค้นหากฎความสัมพันธ์บนข้อมูลซอฟต์แวร์อาร์ไคฟ์สำหรับระบบให้คำแนะนำนักพัฒนาระหว่างการพัฒนาซอฟต์แวร์ ขั้นตอนของการสร้างข้อสอบถามในแต่ละสถานการณ์อธิบายพ้อมยกตัวอย่างได้ดังต่อไปนี้ (Zimmermann et al., 2005; Methanias et al., 2009)

กำหนดให้ ข้อสอบถาม (Query) q ใดๆ ถูกเขียนในรูปแบบ q = (Q, E)

โดยที่ เซต Q คือเซตเหตุการณ์ (Situation) ซึ่งเป็นเซตย่อยของเซตทรานแซคชั่น T

เซต E คือเซตผลลัพธ์ที่คาดไว้ (Expected Result) ซึ่งเป็นเซตย่อยของเซตทรานแซคชั่น T

และเท่ากับเซต T - Q

เซต T คือทรานแซคชั่นที่ประกอบด้วยรายการของการเปลี่ยนแปลงแก้ไขจำนวน |T| รายการ

สถานการณ์การนำทาง (Navigation)

ข้อสอบถามสำหรับสถานการณ์การนำทาง กำหนดให้มีข้อสอบถามทั้งหมด |T| ข้อสอบถามต่อ 1 ทรานแซคชั่น โดยที่ในแต่ละข้อสอบถามนั้นมีสมาชิกในเซตเหตุการณ์ Q เพียง 1 รายการคือรายการ e โดยที่รายการ e เป็นสมาชิกของ ทรานแซคชั่น T และสมาชิกในเซตผลลัพธ์ที่คาดไว้ E ก็คือรายการอีก |T| - 1 รายการที่เหลือที่ไม่ใช่รายการ e (Zimmermann et al., 2005)

สำหรับทรานแซคชั่นชุดทดสอบนั้นประกอบด้วยรายการการเปลี่ยนแปลงแก้ไขทั้งหมด 7 รายการทำให้ได้ข้อสอบถามสำหรับทรานแซคชั่นชุดทดสอบทั้งหมด 7 ข้อสอบถาม โดยที่ในแต่ละ

ข้อสอบถามนั้นมีเซตเหตุการณ์ Q ที่มีสมาชิกเพียง 1 รายการคือ e โดยที่ e เป็นสมาชิกของทรานแซคชั่น T ข้อสอบถามสำหรับการทดสอบสถานการณ์การนำทางจากทรานแซคชั่นชุดทดสอบแสดงดังต่อไปนี้

กำหนดให้ q^n_i คือ ข้อสอบถามสำหรับการทดสอบสถานการณ์การนำทางข้อสอบถามที่ i ในรูป $q^n_i = (Q, E)$

q^n_1 = ({ add_to(Class, IBuffer, (file, IBuffer.java, …)) } — เซตเหตุการณ์

{ alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))) ,
alter(method, add(), (Class, Sale, (file, Sale.java, …))) ,
alter(method, remove(), (Class, Sale, (file, Sale.java, …))) ,
alter(Class, Sale, (file, Sale.java, …)) , — เซตผลลัพธ์ที่คาดไว้
alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))) ,
alter(Class, Shop, (file, Shop.java, …))) }

)

q^n_2 = ({ alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …)))},
{ add_to(Class, IBuffer, (file, IBuffer.java, …)),
alter(method, add(), (Class, Sale, (file, Sale.java, …))),
alter(method, remove(), (Class, Sale, (file, Sale.java, …))),
alter(Class, Sale, (file, Sale.java, …)),
alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))),
alter(Class, Shop, (file, Shop.java, …))) })

q^n_3 = ({ alter(method, add(), (Class, Sale, (file, Sale.java, …)))},

{ alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))),

add_to(Class, IBuffer, (file, IBuffer.java, …)),

alter(method, remove(), (Class, Sale, (file, Sale.java, …))),

alter(Class, Sale, (file, Sale.java, …)),

alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))),

alter(Class, Shop, (file, Shop.java, …))) })

q^n_4 = ({ alter(method, remove(), (Class, Sale, (file, Sale.java, …)))},

{ alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))),

alter(method, add(), (Class, Sale, (file, Sale.java, …))),

add_to(Class, IBuffer, (file, IBuffer.java, …)),

alter(Class, Sale, (file, Sale.java, …)),

alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))),

alter(Class, Shop, (file, Shop.java, …))) })

q^n_5 = ({ alter(Class, Sale, (file, Sale.java, …))},

{ alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))),

alter(method, add(), (Class, Sale, (file, Sale.java, …))),

alter(method, remove(), (Class, Sale, (file, Sale.java, …))),

add_to(Class, IBuffer, (file, IBuffer.java, …)),

alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))),

alter(Class, Shop, (file, Shop.java, …))) })

q^n_6 = ({ alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))))},

{ alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))),

alter(method, add(), (Class, Sale, (file, Sale.java, …))),

alter(method, remove(), (Class, Sale, (file, Sale.java, …))),

alter(Class, Sale, (file, Sale.java, …)),

add_to(Class, IBuffer, (file, IBuffer.java, …)),

alter(Class, Shop, (file, Shop.java, …))) })

q^n_7 = ({ alter(Class, Shop, (file, Shop.java, …))))},

{ alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))),

alter(method, add(), (Class, Sale, (file, Sale.java, …))),

alter(method, remove(), (Class, Sale, (file, Sale.java, …))),

alter(Class, Sale, (file, Sale.java, …)),

alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))),

add_to(Class, IBuffer, (file, IBuffer.java, …)) })

ตัวอย่างระเบียนของตารางชื่อ Queries ในการบันทึกข้อสอบถาม q^n_1

QueryID	QueryType	QAntcSet (Ref:RevisionID)	QConqSet (Ref:RevisionID)
1	Navgation	11	34, 37, 46, 47, 51, 52

สถานการณ์การป้องกันการเกิดข้อผิดพลาด (Error Prevention)

ข้อสอบถามสำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด กำหนดให้มีข้อสอบถามทั้งหมด |T| ข้อสอบถามต่อ 1 ทรานแซคชั่น โดยที่ในแต่ละข้อสอบถามนั้นมีสมาชิกในเซต

เหตุการณ์ Q เท่ากับ |T| - 1 รายการนั่นคือ สมาชิกในเซตเหตุการณ์ Q คือสมาชิกในทรานแซคชั่น T ทุกรายการยกเว้นรายการ e โดยที่รายการ e เป็นสมาชิกของทรานแซคชั่น T และสมาชิกในเซตผลลัพธ์ที่คาดไว้ E มี 1 รายการคือรายการ e นั่นเอง (Zimmermann et al., 2005)

สำหรับทรานแซคชั่นชุดทดสอบนั้นทำให้ได้ข้อสอบถามทั้งหมด 7 ข้อสอบถาม โดยที่ในแต่ละข้อสอบถามนั้นมีเซตเหตุการณ์ Q ที่มีจำนวนสมาชิกเท่ากับ 6 รายการนั่นคือ สมาชิกในเซตเหตุการณ์ Q คือสมาชิกในทรานแซคชั่น T ทุกรายการยกเว้น e โดยที่ e เป็นสมาชิกของทรานแซคชั่น T ข้อสอบถามสำหรับการทดสอบสถานการณ์การนำทางจากทรานแซคชั่นชุดทดสอบแสดงดังต่อไปนี้

กำหนดให้ q^p_i คือ ข้อสอบถามสำหรับการทดสอบสถานการณ์การป้องกันการเกิดข้อผิดพลาดข้อสอบถามที่ i ในรูป $q^p_i = (Q, E)$

q^p_1 = (
{ add_to(Class, IBuffer, (file, IBuffer.java, …)) ,
alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))) ,
alter(method, add(), (Class, Sale, (file, Sale.java, …))) ,
alter(method, remove(), (Class, Sale, (file, Sale.java, …))) ,
alter(Class, Sale, (file, Sale.java, …)) ,
alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))) } ,
— เซตเหตุการณ์
{ alter(Class, Shop, (file, Shop.java, …))) } — เซตผลลัพธ์ที่คาดไว้
)

q^p_2 = ({ add_to(Class, IBuffer, (file, IBuffer.java, …)),
alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))),
alter(method, add(), (Class, Sale, (file, Sale.java, …))),
alter(method, remove(), (Class, Sale, (file, Sale.java, …))),

alter(Class, Sale, (file, Sale.java, …)),

alter(Class, Shop, (file, Shop.java, …))) },

{ alter(method, getDescription(), (Class, Shop, (file, Shop.java, …)))})

q^p_3 = ({ add_to(Class, IBuffer, (file, IBuffer.java, …)),

alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))),

alter(method, add(), (Class, Sale, (file, Sale.java, …))),

alter(method, remove(), (Class, Sale, (file, Sale.java, …))),

alter(Class, Shop, (file, Shop.java, …))),

alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))) },

{ alter(Class, Sale, (file, Sale.java, …))})

q^p_4 = ({ add_to(Class, IBuffer, (file, IBuffer.java, …)),

alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))),

alter(method, add(), (Class, Sale, (file, Sale.java, …))),

alter(Class, Shop, (file, Shop.java, …))),

alter(Class, Sale, (file, Sale.java, …)),

alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))) },

{ alter(method, remove(), (Class, Sale, (file, Sale.java, …)))})

q^p_5 = ({ add_to(Class, IBuffer, (file, IBuffer.java, …)),

alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))),

alter(Class, Shop, (file, Shop.java, …))),

alter(method, remove(), (Class, Sale, (file, Sale.java, …))),

alter(Class, Sale, (file, Sale.java, …)),

alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))) },

{ alter(method, add(), (Class, Sale, (file, Sale.java, …)))})

q^p_6 = ({ add_to(Class, IBuffer, (file, IBuffer.java, …)),

alter(Class, Shop, (file, Shop.java, …))),

alter(method, add(), (Class, Sale, (file, Sale.java, …))),

alter(method, remove(), (Class, Sale, (file, Sale.java, …))),

alter(Class, Sale, (file, Sale.java, …)),

alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))) },

{ alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …)))})

q^p_7 = ({ alter(Class, Shop, (file, Shop.java, …))),

alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))),

alter(method, add(), (Class, Sale, (file, Sale.java, …))),

alter(method, remove(), (Class, Sale, (file, Sale.java, …))),

alter(Class, Sale, (file, Sale.java, …)),

alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))) },

{ add_to(Class, IBuffer, (file, IBuffer.java, …))})

ตัวอย่างระเบียนของตารางชื่อ Queries ในการบันทึกข้อสอบถาม q^p_1

QueryID	QueryType	QAntcSet (Ref:RevisionID)	QConqSet (Ref:RevisionID)
35	Prevention	34, 37, 46, 47, 51, 52	11

สถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (Closure)

ข้อสอบถามสำหรับสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว กำหนดให้มีข้อสอบถามทั้งหมด 1 ข้อสอบถามต่อ 1 ทรานแซคชั่น โดยที่ในแต่ละข้อสอบถามนั้นมีสมาชิกในเซตเหตุการณ์ Q เท่ากับ |T| รายการนั่นคือ สมาชิกในเซตเหตุการณ์ Q คือสมาชิกทุกรายการในทรานแซคชั่น T และเซตผลลัพธ์ที่คาดไว้ E เป็นเซตว่าง (Zimmermann et al., 2005)

สำหรับทรานแซคชั่นชุดทดสอบนั้นทำให้ได้ข้อสอบถามทั้งหมด 1 ข้อสอบถาม โดยที่ในแต่ละข้อสอบถามนั้นมีเซตเหตุการณ์ Q ที่มีจำนวนสมาชิกเท่ากับ 7 รายการนั่นคือ สมาชิกในเซตเหตุการณ์ Q คือสมาชิกทุกรายการในทรานแซคชั่น T ข้อสอบถามสำหรับการทดสอบสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วจากทรานแซคชั่นชุดทดสอบแสดงดังต่อไปนี้

กำหนดให้ q^c_i คือ ข้อสอบถามสำหรับการทดสอบสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วข้อสอบถามที่ i

q^c_1 = ({ add_to(Class, IBuffer, (file, IBuffer.java, …)) ,
alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))),
alter(method, add(), (Class, Sale, (file, Sale.java, …))) , เซตเหตุการณ์
alter(method, remove(), (Class, Sale, (file, Sale.java, …))) ,
alter(Class, Sale, (file, Sale.java, …)) ,
alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))),
alter(Class, Shop, (file, Shop.java, …))) } ,
ø เซตผลลัพธ์ที่คาดไว้

)

ตัวอย่างระเบียนของตารางชื่อ Queries ในการบันทึกข้อสอบถาม q^{p}_{1}

QueryID	QueryType	QAntcSet (Ref:RevisionID)	QConqSet (Ref:RevisionID)
41	Prevention	11, 34, 37, 46, 47, 51, 52	*null*

ข้อสอบถามทั้งหมดที่ได้มานั้นแบ่งออกเป็น 3 ชุด ตามสถานการณ์ต่างกัน 3 สถานการณ์ เพื่อนำไปใช้ในการทดสอบของแต่ละสถานการณ์ในส่วนที่ 4 (หัวข้อ 3.6.4)

3.6.3 การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ทั้ง 2 ตัวแบบ

การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ทั้ง 2 ตัวแบบ คือ ขั้นตอนที่นำทรานแซคชั่นของการเปลี่ยนแปลงแก้ไขทั้งหมดมาทำเหมืองข้อมูลเพื่อค้นหากฎความสัมพันธ์ของการเปลี่ยนแปลงแก้ไข โดยใช้ขั้นตอนวิธีของ Zimmermann และคณะ (Zimmermann et al., 2005) ที่อธิบายอย่างละเอียดไว้ในบทที่ 2 หัวข้อ 2.8.2 มาประยุกต์ใช้สำหรับการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบ

ข้อมูลเข้า คือ ทรานแซคชั่นของการเปลี่ยนแปลงแก้ไข (เฉพาะตารางชื่อ Revisions และตารางชื่อ Transactions)

ข้อมูลออก คือ กฎความสัมพันธ์ที่ได้จากการทำเหมืองข้อมูลทั้ง 2 ตัวแบบรวมถึงค่าสนับสนุนค่าความเชื่อมั่น/ค่าความเชื่อมั่นใหม่ของกฎความสัมพันธ์นั้นๆ (ตารางชื่อ Rules)

กระบวนการทำงานของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ทั้ง 2 ตัวแบบ นั้นคือการนำขั้นตอนวิธีในการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ที่ชื่อว่า ขั้นตอนวิธีอพริโอริ (Apriori algorithm) ที่ถูกนำเสนอโดย Agrawal และ Srikant (Agrawal and Srikant, 1994)

เนื่องจากงานวิจัยนี้สนใจศึกษาว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 หรือการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่

ของ Liu และคณะ (Liu et al., 2008) สามารถเพิ่มประสิทธิภาพของระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ของไอดีอี (IDE: integrated development environment) ได้หรือไม่ ดังนั้นในขั้นตอนการระบุความน่าสนใจของกฎความสัมพันธ์แต่ละกฎในทั้ง 2 ตัวแบบของการค้นหากฎความสัมพันธ์จึงแตกต่างกันดังนี้

- การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 กำหนดให้ใช้การคำนวณค่าสนับสนุนนับ (Support Count) และค่าความเชื่อมั่น (Confidence) ในการระบุความน่าสนใจของกฎความสัมพันธ์ สำหรับการทดสอบนี้ผู้วิจัยกำหนดให้ค่าสนับสนุนนับขั้นต่ำ (Minimum Support Count) เท่ากับ 3 และค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence) เท่ากับ 0.1
- การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 กำหนดให้ใช้การคำนวณค่าสนับสนุนนับและค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008) ในการระบุความน่าสนใจของกฎความสัมพันธ์ สำหรับการทดสอบนี้ผู้วิจัยกำหนดให้ค่าสนับสนุนนับขั้นต่ำเท่ากับ 3 และค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008) ขั้นต่ำเท่ากับ 0.1

สำหรับตัวอย่างของการทำเหมืองข้อมูลด้วยเทคนิคค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบนั้น เนื่องจากผู้วิจัยไม่สามารถสมมุติทรานแซคชั่นทั้งหมดออกมาได้ได้ ผู้วิจัยจึงสมมุติกฎความสัมพันธ์จากการค้นหากฎความสัมพันธ์ ที่เป็นผลลัพธ์ที่ได้มาจากกระบวนการข้างต้น เพื่อประโยชน์ในการอ้างอิงถึงในส่วนต่อไป ดังนี้

กำหนดให้ R คือ เซตของกฎความสัมพันธ์ที่แต่ละรายการอยู่ในรูปแบบ Q -> {x} s;c

Q คือ เซตเหตุการณ์ที่ประกอบด้วยรายการการเปลี่ยนแปลงแก้ไขตั้งแต่ 1 รายการขึ้นไป

x คือ รายการการเปลี่ยนแปลงแก้ไขใดๆ

s คือ ค่าสนับสนุนนับของกฎความสัมพันธ์ Q -> {x}

c คือ ค่าความเชื่อมั่นของกฎความสัมพันธ์ Q -> {x}

R = { alter(method, add(), ...) -> alter(method, getString(), ...) 6:0.5,

alter(method, add(), ...) -> alter(Class, Shop, ...) 5:0.71,

alter(method, getString(), ...) -> alter(method, getString(), ...) 5:0.63,

alter(method, add(), ...) -> alter(method, remove(), ...) 3:0.5,

alter(method, remove(), ...) -> alter(Class, Shop, ...) 2:0.29,

alter(Class, Sale, ...) -> alter(method, remove(), ...) 2:0.5,

alter(method, getDescription(), ...) -> alter(method, getString(), ...)1:0.14

}

ตัวอย่างระเบียนของตารางชื่อ Rules ในการบันทึกกฎความสัมพันธ์แรกในเซตข้างต้น

RuleID	Model	RAntcSet (Ref:RevisionID)	RConqSet (Ref:RevisionID)	Support	Confidence
1	1	11	34	6	0.5

3.6.4 การสร้างเซตของคำแนะนำสำหรับเหตุการณ์

การสร้างเซตของคำแนะนำสำหรับเหตุการณ์ คือ การนำเซตของกฎความสัมพันธ์ R สำหรับเหตุการณ์ Q มาสร้างเป็นเซตของคำแนะนำ (Suggestions) นำเสนอให้กับนักพัฒนาเมื่อนักพัฒนาได้ทำให้เกิดเหตุการณ์ Q ใดๆขึ้นมา เซตของคำแนะนำสำหรับเหตุการณ์ Q สามารถนิยามให้อยู่ในรูปของการยูเนียน (Union) ของเซตรายการที่ตามมาของกฎความสัมพันธ์ R ที่มีเซตรายการที่มาก่อน ตรงกับเซตเหตุการณ์ Q ได้ ดังต่อไปนี้ (Zimmermann et al., 2005)

$$apply_R(Q) = \mathrm{U}_{(\mathrm{Q}\rightarrow\{x_2\})\in R}\, x_2$$

ในงานวิจัยของ Zimmermann และคณะ (Zimmermann et al., 2005) ได้ตั้งข้อสันนิษฐานไว้ว่า การให้คำแนะนำในการเปลี่ยนแปลงแก้ไขกับนักพัฒนานั้น คำแนะนำที่จะได้รับความสนใจก็คือคำแนะนำที่อยู่ใน 10 อันดับแรก ดังนั้นในการสร้างเซตของคำแนะนำจึงควรให้ความสนใจกฎความสัมพันธ์ที่อยู่ใน 10 อันดับแรกโดยเรียงจากค่าสนับสนุนนับและค่าความ

เชื่อมั่นเท่านั้น ดังนั้นผู้วิจัยจึงกำหนดสมการในการสร้างเซตของคำแนะนำดังแสดงในสมการต่อไปนี้

กำหนดให้ q คือ ข้อสอบถาม (Query) ที่ประกอบด้วยเซตเหตุการณ์ (Situation) Q และเซตผลลัพธ์ที่คาดไว้ (Expected Result) E และเขียนให้อยู่ในรูป q = (Q, E)

R คือ เซตของกฎความสัมพันธ์ที่อยู่ในรูปแบบ Q→{x} โดยที่ x คือรายการการเปลี่ยนแปลงแก้ไข และกำหนดให้ R_{10} คือเซตของกฎความสัมพันธ์ที่มีระดับความน่าสนใจสูงสุด 10 กฎแรกซึ่งเรียงลำดับด้วยค่าความเชื่อมั่น โดยที่ $R_{10} \subset R$

A_q คือ เซตของรายการการเปลี่ยนแปลงแก้ไข x ที่ได้จากกฎความสัมพันธ์ในเซต R_{10} ที่สอดคล้องกับเซตเหตุการณ์ Q ของข้อสอบถาม q ซึ่งสามารถเขียนในรูป $A_q = apply_{R_{10}}(Q)$

$$A_q = apply_{R_{10}}(Q)$$

ข้อมูลเข้า คือ เซตของกฎความสัมพันธ์ และเซตเหตุการณ์ของข้อสอบถามชุดทดสอบ (ตารางชื่อ Rules และตารางชื่อ Queries)

ข้อมูลออก คือ เซตของคำแนะนำสำหรับเหตุการณ์ที่นำมาทดสอบ (ตารางชื่อ Suggestions)

กระบวนการทำงานของการสร้างเซตของคำแนะนำสำหรับเหตุการณ์ มีขั้นตอนวิธีดังต่อไปนี้

1) เลือกเหตุการณ์ Q ที่ต้องการนำมาหาคำแนะนำ
2) ค้นหากฎความสัมพันธ์ที่มีเซตรายการที่มาก่อนตรงกับเซตเหตุการณ์ Q มาจากเซตของกฎความสัมพันธ์ทั้งหมด
3) นำเซตรายการที่ตามมาของกฎความสัมพันธ์ที่ได้จากข้อที่ 2 มารวมกัน (Union) เป็นเซตใหม่ให้ชื่อว่า เซตของคำแนะนำสำหรับเหตุการณ์ Q โดยที่แต่ละรายการของเซต

คำแนะนำนี้จะเรียงลำดับตามค่าสนับสนุนนับและค่าความเชื่อมั่นของกฎความสัมพันธ์นั้นๆ

จากเซตของกฎความสัมพันธ์ที่ได้มาจากขั้นตอนที่แล้ว ตัวอย่างของการสร้างเซตของคำแนะนำสำหรับเหตุการณ์ alter(method, add(), (Class, Sale, (file, Sale.java, …))) แสดงได้ดังต่อไปนี้

1) เหตุการณ์ที่เกิดขึ้นคือ เหตุการณ์ alter(method, add(), (Class, Sale, (file, Sale.java, …)))
2) จากเซตของกฎความสัมพันธ์ที่ได้มาจากขั้นตอนที่แล้วมีกฎความสัมพันธ์ที่มีเซตรายการที่มาก่อน เป็น alter(method, add(), (Class, Sale, (file, Sale.java, …))) อยู่ทั้งหมด 3 กฎความสัมพันธ์ดังนี้

   alter(method, add(), ...) -> alter(method, getString(), ...) 6:0.5

   alter(method, add(), ...) -> alter(Class, Shop, ...) 5:0.71

   alter(method, add(), ...) -> alter(method, remove(), ...) 3:0.5

3) นำเซตรายการที่ตามมาของกฎความสัมพันธ์ทั้ง 3 กฎความสัมพันธ์ข้างต้นมารวมกันเป็นเซตของคำแนะนำเมื่อเกิดเห็นการ เหตุการณ์ alter(method, add(), (Class, Sale, (file, Sale.java, …))) ได้ดังต่อไปนี้

   A_q = { alter(method, getString(), ...),
   alter(Class, Shop, ...),
   alter(method, remove(), ...)
   }

ตัวอย่างระเบียนของตารางชื่อ Suggestions ในการบันทึกเซตของคำแนะนำข้างต้น

SuggestionID	ForSituation (Ref:RevisionID)	SuggestionSet (Ref:RevisionID)
1	11	34, 46, 47

3.6.5 การประเมินผลการทดสอบ

เมื่อได้ทำการทดสอบการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์เสร็จสิ้นแล้ว ขั้นตอนต่อไปนี้ก็คือการประเมินผลการทดสอบการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ จุดมุ่งหมายหลักของการทดสอบของงานวิจัยนี้คือ การเปรียบเทียบค่าประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบที่ 1 กับตัวแบบที่ 2 ในสถานการณ์ที่ต่างกัน 3 สถานการณ์ ดังนั้นหลังจากการทดสอบเสร็จสิ้นผู้วิจัยจึงต้องนำผลการทดสอบเหล่านั้นมาคำนวณหาค่าประสิทธิภาพ ซึ่งงานวิจัยนี้จะใช้ค่าเอฟเมสเซอร์ (F-measure) มาเป็นค่าแสดงประสิทธิภาพของสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด และค่าผลสะท้อนกลับ (Feedback) มาเป็นค่าแสดงประสิทธิภาพของสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

ข้อมูลเข้า คือ เซตของคำแนะนำสำหรับเหตุการณ์ของการทดสอบทั้ง 6 การทดสอบ (ตารางชื่อ Suggestions)

ข้อมูลออก คือ ค่าประสิทธิภาพของการทดสอบทั้ง 6 การทดสอบ (แฟ้มข้อมูลประเภทข้อความที่บันทึกค่าประสิทธิภาพของการทดสอบทั้งหมดแยกตามการทดสอบ)

หลังจากขั้นตอนการสร้างเซตของคำแนะนำสำหรับเหตุการณ์แล้ว ผู้วิจัยจะได้เซตของคำแนะนำสำหรับเหตุการณ์ใดๆ ที่สร้างมาจากกฎความสัมพันธ์ของทรานแซคชั่นทั้งหมด เซตของคำแนะนำนั้นจะถูกนำไปเปรียบเทียบกับเซตผลลัพธ์ที่คาดไว้ของชุดทดสอบ สำหรับในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดจะนำมาคำนวณค่าความถูกต้อง (Precision) ค่าเรียกคืน (Recall) และค่าเอฟเมสเซอร์ (F-measure)

ในปี 2005 งานวิจัยของ Zimmermann และคณะ (Zimmermann et al., 2005)) ทำการเปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลบนข้อมูลซอฟต์แวร์อาร์ไคฟว์กับโครงการพัฒนาระบบปฏิบัติการคอมพิวเตอร์ต่างๆ (Operating System) และกล่าวว่า จุดมุ่งหมายของการทดสอบระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์คือค่าความถูกต้องที่สูง (ค่าใกล้เคียง 1) และค่าเรียกคืนที่สูง (ค่าใกล้เคียง 1) นั่นคือต้องการให้ระบบสามารถแนะนำคำแนะนำทั้งหมด (ค่าเรียกคืนเท่ากับ 1) และแต่ละคำแนะนำนั้นถูกต้องหรือตรงกับเซตผลลัพธ์ที่คาดไว้ทั้งหมด (ค่าความถูกต้องเท่ากับ 1) ดังนั้นงานวิจัยของ Zimmermann และคณะจึงใช้

ค่าเฉลี่ยฮาร์โมนินิคของค่าความถูกต้องและค่าเรียกคืน (Harmonic mean of Precision and Recall) หรือค่าเอฟเมสเซอร์ที่ให้น้ำหนักของค่าความถูกต้องและค่าเรียกคืนอย่างสมดุล เป็นค่าประเมินประสิทธิภาพของการทำเหมืองข้อมูล (Zimmermann et al., 2005)

ต่อมาในปี 2009 งานวิจัยของ Methanias และคณะ (Methanias et al., 2009) ทำการเปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลบนข้อมูลซอฟต์แวร์อาร์ไคฟว์ของโครงการซอฟต์แวร์สิ่งแวดล้อมอุตสาหกรรม (Industrial Environment) ใน 3 สถานการณ์เช่นเดียวกัน และแนะนำให้ใช้ค่าเอฟเมสเซอร์ที่ให้น้ำหนักของค่าความถูกต้องและค่าเรียกคืนอย่างสมดุลสำหรับการประเมินประสิทธิภาพในสถานการณ์การนำทางและสถานการณ์การป้องกันข้อผิดพลาด (Methanias et al., 2009)

ดังนั้นงานวิจัยนี้จึงใช้ค่าเอฟเมสเซอร์ที่ $\beta = 1$ หรือค่าเอฟเมสเซอร์ที่ให้น้ำหนักของค่าความถูกต้องและค่าเรียกคืนอย่างสมดุลมาใช้ในการวัดประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด ค่าเอฟเมสเซอร์ (F-measure) ที่ถ่วงน้ำหนักค่าความถูกต้องแต่ค่าเรียกคืนอย่างสมดุล แสดงดังสมการต่อไปนี้

กำหนดให้ F_1 ค่าเอฟเมสเซอร์ (F-measure) ที่ถ่วงน้ำหนักค่าความถูกต้องแต่ค่าเรียกคืนอย่างสมดุล

Precision ค่าความถูกต้อง

Recall ค่าเรียกคืน

q คือ ข้อสอบถาม (Query) ที่ประกอบด้วยเซตเหตุการณ์ (Situation) Q และเซตผลลัพธ์ที่คาดไว้ (Expected Result) E และเขียนให้อยู่ในรูป q = (Q, E)

R คือ เซตของกฎความสัมพันธ์ที่อยู่ในรูปแบบ $Q \rightarrow \{x\}$ โดยที่ x คือรายการการเปลี่ยนแปลงแก้ไข และกำหนดให้ R_{10} คือเซตของกฎความสัมพันธ์ที่มีระดับความน่าสนใจสูงสุด 10 กฎแรกซึ่งเรียงลำดับด้วยค่าความเชื่อมั่น โดยที่ $R_{10} \subset R$

A_q คือ เซตของรายการการเปลี่ยนแปลงแก้ไข x ที่ได้จากกฎความสัมพันธ์ในเซต R_{10} ที่สอดคล้องกับเซตเหตุการณ์ Q ของข้อสอบถาม q ซึ่งสามารถเขียนในรูป $A_q = apply_{R_{10}}(Q)$ เสมอ หรือเรียกว่า เซตของคำแนะนำ

$|A_q \cap E|$ คือ จำนวนรายการการเปลี่ยนแปลงแก้ไขในเซตคำแนะนำที่ตรงกับรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตผลลัพธ์ที่คาดไว้

$|A_q|$ คือ จำนวนรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตของคำแนะนำ

$|E|$ คือ จำนวนรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตผลลัพธ์ที่คาดไว้

$$precision = \frac{|A_q \cap E|}{|A_q|} \quad , \qquad recall = \frac{|A_q \cap E|}{|E|}$$

และ $$F_1 = \frac{2 * precision * recall}{precision + recall}$$

เนื่องจากค่าเอฟเมสเซอร์นั้นจะอยู่ในช่วง 0 ถึง 1 ค่าเอฟเมสเซอร์ที่มีค่าเป็น 0 ในงานวิจัยนี้จะหมายถึงประสิทธิภาพของการค้นหากฎความสัมพันธ์ต่ำหรือรายการการเปลี่ยนแปลงแก้ไขที่ถูกดึงขึ้นมาไม่ตรงกับรายการการเปลี่ยนแปลงแก้ไขใดๆในเซตผลลัพธ์ที่คาดไว้ของข้อสอบถามเลย และค่าเอฟเมสเซอร์ที่มีค่าเท่ากับ 1 ในงานวิจัยนี้จะหมายถึงประสิทธิภาพของการค้นหากฎความสัมพันธ์สูงหรือรายการการเปลี่ยนแปลงแก้ไขที่ถูกดึงขึ้นมาตรงกับเซตผลลัพธ์ที่คาดไว้ของข้อสอบถามทุกรายการ

ตัวอย่างเช่น สมมุติข้อมูลสอบถามชุดทดสอบสำหรับสถานการณ์การนำทาง 1 ข้อสอบถาม และเซตของคำแนะนำสำหรับเหตุการณ์ alter(method, add(), (Class, Sale, (file, Sale.java, …))) ดังนี้

กำหนดให้ q^n คือ ข้อสอบถามชุดทดสอบสำหรับสถานการณ์การนำทาง ในรูป $q^n = (Q, E)$

A_q คือ เซตของคำแนะนำสำหรับเหตุการณ์ alter(method, add(), (Class, Sale, (file, Sale.java, …)))

q^n = ({ alter(method, add(), (Class, Sale, (file, Sale.java, …))) },

{ add_to(Class, IBuffer, (file, IBuffer.java, …)),

alter(method, getString(), (Class, IBuffer, (file, IBuffer.java, …))),

alter(method, remove(), (Class, Sale, (file, Sale.java, …))),

alter(Class, Sale, (file, Sale.java, …)),

alter(method, getDescription(), (Class, Shop, (file, Shop.java, …))),

alter(Class, Shop, (file, Shop.java, …))) })

A_q = { alter(method, getString(), ...),

alter(Class, Shop, ...),

del_from(Class, Product, …),

alter(method, remove(), ...)

}

จากข้อสอบถามชุดทดสอบและเซตของคำแนะนำข้างต้น นำมาคำนวณค่าความถูกต้อง (Precision) ค่าเรียกคืน (Recall) และค่าเอฟเมสเซอร์ (F-measure) ได้ดังนี้

$$precision = \frac{|A_q \cap E|}{|A_q|} = \frac{3}{4} = 0.75$$

$$recall = \frac{|A_q \cap E|}{|E|} = \frac{3}{6} = 0.5$$

$$F_1 = \frac{2 * precision * recall}{precision + recall} = \frac{2 * 0.75 * 0.5}{0.75 + 0.5} = 0.6$$

ค่าเอฟเมสเซอร์ที่คำนวณมาได้นั้นถ้ามีค่าใกล้เคียง 1 มากก็หมายความว่ายิ่งมีประสิทธิภาพมาก

สำหรับสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว เซตของคำแนะนำนั้นถูกนำไปเปรียบเทียบกับเซตผลลัพธ์ที่คาดไว้ของชุดทดสอบและค่าคำนวณผลสะท้อนกลับ (Feedback) ตามสมการต่อไปนี้

กำหนดให้ $|Z^*|$ คือ จำนวนข้อสอบถามที่อยู่ในเซตของข้อสอบถามที่มีเซตของคำแนะนำที่ไม่เป็นเซตว่าง ($|A_q| \neq 0$)

$|Z|$ คือ จำนวนข้อสอบถามทั้งหมด

$$feedback = \frac{|Z^*|}{|Z|}$$

เมื่อนำค่าผลสะท้อนกลับมาใช้วัดในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วจะสามารถแสดงให้เห็นถึงร้อยละของการเกิดการแจ้งเตือนที่ผิด (False alarm) หรือการให้คำแนะนำที่เป็นผลบวกลวง (False Positive) นั่นเอง เนื่องจากค่าผลสะท้อนกลับมีพิสัยอยู่ระหว่าง 0 กับ 1 ค่าผลสะท้อนกลับที่มีค่าเท่ากับ 0 หมายถึงไม่มีข้อสอบถามใดเลยที่ให้คำแนะนำที่เป็นผลบวกลวงออกมาในสถานการณ์นี้นั่นคือมีประสิทธิภาพดีที่สุด และค่าผลสะท้อนกลับที่มีค่าเท่ากับ 1 หมายถึงข้อสอบถามทั้งหมดให้คำแนะนำที่เป็นผลบวกลวงออกมานั่นคือมีประสิทธิภาพไม่ดีที่สุด ดังนั้นการวัดประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วนั้นจะต้องเปรียบเทียบค่าผลสะท้อนกลับและค่าผลสะท้อนกลับที่น้อยกว่าจะมีความหมายว่ามีประสิทธิภาพดีกว่า

ตัวอย่างเช่น สมมุติให้มีข้อสอบถามในชุดทดสอบมีทั้งหมด 100 ข้อสอบถาม และสมมุติให้มีข้อสอบถามที่ได้เซตของคำแนะนำไม่เป็นเซตว่างทั้งหมด 66 ข้อสอบถาม จะสามารถคำนวณค่าคำนวณผลสะท้อนกลับ (Feedback) ได้ดังนี้

$$feedback = \frac{|Z^*|}{|Z|} = \frac{66}{100} = 0.66$$

การพิจารณาประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ สำหรับสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้นสามารถทำได้โดยการนำค่าเอฟเมสเซอร์ที่คำนวณได้มาเปรียบเทียบโดยใช้กราฟในรูปแบบที่เหมาะสม เพื่อง่ายต่อการพิจารณาเปรียบเทียบค่าเอฟเมสเซอร์ของการทำเหมืองข้อมูลด้วยรูปแบบทั้ง 2 ตัวแบบ สำหรับสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วนั้นสามารถทำได้โดยการนำค่าคำนวณผลสะท้อนกลับที่คำนวณได้มาเปรียบเทียบค่ากันได้โดยตรง

3.6.6 การทดสอบสมมติฐาน

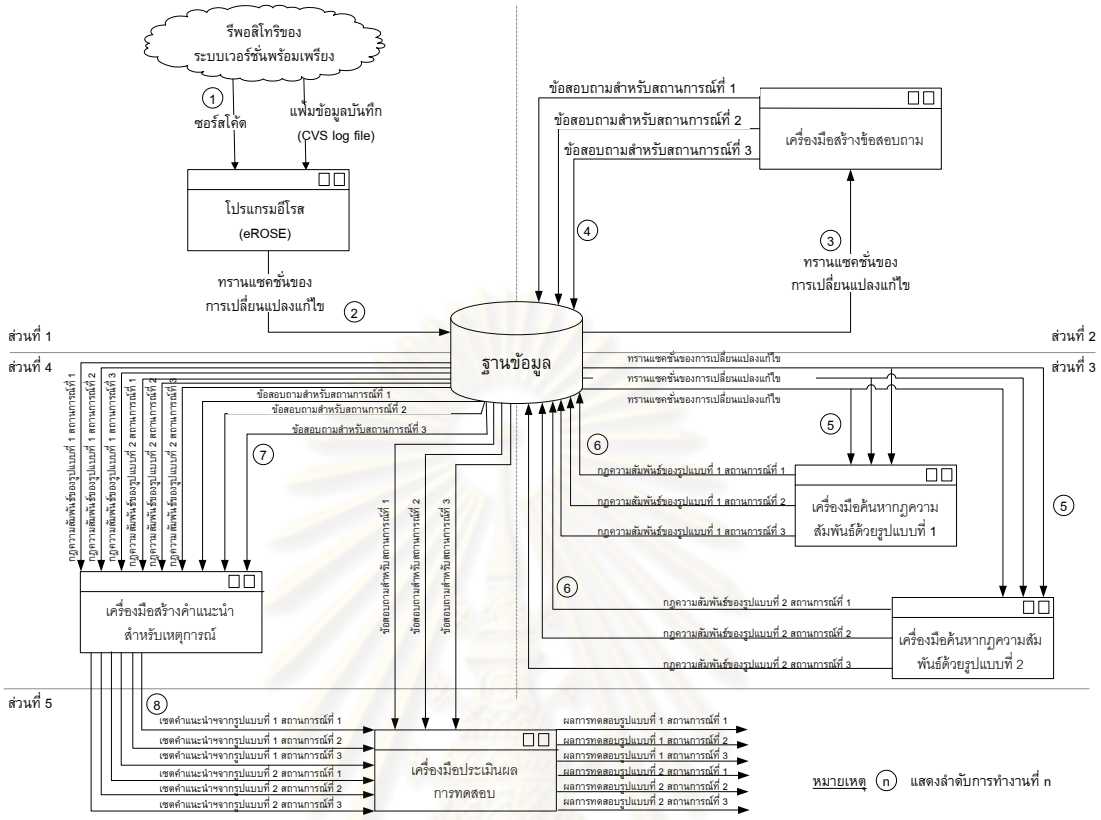
สำหรับกรณีของการทดสอบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้นใช้ค่าเอฟเมสเซอร์เป็นค่าที่แสดงถึงประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ เมื่อการทดสอบประสิทธิภาพเสร็จสิ้นแล้ว ทำให้ได้ค่าเอฟเมสเซอร์ออกมาเท่าจำนวนของข้อสอบถามชุดทดสอบที่สร้างขึ้นในขั้นตอนการสร้างข้อสอบถาม จากนั้นในขั้นตอนแรกจะตรวจสอบการแจกแจงของค่าประสิทธิภาพที่ได้มาว่ามีการแจกแจงปกติหรือไม่ ด้วยการใช้สถิติทดสอบ Kolmogorov-Smirnov เพื่อเลือกทางเลือกในการทดสอบสมมติฐานได้ว่าจะให้การทดสอบสมมติฐานแบบใช้พารามิเตอร์ (Parametric Test)

หรือแบบไม่อิงกับพารามิเตอร์ (Non Parametric Test) ถ้าผลการทดสอบพบว่าประชากรมีการแจกแจงแบบปกติ จึงใช้การวิเคราะห์โดยสถิติทดสอบที (t-test) เพื่อทดสอบสมมติฐานของผลต่างระหว่างค่าเฉลี่ยของค่าเอฟเมสเซอร์ของหน่วยทดลอง 2 กลุ่ม ถ้าค่า Sig. (Significance) ที่คำนวณได้น้อยกว่า 0.05 และค่าสถิติทีมากกว่า 0 จึงจะสามารถปฏิเสธ H_0 ได้ แต่ถ้าผลการแจกแจงประชากรพบว่ามีการแจกแจงไม่ปกติ ต้องใช้วิธีการทดสอบสมมติฐานแบบไม่อิงกับพารามิเตอร์ (Non Parametric Test) ต่อไป โดยในที่นี้คือการวิเคราะห์โดยสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) เพื่อทดสอบสมมติฐานของผลต่างระหว่างค่าเฉลี่ยของค่าเอฟเมสเซอร์ของหน่วยทดลอง 2 กลุ่ม ถ้าค่า Sig. (Significance) ที่คำนวณได้น้อยกว่า 0.05 และค่าสถิติซีมากกว่า 0 ในกรณีที่ผลการวิเคราะห์ตั้งอยู่บนพื้นฐานทางบวก (Based on positive ranks) จึงจะสามารถปฏิเสธ H_0 ได้

สำหรับการทดสอบในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วนั้นใช้ค่าผลสะท้อนกลับ เป็นค่าที่แสดงถึงประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ เมื่อการทดสอบประสิทธิภาพเสร็จสิ้นแล้ว ทำให้ได้ค่าผลสะท้อนกลับออกมาหนึ่งค่าต่อหนึ่งการทดสอบ ค่าผลสะท้อนกลับที่ได้มานั้นแสดงให้เห็นถึงร้อยละของการเกิดการแจ้งเตือนที่ผิด (False Alarm) หรือการให้คำแนะนำที่เป็นผลบวกลวง (False Positive) นั่นเอง ดังนั้นการวัดประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วนั้น ค่าผลสะท้อนกลับที่น้อยกว่าจะมีความหมายว่ามีประสิทธิภาพมากกว่า นั่นคือถ้าค่าผลสะท้อนกลับของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 มากกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 แล้วจึงสามารถปฏิเสธ H_0 ได้

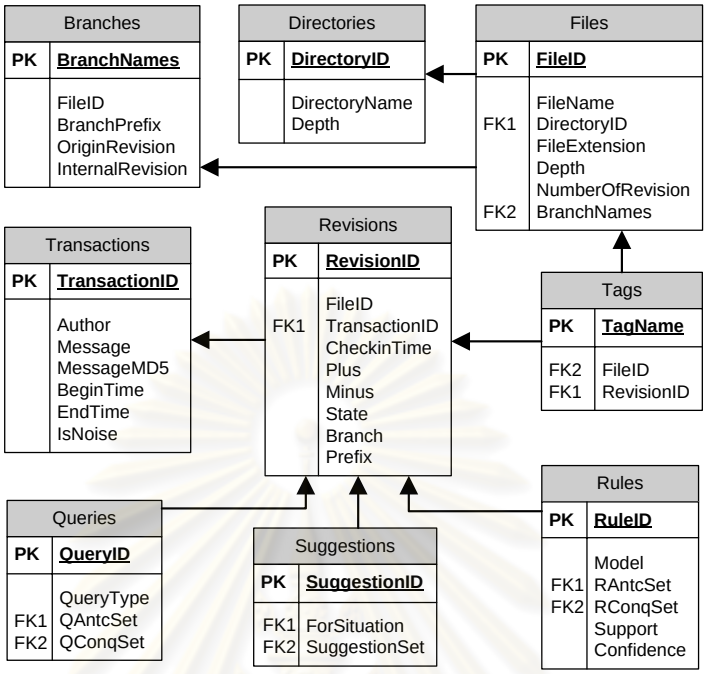
3.7 เครื่องมือที่ใช้ในงานวิจัย

ตามที่ได้กล่าวมาในหัวข้อแนวทางการทำวิจัยแล้วว่าผู้วิจัยได้พัฒนาเครื่องมือทดสอบการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์วทั้ง 2 ตัวแบบใน 3 สถานการณ์ ผู้วิจัยเลือกใช้โปรแกรมประยุกต์อีโรส (eROSE) (Zimmermann et al., 2005) ร่วมกับสคริปท์ (Srcript) ที่ผู้วิจัยสร้างขึ้นมาเองด้วยภาษาพีเอชพี (PHP) และระบบจัดการฐานข้อมูลชื่อพีเอชพีมายแอดมิน (PHPMyAdmin Database Management System) ซึ่งเป็นฐานข้อมูลแบบเปิดสามารถนำมาใช้งานได้โดยไม่เสียค่าใช้จ่ายและเข้ากันได้ดีกับภาษาพีเอชพี (PHP) ผู้วิจัยออกแบบเครื่องมือทดสอบตามขั้นตอนในหัวข้อ 3.6 และแผนภาพในรูป 3-1 ภาพรวมของเครื่องมือการทดสอบทั้งหมดแบ่งออกเป็น 5 ส่วนดังนี้



รูปที่ 3-5 แสดงภาพรวมของเครื่องมือที่ใช้ในการทดสอบทดสอบการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์

การออกแบบฐานข้อมูลที่ใช้ในงานวิจัยนี้ ออกแบบตามฐานข้อมูลของโปรแกรมประยุกต์อีโรส (eROSE) (Zimmermann et al., 2005) เนื่องจากผู้วิจัยใช้บางส่วนของโปรแกรมประยุกต์อีโรสมาใช้ในขั้นตอนแรกของการวิจัย ฐานข้อมูลประกอบด้วยตารางทั้งหมด 9 ตาราง คือ ตารางชื่อ Directories ตารางชื่อ Files ตารางชื่อ Revisions ตารางชื่อ Transactions ตารางชื่อ Branches ตารางชื่อ Tags ตารางชื่อ Queries ตารางชื่อ Rules และตารางชื่อ Suggestions ตัวอย่างระเบียนของแต่ละตารางแสดงในหัวข้อ 3.6 ความสัมพันธ์ของแต่ละตารางแสดงดังแผนภาพต่อไปนี้



รูปที่ 3-6 แสดงแผนภาพอีอาร์ (ER Diagram) ฐานข้อมูลของเครื่องมือที่ใช้ในการทดสอบ

รายละเอียดของเครื่องมือทั้ง 5 เครื่องมือ แสดงดังต่อไปนี้

- **ส่วนของการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์**

ขั้นตอนแรกของการทดสอบคือการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ (Preparing Data for Mining in Software Archives) ของโครงการพัฒนาซอฟต์แวร์ชื่อเคมายมันนี่ (KMyMoney) ที่พัฒนาด้วยภาษาซีพลัสพลัส (C++) ในส่วนนี้ผู้วิจัยเลือกใช้ส่วนการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ซึ่งเป็นส่วนหนึ่งโปรแกรมประยุกต์อีโรส (eROSE) (Zimmermann et al., 2005) เนื่องจากโปรแกรมประยุกต์อีโรสสามารถรับรองการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลของโครงการพัฒนาซอฟต์แวร์ที่พัฒนาด้วยภาษาซีพลัสพลัส (C++) และยังเป็นเครื่องมือที่ถูกนำไปใช้งานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ข้อมูลซอฟต์แวร์อาร์ไคฟว์เช่น งานวิจัยของ Zimmermann และคณะในปี 2005 (Zimmermann et al., 2005) และในงานวิจัยของ Methanias และคณะในปี 2009 (Methanias et al., 2009)

ขั้นตอนวิธีสำหรับส่วนการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟ์ของโปรแกรมประยุกต์อีโรสประกอบด้วย 4 ขั้นตอนคือ 1) การสกัดข้อมูล (Data Extraction) 2) การซ่อมแซมทรานแซคชั่น (Restoring Transactions) 3) การระบุการเปลี่ยนแปลงแก้ไขในระดับเอนทิตี้ (Mapping Changes to Entities) และ 4) การกำจัดสิ่งแปลกปลอม (Data Cleaning) อธิบายไว้อย่างละเอียดหัวข้อที่ 3.6.1 เมื่อได้ทรานแซคชั่นที่สมบูรณ์แล้วเก็บลงฐานข้อมูลของงานวิจัยเพื่อจัดเตรียมข้อมูลไว้ก่อนจะนำข้อมูลเหล่านี้ไปทำการทดสอบในส่วนการสร้างข้อสอบถามสำหรับการทดสอบสำหรับ 3 สถานการณ์ต่อไป

ข้อมูลเข้า คือ ข้อมูลซอฟต์แวร์อาร์ไคฟ์ซึ่งประกอบด้วย แฟ้มข้อมูลซอร์สโค้ดทั้งโครงการทุกเวอร์ชั่น และแฟ้มข้อมูลบันทึกของระบบคอนเคอเรนท์เวอร์ชั่น (CVS Log File) ตัวอย่างของข้อมูลเข้าแสดงดังรูป 3-2 ภายใต้หัวข้อ 3.6.1

ข้อมูลออก คือ ทรานแซคชั่นของการเปลี่ยนแปลงแก้ไข ที่ประกอบด้วยตาราง 6 ตาราง คือตารางชื่อ Files ตารางชื่อ Directories ตารางชื่อ Tags ตารางชื่อ Branches ตารางชื่อ Revisions และตารางชื่อ Transactions ตัวอย่างของข้อมูลออกแสดงในหัวข้อ 3.6.1

- **ส่วนของการสร้างข้อสอบถามสำหรับการทดสอบสำหรับ 3 สถานการณ์**

ส่วนที่ 2 คือส่วนของการสร้างข้อสอบถามสำหรับการทดสอบสำหรับแต่ละสถานการณ์ทั้ง 3 สถานการณ์ การทำงานของส่วนนี้ประกอบด้วยส่วนย่อย 2 ส่วนคือ 1) ส่วนการเลือกทรานแซคชั่นชุดทดสอบ และ 2) ส่วนการสร้างข้อสอบถามแต่ละสถานการณ์

ในส่วนย่อยที่ 1 ผู้วิจัยต้องวิเคราะห์สถิติเชิงพรรณาและสุ่มเลือกทรานแซคชั่นชุดทดสอบโดยใช้โปรแกรมตารางคำนวณไมโครซอฟต์เอ็กซ์เซล (Microsoft Excel) ทั้งหมดไม่ได้พัฒนาเครื่องมือสำหรับส่วนนี้ ขั้นตอนวิธีในการวิเคราะห์และเลือกทรานแซคชั่นชุดทดสอบจากโครงการพัฒนาซอฟต์แวร์ชื่อเคมายมันนี่ (KMyMoney) ที่อธิบายไว้ในหัวข้อ 3.6.2 และภาคผนวก ก

ในส่วนย่อยที่ 2 เป็นส่วนที่ผู้วิจัยพัฒนาเครื่องมือขึ้นมาเองและเรียกว่าเครื่องมือนี้ว่าเครื่องมือสร้างข้อสอบถามสำหรับ 3 สถานการณ์ โดยพัฒนาขึ้นมาในลักษณะของสคริปท์ด้วยภาษาพีเอชพี (PHP) ผู้วิจัยทำการทดสอบความถูกต้องของเครื่องมือนี้โดยการตรวจสอบแบบเดินผ่าน (Walkthrough) กับผลลัพธ์ทั้งหมดของเครื่องมือนี้ นั่นก็คือข้อสอบถามสำหรับ 3 สถานการณ์รายละเอียดแสดงในภาคผนวก ข

ขั้นตอนวิธีในส่วนย่อยที่ 2 การสร้างข้อสอบถามที่อธิบายไว้อย่างละเอียดในหัวข้อ 3.6.2

ข้อมูลเข้า คือ ทรานแซคชั่นของการเปลี่ยนแปลงแก้ไข (เฉพาะตารางชื่อ Revisions และตารางชื่อ Transactions) ตัวอย่างของข้อมูลเข้าแสดงในหัวข้อ 3.6.1

ข้อมูลออก คือ ข้อสอบถามสำหรับสถานการณ์นำทาง ข้อสอบถามสำหรับสถานการณ์ป้องกันการเกิดข้อผิดพลาด และข้อสอบถามสำหรับสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (ตารางชื่อ Queries) ตัวอย่างของข้อมูลออกแสดงในหัวข้อ 3.6.2

- **ส่วนของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์**

ส่วนที่ 3 คือส่วนการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ทั้ง 2 ตัวแบบ นั่นคือ การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 และการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 เครื่องมือที่ผู้วิจัยพัฒนาขึ้นมาสำหรับส่วนนี้เรียกว่าเครื่องมือค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 และเครื่องมือค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 โดยพัฒนาเครื่องมือทั้ง 2 นี้ในลักษณะของสคริปท์ด้วยภาษาพีเอชพี (PHP)

ขั้นตอนวิธีการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 และการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ใช้ขั้นตอนวิธีที่ชื่อว่า วิธีอพริโอริ (Apriori algorithm) อธิบายในหัวข้อ 2.8.2 และกำหนดค่าองค์ประกอบต่างๆตามที่ระบุไว้ในหัวข้อ 3.6.3 เช่นเดียวกับงานวิจัยของ Zimmermann และคณะในปี ค.ศ. 2004 และ 2005 (Zimmermann et al., 2004; Zimmermann et al., 2005) และงานวิจัยของ Michail ในปี ค.ศ. 2000 (Michail, 2000) ผู้วิจัยทำการทดสอบความถูกต้องของเครื่องมือนี้โดยการสุ่มตรวจผลลัพธ์หรือกฎความสัมพันธ์ที่ได้มาจากเครื่องมือทั้ง 2 เครื่องมือรายละเอียดแสดงในภาคผนวก ข

ข้อมูลเข้าของเครื่องมือค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 และเครื่องมือค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 คือ ทรานแซคชั่นทั้งหมดในฐานข้อมูล (เฉพาะเฉพาะตารางชื่อ Revisions และตารางชื่อ Transactions ทั้งหมด) ตัวอย่างของข้อมูลเข้าแสดงในหัวข้อ 3.6.2

ข้อมูลออกของเครื่องมือค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 คือ กฎความสัมพันธ์ด้วยตัวแบบที่ 1 รวมถึงค่าสนับสนุนค่าความเชื่อมั่นของกฎความสัมพันธ์ (ตารางชื่อ Rules) ข้อมูลออกของเครื่องมือค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 คือ กฎความสัมพันธ์ด้วยตัวแบบที่ 2 รวมถึง

ค่าสนับสนุนค่าความเชื่อมั่นของกฎความสัมพันธ์ (ตารางชื่อ Rules) ตัวอย่างของข้อมูลออกแสดงในหัวข้อ 3.6.3

- **ส่วนของการสร้างเซตของคำแนะนำสำหรับเหตุการณ์ในข้อสอบถาม**

ส่วนที่ 4 ของการทดสอบคือส่วนของการสร้างเซตของคำแนะนำสำหรับเหตุการณ์ในข้อสอบถามชุดทดสอบ ในส่วนนี้ผู้วิจัยพัฒนาเครื่องมือขึ้นมาเองและให้ชื่อว่าเครื่องมือสร้างคำแนะนำสำหรับเหตุการณ์ ผู้วิจัยพัฒนาขึ้นมาในลักษณะของสคริปท์ด้วยภาษาพีเอชพี (PHP)

ขั้นตอนวิธีการสร้างเซตของคำแนะนำสำหรับเหตุการณ์ในข้อสอบถามอธิบายไว้ในหัวข้อ 3.6.4 ซึ่งเป็นขั้นตอนวิธีเดียวกันกับที่ใช้ในงานวิจัยของ Zimmermann และคณะในปี 2005 (Zimmermann et al., 2005) และในงานวิจัยของ Methanias และคณะในปี 2009 (Methanias et al., 2009) ผู้วิจัยทำการทดสอบความถูกต้องของเครื่องมือนี้โดยการสุ่มตรวจผลลัพธ์หรือเซตของคำแนะนำสำหรับเหตุการณ์ที่ได้มาจากการใช้เครื่องมือนี้ รายละเอียดแสดงในภาคผนวก ข

ข้อมูลเข้า คือ เซตเหตุการณ์ในข้อสอบถาม (ตารางชื่อ Queries) และกฎความสัมพันธ์ด้วยตัวแบบที่ 1 และกฎความสัมพันธ์ด้วยตัวแบบที่ 2 (ตารางชื่อ Rules) ตัวอย่างของข้อมูลเข้าแสดงในหัวข้อ 3.6.2 และหัวข้อ 3.6.3 ตามลำดับ

ข้อมูลออก คือ เซตของคำแนะนำของการทดสอบทั้ง 6 การทดสอบ (ตารางชื่อ Suggestions) ตัวอย่างของข้อมูลออกแสดงในหัวข้อ 3.6.4

- **ส่วนของการประเมินผลการทดสอบ**

ส่วนสุดท้ายของการทดสอบคือส่วนของการประเมินผลการทดสอบเป็นส่วนที่นำเซตของคำแนะนำสำหรับเหตุการณ์ที่ได้มาจากการทดสอบทั้ง 6 การทดสอบมาคำนวณหาค่าประสิทธิภาพของการทำเหมืองข้อมูล ในส่วนนี้ผู้วิจัยพัฒนาเครื่องมือขึ้นมาเองและให้ชื่อว่าเครื่องมือประเมินผลการทดสอบ โดยการทดสอบที่ทดสอบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้นจะคำนวณค่าเอฟเมสเซอร์ ส่วนการทดสอบที่ทดสอบในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วจะคำนวณค่าสะท้อนกลับ เช่นเดียวกับงานวิจัยของ Zimmermann และคณะในปี 2005 (Zimmermann et al., 2005) และในงานวิจัยของ Methanias และคณะในปี 2009 (Methanias et al., 2009) ผู้วิจัยพัฒนาขึ้นมาในลักษณะของสคริปท์ด้วยภาษาพีเอชพี (PHP) ตามขั้นตอนวิธีในการประเมินผลการทดสอบแต่ละ

การทดสอบอธิบายอย่างละเอียดในหัวข้อ 3.6.5 ผู้วิจัยทำการทดสอบความถูกต้องของเครื่องมือนี้โดยการสุ่มตรวจผลลัพธ์หรือค่าประสิทธิภาพของแต่ละข้อสอบถามในแต่ละการทดสอบที่ได้มาจากการใช้เครื่องมือนี้ รายละเอียดแสดงในภาคผนวก ข

ข้อมูลเข้า คือ เซตของคำแนะนำสำหรับเหตุการณ์ของการทดสอบทั้ง 6 การทดสอบ (ตารางชื่อ Suggestions) และข้อสอบถามของแต่ละสถานการณ์

ข้อมูลออก คือ ค่าประสิทธิภาพของการทดสอบทั้ง 6 การทดสอบที่อยู่ในรูปแบบ (format) ของแฟ้มข้อมูลตัวอักษร (Text file)

ผู้วิจัยจะนำข้อมูลออกที่ได้จากเครื่องมือประเมินผลการทดสอบไปเข้าสู่ขั้นตอนการทดสอบสมมติฐานซึ่งเป็นขั้นตอนสุดท้ายของการวิจัยนี้ ผู้วิจัยต้องใช้การวิเคราะห์และเลือกสถิติทดสอบที่เหมาะสมตามข้อกำหนดที่อธิบายในหัวข้อ 3.6.6 และนำไปวิเคราะห์ด้วยโปรแกรมสถิติเอสพีเอสเอสต่อไป

3.8 ความถูกต้อง (Validity) และค่าความน่าเชื่อถือ (Reliability) ของข้อมูลที่เก็บ

การตอบวัตถุประสงค์ของข้อมูลงานวิจัยให้เชื่อถือได้ (Reliability) และถูกต้อง (Validity) จำเป็นต้องควบคุมปัจจัยที่เกี่ยวข้องอันได้แก่ การเลือกโครงการพัฒนาซอฟต์แวร์ การสร้างข้อสอบถามและการทดสอบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์

เนื่องจากงานวิจัยนี้มีวัตถุประสงค์ในการทดลองเพื่อศึกษาผลกระทบจากตัวแปรต้น คือ การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบต่างๆ กัน ซึ่งตัวแปรต้นนี้เป็นปัจจัยที่ต้องเปลี่ยนค่าไปตามแบบแผนการทดลองเพื่อดูความแตกต่างอันเกิดขึ้นจากการทดลอง นอกจากนั้นยังต้องสามารถควบคุมปัจจัยในด้านต่างๆ ให้มีความเหมือนกันหรือมีความคงที่ภายใต้สภาวะเดียวกัน เพื่อผลการทดลองที่สะท้อนเป็นค่าของตัวแปรต้นของทั้งกลุ่มควบคุมและกลุ่มทดสอบ นั่นคือ การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 และการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 เท่านั้น โดยในการทดลองมีปัจจัยที่ต้องควบคุมดังนี้

1) การเลือกโครงการพัฒนาซอฟต์แวร์ที่นำมาใช้ในการทดสอบ ผู้วิจัยกำหนดให้ข้อมูลซอฟต์แวร์อาร์ไคฟว์และข้อสอบถามที่จะทดสอบนั้นเป็นข้อมูลที่มาจากโครงการพัฒนาซอฟต์แวร์โครงการเดียวกันเพื่อให้ค่าประสิทธิภาพที่วัดออกมานั้นเป็นประสิทธิภาพที่เกิดมาจากตัวแบบของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ที่แตกต่างกันอย่างแท้จริง
2) โครงการพัฒนาซอฟต์แวร์ที่นำมาใช้ในงานวิจัยนี้เป็นโครงการพัฒนาซอฟต์แวร์ชื่อเคมายมันนี่ (KMyMoney) ซึ่งเป็นซอฟต์แวร์ทางการบัญชี ซึ่งเริ่มต้นการเผยแพร่โครงการตั้งแต่ปี ค.ศ. 2000 มีทรานแซคชั่นของการเปลี่ยนแปลงแก้ไขกว่า 28261 ทรานแซคชั่น ทำให้ข้อมูลซอฟต์แวร์อาร์ไคฟว์ของโครงการนี้สามารถเป็นตัวแทนของข้อมูลซอฟต์แวร์อาร์ไคฟว์ของโครงการพัฒนาซอฟต์แวร์ที่มีความหลากหลายได้
3) การกำหนดเซตเหตุการณ์และเซตผลลัพธ์ที่คาดไว้ของข้อสอบถามทั้งหมดถูกกำหนดมาจากข้อมูลซอฟต์แวร์อาร์ไคฟว์ที่นำมาทดสอบเอง จึงสามารถแน่ใจได้ว่าผลลัพธ์ที่จะได้ออกมานั้นมาจากเหตุการณ์ที่เคยเกิดขึ้นมาแล้วจริงๆในอดีต ซึ่งทำให้กระบวนการพิจารณาผลลัพธ์ที่ระบบแสดงออกมาในขั้นตอนการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบทั้ง 2 ตัวแบบในงานวิจัยนี้จะสามารถเชื่อถือความถูกต้องของผลลัพธ์ซึ่งเป็นเซตของคำแนะนำสำหรับเหตุการณ์นั้นๆได้
4) การทดสอบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ในงานวิจัยนี้จะใช้ค่าเอฟเมสเซอร์ (F-measure) ที่เป็นการคำนวณร่วมกันระหว่างค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) เป็นมาตรวัดที่นิยมใช้ในการทดลองในงานวิจัยด้านการค้นคืนสารสนเทศ (Baeza-Yates and Riberio-Neto, 1999) โดยจะเป็นการวัดว่าระบบสามารถให้ผลลัพธ์ของการค้นคืนออกมาได้ถูกต้องหรือไม่
5) เครื่องมือที่ใช้ในการทดสอบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบเป็นเครื่องมือเดียวกันทั้งหมด ต่างกันเพียงเครื่องในการค้นหากฎความสัมพันธ์ที่ใช้ตัวแบบในการระบุความน่าสนใจของกฎความสัมพันธ์ (ตัวแปรต้น) เท่านั้น
6) เครื่องมือที่ใช้ในการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์คือเครื่องมือที่เป็นส่วนหนึ่งโปรแกรมประยุกต์อีโรส (eROSE) (Zimmermann et

al., 2005) ที่ได้รับการยอมรับและถูกนำไปใช้งานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ข้อมูลซอฟต์แวร์อาร์ไคฟ์เช่น งานวิจัยของ Zimmermann และคณะในปี 2005 (Zimmermann et al., 2005) และในงานวิจัยของ Methanias และคณะในปี 2009 (Methanias et al., 2009)

7) เครื่องมือที่ผู้วิจัยพัฒนาขึ้นมาเองได้แก่ เครื่องมือสร้างข้อสอบถามสำหรับ 3 สถานการณ์ เครื่องมือค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 และตัวแบบที่ 2 เครื่องมือสร้างคำแนะนำสำหรับเหตุการณ์ และเครื่องมือประเมินผลการทดสอบ ผู้วิจัยได้ทำการทดสอบความถูกต้องของเครื่องมือทั้งด้วยวิธีการสุ่มตรวจความถูกต้องของผลลัพธ์ที่เป็นตัวแทนของผลลัพธ์ทั้งหมด

3.9 **กรอบการวิเคราะห์ข้อมูล** (Data Analysis Framework)

สำหรับกรณีของการทดสอบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้นใช้ค่าเอฟเมสเซอร์เป็นค่าที่แสดงถึงประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ เมื่อการทดสอบประสิทธิภาพเสร็จสิ้นแล้ว ทำให้ได้ค่าเอฟเมสเซอร์ออกมาเท่าจำนวนของข้อสอบถามชุดทดสอบที่สร้างขึ้นในขั้นตอนการสร้างข้อสอบถาม จากนั้นในขั้นตอนแรกจะตรวจสอบการแจกแจงของค่าประสิทธิภาพที่ได้มาว่ามีการแจกแจงปกติหรือไม่ ด้วยการใช้สถิติทดสอบ Kolmogorov-Smirnov เพื่อเลือกทางเลือกในการทดสอบสมมติฐานได้ว่าจะให้การทดสอบสมมติฐานแบบใช้พารามิเตอร์ (Parametric Test) หรือแบบไม่อิงกับพารามิเตอร์ (Non Parametric Test) ถ้าผลการทดสอบพบว่าประชากรมีการแจกแจงแบบปกติ จึงใช้การวิเคราะห์โดยสถิติทดสอบที (t-test) เพื่อทดสอบสมมติฐานของผลต่างระหว่างค่าเฉลี่ยของค่าเอฟเมสเซอร์ของหน่วยทดลอง 2 กลุ่ม ถ้าค่า Sig. (Significance) ที่คำนวณได้น้อยกว่า 0.05 และค่าสถิติทีมากกว่า 0 จึงจะสามารถปฏิเสธ H_0 ได้ แต่ถ้าผลการแจกแจงประชากรพบว่ามีการแจกแจงไม่ปกติ ต้องใช้วิธีการทดสอบสมมติฐานแบบไม่อิงกับพารามิเตอร์ (Non Parametric Test) ต่อไป โดยในที่นี้คือการวิเคราะห์โดยสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) เพื่อทดสอบสมมติฐานของผลต่างระหว่างค่าเฉลี่ยของค่าเอฟเมสเซอร์ของหน่วยทดลอง 2 กลุ่ม ถ้าค่า Sig. (Significance) ที่คำนวณได้น้อย

กว่า 0.05 และค่าสถิติซีมากกว่า 0 ในกรณีที่ผลการวิเคราะห์ตั้งอยู่บนพื้นฐานทางบวก (Based on positive ranks) จึงจะสามารถปฏิเสธ H_0 ได้

สำหรับการทดสอบในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วนั้นใช้ค่าผลสะท้อนกลับ เป็นค่าที่แสดงถึงประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ เมื่อการทดสอบประสิทธิภาพเสร็จสิ้นแล้ว ทำให้ได้ค่าผลสะท้อนกลับออกมาหนึ่งค่าต่อหนึ่งการทดสอบ ค่าผลสะท้อนกลับที่ได้มานั้นแสดงให้เห็นถึงร้อยละของการเกิดการแจ้งเตือนที่ผิด (False Alarm) หรือการให้คำแนะนำที่เป็นผลบวกลวง (False Positive) นั่นเอง ดังนั้นการวัดประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วนั้น ค่าผลสะท้อนกลับที่น้อยกว่าจะมีความหมายว่ามีประสิทธิภาพมากกว่า นั่นคือถ้าค่าผลสะท้อนกลับของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 มากกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 แล้วจึงสามารถปฏิเสธ H_0 ได้

บทที่ 4
ผลการวิเคราะห์ข้อมูล

4.1 บทนำ

ในบทนี้จะแสดงผลและวิเคราะห์เปรียบเทียบทดสอบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบทั้ง 2 ตัวแบบ เพื่อนำมาตอบวัตถุประสงค์ของงานวิจัยที่กล่าวไปในบทที่ 3 ซึ่งได้แก่ 1) เปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบในสถานการณ์การนำทาง 2) เปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบในสถานการณ์การป้องกันการเกิดข้อผิดพลาด และ 3) เปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว และในส่วนท้ายของบทนี้เป็นการศึกษาและวิเคราะห์ข้อมูลเพิ่มเติม

4.2 ผลการทดลอง

การทดลองนี้มีวัตถุประสงค์เพื่อวัดประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ของตัวแบบ 2 ตัวแบบในสถานการณ์ของการพัฒนาซอฟต์แวร์ที่ต่างกัน 3 สถานการณ์ โดยใช้ค่าเอฟเมสเซอร์เป็นค่าประเมินประสิทธิภาพของการทำเหมืองข้อมูลบนข้อมูลซอฟต์แวร์อาร์ไคฟว์ในสถานการณ์การนำทาง และสถานการณ์การป้องกันข้อผิดพลาด และใช้ค่าผลสะท้อนกลับเป็นค่าประเมินประสิทธิภาพของการทำเหมืองข้อมูลบนข้อมูลซอฟต์แวร์อาร์ไคฟว์ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

ทรานแซคชั่นชุดทดสอบจำนวน 60 ทรานแซคชั่นที่เป็นตัวแทนของทรานแซคชั่นการเปลี่ยนแปลงแก้ไขจากโครงการพัฒนาซอฟต์แวร์ทางการบัญชีชื่อเคมายมันนี่ (KMyMoney) ที่เลือกมาตามขั้นตอนย่อยภายในขั้นตอนการสร้างข้อสอบถามหัวข้อที่ 3.6.2 และภาคผนวก ก สามารถนำมาสร้างเป็นข้อสอบถามได้ทั้งหมด 962 ข้อสอบถาม ประกอบด้วย ข้อสอบถามสำหรับทดสอบในสถานการณ์การนำทางทั้งหมด 451 ข้อสอบถาม ข้อสอบถามสำหรับทดสอบใน

สถานการณ์การป้องกันการเกิดข้อผิดพลาดทั้งหมด 451 ข้อสอบถาม และข้อสอบถามสำหรับทดสอบในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วทั้งหมด 60 ข้อสอบถาม ค่าเอฟเมสเซอร์ที่ได้จากการทดสอบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบสำหรับสถานการณ์การนำทางทั้งหมด 451 ค่าและสำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาดทั้งหมด 451 ค่าแสดงในตารางที่ ข-1 และตารางที่ ข-2 ตามลำดับ ค่าผลสะท้อนกลับที่ได้จากการทดสอบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบสำหรับสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วทั้งหมด 60 ค่าแสดงในตารางที่ ข-3

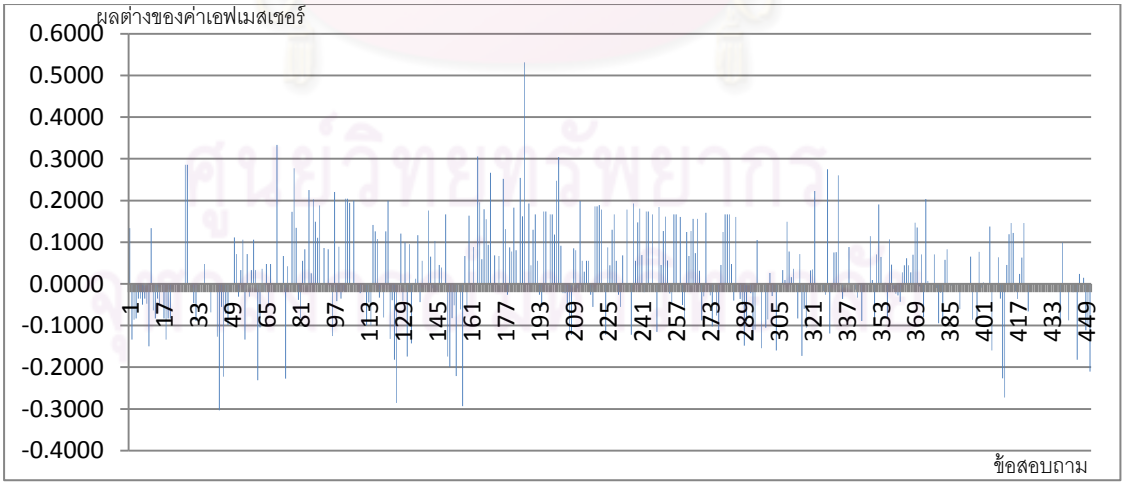
จากการทดสอบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบทั้ง 2 ตัวแบบคือ การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 และ การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 นั้น ในสถานการณ์ของการพัฒนาซอฟต์แวร์ที่ต่างกัน 3 สถานการณ์คือ สถานการณ์การนำทาง สถานการณ์การป้องกันการเกิดข้อผิดพลาด และสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว สามารถสรุปได้ดังนี้

- การทดสอบสถานการณ์การนำทาง
  - ข้อสอบถามที่ได้เซตของคำแนะนำที่ให้ค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มากกว่าค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 มีจำนวน 192 ข้อสอบถาม
  - ข้อสอบถามที่ได้เซตของคำแนะนำที่ให้ค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 เท่ากับค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 มีจำนวน 111 ข้อสอบถาม
  - ข้อสอบถามที่ได้เซตของคำแนะนำที่ให้ค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 น้อยกว่าค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 มีจำนวน 148 ข้อสอบถาม
- การทดสอบสถานการณ์การป้องกันการเกิดข้อผิดพลาด
  - ข้อสอบถามที่ได้เซตของคำแนะนำที่ให้ค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มากกว่าค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 มีจำนวน 46 ข้อสอบถาม

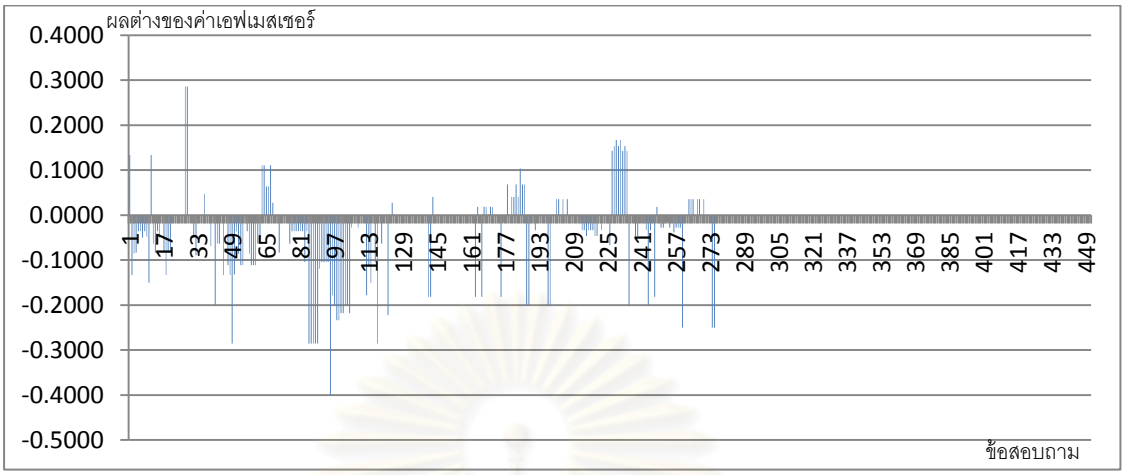
- ข้อสอบถามที่ได้เซตของคำแนะนำที่ให้ค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 เท่ากับค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 มีจำนวน 279 ข้อสอบถาม
- ข้อสอบถามที่ได้เซตของคำแนะนำที่ให้ค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 น้อยกว่าค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 มีจำนวน 126 ข้อสอบถาม

- การทดสอบสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว
  - ข้อสอบถามที่ได้เซตของคำแนะนำเป็นเซตว่าง มีจำนวน 19 ข้อสอบถาม
  - ข้อสอบถามที่ได้เซตของคำแนะนำไม่เป็นเซตว่าง มีจำนวน 41 ข้อสอบถาม

จากผลการทดสอบค่าเอฟเมสเซอร์ของสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด ในตารางที่ ข-1 และ ข-2 ตามลำดับ ผู้วิจัยสามารถแสดงกราฟเพื่อเปรียบเทียบประสิทธิภาพของการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบทั้ง 2 ตัวแบบโดยใช้กราฟแท่ง (Column Chart) ที่แสดงผลต่างค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบทั้ง 2 ตัวแบบของสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดดังรูปที่ 4.1 และ 4.2 ตามลำดับ และตารางที่ 4-1 แสดงตารางสรุปผลการทดสอบของทั้ง 3 สถานการณ์



รูปที่ 4-1 แสดงกราฟผลต่างค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 กับการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การนำทาง



รูปที่ 4-2 แสดงกราฟผลต่างค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 กับการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

ตารางที่ 4-1 แสดงตารางผลการทดสอบทั้ง 3 สถานการณ์

สถานการณ์การนำทาง (ค่าเฉลี่ยของค่าเอฟเมสเซอร์)		สถานการณ์การป้องกันการเกิดข้อผิดพลาด (ค่าเฉลี่ยของค่าเอฟเมสเซอร์)		สถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (ค่าผลสะท้อนกลับ)	
ตัวแบบที่ 1	ตัวแบบที่ 2	ตัวแบบที่ 1	ตัวแบบที่ 2	ตัวแบบที่ 1	ตัวแบบที่ 2
0.3013	0.3245	0.1353	0.1135	$\frac{41}{60}$	$\frac{41}{60}$

จากตารางสรุปข้างต้นแสดงให้เห็นว่าค่าประสิทธิภาพของสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้นมีค่าแตกต่างกัน ค่าประสิทธิภาพของสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วนั้นไม่มีความแตกต่างกัน แต่ไม่สามารถสรุปได้ว่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ในสถานการณ์ต่างๆนั้นแตกต่างกันอย่างมีนัยสำคัญ ผู้วิจัยจึงจำเป็นต้องวิเคราะห์ผลการทดสอบความแตกต่างกันอย่างมีนัยสำคัญ ดังรายละเอียดในหัวข้อต่อไป

4.3 ผลการวิเคราะห์ข้อมูล

สำหรับสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด การตรวจสอบเงื่อนไขพื้นฐานขั้นแรก ผู้วิจัยต้องตรวจสอบการแจกแจงของประชากรว่าการแจกแจง

ปกติหรือไม่ เพื่อเลือกทางเลือกในการทดสอบสมมุติฐานว่าจะใช้วิธีการทดสอบสมมุติฐานแบบอิงพารามิเตอร์ (Parametric Test) หรือแบบไม่อิงพารามิเตอร์ (Non Parametric Test) ถ้าผลการทดสอบพบว่าประชากรมีการแจกแจงแบบปกติ จึงใช้การวิเคราะห์โดยสถิติทดสอบที (t-test) เพื่อทดสอบสมมติฐานของผลต่างระหว่างค่าเฉลี่ยของค่าเอฟเมสเซอร์ของหน่วยทดลอง 2 กลุ่ม ถ้าค่า Sig. (Significance) ที่คำนวณได้น้อยกว่า 0.05 และค่าสถิติทีมากกว่า 0 จึงจะสามารถปฏิเสธ H_0 ได้ แต่ถ้าผลการแจกแจงประชากรพบว่ามีการแจกแจงไม่ปกติ จึงใช้การวิเคราะห์โดยสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) เพื่อทดสอบสมมติฐานของผลต่างระหว่างค่าเฉลี่ยของค่าเอฟเมสเซอร์ของหน่วยทดลอง 2 กลุ่ม ถ้าค่า Sig. (Significance) ที่คำนวณได้น้อยกว่า 0.05 และค่าสถิติซีมากกว่า 0 ในกรณีที่ผลการวิเคราะห์ตั้งอยู่บนพื้นฐานทางบวก (Based on positive ranks) จึงจะสามารถปฏิเสธ H_0 ได้

4.3.1 การวิเคราะห์การแจกแจงข้อมูล

ในงานวิจัยนี้ผู้วิจัยสนใจตัวแปร คือประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ของตัวแบบ 2 ตัวแบบ ดังนั้นจึงตรวจสอบการแจกแจงของข้อมูลที่ได้จากหน่วยทดลอง นั่นคือค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ ซึ่งได้แก่ค่าเอฟเมสเซอร์ที่ได้มาจากผลการทดสอบ 451 ข้อสอบถาม สำหรับสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด ดังนั้นผู้วิจัยจะตรวจสอบว่าค่าเอฟเมสเซอร์ทั้งหมด 451 ค่าของทั้ง 2 สถานการณ์มีการแจกแจงแบบปกติหรือไม่ โดยตั้งสมมติฐานของการทดสอบค่าประสิทธิภาพแต่ละสถานการณ์มีการแจกแจงแบบปกติหรือไม่ภายใต้สมมติฐานทางสถิติ ดังนี้

1) ทดสอบการแจกแจงของข้อมูลค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ตัวแบบที่ 1 สถานการณ์การนำทาง

   H_0: ข้อมูลค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 สถานการณ์การนำทาง มีการแจกแจงแบบปกติ

   H_1: ข้อมูลค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 สถานการณ์การนำทาง ไม่แจกแจงแบบปกติ

2) ทดสอบการแจกแจงของข้อมูลค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 สถานการณ์การนำทาง

H_0: ข้อมูลค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 สถานการณ์การนำทาง มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 สถานการณ์การนำทาง ไม่แจกแจงแบบปกติ

3) ทดสอบการแจกแจงของข้อมูลค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ตัวแบบที่ 1 สถานการณ์การป้องกันการเกิดข้อผิดพลาด

H_0: ข้อมูลค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 สถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 สถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่แจกแจงแบบปกติ

4) ทดสอบการแจกแจงของข้อมูลค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 สถานการณ์การป้องกันการเกิดข้อผิดพลาด

H_0: ข้อมูลค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 สถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 สถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่แจกแจงแบบปกติ

ผู้วิจัยเลือกใช้สถิติทดสอบ Kolmogorov-Sminov เนื่องจากมีขนาดตัวอย่างมากกว่า 50 หน่วย โดยจะยอมรับสมมติฐาน H_0 เมื่อค่า Sig. มีค่ามากกว่าค่า α ซึ่งกำหนดให้เท่ากับ 0.05 ดังตารางต่อไปนี้

ตารางที่ 4-2 แสดงค่าสถิติทดสอบการแจกแจงปกติ (Normality Test) ของค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ทั้ง 2 ตัวแบบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด

สถานการณ์	ตัวแบบที่	Kolmogorov-Smirnov		
		Statistic	df	Sig.
สถานการณ์การนำทาง	1	0.070	451	0.000
	2	0.075	451	0.000
สถานการณ์การป้องกันการเกิดข้อผิดพลาด	1	0.313	451	0.000
	2	0.326	451	0.000

ผลการทดสอบในตารางที่ 4-2 ข้างต้นพบว่าค่า Sig. ของตัวแปรค่าประสิทธิภาพการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดเป็นดังนี้

1) สำหรับสถานการณ์การนำทาง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0
2) สำหรับสถานการณ์การนำทาง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มีค่า Sig. เท่ากับ 0. 000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0
3) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0
4) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

ดังนั้นสรุปได้ว่าการแจกแจงของตัวแปรค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบสำหรับ 2 สถานการณ์นั้นไม่เป็นแบบปกติ

4.3.2 การวิเคราะห์เปรียบเทียบประสิทธิภาพการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบ

จากการวิเคราะห์การแจกแจงข้อมูลข้างต้นพบว่าค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดไม่แจกแจงแบบปกติ ดังนั้นผู้วิจัยจึงเลือกใช้สถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) กับการทดสอบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด โดยที่จะปฏิเสธสมมติฐาน H_0 ได้เมื่อถ้าค่า Sig. (Significance) ที่คำนวณได้น้อยกว่า 0.05 และค่าสถิติ Z มากกว่า 0 โดยที่ผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) และผู้วิจัยเลือกใช้การทดสอบอัตราส่วน 2 กลุ่มด้วยสถิติทดสอบ Z (Two Proportion Z Tests) กับการทดสอบในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว โดยที่จะปฏิเสธสมมติฐาน H_0 ได้เมื่อถ้าค่า Sig. (Significance) ที่คำนวณได้น้อยกว่า 0.05

การวิเคราะห์เปรียบเทียบนี้เป็นการวิเคราะห์เปรียบเทียบประสิทธิภาพระหว่างการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 กับการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์ 3 สถานการณ์คือ สถานการณ์การนำทาง สถานการณ์การป้องกันการเกิดข้อผิดพลาด สถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว ซึ่งสามารถตั้งสมมุติฐานได้ดังนี้

1) วิเคราะห์เปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์วทั้ง 2 ตัวแบบในสถานการณ์การนำทาง ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การนำทาง

M_2 คือ ค่ามัธยฐานของค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การนำทาง

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

2) วิเคราะห์เปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

M_2 คือ ค่ามัธยฐานของค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

3) วิเคราะห์เปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

M_2 คือ ค่ามัธยฐานของค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

การทดสอบสมมติฐานข้อ 1 และ 2 ด้วยสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) แสดงดังตารางที่ 4-3 สำหรับสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว การทดสอบสมมติฐานข้อ 3 ด้วยสถิติทดสอบ Z (Two Proportion Z Tests) แสดงดังตารางที่ 4-4

ตารางที่ 4-3 แสดงค่าสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ของค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 เทียบกับตัวแบบที่ 1 ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด

	ค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 - ค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ตัวแบบที่1	
	สถานการณ์การนำทาง	สถานการณ์การป้องกันการเกิดข้อผิดพลาด
Z	-4.374[a]	-6.055[b]
Asymp. Sig. (2 tailed)	.000	.000

a. Based on negative ranks.

b. Based on positive ranks.

จากตารางที่ 4-3 การวิเคราะห์เปรียบเทียบประสิทธิภาพของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบในสถานการณ์การนำทาง ได้สถิติทดสอบค่า Z เท่ากับ -4.374 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 0.000 / 2 = 0.000 ซึ่งน้อยกว่าค่า α = 0.05 เนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางลบ (Based on negative ranks) ดังนั้นจึงสามารถปฏิเสธสมมติฐาน H_0 ได้ นั่นคือค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มากกว่าค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การนำทาง ที่ระดับนัยสำคัญ 0.05 และสำหรับการวิเคราะห์เปรียบเทียบประสิทธิภาพของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ได้สถิติทดสอบค่า Z เท่ากับ -6.055 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 0.000 / 2 = 0.000 ซึ่งน้อยกว่าค่า α = 0.05 เนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) ดังนั้นจึงไม่สามารถปฏิเสธสมมติฐาน H_0 ได้ นั่นคือค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ไม่ต่างกันหรือน้อยกว่าค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาดที่ระดับนัยสำคัญ 0.05

ตารางที่ 4-4 แสดงการทดสอบอัตราส่วน 2 กลุ่มด้วยสถิติทดสอบ Z ของค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 เทียบกับตัวแบบที่ 1 ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.000[a]	1	1.000		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.000	1	1.000		
Fisher's Exact Test				1.000	.578
N of Valid Cases	120				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 19.00.

จากตารางที่ 4.4 ค่าสถิติ Z เท่ากับรากที่สองของค่า Pearson Chi-Square นั่นค่า 0.000 และมีเครื่องหมายเดียวกับผลต่างของอัตราส่วนที่ 2 กับอัตราส่วนที่ 1 ($\frac{41}{60} - \frac{41}{60} = 0.000$) นั่นค่าเครื่องหมายบวก และจากการตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 1.000 / 2 = 0.500 ซึ่งมากกว่าค่า α = 0.05 ดังนั้นจึงไม่สามารถปฏิเสธสมมติฐาน H_0 ได้ นั่นคือค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ไม่ต่างกับค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วที่ระดับนัยสำคัญ 0.05

4.3.3 สรุปผลการวิเคราะห์ข้อมูล

จากการวิเคราะห์ผลการทดสอบประสิทธิภาพของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด โดยใช้ค่าเอฟเมสเซอร์ และสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว โดยใช้ค่าผลสะท้อนกลับ ผู้วิจัยสามารถสรุปได้ว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มีประสิทธิภาพมากกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การนำทาง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มีประสิทธิภาพไม่ต่างกันหรือน้อยกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด และสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มีประสิทธิภาพไม่ต่างกับการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1

จากผลสรุปดังกล่าวแสดงให้เห็นว่า การใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ในการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์บนข้อมูลซอฟต์แวร์อาร์ไคฟว์แสดงประสิทธิภาพที่ดีกว่าการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นเพียงแค่ในสถานการณ์การนำทางเท่านั้น

4.4 ผลการศึกษาเพิ่มเติม

จากผลทดสอบการเปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นและตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ข้างต้น ผู้วิจัยมีความต้องการทดสอบเพิ่มเติมเพื่ออธิบายสาเหตุที่การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่แสดงประสิทธิภาพที่ดีกว่าเฉพาะในสถานการณ์ของการนำทาง แต่ให้ประสิทธิภาพที่ไม่ต่างกันหรือน้อยกว่าในสถานการณ์การป้องกันการเกิดข้อผิดพลาดและสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว จากการดำเนินการทดสอบข้างต้นผู้วิจัยมีข้อสังเกตหลายประการ เช่น การค้นหากฎความสัมพันธ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่นั้นมักจะให้เซตของคำแนะนำที่ใหญ่กว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบสนับสนุน-ค่าความเชื่อมั่น เนื่องจากกฎความสัมพันธ์ 10 อันดับแรกที่ได้จากการค้นหากฎความสัมพันธ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นนั้นมักจะเป็นกฎความสัมพันธ์ที่มีเซตรายการที่ตามมาขนาด 1 รายการเท่านั้น ในขณะที่กฎความสัมพันธ์ 10 อันดับแรกที่ได้จากการค้นหากฎความสัมพันธ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่มักจะเป็นกฎความสัมพันธ์ที่มีเซตรายการที่ตามมาขนาดมากกว่า 1 รายการ เป็นต้น จากข้อสังเกตดังกล่าวผู้วิจัยคิดว่าเซตของคำแนะนำที่มีขนาดใหญ่อาจเป็นสาเหตุที่ทำให้ค่าความถูกต้อง (Precision) ของการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่มีค่าน้อยและส่งผลให้ค่าเอฟเมเชอร์มีค่าน้อยตามมาด้วย

เนื่องจากขั้นตอนวิธีในการทดสอบของงานวิจัยมีข้อแตกต่างจากขั้นตอนวิธีที่ Zimmermann และคณะกับ Methanias และคณะ (Zimmermann et al., 2005; Methanias et

al., 2009) ใช้อยู่ขั้นตอนหนึ่ง คือนักวิจัยทั้ง 2 คณะนั้นมีการปรับปรุงขั้นตอนวิธีอพริโอริ 2 ประการ เพื่อให้สามารถค้นหากฎความสัมพันธ์ที่รวดเร็วขึ้นตามที่อธิบายไว้ในบทที่ 2 คือ 1) การกำหนดให้ค้นหากฎความสัมพันธ์เฉพาะกฎที่มีเซตรายการที่มาก่อนที่ต้องการเท่านั้น และ 2) การกำหนดให้ทุกกฎความสัมพันธ์ที่ค้นหามีเซตรายการที่ตามมาเพียง 1 รายการเท่านั้น จากข้อกำหนดข้อที่ 2 นั้นทำให้เซตของคำแนะนำที่จะได้มานั้นจะมีขนาดน้อยกว่าหรือเท่ากับ 10 รายการเสมอ แต่ในงานวิจัยนี้ผู้วิจัยไม่ได้ปรับปรุงขั้นตอนวิธีอพริโอริด้วยข้อกำหนดดังกล่าว ขนาดของเซตของคำแนะนำจึงมีความหลากหลายแตกต่างกันออกไป โดยที่เซตของคำแนะนำที่ได้จากการค้นหากฎความสัมพันธ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นนั้นจะมีขนาดน้อยกว่าหรือเท่ากับ 10 รายการเสมอเนื่องจากกฎความสัมพันธ์ 10 อันดับแรกมักจะมีเซตรายการที่ตามมาขนาด 1 รายการหรือเซตรายการที่ตามมาขนาดมากกว่า 1 รายการแต่มีสมาชิกที่ซ้ำกับเซตรายการที่ตามมาของกฎความสัมพันธ์ที่อยู่อันดับที่สูงกว่า แต่สำหรับเซตของคำแนะนำที่ได้จากการค้นหากฎความสัมพันธ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่นั้นมักมีขนาดมากกว่า 10 เสมอ เนื่องจากเนื่องจากกฎความสัมพันธ์ 10 อันดับแรกมักจะมีเซตรายการที่ตามมาขนาดมากกว่า 1 รายการและสมาชิกของเซตรายการที่ตามมาของแต่ละกฎความสัมพันธ์ใน 10 อันดับแรกมักจะไม่ซ้ำกันด้วย

ด้วยเหตุนี้ผู้วิจัยจึงมีความต้องการที่จะทดสอบเพิ่มเติมเพื่ออธิบายสาเหตุของผลการทดสอบข้างต้น โดยที่การทดสอบเพิ่มเติมนี้ผู้วิจัยทำการปรับปรุงขั้นตอนวิธีในการสร้างเซตของคำแนะนำใหม่ จากคุณลักษณะของกฎความสัมพันธ์ที่อยู่ใน 10 อันดับของการค้นหากฎความสัมพันธ์ด้วยตัวแบบทั้ง 2 นั้น ผู้วิจัยเห็นว่าไม่สามารถนำเซตรายการที่ตามมาของกฎความสัมพันธ์ทั้ง 10 อันดับแรกมาสร้างเป็นเซตของคำแนะนำได้โดยตรง และผู้วิจัยก็เห็นว่าไม่สามารถกำหนดให้พิจารณาเฉพาะกฎความสัมพันธ์ที่ค้นหามีเซตรายการที่ตามมา 1 รายการเท่านั้นเช่นเดียวกับ Zimmermann และคณะกับ Methanias และคณะ (Zimmermann et al., 2005; Methanias et al., 2009) ด้วย ผู้วิจัยเห็นว่าการสร้างเซตของคำแนะนำที่ได้มาจากการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบเพื่อวัตถุประสงค์ในการทดสอบเปรียบเทียบควรจะต้องสร้างเซตของคำแนะนำมาจากการยูเนียน (Union) ของรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตรายการที่ตามมาของกฎความสัมพันธ์อันดับสูงสุดจำนวน 10 รายการแรก ตัวอย่างเช่น ตารางด้านล่างนี้เป็นตารางแสดงกฎความสัมพันธ์ 10 อันดับแรกที่ได้จากการค้นหากฎความสัมพันธ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่เรียงลำดับตามค่าความเชื่อมั่นใหม่ของกฎความสัมพันธ์ที่มีเซตรายการที่มาก่อนคือ {15}

ตารางที่ 4-5 แสดงตัวอย่างการสร้างเซตของคำแนะนำจากการยูเนียน (Union) ของรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตรายการที่ตามมาของกฎความสัมพันธ์อันดับสูงสุดจำนวน 10 รายการแรก

	กฎความสัมพันธ์	ค่าสนับสนุน	ค่าความเชื่อมั่นใหม่
1	15 -> 29, 189	25	0.52
2	15 -> 365, 520, 521	16	0.51
3	15 -> 798	42	0.48
4	15 -> 646, 798, 799	16	0.46
5	15 -> 179, 465, 466	15	0.45
6	15 -> 646, 798, 804	15	0.45
7	15 -> 29, 678	22	0.43
8	15 -> 798, 799	22	0.43
9	15 -> 189	39	0.43
10	15 -> 225	56	0.41
เซตของคำแนะนำ คือ {29, 179, 189, 365, 465, 520, 521, 646, 798, 799 }			

ในตารางข้างต้นแสดงตัวอย่างของกฎความสัมพันธ์ ค่าสนับสนุน และค่าความเชื่อมั่นใหม่ โดยในหลัก (column) ของกฎความสัมพันธ์นั้นใช้ตัวเลขต่างๆ แสดงแทนการเปลี่ยนแปลงแก้ไขต่างๆ เพื่อความสะดวกในการทำความเข้าใจ

ดังนั้นผู้วิจัยจึงมีความต้องการทำการทดสอบเพิ่มเติม โดยเริ่มจาก การเปรียบเทียบค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด ต่อด้วยการเปรียบเทียบค่าประสิทธิภาพ ค่าความถูกต้อง และค่าเรียกคืน ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบโดยสร้างเซตของคำแนะนำจากการยูเนียน (Union) ของรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตรายการที่ตามมาของกฎความสัมพันธ์อันดับสูงสุดจำนวน 10 รายการแรก ตามลำดับ เพื่อตอบข้อสังเกตที่ว่าค่าความ

ถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่มีค่าน้อยกว่าแต่ค่าเรียกคืนมีค่ามากกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น

4.4.1 การเปรียบเทียบค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด

ผลการทดสอบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบในการทดสอบข้างต้นนั้นเป็นการเปรียบประสิทธิภาพจากการใช้ค่าเอฟเมเซอร์ (F-measure) เท่านั้น ผู้วิจัยจึงมีความต้องการศึกษาเพิ่มเติมโดยการทดสอบเปรียบเทียบค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด ผลการทดสอบแสดงดังตารางต่อไปนี้

ตารางที่ 4-6 แสดงตารางค่าความถูกต้องและค่าเรียกคืนของการทดสอบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด

สถานการณ์การนำทาง				สถานการณ์การป้องกันการเกิดข้อผิดพลาด			
ตัวแบบที่ 1		ตัวแบบที่ 2		ตัวแบบที่ 1		ตัวแบบที่ 2	
P_μ	R_μ	P_μ	R_μ	P_μ	R_μ	P_μ	R_μ
0.4750	0.2955	0.4554	0.3530	0.4758	0.5188	0.4620	0.4922

จากตารางสรุปข้างต้นแสดงให้เห็นว่าค่าเฉลี่ยของค่าความถูกต้อง (Precision) หรือ P_μ และค่าเฉลี่ยของค่าเรียกคืน (Recall) หรือ R_μ ของการทดสอบข้างต้นนั้นมีค่าแตกต่างกันทั้ง 2 ตัวแบบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด แต่ไม่สามารถสรุปได้ว่าค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการค้นหากฎความสัมพันธ์ของตัวแบบทั้ง 2 ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้นแตกต่างกันอย่างมีนัยสำคัญ ผู้วิจัยจึงจำเป็นต้องวิเคราะห์ผลการทดสอบความแตกต่างกันอย่างมีนัยสำคัญ ดังรายละเอียดต่อไปนี้

การวิเคราะห์การแจกแจงปกติ

เนื่องจากผู้วิจัยต้องการเปรียบเทียบค่าความถูกต้อง (Precision) หรือค่าเรียกคืน (Recall) ของการค้นหากฎความสัมพันธ์ของ 2 ตัวแบบสำหรับสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด ดังนั้นผู้วิจัยจะตรวจสอบว่าค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ทั้งหมดของทั้ง 2 สถานการณ์มีการแจกแจงแบบปกติหรือไม่ โดยตั้งสมมติฐานของการทดสอบค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) แต่ละสถานการณ์มีการแจกแจงแบบปกติหรือไม่ภายใต้สมมติฐานทางสถิติ ดังนี้

1) ทดสอบการแจกแจงของข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 ในสถานการณ์การนำทาง

   H_0: ข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 ในสถานการณ์การนำทาง มีการแจกแจงแบบปกติ

   H_1: ข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 ในสถานการณ์การนำทาง ไม่แจกแจงแบบปกติ

2) ทดสอบการแจกแจงของข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 ในสถานการณ์การนำทาง

   H_0: ข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 ในสถานการณ์การนำทาง มีการแจกแจงแบบปกติ

   H_1: ข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 ในสถานการณ์การนำทาง ไม่แจกแจงแบบปกติ

3) ทดสอบการแจกแจงของข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

   H_0: ข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

   H_1: ข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่แจกแจงแบบปกติ

4) ทดสอบการแจกแจงของข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

H_0: ข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่แจกแจงแบบปกติ

5) ทดสอบการแจกแจงของข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 ในสถานการณ์การนำทาง

H_0: ข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 ในสถานการณ์การนำทาง มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 ในสถานการณ์การนำทาง ไม่แจกแจงแบบปกติ

6) ทดสอบการแจกแจงของข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 ในสถานการณ์การนำทาง

H_0: ข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 ในสถานการณ์การนำทาง มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 ในสถานการณ์การนำทาง ไม่แจกแจงแบบปกติ

7) ทดสอบการแจกแจงของข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

H_0: ข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่แจกแจงแบบปกติ

8) ทดสอบการแจกแจงของข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

H_0: ข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่แจกแจงแบบปกติ

ในการตรวจสอบการแจกแจงของข้อมูลว่าเป็นแบบปกติโดยใช้สถิติทดสอบนั้น มีสถิติทดสอบที่ใช้คือ Kolmogorov-Sminov สำหรับหน่วยทดลองมากกว่า 50 หน่วย และ Shapiro-Wilk สำหรับหน่วยทดลองน้อยกว่า 50 หน่วย โดยจะยอมรับสมมติฐาน H_0 เมื่อค่า Sig. มีค่ามากกว่าค่า α ซึ่งกำหนดให้เท่ากับ 0.05 ผลการทดสอบแสดงดังตารางต่อไปนี้

ตารางที่ 4-7 แสดงค่าสถิติทดสอบการแจกแจงปกติ (Normality Test) ของค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการค้นกฎความสัมพันธ์ทั้ง 2 ตัวแบบ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด

		ตัวแบบที่	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
			Statistic	df	Sig.	Statistic	df	Sig.
สถานการณ์การนำทาง	P_μ	1	0.088	451	0.000	0.951	451	0.000
		2	0.097	451	0.000	0.945	451	0.000
	R_μ	1	0.171	451	0.000	0.852	451	0.000
		2	0.115	451	0.000	0.901	451	0.000
สถานการณ์การป้องกันการเกิดข้อผิดพลาด	P_μ	1	0.296	451	0.000	0.708	451	0.000
		2	0.314	451	0.000	0.699	451	0.000
	R_μ	1	0.351	451	0.000	0.636	451	0.000
		2	0.345	451	0.000	0.636	451	0.000

ผลการทดสอบในตารางที่ 4-7 ข้างต้นพบว่าค่า Sig. ของตัวแปรค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดเป็นดังนี้

1) สำหรับสถานการณ์การนำทาง ค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

2) สำหรับสถานการณ์การนำทาง ค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0
3) สำหรับสถานการณ์การนำทาง ค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0
4) สำหรับสถานการณ์การนำทาง ค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0
5) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด ค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0
6) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด ค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่ามากกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0
7) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด ค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0
8) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด ค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่ามากกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

ดังนั้นสรุปได้ว่าการแจกแจงของตัวแปรค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบ ใน 2 สถานการณ์นั้นไม่แจกแจงแบบปกติ

ผลการทดสอบ

จากการวิเคราะห์การแจกแจงข้อมูลข้างต้นพบว่าค่าความถูกต้องของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดไม่มีการแจกแจงแบบปกติ ดังนั้นผู้วิจัยจึงเลือกใช้การทดสอบสมมติฐานแบบไม่อิงพารามิเตอร์ (Non Parametic Test) นั่นคือสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) กับการทดสอบต่อไปนี้ โดยที่จะปฏิเสธสมมติฐาน H_0 ได้เมื่อถ้าค่า Sig. (Significance) ที่คำนวณได้น้อยกว่า 0.05 และค่าสถิติ Z มากกว่า 0 โดยที่ผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks)

การวิเคราะห์เปรียบเทียบนี้เป็นการวิเคราะห์เปรียบเทียบค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ระหว่างการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 กับการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การนำทาง และสถานการณ์การป้องกันการเกิดข้อผิดพลาด ซึ่งสามารถตั้งสมมุติฐานได้ดังนี้

1) วิเคราะห์เปรียบเทียบค่าความถูกต้องของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ทั้ง 2 ตัวแบบ ในสถานการณ์การนำทาง ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การนำทาง

M_2 คือ ค่ามัธยฐานของค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การนำทาง

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

2) วิเคราะห์เปรียบเทียบค่าเรียกคืนของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ทั้ง 2 ตัวแบบโดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัว

แบบที่ 1 ในสถานการณ์การนำทาง

M_2 คือ ค่ามัธยฐานของค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การนำทาง

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

3) วิเคราะห์เปรียบเทียบค่าความถูกต้องของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

M_2 คือ ค่ามัธยฐานของค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

$H_0 : \mu_2 \leq \mu_1$

$H_1 : \mu_2 > \mu_1$

4) วิเคราะห์เปรียบเทียบค่าเรียกคืนของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

M_2 คือ ค่ามัธยฐานของค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

ผู้วิจัยใช้การทดสอบสมมติฐานโดยสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) กับการทดสอบต่อไปนี้ ผลการทดสอบแสดงดังตารางต่อไปนี้

ตารางที่ 4-8 แสดงสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ของค่าความถูกต้องและค่าเรียกคืนของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 เทียบกับตัวแบบที่ 1 ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด

	ค่าความถูกต้องของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 - ค่าความถูกต้องของการค้นหากฎความสัมพันธ์ตัวแบบที่ 1		ค่าเรียกคืนของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 - ค่าเรียกคืนของการค้นหากฎความสัมพันธ์ตัวแบบที่ 1	
	สถานการณ์การนำทาง	สถานการณ์การป้องกันการเกิดข้อผิดพลาด	สถานการณ์การนำทาง	สถานการณ์การป้องกันการเกิดข้อผิดพลาด
Z	-2.669[a]	-6.054[a]	-9.234[b]	-2.000[a]
Asymp. Sig. (2-tailed)	0.008	0.000	0.000	0.046

a. Based on positive ranks.

b. Based on negative ranks.

จากตารางที่ 4-8 ได้ผลทดสอบดังนี้

1) การเปรียบเทียบค่าความถูกต้องในสถานการณ์การนำทาง ได้สถิติทดสอบค่า Z เท่ากับ -2.669 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 0.008 / 2 = 0.004 ซึ่งน้อยกว่าค่า α = 0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) ดังนั้นจึงไม่สามารถปฏิเสธสมมติ H_0 ได้ นั่นคือค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ไม่ต่างกันหรือน้อยกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การนำทาง ที่ระดับนัยสำคัญ 0.05

2) การเปรียบเทียบค่าความถูกต้องในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ได้สถิติทดสอบค่า Z เท่ากับ -6.054 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 0.000 / 2 = 0.000 ซึ่งน้อยกว่าค่า α = 0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) ดังนั้นจึงไม่สามารถปฏิเสธสมมติ H_0 ได้ นั่นคือค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ไม่ต่างกันหรือน้อยกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาดที่ระดับนัยสำคัญ 0.05

3) การเปรียบเทียบค่าเรียกคืนในสถานการณ์การนำทาง ได้สถิติทดสอบค่า Z เท่ากับ -9.234 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 0.000 / 2 = 0.000 ซึ่งน้อยกว่าค่า α = 0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางลบ (Based on negative ranks) ดังนั้นจึงสามารถปฏิเสธสมมติ H_0 ได้ นั่นคือค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มากกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การนำทาง ที่ระดับนัยสำคัญ 0.05

4) การเปรียบเทียบค่าเรียกคืนในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ได้สถิติทดสอบค่า Z เท่ากับ -2.000 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 0.000 / 2 = 0.000 ซึ่งน้อยกว่าค่า α = 0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) ดังนั้นจึงไม่สามารถปฏิเสธสมมติ H_0 ได้ นั่นคือค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ไม่ต่างกันหรือน้อยกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาดที่ระดับนัยสำคัญ 0.05

สรุปผลการทดสอบ

จากการวิเคราะห์ผลการเปรียบเทียบค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด ผู้วิจัยสามารถสรุปได้ว่าในสถานการณ์การนำทางการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ให้ค่าเรียกคืน (Recall) มากกว่าแต่ให้ค่าความถูกต้อง (Precision) ที่น้อยกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 สำหรับสถานการณ์การป้องกันการเกิด

ข้อผิดพลาดการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ให้ค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ไม่ต่างกันหรือน้อยกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1

4.4.2 การเปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบโดยเปลี่ยนข้อกำหนดของการสร้างเซตของคำแนะนำ

การเปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบโดยสร้างเซตของคำแนะนำจากการยูเนียน (Union) ของรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตรายการที่ตามมาของกฎความสัมพันธ์อันดับสูงสุดจำนวน 10 รายการแรก ผู้วิจัยกำหนดให้วิธีการสร้างเซตของคำแนะนำดังกล่าวเรียกว่า “การสร้างเซตของคำแนะนำแบบใหม่” ผลการทดสอบดังกล่าวแสดงดังต่อไปนี้

ตารางที่ 4-9 แสดงตารางผลการทดสอบเพิ่มเติมทั้ง 3 สถานการณ์

สถานการณ์การนำทาง (ค่าเฉลี่ยของค่าเอฟเมสเซอร์)		สถานการณ์การป้องกันการเกิดข้อผิดพลาด (ค่าเฉลี่ยของค่าเอฟเมสเซอร์)		สถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (ค่าผลสะท้อนกลับ)	
ตัวแบบที่ 1	ตัวแบบที่ 2	ตัวแบบที่ 1	ตัวแบบที่ 2	ตัวแบบที่ 1	ตัวแบบที่ 2
0.3195	0.3335	0.1152	0.1057	$\frac{41}{60}$	$\frac{41}{60}$

จากตารางสรุปข้างต้นแสดงให้เห็นว่าค่าประสิทธิภาพของสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้นมีค่าแตกต่างกัน ค่าประสิทธิภาพของสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วนั้นไม่มีความแตกต่างกัน แต่ไม่สามารถสรุปได้ว่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ในสถานการณ์ต่างๆนั้นแตกต่างกันอย่างมีนัยสำคัญ ผู้วิจัยจึงจำเป็นต้องวิเคราะห์ผลการทดสอบความแตกต่างกันอย่างมีนัยสำคัญ ดังรายละเอียดต่อไปนี้

การวิเคราะห์การแจกแจงข้อมูล

เนื่องจากผู้วิจัยต้องการเปรียบเทียบประสิทธิภาพของการค้นหากฎความสัมพันธ์ของ 2 ตัวแบบสำหรับสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด โดยการ

สร้างเซตของคำแนะนำจากการยูเนียน (Union) ของรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตรายการที่ตามมาของกฎความสัมพันธ์อันดับสูงสุดจำนวน 10 รายการแรก ดังนั้นผู้วิจัยจะตรวจสอบว่าค่าเอฟเมเซอร์ทั้งหมดมีการแจกแจงแบบปกติหรือไม่ เฉพาะในสำหรับสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด โดยตั้งสมมติฐานของการทดสอบค่าเอฟเมเซอร์แต่ละการทดสอบมีการแจกแจงแบบปกติหรือไม่ภายใต้สมมติฐานทางสถิติ ดังนี้

1) ทดสอบการแจกแจงของข้อมูลค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง

   H_0: ข้อมูลค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง มีการแจกแจงแบบปกติ

   H_1: ข้อมูลค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง ไม่แจกแจงแบบปกติ

2) ทดสอบการแจกแจงของข้อมูลค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง

   H_0: ข้อมูลค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง มีการแจกแจงแบบปกติ

   H_1: ข้อมูลค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง ไม่แจกแจงแบบปกติ

3) ทดสอบการแจกแจงของข้อมูลค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

   H_0: ข้อมูลค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

   H_1: ข้อมูลค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่แจกแจงแบบปกติ

4) ทดสอบการแจกแจงของข้อมูลค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

H_0: ข้อมูลค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่แจกแจงแบบปกติ

ในการตรวจสอบการแจกแจงของข้อมูลว่าเป็นแบบปกติโดยใช้สถิติทดสอบนั้น มีสถิติทดสอบที่ใช้คือ Kolmogorov-Sminov สำหรับหน่วยทดลองมากกว่า 50 หน่วย และ Shapiro-Wilk สำหรับหน่วยทดลองน้อยกว่า 50 หน่วย โดยจะยอมรับสมมติฐาน H_0 เมื่อค่า Sig. มีค่ามากกว่าค่า α ซึ่งกำหนดให้เท่ากับ 0.05 ผลการทดสอบแสดงดังตารางต่อไปนี้

ตารางที่ 4-10 แสดงค่าสถิติทดสอบการแจกแจงปกติ (Normality Test) ของค่าประสิทธิภาพของการค้นกฎความสัมพันธ์ทั้ง 2 ตัวแบบ โดยการสร้างเซตของคำแนะนำจากการยูเนียน (Union) ของรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตรายการที่ตามมาของกฎความสัมพันธ์อันดับสูงสุด จำนวน 10 รายการแรก ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด

สถานการณ์	ตัวแบบที่	Kolmogorov-Smirnov		
		Statistic	df	Sig.
สถานการณ์การนำทาง	1	0.071	451	0.000
	2	0.070	451	0.000
สถานการณ์การป้องกันการเกิดข้อผิดพลาด	1	0.296	451	0.000
	2	0.325	451	0.000

ผลการทดสอบในตารางที่ 4-10 ข้างต้นพบว่าค่า Sig. ของตัวแปรค่าประสิทธิภาพการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดเป็นดังนี้

1) สำหรับสถานการณ์การนำทาง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

2) สำหรับสถานการณ์การนำทาง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ มีค่า Sig. เท่ากับ 0. 000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0
3) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0
4) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0

ดังนั้นสรุปได้ว่าการแจกแจงของตัวแปรค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบสำหรับ 2 สถานการณ์นั้นไม่เป็นแบบปกติ

การวิเคราะห์เปรียบเทียบประสิทธิภาพการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบ

จากการวิเคราะห์การแจกแจงข้อมูลข้างต้นพบว่าค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดโดยการสร้างเซตของคำแนะนำจากการยูเนียน (Union) ของรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตรายการที่ตามมาของกฎความสัมพันธ์อันดับสูงสุดจำนวน 10 รายการแรก ไม่แจกแจงแบบปกติ ดังนั้นผู้วิจัยจึงเลือกใช้สถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) กับการทดสอบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด โดยที่จะปฏิเสธสมมติฐาน H_0 ได้เมื่อถ้าค่า Sig. (Significance) ที่คำนวณได้น้อยกว่า 0.05 และค่าสถิติ Z มากกว่า 0 โดยที่ผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) และผู้วิจัยเลือกใช้การทดสอบอัตราส่วน 2 กลุ่มด้วยสถิติทดสอบ Z (Two Proportion Z Tests) กับการทดสอบในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว โดยที่จะปฏิเสธสมมติฐาน H_0 ได้เมื่อถ้าค่า Sig. (Significance) ที่คำนวณได้น้อยกว่า 0.05

การวิเคราะห์เปรียบเทียบนี้เป็นการวิเคราะห์เปรียบเทียบประสิทธิภาพระหว่างการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 กับการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำจากการยูเนียน (Union) ของรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตรายการที่

ตามมาของกฎความสัมพันธ์อันดับสูงสุดจำนวน 10 รายการแรก ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด ในสถานการณ์ 3 สถานการณ์คือ สถานการณ์การนำทาง สถานการณ์การป้องกันการเกิดข้อผิดพลาด สถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว ซึ่งสามารถตั้งสมมุติฐานได้ดังนี้

1) วิเคราะห์เปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์วทั้ง 2 ตัวแบบโดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การนำทาง

M_2 คือ ค่ามัธยฐานของค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การนำทาง

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

2) วิเคราะห์เปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์วทั้ง 2 ตัวแบบโดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

M_2 คือ ค่ามัธยฐานของค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

3) วิเคราะห์เปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์วทั้ง 2 ตัวแบบโดยการสร้างเซตของ

คำแนะนำแบบใหม่ ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

M_2 คือ ค่ามัธยฐานของค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

ผู้วิจัยทดสอบสมมติฐานข้อ 1 และ 2 ด้วยสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) แสดงดังตารางที่ 4-11 สำหรับสมมติฐานข้อ 3 ผู้วิจัยเลือกใช้การทดสอบอัตราส่วน 2 กลุ่มด้วยสถิติทดสอบ Z (Two Proportion Z Tests) แสดงดังตารางที่ 4-12

ตารางที่ 4-11 แสดงค่าสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ของค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 เทียบกับตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด

	ค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 - ค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ตัวแบบที่1	
	สถานการณ์การนำทาง	สถานการณ์การป้องกันการเกิดข้อผิดพลาด
Z	-3.159[a]	-3.878[b]
Asymp. Sig. (2 tailed)	.000	.000

a. Based on negative ranks.

b. Based on positive ranks.

จากตารางที่ 4-11 การวิเคราะห์เปรียบเทียบประสิทธิภาพของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบในสถานการณ์การนำทาง ได้สถิติทดสอบค่า Z เท่ากับ -3.159 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 0.000 / 2 = 0.000 ซึ่งน้อยกว่าค่า α = 0.05 เนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางลบ (Based on negative ranks) ดังนั้นจึงสามารถปฏิเสธสมมติฐาน H_0 ได้ นั่นคือค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มากกว่าค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การนำทาง ที่ระดับนัยสำคัญ 0.05 และสำหรับการวิเคราะห์เปรียบเทียบประสิทธิภาพของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ได้สถิติทดสอบค่า Z เท่ากับ -3.878 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 0.000 / 2 = 0.000 ซึ่งน้อยกว่าค่า α = 0.05 เนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) ดังนั้นจึงไม่สามารถปฏิเสธสมมติฐาน H_0 ได้ นั่นคือค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ไม่ต่างกันหรือน้อยกว่าค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาดที่ระดับนัยสำคัญ 0.05

ตารางที่ 4-12 แสดงการทดสอบอัตราส่วน 2 กลุ่มด้วยสถิติทดสอบ Z ของค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 เทียบกับตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.000[a]	1	1.000		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.000	1	1.000		
Fisher's Exact Test				1.000	.578
N of Valid Cases	120				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 19.00.

จากตารางที่ 4-12 ค่าสถิติ Z เท่ากับรากที่สองของค่า Pearson Chi-Square นั่นค่า 0.000 และมีเครื่องหมายเดียวกับผลต่างของอัตราส่วนที่ 2 กับอัตราส่วนที่ 1 ($\frac{41}{60} - \frac{41}{60} = 0.000$) นั่นค่าเครื่องหมายบวก และจากการตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 1.000 / 2 = 0.500 ซึ่งมากกว่าค่า α = 0.05 ดังนั้นจึงไม่สามารถปฏิเสธสมมติฐาน H_0 ได้ นั่นคือค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ไม่ต่างกับค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วที่ระดับนัยสำคัญ 0.05

สรุปผลการวิเคราะห์ข้อมูล

จากการวิเคราะห์ผลการทดสอบประสิทธิภาพของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบโดยการสร้างเซตของคำแนะนำจากการยูเนียน (Union) ของรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตรายการที่ตามมาของกฎความสัมพันธ์อันดับสูงสุดจำนวน 10 รายการแรก ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด โดยใช้ค่าเอฟเมสเซอร์และสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว โดยใช้ค่าผลสะท้อนกลับ ผู้วิจัยสามารถสรุปได้ว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มีประสิทธิภาพมากกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การนำทาง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มีประสิทธิภาพไม่ต่างกันหรือน้อยกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด และสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มีประสิทธิภาพไม่ต่างกับการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1

ผลสรุปดังกล่าวไม่ได้แสดงให้เห็นว่า การกำหนดการสร้างเซตของ คำแนะนำจากการยูเนียน (Union) ของรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตรายการที่ตามมาของกฎความสัมพันธ์อันดับสูงสุดจำนวน 10 รายการแรกนั้นสามารถเพิ่มค่าความถูกต้อง (Precision) ให้กับการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 อย่างที่ผู้วิจัยได้ตั้งข้อสังเกตไว้หรือไม่ ดังนั้นผู้วิจัยจึงมีความต้องการศึกษาเพิ่มเติมคือ การเปรียบเทียบค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบทั้ง 2 ตัวแบบ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด เพื่อทดสอบว่ากำหนดการสร้างเซตของ

คำแนะนำดังกล่าวสามารถเพิ่มค่าความถูกต้อง (Precision) หรือค่าเรียกคืน (Recall) ในสถานการณ์ทั้ง 2 ได้หรือไม่ การทดสอบดังกล่าวแสดงในหัวต่อไปนี้

4.4.3 การเปรียบเทียบค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบโดยเปลี่ยนข้อกำหนดของการสร้างเซตของคำแนะนำ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด

ผลการทดสอบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบโดยการสร้างเซตของคำแนะนำจากการยูเนียน (Union) ของรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตรายการที่ตามมาของกฎความสัมพันธ์อันดับสูงสุดจำนวน 10 รายการแรกหรือการสร้างเซตของคำแนะนำแบบใหม่ในการทดสอบที่แล้วนั้นเป็นการเปรียบประสิทธิภาพจากการใช้ค่าเอฟเมเซอร์ (F-measure) เท่านั้น ผู้วิจัยจึงมีความต้องการศึกษาเพิ่มเติมโดยการทดสอบเปรียบเทียบค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด ผลการทดสอบแสดงดังตารางต่อไปนี้

ตารางที่ 4-13 แสดงตารางค่าความถูกต้องและค่าเรียกคืนของการทดสอบเพิ่มเติมโดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด

สถานการณ์การนำทาง				สถานการณ์การป้องกันการเกิดข้อผิดพลาด			
ตัวแบบที่ 1		ตัวแบบที่ 2		ตัวแบบที่ 1		ตัวแบบที่ 2	
P_μ	R_μ	P_μ	R_μ	P_μ	R_μ	P_μ	R_μ
0.4034	0.3856	0.4191	0.3985	0.4621	0.5477	0.4568	0.4945

จากตารางสรุปข้างต้นแสดงให้เห็นว่าค่าเฉลี่ยของค่าความถูกต้อง (Precision) หรือ P_μ และค่าเฉลี่ยของค่าเรียกคืน (Recall) หรือ R_μ ของการทดสอบข้างต้นนั้นมีค่าแตกต่างกันทั้ง 2 ตัวแบบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด แต่ไม่สามารถสรุปได้ว่าค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการค้นหากฎความสัมพันธ์ของตัว

แบบทั้ง 2 ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้นแตกต่างกันอย่างมีนัยสำคัญ ผู้วิจัยจึงจำเป็นต้องวิเคราะห์ผลการทดสอบความแตกต่างกันอย่างมีนัยสำคัญ ดังรายละเอียดต่อไปนี้

การวิเคราะห์การแจกแจงปกติ

เนื่องจากผู้วิจัยต้องการเปรียบเทียบค่าความถูกต้อง (Precision) หรือค่าเรียกคืน (Recall) ของการค้นหากฎความสัมพันธ์ของ 2 ตัวแบบโดยการสร้างเซตของคำแนะนำแบบใหม่ สำหรับสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด ดังนั้นผู้วิจัยจะตรวจสอบว่าค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ทั้งหมดของทั้ง 2 สถานการณ์มีการแจกแจงแบบปกติหรือไม่ โดยตั้งสมมติฐานของการทดสอบค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) แต่ละสถานการณ์มีการแจกแจงแบบปกติหรือไม่ภายใต้สมมติฐานทางสถิติ ดังนี้

1) ทดสอบการแจกแจงของข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง

   H_0: ข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง มีการแจกแจงแบบปกติ

   H_1: ข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง ไม่แจกแจงแบบปกติ

2) ทดสอบการแจกแจงของข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง

   H_0: ข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง มีการแจกแจงแบบปกติ

   H_1: ข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง ไม่แจกแจงแบบปกติ

3) ทดสอบการแจกแจงของข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

   H_0: ข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่แจกแจงแบบปกติ

4) ทดสอบการแจกแจงของข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

H_0: ข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าความถูกต้องของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่แจกแจงแบบปกติ

5) ทดสอบการแจกแจงของข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง

H_0: ข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง ไม่แจกแจงแบบปกติ

6) ทดสอบการแจกแจงของข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง

H_0: ข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง ไม่แจกแจงแบบปกติ

7) ทดสอบการแจกแจงของข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

H_0: ข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่แจกแจงแบบปกติ

8) ทดสอบการแจกแจงของข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

H_0: ข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเรียกคืนของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่แจกแจงแบบปกติ

ในการตรวจสอบการแจกแจงของข้อมูลว่าเป็นแบบปกติโดยใช้สถิติทดสอบนั้น มีสถิติทดสอบที่ใช้คือ Kolmogorov-Sminov สำหรับหน่วยทดลองมากกว่า 50 หน่วย และ Shapiro-Wilk สำหรับหน่วยทดลองน้อยกว่า 50 หน่วย โดยจะยอมรับสมมติฐาน H_0 เมื่อค่า Sig. มีค่ามากกว่าค่า α ซึ่งกำหนดให้เท่ากับ 0.05 ผลการทดสอบแสดงดังตารางต่อไปนี้

ตารางที่ 4-14 แสดงค่าสถิติทดสอบการแจกแจงปกติ (Normality Test) ของค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการค้นกฎความสัมพันธ์ทั้ง 2 ตัวแบบโดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด

		ตัวแบบที่	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
			Statistic	df	Sig.	Statistic	df	Sig.
สถานการณ์การนำทาง	P_μ	1	0.112	451	0.000	0.955	451	0.000
		2	0.111	451	0.000	0.949	451	0.000
	R_μ	1	0.113	451	0.000	0.906	451	0.000
		2	0.09	451	0.000	0.927	451	0.000
สถานการณ์การป้องกันการเกิดข้อผิดพลาด	P_μ	1	0.339	451	0.000	0.673	451	0.000
		2	0.333	451	0.000	0.683	451	0.000
	R_μ	1	0.366	451	0.000	0.633	451	0.000
		2	0.344	451	0.000	0.636	451	0.000

ผลการทดสอบในตารางที่ 4-14 ข้างต้นพบว่าค่า Sig. ของตัวแปรค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบโดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดเป็นดังนี้

1) สำหรับสถานการณ์การนำทาง ค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0
2) สำหรับสถานการณ์การนำทาง ค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0
3) สำหรับสถานการณ์การนำทาง ค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0
4) สำหรับสถานการณ์การนำทาง ค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0
5) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด ค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0
6) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด ค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่ามากกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0

7) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด ค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0
8) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด ค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่ามากกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

ดังนั้นสรุปได้ว่าการแจกแจงของตัวแปรค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบโดยการสร้างเซตของคำแนะนำแบบใหม่ ในทั้ง 2 สถานการณ์นั้นไม่แจกแจงแบบปกติ

ผลการทดสอบ

จากการวิเคราะห์การแจกแจงข้อมูลข้างต้นพบว่าค่าความถูกต้องของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบโดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดไม่มีการแจกแจงแบบปกติ ดังนั้นผู้วิจัยจึงเลือกใช้การทดสอบสมมติฐานแบบไม่อิงพารามิเตอร์ (Non Parametic Test) นั่นคือสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) กับการทดสอบต่อไปนี้ โดยที่จะปฏิเสธสมมติฐาน H_0 ได้เมื่อถ้าค่า Sig. (Significance) ที่คำนวณได้น้อยกว่า 0.05 และค่าสถิติ Z มากกว่า 0 โดยที่ผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks)

การวิเคราะห์เปรียบเทียบนี้เป็นการวิเคราะห์เปรียบเทียบค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ระหว่างการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 กับการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด ซึ่งสามารถตั้งสมมุติฐานได้ดังนี้

1) วิเคราะห์เปรียบเทียบค่าความถูกต้องของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ทั้ง 2 ตัวแบบโดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง

M_2 คือ ค่ามัธยฐานของค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

2) วิเคราะห์เปรียบเทียบค่าเรียกคืนของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ทั้ง 2 ตัวแบบโดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง

M_2 คือ ค่ามัธยฐานของค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทาง

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

3) วิเคราะห์เปรียบเทียบค่าความถูกต้องของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ทั้ง 2 ตัวแบบโดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

M_2 คือ ค่ามัธยฐานของค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

$H_0 : \mu_2 \leq \mu_1$

$H_1 : \mu_2 > \mu_1$

4) วิเคราะห์เปรียบเทียบค่าเรียกคืนของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ทั้ง 2 ตัวแบบโดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

M_2 คือ ค่ามัธยฐานของค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

ผู้วิจัยจึงใช้การทดสอบสมมติฐานแบบไม่อิงพารามิเตอร์ (Non Parametic Test) นั่นคือ สถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) กับการทดสอบต่อไปนี้ โดยที่ผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) ผลการทดสอบแสดงดังตารางต่อไปนี้

ตารางที่ 4-15 แสดงสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ของค่าความถูกต้องและค่าเรียกคืนของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 เทียบกับตัวแบบที่ 1 โดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด

	ค่าความถูกต้องของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 - ค่าความถูกต้องของการค้นหากฎความสัมพันธ์ตัวแบบที่ 1		ค่าเรียกคืนของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 - ค่าเรียกคืนของการค้นหากฎความสัมพันธ์ตัวแบบที่ 1	
	สถานการณ์การนำทาง	สถานการณ์การป้องกันการเกิดข้อผิดพลาด	สถานการณ์การนำทาง	สถานการณ์การป้องกันการเกิดข้อผิดพลาด
Z	-3.342[a]	-3.878[b]	-3.290[a]	-4.116[b]
Asymp. Sig. (2-tailed)	0.001	0.000	0.001	0.000

a. Based on negative ranks.

b. Based on positive ranks.

จากตารางที่ 4-15 ได้ผลทดสอบดังนี้

1) การเปรียบเทียบค่าความถูกต้องในสถานการณ์การนำทาง ได้สถิติทดสอบค่า Z เท่ากับ -3.342 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 0.001 / 2 = 0.0005 ซึ่งน้อยกว่าค่า α = 0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางลบ (Based on negative ranks) ดังนั้นจึงสามารถปฏิเสธสมมติ H_0 ได้ นั่นคือค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มากกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การนำทาง ที่ระดับนัยสำคัญ 0.05

2) การเปรียบเทียบค่าความถูกต้องในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ได้สถิติทดสอบค่า Z เท่ากับ -3.878 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 0.000 / 2 = 0.000 ซึ่งน้อยกว่าค่า α = 0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) ดังนั้นจึงไม่สามารถปฏิเสธสมมติ H_0 ได้ นั่นคือค่าความถูกต้องของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ไม่ต่างกันหรือน้อยกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาดที่ระดับนัยสำคัญ 0.05

3) การเปรียบเทียบค่าเรียกคืนในสถานการณ์การนำทาง ได้สถิติทดสอบค่า Z เท่ากับ -3.290 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 0.001 / 2 = 0.0005 ซึ่งน้อยกว่าค่า α = 0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางลบ (Based on negative ranks) ดังนั้นจึงสามารถปฏิเสธสมมติ H_0 ได้ นั่นคือค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มากกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การนำทาง ที่ระดับนัยสำคัญ 0.05

4) การเปรียบเทียบค่าเรียกคืนในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ได้สถิติทดสอบค่า Z เท่ากับ -4.116 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 0.000 / 2 = 0.000 ซึ่งน้อยกว่าค่า α = 0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) ดังนั้นจึงไม่สามารถปฏิเสธสมมติ H_0 ได้ นั่นคือค่าเรียกคืนของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ไม่ต่างกันหรือน้อยกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในสถานการณ์การป้องกันการเกิดข้อผิดพลาดที่ระดับนัยสำคัญ 0.05

สรุปผลการทดสอบ

จากการวิเคราะห์ผลการเปรียบเทียบค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบโดยการสร้างเซตของคำแนะนำแบบใหม่ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด ผู้วิจัยสามารถสรุปได้ว่าในสถานการณ์การนำทางการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ให้ค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) มากกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาดการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ให้ค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ไม่ต่างกันหรือน้อยกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1

จากผลการทดสอบนี้และผลการทดสอบในหัวข้อ 4.4.1 แสดงให้เห็นว่าการกำหนดการสร้างเซตของคำแนะนำแบบใหม่หรือการสร้างเซตของคำแนะนำจากการยูเนียน (Union) ของรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตรายการที่ตามมาของกฎความสัมพันธ์อันดับสูงสุดจำนวน 10 รายการแรกนั้น ทำให้ค่าความถูกต้อง (Precision) ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในสถานการณ์การนำทางเพิ่มขึ้นได้ตามที่ผู้วิจัยได้ตั้งข้อสังเกตไว้ แต่สำหรับใน

สถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้นการกำหนดการสร้างเซตของคำแนะนำแบบใหม่ไม่ได้ทำให้ค่าความถูกต้อง (Precision) ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ดีขึ้น ด้วยเหตุนี้ผู้วิจัยจึงตั้งข้อสังเกตต่อไปว่าสาเหตุที่ทำให้ผลการทดสอบออกมาเป็นเช่นนั้นเพราะในสถานการณ์การนำทางกฎความสัมพันธ์ที่ได้การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในอันดับต้นๆนั้นมีกฎความสัมพันธ์ที่เป็นผลบวกลวงปะปนอยู่มากในขณะที่กฎความสัมพันธ์ที่ได้การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในอันดับต้นๆนั้นมีกฎความสัมพันธ์ที่เป็นผลบวกลวงน้อยกว่า แต่ในสถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้นกฎความสัมพันธ์ที่ได้การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในอันดับต้นๆนั้นมีกฎความสัมพันธ์ที่เป็นผลบวกลวงปะปนอยู่น้อย ในขณะที่กฎความสัมพันธ์ที่ได้การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในอันดับต้นๆนั้นมีกฎความสัมพันธ์ที่เป็นผลบวกลวงมากกว่า

ด้วยเหตุนี้ผู้วิจัยจึงมีความต้องการที่จะทดสอบเพิ่มเติมตามข้อสังเกตข้างต้นเพื่ออธิบายสาเหตุของผลการทดสอบหลักของงานวิจัยนี้ ผู้วิจัยกำหนดการทดสอบเพิ่มเติมโดยการเปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบโดยปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็น 7 อันดับแรก 5 อันดับแรก และ 3 อันดับแรก ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด ตามลำดับ

4.4.4 การเปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบโดยปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำ

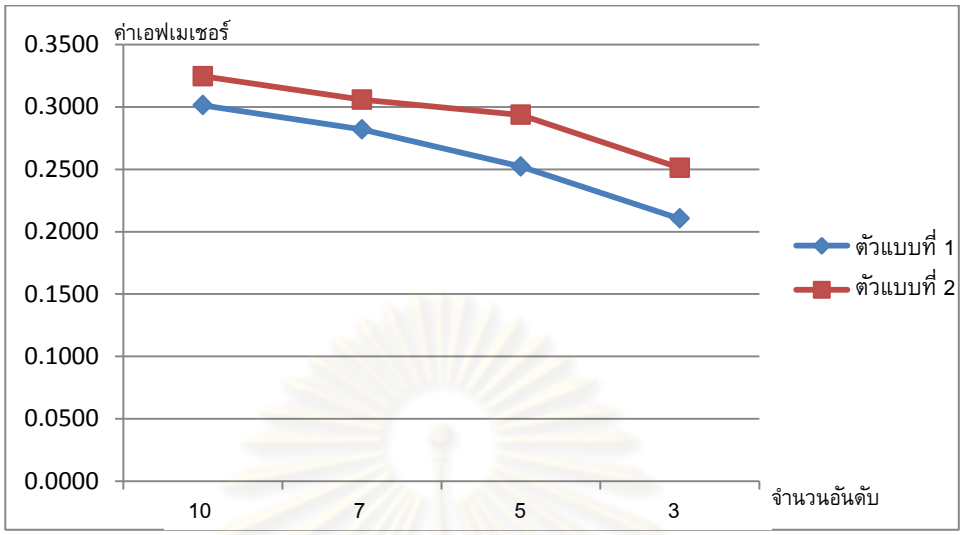
ผู้วิจัยยังตั้งข้อสังเกตอีกว่ากฎความสัมพันธ์ทั้ง 10 อันดับแรกที่ได้จากการค้นหากฎความสัมพันธ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นนั้น กฎความสัมพันธ์ที่ตรงกับเซตของผลลัพธ์ที่คาดไว้มักจะไม่ใช่กฎความสัมพันธ์ที่อยู่ในอันดับแรกๆ แต่สำหรับการค้นหากฎความสัมพันธ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่นั้น กฎความสัมพันธ์ที่ตรงกับเซตของผลลัพธ์ที่คาดไว้มักจะเป็นกฎความสัมพันธ์ที่อยู่ในอันดับแรกๆ เหตุผลที่เป็นเช่นนั้นผู้วิจัยคิดว่าอาจเป็นไปได้ว่ากฎความสัมพันธ์ที่ได้จากการค้นหากฎความสัมพันธ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นในอันดับแรกๆมักเป็นผลบวกลวง ผู้วิจัยจึงสนใจวิเคราะห์เปรียบเทียบประสิทธิภาพ

ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบในสถานการณ์นำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด โดยทดลองปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ คือ 1) สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก 2) สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก และ 3) สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ผลการทดสอบแสดงดังตารางต่อไปนี้

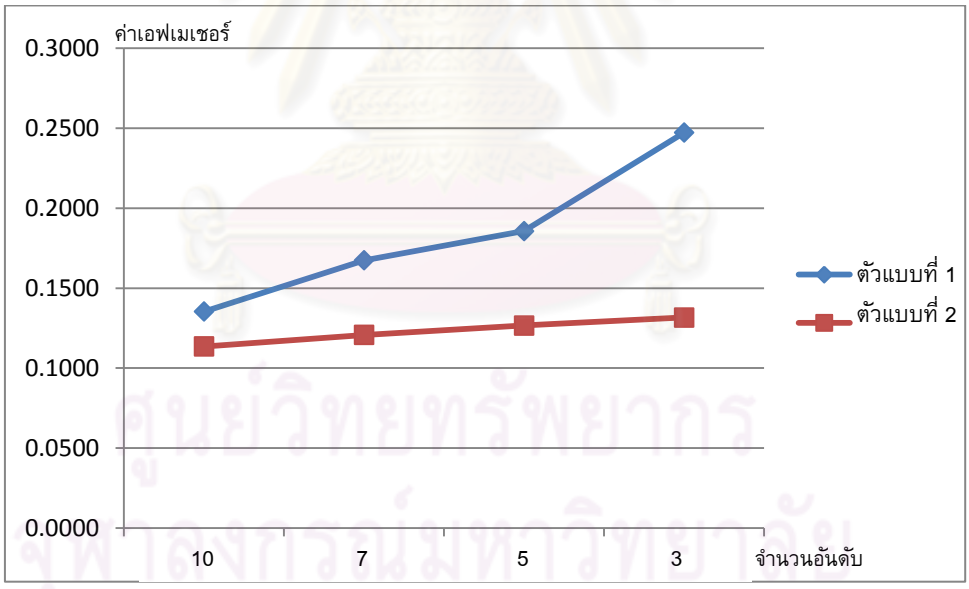
ตารางที่ 4-16 แสดงตารางค่าเฉลี่ยของค่าเอฟเมสเซอร์ของการทดสอบการปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด

	ตัวแบบที่	ค่าเฉลี่ยของค่าเอฟเมสเซอร์			
		10 อันดับแรก	7 อันดับแรก	5 อันดับแรก	3 อันดับแรก
สถานการณ์การนำทาง	1	0.3013	0.2820	0.2522	0.2105
	2	0.3245	0.3058	0.2936	0.2511
สถานการณ์การป้องกันการเกิดข้อผิดพลาด	1	0.1353	0.1673	0.1857	0.2472
	2	0.1135	0.1207	0.1267	0.1316

จากตารางที่ 4-16 ข้างบนสามารถนำมาแสดงในรูปแบบข้างกราฟเส้นเพื่อให้เห็นแนวโน้มค่าเฉลี่ยของค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบ เมื่อปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด ได้ดังนี้



รูปที่ 4-3 แสดงกราฟเส้นค่าเฉลี่ยของค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบ เมื่อปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ ในสถานการณ์การนำทาง



รูปที่ 4-4 แสดงกราฟเส้นค่าเฉลี่ยของค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบ เมื่อปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

จากรูปที่ 4-3 จะเห็นว่าค่าเฉลี่ยของค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มีค่ามากกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในทุกๆ การปรับจำนวนของกฎ

ความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ ในสถานการณ์การนำทาง นอกจากนั้นจะเห็นว่าค่าเฉลี่ยของค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบมีแนวโน้มลดลงเรื่อยๆเมื่อมีการปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าที่น้อยลง แสดงให้เห็นว่าการค้นหากฎความสัมพันธ์ในสถานการณ์การนำทางนั้น กฎความสัมพันธ์ที่ถูกต้องจริงๆนั้นมันจะเป็นกฎความสัมพันธ์ที่อยู่ในอันดับท้ายๆ และกฎความสัมพันธ์ที่อยู่ในอันดับต้นๆนั้นเป็นผลบวกลวงในทั้ง 2 ตัวแบบ

จากรูปที่ 4-4 จะเห็นว่าค่าเฉลี่ยของค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 มีค่ามากกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในทุกๆ การปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ นอกจากนั้นจะเห็นว่าค่าเฉลี่ยของค่าเอฟเมสเซอร์ของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบมีแนวโน้มสูงขึ้นเรื่อยๆเมื่อมีการปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าที่น้อยลง สาเหตุมาจากค่าเรียกคืนที่ยังคงที่เสมอในขณะที่ค่าความถูกต้องมีค่ามากขึ้นเรื่อยๆ (ขนาดเซตของคำแนะนำที่ลดลงทำให้ค่าความถูกต้องสูงขึ้น ในขณะที่มีกฎความสัมพันธ์ที่ถูกต้องเท่าเดิมหรือไม่ลดลงจึงทำให้ค่าเรียกคืนเท่าเดิม) แสดงให้เห็นว่าการค้นหากฎความสัมพันธ์ในสถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้น กฎความสัมพันธ์ที่ถูกต้องจริงๆนั้นมันจะเป็นกฎความสัมพันธ์ที่อยู่ในอันดับต้นๆในทั้ง 2 ตัวแบบ

จากตารางและกราฟข้างต้นแสดงให้เห็นว่าค่าเอฟเมสเซอร์หรือค่าประสิทธิภาพของการทดสอบข้างต้นนั้นมีค่าแตกต่างกันทั้ง 2 ตัวแบบเมื่อปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด แต่ไม่สามารถสรุปได้ว่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ของตัวแบบทั้ง 2 เมื่อปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้นแตกต่างกันอย่างมีนัยสำคัญ ผู้วิจัยจึงจำเป็นต้องวิเคราะห์ผลการทดสอบความแตกต่างกันอย่างมีนัยสำคัญ ดังรายละเอียดต่อไปนี้

การวิเคราะห์การแจกแจงปกติ

เนื่องจากผู้วิจัยต้องการเปรียบเทียบประสิทธิภาพของการค้นหากฎความสัมพันธ์ของ 2 ตัวแบบสำหรับสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด โดยทดลองปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ ดังนั้น ผู้วิจัยจะตรวจสอบว่าค่าเอฟเมเซอร์ทั้งหมดมีการแจกแจงแบบปกติหรือไม่ โดยตั้งสมมติฐานของการทดสอบค่าเอฟเมเซอร์แต่ละการทดสอบมีการแจกแจงแบบปกติหรือไม่ภายใต้สมมติฐานทางสถิติ ดังนี้

1) ทดสอบการแจกแจงของข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การนำทาง

   H_0: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การนำทาง มีการแจกแจงแบบปกติ

   H_1: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การนำทาง ไม่มีการแจกแจงแบบปกติ

2) ทดสอบการแจกแจงของข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การนำทาง

   H_0: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การนำทาง มีการแจกแจงแบบปกติ

   H_1: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การนำทาง ไม่มีการแจกแจงแบบปกติ

3) ทดสอบการแจกแจงของข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การนำทาง

   H_0: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การนำทาง มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การนำทาง ไม่มีการแจกแจงแบบปกติ

4) ทดสอบการแจกแจงของข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การนำทาง

H_0: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การนำทาง มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การนำทาง ไม่มีการแจกแจงแบบปกติ

5) ทดสอบการแจกแจงของข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ในสถานการณ์การนำทาง

H_0: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ในสถานการณ์การนำทาง มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ในสถานการณ์การนำทาง ไม่มีการแจกแจงแบบปกติ

6) ทดสอบการแจกแจงของข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ในสถานการณ์การนำทาง

H_0: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ในสถานการณ์การนำทาง มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ในสถานการณ์การนำทาง ไม่มีการแจกแจงแบบปกติ

7) ทดสอบการแจกแจงของข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

H_0: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่มีการแจกแจงแบบปกติ

8) ทดสอบการแจกแจงของข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

H_0: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่มีการแจกแจงแบบปกติ

9) ทดสอบการแจกแจงของข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

H_0: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่มีการแจกแจงแบบปกติ

10) ทดสอบการแจกแจงของข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

H_0: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่มีการแจกแจงแบบปกติ

11) ทดสอบการแจกแจงของข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

H_0: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 1 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่มีการแจกแจงแบบปกติ

12) ทดสอบการแจกแจงของข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

H_0: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ตัวแบบที่ 2 โดยสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่มีการแจกแจงแบบปกติ

ในการตรวจสอบการแจกแจงของข้อมูลว่าเป็นแบบปกติโดยใช้สถิติทดสอบนั้น มีสถิติทดสอบที่ใช้คือ Kolmogorov-Sminov สำหรับหน่วยทดลองมากกว่า 50 หน่วย และ Shapiro-Wilk สำหรับหน่วยทดลองน้อยกว่า 50 หน่วย โดยจะยอมรับสมมติฐาน H_0 เมื่อค่า Sig. มีค่ามากกว่าค่า α ซึ่งกำหนดให้เท่ากับ 0.05 ผลการทดสอบแสดงดังตารางต่อไปนี้

ตารางที่ 4-17 แสดงค่าสถิติทดสอบการแจกแจงปกติ (Normality Test) ของค่าเอฟเมเซอร์ของการค้นกฎความสัมพันธ์ทั้ง 2 ตัวแบบ โดยปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด

สถานการณ์	จน.กฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำ	ตัวแบบที่	Kolmogorov-Smirnov			Shapiro-Wilk		
			Statistic	df	Sig.	Statistic	df	Sig.
สถานการณ์การนำทาง	3	1	.127	451	.000	.886	451	.000
		2	.130	451	.000	.929	451	.000
	5	1	.087	451	.000	.952	451	.000
		2	.085	451	.000	.957	451	.000
	7	1	.073	451	.000	.960	451	.000
		2	.081	451	.000	.960	451	.000
สถานการณ์การป้องกันข้อผิดพลาด	3	1	.372	451	.000	.700	451	.000
		2	.409	451	.000	.674	451	.000
	5	1	.350	451	.000	.732	451	.000
		2	.361	451	.000	.752	451	.000
	7	1	.311	451	.000	.798	451	.000
		2	.340	451	.000	.779	451	.000

ผลการทดสอบในตารางที่ 4-17 ข้างต้นพบว่าค่า Sig. ของตัวแปรค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดเป็นดังนี้

1) สำหรับสถานการณ์การนำทาง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0
2) สำหรับสถานการณ์การนำทาง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0
3) สำหรับสถานการณ์การนำทาง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0
4) สำหรับสถานการณ์การนำทาง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0
5) สำหรับสถานการณ์การนำทาง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0
6) สำหรับสถานการณ์การนำทาง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0
7) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0
8) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่ามากกว่าค่าระดับนัยสำคัญ $\alpha = 0.05$ ดังนั้นจึงปฏิเสธสมมติฐาน H_0
9) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก มีค่า Sig.

เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

10) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่ามากกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

11) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

12) สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่ามากกว่าค่าระดับนัยสำคัญ α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

ดังนั้นสรุปได้ว่าการแจกแจงของตัวแปรค่าประสิทธิภาพของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบโดยปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ ในทั้ง 2 สถานการณ์นั้นไม่เป็นแบบปกติ

ผลการทดสอบ

จากการวิเคราะห์การแจกแจงข้อมูลข้างต้นพบว่าค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดไม่มีการแจกแจงแบบปกติ ดังนั้นผู้วิจัยจึงเลือกใช้การทดสอบสมมติฐานแบบไม่อิงพารามิเตอร์ (Non Parametic Test) นั่นคือสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) กับการทดสอบต่อไปนี้ โดยที่จะปฏิเสธสมมติฐาน H_0 ได้เมื่อถ้าค่า Sig. (Significance) ที่คำนวณได้น้อยกว่า 0.05 และค่าสถิติ Z มากกว่า 0 โดยที่ผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks)

การวิเคราะห์เปรียบเทียบนี้เป็นการวิเคราะห์เปรียบเทียบค่าเอฟเมเซอร์ระหว่างการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 กับการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 โดยปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด ซึ่งสามารถตั้งสมมุติฐานได้ดังนี้

1) วิเคราะห์เปรียบเทียบค่าเอฟเมเซอร์ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ทั้ง 2 ตัวแบบ ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การนำทาง ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การนำทาง

M_2 คือ ค่ามัธยฐานของค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การนำทาง

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

2) วิเคราะห์เปรียบเทียบค่าเอฟเมเซอร์ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบ ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การนำทาง ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การนำทาง

M_2 คือ ค่ามัธยฐานของค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การนำทาง

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

3) วิเคราะห์เปรียบเทียบค่าเอฟเมเซอร์ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบ ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การนำทาง ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การนำทาง

M_2 คือ ค่ามัธยฐานของค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การนำทาง

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

4) วิเคราะห์เปรียบเทียบค่าเอฟเมเซอร์ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบ ที่สร้างเซตของคำแนะนำ

จากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

M_2 คือ ค่ามัธยฐานของค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

5) วิเคราะห์เปรียบเทียบค่าเอฟเมเซอร์ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบ ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

M_2 คือ ค่ามัธยฐานของค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

6) วิเคราะห์เปรียบเทียบค่าเอฟเมเซอร์ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบ ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ว่ามีความแตกต่างกันหรือไม่

กำหนดให้ M_1 คือ ค่ามัธยฐานของค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

M_2 คือ ค่ามัธยฐานของค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

$H_0 : M_2 \leq M_1$

$H_1 : M_2 > M_1$

ผู้วิจัยใช้การทดสอบสมมติฐานโดยสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) กับการทดสอบต่อไปนี้ ผลการทดสอบแสดงดังตารางต่อไปนี้

ตารางที่ 4-18 แสดงสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ของค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบ โดยปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ

	ค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ตัวแบบที่ 2 - ค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ตัวแบบที่ 1					
	สถานการณ์การนำทาง			สถานการณ์การป้องกันข้อผิดพลาด		
จน.กฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำ	3	5	7	3	5	7
Z	-4.795[a]	-5.815[a]	-3.973[a]	-8.766[b]	-8.280[b]	-9.852[b]
Asymp. Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000

a. Based on negative ranks.

b. Based on positive ranks.

จากตารางที่ 4-18 ได้ผลทดสอบดังนี้

1) ในสถานการณ์การนำทาง ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ได้สถิติทดสอบค่า Z เท่ากับ -4.795 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียว ดังนั้นค่า Sig. จึงเท่ากับ 0.000 / 2 = 0.000 ซึ่งน้อยกว่าค่า α = 0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางลบ (Based on negative ranks) ดังนั้นจึงสามารถปฏิเสธ

สมมติ H_0 ได้ นั่นคือค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มากกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรกในสถานการณ์การนำทาง ที่ระดับนัยสำคัญ 0.05

2) ในสถานการณ์การนำทาง ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ได้สถิติทดสอบค่า Z เท่ากับ -5.815 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียว ดังนั้นค่า Sig. จึงเท่ากับ 0.000 / 2 = 0.000 ซึ่งน้อยกว่าค่า α = 0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางลบ (Based on negative ranks) ดังนั้นจึงสามารถปฏิเสธสมมติ H_0 ได้ นั่นคือค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มากกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรกในสถานการณ์การนำทาง ที่ระดับนัยสำคัญ 0.05

3) ในสถานการณ์การนำทาง ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ได้สถิติทดสอบค่า Z เท่ากับ -3.973 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียว ดังนั้นค่า Sig. จึงเท่ากับ 0.000 / 2 = 0.000 ซึ่งน้อยกว่าค่า α = 0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางลบ (Based on negative ranks) ดังนั้นจึงสามารถปฏิเสธสมมติ H_0 ได้ นั่นคือค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มากกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรกในสถานการณ์การนำทาง ที่ระดับนัยสำคัญ 0.05

4) ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรก ได้สถิติทดสอบค่า Z เท่ากับ -8.766 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 0.000 / 2 = 0.000 ซึ่งน้อยกว่าค่า α = 0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) ดังนั้นจึงไม่สามารถปฏิเสธสมมติ H_0 ได้ นั่นคือค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ไม่ต่างกันหรือน้อยกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 7 อันดับแรกในสถานการณ์การป้องกันการเกิดข้อผิดพลาดที่ระดับนัยสำคัญ 0.05

5) ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก ได้สถิติทดสอบค่า Z เท่ากับ -8.280 ซึ่งน้อยกว่า 0 และจากการ

ตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 0.000 / 2 = 0.000 ซึ่งน้อยกว่าค่า α = 0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) ดังนั้นจึงไม่สามารถปฏิเสธสมมติ H_0 ได้ นั่นคือค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ไม่ต่างกันหรือน้อยกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรกในสถานการณ์การป้องกันการเกิดข้อผิดพลาดที่ระดับนัยสำคัญ 0.05

6) ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก ได้สถิติทดสอบค่า Z เท่ากับ -9.852 ซึ่งน้อยกว่า 0 และจากการตั้งสมมติฐานเป็นแบบทางเดียวดังนั้นค่า Sig. จึงเท่ากับ 0.000 / 2 = 0.000 ซึ่งน้อยกว่าค่า α = 0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) ดังนั้นจึงไม่สามารถปฏิเสธสมมติ H_0 ได้ นั่นคือค่าเอฟเมเซอร์ของการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ไม่ต่างกันหรือน้อยกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ที่สร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรกในสถานการณ์การป้องกันการเกิดข้อผิดพลาดที่ระดับนัยสำคัญ 0.05

สรุปผลการทดสอบ

จากการวิเคราะห์ผลการเปรียบเทียบค่าเอฟเมเซอร์หรือประสิทธิภาพของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบ ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด ของผลการทดสอบนี้และผลการทดสอบหลักในหัวข้อ 4.3 ผู้วิจัยสามารถสรุปได้ว่า ในสถานการณ์การนำทาง การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มีค่าเอฟเมเซอร์เฉลี่ยสูงกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ไม่ว่าจะกำหนดจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเท่าใดก็ตาม โดยการสร้างเซตของคำแนะนำจากกฎความสัมพันธ์จำนวนต่างๆกันที่ให้ผลต่างค่าเอฟเมเซอร์หรือประสิทธิภาพของการค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบสูงที่สุดในสถานการณ์การนำทางคือ การสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 5 อันดับแรก 3 อันดับแรก 10 อันดับแรก และ 7 อันดับแรก ตามลำดับ สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาด การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มีค่าเอฟเมเซอร์เฉลี่ยไม่ต่างกันหรือน้อยกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ไม่ว่าจะกำหนดจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเท่าไรก็ตาม โดยการสร้างเซตของคำแนะนำจากกฎความสัมพันธ์จำนวนต่างๆกันที่ให้ผลต่างค่าเอฟเมเซอร์หรือประสิทธิภาพของการค้นหากฎ

ความสัมพันธ์ทั้ง 2 ตัวแบบมากที่สุดในสถานการณ์การป้องกันการเกิดข้อผิดพลาดคือ การสร้างเซตของคำแนะนำจากกฎความสัมพันธ์ 3 อันดับแรก 5 อันดับแรก 7 อันดับแรก และ 10 อันดับแรก ตามลำดับ

สรุปผลการทดลองข้างต้นสามารถตอบข้อสังเกตของผู้วิจัยที่ตั้งไว้ได้คือ ในสถานการณ์การนำทาง การค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบมีกฎความสัมพันธ์ที่เป็นผลบวกลวงปะปนอยู่ในอันดับต้นๆทั้งคู่ แต่กฎความสัมพันธ์ที่ได้การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในอันดับต้นๆนั้นมีกฎความสัมพันธ์ที่เป็นผลบวกลวงปะปนอยู่มากกว่ากฎความสัมพันธ์ที่ได้การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในอันดับต้นๆ แต่สำหรับในสถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้น การค้นหากฎความสัมพันธ์ทั้ง 2 ตัวแบบมีกฎความสัมพันธ์ที่เป็นผลบวกลวงปะปนอยู่ในอันดับต้นๆในปริมาณที่น้อยทั้งคู่ โดยที่กฎความสัมพันธ์ที่ได้การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 ในอันดับต้นๆนั้นมีกฎความสัมพันธ์ที่เป็นผลบวกลวงปะปนอยู่น้อยกว่ากฎความสัมพันธ์ที่ได้การค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 ในอันดับต้นๆนั้นมีกฎความสัมพันธ์ที่เป็นผลบวกลวง

ผู้วิจัยสังเกตเห็นว่าในสถานการณ์การนำทางนั้น มักจะเกิดกฎความสัมพันธ์ที่เป็นผลบวกลวงขึ้นในอันดับต้นๆเสมอไม่ว่าใช้ตัวแบบที่ 1 หรือตัวแบบที่ 2 แต่สำหรับในสถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้น กฎความสัมพันธ์ในอันดับต้นๆนั้นค่อนข้างถูกต้องแม่นยำและมีกฎความสัมพันธ์ที่เป็นผลบวกลวงปะปนอยู่น้อยไม่ว่าใช้ตัวแบบที่ 1 หรือตัวแบบที่ 2 ด้วยเหตุนี้ผู้วิจัยจึงสนใจที่จะวิเคราะห์ข้อสังเกตดังกล่าวในหัวข้อถัดไป

4.4.5 การวิเคราะห์ค่าประเมินระดับความน่าสนใจของกฎความสัมพันธ์ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด

จากการทดสอบเพิ่มเติมในหัวข้อที่ 4.4.4 ทำให้ผู้วิจัยทราบว่าในสถานการณ์การนำทางนั้น มักจะเกิดกฎความสัมพันธ์ที่เป็นผลบวกลวงขึ้นในอันดับต้นๆเสมอไม่ว่าใช้ตัวแบบที่ 1 หรือตัวแบบที่ 2 แต่สำหรับในสถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้น กฎความสัมพันธ์ในอันดับต้นๆนั้นค่อนข้างถูกต้องแม่นยำและมีกฎความสัมพันธ์ที่เป็นผลบวกลวงปะปนอยู่ในอันดับต้นๆน้อยไม่ว่าจะใช้ตัวแบบที่ 1 หรือตัวแบบที่ 2 ผู้วิจัยเห็นว่าสาเหตุที่เป็นเช่นนั้นเพราะการคำนวณค่าความเชื่อมั่นใหม่นั้นแตกต่างจากการคำนวณค่าความเชื่อมั่นใหม่ตรงที่ค่าความเชื่อมั่นใหม่มีการ

นำเอาค่าความน่าจะเป็นในการไม่พบเซตรายการที่ตามมาเข้ามาพิจารณาด้วย นอกจากนั้นผู้วิจัยเห็นว่าสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดมีลักษณะเฉพาะตัวที่แตกต่างกันที่ทำให้ผลของการนำเอาค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่มาก่อนแต่ไม่มีเซตรายการที่ตามมาเข้ามาพิจารณามีผลแตกต่างกัน ผู้วิจัยจึงสนใจที่จะวิเคราะห์การคำนวณค่าประเมินระดับความน่าสนใจของกฎความสัมพันธ์ทั้ง 2 ค่า รวมถึงการวิเคราะห์ลักษณะเฉพาะของสถานการณ์ทั้ง 2 สถานการณ์ด้วย

ตารางที่ 4-19 แสดงการเปรียบเทียบสูตรคำนวณและพิสัยของค่าความเชื่อมั่นและค่าความเชื่อมั่นใหม่

	ค่าความเชื่อมั่น	ค่าความเชื่อมั่นใหม่
สูตรคำนวณ	$Conf(X \rightarrow Y) = \frac{P(X \text{ and } Y)}{P(X)}$	$NConf(X \rightarrow Y) = \frac{P(X \text{ and } Y)}{P(Y)} - \frac{P(X \text{ and } \bar{Y})}{P(\bar{Y})}$
พิสัย	[0,1]	[-1,1]

จากตารางข้างบนแสดงให้เห็นว่าความแตกต่างของค่าความเชื่อมั่นกับค่าความเชื่อมั่นใหม่คือการนำเอาค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่มาก่อนแต่ไม่มีเซตรายการที่ตามมาเข้ามาพิจารณาด้วย ในสูตรคำนวณของค่าความเชื่อมั่นใหม่จะเห็นว่าถ้าพจน์หลังของสูตรมีค่ามากจะทำให้ค่าความเชื่อมั่นใหม่มีค่าน้อย นั่นคือถ้าค่าความน่าจะเป็นในการพบเซตรายการที่มาก่อนในทรานแซคชั่นที่ไม่มีเซตรายการที่ตามมาอยู่มากค่ามากแล้วกฎความสัมพันธ์นั้นจะมีค่าความเชื่อมั่นใหม่น้อย ในขณะที่สูตรคำนวณของค่าความเชื่อมั่นไม่ได้นำเอาค่าความน่าจะเป็นในการไม่พบเซตรายการที่ตามมาเข้ามาพิจารณาด้วย แต่กฎความสัมพันธ์ที่มีค่าความเชื่อมั่นมากอาจจะมีค่าความน่าจะเป็นในการพบเซตรายการที่มาก่อนในทรานแซคชั่นที่ไม่มีเซตรายการที่ตามมาอยู่ที่มากก็เป็นไปได้ ซึ่งถ้าเป็นเช่นนั้นกฎความสัมพันธ์ดังกล่าวก็คือกฎความสัมพันธ์ที่ถูกแนะนำเป็นอันดับต้นๆแต่จริงๆแล้วเป็นผลบวกลวงนั่นเอง (Liu et al., 2008)

ตารางที่ 4-20 แสดงการเปรียบเทียบขนาดของเซตรายการที่มาก่อนและเซตรายการที่ตามมาของกฎความสัมพันธ์ขนาด n ในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด

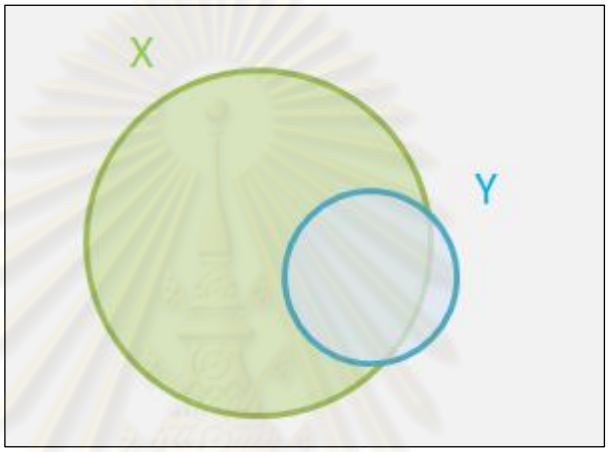
	สถานการณ์การนำทาง	สถานการณ์การป้องกันการเกิดข้อผิดพลาด
ขนาดของเซตรายการที่มาก่อน	1	n-1
ขนาดของเซตรายการที่ตามมา	n-1	1

จากตารางข้างบนแสดงให้เห็นว่าความแตกต่างของกฎความสัมพันธ์สถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดก็คือ ในสถานการณ์การนำทางขนาดของเซตรายการที่ตามมาจะมีขนาดใหญ่ขึ้นเป็นเชิงเส้นเมื่อขนาดของกฎความสัมพันธ์ใหญ่ขึ้นโดยที่มีขนาดของเซตรายการที่มาก่อนเป็น 1 เสมอ สำหรับสถานการณ์การป้องกันการเกิดข้อผิดพลาดขนาดของเซตรายการที่ตามมาเป็น 1 เสมอโดยที่มีขนาดของเซตรายการที่มาก่อนจะมีขนาดใหญ่ขึ้นเป็นเชิงเส้นเมื่อขนาดของกฎความสัมพันธ์ใหญ่ขึ้น

ในปีค.ศ. 2004 Lenca และคณะ (Lenca et al, 2004) ก็เสนอคุณสมบัติ 5 ข้อที่ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ควรจะมี หลังจากนั้นในปี ค.ศ. 2007 Lenca และคณะได้ปรับปรุงคุณสมบัติทั้ง 5 ข้อนั้นเล็กน้อยเพื่อให้มีความยืดหยุ่นมากขึ้น (Lenca et al, 2007) คุณสมบัติทั้ง 5 ข้อของ Lenca และคณะนั้นได้แก่คุณสมบัติ Q1 Q2 Q3 Q4 และ Q5 ที่อธิบายไว้ในบทที่ 2 คุณสมบัติ Q1 นั้นกล่าวไว้ว่า กฎความสัมพันธ์นั้นมีความน่าสนใจสูงที่สุด เมื่อค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่มาก่อนแต่ไม่มีเซตรายการที่ตามมาในฐานข้อมูลเท่ากับ 0 และค่าประเมินความน่าสนใจของกฎความสัมพันธ์ควรจะเป็นค่าคงที่ค่าใดค่าหนึ่งหรือเป็นค่าอนันต์เพื่อสื่อความหมายอย่างชัดเจนว่ากฎความสัมพันธ์นั้นน่าสนใจสูงที่สุดด้วย

จากคุณสมบัตินี้จะเห็นว่าการประเมินว่ากฎความสัมพันธ์ใดมีความน่าสนใจสูงนั้น ไม่สามารถพิจารณาเพียงแค่ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่มาก่อนและเซตรายการที่ตามมาเท่านั้น แต่จะต้องพิจารณาที่ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่มาก่อนแต่ไม่มีเซตรายการที่ตามมาควบคู่ด้วย หรือเพียงพิจารณาที่ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่มาก่อนแต่ไม่มีเซตรายการที่ตามมาเป็นหลัก สาเหตุที่

เป็นเช่นนั้นก็เพราะในบางกรณีที่กฎความสัมพันธ์มีค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่มาก่อนและเซตรายการที่ตามมามีค่าสูง แต่ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่มาก่อนแต่ไม่มีเซตรายการที่ตามมามีค่าสูงกว่า หรือกล่าวคือในทรานแซคชั่นที่มีเซตรายการที่มาก่อนอยู่มักจะไม่มีเซตรายการที่ตามมาเป็นส่วนใหญ่ แต่ในทรานแซคชั่นที่มีเซตรายการที่ตามมาอยู่มักจะมีเซตรายการที่มาก่อนเป็นส่วนใหญ่ ตามรูปด้านล่างนี้



รูปที่ 4-5 แสดงที่กฎความสัมพันธ์มีค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่มาก่อนและเซตรายการที่ตามมามีค่าสูง แต่ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่มาก่อนแต่ไม่มีเซตรายการที่ตามมามีค่าสูงกว่า

กฎความสัมพันธ์ที่พบว่า ทรานแซคชั่นส่วนใหญ่นั้นที่มีเซตรายการที่มาก่อนอยู่มักจะไม่มีเซตรายการที่ตามมา สามารถกล่าวได้ว่าเป็นกฎความสัมพันธ์เซตรายการที่มาก่อนกับเซตรายการที่ตามมามีความสัมพันธ์เชิงลบต่อกัน ถ้ากฎความสัมพันธ์นี้ถูกจัดอันดับว่ามีความน่าสนใจในอันดับต้นๆ กฎความสัมพันธ์นี้ก็คือผลบวกลวงนั่นเอง (Liu et al., 2008) ดังนั้นการใช้ค่าประเมินความน่าสนใจที่ขึ้นกับค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่มาก่อนและเซตรายการที่ตามมาเป็นหลักเพียงอย่างเดียว อย่างเช่น ค่าสนับสนุน ค่าความเชื่อมั่น ค่าลิฟท์ เป็นต้น จึงมีโอกาสที่จะทำให้เกิดกฎความสัมพันธ์ที่เป็นผลบวกลวงได้ง่าย

กรณีที่ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่มาก่อนและเซตรายการที่ตามมามีค่าสูง แต่ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่มาก่อนแต่ไม่มีเซตรายการที่ตามมามีค่าสูงกว่าดังรูปด้านบนนั้น จะมีโอกาสเกิดขึ้นได้ง่ายเมื่อพบว่าค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่มาก่อนมีสูงมากๆ ในขณะที่ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่ตามมามีน้อยกว่ามากๆ เมื่อพิจารณาที่รูปแบบกฎความสัมพันธ์

ในสถานการณ์การนำทาง จะเห็นว่าเซตรายการที่มาก่อนที่มีขนาดเท่ากับ 1 เสมอนั้นมีโอกาสสูงที่ค่าความน่าเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่มาก่อนมีค่าสูงมากๆ ในขณะที่เซตรายการที่ตามมามีขนาดเพิ่มขึ้นเป็นเชิงเส้นตามขนาดของกฎความสัมพันธ์นั้นมีโอกาสสูงที่ค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่ตามมาจะมีน้อยมากๆ ด้วยเหตุนี้กฎความสัมพันธ์ในสถานการณ์การนำทางที่ใช้ค่าประเมินความน่าสนใจของกฎความสัมพันธ์ที่ขึ้นกับค่าความน่าจะเป็นในการพบทรานแซคชั่นที่มีเซตรายการที่มาก่อนและเซตรายการที่ตามมาเป็นหลักเพียงอย่างเดียวจึงมีโอกาสเกิดผลบวกลวงได้สูง

เมื่อโอกาสเกิดกฎความสัมพันธ์ที่เป็นผลบวกลวงในสถานการณ์การนำทางมีมาก การนำเอาค่าความน่าจะเป็นในการไม่พบเซตรายการที่ตามมาเข้ามาพิจารณาด้วยของค่าความเชื่อมั่นใหม่จึงสามารถลดระดับความน่าสนใจของกฎความสัมพันธ์ที่เป็นผลบวกลวงได้มากเช่นกัน สำหรับในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ค่าความน่าจะเป็นในการไม่พบเซตรายการที่ตามมาที่น้อยส่งผลกฎความสัมพันธ์ที่มีค่าความเชื่อมั่นหรือค่าเชื่อมั่นใหม่สูงมีกฎความสัมพันธ์ที่เป็นผลบวกลวงปะปนอยู่น้อย การนำเอาค่าความน่าจะเป็นในการไม่พบเซตรายการที่ตามมาเข้ามาพิจารณาด้วยของค่าความเชื่อมั่นใหม่จึงสามารถช่วยลดระดับความน่าสนใจของกฎความสัมพันธ์ที่เป็นผลบวกลวงได้น้อยด้วย

4.4.6 สรุปผลการศึกษาเพิ่มเติม

การทดสอบการเปรียบเทียบค่าประสิทธิภาพหรือค่าเอฟเมเซอร์ (F-measure) ค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาดโดยการสร้างเซตของคำแนะนำแบบปกติกับการสร้างเซตของคำแนะนำจากการยูเนียน (Union) ของรายการการเปลี่ยนแปลงแก้ไขที่อยู่ในเซตรายการที่ตามมาของกฎความสัมพันธ์อันดับสูงสุดจำนวน 10 รายการแรก ทำให้ผู้วิจัยทราบว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 มีประสิทธิภาพเพิ่มขึ้นเมื่อพิจารณาเพียงกฎความสัมพันธ์ที่อยู่ในอันดับต้นๆ ผู้วิจัยจึงศึกษาเพิ่มเติมโดยการทดสอบการเปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบโดยปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆคือ 7 อันดับแรก 5 อันดับแรกและ 3 อันดับแรก จากการทดสอบนี้ทำให้ผู้วิจัยทราบถึงแนวโน้มการลดลงหรือเพิ่มของประสิทธิภาพของการค้นหากฎความสัมพันธ์ของตัวแบบทั้ง 2 ที่แตกต่างกันในสถานการณ์การนำทางและสถานการณ์การ

ป้องกันการเกิดข้อผิดพลาดเมื่อมีการปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ ผู้วิจัยจึงวิเคราะห์สาเหตุของแนวโน้มที่แตกต่างกันใน 2 สถานการณ์นั้น

การทดสอบเพิ่มทั้งหมดนี้เพื่อศึกษาหาสาเหตุของผลการทดสอบหลังของงานวิจัยนี้นั่นคือการเปรียบเทียบประสิทธิภาพของการประยุกต์เทคนิคการค้นหากฎความสัมพันธ์บนข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นและตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ ผลการทดสอบแสดงให้เห็นว่าตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ช่วยเพิ่มประสิทธิภาพให้กับการประยุกต์เทคนิคการค้นหากฎความสัมพันธ์บนข้อมูลซอฟต์แวร์อาร์ไคฟ์ได้เฉพาะในสถานการณ์การนำทางเท่านั้น และผลการทดสอบเพิ่มแสดงให้เห็นว่าสาเหตุที่ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ช่วยเพิ่มประสิทธิภาพในสถานการณ์การนำทางเนื่องมาจากการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นนั้นสามารถทำให้เกิดกฎความสัมพันธ์ที่เป็นผลบวกลวงได้มาก แต่การใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่สามารถลดการเกิดผลบวกลวงเหล่านั้นได้ แต่สำหรับในสถานการณ์การป้องกันการเกิดข้อผิดพลาดนั้นการใช้ตัวแบบทั้ง 2 ให้กฎความสัมพันธ์ที่เป็นผลบวกลวงค่อนข้างน้อย และการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นให้ค่าความถูกต้องและค่าเรียกคืนที่ดีกว่าการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่จึงทำให้ประสิทธิภาพของการค้นหากฎความสัมพันธ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นดีกว่าการค้นหากฎความสัมพันธ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่

บทที่ 5
สรุปผลการวิจัย

5.1 บทนำ

บทนี้จะแสดงการสรุปผลของงานวิจัยและปัญหาที่เกิดขึ้นในการวิจัย สุดท้ายเป็นข้อเสนอแนะของงานวิจัย เพื่อปรับเปลี่ยนรูปแบบของงานวิจัยหรือพัฒนาการทดลองให้มีประสิทธิภาพยิ่งขึ้น

5.2 การออกแบบการวิจัยและลักษณะของข้อมูลที่นำมาใช้

งานวิจัยนี้เป็นการวิจัยเชิงทดลอง (Experimental Research) โดยใช้ทรานแซคชั่นของการเปลี่ยนแปลงแก้ไขและข้อสอบถามที่สร้างมาจากข้อมูลซอฟต์แวร์อาร์ไคฟ์ของโครงการพัฒนาซอฟต์แวร์ทางการบัญชีชื่อเคมายมันนี่ (KMyMoney) มาทดสอบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นและประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008) ซึ่งมีทรานแซคชั่นของการเปลี่ยนแปลงแก้ไขทั้งหมด 5458 ทรานแซคชั่นและเลือกทรานแซคชั่นชุดทดสอบทั้งหมด 60 ทรานแซคชั่น (ทรานแซคชั่นชุดทดสอบแสดงอยู่ในภาคผนวก ก) ซึ่งทำให้ได้ข้อสอบถามที่ใช้ในการทดสอบทั้งหมด 962 ข้อสอบถาม

การทดสอบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ในงานวิจัยนี้เป็นการทดสอบประสิทธิภาพในแง่ของการประยุกต์การทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ โดยใช้วิธีการทดสอบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ที่เสนอโดย Zimmermann และคณะในปี 2005 (Zimmermann et al., 2005) ซึ่งแบ่งการทดสอบประสิทธิภาพออกเป็น 3 สถานการณ์ของการพัฒนาซอฟต์แวร์ได้แก่ 1) สถานการณ์การนำทาง (Navigation) 2) สถานการณ์การป้องกันการเกิดข้อผิดพลาด (Error Prevention) และ 3) สถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว (Closure)

5.3 สรุปผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นกับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008) โดยแบ่งการทดสอบเปรียบเทียบประสิทธิภาพของตัวแบบทั้ง 2 ตัวแบบเป็น 3 สถานการณ์ดังนี้

1) เปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นกับตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ ในสถานการณ์การนำทาง
2) เปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นกับตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด
3) เปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นกับตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

การทดสอบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ ผู้วิจัยใช้ข้อมูลซอฟต์แวร์อาร์ไคฟว์ของโครงการพัฒนาซอฟต์แวร์ทางการบัญชีชื่อเคมายมันนี่ (KMyMoney) กับเครื่องมือทดสอบที่ผู้วิจัยพัฒนาขึ้นมา โดยประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ในงานวิจัยนี้ใช้ค่าเอฟเมเซอร์สำหรับการทดสอบในสถานการณ์การนำทางและสถานการณ์การป้องกันการเกิดข้อผิดพลาด และใช้ค่าผลสะท้อนกลับสำหรับการทดสอบในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว ผลการทดสอบประสิทธิภาพและการเปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ 2 ตัวแบบใน 3 สถานการณ์แสดงไว้ในภาคผนวก ค สามารถสรุปได้ดังนี้

5.3.1 ผลการเปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นกับตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ ในสถานการณ์การนำทาง

การทดสอบในสถานการณ์การนำทางนี้ใช้ข้อสอบถามทั้งหมด 451 ข้อสอบถาม ผลการทดสอบผู้วิจัยพบว่าค่าประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่มากกว่าการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น และผลการทดสอบความแตกต่างค่าประสิทธิภาพของทั้ง 2 กลุ่มด้วยสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) สรุปได้ว่าการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่มีประสิทธิภาพที่ดีกว่าการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นในสถานการณ์การนำทาง กล่าวคือการนำตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่มาประยุกต์สามารถเพิ่มประสิทธิภาพให้กับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์สำหรับระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ในสถานการณ์การนำทางได้อย่างมีนัยสำคัญ

ผู้วิจัยได้ทดสอบเพิ่มเติมเพื่อเปรียบเทียบค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบ ผลการทดสอบผู้วิจัยพบว่าการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ให้ค่าความถูกต้องเฉลี่ย (Average Precision) ที่ต่ำกว่าและให้ค่าเรียกคืนเฉลี่ย (Average Recall) ที่สูงกว่าการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นอย่างมีนัยสำคัญด้วย เนื่องจากผู้วิจัยสังเกตเห็นว่าขนาดเซตของคำแนะนำที่ได้มาจากการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ที่มีขนาดใหญ่กว่าเซตของคำแนะนำที่ได้มาจากการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นเสมอ (ค่าความถูกต้องแปรผกผันกับขนาดเซตของคำแนะนำ) กฎความสัมพันธ์ที่ได้มาจากการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นที่อยู่ในอันดับต้นๆจะเป็นกฎความสัมพันธ์ที่มีเซตรายการที่ตามมาเพียง 1 รายการเสมอ (ค่าความเชื่อมั่นของกฎความสัมพันธ์แปรผันตรงกับค่าสนับสนุนของกฎความสัมพันธ์ และกฎความสัมพันธ์ที่มีเซตรายการที่ตามมาเพียง 1 รายการมักจะทำให้ค่าสนับสนุนของกฎ

ความสัมพันธ์มีค่ามากกว่ากฎความสัมพันธ์ที่มีเซตรายการที่ตามมาหลายรายการ) สำหรับกฎความสัมพันธ์ที่ได้มาจากการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ที่มักจะมีเซตรายการที่ตามมีขนาดใหญ่นั้น ผู้วิจัยเห็นว่าอาจเป็นเพราะข้อมูลซอฟต์แวร์อาร์ไคฟ์ของโครงการพัฒนาซอฟต์แวร์เคมายมันนี่ (KMyMoney) นั้นแต่ละแฟ้มข้อมูลในโครงการมีความสัมพันธ์เชื่อมโยงกันมาก จึงทำให้เกิดกฎความสัมพันธ์ที่ได้มาจากการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ 10 อันดับแรกมีความหลากหลาย

นอกจากนั้นผู้วิจัยยังได้ทดสอบเพิ่มเติมโดยการปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ ผลการทดสอบซึ่งพบว่าการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ให้ประสิทธิภาพที่ดีกว่าการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นเสมอในสถานการณ์การนำทาง ไม่ว่าจะปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าใดๆก็ตาม

5.3.2 ผลการเปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นกับตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ ในสถานการณ์การป้องกันการเกิดข้อผิดพลาด

การทดสอบในสถานการณ์การป้องกันการเกิดข้อผิดพลาดนี้ใช้ข้อสอบถามทั้งหมด 451 ข้อสอบถาม ผลการทดสอบผู้วิจัยพบว่าค่าประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ต่ำกว่าการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น และผลการทดสอบความแตกต่างค่าประสิทธิภาพของทั้ง 2 กลุ่มด้วยสถิติทดสอบเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) สรุปได้ว่าการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่มีประสิทธิภาพที่ด้อยกว่าการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นในสถานการณ์การป้องกันการเกิดข้อผิดพลาด กล่าวคือการนำตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่มาประยุกต์ใช้ไม่สามารถเพิ่มประสิทธิภาพให้กับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎ

ความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์สำหรับระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ในสถานการณ์การป้องกันการเกิดข้อผิดพลาดได้

ผู้วิจัยได้ทดสอบเพิ่มเติมเพื่อเปรียบเทียบค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ของทั้ง 2 ตัวแบบ ผลการทดสอบแสดงให้เห็นว่าการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ให้ค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) ที่ต่ำกว่าการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นอย่างมีนัยสำคัญด้วย โดยสาเหตุของผลการทดสอบนี้เป็นเช่นเดียวกับที่ได้กล่าวไปในหัวข้อที่แล้ว

นอกจากนั้นผู้วิจัยยังได้ทดสอบเพิ่มเติมโดยการปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าต่างๆ ผลการทดสอบผู้วิจัยพบว่าการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ให้ประสิทธิภาพที่ด้อยกว่าการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นเสมอในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ไม่ว่าจะปรับจำนวนของกฎความสัมพันธ์ที่นำมาสร้างเป็นเซตของคำแนะนำเป็นค่าใดๆก็ตาม

5.3.3 ผลการเปรียบเทียบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นกับตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

การทดสอบในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วนี้ใช้ข้อสอบถามทั้งหมด 60 ข้อสอบถาม ผลการทดสอบผู้วิจัยพบว่าค่าประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่เท่ากับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่น และผลการทดสอบอัตราส่วน 2 กลุ่มด้วยสถิติทดสอบ Z (Two Proportion Z Tests) สรุปได้ว่าการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่มีประสิทธิภาพที่ไม่ต่างกับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว กล่าวคือการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูล

ซอฟต์แวร์อาร์ไคฟ์สำหรับระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วสามารถนำตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นหรือตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ไปประยุกต์ใช้ได้โดยให้ประสิทธิภาพที่ไม่ต่างกัน

ผลการทดสอบประสิทธิภาพการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ข้างต้นเป็นผลที่เกิดจากการใช้ข้อมูลซอฟต์แวร์อาร์ไคฟ์ของโครงการพัฒนาซอฟต์แวร์ทางการบัญชีชื่อเคมายมันนี่ (KMyMoney) ที่เป็นโครงการพัฒนาซอฟต์แวร์แบบเปิด (Open Source) และมีรายละเอียดของโครงการตามที่อธิบายไว้ในหัวข้อที่ 3.4 จากการทดสอบเพิ่มเติมในหัวข้อ 4.4.1 ถึง 4.4.4 และการวิเคราะห์เพิ่มเติมในหัวข้อ 4.4.5 ผู้วิจัยสังเกตว่าลักษณะของการเปลี่ยนแปลงแก้ไขที่เกิดขึ้นของนักพัฒนาในโครงการนี้มีลักษณะไม่ค่อยมีการแบ่งกลุ่มแฟ้มข้อมูลที่แต่ละคนแก้ไขอย่างชัดเจน แสดงให้เห็นว่านักพัฒนาภายในโครงการนี้อาจไม่มีการแบ่งหน้าที่การทำงานของนักพัฒนาแต่ละคน ซึ่งก็เป็นลักษณะที่สามารถเกิดขึ้นได้ทั่วไปในโครงการพัฒนาซอฟต์แวร์แบบเปิด ลักษณะดังกล่าวนี้อาจแตกต่างจากโครงการพัฒนาซอฟต์แวร์เชิงพาณิชย์ที่มีการแบ่งหน้าที่การทำงานที่ชัดเจน ผู้วิจัยจึงเห็นว่าลักษณะที่แตกต่างกันดังกล่าวสามารถส่งผลกระทบต่อประสิทธิภาพของระบบให้คำแนะนำนักพัฒนาได้และเป็นประเด็นที่น่าสนใจสำหรับการศึกษาในอนาคต

5.4 การนำงานวิจัยไปประยุกต์ใช้

ในงานวิจัยนี้สามารถใช้เป็นแนวทางในการศึกษาต่อไปหรือนำไปประยุกต์ใช้ในการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลประเภทอื่นๆ โดยผู้วิจัยแบ่งข้อเสนอไว้ดังต่อไปนี้

5.4.1 การนำงานวิจัยไปใช้ในเชิงทฤษฎี

งานวิจัยในอดีตมีการทดสอบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นสำหรับระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ให้ผลประสิทธิภาพที่ดี (Zimmermann et al., 2005; Methanias et al., 2009) ดังนั้นผู้วิจัยจึงสนใจเพิ่มประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ดังกล่าว

ด้วยการนำตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ของ Liu และคณะ (Liu et al., 2008) มาประยุกต์ใช้ เนื่องจากตัวแบบนี้มีคุณสมบัติที่ดีกว่าตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นเดิมและยังมีคุณสมบัติที่น่าสนใจคือสามารถลดจำนวนของกฎความสัมพันธ์ที่เป็นผลบวกลวงลงได้ งานวิจัยนี้จึงสามารถเป็นแนวทางให้กับนักพัฒนาที่สนใจเพิ่มประสิทธิภาพให้กับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์สำหรับระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ต่อไปได้ ตัวอย่างเช่น การนำวิธีการทดสอบประสิทธิภาพของงานวิจัยไปเป็นแนวทางในทดสอบเปรียบเทียบประสิทธิภาพกับตัวแบบของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ตัวแบบอื่นๆต่อไป เป็นต้น

เนื่องจากงานวิจัยนี้มีข้อจำกัดในการเลือกโครงการพัฒนาซอฟต์แวร์ที่นำมาทดสอบ จึงจำเป็นต้องใช้โครงการพัฒนาซอฟต์แวร์ที่เป็นแบบเปิด (Open source) ผู้วิจัยเห็นว่าโครงการพัฒนาซอฟต์แวร์ที่เป็นแบบเปิดอาจมีลักษณะหลายๆอย่างที่แตกต่างจากโครงการพัฒนาซอฟต์แวร์เชิงพาณิชย์ซึ่งอาจมีผลต่อประสิทธิภาพของระบบให้คำแนะนำนักพัฒนาได้ งานวิจัยนี้จึงสามารถเป็นแนวทางกับงานวิจัยที่สนใจทดสอบประสิทธิภาพกับโครงการพัฒนาซอฟต์แวร์เชิงพาณิชย์ได้ รวมถึงงานวิจัยที่สนใจเปรียบเทียบความแตกต่างของการทดสอบประสิทธิภาพกับโครงการพัฒนาซอฟต์แวร์ที่เป็นแบบเปิดและโครงการพัฒนาซอฟต์แวร์เชิงพาณิชย์ด้วย

5.4.2 การนำงานวิจัยไปใช้ในเชิงประยุกต์

จากผลการทดสอบเปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบทั้ง 2 ตัวแบบ ในสถานการณ์การนำทางที่ใช้ค่าเอฟเมเซอร์เป็นค่าประสิทธิภาพนั้นแสดงให้เห็นว่าตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ให้ประสิทธิภาพที่ดีกว่า ดังนั้นถ้าผู้ใช้ให้ความสำคัญกับคำแนะนำในสถานการณ์การนำทางมากที่สุด ผู้ใช้สามารถนำตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ไปประยุกต์ใช้กับระบบให้คำแนะนำนักพัฒนาระหว่างการพัฒนาซอฟต์แวร์ สำหรับในสถานการณ์การป้องกันการเกิดข้อผิดพลาดที่ใช้ค่าเอฟเมเซอร์เป็นค่าประสิทธิภาพเช่นกันนั้นแสดงให้เห็นว่าตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นให้ประสิทธิภาพที่ดีกว่า ดังนั้นถ้าผู้ใช้ให้ความสำคัญกับคำแนะนำในสถานการณ์การป้องกันการเกิดข้อผิดพลาด ผู้ใช้ควรนำตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นไปประยุกต์ใช้กับระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ และสำหรับในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้วที่ใช้ค่าผลสะท้อนกลับเป็นค่าประสิทธิภาพนั้นแสดงให้เห็นว่าตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นและตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ให้ประสิทธิภาพที่

ไม่ต่างกัน ถ้าผู้ใช้ให้ความสำคัญกับคำแนะนำในสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว ผู้ใช้สามารถเลือกนำตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นหรือตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ไปประยุกต์ใช้กับระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ได้

นอกจากนั้น จากผลการทดสอบเพิ่มเติม แสดงให้เห็นว่าการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นให้ค่าความถูกต้อง (Precision) ที่สูงกว่าการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่แต่การใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นให้เรียกคืน (Recall) ที่ต่ำกว่าการใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ ถ้าผู้ใช้ให้ความสำคัญกับค่าความถูกต้องมากกว่าค่าเรียกคืน ผู้ใช้ควรเลือกนำตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นไปประยุกต์ใช้กับระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์ แต่ถ้าผู้ใช้ให้ความสำคัญกับค่าเรียกคืนมากกว่าค่าความถูกต้อง ผู้ใช้ควรเลือกนำตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่ไปประยุกต์ใช้กับระบบให้คำแนะนำนักพัฒนาในระหว่างการพัฒนาซอฟต์แวร์

เนื่องจากผลของงานวิจัยนี้เป็นผลมาจากการทดสอบกับโครงการพัฒนาซอฟต์แวร์แบบเปิด (Open source) ดังนั้นการนำงานวิจัยไปใช้ในเชิงประยุกต์จึงต้องใส่ใจในประเด็นของลักษณะของโครงการที่นำไปประยุกต์ใช้ด้วย ผู้วิจัยเห็นว่าผู้ใช้สามารถนำคำแนะนำในการประยุกต์ใช้ที่กล่าวในข้างต้นไปใช้กับโครงการพัฒนาซอฟต์แวร์แบบเปิดและโครงการพัฒนาซอฟต์แวร์เชิงพาณิชย์แต่อยู่ภายใต้เงื่อนไขของทีมพัฒนาย่อยที่มีขนาดของทีมไม่มากนักและมีความอิสระภายในทีมสูง

5.5 ข้อจำกัดของงานวิจัย

จากการทดสอบเปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบทั้ง 2 ตัวแบบในงานวิจัยนี้ มีข้อจำกัดบางประการดังนี้

1) ผลการทดสอบเปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบทั้ง 2 ตัวแบบในงานวิจัยนี้ เป็นผลจากการทดสอบกับข้อมูลซอฟต์แวร์อาร์ไคฟ์ของโครงการพัฒนาซอฟต์แวร์เคมายมันนี่ (KMyMoney) เท่านั้น สำหรับการทดสอบเปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูล

ซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบทั้ง 2 ตัวแบบด้วยข้อมูลซอฟต์แวร์อาร์ไคฟว์ของโครงการพัฒนาซอฟต์แวร์อื่นๆ ที่มีคุณลักษณะแตกต่างไปจากโครงการพัฒนาซอฟต์แวร์เคมายมันนี่ (KMyMoney) เช่น ขนาด อัตราการเปลี่ยนแปลงแก้ไข ภาษาที่ใช้ในการพัฒนา รวมถึงลักษณะของโครงการพัฒนาซอฟต์แวร์แบบเปิด (Open Souce) หรือโครงการพัฒนาซอฟต์แวร์เชิงพาณิชย์ (Commercial) อาจให้ผลที่แตกต่างกันออกไป

2) การทดสอบเปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ด้วยตัวแบบทั้ง 2 ตัวแบบในงานวิจัยนี้เป็นผลจากการทดสอบที่ผู้วิจัยกำหนดค่าต่างๆ ในเครื่องมือที่ใช้ในการทดสอบดังนี้

- ค่าสนับสนุนนับขั้นต่ำ (Minimum Support Count) เท่ากับ 3
- ค่าความเชื่อมั่นขั้นต่ำ/ค่าความเชื่อมั่นใหม่ขั้นต่ำ (Mimimum Confidence / New Confidence) เท่ากับ 0.1
- การกำหนดค่าน้ำหนักของค่าเอฟเมอร์แบบสมดุล กล่าวคือให้น้ำหนักกับค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) อย่างละ 0.5 เท่ากัน

การกำหนดค่าต่างๆ ในเครื่องมือที่ใช้ในการทดสอบที่แตกต่างกันออกไปนี้ อาจทำให้ได้ผลที่แตกต่างออกไปดังนี้

- การกำหนดค่าสนับสนุนนับขั้นต่ำ (Minimum Support Count) มากกว่าหรือน้อยกว่า 3 อาจทำให้เซตของคำแนะนำที่ได้แตกต่างออกไป ซึ่งมีผลให้ทำให้ค่าความถูกต้อง ค่าเรียกคืนและค่าเอฟเมเซอร์แตกต่างออกไปได้
- การกำหนดค่าความเชื่อมั่นขั้นต่ำ/ค่าความเชื่อมั่นใหม่ขั้นต่ำ (Mimimum Confidence / New Confidence) ที่มากกว่า 0.1 อาจทำให้เซตของคำแนะนำที่ได้แตกต่างออกไป ซึ่งมีผลให้ทำให้ค่าความถูกต้อง ค่าเรียกคืนและค่าเอฟเมเซอร์แตกต่างออกไปได้
- การกำหนดค่าน้ำหนักของค่าเอฟเมอร์แบบสมดุล กล่าวคือให้น้ำหนักกับค่าความถูกต้อง (Precision) และค่าเรียกคืน (Recall) อย่างละ 0.5 เท่ากัน

5.6 แนวทางการศึกษาต่อเนื่อง

จากข้อจำกัดของงานวิจัย ผู้ที่สนใจศึกษาต่อเนื่องอาจใช้เป็นแนวทางดังต่อไปนี้ในการศึกษาได้

1) ผู้ที่สนใจสามารถทดสอบกับข้อมูลซอฟต์แวร์และข้อสอบถามชุดอื่นที่นอกเหนือจากข้อมูลซอฟต์แวร์อาร์ไคฟ์ของโครงการพัฒนาซอฟต์แวร์เคมายมันนี่ (KMyMoney) เพื่อให้ผลการทดสอบครอบคลุมโครงการพัฒนาซอฟต์แวร์ทุกๆคุณลักษณะ
2) ผู้ที่สนใจสามารถทดสอบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์โดยกำหนดค่าต่างๆของเครื่องมือทดสอบที่เหมาะสม ได้แก่ ค่าสนับสนุนนับขั้นต่ำ (Minimum Support Count) ค่าความเชื่อมั่นขั้นต่ำ / ค่าความเชื่อมั่นใหม่ขั้นต่ำ (Mimimum Confidence / New Confidence) และกำหนดค่าน้ำหนักของค่าเอฟเมอร์
3) ผู้สนใจการทดสอบเปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ด้วยตัวแบบทั้ง 2 ตัวแบบในงานวิจัยสามารถเปลี่ยนการวัดประสิทธิภาพจากการใช้ค่าเอฟเมเซอร์ไปเป็นการวัดค่าอื่นๆแทน เช่น การวัดประสิทธิภาพจากนับจำนวนของกฎความสัมพันธ์ที่เป็นผลบวกลวง
4) ผู้สนใจสามารถนำตัวแบบประเมินระดับความน่าสนใจของกฎความสัมพันธ์อื่นๆมาประยุกต์ใช้กับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์ เช่น ค่าลิฟท์ (Lift) ค่าเลฟเวอเรจ (Leverage) ค่าคัฟเวอเรจ (Coverage) ค่าสหสัมพันธ์ (Correlation) และ ค่าอัตราส่วนออดส์ (Odds Ratio) เป็นต้น
5) ผู้สนใจการทดสอบเปรียบเทียบประสิทธิภาพของการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ด้วยตัวแบบทั้ง 2 ตัวแบบสามารถทดลองกับข้อมูลซอฟต์แวร์อาร์ไคฟ์ของโครงการพัฒนาซอฟต์แวร์เชิงพาณิชย์หรือโครงการพัฒนาซอฟต์แวร์แบบเปิดที่มีกฎระเบียบที่เคร่งครัด
6) ผู้สนใจสามารถประยุกต์ใช้ตัวแบบค่าสนับสนุน-ค่าความเชื่อมั่นใหม่กับการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟ์เพื่อวัตถุประสงค์อื่นๆได้ เช่น การประยุกต์ใช้กฎความสัมพันธ์ปฏิเสธ (Nagative Association Rule) กับการสร้างระบบแจ้งเตือนนักพัฒนาว่า ไม่ควรแก้ไขแฟ้มข้อมูลนี้ต่อจากแฟ้มข้อมูลที่แก้ไขไปก่อนหน้า เป็นต้น

รายการอ้างอิง

Agrawal, R., Imielinski T., and Swami A. 1993. Mining Association Rules between sets of items in large databases. In Proceedings of the ACM SIGMOD Conference on Management of Data : 207-216.

Agrawal, R., and Srikant, R. 1994. Fast Algorithms for Mining Association Rules. Proc. 20th Very Large Data Bases Conf. (VLDB) : 487-499.

Agrawal, R., and Srikant, R. 1995. Mining Sequential Patterns. In Proc. of the 11th Int'l Conference on Data Engineering.

Agrawal, R., Arning, A., Bollinger, T., Mehta, M., Shafer, J., and Srikant, R. 1996. The quest Data Mining system. In Proceedings of KDD'96 : 244-249.

Ambriola, V., Bendix, L., and Ciancarini, P. 1990. The Evolution of configuration management and version control. Software Engineering Journal. 5, 6: 303-310.

Baeza-Yates, R., and Ribeiro-Neto, B. 1999. Modern Information Retrieval, first ed. Addison-Wesley-Longman.

Ball, T., Kim, J.-M., Porter, A.A., and Siy, H.P. 1997. If Your Version Control System Could Talk. Proc. ICSE Workshop Process Modelling and Empirical Studies of Software Eng.

Baudis, P. 2009. Current Concepts in Version Control Systems.

Beil, F., Ester, M. and Xu, X. 2002. Frequent Term-Based Text Clustering. Proc. Eighth Int'l Conf. Knowledge Discovery and Data Mining (KDD 2002).: 436-442.

Bieman, J.M., Andrews, A.A., and Yang, H.J. 2003. Understanding Change-Proneness In OO Software through Visualization. Proc. 11th Int'l Workshop Program Comprehension. : 44-53.

Bird, C., Gourley, A., Devanbu, P.T., Gertz, M., and Swaminathan, A. 2006. Mining email social networks. In Proceedings of Int. Workshop on Mining Software Repositories MSR.

Breu, S., and Zimmermann, T. 2006. Mining Aspects from Version History. Proceedings of the 21st IEEE/ACM International Conference on Automated Software Engineering.: 221-230.

Brin, S., Motwani, R., Ullman J. 1997. Dynamic Itemsets Counting and implication rules for market basket data. In proc. of 997 ACM-SIGMOD Int. Conf. on Management of Data.

Burch, M., Diehl, S., and Weißgerber, P. Visual Data Mining in Software Archives. In Proc. ACM Symposium on Software Visualization SOFTVIS, St. Louis, Missouri, USA, May 2005.

Cederqvist, P. 2006. Version Management with CVS. Network Theory Ltd.

Chadd, W., Jaime, S. 2008. Branching and merging in the repository. Proceedings of the 2008 international working conference on Mining software repositories.

CMMI Product Team. 2006. CMMI for Development, Version 1.2. Software Engineering Institute, Carnegie Mellon University.: CMU/SEI-2006-TR-008.

Conradi, R., Westfechtel B. 1998. Version Models for software configuration management. ACM Computing Surveys (CSUR). 30, 2: 232-282.

Freitas, A. 1999. On rule interestingness measures. Knowledge-Based Systems journal.: 309–315.

Gall, H., Hajek, K., and Jazayeri, M. 1998. Detection of Logical Coupling Based on Product Release History. In Proceedings of the 26th International Conference on Software Maintenance (ICSM '98). : 190-198.

Gall, H., Jazayeri, M., and Krajewski, J. 2003. CVS release history data for detecting logical Couplings. In IWPSE 2003.

Geng. L., Hamilton, HJ. 2006. Interestingness measures for Data Mining - a survey. ACM Comput Surveys 2006. 38(3), article 9

Geyer-Schulz, A., and Hahsler, M. 2002. Evaluation of recommender algorithms for an internet information broker based on simple Association Rules and on the repeat-buying theory. Proceedings of Fourth WebKDD Workshop: Web Mining for Usage Patterns & User Profiles.: 100–114.

Grune, D., Berliner, B. 2006. CVS [On-Line]. Available from: http://www.nongnu.org/cvs/

Hahsler, M. 2009. A Comparison of Commonly Used Interest Measures for Association Rules. Available from: http://www.ai.wu-wien.ac.at/~hahsler/research/association_rules/measures.html

Heravi, M.J. 2009. A study on Interestingness Measures for Associative Classifiers. Master's Thesis, Department of Computing Science, Faculty of Science, University of Alberta.

Huzefa, K., Michael, C., Jonathan, M. 2007. Comparing Approaches to Mining Source Code for Call-Usage Patterns. Proceedings of the Fourth International Workshop on Mining Software Repositories. :20.

Junqueira, D., Bittar, T., Fortes, R. 2008. A fine-grained and flexible version control for software. SIGDOC'08.

Kim, M., Sazawal, V., Notkin, D., Murphy, G. 2005. An empirical study of code clone genealogies. Proceedings of the 10th European software engineering conference held jointly with 13th ACM SIGSOFT international symposium on Foundations of software engineering.

Kotsiantis, S., Kanellopoulos, D. 2006. Association Rules Mining: A Recent Overview. GESTS International Transactions on Computer Science and Engineering. 32, 1: 71-82.

Lenca, P., Meyer, P., Vaillant, B., and Lallich, S. 2004. A multicriteria decision aid for interestingness measure selection. Technical Report LUSSI-TR-2004-01-EN, LUSSI Department, GET/ENST, Bretagne, France.

Lenca, P., Vaillant, B., Meyer P., and Lallich, S. 2007. Association Rule interestingness measures: Experimental and theoretical studies. In Quality Measures in Data Mining.: pages 51–76.

Li, H., Duo, Z., Jian, H., Hua-Jun, Z., Zheng, C. 2007. Finding keyword from online broadcasting content for targeted advertising. Proceedings of the 1st international workshop on Data Mining and audience intelligence for advertising. : 55-62.

Li, Z., and Zhou, Y. 2005. PR-Miner: Automatically Extracting Implicit Programming Rules and Detecting Violations in Large Software Code. In Proceedings of 13th International Symposium on Foundations of Software Engineering (ESEC/FSE'05).

Liguo, Yu. 2007. Understanding component co-Evolution with a study on Linux. Empirical Software Engineering. 12, 2: 123-141.

Liu, J., Xiaoping, F., Zhihua, Q. 2008. A New Interestingness Measure of Association Rules. Genetic and Evolutionary Computing 2008. WGEC '08. Second International Conference. :393-397.

Livshits, B., and Zimmermann, T. 2005. DyanMine: Finding Common Error Patterns by Mining Software Revision Histories. In Proceedings of 13th International Symposium on Foundations of Software Engineering (ESEC/FSE'05).

Löh, A., Swierstra, W. Leijen D. 2007. A Principled Approach to Version Control [On-Line]. Available from: http://people.cs.uu.nl/andres/VersionControl.html

Lucian, V., Alex, T., Jarke, W. 2005. CVSscan: visualization of code Evolution. Proceedings of the 2005 ACM symposium on Software visualization.

Lucian, V., Alexandru, T. 2006. An open framework for CVS repository Querying, analysis and visualization. Proceedings of the 2006 international workshop on Mining software repositories.

Major, J.A., and Mangano, J.J. 1995. Selecting among rules induced from a hurricane database. Journal of Intelligent Information systems.: 4:39–52.

Mcgarry, K. 2005. A survey of interestingness measures for knowledge discovery. Knowl. Eng. Review 20, 1, 39–61.

Methanias C.J., Manoel M., Francisco R.. 2009. Mining software change history in an industrial environment. XXIII Brazilian Symposium on Software Engineering.

Michael Fischer, Martin Pinzger, Harald Gall. 2003. Populating a Release History Database from Version Control and Bug Tracking Systems. Proceedings of the International Conference on Software Maintenance. : 23

Michail, A. 1999. Data Mining Library Reuse Patterns in Userselected Applications. In 14th IEEE International Conference on Automated Software Engineering. : 24–33.

Michail, A. 2000. Data Mining Library Reuse Patterns Using Generalized Association Rules. In Proceedings of 22nd International Conference on Software Engineering (ICSE'00). : 167-176.

Miller, W., and Myers, E.W. 1985. A file comparison program. Software Practice and Experience. 15, 11: 1025–1040.

Nayyeri, A., and Oroumchian, F. 2006. Fufair: a fuzzy farsi information retrieval system. in proceedings of the 4th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-06).

Object Technology International. Eclipse Platform Technical Overview, Feb. 2003. Available at www.eclipse.org

Olivier, C., Vincent, D., Tienté, H. Engelbert Mephu Nguifo. 2008. Optimizing Occlusion Appearances In 3D Association Rules Visualization. Intelligent Systems 2008. 2: 15-42-15-49.

O'Sullivan, B. 2009. Making Sense of Revision-control Systems. ACM Queue.

Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal U., and Hsu, M. 2004. Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach. IEEE Transaction on Knowledge and Data Engineering. 16: 10.

Piatetsky-Shapiro, G., 1991. Discovery Analysis and Presentation of strong rules. in: Knowledge Discovery in Databases, AAAI/MIT Press. : p. 229-248.

Rijsbergen, C.J. 1979. Information Retrieval. London: Butterworth

Rochkind, M.J. 1975. The Source Code Control System. IEEE Transactions on Software Engineering. SE-1: 4364-370.

Sheikh, L.m., Tanveer, B., Hamdani, M.A. 2004. Interesting measures for Mining Association Rules, 8th International Multitopic Conference.

Sheykh E.K., Abolhassani, H., Neshati M., Behrangi, E., Rostami, A., Mohammadi, M. 2007. Mahak: A Test Collection for Evaluation of Farsi Information Retrieval Systems. IEEE/ACS International Conference on Computer Systems and Applications.

Srikant, R. and Agrawal, R. 1995. Mining Sequential Patterns: Generalizations and Performance Improvements. Research Report RJ 9994, IBM Almaden Research Center.

Srikant, R., Vu, Q., and Agrawal, R. 1997. Mining Association Rules with Item Constraints. Proc. Third Int'l Conf. KDD and Data Mining (KDD '97).

Tan, P.n., Kumar, V. 2002. Selecting the Right Interestingness Measure for Association, In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data.

Tan, P., Kumar, V., and Srivastava, J. 2002. Selecting the right interestingness measure for association patterns. In Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (KDD 2002). Edmonton, Canada.: 32–41.

Tichy, W. 1985. RCS: A system for version control. Software-Practice and Experience. 15, 7: 637-654.

Tichy, W.F. 1982. Design, implementation, and evaluation of a revision control system. In ICSE '82: Proceedings of the 6th international conference on Software engineering. : 58-67.

Tsunenori I. 2003. Evaluation of Criteria for Information Retrieval. International Conference on Web Intelligence. IEEE Computer Society.

Ubranic´, D.Cˇ., and Murphy, G.C. 2003. Hipikat: Recommending pertinent software development artifacts. In Proc. 25th International Conference on Software Engineering (ICSE). : 408–418.

Williams, C.C., and Hollingsworth, J.K. 2005. Automatic Mining of Source Code Repositories to Improve Bug Finding Techniques. IEEE Trans. Software Eng.: vol. 31, no. 6, pp. 466-480.

Williams, C.C., and Hollingsworth J.K. 2005. Recovering System Specific Rules from Software Repositories. In Proceedings of 2nd International Workshop on Mining Software Repositories (MSR'05). : 7-11

Weißgerber, P., Leo, K., Burch, M., Diehl, M. 2005. Exploring Evolutionary Coupling in Eclipse. Proceedings of the 2005 OOPSLA workshop on Eclipse technology eXchange. : 31-34.

Weißgerber, P., and Diehl, S. 2006. Identifying Refactorings from Source-Code Changes. Proceedings of the 21st IEEE International Conference on Automated Software Engineering (ASE'06).: p.231-240.

Weissgerber, P., Mathias, P., Michael, B. 2007. Visual Data Mining in Software Archives to Detect How Developers Work Together. Proceedings of the Fourth International Workshop on Mining Software Repositories, : 9.

Ying, A., Murphy, G., Raymond N., Chu-Carroll, M. 2004. Predicting Source Code Changes by Mining Change History. IEEE Transactions on Software Engineering, 30, 9: 574-586.

Zimmermann, T., Diehl, S., and Zeller Andreas. 2003. How history justifies system architecture (or not). In IWPSE 2003.

Zimmermann, T., Weißgerber, P.. 2004. Preprocessing CVS Data For Fine-Grained Analysis. Proc. Mining Software Repositories. : 2-6.

Zimmermann, T., Weisgerber, P., Diehl, S., Zeller, A. 2005. Mining Version Histories to Guide Software Changes. Proceedings of the 26th International Conference on Software Engineering. : 563-572.

ภาคผนวก

ภาคผนวก ก
การเลือกทรานแซคชั่นชุดทดสอบ.

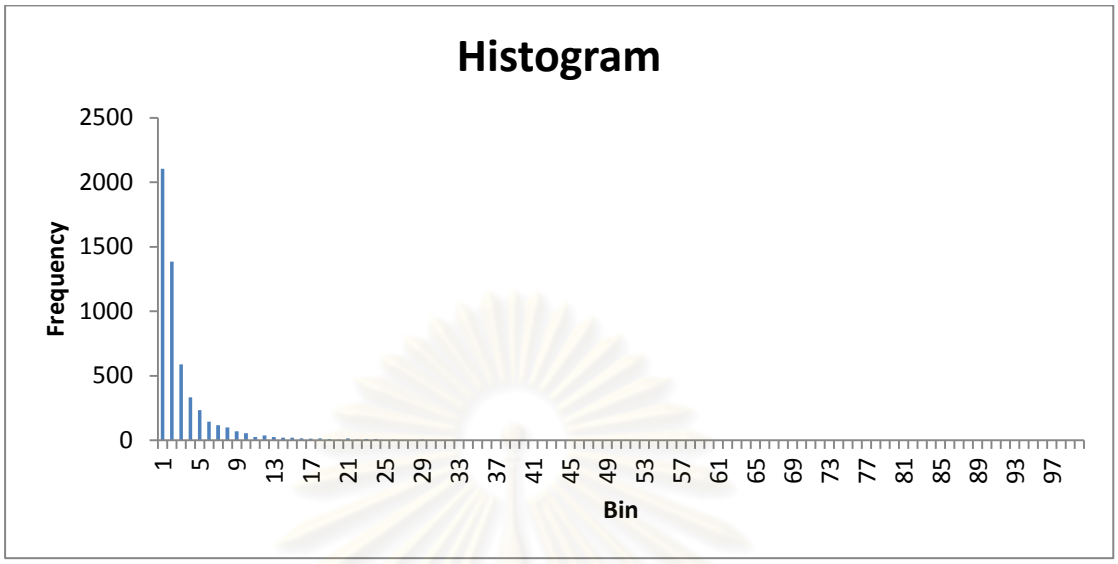
ขั้นตอนการเลือกทรานแซคชั่นชุดทดสอบนี้ เป็นขั้นตอนย่อยขั้นตอนหนึ่งภายในขั้นตอนการสร้างข้อสอบถามที่อธิบายอย่างละเอียดในหัวข้อ 3.6.2 แต่อธิบายในภาคผนวกเนื่องจากเป็นการเลือกทรานแซคชั่นชุดทดสอบที่เฉพาะเจาะจงกับโครงการเคมายมันนี่ (KMyMoney) เท่านั้น

เริ่มต้นผู้วิจัยต้องกำหนดว่าทรานแซคชั่นขนาดสั้น กลางและยาวนั้นมีขนาดอยู่ช่วงไหนบ้าง ผู้วิจัยนำข้อมูลขนาดของทรานแซคชั่นทั้งหมดในฐานข้อมูลมาคำนวณหาค่าทางสถิติเชิงพรรณนา รวมถึงนำไปสร้างเป็นแผนภูมิแท่งแจกแจงความถี่ของขนาดทรานแซคชั่น โดยใช้โปรแกรมตารางคำนวณไมโครซอฟต์เอ็กซ์เซล (Microsoft Excel) มีข้อมูลเข้าเป็นขนาดของแต่ละทรานแซคชั่นทั้งหมด 5,458 ทรานแซคชั่น ได้ข้อมูลออกเป็นตารางค่าทางสถิติเชิงพรรณนาของขนาดทรานแซคชั่นและแผนภูมิแท่งแจกแจงความถี่ของขนาดทรานแซคชั่น ดังตารางและรูปด้านล่างนี้

ตารางที่ ก-1 แสดงค่าทางสถิติเชิงพรรณนาของขนาดทรานแซคชั่นในฐานข้อมูลซอฟต์แวร์อาร์ไคฟ์โครงการเคมายมันนี่ (KMyMoney)

ค่าเฉลี่ยเลขคณิตของขนาดทรานแซคชั่น	3.91
ส่วนเบี่ยงเบนมาตรฐานของขนาดทรานแซคชั่น	9.81
พิสัยของขนาดทรานแซคชั่น	300
ขนาดของทรานแซคชั่นต่ำสุด	1
ขนาดของทรานแซคชั่นสูงสุด	301
จำนวนทรานแซคชั่นทั้งหมด (ทรานแซคชั่น)	5,458
จำนวนการเปลี่ยนแปลงแก้ไขทั้งหมด (รายการ)	21,358

ข้อมูลที่แสดงอยู่ในตารางข้างต้นนี้เป็นข้อมูลทางสถิติของข้อมูลซอฟต์แวร์อาร์ไคฟ์โครงการเคมายมันนี่ (KMyMoney) ภายหลังจากผ่านขั้นตอนการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟ์ (หัวข้อ 3.6.1) เรียบร้อยแล้ว



รูปที่ ก-1 แสดงแผนภูมิแท่งแจกแจงความถี่ตามขนาดทรานแซคชั่นในฐานข้อมูลซอฟต์แวร์อาร์ไควฟ์โครงการเคมายมันนี่ (KMyMoney)

เนื่องจากทรานแซคชั่นชุดทดสอบที่จะเลือกขึ้นมาจะต้องถูกนำไปใช้สร้างข้อสอบถามสำหรับทุกสถานการณ์ของการทดสอบ ดังนั้นทรานแซคชั่นชุดทดสอบที่จะเลือกขึ้นมาจะต้องเป็นไปตามข้อกำหนดของการสร้างข้อสอบถามทั้ง 3 สถานการณ์นั่นคือ ต้องเป็นทรานแซคชั่นที่มีขนาดมากกว่าหรือเท่ากับ 2 รายการ จากตารางที่ ก-1 แสดงค่าทางสถิติเชิงพรรณนาทำให้ทราบว่าขนาดโดยเฉลี่ยของทรานแซคชั่นคือ 4 รายการต่อ 1 ทรานแซคชั่น ผู้วิจัยจึงเลือกให้ทรานแซคชั่นที่มีขนาด 2 รายการเพียงขนาดเดียวเป็นทรานแซคชั่นที่อยู่ในกลุ่มขนาดสั้น เนื่องจากเป็นกลุ่มที่มีขนาดสั้นกว่าค่าเฉลี่ยและเป็นกลุ่มที่มีปริมาณเยอะมากจากแผนภูมิแท่งในรูปที่ ก-1 ข้างต้น และผู้วิจัยเลือกให้ทรานแซคชั่นที่มีขนาดมากกว่า 12 รายการเป็นทรานแซคชั่นที่อยู่ในกลุ่มขนาดยาว เนื่องจากเป็นกลุ่มที่มีขนาดยาวกว่าค่าเฉลี่ยและเป็นกลุ่มที่แยกจากกลุ่มทรานแซคชั่นที่มีขนาดอยู่ระหว่าง 3 - 11 รายการที่ผู้ใช้กำหนดให้เป็นกลุ่มขนาดกลางอย่างชัดเจน (กลุ่มทรานแซคชั่นที่มีขนาดอยู่ระหว่าง 3 - 11 รายการมีแนวโน้มลดลงเรื่อยๆจนถึงทรานแซคชั่นที่มีขนาด 12 รายการซึ่งเป็นจุดเปลี่ยนของแนวโน้ม ดังแสดงในแผนภูมิแท่งในรูปที่ ก-1 ข้างต้น)

หลังจากที่ได้ช่วงขนาดของทรานแซคชั่นในแต่ละกลุ่มแล้ว ต่อไปผู้วิจัยจะต้องกำหนดว่าการพบทรานแซคชั่นรูปแบบหนึ่งจำนวนกี่ครั้งจึงจะอยู่ในกลุ่มพบบ่อย และจำนวนกี่ครั้งจึงจะอยู่ในกลุ่มพบไม่บ่อย โดยผู้วิจัยได้ทำการสร้างรูปแบบที่เป็นไปได้ทั้งหมดของทรานแซคชั่นที่มีขนาดต่างๆออกมาแล้วทำการนับจำนวนการปรากฏของรูปแบบที่เป็นไปได้เหล่านั้นในฐานข้อมูลทั้งหมด

เนื่องจากฐานข้อมูลซอฟต์แวร์อาร์ไคฟ์ที่นำมาใช้มีจำนวนของการเปลี่ยนแปลงแก้ไขทั้งหมด 21,358 รายการ (นับมาจากการเปลี่ยนแปลงแก้ไขที่เกิดขึ้นจริงทั้งหมด ดังนั้นการเปลี่ยนแปลงแก้ไขที่เคยเกิดขึ้นมากกว่า 1 ครั้งจะถูกนับซ้ำด้วย) ในการเปลี่ยนแปลงแก้ไขเหล่านั้นมีการเปลี่ยนแปลงแก้ไขที่แตกต่างกันทั้งหมด 3,162 รายการ การเปลี่ยนแปลงแก้ไขทั้ง 3,162 รายการนี้จะถูกนำไปสร้างเป็นรูปแบบทรานแซคชั่นที่เป็นไปได้ทั้งหมดในทุกขนาดต่างๆกัน (ตัวอย่างเช่น การสร้างทรานแซคชั่นขนาด 2 รายการที่เป็นไปได้ทั้งหมดจะได้ $\frac{3162!}{2!\,3160!}$ = 4,997,541 รูปแบบทรานแซคชั่น เป็นต้น) แล้วนับจำนวนว่าแต่ละรูปแบบมีจำนวนการพบในฐานข้อมูลทั้งหมดกี่ครั้ง รูปแบบทรานแซคชั่นที่ไม่เคยปรากฏเลยในฐานข้อมูล (รูปแบบทรานแซคชั่นที่มีจำนวนการพบเท่ากับ 0 ครั้ง) จะถูกตัดออก การนับจำนวนการพบหรือการปรากฏของรูปแบบทรานแซคชั่นต่างๆ ในแต่ละขนาดแสดงดังตารางต่อไปนี้

ตารางที่ ก-2 แสดงจำนวนของรูปแบบของทรานแซคชั่นในขนาดต่างๆและจำนวนการปรากฏต่างๆ

ขนาดของรูปแบบของทรานแซคชั่น	จำนวนครั้งที่ปรากฏ																												
	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
2	2884	1419	844	486	299	191	162	115	83	52	50	40	24	29	25	19	13	17	12	19	12	11	7	6	9	5	8	7
3	18712	5653	2214	1001	514	291	193	125	99	71	48	41	25	18	19	8	10	7	6	3	4	2	6	5	6	7	2	1
4	65434	11085	2853	1082	447	207	134	87	61	37	18	11	6	6	3	1	0	2	0	0	0	2	2	3	2	2	0	0
5	161263	14374	2238	728	205	91	57	29	11	10	4	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	381140	18030	1271	322	52	26	13	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	980881	16243	502	85	5	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1332620	4932	131	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1204311	2002	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	898238	595	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	226510	123	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	71633	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	60521	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	30201	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	5630	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	639	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	217	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

ตารางข้างต้นแสดงจำนวนของรูปแบบของทรานแซคชั่นในขนาดต่างๆตั้งแต่ขนาด 2 จนถึงขนาด 20 รายการและจำนวนการปรากฏต่างๆตั้งแต่ 3 ครั้งจนถึง 30 ครั้งซึ่งเป็นเพียงส่วนหนึ่งของการสร้างและนับจำนวนรูปแบบทรานแซคชั่นที่เป็นไปได้ แถวของตารางแสดงขนาดของรูปแบบทรานแซคชั่น หลักของตารางแสดงจำนวนครั้งที่ปรากฏ จำนวนที่ปรากฏภายในตารางคือจำนวนของรูปแบบที่มีขนาดเท่ากับขนาดของแถวนั้นและปรากฏในฐานข้อมูลเท่ากับจำนวนของหลักนั้น สาเหตุที่ตารางเริ่มที่จำนวนการปรากฏ 3 ครั้งเนื่องจากการปรากฏ 1 ครั้งเป็นการปรากฏขั้นต่ำอยู่แล้ว ส่วนการปรากฏ 2 ครั้งนั้นมีจำนวนมากในหลักหมื่น แสนและล้านผู้วิจัยจึงไม่นำมาแสดง

จากจำนวนที่แสดงในตารางข้างต้นทำให้ผู้วิจัยสามารถกำหนดขอบบนให้กับทรานแซคชั่นชุดทดสอบกลุ่มที่มีขนาดยาวได้ ขอบบนนั้นคือขนาด 19 รายการ เนื่องจากในขนาด 20 รายการนั้นมีการปรากฏมากที่สุดแค่เพียง 2 ครั้ง ถ้านำมาแยกต่อเป็นพบบ่อยกับพบไม่บ่อยจะได้ว่าพบไม่บ่อยคือพบ 1 ครั้งส่วนพบบ่อยคือพบเพียง 2 ครั้ง ซึ่งในความเป็นการพบ 1 และ 2 ครั้งนั้นแตกต่างกันน้อยมาก

จากตาราง ทำให้ผู้วิจัยต้องกำหนดจำนวนที่จะเรียกว่าพบบ่อยและพบไม่บ่อยของแต่ละขนาดแตกต่างกันออกไป โดยผู้วิจัยเลือกแบ่งกึ่งกลางระหว่างจำนวนการปรากฏที่น้อยที่สุด (ปรากฏ 1 ครั้งเสมอ) กับจำนวนการปรากฏที่มากที่สุดของแถวนั้นๆ ตัวอย่างเช่น ทรานแซคชั่นขนาด 5 รายการมีจำนวนการปรากฏที่มากที่สุดคือ 15 ครั้ง ดังนั้นจุดแบ่งคือ (1+15)/2 = 8 จะได้ว่าการพบทรานแซคชั่นขนาด 5 รายการที่พบน้อยกว่า 8 ครั้งเป็นกลุ่มที่พบไม่บ่อยและการพบทรานแซคชั่นขนาด 5 รายการที่พบมากกว่าหรือเท่ากับ 8 ครั้งเป็นกลุ่มที่พบบ่อย ด้วยวิธีการนี้จะทำให้ได้กลุ่มพบบ่อยและกลุ่มพบไม่บ่อยของแต่ละขนาดดังนี้

ตารางที่ ก-3 แสดงการแบ่งกลุ่มจำนานการปรากฏของทรานแซคชั่นแต่ละขนาด

ขนาด \ การปรากฏ	จำนวนการปรากฏ (ครั้ง)	
	กลุ่มพบไม่บ่อย	กลุ่มพบบ่อย
2	1-24	25-49
3	1-24	25-48
4	1-14	15-29
5	1-7	8-15
6	1-6	7-13
7	1-4	5-9
8	1-3	4-6
9	1-2	3-5
10	1-2	3-5
11	1-2	3-4
12	1-2	3-4
13	1-2	3-4
14	1	2-3
15	1	2-3
16	1	2-3
17	1	2-3
18	1	2-3
19	1	2-3

เมื่อได้ขอบเขตของแต่ละกลุ่มในการเลือกทรานแซคชั่นขึ้นมาสร้างเป็นข้อสอบถามแล้ว ผู้วิจัยใช้วิธีการสุ่มเลือก ทรานแซคชั่นออกมาตามกลุ่มนั้นๆจำนวนกลุ่มละ 10 ทรานแซคชั่นทำให้ได้ทรานแซคชั่นที่เป็นตัวแทนของแต่ละกลุ่ม รวมทั้งสิ้น 60 ทรานแซคชั่นมาเป็นตัวแทนของทรานแซคชั่นทั้งหมด และเรียกทรานแซคชั่นทั้ง 60 ทรานแซคชั่นนี้ว่า ทรานแซคชั่นชุดทดสอบ การสุ่มเลือกดังกล่าวใช้โปรแกรมตารางคำนวณไมโครซอฟต์เอ็กซ์เซล (Microsoft Excel) โดยมีข้อมูลเข้าเป็นหมายเลขของทรานแซคชั่น ขนาด จำนานการปรากฏทั้งหมดและข้อกำหนดในตารางข้างต้น

ข้อมูลออกคือหมายเลขทรานแซคชั่นที่ได้รับการสุ่ม ทรานแซคชั่นชุดทดสอบทั้งหมดแสดงไว้ในตารางต่อไปนี้

กำหนดให้ alter(*x*) คือ การเปลี่ยนแปลงแก้ไขในมิติแก้ไข (alter) เอนทิตี้ *x* โดยที่ *x* เป็นเอนทิตี้ในระดับของแฟ้มข้อมูล แทนการเขียนแบบเต็ม alter(file, *x*, …) เพื่อความสะดวกในการแสดงข้อมูล

ตารางที่ ก-4 แสดงทรานแซคชั่นชุดทดสอบ

	ทรานแซคชั่น
1	{alter(ChangeLog), alter(mymoneyaccount.cpp)}
2	{alter(mymoneyfile.h), alter(mymoneyseqaccessmgr.h)}
3	{alter(kmymoneyutils.cpp), alter(kmymoneyutils.h)}
4	{alter(mymoneyfile.cpp), alter(mymoneyseqaccessmgr.cpp)}
5	{alter(kledgerviewcheckings.cpp), alter(kledgerviewloan.cpp)}
6	{alter(ChangeLog), alter(kmymoneyview.cpp)}
7	{alter(imymoneystorage.h), alter(mymoneyseqaccessmgr.cpp)}
8	{alter(kgloballedgerview.cpp), alter(kmymoneyview.cpp)}
9	{alter(kmymoney2.h), alter(kmymoney2ui.rc)}
10	{alter(imymoneystorage.h), alter(mymoneyseqaccessmgr.h)}
11	{alter(knewaccountwizard.cpp), alter(kmymoneyedit.cpp)}
12	{alter(Makefile.am), alter(kreportsview.cpp)}
13	{alter(knewaccountwizard.cpp), alter(kbanklistitem.cpp)}
14	{alter(mymoneyreport.cpp), alter(querytable.cpp)}
15	{alter(kmymoneyview.cpp), alter(kmymoneyregister.cpp)}
16	{alter(keditequityentrydlg.cpp), alter(keditequityentrydlg.h)}
17	{alter(knewaccountdlg.cpp), alter(kledgerview.cpp)}
18	{alter(mymoneyfile.cpp), alter(mymoneystoragedump.cpp)}
19	{alter(kledgerview.cpp), alter(kmymoneyregister.cpp)}
20	{alter(Makefile.am), alter(kmymoneyview.cpp)}
21	{alter(kmymoney2.cpp), alter(kmymoney2.h), alter(kmymoneyview.h)}

22	{alter(kledgerview.cpp), alter(kledgerview.h), alter(kledgerviewcheckings.cpp)}
23	{alter(ChangeLog), alter(kmymoney2.cpp), alter(mymoneyaccount.cpp), alter(mymoneyfile.h)}
24	{alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(imymoneyserialize.h), alter(mymoneyseqaccessmgr.cpp), alter(mymoneyseqaccessmgr.h)}
25	{alter(mymoneyfile.cpp), alter(imymoneyserialize.h), alter(imymoneystorage.h), alter(mymoneyseqaccessmgr.cpp), alter(mymoneyseqaccessmgr.h), alter(mymoneyseqaccessmgrtest.cpp)}
26	{alter(ChangeLog), alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(imymoneystorage.h), alter(mymoneyseqaccessmgr.cpp), alter(mymoneyseqaccessmgr.h), alter(kmymoneyview.cpp)}
27	{alter(kmymoney2.cpp), alter(kgloballedgerview.cpp), alter(kmymoneyview.cpp), alter(register.cpp), alter(register.h), alter(transaction.cpp), alter(transaction.h)}
28	{alter(kmymoney2.cpp), alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(imymoneyserialize.h), alter(mymoneyseqaccessmgr.cpp), alter(mymoneyseqaccessmgr.h), alter(mymoneystoragedump.cpp), alter(mymoneystoragexml.cpp), alter(kmymoneyview.cpp)}
29	{alter(kendingbalancedlg.cpp), alter(kexportdlg.cpp), alter(kfindtransactiondlg.cpp), alter(kimportdlg.cpp), alter(knewaccountdlg.cpp), alter(knewbankdlg.cpp), alter(kreconciledlg.cpp), alter(kcategoriesview.cpp), alter(kmymoneyview.cpp), alter(kpayeesview.cpp)}

30	{alter(kmymoney2.cpp), alter(ieditscheduledialog.cpp), alter(knewaccountdlg.cpp), alter(mymoneyfile.cpp), alter(mymoneyfiletest.cpp), alter(mymoneystorage.h), alter(mymoneyseqaccessmgr.cpp), alter(kcategoriesview.cpp), alter(khomeview.cpp), alter(kpayeesview.cpp)}
31	{alter(imymoneystorage.h), alter(mymoneystoragexml.cpp), alter(kmymoneyview.cpp)}
32	{alter(kmymoney2.cpp), alter(mymoneystoragexml.cpp), alter(kmymoneyview.cpp)}
33	{alter(kmymoney2.cpp), alter(imymoneystorage.h), alter(mymoneyseqaccessmgr.cpp), alter(mymoneyseqaccessmgr.h)}
34	{alter(kmymoney2.h), alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(kcategoriesview.cpp), alter(kmymoneyview.cpp)}
35	{alter(kmainview.cpp), alter(kmymoneyview.cpp), alter(kmymoneyview.h), alter(knewbankdlg.cpp), alter(mymoneyaccount.h), alter(mymoneyfile.h)}
36	{alter(config.h.in), alter(kmymoney2.cpp), alter(knewaccountdlg.cpp), alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(kbanksview.cpp), alter(kcategoriesview.cpp)}
37	{alter(config.h.in), alter(Makefile.common), alter(acinclude.m4.in), alter(am_edit), alter(conf.change.pl), alter(config.pl), alter(config.sub)}
38	{alter(kmymoney2.cpp), alter(kcurrencycalculator.cpp), alter(kcurrencyeditdlg.cpp), alter(knewaccountdlg.cpp), alter(ksplittransactiondlg.cpp), alter(kmymoneyaccountselector.cpp), alter(kmymoneysplittable.cpp)}
39	{alter(kmymoney2.kdevprj), alter(Makefile.am), alter(mymoneyaccount.h), alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(mymoneytransaction.cpp), alter(mymoneytransaction.h)}

40	{alter(kmymoney2.kdevprj), alter(kmymoney2.cpp), alter(kmymoney2.h), alter(Makefile.am), alter(kbanklistitem.cpp), alter(kbanklistitem.h), alter(kbanksview.cpp), alter(kbanksview.h), alter(kmymoneyview.cpp), alter(kmymoneyview.h)}
41	{alter(kmymoney2.cpp), alter(kcsvprogressdlg.cpp), alter(kendingbalancedlg.cpp), alter(kexportdlg.cpp), alter(kfindtransactiondlg.cpp), alter(kimportdlg.cpp), alter(knewaccountdlg.cpp), alter(knewbankdlg.cpp), alter(kreconciledlg.cpp), alter(kcategoriesview.cpp), alter(kmymoneyview.cpp), alter(kpayeesview.cpp)}
42	{alter(kmymoney2.cpp), alter(kcurrencycalculator.cpp), alter(kcurrencyeditdlg.cpp), alter(kendingbalancedlg.cpp), alter(knewaccountdlg.cpp), alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(kcategoriesview.cpp), alter(kmymoneyview.cpp), alter(Makefile.am), alter(kmymoneypriceview.cpp), alter(kmymoneypriceview.h)}
43	{alter(kmymoney2.cpp), alter(knewaccountdlg.cpp), alter(mymoneyfile.cpp,mymoneyfile.h), alter(mymoneyfiletest.cpp), alter(imymoneystorage.h), alter(mymoneyseqaccessmgr.cpp), alter(mymoneyseqaccessmgr.h), alter(kcategoriesview.cpp), alter(khomeview.cpp), alter(kmymoneyview.cpp), alter(kpayeesview.cpp)}
44	{alter(ieditscheduledialog.cpp), alter(keditloanwizard.cpp), alter(kenterscheduledialog.cpp), alter(knewaccountwizard.cpp), alter(kledgerview.cpp), alter(kledgerviewcheckings.cpp), alter(kledgerviewinvestments.cpp), alter(kledgerviewloan.cpp), alter(kmymoneyview.cpp), alter(kpayeesview.cpp), alter(kmymoneyregistercheckings.cpp), alter(kmymoneyregistersearch.cpp)}

45	{alter(ieditscheduledialog.cpp), alter(knewaccountdlg.cpp), alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(mymoneyfiletest.cpp), alter(imymoneystorage.h), alter(mymoneyseqaccessmgr.cpp), alter(mymoneyseqaccessmgr.h), alter(kcategoriesview.cpp), alter(khomeview.cpp), alter(kmymoneyview.cpp), alter(kpayeesview.cpp)}
46	{alter(kmymoney2.cpp), alter(ieditscheduledialog.cpp), alter(knewaccountdlg.cpp), alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(mymoneyfiletest.cpp), alter(imymoneystorage.h), alter(mymoneyseqaccessmgr.cpp), alter(mymoneyseqaccessmgr.h), alter(khomeview.cpp), alter(kmymoneyview.cpp), alter(kpayeesview.cpp)}
47	{alter(kmymoney2.cpp), alter(ieditscheduledialog.cpp), alter(knewaccountdlg.cpp), alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(mymoneyfiletest.cpp), alter(imymoneystorage.h), alter(mymoneyseqaccessmgr.cpp), alter(kcategoriesview.cpp), alter(khomeview.cpp), alter(kmymoneyview.cpp), alter(kpayeesview.cpp)}
48	{alter(kmymoney2.cpp), alter(ieditscheduledialog.cpp), alter(knewaccountdlg.cpp), alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(mymoneyfiletest.cpp), alter(imymoneystorage.h), alter(mymoneyseqaccessmgr.cpp), alter(mymoneyseqaccessmgr.h), alter(kcategoriesview.cpp), alter(khomeview.cpp), alter(kmymoneyview.cpp), alter(kpayeesview.cpp)}
49	{alter(ChangeLog), alter(mymoneyqifreader.cpp), alter(mymoneyfile.h), alter(mymoneyfiletest.cpp), alter(imymoneyserialize.h), alter(imymoneystorage.h), alter(mymoneyseqaccessmgr.cpp), alter(mymoneyseqaccessmgr.h), alter(mymoneyseqaccessmgrtest.cpp), alter(mymoneystoragedump.cpp), alter(kcategoriesview.cpp), alter(khomeview.cpp), alter(kmymoneyview.cpp), alter(kpayeesview.cpp)}

50	{alter(ChangeLog), alter(kmymoney2.cpp), alter(mymoneyqifreader.cpp), alter(ieditscheduledialog.cpp), alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(mymoneyfiletest.cpp), alter(imymoneystorage.h), alter(mymoneyseqaccessmgr.h), alter(mymoneyseqaccessmgrtest.cpp), alter(kcategoriesview.cpp,khomeview.cpp), alter(kmymoneyview.cpp), alter(kpayeesview.cpp)}
51	{alter(ChangeLog), alter(configure.in.in), alter(acinclude.m4.in), alter(kmymoney2.cpp), alter(kmymoney2.h), alter(kmymoneytest.cpp), alter(Makefile.am), alter(Makefile.am), alter(mymoneyofxstatement.cpp), alter(mymoneyofxstatement.h), alter(mymoneystatementreader.cpp), alter(mymoneystatementreader.h), alter(mymoneystatement.cpp), alter(mymoneystatement.h)}
52	{alter(kexportdlgdecl.ui), alter(kimportdlgdecl.ui), alter(knewloanwizarddecl.ui), alter(kofxdirectconnectdlgdecl.ui), alter(konlinequoteconfigurationdecl.ui), alter(ksplitcorrectiondlg.ui), alter(mymoneyqifprofileeditordecl.ui), alter(mymoneybudget.cpp), alter(mymoneybudget.h), alter(kbudgetview.cpp), alter(kbudgetview.h), alter(kbudgetviewdecl.ui), alter(kscheduledviewdecl.ui), alter(Makefile.am), alter(kmymoneyaccounttree.cpp), alter(kmymoneyaccounttree.h), alter(kmymoneyaccounttreebudget.cpp), alter(kmymoneyaccounttreebudget.h)}
53	{alter(mymoneyaccount.cpp), alter(mymoneyaccount.h), alter(mymoneycheckingaccount.cpp), alter(mymoneycheckingaccount.h), alter(mymoneycheckingaccounttest.h), alter(mymoneycheckingtransaction.cpp), alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(mymoneyfiletest.h), alter(mymoneyinstitution.h), alter(mymoneyinstitutiontest.h), alter(mymoneymoneytest.h), alter(mymoneytransaction.h), alter(mymoneytransactiontest.h), alter(mymoneycheckingtransactiontest.h)}

54	{alter(ChangeLog), alter(kmymoney2.cpp), alter(knewuserwizard.cpp), alter(knewuserwizard.h), alter(knewuserwizard_p.h), alter(kpreferencepagedecl.ui), alter(userinfo.cpp), alter(userinfo.h), alter(userinfodecl.ui), alter(mymoneyreport.cpp), alter(khomeview.cpp), alter(kmymoneyview.cpp), alter(kmymoneyview.h)}
55	{alter(Makefile.am), alter(kmymoney2.kdevprj), alter(Makefile.am), alter(kmymoney2.h), alter(kmymoney2ui.rc), alter(kfindtransactiondlg.cpp), alter(kupdatestockpricedlgdecl.ui), alter(Makefile.am), alter(mymoneyaccount.cpp), alter(mymoneyaccount.h), alter(mymoneyequity.cpp), alter(mymoneyequity.h), alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(imymoneyserialize.h), alter(imymoneystorage.h), alter(mymoneyseqaccessmgr.cpp), alter(mymoneyseqaccessmgr.h), alter(mymoneyseqaccessmgrtest.cpp)}
56	{alter(kmymoneyutils.cpp), alter(kmymoneyutils.h), alter(main.cpp), alter(knewaccountdlg.cpp), alter(knewaccountwizard.cpp), alter(knewfiledlg.cpp), alter(ksettingsdlg.cpp), alter(home.html), alter(home_de.de.html), alter(home_fr.fr.html), alter(kcategoriesview.cpp), alter(kledgerview.cpp), alter(kledgerviewcheckings.cpp), alter(kledgerviewloan.cpp), alter(kmymoneyview.cpp), alter(kmymoneyview.h), alter(kmymoneyregister.cpp), alter(de.po), alter(fr.po)}
57	{alter(knewaccountwizard.cpp), alter(knewloanwizard.cpp), alter(transactioneditor.cpp), alter(mymoneybudget.h), alter(mymoneyscheduled.cpp), alter(mymoneyscheduled.h), alter(mymoneyscheduletest.cpp), alter(mymoneyscheduletest.h), alter(mymoneyseqaccessmgr.cpp), alter(mymoneyseqaccessmgrtest.cpp), alter(mymoneystoragedump.cpp), alter(mymoneystoragesql.cpp), alter(khomeview.cpp), alter(kscheduledlistitem.cpp), alter(kscheduledview.cpp), alter(register.cpp), alter(transaction.cpp), alter(keditscheduledlg.cpp), alter(keditscheduledlg.h), alter(ieditscheduledialog.cpp)}

58	{alter(kmymoney2.cpp), alter(kmymoney2.h), alter(mymoneyaccount.h), alter(mymoneyaccounttest.cpp), alter(mymoneyaccounttest.h), alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(mymoneyfiletest.cpp), alter(imymoneyserialize.h), alter(mymoneyseqaccessmgr.cpp), alter(mymoneyseqaccessmgr.h), alter(mymoneystoragedump.cpp), alter(mymoneystoragexml.cpp), alter(mymoneystoragexml.h), alter(kmymoneyview.cpp), alter(kmymoneyview.h)}
59	{alter(bottomleft.png), alter(topleft.png), alter(check-16.png), alter(check-20.png,frozen.png), alter(kmm-frozen.png), alter(lock-16.png), alter(newsplashcvs.png), alter(newsplashcvs2.png), alter(paperclip-diag.png), alter(paperclip-diag16.png), alter(paperclip-vert.png), alter(reconciled-frozen.png), alter(reconciled-frozen2.png)}
60	{alter(details-impexp.docbook), alter(details-investments.docbook), alter(gnucash-import_options.png), alter(Makefile.am), alter(mymoneygncreader.cpp), alter(mymoneygncreader.h), alter(webpricequote.cpp), alter(webpricequote.h), alter(kequitypriceupdatedlg.cpp), alter(kgncimportoptionsdlg.cpp), alter(kgncimportoptionsdlg.h), alter(kgncimportoptionsdlgdecl.ui), alter(knewinvestmentwizard.cpp), alter(knewinvestmentwizard.h)}

ทรานแซคชั่นชุดทดสอบทั้งหมดจะถูกนำไปสร้างเป็นข้อสอบถามสำหรับสถานการณ์ต่างๆ 3 สถานการณ์ตามขั้นตอนวิธีการสร้างข้อสอบถามที่อธิบายไว้ในบทที่ 3 จากทรานแซคชั่นข้างต้นสามารถนำมาแสดงตัวอย่างของข้อสอบถามสำหรับแต่ละสถานการณ์ได้ดังนี้

ตัวอย่างข้อสอบถามสำหรับสถานการณ์การนำทาง

จากทรานแซคชั่น {alter(kmymoney2.cpp), alter(kmymoney2.h), alter(kmymoneyview.h)} สามารถนำมาสร้างข้อสอบถามได้ทั้งหมด 2 ข้อสอบถามดังนี้

- ข้อสอบถามที่ 1 { {alter(kmymoney2.cpp)} , {alter(kmymoney2.h), alter(kmymoneyview.h)} } โดยที่ {alter(kmymoney2.cpp)} คือเซตเหตุการณ์ และ {alter(kmymoney2.h), alter(kmymoneyview.h)} คือเซตของผลลัพธ์ที่คาดไว้
- ข้อสอบถามที่ 2 { {alter(kmymoney2.h)} , {alter(kmymoney2.cpp), alter(kmymoneyview.h)} } โดยที่ {alter(kmymoney2.h)} คือเซตเหตุการณ์ และ {alter(kmymoney2.cpp), alter(kmymoneyview.h)} คือเซตของผลลัพธ์ที่คาดไว้
- ข้อสอบถามที่ 3 { {alter(kmymoneyview.h)} , {alter(kmymoney2.cpp), alter(kmymoney2.h)} } โดยที่ {alter(kmymoneyview.h)} คือเซตเหตุการณ์ และ {alter(kmymoney2.cpp), alter(kmymoney2.h)} คือเซตของผลลัพธ์ที่คาดไว้

ตัวอย่างข้อสอบถามสำหรับสถานการณ์การป้องกันข้อผิดพลาด

จากทรานแซคชั่น {alter(kmymoney2.cpp), alter(kmymoney2.h), alter(kmymoneyview.h)} สามารถนำมาสร้างข้อสอบถามได้ทั้งหมด 3 ข้อสอบถามดังนี้

- ข้อสอบถามที่ 1 { {alter(kmymoney2.cpp), alter(kmymoney2.h)} , {alter(kmymoneyview.h)} } โดยที่ {alter(kmymoney2.cpp), alter(kmymoney2.h)} คือเซตเหตุการณ์ และ {alter(kmymoneyview.h)} คือเซตของผลลัพธ์ที่คาดไว้
- ข้อสอบถามที่ 2 { {alter(kmymoney2.cpp), alter(kmymoneyview.h)} , {alter(kmymoney2.cpp) } } โดยที่ {alter(kmymoney2.h), alter(kmymoneyview.h)} คือเซตเหตุการณ์ และ {alter(kmymoney2.cpp)} คือเซตของผลลัพธ์ที่คาดไว้
- ข้อสอบถามที่ 3 { {alter(kmymoneyview.h), alter(kmymoney2.cpp)} , {alter(kmymoney2.h)} } โดยที่ {alter(kmymoneyview.h), alter(kmymoney2.cpp)} คือเซตเหตุการณ์ และ {alter(kmymoney2.h)} คือเซตของผลลัพธ์ที่คาดไว้

ตัวอย่างข้อสอบถามสำหรับสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

จากทรานแซคชั่น {alter(kmymoney2.cpp), alter(kmymoney2.h), alter(kmymoneyview.h)} สามารถนำมาสร้างข้อสอบถามได้ทั้งหมด 1 ข้อสอบถามดังนี้

- ข้อสอบถามที่ 1 { {alter(kmymoney2.cpp), alter(kmymoney2.h), alter(kmymoneyview.h)}, {} } โดยที่ {alter(kmymoney2.cpp), alter(kmymoney2.h), alter(kmymoneyview.h)} คือเซตเหตุการณ์ และเซตของผลลัพธ์ที่คาดไว้คือเซตว่าง

ภาคผนวก ข
ประเด็นความถูกต้องและน่าเชื่อถือของเครื่องมือทดสอบ

งานวิจัยนี้มีขั้นตอนในการทดสอบทั้งหมด 6 ขั้นตอนดังรายละเอียดในหัวข้อ 3.6 และแสดงในรูปที่ 3-1 ในขั้นตอนทั้ง 6 ขั้นตอนดังกล่าว 5 ขั้นตอนแรกจำเป็นต้องใช้เครื่องมือต่างๆเข้ามาช่วยในการทดสอบซึ่งแสดงรายละเอียดในหัวข้อที่ 3.7 และรูปที่ 3-5 ขั้นตอนแรกหรือขั้นตอนการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ ผู้วิจัยเลือกใช้ส่วนการจัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูลกับข้อมูลซอฟต์แวร์อาร์ไคฟว์ของโปรแกรมประยุกต์อีโรส (eROSE) (Zimmermann et al., 2005) ส่วนขั้นตอนอื่นๆได้แก่ ขั้นตอนการสร้างข้อสอบถามสำหรับการทดสอบ 3 สถานการณ์ ขั้นตอนการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์กับข้อมูลซอฟต์แวร์อาร์ไคฟว์ทั้ง 2 ตัวแบบสำหรับ 3 สถานการณ์ ขั้นตอนการสร้างเซตของคำแนะนำสำหรับเหตุการณ์ และขั้นตอนการประเมินผลการทดสอบ ผู้วิจัยจำเป็นต้องพัฒนาเครื่องมือทดสอบขึ้นมาเองโดยใช้ภาษาพีเอชพี (PHP) รวมกับระบบจัดการฐานข้อมูลชื่อพีเอชพีมายแอดมิน (PHPMyAdmin Database Management System) ผู้วิจัยจำเป็นต้องคำนึงถึงประเด็นความถูกต้องและน่าเชื่อถือของเครื่องมือทดสอบเหล่านั้นด้วย ผู้วิจัยจึงต้องตรวจสอบความถูกต้องและแม่นยำของเครื่องมือเหล่านั้นในภาคผนวกนี้

ประเด็นความถูกต้องและน่าเชื่อถือของเครื่องมือทดสอบที่ผู้วิจัยพัฒนาขึ้นมาเองทั้งหมด 4 เครื่องมือ แสดงดังต่อไปนี้

เครื่องมือสร้างข้อสอบถามสำหรับ 3 สถานการณ์

ข้อมูลออกผลลัพธ์ของเครื่องมือนี้ คือ ข้อสอบถามสำหรับ 3 สถานการณ์ แสดงในหัวข้อ 3.6.2 ผลลัพธ์ที่ออกมารวม 3 สถานการณ์จะได้ทั้งหมด 962 ข้อถามสอบ แบ่งเป็นสถานการณ์การนำทาง 451 ข้อสอบถาม สถานการการณ์การป้องกันการเกิดข้อผิดพลาด 451 ข้อสอบถาม และสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว 60 ข้อสอบถาม ในส่วนนี้ผู้วิจัยสามารถใช้การตรวจสอบแบบเดินผ่าน (Walkthrough) หรือการตรวจสอบเฉพาะผลลัพธ์อย่างไม่เป็นทางการกับผลลัพธ์ทั้งหมด 962 ข้อสอบถามแยกตามข้อกำหนดของแต่ละสถานการณ์ได้ ผลการตรวจสอบแสดงให้เห็นว่าเครื่องมือสร้างข้อสอบถามสำหรับ 3 สถานการณ์ให้ผลลัพธ์ที่มีความถูกต้องและแม่นยำทั้งหมด

เครื่องมือค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 และเครื่องมือค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2

ข้อมูลออกผลลัพธ์ของเครื่องมือทั้ง 2 เครื่องมือนี้ คือ กฎความสัมพันธ์ด้วยตัวแบบที่ 1 รวมถึงค่าสนับสนุนค่าความเชื่อมั่นของกฎความสัมพันธ์และกฎความสัมพันธ์ด้วยตัวแบบที่ 2 รวมถึงค่าสนับสนุนค่าความเชื่อมั่นใหม่ของกฎความสัมพันธ์ แสดงในหัวข้อ 3.6.3 กฎความสัมพันธ์ผลลัพธ์ที่สร้างออกมาคือกฎความสัมพันธ์ทั้งหมดที่มีค่าความน่าสนใจสูงกว่าค่าที่กำหนดไว้ ผู้วิจัยจึงจำเป็นต้องเลือกตัวแทนของกฎความสัมพันธ์ทั้งหมดขึ้นมาตรวจสอบหรือการตรวจสอบแบบสุ่มตรวจ (Random Inspection) กับกฎความสัมพันธ์ด้วยตัวแบบที่ 1 และกฎความสัมพันธ์ด้วยตัวแบบที่ 2 การสุ่มจะเป็นการสุ่มเลือกเซตรายการที่มาก่อนที่เป็นตัวแทนของเซตรายการที่มาก่อนขนาดต่างๆกัน (สั้น กลางและยาว) และความถี่ในการปรากฏต่างๆกัน (พบบ่อย และพบไม่บ่อย) กฎความสัมพันธ์ที่มีเซตรายการที่มาก่อนเหมือนกับเซตรายการที่มาก่อนที่กำหนดทั้งหมดจะถูกเลือกออกมา เซตรายการที่มาก่อนที่ถูกเลือกขึ้นมาเป็นตัวแทนทั้งหมด 10 เซตรายการดังนี้

กำหนดให้ alter(*x*) คือ การเปลี่ยนแปลงแก้ไขในมิติแก้ไข (alter) เอนทิตี้ *x* โดยที่ *x* เป็นเอนทิตี้ในระดับของแฟ้มข้อมูล แทนการเขียนแบบเต็ม alter(file, *x*, …) เพื่อความสะดวกในการแสดงข้อมูล

1. {alter(mymoneyfile.h)}
2. {alter(knewaccountwizard.cpp)}
3. {alter(kmymoney2.cpp), alter(kmymoney2.h)}
4. {alter(ChangeLog), alter(kmymoney2.cpp), alter(mymoneyaccount.cpp)}
5. {alter(imymoneystorage.h), alter(kmymoneyview.cpp)}
6. {alter(config.h.in), alter(acinclude.m4.in), alter(am_edit), alter(conf.change.pl), alter(config.pl), alter(config.sub)}
7. {alter(knewaccountdlg.cpp), alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(mymoneyfiletest.cpp), alter(imymoneystorage.h), alter(mymoneyseqaccessmgr.cpp), alter(mymoneyseqaccessmgr.h), alter(kcategoriesview.cpp), alter(khomeview.cpp),

alter(kmymoneyview.cpp), alter(kpayeesview.cpp)}

8. {alter(kmymoney2.cpp), alter(knewaccountdlg.cpp), alter(mymoneyfile.cpp), alter(mymoneyfile.h), alter(mymoneyfiletest.cpp), alter(imymoneystorage.h), alter(mymoneyseqaccessmgr.cpp), alter(mymoneyseqaccessmgr.h), alter(kcategoriesview.cpp), alter(khomeview.cpp), alter(kmymoneyview.cpp), alter(kpayeesview.cpp)}
9. {alter(ChangeLog), alter(configure.in.in), alter(kmymoney2.cpp), alter(kmymoney2.h), alter(kmymoneytest.cpp), alter(Makefile.am), alter(Makefile.am), alter(mymoneyofxstatement.cpp), alter(mymoneyofxstatement.h), alter(mymoneystatementreader.cpp), alter(mymoneystatementreader.h), alter(mymoneystatement.cpp), alter(mymoneystatement.h)}
10. {alter(kexportdlgdecl.ui), alter(kimportdlgdecl.ui), alter(kofxdirectconnectdlgdecl.ui), alter(konlinequoteconfigurationdecl.ui), alter(ksplitcorrectiondlg.ui), alter(mymoneyqifprofileeditordecl.ui), alter(mymoneybudget.cpp), alter(mymoneybudget.h), alter(kbudgetview.cpp), alter(kbudgetview.h), alter(kbudgetviewdecl.ui), alter(kscheduledviewdecl.ui), alter(Makefile.am), alter(kmymoneyaccounttree.cpp), alter(kmymoneyaccounttree.h), alter(kmymoneyaccounttreebudget.cpp), alter(kmymoneyaccounttreebudget.h)}

เมื่อได้กฎความสัมพันธ์ที่มีเซตรายการที่มาก่อนเหมือนเซตรายการข้างต้นนี้แล้ว นำกฎความสัมพันธ์เหล่านั้นมาจัดเป็นกลุ่มตามเซตรายการที่มีก่อนที่เหมือนกัน และตรวจสอบความถูกต้องของรายละเอียดต่อไปนี้

1) ตรวจสอบค่าสนับสนุนนับของเซตรายการที่มาก่อนภายในกลุ่มเดียวกันจะต้องเท่ากันทั้งหมด

2) ตรวจสอบภายในกลุ่มเดียวกันว่าไม่มีกฎความสัมพันธ์ที่มีเซตรายการที่ตามมาเหมือนกัน
3) ตรวจสอบค่าสนับสนุนนับของกฎความสัมพันธ์แต่ละตัวในกลุ่มว่าคำนวณถูกต้องหรือไม่
4) ตรวจสอบค่าความเชื่อมั่น/ค่าความเชื่อมั่นใหม่ของกฎความสัมพันธ์แต่ละตัวในกลุ่มว่าคำนวณถูกต้องหรือไม่

ผลการตรวจสอบแสดงให้เห็นว่าเครื่องมือค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 และเครื่องมือค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 เครื่องมือละ 2 ครั้งให้ผลลัพธ์ที่มีความถูกต้องและแม่นยำทั้งหมด

เครื่องมือสร้างคำแนะนำสำหรับเหตุการณ์

ข้อมูลออกผลลัพธ์ของเครื่องมือนี้ คือ เซตของคำแนะนำสำหรับเหตุการณ์ที่เป็นข้อสอบถามที่ได้มาจากเครื่องมือสร้างข้อสอบถามสำหรับ 3 สถานการณ์ แสดงในหัวข้อ 3.6.4 เครื่องมือนี้มีความต่อเนื่องมาจากเครื่องมือก่อนหน้านี้โดยตรง กล่าวคือกฎความสัมพันธ์ที่ได้มาจากเครื่องมือค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 1 และเครื่องมือค้นหากฎความสัมพันธ์ด้วยตัวแบบที่ 2 จะถูกนำมาค้นหากฎความสัมพันธ์ที่มีเซตรายการที่มาก่อนเหมือนกับข้อสอบถาม เมื่อได้มาแล้วกฎความสัมพันธ์เหล่านั้นจะถูกนำมาจัดอันดับความน่าสนใจตามค่าความเชื่อมั่น/ค่าความเชื่อมั่นใหม่ของกฎความสัมพันธ์ เซตรายการที่ตามของกฎความสัมพันธ์ 10 อันดับแรกจะถูกนำมายูเนียนกันเป็นเซตของกฎความสัมพันธ์ ผู้วิจัยตรวจสอบความถูกต้องของเครื่องมือนี้โดยการตรวจสอบแบบสุ่มตรวจ (Random Inspection) ผู้วิจัยใช้เซตรายการ 10 เซตที่สุ่มมาจากหัวข้อที่แล้วมาเป็นข้อสอบถาม เซตของคำแนะนำที่ได้มาจากเครื่องมือนี้จะถูกมาตรวจสอบดังต่อไปนี้

1) ตรวจสอบสมาชิกของเซตคำแนะนำว่าสมาชิกที่ซ้ำกันหรือไม่
2) ตรวจสอบสมาชิกของเซตคำแนะนำว่าสมาชิกทุกตัวมาเซตรายการที่ตามของกฎความสัมพันธ์ที่อยู่ใน 10 อันดับแรกหรือไม่
3) ตรวจสอบสมาชิกของเซตรายการที่ตามมาของกฎความสัมพันธ์ 10 อันดับแรกว่ามีสมาชิกตัวใดไม่ได้อยู่ในเซตคำแนะนำ

4) กรณีที่มีกฎความสัมพันธ์ที่ผ่านค่าความเชื่อมั่นขั้นต่ำ/ค่าความเชื่อมั่นใหม่ขั้นต่ำ น้อยกว่า 10 กฎ สมาชิกของเซตคำแนะนำก็จะต้องมาจากกฎความสัมพันธ์น้อยกว่า 10 กฎนั้น

ผลการตรวจสอบแสดงให้เห็นว่าเครื่องมือสร้างคำแนะนำสำหรับเหตุการณ์ทั้งหมด 2 ครั้ง ให้ผลลัพธ์ที่มีความถูกต้องและแม่นยำทั้งหมด

เครื่องมือประเมินผลการทดสอบ

ข้อมูลออกผลลัพธ์ของเครื่องมือนี้ คือ ค่าประสิทธิภาพของการทดสอบทั้ง 6 การทดสอบ ได้แก่ค่าเอฟเมเซอร์และค่าผลสะท้อนกลับ แสดงในหัวข้อ 3.6.5 เครื่องมือนี้มีความต่อเนื่องมาจากเครื่องมือก่อนหน้านี้โดยตรง กล่าวคือเซตคำแนะนำที่ได้มาจากเครื่องมือสร้างคำแนะนำสำหรับเหตุการณ์จะถูกนำมาประเมินกับเซตผลลัพธ์ที่คาดไว้ของข้อสอบถามที่ได้มาจากเครื่องมือสร้างข้อสอบถามสำหรับ 3 สถานการณ์ ผู้วิจัยตรวจสอบความถูกต้องของเครื่องมือนี้โดยการตรวจสอบแบบสุ่มตรวจ (Random Inspection) ผู้วิจัยใช้เซตรายการ 10 เซตที่สุ่มมาจากหัวข้อที่แล้วมาเป็นข้อสอบถาม ค่าประสิทธิภาพของแต่ละข้อสอบถามจะถูกนำมาตรวจสอบความถูกต้องในการคำนวณดังต่อไปนี้

1) ค่าประสิทธิภาพที่ได้อยู่ในช่วงพิสัยที่เป็นไปได้ของค่าประสิทธิภาพนั้นหรือไม่
2) ในกรณีที่เซตคำแนะนำเป็นเซตว่าง ทำให้การคำนวณค่าความถูกต้องมีส่วนเป็น 0 ค่าความถูกต้องนั้นถูกกำหนดให้มีค่าเป็น 1 หรือไม่
3) ในกรณีที่เซตของผลลัพธ์ที่คาดไว้เป็นเซตว่าง ทำให้การคำนวณค่าเรียกคืนมีส่วนเป็น 0 ค่าเรียกคืนนั้นถูกกำหนดให้มีค่าเป็น 1 หรือไม่
4) การคำนวณค่าประสิทธิภาพถูกต้องหรือไม่

ผลการตรวจสอบแสดงให้เห็นว่าเครื่องมือประเมินผลการทดสอบทั้งหมด 2 ครั้งให้ผลลัพธ์ที่มีความถูกต้องและแม่นยำทั้งหมด

ภาคผนวก ค

ตารางผลการทดสอบ

ตารางที่ ค-1 แสดงค่าเอฟเมสเซอร์ของการทดสอบสถานการณ์การนำทาง

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
1	0.0000	0.1333
2	0.3333	0.2000
3	0.2857	0.2000
4	0.3333	0.2500
5	0.2857	0.2500
6	0.2857	0.2500
7	0.2500	0.2000
8	0.2857	0.2500
9	0.3333	0.2857
10	0.4000	0.2500
11	0.0000	0.1333
12	0.2857	0.2222
13	0.3333	0.2500
14	0.2857	0.2500
15	0.2500	0.2000
16	0.2857	0.2857
17	0.3333	0.2500
18	0.3333	0.2000
19	0.3333	0.2500
20	0.3333	0.2500
21	0.0000	0.0000
22	0.0000	0.0000
23	0.0000	0.0000
24	0.0000	0.0000
25	0.0000	0.0000
26	0.0000	0.0000
27	0.0000	0.2857
28	0.0000	0.2857
29	0.0000	0.0000
30	0.0000	0.0000

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
31	0.2000	0.1538
32	0.2222	0.1333
33	0.0000	0.0000
34	0.0000	0.0000
35	0.0000	0.0000
36	0.2857	0.3333
37	0.0000	0.0000
38	0.2857	0.2857
39	0.2500	0.1818
40	0.0000	0.0000
41	0.5714	0.5714
42	0.5714	0.4444
43	0.6667	0.3636
44	0.5000	0.4444
45	0.6667	0.4444
46	0.5714	0.5000
47	0.3333	0.2353
48	0.2500	0.2500
49	0.5000	0.5000
50	0.2222	0.3333
51	0.5000	0.5714
52	0.8000	0.7692
53	0.8000	0.8333
54	0.7273	0.8333
55	0.8000	0.6667
56	0.5000	0.5714
57	0.8000	0.7692
58	0.8000	0.8333
59	0.7273	0.8333
60	0.8000	0.8333

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
61	1.0000	0.7692
62	0.2000	0.0952
63	0.7143	0.7500
64	0.7692	0.7500
65	0.6667	0.7143
66	0.7692	0.7143
67	0.6667	0.7143
68	0.1538	0.1538
69	0.1818	0.1818
70	0.3333	0.6667
71	0.3333	0.3333
72	0.7273	0.7273
73	0.6000	0.6667
74	0.7273	0.5000
75	0.7273	0.7692
76	0.1538	0.1538
77	0.5333	0.7059
78	0.4286	0.7059
79	0.6154	0.7500
80	0.5714	0.5333
81	0.6154	0.5333
82	0.7143	0.7692
83	0.6667	0.7500
84	0.1429	0.1429
85	0.4000	0.6250
86	0.1176	0.1429
87	0.3529	0.5556
88	0.1176	0.2667
89	0.1250	0.2353
90	0.4000	0.5882

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
91	0.6667	0.6667
92	0.2667	0.3529
93	0.0000	0.0000
94	0.2500	0.3333
95	0.0000	0.0000
96	0.1250	0.0000
97	0.2500	0.4706
98	0.3750	0.3333
99	0.2857	0.3750
100	0.2857	0.2500
101	0.2667	0.2500
102	0.2667	0.4706
103	0.2667	0.4706
104	0.2500	0.4444
105	0.0000	0.0000
106	0.0000	0.2000
107	0.0000	0.0000
108	0.2857	0.2857
109	0.2222	0.2000
110	0.2500	0.2500
111	0.0000	0.0000
112	0.5000	0.4000
113	0.4444	0.4000
114	0.5000	0.4000
115	0.2222	0.3636
116	0.1818	0.3077
117	0.2000	0.3077
118	0.2000	0.1667
119	0.2000	0.2000
120	0.4444	0.3636

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
121	0.1818	0.3077
122	0.2000	0.4000
123	0.2857	0.1538
124	0.1818	0.1429
125	0.1818	0.0000
126	0.2857	0.0000
127	0.0000	0.0000
128	0.3077	0.4286
129	0.3077	0.2667
130	0.1667	0.2667
131	0.3077	0.1333
132	0.3333	0.4286
133	0.1429	0.0000
134	0.5455	0.5455
135	0.5333	0.5455
136	0.4286	0.5455
137	0.6154	0.5714
138	0.7143	0.7692
139	0.5455	0.5455
140	0.0000	0.0000
141	0.2857	0.4615
142	0.3636	0.4286
143	0.1538	0.1429
144	0.2500	0.3529
145	0.1333	0.1176
146	0.3077	0.3529
147	0.1429	0.1818
148	0.1667	0.1667
149	0.5000	0.6667
150	0.3077	0.1333

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
151	0.3333	0.1333
152	0.6154	0.5333
153	0.6667	0.6154
154	0.3750	0.1538
155	0.4615	0.4615
156	0.4615	0.4000
157	0.7143	0.4211
158	0.3333	0.4000
159	0.5333	0.4706
160	0.4615	0.6250
161	0.4286	0.4286
162	0.5000	0.5882
163	0.1250	0.1250
164	0.2500	0.5556
165	0.4706	0.6667
166	0.3158	0.3750
167	0.4211	0.6000
168	0.3158	0.4706
169	0.2222	0.3158
170	0.4706	0.7368
171	0.7059	0.7059
172	0.3529	0.4211
173	0.1176	0.1176
174	0.3333	0.4000
175	0.1250	0.1250
176	0.5263	0.7778
177	0.5000	0.6316
178	0.4706	0.4444
179	0.3333	0.4211
180	0.2222	0.3000

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
181	0.1176	0.3000
182	0.2353	0.3158
183	0.1176	0.1176
184	0.2000	0.4545
185	0.6000	0.7619
186	0.4211	0.9524
187	0.1250	0.1250
188	0.3333	0.5263
189	0.5556	0.6000
190	0.4706	0.6000
191	0.5000	0.6667
192	0.5000	0.5556
193	0.4706	0.4444
194	0.5000	0.4444
195	0.3529	0.5263
196	0.3529	0.5263
197	0.1176	0.1176
198	0.3333	0.5000
199	0.3333	0.5000
200	0.7000	0.8182
201	0.3529	0.6000
202	0.2222	0.5263
203	0.3529	0.4444
204	0.3750	0.3529
205	0.3750	0.3529
206	0.4000	0.3333
207	0.1176	0.0000
208	0.2222	0.1000
209	0.4706	0.5556
210	0.6250	0.7059

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
211	0.1111	0.1000
212	0.2222	0.4211
213	0.4444	0.5000
214	0.4706	0.5000
215	0.5000	0.5556
216	0.5000	0.5556
217	0.4706	0.4444
218	0.5000	0.4444
219	0.2353	0.4211
220	0.2353	0.4211
221	0.1111	0.3000
222	0.2222	0.4000
223	0.1250	0.1250
224	0.2222	0.1000
225	0.3333	0.4211
226	0.5556	0.6000
227	0.4706	0.6000
228	0.5000	0.6667
229	0.5000	0.5556
230	0.4706	0.4444
231	0.5000	0.4444
232	0.3529	0.4211
233	0.1176	0.1176
234	0.2222	0.4000
235	0.1250	0.1250
236	0.2222	0.1000
237	0.3333	0.5263
238	0.4444	0.5000
239	0.3529	0.5000
240	0.3750	0.5556

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
241	0.3750	0.4444
242	0.3529	0.3333
243	0.3529	0.5263
244	0.3529	0.5263
245	0.1176	0.1176
246	0.3333	0.5000
247	0.1176	0.1176
248	0.2105	0.0952
249	0.3158	0.5000
250	0.5263	0.5714
251	0.4444	0.5714
252	0.4706	0.6316
253	0.4706	0.5263
254	0.4444	0.4211
255	0.4706	0.4211
256	0.3333	0.5000
257	0.3333	0.5000
258	0.1111	0.1111
259	0.3158	0.4762
260	0.1250	0.0741
261	0.2857	0.2000
262	0.4211	0.5455
263	0.3333	0.4000
264	0.4444	0.5714
265	0.4444	0.6000
266	0.5263	0.6000
267	0.4444	0.6000
268	0.4444	0.4762
269	0.4211	0.3333
270	0.3158	0.2857

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
271	0.2105	0.3810
272	0.1053	0.1053
273	0.3000	0.2727
274	0.2500	0.1481
275	0.2222	0.2222
276	0.4762	0.4000
277	0.3000	0.1818
278	0.5000	0.5455
279	0.4211	0.5455
280	0.3333	0.5000
281	0.3333	0.5000
282	0.3333	0.5000
283	0.3333	0.3810
284	0.4211	0.3810
285	0.3158	0.4762
286	0.2105	0.2105
287	0.4000	0.3636
288	0.1250	0.0741
289	0.2353	0.0870
290	0.0909	0.0000
291	0.2222	0.2222
292	0.2222	0.2000
293	0.4211	0.2857
294	0.3000	0.1905
295	0.5263	0.6316
296	0.3750	0.3750
297	0.5714	0.4167
298	0.2667	0.2667
299	0.1053	0.0000
300	0.5556	0.4706

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
301	0.4444	0.4706
302	0.1667	0.1379
303	0.0870	0.0769
304	0.1600	0.0000
305	0.0000	0.0000
306	0.0000	0.0000
307	0.3478	0.3810
308	0.1905	0.2000
309	0.2857	0.4348
310	0.2857	0.3636
311	0.3478	0.3636
312	0.3636	0.4000
313	0.3636	0.3636
314	0.0833	0.0000
315	0.0000	0.0714
316	0.2500	0.0769
317	0.5600	0.5000
318	0.1053	0.0000
319	0.0000	0.0000
320	0.3158	0.3478
321	0.4000	0.4348
322	0.4444	0.6667
323	0.6667	0.6667
324	0.6667	0.6667
325	0.6667	0.6667
326	0.1905	0.1739
327	0.2000	0.1739
328	0.4211	0.6957
329	0.5000	0.3810
330	0.4000	0.4000

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
331	0.2727	0.3478
332	0.4000	0.4762
333	0.3000	0.5600
334	0.0000	0.0000
335	0.2667	0.2308
336	0.3529	0.3529
337	0.5000	0.5000
338	0.5000	0.5882
339	0.5000	0.5000
340	0.0000	0.0000
341	0.1250	0.1111
342	0.3333	0.3000
343	0.3333	0.3333
344	0.3750	0.2857
345	0.0000	0.0000
346	0.0000	0.0000
347	0.0000	0.0000
348	0.2857	0.4000
349	0.0741	0.0833
350	0.1600	0.0800
351	0.0833	0.1538
352	0.1667	0.3571
353	0.0000	0.0645
354	0.0000	0.0000
355	0.0690	0.0714
356	0.3200	0.2400
357	0.2500	0.3571
358	0.2400	0.2857
359	0.3200	0.3077
360	0.3200	0.2963

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
361	0.3846	0.3571
362	0.4000	0.3571
363	0.4167	0.4444
364	0.4167	0.4615
365	0.4000	0.4615
366	0.4167	0.4615
367	0.4167	0.4444
368	0.2500	0.3200
369	0.3333	0.4800
370	0.0870	0.2222
371	0.1600	0.1538
372	0.1600	0.2308
373	0.1600	0.0800
374	0.1667	0.3704
375	0.1667	0.1739
376	0.1818	0.1818
377	0.1739	0.1739
378	0.0833	0.1538
379	0.2500	0.2400
380	0.2609	0.1667
381	0.1818	0.1600
382	0.1667	0.0833
383	0.0909	0.1481
384	0.1667	0.2500
385	0.0870	0.0870
386	0.0909	0.0870
387	0.0000	0.0000
388	0.0000	0.0000
389	0.0000	0.0000
390	0.2143	0.1538

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
391	0.0000	0.0000
392	0.0714	0.0714
393	0.0000	0.0000
394	0.0000	0.0000
395	0.2308	0.2963
396	0.2400	0.1538
397	0.2308	0.2308
398	0.2500	0.1600
399	0.0000	0.0769
400	0.0833	0.0741
401	0.0800	0.0833
402	0.0870	0.0741
403	0.0000	0.0000
404	0.0769	0.2143
405	0.1600	0.0000
406	0.0000	0.0000
407	0.3000	0.3000
408	0.3000	0.3636
409	0.2857	0.2500
410	0.4762	0.2500
411	0.2727	0.0000
412	0.4545	0.5000
413	0.3810	0.5000
414	0.4000	0.5455
415	0.4000	0.5217
416	0.3810	0.3636
417	0.4000	0.3636
418	0.4762	0.5000
419	0.5455	0.6087
420	0.4000	0.5455

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
421	0.2857	0.2857
422	0.3158	0.2500
423	0.0000	0.0000
424	0.0000	0.0000
425	0.0000	0.0000
426	0.0000	0.0000
427	0.0000	0.0000
428	0.0000	0.0000
429	0.0000	0.0000
430	0.0000	0.0000
431	0.0000	0.0000
432	0.0000	0.0000
433	0.0000	0.0000
434	0.0000	0.0000
435	0.0000	0.0000
436	0.0000	0.0000
437	0.3333	0.2400
438	0.1739	0.2727
439	0.0000	0.0000
440	0.0000	0.0000
441	0.4211	0.3333
442	0.3333	0.3333
443	0.2105	0.1905
444	0.2222	0.2222
445	0.1818	0.0000
446	0.4762	0.5000
447	0.1333	0.1333
448	0.2353	0.2500
449	0.3333	0.2222
450	0.3529	0.2500
451	0.2105	0.0000

ตารางที่ ค-2 แสดงค่าเอฟเมสเซอร์ของการทดสอบสถานการณ์การป้องกันการเกิดข้อผิดพลาด

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
1	0.0000	0.1333
2	0.3333	0.2000
3	0.2857	0.2000
4	0.3333	0.2500
5	0.2857	0.2500
6	0.2857	0.2500
7	0.2500	0.2000
8	0.2857	0.2500
9	0.3333	0.2857
10	0.4000	0.2500
11	0.0000	0.1333
12	0.2857	0.2222
13	0.3333	0.2500
14	0.2857	0.2500
15	0.2500	0.2000
16	0.2857	0.2857
17	0.3333	0.2500
18	0.3333	0.2000
19	0.3333	0.2500
20	0.3333	0.2500
21	0.0000	0.0000
22	0.0000	0.0000
23	0.0000	0.0000
24	0.0000	0.0000
25	0.0000	0.0000
26	0.0000	0.0000
27	0.0000	0.2857
28	0.0000	0.2857
29	0.0000	0.0000
30	0.0000	0.0000

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
31	0.2000	0.1538
32	0.2222	0.1333
33	0.0000	0.0000
34	0.0000	0.0000
35	0.0000	0.0000
36	0.2857	0.3333
37	0.0000	0.0000
38	0.2857	0.2857
39	0.2500	0.1818
40	0.0000	0.0000
41	0.4000	0.2000
42	0.2857	0.2222
43	0.2857	0.2222
44	0.2857	0.2857
45	0.3333	0.2000
46	0.2857	0.2857
47	0.3333	0.2222
48	0.3333	0.2000
49	0.2857	0.0000
50	0.2857	0.1538
51	0.2857	0.2500
52	0.2857	0.2000
53	0.3333	0.2222
54	0.3333	0.2222
55	0.0000	0.0000
56	0.2857	0.2500
57	0.2857	0.2000
58	0.3333	0.2222
59	0.3333	0.2222
60	0.3333	0.2222

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
61	0.2857	0.2222
62	0.2500	0.2000
63	0.2222	0.3333
64	0.2222	0.3333
65	0.2222	0.2857
66	0.2222	0.2857
67	0.2222	0.3333
68	0.2222	0.2500
69	0.3333	0.3333
70	0.3333	0.3333
71	0.3333	0.2500
72	0.3333	0.3333
73	0.3333	0.3333
74	0.3333	0.3333
75	0.3333	0.3333
76	0.2857	0.2222
77	0.2857	0.2500
78	0.2857	0.2500
79	0.2857	0.2500
80	0.2857	0.2500
81	0.2857	0.2500
82	0.2857	0.2500
83	0.2857	0.1818
84	0.2857	0.2500
85	0.2857	0.0000
86	0.2857	0.0000
87	0.2857	0.0000
88	0.2857	0.0000
89	0.2857	0.0000
90	0.2857	0.1667

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
91	0.2857	0.1818
92	0.2857	0.1818
93	0.2857	0.1818
94	0.2857	0.1818
95	0.4000	0.0000
96	0.4000	0.2222
97	0.4000	0.2000
98	0.4000	0.1667
99	0.4000	0.1667
100	0.4000	0.1818
101	0.4000	0.1818
102	0.4000	0.2000
103	0.4000	0.2000
104	0.4000	0.1818
105	0.1818	0.1538
106	0.0000	0.0000
107	0.0000	0.0000
108	0.2500	0.2222
109	0.0000	0.0000
110	0.2857	0.2857
111	0.0000	0.0000
112	0.4000	0.2222
113	0.4000	0.2857
114	0.4000	0.2500
115	0.0000	0.0000
116	0.2857	0.1818
117	0.2857	0.0000
118	0.0000	0.0000
119	0.2857	0.2222
120	0.2222	0.2222

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
121	0.2222	0.2222
122	0.2222	0.0000
123	0.2222	0.2222
124	0.2222	0.2500
125	0.2222	0.2222
126	0.0000	0.0000
127	0.3333	0.3333
128	0.3333	0.3333
129	0.3333	0.3333
130	0.3333	0.3333
131	0.3333	0.3333
132	0.3333	0.3333
133	0.0000	0.0000
134	0.0000	0.0000
135	0.0000	0.0000
136	0.0000	0.0000
137	0.0000	0.0000
138	0.0000	0.0000
139	0.0000	0.0000
140	0.0000	0.0000
141	0.1818	0.0000
142	0.1818	0.0000
143	0.1818	0.2222
144	0.0000	0.0000
145	0.1818	0.1818
146	0.1818	0.1818
147	0.5000	0.5000
148	0.3333	0.3333
149	0.5000	0.5000
150	0.5000	0.5000

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
151	0.5000	0.5000
152	0.5000	0.5000
153	0.5000	0.5000
154	0.0000	0.0000
155	0.0000	0.0000
156	0.0000	0.0000
157	0.0000	0.0000
158	0.0000	0.0000
159	0.0000	0.0000
160	0.0000	0.0000
161	0.0000	0.0000
162	0.0000	0.0000
163	0.1818	0.0000
164	0.1818	0.2000
165	0.1818	0.1818
166	0.1818	0.0000
167	0.1818	0.2000
168	0.1818	0.2000
169	0.1818	0.1818
170	0.1818	0.2000
171	0.1818	0.2000
172	0.1818	0.1818
173	0.1818	0.1818
174	0.1818	0.1818
175	0.1818	0.0000
176	0.1818	0.1667
177	0.1818	0.1667
178	0.1818	0.2500
179	0.1818	0.1818
180	0.1818	0.2222

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
181	0.1818	0.2222
182	0.1818	0.2500
183	0.1818	0.2222
184	0.1818	0.2857
185	0.1818	0.2500
186	0.1818	0.2500
187	0.2000	0.0000
188	0.2000	0.0000
189	0.2000	0.2000
190	0.2000	0.2000
191	0.2000	0.1667
192	0.2000	0.2000
193	0.2000	0.2000
194	0.2000	0.2000
195	0.2000	0.2000
196	0.2000	0.2000
197	0.2000	0.0000
198	0.2000	0.0000
199	0.2500	0.2500
200	0.2500	0.2500
201	0.2500	0.2857
202	0.2500	0.2857
203	0.2500	0.2500
204	0.2500	0.2857
205	0.2500	0.2500
206	0.2500	0.2857
207	0.2500	0.2500
208	0.2500	0.2500
209	0.2500	0.2500
210	0.2500	0.2500

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
211	0.2000	0.2000
212	0.2000	0.1818
213	0.2000	0.1667
214	0.2000	0.1667
215	0.2000	0.1538
216	0.2000	0.1667
217	0.2000	0.1667
218	0.2000	0.1667
219	0.2000	0.1538
220	0.2000	0.1538
221	0.0000	0.0000
222	0.2000	0.1667
223	0.0000	0.0000
224	0.2000	0.2000
225	0.2000	0.1818
226	0.2000	0.1333
227	0.0000	0.1429
228	0.0000	0.1538
229	0.0000	0.1667
230	0.0000	0.1538
231	0.0000	0.1667
232	0.0000	0.1429
233	0.0000	0.1538
234	0.0000	0.1429
235	0.2000	0.0000
236	0.2000	0.2000
237	0.2000	0.2000
238	0.2000	0.1538
239	0.2000	0.1538
240	0.2000	0.2000

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
241	0.2000	0.1818
242	0.2000	0.1818
243	0.2000	0.1667
244	0.2000	0.0000
245	0.2000	0.1667
246	0.2000	0.1818
247	0.1818	0.0000
248	0.1818	0.2000
249	0.1818	0.1667
250	0.1818	0.1538
251	0.1818	0.1538
252	0.1818	0.1667
253	0.1818	0.1667
254	0.1818	0.1538
255	0.1818	0.1667
256	0.1818	0.1429
257	0.1818	0.1538
258	0.1818	0.1538
259	0.1818	0.1538
260	0.2500	0.0000
261	0.2500	0.2500
262	0.2500	0.2500
263	0.2500	0.2857
264	0.2500	0.2857
265	0.2500	0.2857
266	0.2500	0.2500
267	0.2500	0.2857
268	0.2500	0.2857
269	0.2500	0.2500
270	0.2500	0.2857

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
271	0.2500	0.2500
272	0.2500	0.2500
273	0.2500	0.2500
274	0.2500	0.0000
275	0.2500	0.0000
276	0.2500	0.2500
277	0.2500	0.2500
278	0.2500	0.2500
279	0.2500	0.2500
280	0.2500	0.2500
281	0.2500	0.2500
282	0.2500	0.2500
283	0.2500	0.2500
284	0.2500	0.2500
285	0.2500	0.2500
286	0.2500	0.2500
287	0.2500	0.2500
288	0.0000	0.0000
289	0.0000	0.0000
290	0.0000	0.0000
291	0.0000	0.0000
292	0.0000	0.0000
293	0.0000	0.0000
294	0.0000	0.0000
295	0.0000	0.0000
296	0.0000	0.0000
297	0.0000	0.0000
298	0.0000	0.0000
299	0.0000	0.0000
300	0.0000	0.0000

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
301	0.0000	0.0000
302	0.0000	0.0000
303	0.0000	0.0000
304	0.0000	0.0000
305	0.0000	0.0000
306	0.0000	0.0000
307	0.0000	0.0000
308	0.0000	0.0000
309	0.0000	0.0000
310	0.0000	0.0000
311	0.0000	0.0000
312	0.0000	0.0000
313	0.0000	0.0000
314	0.0000	0.0000
315	0.0000	0.0000
316	0.0000	0.0000
317	0.0000	0.0000
318	0.0000	0.0000
319	0.0000	0.0000
320	0.0000	0.0000
321	0.0000	0.0000
322	0.0000	0.0000
323	0.0000	0.0000
324	0.0000	0.0000
325	0.0000	0.0000
326	0.0000	0.0000
327	0.0000	0.0000
328	0.0000	0.0000
329	0.0000	0.0000
330	0.0000	0.0000

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
331	0.0000	0.0000
332	0.0000	0.0000
333	0.0000	0.0000
334	0.0000	0.0000
335	0.0000	0.0000
336	0.0000	0.0000
337	0.0000	0.0000
338	0.0000	0.0000
339	0.0000	0.0000
340	0.0000	0.0000
341	0.0000	0.0000
342	0.0000	0.0000
343	0.0000	0.0000
344	0.0000	0.0000
345	0.0000	0.0000
346	0.0000	0.0000
347	0.0000	0.0000
348	0.0000	0.0000
349	0.0000	0.0000
350	0.0000	0.0000
351	0.0000	0.0000
352	0.0000	0.0000
353	0.0000	0.0000
354	0.0000	0.0000
355	0.0000	0.0000
356	0.0000	0.0000
357	0.0000	0.0000
358	0.0000	0.0000
359	0.0000	0.0000
360	0.0000	0.0000

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
361	0.0000	0.0000
362	0.0000	0.0000
363	0.0000	0.0000
364	0.0000	0.0000
365	0.0000	0.0000
366	0.0000	0.0000
367	0.0000	0.0000
368	0.0000	0.0000
369	0.0000	0.0000
370	0.0000	0.0000
371	0.0000	0.0000
372	0.0000	0.0000
373	0.0000	0.0000
374	0.0000	0.0000
375	0.0000	0.0000
376	0.0000	0.0000
377	0.0000	0.0000
378	0.0000	0.0000
379	0.0000	0.0000
380	0.0000	0.0000
381	0.0000	0.0000
382	0.0000	0.0000
383	0.0000	0.0000
384	0.0000	0.0000
385	0.0000	0.0000
386	0.0000	0.0000
387	0.0000	0.0000
388	0.0000	0.0000
389	0.0000	0.0000
390	0.0000	0.0000

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
391	0.0000	0.0000
392	0.0000	0.0000
393	0.0000	0.0000
394	0.0000	0.0000
395	0.0000	0.0000
396	0.0000	0.0000
397	0.0000	0.0000
398	0.0000	0.0000
399	0.0000	0.0000
400	0.0000	0.0000
401	0.0000	0.0000
402	0.0000	0.0000
403	0.0000	0.0000
404	0.0000	0.0000
405	0.0000	0.0000
406	0.0000	0.0000
407	0.0000	0.0000
408	0.0000	0.0000
409	0.0000	0.0000
410	0.0000	0.0000
411	0.0000	0.0000
412	0.0000	0.0000
413	0.0000	0.0000
414	0.0000	0.0000
415	0.0000	0.0000
416	0.0000	0.0000
417	0.0000	0.0000
418	0.0000	0.0000
419	0.0000	0.0000
420	0.0000	0.0000

	ค่าเอฟเมสเซอร์	
	ตัวแบบที่ 1	ตัวแบบที่ 2
421	0.0000	0.0000
422	0.0000	0.0000
423	0.0000	0.0000
424	0.0000	0.0000
425	0.0000	0.0000
426	0.0000	0.0000
427	0.0000	0.0000
428	0.0000	0.0000
429	0.0000	0.0000
430	0.0000	0.0000
431	0.0000	0.0000
432	0.0000	0.0000
433	0.0000	0.0000
434	0.0000	0.0000
435	0.0000	0.0000
436	0.0000	0.0000
437	0.0000	0.0000
438	0.0000	0.0000
439	0.0000	0.0000
440	0.0000	0.0000
441	0.0000	0.0000
442	0.0000	0.0000
443	0.0000	0.0000
444	0.0000	0.0000
445	0.0000	0.0000
446	0.0000	0.0000
447	0.0000	0.0000
448	0.0000	0.0000
449	0.0000	0.0000
450	0.0000	0.0000
451	0.0000	0.0000

ตารางที่ ค-3 แสดงข้อสอบถามที่ได้เซตของคำแนะนำเป็นเซตว่างในการทดสอบสถานการณ์การเปลี่ยนแปลงแก้ไขที่สมบูรณ์แล้ว

	เซตของคำแนะนำที่เป็นเซตว่าง	
	ตัวแบบที่ 1	ตัวแบบที่ 2
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12	✓	✓
13		
14		
15		
16		
17		
18		
19		
20		

	เซตของคำแนะนำที่เป็นเซตว่าง	
	ตัวแบบที่ 1	ตัวแบบที่ 2
21		
22		
23		
24		
25		
26		
27		
28		
29		
30		
31		
32		
33		
34		
35		
36		
37	✓	✓
38		
39		
40	✓	✓

	เซตของคำแนะนำที่เป็นเซตว่าง	
	ตัวแบบที่ 1	ตัวแบบที่ 2
41	✓	✓
42	✓	✓
43		
44	✓	✓
45		
46		
47		
48	✓	✓
49	✓	✓
50	✓	✓
51	✓	✓
52	✓	✓
53	✓	✓
54	✓	✓
55	✓	✓
56	✓	✓
57	✓	✓
58	✓	✓
59	✓	✓
60	✓	✓

เครื่องหมาย ✓ ในตารางที่ ข-3 หมายถึงข้อสอบถามในแถวนั้นทำให้ได้เซตของคำแนะนำเป็นเซตว่างเมื่อนำไปสอบถามระบบให้คำแนะนำที่ใช้การค้นหากฎความสัมพันธ์ด้วยตัวแบบในหลักนั้น

ประวัติผู้เขียนวิทยานิพนธ์

นาย สัญชัย พิทักษ์ชลทรัพย์ เกิดวันที่ 25 ตุลาคม พ.ศ. 2527 สำเร็จการศึกษา วิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ จากภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปี พ.ศ. 2549 จากนั้นได้เข้าศึกษาต่อในระดับปริญญาโท สาขาการพัฒนาซอฟต์แวร์ด้านธุรกิจ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย