รายงานการวิจัยฉบับสมบูรณ์

โครงการ การค้นหารูปแบบที่มีค่าคุณประโยชน์สูงที่ปรากฏอย่างสม่ำเสมอเพื่อการวิเคราะห์พฤติกรรมผู้บริโภค (Mining high utility patterns with regular occurrence for customers' behavior analysis)

คณะผู้วิจัย

นายโกเมศ	อัมพวัน	หัวหน้าโครงการ
นายอรรถสิทธิ์	สรุฤกษ์	ผู้ร่วมวิจัย
นายอนุชิต	จิตพัฒนกุล	ผู้ร่วมวิจัย

โครงการวิจัยประเภทงบประมาณเงินรายได้

จากเงินอุดหนุนรัฐบาล (งบประมาณแผ่นดิน)

ปีงบประมาณ พ.ศ. ๒๕๕๘

มหาวิทยาลัยบูรพา

รหัสโครงการ 2558A10802281

สัญญาเลขที่ 19/2558

รายงานการวิจัยฉบับสมบูรณ์

โครงการ การค้นหารูปแบบที่มีค่าคุณประโยชน์สูงที่ปรากฏอย่างสม่ำเสมอเพื่อการวิเคราะห์พฤติกรรมผู้บริโภค (Mining high utility patterns with regular occurrence for customers' behavior analysis)

คณะผู้วิจัย

นายโกเมศ	อัมพวัน	หัวหน้าโครงการ*
นายอรรถสิทธิ์	สรุฤกษ์	ผู้ร่วมวิจัย**
นายอนุชิต	จิตพัฒนกุล	ผู้ร่วมวิจัย***

*ห้องปฏิบัตการวิจัยนวัตกรรมการประมวลผล คณะวิทยาการสารสารเทศ มหาวิทยาลัยบูรพา

**ห้องปฏิบัติการทางวิศวกรรมระบบนับได้เชิงทฤษฎี คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

*** คณะวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

สิงหาคม 2559

บทคัดย่อ

การค้นหารูปแบบที่มีค่าคุณประโยชน์สูงเป็นหัวข้องานวิจัยหนึ่งภายใต้การทำเมืองข้อมูลที่น่าสนใจ การค้นหารูปแบบดังกล่าวสามารถประยุกต์ใช้ในแอพพลิเคชันต่างๆอย่างแพร่หลาย ตัวอย่างเช่น การประยุกต์ใช้ในธุรกิจค้าปลีกเพื่อทำการค้นหาเซตของสินค้าที่ถูกซื้อจากลูกค้า โดยเซตของสินค้าดังกล่าวจะเป็นรายการสินค้าต่างๆที่ถูกซื้อร่วมกันที่จะให้ผลกำไรสูงหรือต้นทุนที่ต่ำเป็นต้น แต่อย่างไรก็ตาม การค้นหารูปแบบที่มีค่าคุณประโยชน์สูงจะทำการพิจารณาเพียงแค่ค่าคุณประโยชน์ของรายการต่างๆเท่านั้นที่ซึ่งการดำเนินการดังกล่าวอาจไม่เพียงพอต่อการสังเกตุ/วิเคราะห์พฤติกรรมการซื้อสินค้าของผู้บริโภค ด้วยเหตุนี้ งานวิจัยนี้จึงมุ่งเน้นที่จะทำการเพิ่มเติมเงื่อนไขการพิจารณารูปแบบโดยจะทำการเพิ่มเติมเงื่อนไขของการปรากฏอย่างสม่ำเสมอร่วมกับการพิจารณาค่าคุณประโยชน์ของรายการต่างๆ ภายใต้แนวคิดข้างต้น รูปแบบที่น่าสนใจจะเป็นรูปแบบที่มีค่าคุณประโยชน์สูงและปรากฏขึ้นในชุดข้อมูลอย่างสม่ำเสมอ

ในการค้นหารูปแบบใหม่ที่นำเสนอ ผู้วิจัยได้เสนอขั้นตอนวิธีที่มีประสิทธิภาพที่ชื่อว่า HURI-UL ที่ซึ่งจะทำการอ่านข้อมูลจากฐานข้อมูลเพียงครั้งเดียว และทำการประยุกต์ใช้แนวความคิดเกี่ยวกับค่าคุณประโยชน์ที่หลงเหลือและค่าคุณประโยชน์แบบประมาณเพื่อช่วยลดทอนปริภูมิสถานะของการค้นหารูปแบบ นอกจากนั้นยังประยุกต์ใช้โครงสร้างลิสต์คุณประโยชน์เพื่อใช้ในการจัดเก็บค่าคุณประโยชน์และข้อมูลการปรากฏขึ้นของรูปแบบหนึ่งๆ โดยในการทดสอบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอ เราจะสังเกตุได้ว่าขั้นตอนวิธีที่นำเสนอสามารถค้นหารูปแบบที่มีค่าคุณประโยชน์สูงและปรากฏอย่างสม่ำเสมอได้อย่างมีประสิทธิภาพ

Abstract

High utility itemsets mining (HUIM) is an interesting topic in data mining which can be applied in a wide range of applications, for example on retail marketing to find sets of sold products that give high profit, low cost, etc. However, HUIM only considers utility values of items/itemsets which may be insufficient to observe buying behavior of customers. To address this issue, we here introduce an approach on pushing regularity constraint on high utility itemsets mining to observe occurrence behavior of high utility itemsets. Based on this approach, sets of co-occurrence items with (*i*) high utility values and (*ii*) regular occurrence, called *high utility-regular itemsets* (*HURI*s), are regarded as interesting. To mine HURIs, an efficient single-pass algorithm, called HURI-UL, is proposed. HURI-UL applies the concept of remaining and overestimated utilities of itemsets to early prune search space and also utilizes utility list structure to efficiently maintain utility values and occurrence information of itemsets. Experimental results on real datasets show that our proposed approach is efficient to discover high utility itemsets with regular occurrence.

สารบัญ

สารบัญรูปภาพ

สารบัญตาราง

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในยุคปัจจุบันเป็นยุคที่การดำเนินธุรกิจการค้ามีการแข่งขันกันค่อนข้างสูง และมีธุรกิจขนาดกลางและขนาดเล็กก่อตั้งขึ้นเป็นจำนวนมาก ด้วยเหตุนี้จึงเป็นเหตุให้หลายธุรกิจให้ความสนใจกับการประยุกต์ใช้ข้อมูลข่าวสารที่เป็นข้อมูลเชิงกลยุทธ์เพื่อประกอบการตัดสินใจในการดำเนินธุรกิจและการประยุกต์ใช้เทคโนโลยีคอมพิวเตอร์หรือระบบพื้นฐานต่างๆเพื่อช่วยในการดำเนินธุรกิจ จากความต้องการดังกล่าวจึงเป็นเหตุให้มีนักวิจัยพยายามที่จะทำการศึกษารูปแบบพฤติกรรมผู้บริโภคด้วยการค้นหารูปแบบที่ปรากฏบ่อย (Frequent pattern mining) เพื่อบ่งบอกถึงความสัมพันธ์ของสิ่งของหรือเหตุการณ์ที่ปรากฏขึ้นพร้อมกันบ่อยๆ ตัวอย่างเช่น ในธุรกิจห้างสรรพสินค้าหรือธุรกิจค้าปลีกจะทำการหาความสัมพันธ์ของรายการสินค้าที่ถูกซื้อร่วมกันบ่อยๆ เพื่อช่วยในการจัดทำโปรโมชันสินค้า ช่วยในการจัดชั้นวางสินค้าให้สินค้าที่ถูกซื้อร่วมกันบ่อยๆให้อยู่ในพื้นที่ใกล้ๆกันเพื่ออำนวยความสะดวกให้แก่ลูกค้าและยังช่วยกระตุ้นการจับจ่ายใช้สอยของลูกค้า นอกจากนั้นยังช่วยในการจัดทำแคตาล็อกสินค้าให้สินค้าที่ถูกซื้อร่วมกันบ่อยๆให้อยู่ใกล้ๆกัน

แนวความคิดเบื้องต้นของการค้นหารูปแบบที่ปรากฏบ่อยจะประยุกต์ใช้ค่าสนับสนุน (ค่าความถี่หรือจำนวนครั้งในการเกิดขึ้นของรูปแบบนั้นๆ) เป็นตัวชี้วัดความสำคัญหรือความน่าสนใจของรูปแบบ แต่อย่างไรก็ตามการใช้เพียงแค่ค่าสนับสนุนอาจจะไม่เพียงพอต่อการค้นหารูปแบบที่มีความหลากหลาย โดยแนวความคิดนี้ถูกพัฒนาอย่างต่อเนื่องในหลายๆแง่มุม อาทิเช่น การค้นหารูปแบบที่มีการเรียงลำดับที่ปรากฏบ่อย (Frequent sequential pattern mining) การค้นหารูปแบบที่ปรากฏบ่อยภายใต้ค่าน้ำหนักของแต่ละรายการ (Frequent weighted pattern mining) การค้นหารูปแบบที่มีค่าคุณประโยชน์สูง (High utility pattern mining) การค้นหารูปแบบที่ปรากฏบ่อยและปรากฏอย่างสม่ำเสมอ (Frequent-regular pattern mining) และ อื่นๆ

จากงานวิจัยข้างต้น มีหัวข้องานวิจัยหนึ่งที่ทำการพิจารณาค่าคุณประโยชน์ของรูปแบบ (utility of patterns) ที่ซึ่งค่าคุณประโยชน์ของรูปแบบนั้นอาจหมายถึง ผลกำไรที่ได้รับจากการขายสินค้าชิ้นหนึ่งๆของรายการสินค้าหนึ่งๆหรือการบริการหนึ่งๆ เมื่อเราทำการคำนวณค่าคุณประโยชน์ทั้งหมดของรายการสินค้านั้นจะทำให้เราสามารถทราบได้ถึงจำนวนผลกำไร/ขาดทุนที่ได้จากรายการสินค้านั้นๆ และยังทราบถึงรายการสินค้าที่ให้ผลตอบแทนที่สูง แต่อย่างไรก็ดีการค้นหารูปแบบที่มีค่าคุณประโยชน์สูงไม่ได้ทำการพิจารณาถึงพฤติกรรมการปรากฏขึ้นของรูปแบบหรือรายการนั้นๆ ว่ามีพฤติกรรมการปรากฏขึ้นอย่างสม่ำเสมอ ไม่สม่ำเสมอ สม่ำเสมอในบางช่วงเวลาหรือไม่ ซึ่งการตรวจสอบหรือการค้นหาพฤติกรรมการปรากฏขึ้นของรูปแบบอาจช่วยให้เราทราบถึง

ช่วงเวลาที่รายการสินค้าได้รับความนิยม อันนำมาซึ่งการตัดสินใจในการเพิ่ม-ลดกำลังการผลิต การเตรียมวัตถุดิบเพื่อให้สอดคล้องกับกำลังการผลิต การจัดทำโปรโมชันให้กับรายการสินค้าที่พิจารณาเพื่อช่วยกระตุ้นยอดขาย และอื่นๆ เป็นต้น

ดังนั้น ในงานวิจัยนี้ได้นำเสนอกรอบความคิดที่จะทำการค้นหารูปแบบที่มีประโยชน์สูงที่ปรากฏขึ้นอย่างสม่ำเสมอที่จะสามารถทำให้ผู้ที่ทำการวิเคราะห์หรือต้องการค้นหารูปแบบดังกล่าวทราบถึงพฤติกรรมการเกิดขึ้นของรูปแบบที่มีค่าคุณประโยชน์สูงว่ามีพฤติกรรมการปรากฏขึ้นของรูปแบบนั้นๆอย่างสม่ำเสมอหรือไม่ ที่ซึ่งจะทำให้ธุรกิจต่างๆสามารถประยุกต์ใช้รูปแบบดังกล่าวไปวิเคราะห์หาสาเหตุของการเกิดขึ้นของรูปแบบนั้นๆ อันนำไปซึ่งการปรับปรุงและการพัฒนาผลิตภัณฑ์ หรือ กระบวนการดำเนินธุรกิจที่จะทำให้ได้ผลประกอบการที่ดียิ่งขึ้น

1.2 วัตถุประสงค์ของโครงการวิจัย

1. เพื่อศึกษาการวิเคราะห์พฤติกรรมหรือรูปแบบการบริโภคของผู้บริโภคภายใต้กรอบแนวคิดเกี่ยวกับรูปแบบที่มีค่าคุณประโยชน์สูงและการปรากฏขึ้นของรูปแบบนั้นๆอย่างสม่ำเสมอ ที่จะนำไปสู่การค้นหาสาเหตุของการเกิดขึ้นของพฤติกรรมหรือรูปแบบการบริโภคเหล่านั้นที่ซึ่งจะสามารถนำไปเป็นส่วนหนึ่งในการวิเคราะห์สำหรับการพัฒนาผลิตภัณฑ์หรือขั้นตอนการดำเนินธุรกิจสืบไป
2. เพื่อสร้างนวัตกรรมใหม่ในการตรวจสอบพฤติกรรมของการบริโภค
3. เพื่อให้ผู้ที่สนใจสามารถนำแนวคิดที่นำเสนอ ไปศึกษาเพื่อทำการพัฒนาหรือประยุกต์ใช้ในงานวิจัยหรือประยุกต์ใช้ในการดำเนินธุรกิจของตนเองต่อไป

1.3 ขอบเขตของโครงการวิจัย

การวิจัยครั้งนี้มุ่งที่จะศึกษาและพัฒนาการค้นหารูปแบบที่มีคุณประโยชน์สูงที่ปรากฏอย่างสม่ำเสมอ โดยมีขอบเขตดังนี้

1. ผู้ที่ต้องการค้นหารูปแบบที่มีคุณประโยชน์สูงและปรากฏขึ้นอย่างสม่ำเสมอจะต้องกำหนดค่าพารามิเตอร์สองค่าด้วยกันคือ 1) ค่าขีดแบ่งคุณประโยชน์ (Utility threshold) และ 2) ค่าขีดแบ่งความสม่ำเสมอ (Regularity threshold) เพื่อใช้เป็นมาตรวัดความสำคัญของรูปแบบที่จะทำการค้นหา
2. ฐานข้อมูลที่ใช้ในการค้นหารูปแบบจะต้องมีจำนวนรายการที่ปรากฏขึ้นในแต่ละทรานแซกชันแนบอยู่ด้วย และแต่ละรายการจะต้องมีค่าคุณประโยชน์ที่กำหนดไว้ก่อนหน้า (โดยค่าคุณประโยชน์อาจหมายถึงยอดขาย ต้นทุน หรือผลกำไรที่จะได้รับต่อการขายสินค้าชิ้นหนึ่งๆ)
3. การทดสอบประสิทธิภาพของขั้นตอนวิธีในการค้นหารูปแบบที่มีค่าคุณประโยชน์สูงที่ปรากฏอย่างสม่ำเสมอจะวัดในเชิงเวลาและจำนวนหน่วยความจำที่ใช้ในการคำนวณ

1.4 ประโยชน์ที่ได้รับ

1. ได้แนวทางในการวิเคราะห์รูปแบบที่มีประโยชน์สูงและปรากฏขึ้นอย่างสม่ำเสมอ ที่สามารถนำไปวิเคราะห์หาสาเหตุของการเกิดขึ้นของรูปแบบ ที่ซึ่งจะช่วยให้ผู้บริหารกิจการ บริษัท หรือเจ้าของธุรกิจ จะสามารถทำการปรับเปลี่ยนวิธีหรือกลยุทธ์ในการดำเนินธุรกิจเพื่อให้ธุรกิจที่ทำอยู่สามารถดำเนินไปได้ด้วยดี
2. ได้ขั้นตอนวิธีต้นแบบในการค้นหารูปแบบที่มีประโยชน์สูงและปรากฏขึ้นอย่างสม่ำเสมอ
3. สามารถนำขั้นตอนวิธีข้างต้นไปพัฒนาระบบซอฟต์แวร์ เพื่อใช้ในการวิเคราะห์พฤติกรรมของลูกค้าหรือผู้บริโภค ที่จะทำให้กิจการ บริษัท ห้างร้านต่างๆสามารถปรับตัวตามพฤติกรรมของผู้บริโภคได้
4. ขั้นตอนวิธีที่นำเสนอสามารถถูกใช้เป็นต้นแบบในการศึกษาและวิธีขั้นสูงต่อไป

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

งานวิจัยนี้เกิดจากการนำสองแนวความคิดหลักมาผสมผสานกันเพื่อทำการค้นหารูปแบบที่มีประโยชน์สูงและปรากฏขึ้นอย่างสม่ำเสมอ คือ 1) การค้นหารูปแบบที่มีประโยชน์สูง และ 2) การค้นหารูปแบบที่ปรากฏขึ้นสม่ำเสมอ ที่ซึ่งเป็นการพัฒนาต่อยอดมาจากการค้นหารูปแบบที่ปรากฏบ่อยที่ถูกพัฒนาในวงกว้าง โดยทั้งสามแนวคิดหลักสามารถอธิบายได้ดังนี้

2.1.1 การค้นหารูปแบบที่ปรากฏบ่อย (Mining frequent patterns from transactional databases)

การค้นหารูปแบบที่ปรากฏบ่อยเป็นการค้นหารูปแบบโดยมุ่งเน้นที่จะพิจารณาจำนวนครั้ง/ความบ่อย/ความถี่ในการปรากฏขึ้นของรูปแบบเหล่านั้นในฐานข้อมูล โดยปัญหาการค้นหารูปแบบที่ปรากฏบ่อยสามารถนิยามได้ดังนี้

กำหนดให้ เซต $\boldsymbol{I} = \{i_1, i_2, \ldots, i_m\}$ เป็นเซตของรายการ (items) ที่อาจหมายถึงสิ่งของหรือเหตุการณ์ที่ต้องการหาความสัมพันธ์ เซต $X = \{i_p, i_{p+1}, \ldots, i_q\} \subseteq \boldsymbol{I}$ จะเรียกว่าว่าเป็น เซตรายการ (set of items, an itemset หรือ a pattern) และ จะเรียกว่า k-itemset หรือ k-patterns เมื่อ เซต X ประกอบไปด้วยรายการทั้งสิ้น k รายการ กำหนดให้ *TDB* $= \{t_1, t_2, \ldots, t_n\}$ คือ ฐานข้อมูลรายการหรือฐานข้อมูลแบบทรานแซกชัน (transactional database) ที่ซึ่งแต่ละทรานแซกชัน $t_j \in$ *TDB* จะประกอบด้วยหมายเลขกำกับทรานแซกชัน (unique transaction identifier, *tid*) tid = j และ เซตของรายการ Y ที่ถูกบรรจุอยู่ในทรานแซกชันนั้นๆ (ดังแสดงตัวอย่างในตารางที่ 1) ถ้าเซตรายการ $X \subseteq Y$ เราจะสามารถสรุปได้ว่าเซตรายการ X ปรากฏขึ้นในทรานแซกชัน t_j หรือทรานแซกชัน t_j มี X บรรจุอยู่ เราสามารถเขียนสัญลักษณ์แทนได้ว่า t^X_j ดังนั้นเมื่อทำการตรวจสอบรูปแบบ X ว่าปรากฏขึ้นในทรานแซกชันใดบ้างในฐานข้อมูล *TDB* เราจะทราบถึง $T^X = \{t^X_j, t^X_{j+1}, \ldots, t^X_k\}$ ซึ่งก็คือเซตของหมายเลขทรานแซกชัน (tid) ที่ถูกเรียงลำดับที่ซึ่งมี X อยู่ในทรานแซกชันเหล่านั้น (สามารถเขียนย่อๆได้เป็น tidset) ดังนั้นเราจะสามารถทราบถึงจำนวนครั้งในการปรากฏขึ้นของรูปแบบ X ในฐานข้อมูล (ค่าความถี่หรือ

ค่าสนับสนุน) โดยสามารถคำนวณได้เป็น $s^X = |T^X|$ จากนิยามข้างต้น ปัญหาการค้นหารูปแบบที่ปรากฏบ่อยจะเป็นการค้นหารูปแบบที่มีค่าความถี่มากกว่าหรือเท่ากับค่าขีดแบ่งสนับสนุน (support threshold) ที่ผู้ใช้กำหนด

ตารางที่ 1 ตัวอย่างฐานข้อมูลรายการที่ประกอบไปด้วยหมายเลขทรานแซกชันและเซตรายการที่ปรากฏในทรานแซกชัน

หมายเลขทรานแซกชัน (tid)	เซตรายการที่ปรากฏในทรานแซกชัน (a set of items or an itemset)
1	a b c d
2	a c d
3	a b d
4	b c d e
5	a b c e
6	a e
7	a b c
8	b c d e
9	a b d e
10	a e

2.1.2 การค้นหารูปแบบที่มีประโยชน์สูง (Mining high utility patterns from transactional databases)

การค้นหารูปแบบที่มีประโยชน์สูงนั้นถูกพัฒนามาจากการค้นหารูปแบบที่ปรากฏบ่อยที่ซึ่งจะสามารถนิยามได้ดังนี้

กำหนดให้ เซต $I = \{i_1, i_2, ..., i_m\}$ เป็นเซตของรายการ (items) โดยแต่ละรายการ i_p ($1 \leq p \leq m$) ในเซต I จะมีค่าคุณประโยชน์ (utility of item) แนบอยู่ด้วย (เขียนแทนด้วย $p(i_p)$) โดยค่าคุณประโยชน์อาจหมายถึงค่าผลกำไร คุณค่าของสิ่งของเหล่านั้น และ อื่นๆ ดังแสดงในตารางที่ 2 รายการ ‘a’ จะมีค่าคุณประโยชน์เท่ากับ 10 ดังนั้นเมื่อมีรายการ ‘a’ ปรากฏขึ้นหนึ่งครั้งในฐานข้อมูล จะทำให้เราทราบว่า a มีค่าคุณประโยชน์ที่ปรากฏขึ้นในฐานข้อมูลเท่ากับ 10เซต $X = \{i_p, i_{p+1}, ...,i_q\} \subseteq I$ จะเรียกว่าว่าเป็น เซตรายการ (set of items or an itemset) และ จะเรียกว่า k-itemset หรือ k-patterns เมื่อ เซต X ประกอบไปด้วยรายการทั้งสิ้น k รายการ นอกจากนี้ยังกำหนดให้ *TDB* = $\{t_1, t_2, ..., t_n\}$ คือ ฐานข้อมูลรายการหรือฐานข้อมูลแบบทรานแซกชัน (transactional database) ที่ซึ่งแต่ละทรานแซกชัน $t_j \in TDB$ จะประกอบด้วยหมายเลขกำกับทรานแซกชัน

(unique transaction identifier, *tid*) tid = j และ เซตของรายการที่ถูกบรรจุอยู่ในทรานแซกชันนั้นๆ ดังแสดงในตารางที่ 3 ทรานแซกชันที่ 1 กล่าวคือ t_1 = {a(3), b(6)} จะประกอบด้วย 2 รายการคือ ‘a’ และ ‘b’ โดยทรานแซกชันที่ 1 จะมีรายการ ‘a’ ปรากฏขึ้นทั้งหมด 3 ครั้ง และ รายการ ‘b’ ปรากฏขึ้นทั้งหมด 6 ครั้ง ตามลำดับ จากตัวอย่างเราสามารถพูดได้ว่าเซตรายการ ‘ab’ ปรากฏขึ้นหรือถูกบรรจุอยู่ในทรานแซกชันที่ 1 โดยที่เซตรายการ ‘ab’ จะมีค่าคุณประโยชน์ที่ปรากฏขึ้นในทรานแซกชันที่ 1 เท่ากับ (3×10) + (6×5) = 60 เป็นต้น

จากนิยามข้างต้น ปัญหาการค้นหารูปแบบที่มีประโยชน์สูง อาจหมายถึง การค้นหารูปแบบรายการสินค้าที่มีผลกำไรสูง ที่จะทำการพิจารณาค่าคุณประโยชน์ของแต่ละรายการและจำนวนครั้งที่เกิดขึ้นของแต่ละรายการในแต่ละทรานแซกชัน โดยที่เซตรายการ X ใดๆจะเป็นเซตรายการที่มีประโยชน์สูงก็ต่อเมื่อ X มีค่าคุณประโยชน์มากกว่าหรือเท่ากับค่าขีดแบ่งคุณประโยชน์ (utility threshold) ที่ผู้ที่ต้องทำการค้นหารูปแบบเป็นผู้กำหนด โดยค่าคุณประโยชน์ของเซตรายการ X หาได้จากผลรวมของค่าคุณประโยชน์ของแต่ละรายการที่เป็นสมาชิกของ X โดยค่าคุณประโยชน์ของแต่ละการ $x_i \in X$ จะสามารถคำนวณได้จากผลรวมของจำนวนทั้งหมดที่ตาละรายการ x_i ปรากฏขึ้นในฐานข้อมูลคูณกับค่าคุณประโยชน์ของ x_i นั้นๆ ดังนั้น เราสามารถสรุปได้ว่าปัญหาการค้นหารูปแบบที่มีประโยชน์สูงคือ การค้นหารูปแบบหรือเซตรายการที่มีค่าคุณประโยชน์มากกว่าหรือเท่ากับค่าขีดแบ่งคุณประโยชน์

ตารางที่ 2 ตัวอย่างตารางแสดงค่าคุณประโยชน์ของแต่ละรายการ

รายการ	a	b	c	d	e
ค่าคุณประโยชน์	10	5	3	2	7

ตารางที่ 3 ตัวอย่างฐานข้อมูลรายการที่ประกอบไปด้วยหมายเลขทรานแซกชันและเซตรายการที่ปรากฏในทรานแซกชันที่มีการจำนวนของการปรากฏขึ้นของแต่ละรายการ

หมายเลขทรานแซกชัน (tid)	เซตรายการที่ปรากฏในทรานแซกชัน (a set of items or an itemset)
1	a(3) b(6)
2	a(2) c(1) d(3)
3	a(7) b(1) d(5)
4	b(2) c(1) d(3) e(2)
5	a(1) b(1) c(2) e(2)
6	a(2) e(2)
7	a(3) b(2) c(4)

8	b(4) c(1) d(3) e(2)
9	a(3) b(2) d(4) e(1)
10	a(2) e(7)

2.1.3 การค้นหารูปแบบที่ปรากฏบ่อยและปรากฏอย่างสม่ำเสมอ (Mining frequent-regular patterns from transactional databases)

การค้นหารูปแบบที่ปรากฏบ่อยและปรากฏอย่างสม่ำเสมอจะเป็นการค้นหารูปแบบที่พัฒนาต่อยอดจากการค้นหารูปแบบที่ปรากฏบ่อย โดยพัฒนาการวัดความน่าสนใจของรูปแบบที่จากเดิมจะวัดความน่าสนใจโดยพิจารณาจากความถี่ในการปรากฏขึ้นของรูปแบบไปเป็นการวัดความสนใจโดยการตรวจสอบจากพฤติกรรมการปรากฏขึ้นของรูปแบบ โดยจะทำการพิจารณาความถี่และความสม่ำเสมอของการปรากฏขึ้นของรูปแบบ ที่ซึ่งสามารถนิยามได้ดังนี้

กำหนดให้ เซต $I = \{i_1, i_2, ..., i_m\}$ เป็นเซตของรายการ (items) เซต $X = \{i_p, i_{p+1}, ...,i_q\} \subseteq I$ จะเรียกว่าว่าเป็น เซตรายการ (set of items or an itemset) และ จะเรียกว่า k-itemset หรือ k-patterns เมื่อ เซต X ประกอบไปด้วยรายการทั้งสิ้น k รายการ กำหนดให้ *TDB* = $\{t_1, t_2, ..., t_n\}$ คือ ฐานข้อมูลรายการหรือฐานข้อมูลแบบทรานแซกชัน (transactional database) ที่ซึ่งแต่ละทรานแซกชัน $t_j \in TDB$ จะประกอบด้วยหมายเลขกำกับทรานแซกชัน (unique transaction identifier, *tid*) tid = j และ เซตของรายการ Y ที่ถูกบรรจุอยู่ในทรานแซกชันนั้นๆ ถ้า $X \subseteq Y$ เราจะสามารถสรุปได้ว่าเซตรายการ X ปรากฏขึ้นในทรานแซกชัน t_j หรือ t_j มี X อยู่ในทรานแซกชัน ที่ซึ่งเราสามารถเขียนสัญลักษณ์แทนได้ว่า t^X_j ดังนั้นเมื่อทำการตรวจสอบรูปแบบ X ว่าปรากฏขึ้นในทรานแซกชันใดบ้างในฐานข้อมูล *TDB* เราจะทราบถึง $T^X = \{t^X_j, t^X_{j+1}, ..., t^X_k\}$ ซึ่งก็คือ เซตของหมายเลขทรานแซกชัน (tid) ที่ถูกเรียงลำดับที่ซึ่งมี X อยู่ในทรานแซกชัน (สามารถเขียนย่อๆได้เป็น tidset) ดังนั้นเราจะสามารถทราบถึงจำนวนครั้งในการปรากฏขึ้นของรูปแบบ X ในฐานข้อมูล โดยสามารถคำนวณได้เป็น $s^X = |T^X|$

ในการที่จะศึกษาถึงพฤติกรรมการปรากฏขึ้นของรูปแบบว่ามีพฤติกรรมการปรากฏขึ้นอย่างสม่ำเสมอหรือไม่ เราจะต้องทำการพิจารณาเซตของหมายเลขทรานแซกชันที่มี X ปรากฏขึ้น โดยเริ่มจากการพิจารณาแต่ละคู่ของหมายเลขทรานแซกชัน t^X_j และ t^X_{j+1} ที่อยู่ในลำดับติดกัน แล้วทำการหาจำนวนทรานแซกชันที่ไม่มี X ปรากฏระหว่างสองทรานแซกชันนั้นๆ ที่ซึ่งสามารถคำนวณได้เป็น $rtt^X_j = t^X_{j+1} - t^X_j$ แต่สำหรับการปรากฏขึ้นในครั้งแรกและครั้งสุดท้ายของ X จะมีวิธีการคำนวณที่แตกต่างจากการปรากฏขึ้นครั้งอื่นๆที่ซึ่งสามารถคำนวณได้เป็น $fr^X = t^X_1$ (เมื่อ t^X_1 คือ หมายเลขทรานแซกชันที่ X ปรากฏขึ้นครั้งแรก) และ $lr^X = n - t^X_{|T^X|}$ (เมื่อ n คือจำนวนทรานแซกชันทั้งหมดในฐานข้อมูล และ $t^X_{|T^X|}$ คือ หมายเลขทรานแซกชันที่ X ปรากฏขึ้นครั้งสุดท้าย) จากที่กล่าวข้างต้น เราสามารถหาจำนวนทรานแซกชันที่ติดกันสูงที่สุดที่ไม่มี X ปรากฏ โดยสามารถคำนวณได้เป็น $r^X =$

max (fr^X, rtt^X_1, rtt^X_2, ..., lr^X) ที่ซึ่งสามารถบอกได้ถึงช่วงเวลาที่ยาวนานที่สุดที่ไม่มี X ปรากฏขึ้นในฐานข้อมูล และยังสามารถการันตีได้ว่าเซตรายการ X จะปรากฏขึ้นอย่างน้อยหนึ่งครั้งในทุกๆ r^X ทรานแซกชันที่เรียงต่อกัน ซึ่งจากการพิจารณาค่า r^X จะทำให้ทราบถึงพฤติกรรมการปรากฏขึ้นของ X ได้

จากนิยามที่กล่าวมาข้างต้น ปัญหาการค้นหารูปแบบที่ปรากฏบ่อยและปรากฏอย่างสม่ำเสมอจะเป็นการค้นหารูปแบบที่มีค่าความถี่ในการปรากฏมากกว่าหรือเท่ากับค่าขีดแบ่งสนับสนุน (support/ frequency threshold) และมีการช่วงเวลาที่ยาวนานที่สุดที่รูปแบบนั้นๆไม่ปรากฏขึ้นในฐานข้อมูลไม่มากไปกว่าค่าขีดแบ่งความสม่ำเสมอ (regularity threshold)

2.2 งานวิจัยที่เกี่ยวข้อง

S.K. Tanbeer และ คณะ (Tanbeer, 2009) นำเสนองานวิจัยเรื่อง “Discovering periodic-frequent patterns in transactional databases” ที่ชี้ให้เห็นว่าการค้นหารูปแบบปรากฏบ่อยจากฐานข้อมูลโดยใช้ค่าสนับสนุน (จำนวนครั้งของการเกิดขึ้นของรูปแบบเหล่านั้นในฐานข้อมูล) อาจจะไม่เพียงพอต่อการค้นหารูปแบบที่น่าสนใจ จึงได้ทำการเสนอแนวความคิดในการค้นหารูปแบบที่ปรากฏบ่อยและปรากฏอย่างสม่ำเสมอ เพื่อที่จะได้ทราบถึงพฤติกรรมการเกิดขึ้นของรูปแบบเหล่านั้น โดยแนวความคิดนี้สามารถนำไปประยุกต์ใช้ได้ในงานหลายๆด้าน อาทิเช่น ผู้จัดการหรือผู้บริหารของธุรกิจค้าปลีกอาจจะสนใจรายการสินค้าที่ถูกซื้อบ่อยๆและถูกซื้ออย่างสม่ำเสมอมากกว่ารายการสินค้าที่ถูกซื้อบ่อยๆเพียงอย่างเดียว เพื่อที่จะทำการจัดเตรียมสินค้าให้พอเหมาะกับความต้องการของผู้บริโภค และยังสามารถช่วยในการจัดทำโปรโมชันสำหรับสินค้าที่ถูกซื้อบ่อยๆร่วมกับสินค้าที่ถูกซื้อไม่บ่อยได้อีกด้วย ในส่วนของการพัฒนาการออกแบบเว็บไซต์หรือการดูแลรักษาเว็บไซต์ ผู้ดูแลเว็บไซต์อาจจะสนใจความสม่ำเสมอของการคลิกเพื่อเรียกดูข้อมูลในเว็บเพจที่ต่อเนื่องกันเพื่อนำไปปรับปรุงข้อความหรือเนื้อหาของเว็บไซต์ให้มีความน่าสนใจยิ่งขึ้น ในส่วนของการวิเคราะห์ข้อมูลทางพันธุกรรม กลุ่มของยีนส์ที่ปรากฏบ่อยและสม่ำเสมออาจบ่งบอกถึงข้อมูลที่สำคัญให้แก่นักวิทยาศาสตร์ได้ ในส่วนตลาดหุ้น กลุ่มของหุ้นที่มีดัชนีที่มีการเพิ่มขึ้นอย่างสม่ำเสมออาจจะได้รับความน่าสนใจจากนักลงทุนต่างๆ และ อื่นๆ

ในการหารูปแบบที่ปรากฏบ่อยและปรากฏอย่างสม่ำเสมอ ผู้ใช้จะต้องทำการกำหนดค่าพารามิเตอร์ 2 ค่าด้วยกันคือ 1) ค่าขีดแบ่งสนันสนุน และ 2) ค่าขีดแบ่งความสม่ำเสมอ เพื่อใช้วัดความน่าสนใจหรือความสำคัญของรูปแบบภายใต้พฤติกรรมการเกิดขึ้นของรูปแบบเหล่านั้น แต่อย่างไรก็ดี เป็นที่ทราบกันดีว่า “ถ้าเราไม่ได้มีความรู้ในข้อมูลมาก่อน การกำหนดค่าขีดแบ่งสนับสนุนเพื่อที่จะได้รับรูปแบบที่น่าสนใจและมีความสำคัญมากที่สุดจะเป็นเรื่องที่ยุ่งยากและลำบาก” โดยที่ถ้าเรากำหนดค่าขีดแบ่งสนับสนุนสูงเกินไป อาจทำให้เราได้ผลลัพธ์เป็นจำนวนน้อยหรืออาจจะไม่ได้ผลลัพธ์เลย ในกรณีนี้ เราจำเป็นต้องคาดเดาค่าขีดแบ่งให้มีค่าน้อยลงแล้วทำการค้นหาผลลัพธ์ใหม่อีกครั้งที่ซึ่งอาจจะได้รับหรือไม่ได้รับผลลัพธ์ที่ดีขึ้นก็เป็นได้ แต่ในกรณีที่ค่าขีดแบ่งถูกกำหนดให้มีค่า

น้อย อาจทำให้เราได้ผลลัพธ์ออกมาเป็นจำนวนมากเกินกว่าที่เราจะทำการพิจารณาองค์ความรู้ได้ และการค้นหาผลลัพธ์จะใช้เวลาค่อนข้างมากอีกด้วย

จากปัญหาข้างต้นดังกล่าว จึงมีงานวิจัยที่ทำการพัฒนาต่อยอดจากงานของ Tanbeer โดยมีวัตถุประสงค์ที่จะหลีกเลี่ยงการกำหนดค่าขีดแบ่งสนับสนุน โดยการกำหนดให้ผู้ใช้ทำการกำหนดจำนวนผลลัพธ์ (รูปแบบ) ที่ต้องการแทนด้วยการนำแนวคิดของการค้นหารูปแบบที่ปรากฏขึ้นบ่อยสุดเคอันดับแรกมาประยุกต์ใช้ (Fu, 2000)(Wang, 2005)(Yang, 2008)(Li, 2009)(Ke, 2009)(Fournier-Viger, 2013) เป็นต้น โดยในงานวิจัยนั้นได้เสนอปัญหาการค้นหารูปแบบที่ปรากฏขึ้นบ่อยและสม่ำเสมอเคอันดับแรก (Mining top-k frequent-regular pattern) (Amphawan, 2009) เพื่อทำการหารูปแบบทั้งสิ้นเครูปแบบ ที่ซึ่งปรากฏในฐานข้อมูลอย่างสม่ำเสมอและปรากฏบ่อยที่สุด ภายใต้ปัญหานี้ ผู้ที่ต้องการค้นหารูปแบบจะต้องทำการกำหนดค่าพารามิเตอร์ 2 ค่าด้วยกัน คือ 1) ค่าขีดแบ่งความสม่ำเสมอ (σ_r) และ 2) จำนวนผลลัพธ์ที่ต้องการ (k) โดยในการค้นหารูปแบบดังกล่าวได้อย่างรวดเร็ว ผู้วิจัยได้เสนอ 3 อัลกอริทึมที่มีประสิทธิภาพ ได้แก่ MTKPP (Amphawan, 2009), TR-CT(Amphawan, 2011) และ TKRIMPE(Amphawan, 2012) ตามลำดับ นอกเหนือจากหลีกเลี่ยงความยุ่งยากในการกำหนดค่าขีดแบ่งสนับสนุน การค้นหารูปแบบที่ปรากฏบ่อยและปรากฏอย่างสม่ำเสมอได้ถูกพัฒนาอย่างต่อเนื่องในหลายๆแง่มุม อาทิเช่น การค้นหารูปแบบที่ปรากฏบ่อยและสม่ำเสมอจากฐานข้อมูลที่มีการเพิ่มเติมข้อมูล (Mining frequent-regular patterns on incremental transactional databases)(Tanbeer, 2010) และจากฐานข้อมูลที่เป็นแบบสายข้อมูล (Mining frequent-regular patterns on data stream)(Tanbeer, 2010) การหารูปแบบที่เกิดขึ้นอย่างสม่ำเสมอที่ประกอบด้วยรูปแบบที่ปรากฏบ่อยและปรากฏไม่บ่อย (Mining frequent-regular patterns consisting of both frequent and rare items)(Surana, 2012), การค้นหารูปแบบที่ปรากฏบ่อยและสม่ำเสมอด้วยการกำหนดเงื่อนไขเกี่ยวกับค่าสนับสนุน (Mining periodic-frequent patterns with maximum items' support constraints)(Kiran, 2010), และ อื่นๆ

ในส่วนของการค้นหารูปแบบที่มีค่าคุณประโยชน์สูงจะเริ่มจาก H. Yao และชาวคณะ (Yao, 2004) ได้นำเสนอปัญหาและนิยามการค้นหารูปแบบที่มีค่าคุณประโยชน์สูงที่ซึ่งจะทำการวัดความน่าสนใจหรือความสำคัญของรูปแบบจากความมีประโยชน์ของรูปแบบนั้นๆ โดยความมีประโยชน์อาจเกี่ยวข้องกับ ผลกำไร ยอดขาย หรือต้นทุน เป็นต้น โดยรูปแบบที่มีค่าคุณประโยชน์สูงจะมีบทบาทสำคัญในการตัดสินใจต่างๆในการดำเนินธุรกิจ อาทิเช่น การเพิ่มรายได้ การลดค่าใช้จ่ายด้านการตลาดและการจัดการคลังสินค้า และอื่นๆ นอกเหนือจากการประยุกต์ใช้ในแวดวงธุรกิจ การค้นหารูปแบบในลักษณะนี้ยังสามารถนำไปประยุกต์ใช้กับงานในหลายๆด้าน อาทิเช่น การวิเคราะห์ทางพันธุกรรม(Biological gene analysis) การเข้าถึงเว็บแบบมีลำดับ(Web-click sequence analysis) การตรวจสอบการขึ้นลงของดัชนีหุ้น การวัดประสิทธิภาพของการจราจร การตรวจสอบการเข้าใช้งานเซิร์ฟเวอร์ การวิเคราะห์ข้อมูลที่ได้จากเซ็นเซอร์เน็ตเวิร์ค และ การวิเคราะห์ข้อมูลการใช้โทรศัพท์ทางไกล เป็นต้น โดยในตอนเริ่มต้น (Yao, 2004)(Yao, 2006)(Liu, 2005) ได้พยายามที่จะคิดต้นขั้นตอนวิธีที่จะทำการค้นหา

รูปแบบที่มีค่าคุณประโยชน์สูงได้อย่างมีประสิทธิภาพ โดยพยายามที่จะคิดค้นขั้นตอนวิธีที่จะลดทอนปริมาณข้อมูลหรือจำนวนรูปแบบที่ต้องทำการพิจารณา (itemset lattice หรือ search space of itemsets) แต่อย่างไรก็ตามขั้นตอนที่คิดค้นขึ้นยังคงทำการอ่านข้อมูลจากฐานข้อมูลอยู่หลายครั้ง ที่ซึ่งทำให้ใช้เวลาในการคำนวณค่อนข้างมาก ต่อมา (Tseng, 2010) ได้ทำการคิดค้นโครงสร้างต้นไม้ที่มีชื่อว่า UP-tree (Utility Pattern Tree) เพื่อใช้ในการจัดเก็บข้อมูลรูปแบบมีค่าคุณประโยชน์สูงระหว่างการคำนวณ โดยที่การใช้โครงสร้างต้นไม้ดังกล่าวจะสามารถลดจำนวนครั้งในการอ่านฐานข้อมูลเหลือเพียง 3 ครั้งเท่านั้น แต่อย่างไรก็ดีในปี 2012 (Liu, 2012) ได้เสนอโครงสร้างข้อมูลที่มีชื่อว่า Utility-list ที่ซึ่งเป็นลิงค์ลิสที่ใช้ในการจัดเก็บรูปแบบมีค่าคุณประโยชน์สูงในระหว่างการคำนวณ ซึ่งจากการประยุกต์ใช้ utility-list จะทำให้สามารถลดการอ่านฐานข้อมูลเหลือเพียง 2 ครั้งเท่านั้น

จากที่ได้กล่าวมาทั้งหมดการค้นหารูปแบบที่มีคุณประโยชน์สูงยังคงได้รับความสนใจจากนักวิจัยเป็นจำนวนมากที่ซึ่งพยายามพัฒนาขั้นตอนวิธีสำหรับข้อมูลที่มีการเพิ่มข้อมูลทรานแซกชัน (Incremental transactional database) และ ฐานข้อมูลแบบสตรีม (Data stream) เป็นต้น

บทที่ 3
วิธีดำเนินการวิจัย

จากที่กล่าวข้างต้น การค้นหารูปแบบปรากฏบ่อยและปรากฏสม่ำเสมอจะไม่สามารถบ่งบอกถึงความสำคัญหรือคุณประโยชน์ของรูปแบบได้ แต่สำหรับการค้นหารูปแบบที่มีค่าคุณประโยชน์สูงจะไม่สามารถบ่งบอกถึงพฤติกรรมการปรากฏขึ้นของรูปแบบได้ ด้วยเหตุนี้ ในบทนี้จะนำเสนอการค้นหารูปแบบที่มีค่าคุณประโยชน์สูงและปรากฏสม่ำเสมอจากฐานข้อมูลรายการ (mining high-utility regular itemsets from transactional database) ที่ซึ่งจะทำการพิจารณาค่าคุณประโยชน์พร้อมกับพฤติกรรมการปรากฏของรูปแบบภายใต้ความสม่ำเสมอ โดยรูปแบบในลักษณะนี้จะสามารถบ่งบอกถึงพฤติกรรมผู้บริโภคที่เกี่ยวเนื่องกับการซื้อสินค้าที่ให้ผลกำไรสูงโดยสินค้าเหล่านั้นถูกซื้ออย่างสม่ำเสมอ จากองค์ความรู้ดังกล่าว จะทำให้ทราบถึงความต้องการสินค้าของผู้บริโภค และมีส่วนช่วยเป็นข้อมูลประกอบการตัดสินใจเกี่ยวกับการบริหารจัดการคลังสินค้า การจัดทำโปรโมชันเพื่อส่งเสริมการขาย และอื่นๆ

3.1 นิยามที่เกี่ยวข้องกับการค้นหารูปแบบที่มีค่าคุณประโยชน์สูงและปรากฏสม่ำเสมอ

กำหนดให้

- เซต $I = \{i_1, i_2, ..., i_m\}$ เป็นเซตของรายการ (items)
- แต่ละรายการ $i_j \in I$ จะมีค่าคุณประโยชน์ต่อ i_j หนึ่งๆ (เรียกว่า external utility) อาทิเช่น ผลกำไรจากการสินค้าชิ้นหนึ่งๆ ต้นทุนของสินค้าชิ้นหนึ่งๆ หรือ อื่นๆ ซึ่งจะแทนด้วยสัญลักษณ์ $eu(i_j)$
- เซต $X = \{i_p, i_{p+1}, ...,i_q\} \subseteq I$ จะเรียกว่าว่าเป็น เซตรายการ (set of items or an itemset) และจะเรียกว่า k-itemset หรือ k-patterns เมื่อ เซต X ประกอบไปด้วยรายการทั้งสิ้น k รายการ
- *TDB* = $\{t_1, t_2, ..., t_n\}$ คือ ฐานข้อมูลรายการหรือฐานข้อมูลแบบทรานแซกชัน (transactional database) ที่ซึ่งแต่ละทรานแซกชัน $t_p \in TDB$ จะประกอบด้วยหมายเลขกำกับทรานแซกชัน (unique transaction identifier, *tid*) tid = p และ เซตของรายการ Y ที่ถูกบรรจุอยู่ในทรานแซกชันนั้นๆ โดยที่แต่ละ $i_j \in Y$ จะมีจำนวนชิ้นของ i_j ที่ปรากฏในทรานแซกชันนั้นๆ (เรียกว่า internal utility) สามารถแทนได้ด้วย $iu(i_j, t_p)$

- ถ้า $X \subseteq Y$ ของ t_p เราจะสามารถสรุปได้ว่าเซตรายการ X ปรากฏขึ้นในทรานแซกชัน t_p หรือ t_p มี X อยู่ในทรานแซกชัน ที่ซึ่งเราสามารถเขียนสัญลักษณ์แทนได้ว่า t^X_p ดังนั้นเมื่อทำการตรวจสอบรูปแบบ X ว่าปรากฏขึ้นในทรานแซกชันใดบ้างในฐานข้อมูล *TDB* เราจะทราบถึง $T^X = \{t^X_p, t^X_{p+1}, ..., t^X_q\}$ ซึ่งก็คือ เซตของหมายเลขทรานแซกชัน (tid) ที่ถูกเรียงลำดับที่ซึ่งมี X อยู่ในทรานแซกชัน (สามารถเขียนย่อๆได้เป็น tidset)
- จำนวนครั้งในการปรากฏขึ้นของรูปแบบ X ในฐานข้อมูล โดยสามารถคำนวณได้เป็น $s^X = |T^X|$
- ค่าคุณประโยชน์ของรายการ i_j ที่ปรากฏในทรานแซกชัน t_p จะเป็นผลคูณระหว่าง จำนวนชิ้นของ i_j ที่ปรากฏในทรานแซกชัน t_p กับค่าคุณประโยชน์ต่อ i_j สามารถคำนวณและแทนด้วยสัญลักษณ์ $u(i_j, t_p) = iu(i_j, t_p) \times eu(i_j)$
- ค่าคุณประโยชน์ของเซตรายการ X ที่ปรากฏในทรานแซกชัน t_p จะเป็นผลรวมของค่าคุณประโยชน์ของทุกรายการที่เป็นสมาชิกของเซตรายการ X ที่ปรากฏในทรานแซกชัน t_p สามารถคำนวณและแทนด้วยสัญลักษณ์ $\mathrm{u}(\mathrm{X}, \mathrm{t}_p) = \sum_{i_j \in X} iu(i_j, t_p) \times \mathrm{eu}(i_j)$
- ค่าคุณประโยชน์ของเซตรายการ X ที่ปรากฏในฐานข้อมูลรายการ *TDB* จะเป็นผลรวมของค่าคุณประโยชน์ของเซตรายการ X ที่ปรากฏในทุกทรานแซกชันของฐานข้อมูลรายการ *TDB* สามารถคำนวณและแทนด้วยสัญลักษณ์ $\mathrm{u}(\mathrm{X}) = \sum_{i_j \in X, X \subseteq t_p} \mathrm{u}(\mathrm{X}, \mathrm{t}_p)$

จากนิยามและสัญลักษณ์ทั้งหมดข้างต้น ปัญหาการค้นหารูปแบบที่มีค่าคุณประโยชน์สูงจะเป็นการค้นหารูปแบบ X ใดๆ ที่มีค่าคุณประโยชน์ $\mathrm{u}(\mathrm{X})$ มากกว่าหรือเท่ากับค่าขีดแบ่งคุณประโยชน์ σ_u ที่ผู้ใช้กำหนด แต่อย่างไรก็ตาม การค้นหารูปแบบตามที่มีคุณสมบัติข้างต้นจะมีอุปสรรคตรงที่ไม่สามารถประยุกต์ใช้ downward closure property ในการลดทอนปริภูมิสถานะ เนื่องจากรูปแบบที่เป็นซุปเปอร์เซตของรูปแบบที่มีค่าคุณประโยชน์ต่ำอาจมีค่าคุณประโยชน์สูง นี้จึงเป็นเหตุเราไม่สามารถตัดรูปแบบที่มีค่าคุณประโยชน์ต่ำออกจากการพิจารณาได้ ด้วยเหตุนี้ Liu (Liu, 2005) จึงได้เสนอแนวคิดเกี่ยวกับ “transaction-weighted utility (twu)” ที่ซึ่งจะเป็นค่าประมาณของค่าคุณประโยชน์ที่ทำให้สามารถประยุกต์ใช้สามารถประยุกต์ใช้ downward closure property ในการลดทอนปริภูมิสถานะได้ โดย twu จะสามารถนิยามได้ดังนี้

- ค่าคุณประโยชน์ของทรานแซกชัน t_p จะเป็นผลรวมของค่าคุณประโยชน์ของทุกๆรายการที่ปรากฏในทรานแซกชัน t_p สามารถคำนวณและแทนด้วยสัญลักษณ์ $\mathrm{tu}(\mathrm{t}_p) = \sum_{i_j \in t_p} \mathrm{u}(i_j, \mathrm{t}_p)$
- ค่า twu ของเซตรายการ X ในฐานข้อมูลรายการ *TDB* จะเป็นค่าประมาณคุณประโยชน์ของ X ที่เกิดจากผลรวมของค่าคุณประโยชน์ของทุกทรานแซกชันในฐานข้อมูลรายการ *TDB* ที่มี X ปรากฏ สามารถคำนวณและแทนด้วยสัญลักษณ์ $\mathrm{twu}(\mathrm{X}) = \sum_{X \in D, X \subseteq t_p} \mathrm{tu}(\mathrm{t}_p)$

จากนิยามเกี่ยวกับ twu ของรูปแบบหนึ่งๆ จะทำให้เราสามารถประยุกต์ใช้ downward closure property ในการลดทอนปริภูมิสถานะได้ ก็ต่อเมื่อ เซตรายการ X ใดๆ มีค่า twu(X) น้อยกว่าค่าขีดแบ่งคุณประโยชน์แล้ว จะทำให้ทุกๆเซตรายการที่เป็นซุปเปอร์เซตของ X จะมีค่าคุณประโยชน์น้อยกว่าค่าขีดแบ่งคุณประโยชน์ด้วยเช่นกัน นี้จึงเป็นเหตุให้ เราสามารถตัดการพิจารณาเซตรายการ X และเซตรายการที่เป็นซุปเปอร์เซตของ X ออกจากการพิจารณาได้ เนื่องจาก เซตรายการ X และ ทุกๆเซตรายการที่เป็นซุปเปอร์เซตของ X จะมีค่าคุณประโยชน์น้อย

แม้ว่า twu ของรูปแบบจะมีส่วนช่วยในการลดทอนปริภูมิสถานะได้ แต่อย่างไรก็ตาม twu ของรูปแบบหนึ่งๆจะเป็นค่าประมาณคุณประโยชน์ที่มีค่าสูงกว่าค่าคุณประโยชน์จริงค่อนข้างมาก ด้วยเหตุนี้จึงเป็นเหตุให้ Liu (Liu, 2012) คิดค้นแนวความคิดเกี่ยวกับค่าประมาณคุณประโยชน์ที่มีความกระชับ (มีค่าเกินจริงน้อยกว่าค่า twu) ที่ซึ่งสามารถนิยามได้ดังนี้

- กำหนดให้ $>$ แสดงถึงลำดับของรายการในเซตรายการ I
- ค่าคุณประโยชน์ส่วนที่เหลือ (remaining utility) ของเซตรายการ X ในทรานแซกชัน t_p หมายถึงผลรวมของค่าคุณประโยชน์ของทุกรายการที่ปรากฏในทรานแซกชัน t_p และรายการเหล่านั้นมีลำดับหลังจาก X สามารถคำนวณและแทนด้วยสัญลักษณ์ $\mathrm{ru}(\mathrm{X}, \mathrm{t}_p) = \sum_{i_j \in t_p, X<i_j} \mathrm{u}(i_j, \mathrm{t}_p)$
- ค่าคุณประโยชน์ส่วนที่เหลือของเซตรายการ X ในฐานข้อมูลรายการ *TDB* จะเป็นค่าผลรวมของค่าคุณประโยชน์ส่วนที่เหลือของเซตรายการ X ในทุกทรานแซกชัน ที่มี X ปรากฏ สามารถคำนวณและแทนด้วยสัญลักษณ์ $\mathrm{ru}(\mathrm{X}) = \sum_{X \in D, X \subseteq t_p} \mathrm{ru}(\mathrm{X}, \mathrm{t}_p)$
- ค่าประมาณคุณประโยชน์แบบกระชับของเซตรายการ X ในฐานข้อมูล *TDB* จะเป็นค่าผลรวมระหว่างค่าคุณประโยชน์จริงของเซตรายการ X กับค่าคุณประโยชน์ส่วนที่เหลือของเซตรายการ X ในฐานข้อมูลรายการ *TDB* สามารถคำนวณและแทนด้วยสัญลักษณ์ $\mathrm{ou}(\mathrm{X}) = \mathrm{u}(\mathrm{X}) + \mathrm{ru}(\mathrm{X})$

จากแนวคิดและนิยามข้างต้นเราสามารถบอกได้ว่า ถ้าเซตรายการ X มีค่าประมาณคุณประโยชน์แบบกระชับน้อยกว่าค่าขีดแบ่งคุณประโยชน์แล้ว เซตรายการใดที่เกิดจากการรวมกันระหว่างเซตรายการ X และรายการ i_j ใดๆที่มีลำดับหลังจาก X จะมีค่าคุณประโยชน์น้อยกว่าค่าขีดแบ่งคุณประโยชน์เสมอ ซึ่งจากข้อสรุปดังกล่าวจะทำให้เราสามารถลดทอนการพิจารณาเซตรายการ X และซุปเปอร์เซตของ X ที่เกิดจากการรวมกันระหว่างเซตรายการ X และรายการ i_j ใดๆที่มีลำดับหลังจาก X ได้ นี้อันนำมาซึ่งการลดทอนปริภูมิสถานะ

ในการที่จะศึกษาถึงพฤติกรรมการปรากฏขึ้นของรูปแบบว่ามีพฤติกรรมการปรากฏขึ้นอย่างสม่ำเสมอหรือไม่ เราจะต้องทำการพิจารณาเซตของหมายเลขทรานแซกชันที่มี X ปรากฏขึ้น โดยเริ่มจากการพิจารณาแต่ละคู่ของหมายเลขทรานแซกชัน t^X_j และ t^X_{j+1} ที่อยู่ในลำดับติดกันในเซต T^X แล้วทำการหาจำนวนทรานแซกชันที่ไม่มี

X ปรากฏระหว่างสองทรานแซกชันนั้นๆ ที่ซึ่งสามารถคำนวณได้เป็น $rtt^X_j = t^X_{j+1} - t^X_j$ แต่สำหรับการปรากฏขึ้นในครั้งแรกและครั้งสุดท้ายของ X จะมีวิธีการคำนวณที่แตกต่างจากการปรากฏขึ้นครั้งอื่นๆที่ซึ่งสามารถคำนวณได้เป็น $fr^X = t^X_1$ (เมื่อ t^X_1 คือ หมายเลขทรานแซกชันที่ X ปรากฏขึ้นครั้งแรก) และ $lr^X = n - t^X_{|T^X|}$ (เมื่อ n คือจำนวนทรานแซกชันทั้งหมดในฐานข้อมูล และ $t^X_{|T^X|}$ คือ หมายเลขทรานแซกชันที่ X ปรากฏขึ้นครั้งสุดท้าย) จากที่กล่าวข้างต้น เราสามารถหาจำนวนทรานแซกชันที่ติดกันสูงที่สุดที่ไม่มี X ปรากฏ โดยสามารถคำนวณได้เป็น $r^X = \max(fr^X, rtt^X_1, rtt^X_2, ..., lr^X)$ ที่ซึ่งสามารถบอกได้ถึงช่วงเวลาที่ยาวนานที่สุดที่ไม่มี X ปรากฏขึ้นในฐานข้อมูล และยังสามารถการันตีได้ว่าเซตรายการ X จะปรากฏขึ้นอย่างน้อยหนึ่งครั้งในทุกๆ r^X ทรานแซกชันที่เรียงต่อกัน ซึ่งจากการพิจารณาค่า r^X จะทำให้ทราบถึงพฤติกรรมการปรากฏขึ้นของ X ได้ ดังนั้น เซตรายการ X ใดๆจะเป็นเซตรายการที่ปรากฏสม่ำเสมอก็ต่อเมื่อ ค่าความสม่ำเสมอ r^X มีค่าน้อยกว่าหรือเท่ากับค่าขีดแบ่งความสม่ำเสมอ σ_r ที่ผู้ใช้กำหนด

จากนิยามทั้งหมดข้างต้น รูปแบบที่มีค่าคุณประโยชน์สูงและปรากฏสม่ำเสมอสามารถนิยามได้ดังนี้

> **นิยาม** เซตรายการ X หนึ่งๆจะเป็นเซตรายการที่มีค่าคุณประโยชน์สูงและปรากฏสม่ำเสมอก็ต่อเมื่อ X มีค่าคุณประโยชน์ $u(X)$ ไม่น้อยกว่า σ_u และ X มีค่าความสม่ำเสมอ r^X ไม่เกินกว่า σ_r ที่ผู้ใช้กำหนด

3.2 ขั้นตอนวิธีที่นำเสนอ

จากปัญหาการค้นหารูปแบบที่มีค่าคุณประโยชน์สูงและปรากฏสม่ำเสมอจากฐานข้อมูลรายการภายใต้ค่าขีดแบ่ง σ_u และ σ_r ที่ผู้ใช้กำหนด ผู้วิจัยได้นำเสนอขั้นตอนวิธีที่มีชื่อว่า HURI-UL ที่ซึ่งสามารถลดทอนการอ่านข้อมูลจากฐานข้อมูล (วิธีอื่นๆต้องอ่านข้อมูลจากฐานข้อมูล 2 ครั้ง) ให้ทำการอ่านข้อมูลเพียงครั้งเดียว โดยทำการจัดเก็บค่าคุณประโยชน์ของทุกทรานแซกชันในฐานข้อมูลไว้ในอะเรย์ที่เรียกว่า tu-List นอกจากนั้น HURI-UL ได้ประยุกต์ใช้แนวคิดค่าประมาณคุณประโยชน์แบบกระชับของเซตรายการในการลดทอนและประยุกต์ใช้โครงสร้างข้อมูล utility list เพื่อทำการจัดเก็บข้อมูลการปรากฏพร้อมกับค่าคุณประโยชน์ของรายการหนึ่งๆที่ปรากฏในทรานแซกชันหนึ่งๆ จากการประยุกต์ใช้โครงสร้างข้อมูลดังกล่าว จะทำให้ HURI-UL สามารถค้นหารูปแบบที่มีค่าคุณประโยชน์สูงและปรากฏสม่ำเสมอจากฐานข้อมูลรายการได้อย่างมีประสิทธิภาพ

3.2.1 โครงสร้างข้อมูล Utility list

จากการกำหนดลำดับของรายการที่เป็นสมาชิกของเซตรายการ $I = \{i_1, i_2, ..., i_n\}$ เราสามารถสร้าง utility list ของเซตรายการ X จากฐานข้อมูล TDB ได้เป็นลิสต์ของ 3-tuple คือ $<p, u(X, t_p), ru(X, t_p)>$ เมื่อ 1) p คือหมายเลขทรานแซกชันที่มีเซตรายการ X ปรากฏ 2) $u(X, t_p)$ คือ ค่าคุณประโยชน์ของเซตรายการ X ในทรานแซกชัน t_p และ 3) $ru(X, t_p)$ คือ ค่าคุณประโยชน์ส่วนที่เหลือ (remaining utility) ของเซตรายการ X ในทรานแซกชัน

t_p ตัวอย่างเช่น พิจารณารายการ ‘a’ ที่มีค่าคุณประโยชน์ในตารางที่ 4 เท่ากับ 3 และปรากฏในฐานข้อมูลในตารางที่ 5 เป็น T^a = {t_1, t_3, t_5, t_7, t_8} ตามลำดับ ถ้าลำดับของรายการทั้งหมดเป็น a < b < c < d < e < f < g < f เราจะสามารถสร้าง utility list ของรายการ ‘a’ ได้เป็น {<1, 9, 73>, <3, 6, 64>, <5, 6, 19>, <7, 15, 42>, <8, 9, 41>} โดยที่สมาชิกอันดับแรกของ utility list จะบ่งบอกได้ว่า รายการ ‘a’ ปรากฏในทรานแซกชันที่ 1 รายการ ‘a’ มีค่าคุณประโยชน์ในทรานแซกชันที่ 1 เท่ากับ 9 และมีค่าคุณประโยชน์ส่วนเหลือในทรานแซกชันที่ 1 เท่ากับ 73 ตามลำดับ แต่สำหรับสมาชิกลำดับอื่นๆของ utility list ก็จะบ่งบอกถึงการปรากฏขึ้นของรายการ ‘a’ หลังจากทรานแซกชันที่ 1 ตามลำดับ

ตารางที่ 4 ตัวอย่างตารางแสดงค่าคุณประโยชน์ของแต่ละรายการ

รายการ	a	b	c	d	e	f	g	f
ค่าคุณประโยชน์	3	2	1	30	5	3	4	15

ตารางที่ 5 ตัวอย่างฐานข้อมูลรายการที่ประกอบไปด้วยหมายเลขทรานแซกชันและเซตรายการที่ปรากฏในทรานแซกชันที่มีการจำนวนของการปรากฏขึ้นของแต่ละรายการ

หมายเลขทรานแซกชัน (tid)	เซตรายการที่ปรากฏในทรานแซกชัน (a set of items or an itemset)
1	a(3), c(8), d(2), e(1)
2	b(5), f(3), g(5), h(20)
3	a(2), c(4), d(1)
4	c(5), e(1), f(1)
5	a(2), b(3), c(1), f(4)
6	d(1), g(5), h(1)
7	a(5), b(1) c(4)
8	b(4) c(1) d(3) e(2)
9	a(3) b(2) d(4) e(1)
10	a(2) e(7)

3.2.2 ขั้นตอนวิธี HURI-UL

ดังที่กล่าวข้างต้น ขั้นตอนวิธี HURI-UL ไม่เพียงแต่ประยุกต์แนวความคิดค่าคุณประโยชน์ส่วนเหลือ และค่าประมาณคุณประโยชน์แบบกระชับเพื่อทำการลดทอนปริภูมิสถานะ แต่ยังทำการประยุกต์ใช้โครงสร้าง utility

list เพื่อใช้ในการจัดเก็บข้อมูลเกี่ยวกับการปรากฏขึ้นของรายการ/เซตรายการหนึ่งๆ พร้อมกับค่าคุณประโยชน์ของรายการ/เซตรายการนั้นๆที่ปรากฏในทรานแซกชันหนึ่งๆ ระหว่างการค้นหาผลลัพธ์ นอกจากนั้น HURI-UL ยังทำการลดทอนการอ่านฐานข้อมูลให้เหลือเพียงครั้งเดียวด้วยการใช้อะเรย์หนึ่งมิติสำหรับจัดเก็บค่าคุณประโยชน์ของทุกทรานแซกชัน (เรียกว่า tu-List) โดยในการหาผลลัพธ์จากขั้นตอนวิธี HURI-UL จะประกอบไปด้วย 2 ขั้นตอนการคำนวณหลักคือ

1) การระบุถึงรายการที่ปรากฏสม่ำเสมอและคาดว่าน่าจะมีค่าคุณประโยชน์สูง (เมื่อรวมกับรายการอื่นๆ) ที่ซึ่งจะทำการสร้างลิสต์หลายมิติเพื่อใช้เก็บรายการและเซตของรายการที่มีคุณสมบัติข้างต้น (เรียกว่า i-List) โดยการระบุถึงรายการที่ปรากฏสม่ำเสมอและคาดว่าน่าจะมีค่าคุณประโยชน์สูง (เมื่อรวมกับรายการอื่นๆ)จากการอ่านฐานข้อมูลหนึ่งครั้งจะถูกจัดเก็บอยู่ในลิสต์มิติแรก (เรียกว่า 1-List)
2) การค้นหาผลลัพธ์จากลิสต์ที่สร้างขึ้นในขั้นตอนแรก

ดังแสดงในรูปที่ 1 การระบุถึงรายการที่ปรากฏสม่ำเสมอและคาดว่าน่าจะมีค่าคุณประโยชน์สูง (เมื่อรวมกับรายการอื่นๆ) จะเริ่มจากการสร้างอะเรย์ tu-List เพื่อใช้ในการจัดเก็บค่าคุณประโยชน์ของทุกทรานแซกชันในฐานข้อมูล ทำการสร้างลิสต์หลายมิติ i-List และทำการสร้างลิสต์ที่ใช้สำหรับจัดเก็บรายการที่ปรากฏสม่ำเสมอและคาดว่าน่าจะมีค่าคุณประโยชน์สูง (เมื่อรวมกับรายการอื่นๆ) 1-List (บรรทัดที่ 1) จากนั้นแต่ละทรานแซกชัน t_p ในฐานข้อมูลจะถูกอ่านเพื่อทำการคำนวณค่าคุณประโยชน์ $tu(t_p)$ และ จัดเก็บ $tu(t_p)$ ใน tu-List ต่อมา HURI-UL จะทำการพิจารณาแต่ละรายการ i_j ที่ปรากฏละทรานแซกชัน t_p และทำการอัพเดท utility list ของรายการ i_j ที่ถูกจัดเก็บใน 1-List ด้วย $<p, u(i_j, t_p), 0>$ พร้อมกับทำการคำนวณค่าความสม่ำเสมอจากการปรากฏขึ้นของ i_j ในทรานแซกชัน t_p (บรรทัดที่ 2 -4)

ขั้นตอนต่อไปจะเป็นการตรวจสอบการปรากฏขึ้นของแต่ละรายการ i_j ที่ถูกจัดเก็บใน 1-List ว่ามีความสม่ำเสมอหรือไม่? โดยทำการตรวจสอบค่าความสม่ำเสมอ $\boldsymbol{r^{i_j}}$ ของรายการ i_j ว่ามีค่ามากกว่าค่าขีดบางความสม่ำเสมอ $\boldsymbol{\sigma}_r$ ที่ผู้ใช้กำหนดหรือไม่ ถ้าค่าความสม่ำเสมอ $\boldsymbol{r^{i_j}}$ มีค่ามากกว่า $\boldsymbol{\sigma}_r$ HURI-UL จะลบรายการ i_j ออกจากการพิจารณา แต่ก่อนที่จะทำการลบ i_j ออกจากการพิจารณาจะทำการลดทอนค่าคุณประโยชน์ของแต่ละทรานแซกชันที่มีรายการ i_j ปรากฏเพื่อทำการลดทอนค่าประมาณคุณประโยชน์ โดยทำการพิจารณาแต่ละสมาชิกใน utility list ของรายการ i_j ที่ซึ่งมีลักษณะเป็น $<p, u(i_j, t_p), 0>$ และทำการลดทอนค่า $tu(t_p)$ ที่ถูกจัดเก็บใน tu-List ด้วยค่า $u(i_j, t_p)$ และเมื่อทำการพิจารณาทุกสมาชิกใน utility list ของรายการ i_j แล้ว จะสามารถลบข้อมูลทั้งหมดของ i_j ที่ถูกจัดเก็บอยู่ใน 1-list ออกจากหน่วยความจำและการพิจารณาได้ (บรรทัดที่ 6 – 10)

Algorithm 1 1-HURIs identification

Input: D : transactional database, σ_u : a minimum utility threshold, σ_r : a maximum regularity threshold

Output: *i-List* : a two-dimension list containing 1-HURIs and non 1-HURIs (with potentail to be HURIs with other items) in *1-List*

• create *tu-List*, *i-List* and then initial *1-List* with all single items
for each transaction t in database D **do**
  • compute $tu(t)$ and then collect $tu(t)$ in *tu-List*(t)
  for each item i in t **do**
    • update utility list of i in *1-List* on utiltiy and regularity value
for each item i in *1-List* **do**
  if $r^i > \sigma_r$ **then**
    for each entry e in utility list of i **do**
      • decrease utility value in *tu-List* by e's tid and e's $u(i, t_{tid})$
      • remove i and all of its information from *1-List*
• sort *1-List* based on order of items $\succ$
for each item i in *1-List* **do**
  • set $u(i)$ and $ru(i)$ to be zero
  for each entry $e = \langle tid_e, u(i)_e, ru(i)_e \rangle$ in utility list of item i **do**
    • update *tu-List*(tid_e) by *tu-List*$(tid_e) - u(i)_e$
    • set $ru(i)_e$ to be equal to *tu-List*$(tid_e) - u(i)_e$
    • increase $ru(i)$ by $ru(i)_e$
    • increase $u(i)$ by $u(i, t_{tid})$
  if $u(i) \geq \sigma_u$ **then**
    • HURIs = HURIs $\cup i$

รูปที่ 1 ขั้นตอนการระบุรายการที่มีค่าคุณประโยชน์สูงและปรากฏสม่ำเสมอ

ขั้นตอนต่อไปจะเป็นการคำนวณค่าคุณประโยชน์ส่วนเหลือในแต่ละสมาชิกใน utility list ของแต่ละรายการ i_j รวมถึงทำการคำนวณหาค่าคุณประโยชน์ที่แท้จริง $u(i_j)$ ของรายการ i_j และ ทำการคำนวณค่าคุณประโยชน์ส่วนเหลือ $ru(i_j)$ ของรายการ i_j (บรรทัดที่ 12 – 20) โดยในการพิจารณาจะทำการพิจารณาทีละรายการ i_j (จากลำดับของรายการที่ทราบก่อนหน้าแล้ว) จากนั้นทำการพิจารณาแต่ละสมาชิก <p, $u(i_j, t_p)$, 0> ใน utility list ของรายการ i_j จากนั้นทำการอัพเดทค่าคุณประโยชน์ของทรานแซกชัน t_p ที่ถูกจัดเก็บอยู่ใน tu-List ด้วย $tu(i_j, t_p) = tu(i_j, t_p) - u(i_j, t_p)$ เพื่อที่จะทราบถึงค่าคุณประโยชน์ของทุกรายการในทรานแซกชันที่อยู่ในลำดับถัดไปจาก i_j ซึ่งค่า $tu(i_j, t_p)$ หลังการอัพเดทจะหมายถึงค่า $ru(i_j, t_p)$ โดยเมื่อทราบถึงค่าดังกล่าว จะทำการอัพเดทสมาชิกใน utility list ของรายการ i_j ที่พิจารณาให้มีค่าเป็น <p, $u(i_j, t_p)$, $ru(i_j, t_p)$> และทำการอัพเดทค่า $u(i_j)$ และ $ru(i_j)$ ของรายการ i_j ด้วย $u(i_j, t_p)$ และ $ru(i_j, t_p)$ ตามลำดับ และท้ายสุด ถ้าค่า $u(i_j)$ ของรายการ i_j มีค่าไม่น้อยกว่าค่าขีดแบ่งคุณประโยชน์ที่ผู้ใช้กำหนด HURI-UL จะทำการระบุและจัดเก็บรายการ i_j ว่าเป็นรูปแบบที่มีค่าคุณประโยชน์สูงและปรากฏอย่างสม่ำเสมอ

หลังจากทำการสร้างลิสต์ของรายการที่ปรากฏอย่างสม่ำเสมอและมีแนวโน้มที่จะมีค่าคุณประโยชน์สูง (เมื่อรวมกับรายการอื่นๆ) ที่ซึ่งเก็บไว้ใน 1-List แล้ว ขั้นตอนต่อไปจะเป็นการค้นหารูปแบบที่มีค่าคุณประโยชน์สูงและปรากฏสม่ำเสมอทั้งหมดจาก 1-List ที่สร้างขึ้น (ดังแสดงในขั้นตอนวิธีในรูปที่ 2) โดยในขั้นแรก จะเป็นการการค้นหารูปแบบที่มีค่าคุณประโยชน์สูงและปรากฏสม่ำเสมอที่ประกอบไปด้วย 2 รายการ โดยเริ่มทำการพิจารณาแต่ละรายการ i_j ที่ถูกจัดเก็บใน 1-List (โดยในการพิจารณาจะพิจารณาแบบเรียงลำดับตามลำดับของรายการที่เป็นสมาชิกของเซต I) จากนั้นจะทำการสร้าง 2-List เพื่อใช้ในการจัดเก็บเซตรายการที่ประกอบด้วย 2 รายการที่ซึ่งเกิดจากการรวมกันระหว่างรายการ i_j กับรายการอื่นๆที่อยู่ในลำดับถัดๆไป โดยในตอนเริ่มแรกจะกำหนดให้ 2-List ไม่มีข้อมูลใดๆ ต่อมาจะเป็นการรวมรายการ i_j เข้ากับรายการ i_k ที่อยู่ในลำดับถัดๆไปจากรายการ i_j เพื่อสร้างการพิจารณาเซตรายการ $i_j \cup i_k$ จากนั้นจะทำการอินเทอร์เซกชัน utility list ของรายการ i_j กับ utility list ของรายการ i_k เพื่อทำการจัดเก็บข้อมูลการปรากฏขึ้น ค่าคุณประโยชน์ และค่าคุณประโยชน์ส่วนเหลือสำหรับทรานแซกชันหนึ่งๆ ที่เซตรายการ $i_j \cup i_k$ ปรากฏ หลังจากขั้นตอนอินเทอร์เซกชันเสร็จสิ้น เราจะได้ utility list ของเซตรายการ $i_j \cup i_k$ ที่จะสามารถใช้ utility list ดังกล่าวในการคำนวณค่าคุณประโยชน์ ค่าความสม่ำเสมอ และค่าประมาณคุณประโยชน์แบบกระชับของเซตรายการ $i_j \cup i_k$ ได้

หลังจากคำนวณค่าคุณประโยชน์ ค่าความสม่ำเสมอ และค่าประมาณคุณประโยชน์แบบกระชับของเซตรายการ $i_j \cup i_k$ แล้ว จะทำการตรวจสอบว่า เซตรายการ $i_j \cup i_k$ ปรากฏอย่างสม่ำเสมอหรือไม่ และทำการตรวจสอบว่า เซตรายการ $i_j \cup i_k$ มีแนวโน้มที่จะมีค่าคุณประโยชน์สูงหรือไม่ ถ้าเซตรายการ เซตรายการ $i_j \cup i_k$ ผ่านเงื่อนไขทั้งสองข้อข้างต้น HURI-UL จะทำการจัดเก็บเซตรายการ เซตรายการ $i_j \cup i_k$ พร้อมทั้งข้อมูลทั้งหมดที่เกี่ยวข้องกับเซตรายการ $i_j \cup i_k$ ไว้ใน 2-List เพื่อทำการพิจารณาเซตรายการที่ประกอบด้วย 3 รายการต่อไป โดยหลังจากทำการรวมรายการ i_j เข้ากับ i_k แล้ว ขั้นตอนวิธี HURI-UL จะดำเนินการรวมรายการ i_j เข้ากับ i_l ที่ถูกบรรจุอยู่ใน 1-List ในลำดับหลังจากรายการ i_k และจะดำเนินการตามกระบวนการต่างๆข้างต้น เมื่อทำการรวมรายการ i_j เข้ากับทุกรายการใน 1-List แล้ว เราจะได้ 2-List ที่บรรจุไปด้วยเซตรายการที่ประกอบไปด้วย 2 รายการ และเป็นเซตรายการ i_j เป็นรายการขึ้นต้น จากนั้นจะทำการตรวจสอบจำนวนเซตรายการที่บรรจุอยู่ใน 2-List ซึ่งถ้ามีจำนวนเซตรายการมากกว่า 1 จะทำการวนซ้ำการทำงานเพื่อหาเซตรายการที่ประกอบด้วย 3 รายการ โดยการดำเนินการจะดำเนินการเช่นเดียวกับการพิจารณาเซตรายการที่ประกอบไปด้วย 2 รายการ (ส่วนของ Procedure Gen-Longer-Itemsets)

Algorithm 2 Mining HURIs

Input: *i-List* : a list of itemsets
Output: HURIs : a complete set of HURIs

for each item i in *1-List* **do**
- initial *2-List* to be empty

for each item j in *1-List* ($i \prec j$) **do**
- merge i and j to be itemset Z
- intersect utility lists of u and v to compute r^Z, $u(Z)$, $ru(Z)$, $ou(Z)$ and then to collect occurrence information and utilities in UL^Z

if $r^Z \leq \sigma_r$ or $ou(Z) \geq \sigma_u$ **then**
- create an entry of itemset Z in *2-List with* r^Z, $u(Z)$, $ru(Z)$, $ou(Z)$ and UL^Z

if $u(Z) \geq \sigma_u$ **then**
- HURIs $=$ HURIs $\cup Z$

if |2-List| > 1 **then**
- Gen-Longer-Itemsets(2, *i-List*)

Procedure Gen-Longer-Itemsets(k, *i-List*)
- initial *(k+1)-List* to be empty

for each entry u in *k-List* **do**

for each entry v in *k-List* ($u \prec v$) **do**
- merge itemsets in entry u and v to create itemset Z
- intersect utility lists of u and v to compute r^Z, $u(Z)$, $ru(Z)$, $ou(Z)$ and then to collect occurrence information and utilities in UL^Z

if $r^Z \leq \sigma_r$ or $ou(Z) \geq \sigma_u$ **then**
- create an entry for itemset Z in *(k+1)-List* with r^Z, $u(Z)$, $ru(Z)$, $ou(Z)$ and UL^Z

if $u(Z) \geq \sigma_u$ **then**
- HURIs $=$ HURIs $\cup Z$

if |*(k+1)-List*| > 1 **then**
- Gen-Long-Itemsets(k+1, *i-List*)

รูปที่ 2 ขั้นตอนการหารูปแบบทั้งหมดที่มีค่าคุณประโยชน์สูงและปรากฏสม่ำเสมอ

บทที่ 4
ผลการทดลอง

ในบทนี้จะนำเสนอการทดสอบประสิทธิภาพการค้นหารูปแบบที่ปรากฏอย่างสม่ำเสมอและมีค่าคุณประโยชน์สูงที่ค้นหาจากขั้นตอนวิธี HURI-UL โดยจากการศึกษางานวิจัยที่เกี่ยวข้องพบว่ายังไม่มีงานวิจัยใดที่ทำการพิจารณาความน่าสนใจของรูปแบบภายใต้การพิจารณาความสม่ำเสมอของการปรากฏร่วมกับคุณประโยชน์ของรูปแบบนั้นๆ ดังนั้น การดำเนินการทดลองจะไม่มีการเปรียบเทียบประสิทธิภาพของขั้นตอนวิธี HURI-UL กับขั้นตอนวิธีอื่นๆจากงานวิจัยที่เกี่ยวข้อง เนื่องจากผลลัพธ์ที่ทำการพิจารณาในงานวิจัยนี้มีความแตกต่างกับผลลัพธ์ที่ทำการค้นหาจากงานวิจัยก่อนหน้า นี้จึงเป็นเหตุให้ผู้วิจัยไม่สามารถเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอกับขั้นตอนวิธีก่อนหน้าได้

แต่อย่างไรก็ตาม ในการดำเนินการทดลอง ผู้วิจัยได้กำหนดพารามิเตอร์ให้มีความใกล้เคียงกับงานวิจัยที่เกี่ยวข้อง กล่าวคือ ค่าขีดแบ่งความสม่ำเสมอจะกำหนดให้มีค่าอยู่ระหว่าง 1 – 10% ของจำนวนทรานแซกชันทั้งหมดในฐานข้อมูล (กล่าวคือ รูปแบบจะเป็นรูปแบบที่ปรากฏอย่างสม่ำเสมอ จะต้องมีค่าความสม่ำเสมอไม่เกิน 1 – 10% ของจำนวนทรานแซกชันทั้งหมดในฐานข้อมูล) และ ค่าขีดแบ่งคุณประโยชน์จะกำหนดให้มีค่าระหว่าง 0.1 – 1% ของค่าคุณประโยชน์ทั้งหมดของรูปแบบที่ปรากฏในฐานข้อมูล (กล่าวคือ รูปแบบจะเป็นรูปแบบที่มีค่าคุณประโยชน์สูง จะต้องมีค่าคุณประโยชน์ไม่น้อยกว่า 0.1 – 1% ของค่าคุณประโยชน์ทั้งหมดของรูปแบบที่ปรากฏในฐานข้อมูล) ตามลำดับ โดยในการทดสอบประสิทธิภาพของขั้นตอนวิธี HURI-UL ผู้วิจัยได้ทำการเขียนโปรแกรมการคำนวณตามขั้นตอนวิธี HURI-UL ด้วยภาษาซี และทำการทดสอบประสิทธิภาพในเครื่อง Xeon® 2.4 GHz ที่มีปริมาณหน่วยความจำ 64 GB

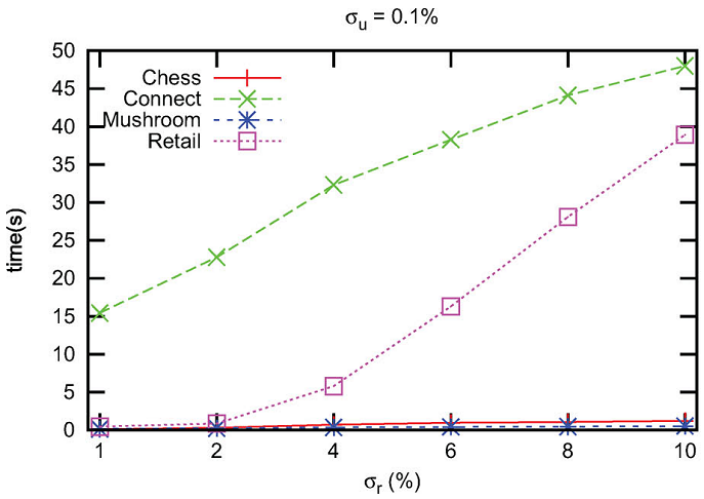
ในการทดสอบประสิทธิภาพจะดำเนินการทดสอบกับชัดข้อมูลจริง 4 ชุดข้อมูล โดยชุดข้อมูลที่ใช้ทดสอบสามารถดาว์นโหลดได้จาก P. F. Viger, "SPMF: An Open-Source Data Mining Library" โดยแต่ละชุดข้อมูลจะมีรายละเอียด ดังแสดงในตารางที่ 6

ตารางที่ 6 คุณลักษณะของชุดข้อมูลที่ใช้ในการทดสอบประสิทธิภาพของขั้นตอนวิธี HURI-UL

ชื่อฐานข้อมูล	จำนวนรายการที่ปรากฏ	จำนวนทรานแซกชัน	ความยาวเฉลี่ยของทรานแซกชัน	ชนิดของฐานข้อมูล
Chess	75	3,196	37	หนาแน่น
Connect	129	67,557	43	หนาแน่น
Mushroom	119	8,124	23	หนาแน่น
Retail	16,469	88,162	10.3	เบาบาง

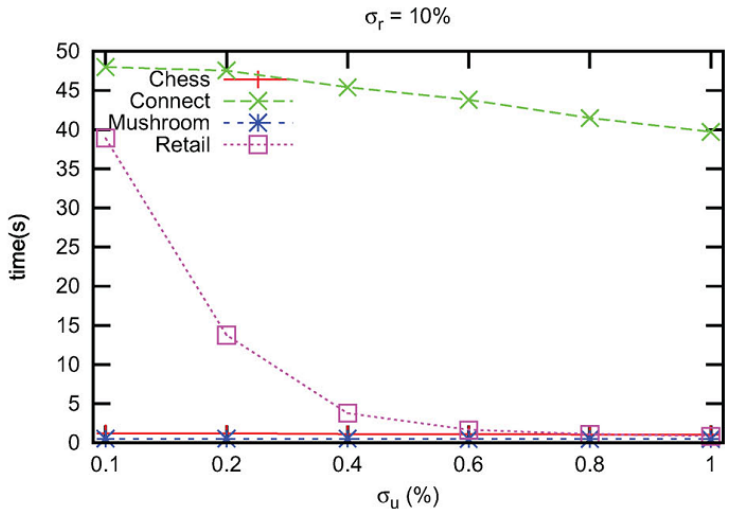
การทดลองที่ผู้วิจัยได้ทำการทดสอบประสิทธิภาพของขั้นตอนวิธี HURI-UL สามารถแบ่งได้เป็น 4 กรณี คือ 1) การทดสอบเวลาในการคำนวณเมื่อทำการกำหนดค่าขีดแบ่งคุณประโยชน์แบบตายตัวและกำหนดค่าขีดแบ่งความสม่ำเสมอแบบแปรปรวน 2) การทดสอบเวลาในการคำนวณเมื่อทำการกำหนดค่าขีดแบ่งคุณประโยชน์แบบแปรปรวนและกำหนดค่าขีดแบ่งความสม่ำเสมอแบบตายตัว 3) การพิจารณาจำนวนผลลัพธ์ที่ขั้นตอนวิธี HURI-UL สามารถค้นหาได้เมื่อทำการกำหนดค่าขีดแบ่งคุณประโยชน์แบบตายตัวและกำหนดค่าขีดแบ่งความสม่ำเสมอแบบแปรปรวน และ 4) การพิจารณาจำนวนผลลัพธ์ที่ขั้นตอนวิธี HURI-UL สามารถค้นหาได้เมื่อทำการกำหนดค่าขีดแบ่งคุณประโยชน์แบบแปรปรวนและกำหนดค่าขีดแบ่งความสม่ำเสมอแบบตายตัว ตามลำดับ

รูปที่ 3 แสดงเวลาที่ใช้ในการประมวลผลของขั้นตอนวิธี HURI-UL กับ 4 ชุดข้อมูลภายใต้การกำหนดค่าขีดแบ่งคุณประโยชน์เท่ากับ 0.1% ซึ่งเป็นค่าต่ำสุดของค่าขีดแบ่งคุณประโยชน์ที่ทำการพิจารณาที่ซึ่งจะทำให้มีรูปแบบเป็นจำนวนมากมีค่าคุณประโยชน์มากกว่าหรือเท่ากับค่าขีดแบ่งคุณประโยชน์ โดยค่าดังกล่าวจะทำให้การค้นหารูปแบบใช้เวลามาก ซึ่งจะสะท้อนให้เห็นถึงประสิทธิภาพของขั้นตอนวิธี HURI-UL เมื่อต้องพิจารณารูปแบบเป็นจำนวนมาก นอกจากนั้นการทดลองในรูปข้างต้นจะมีการแปรปรวนของค่าขีดแบ่งความสม่ำเสมอเพื่อทำการศึกษาถึงผลกระทบของการกำหนดค่าขีดแบ่งความสม่ำเสมอกับประสิทธิภาพการคำนวณของขั้นตอนวิธี HURI-UL โดยจากรูป จะสังเกตุได้ว่าเมื่อค่าขีดแบ่งความสม่ำเสมอมีค่าเพิ่มขึ้นจะทำให้ขั้นตอนวิธี HURI-UL ใช้เวลาในการคำนวณเพิ่มขึ้น เนื่องจากเมื่อค่าขีดแบ่งความสม่ำเสมอมีค่าเพิ่มขึ้นจะทำให้มีรูปแบบเป็นจำนวนมากขึ้นมีค่าความสม่ำเสมอน้อยกว่าหรือเท่ากับค่าขีดแบ่งความสม่ำเสมอ ซึ่งจะส่งผลให้ขั้นตอนวิธี HURI-UL ใช้เวลาในการพิจารณารูปแบบต่างๆเพิ่มขึ้นด้วยเช่นกัน



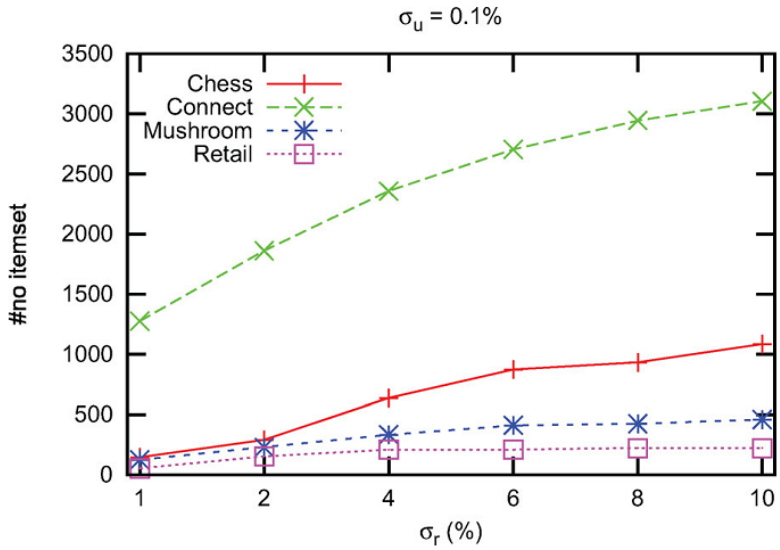
รูปที่ 3 เวลาที่ใช้ในการคำนวณของขั้นตอนวิธี HURI-UL เมื่อทำการเปลี่ยนแปลงค่าขีดแบ่งความสม่ำเสมอ

ในทางตรงกันข้าม รูปที่ 4 จะแสดงเวลาที่ใช้ในการประมวลผลของขั้นตอนวิธี HURI-UL กับ 4 ชุดข้อมูล ภายใต้การกำหนดค่าขีดแบ่งความสม่ำเสมอแบบตายตัวให้มีค่าเท่ากับ 10% โดยจากการกำหนดดังกล่าวจะทำให้มีรูปแบบเป็นจำนวนมากมีค่าความสม่ำเสมอน้อยกว่าหรือเท่ากับค่าขีดแบ่งสม่ำเสมอที่กำหนด ซึ่งจะทำให้เราได้ทราบถึงเวลาที่ใช้ในการประมวลผลของขั้นตอนวิธี HURI-UL เมื่อต้องทำการพิจารณารูปแบบเป็นจำนวนมาก นอกจากนั้นการทดลองในรูปข้างต้นยังเป็นการทดลองภายใต้การกำหนดค่าขีดแบ่งคุณประโยชน์แบบแปรปรวนที่มีการกำหนดค่าขีดแบ่งคุณประโยชน์ตั้งแต่ 0.1 – 1% โดยจากรูปจะสามารถสังเกตุได้ว่า เมื่อค่าขีดแบ่งคุณประโยชน์มีค่าเพิ่มขึ้นจะทำให้เวลาที่ใช้ในการคำนวณของขั้นตอนวิธี HURI-UL น้อยลง เนื่องจากเมื่อค่าขีดแบ่งคุณประโยชน์มีค่าสูงขึ้นที่ซึ่งจะทำให้มีรูปแบบเป็นจำนวนน้อยมีค่าคุณประโยชน์สูงกว่าหรือเท่ากับค่าขีดแบ่งคุณประโยชน์ และเมื่อจำนวนรูปแบบที่ต้องทำการพิจารณามีจำนวนน้อยลง ขั้นตอนวิธี HURI-UL จึงใช้เวลาในการคำนวณน้อยลงด้วยเช่นกัน

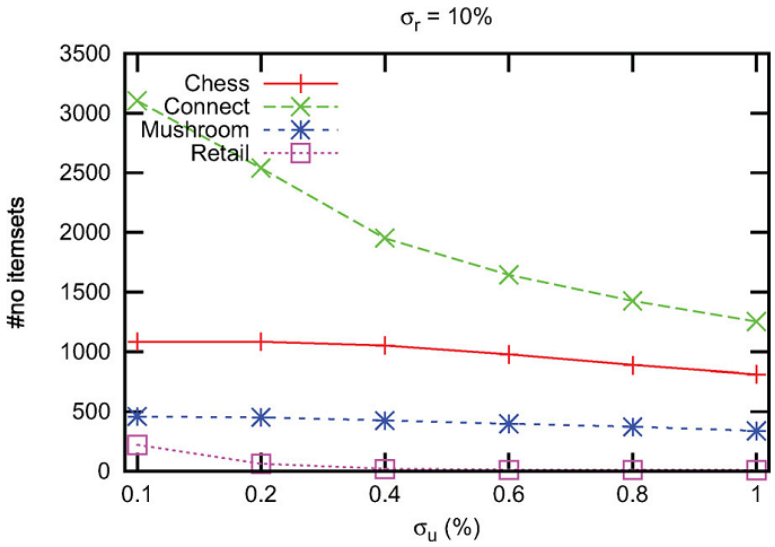


รูปที่ 4 เวลาที่ใช้ในการคำนวณของขั้นตอนวิธี HURI-UL เมื่อทำการเปลี่ยนแปลงค่าขีดแบ่งคุณประโยชน์

ในส่วนของการพิจารณาจำนวนผลลัพธ์ที่สามารถค้นหาได้จากขั้นตอนวิธี HURI-UL ภายใต้การกำหนดค่าขีดแบ่งคุณประโยชน์แบบตายตัวและกำหนดค่าขีดแบ่งความสม่ำเสมอแบบแปรปรวน และการกำหนดค่าขีดแบ่งคุณประโยชน์แบบแปรปรวนและกำหนดค่าขีดแบ่งความสม่ำเสมอแบบตายตัว จะแสดงได้ดังรูปที่ 5 และ 6 ที่จะสังเกตุได้ว่าจำนวนผลลัพธ์จะมีจำนวนเพิ่มขึ้นอย่างมีนัยสำคัญเมื่อทำการกำหนดค่าขีดแบ่งความสม่ำเสมอเพิ่มขึ้น และจะมีจำนวนลดลงเมื่อทำการกำหนดค่าขีดแบ่งคุณประโยชน์เพิ่มขึ้น



รูปที่ 5 จำนวนผลลัพธ์ที่ค้นหาได้จากขั้นตอนวิธี HURI-UL เมื่อทำการเปลี่ยนแปลงค่าขีดแบ่งความสม่ำเสมอ



รูปที่ 6 จำนวนผลลัพธ์ที่ค้นหาได้จากขั้นตอนวิธี HURI-UL เมื่อทำการเปลี่ยนแปลงค่าขีดแบ่งคุณประโยชน์

บทที่ 5

สรุปผลการวิจัย

งานวิจัยนี้ได้นำเสนอแนวความคิดในการค้นหารูปแบบที่น่าสนใจภายใต้การพิจารณาความสม่ำเสมอของการปรากฏและค่าคุณประโยชน์ของรูปแบบนั้นๆ (อาจอยู่ในรูปแบบของผลกำไร ต้นทุน หรือ อื่นๆ) ซึ่งการค้นหารูปแบบภายใต้การพิจารณาข้างต้นจะมีส่วนช่วยในการสังเกตพฤติกรรมการซื้อของลูกค้าที่ทำการซื้อสินค้าที่ให้ผลกำไรสูงที่ซึ่งจะเป็นข้อมูลที่จะช่วยให้ผู้บริหารสามารถวิเคราะห์ความต้องการลูกค้า บริหารจัดการคลังสินค้า และสามารถวางแผนการตลาดได้ดียิ่งขึ้น จากกรอบความคิดข้างต้น รูปแบบหนึ่งๆ จะเป็นรูปแบบที่น่าสนใจก็ต่อเมื่อรูปแบบนั้นๆ เป็นรูปแบบที่ปรากฏ (ถูกซื้อ) อย่างสม่ำเสมอ (กล่าวคือ รูปแบบมีค่าความสม่ำเสมอไม่มากกว่าค่าขีดแบ่งความสม่ำเสมอ) และ เป็นรูปแบบที่มีค่าคุณประโยชน์ที่ได้รับจากการปรากฏขึ้นของรูปแบบ (กล่าวคือ รูปแบบมีค่าคุณประโยชน์ไม่น้อยกว่าค่าขีดแบ่งคุณประโยชน์)

ในการค้นหารูปแบบอย่างมีประสิทธิภาพ ผู้วิจัยได้นำเสนอขั้นตอนวิธีที่เรียกว่า HURI-UL (High Utility Regular Itemsets using Utility-List) ที่ทำการอ่านข้อมูลจากฐานข้อมูลเพียงครั้งเดียว ขั้นตอนวิธี HURI-UL ที่นำเสนอยังได้ประยุกต์ใช้แนวความคิดเกี่ยวกับ ค่าคุณประโยชน์คงเหลือ (remaining utility) และ ค่าประมาณคุณประโยชน์ (overestimated utility) เพื่อทำการลดทอนปริภูมิสถานะ นอกจากนั้น HURI-UL ประยุกต์ใช้โครงสร้างข้อมูลในรูปแบบลิสต์ (utility list) เพื่อทำการจัดเก็บข้อมูลการปรากฏขึ้นและค่าคุณประโยชน์ของรูปแบบหนึ่ง และ ท้ายสุด การค้นหาผลลัพธ์แบบเชิงลึก (depth first search) ได้ถูกประยุกต์ใช้เพื่อทำการค้นหาผลลัพธ์ได้อย่างมีประสิทธิภาพ

ผู้วิจัยได้ทำการทดลองเพื่อทดสอบประสิทธิภาพของขั้นตอนวิธีที่เสนอกับข้อมูลจริง 4 ชุดข้อมูล โดยทำการทดลองในสองแง่มุมด้วยกันคือ เวลาที่ใช้ในการหาผลลัพธ์ และจำนวนผลลัพธ์ที่สามารถค้นหาได้ โดยการทดลองชี้ให้เห็นว่า ขั้นตอนวิธี HURI-UL สามารถค้นหารูปแบบที่ปรากฏอย่างสม่ำเสมอและมีค่าคุณประโยชน์สูงได้อย่างมีประสิทธิภาพ

บรรณานุกรม

1. Tanbeer, S.K., Ahmed, C.F., Jeong, B.S., and Lee, Y.K. (2009). Discovering periodic-frequent patterns in transactional databases. Proceedings of PAKDD.
2. Fu, A.W.C., w. Kwong, R.W., and Tang, J. (2000). Mining n-most interesting itemsets. Proceedings of the 12th International Symposium on Foundations of Intelligent Systems, Springer-Verlag.
3. Wang, J., Han, J., Lu, Y., and Tzvetkov, P. 2005. Tfp: an efficient algorithm for mining top-k frequent closed itemsets. Proceedings of the IEEE Transactions on Knowledge and Data Engineering. Volume 17.
4. Yang, B., Huang, H., Wu, Z. 2008. Topsis: Finding top-k significant n-itemsets in sliding windows adaptively. Journal of Knowledge-Based Systems. 21(6).
5. Li, H.F. 2009. Mining top-k maximal reference sequences from streaming web click sequences with a damped sliding window. Journal of Expert Systems with Applications 36(8).
6. Ke, Y., Cheng, J., and Yu, J.X. 2009. Top-k correlative graph mining. Proceedings of SDM, SIAM.
7. Fournier-Viger, P., and Tseng, V.S. 2013. Tns: mining top-k non-redundant sequential rules. Proceedings of SAC, ACM.
8. Amphawan, K., Lenca, P., and Surarerks, A. 2009. Mining top-k periodic-frequent patterns without support threshold. Proceeding of IAIT. Volume 55 of CCIS., Springer.
9. Amphawan, K., Lenca, P., and Surarerks, A. 2011. Efficient mining top-k regular-frequent itemset using compressed tidsets. Proceedings of New Frontiers in Applied Data Mining. Volume 7104 of Lecture Notes in Computer Science.
10. Amphawan, K., Lenca, P., and Surarerks, A. 2012. Mining top-k regular-frequent itemsets using database partitioning and support estimation. Journal of Expert Systems with Applications 39(2).
11. Tanbeer, S.K., Ahmed, C.F., and Jeong, B.S. 2010. Mining regular patterns in incremental transactional databases. Proceedings of Int. Asia-Pacific Web Conference, IEEE Computer Society.
12. Tanbeer, S.K., Ahmed, C.F., and Jeong, B.S. 2010. Mining regular patterns in data streams. Proceedings of DASFAA. Volume 5981 of LNCS., Springer.
13. Surana, A., Kiran, R.U., and, Reddy, P.K. 2012. An efficient approach to mine periodic-frequent patterns in transactional databases. Proceedings of PAKDD Workshops.
14. Kiran, R.U., and Reddy, P.K. 2010. Mining periodic-frequent patterns with maximum items' support constraints. Proceedings of the Third Annual ACM Bangalore Conference. COMPUTE '10.
15. Yao, H., Hamilton, H. J. and Butz, C. J. 2004. A Foundational Approach to Mining Itemset Utilities from Databases. Proceedings of the 4th SIAM International Conference on Data Mining.

16. Yao, H., and Hamilton, H.J. 2006. Mining itemset utilities from transaction databases. Journal of Data & Knowledge Engineering.
17. Liu, Y., Liao, W.-K., and Choudhary, A. 2005. A Two Phase algorithm for fast discovery of High Utility of Itemsets. Proceedings of PAKDD.
18. Tseng, V., Wu, C. W., Shie, B. E. and Yu, P. 2010. UP-Growth: An Efficient Algorithm for High Utility Itemsets Mining. Proceedings of the 16th ACM SIG KDD Conference on Knowledge Discovery and Data Mining.
19. Liu, M. and Qu, J. 2012. Mining High Utility Itemsets without Candidate Generation. Proceedings of the ACM international conference on Information and Knowledge Management (CIKM).

ภาคผนวก

ICAICTA 2015

The 2015 International Conference On Advanced Informatics: Concepts, Theory And Application

August 19-22, 2015

The Tide Resort, Bang Saen Beach,

Chonburi, Thailand

ISBN : 978-1-74799-1966-6

Presentation Program, Day 1: August 20 Caribbean I

Time	Caribbean I
08:45 - 09:20	Opening Ceremony
09:20 - 10:10	Keynote: Dr. Ayu Purwarianti (ITB)
10:10 - 10:30	Coffee break
10:30 - 10:50	Relation between EMG Signal Activation & Time Lags using Feature Analysis during Dynamic Contraction *Nizam Uddin Ahamed, S.A.M MatiurRahaman, Mahdi Alqahtani, Omar Altwijri, Nasim Ahmed, Kenneth Sundaraj*
10:50 - 11:10	Acoustic and Language Models Adaptation for Indonesian Spontaneous Speech Recognition *Dessi Puji Lestari and Angela Irfani*
11:10 - 11:30	Automatic Multilabel Classification for Indonesian News Articles *Dyah Rahmawati and Masayu Leylia Khodra*
11:30 - 11:50	Combining Temporal and Content Aware Features for Microblog Retrieval *Abu Nowshed Chy, Md Zia Ullah and Masaki Aono*
11:50 - 12:10	Simulating Crowd Movement in Agent-based Model of Large-Scale Flood *Suvalak Vijitpornkul and Worawan Marurngsith*
12:10 - 13:30	Luncheon
13:30 - 13:50	A Framework of Fundamental News Summarization To determine the direction of Foreign Excange Rate *Yusran T. Samuel, Dwi H. Widyantoro and Aciek I. Wuryandari*
13:50 - 14:10	Compilation and Evaluation of Paraphrase Representation List of Compound Verbs Toward Development of "Control Language for Action" *Tomoya Shirai, Hirofumi Yabumoto, Kyoko Kanzaki and Hitoshi Isahara*
14:10 - 14:30	Incorporating Text Information on Presentation Slides for Spoken Lecture Retrieval *Kosuke Yamauchi and Tomoyoshi Akiba*
14:30 - 14:50	Event Information Extraction from Indonesian Tweets using Conditional Random Field *Fawwaz Muhammad and Masayu Leylia Khodra*
14:50 - 15:10	Coffee break
15:10 - 16:00	Keynote: Dr. Keiji Yamada (NEC Asia Pacific Pte. Ltd.)
16:20 - 16:40	Elliptic Curve Cryptography: Algorithms and Implementation Analysis over Coordinate Systems *Iskandar Setiadi, Achmad Imam Kistijantoro and Atsuko Miyaji*
16:40 - 17:00	Integration of Diesel Engine Monitoring System and Recommendation System *Nur Ulfa Maulidevi, Mahdan Ahmad Fauzi Al-Hasan and Masayu Leylia Khodra*
17:00 - 17:20	Confidence Interval of Probability Estimator of Laplace Smoothing *Masato Kikuchi, Mitsuo Yoshida, Masayuki Okabe and Kyoji Umemura*
18:00 - 20:30	Banquet @ Poolside Le Casa Hotel

Presentation Program, Day 1: August 20 — Caribbean II

Time	Caribbean II
08:45 - 09:20	Opening Ceremony
09:20 - 10:10	Keynote: Dr. Ayu Purwarianti (ITB)
10:10 - 10:30	Coffee break
10:30 - 10:50	Pushing Regularity Constraint on High Utility Itemsets Mining *Komate Amphawan and Athasit Surarerks*
10:50 - 11:10	Event Extraction on Indonesian News Article Using Multiclass Categorization *Masayu Leylia Khodra*
11:10 - 11:30	Improving Emotion Classification in Imbalanced YouTube Dataset Using SMOTE Algorithm *Phakhawat Sarakit, Thanaruk Theeramunkong and Choochart Haruechaiyasak*
11:30 - 11:50	A Framework for Laptop Review Analysis *Thanapat Chatchaithanawat and Pakawan Pugsee*
11:50 - 12:10	A New Feature Selection Based on Class Dependency and Feature Dissimilarity *Niphat Claypo and Saichon Jaiyen*
12:10 - 13:30	Luncheon
13:30 - 13:50	Suggestion Analysis for Food Recipe Improvement *Pakawan Pugsee and Monsinee Niyomvanich*
13:50 - 14:10	SAE Syntactic-based Aspect and Opinion Extraction from Product Reviews *Warih Maharani, Dwi H. Widyantoro and Masayu L. Khodra*
14:10 - 14:30	Effectiveness of Social Media Text Classification by Utilizing the Online News Category *Phat Jotikabukkana, Virach Sornlertlamvanich, Okumura Manabu and Choochart Haruechaiyasak*
14:30 - 14:50	Automatically Relation Modeling On Spatial Relationship As Self-Adaptation Ability *Iping Supriana Suwardi and Aradea*
14:50 - 15:10	Coffee break
15:10 - 16:00	Keynote: Dr. Keiji Yamada (NEC Asia Pacific)
16:20 - 16:40	The Monitoring Management System for Tangible Cultural Heritage Surveillance *Chaowalit Nalad, Thatsanee Charoenporn and Nattapong Tongtep*
16:40 - 17:00	Instruments Measurement Design of Human Behavior in Collaborative Software Construction *Tien Fabrianti Kusumasari, Kridanto Surendro, Husni Sastramihardja and Iping Supriana*
17:00 - 17:20	Developing a Part-Time Lecturer Appointment Notification System using UCD *Suchanon Sengchuan, Wittawas Pantumchinda and Athita Onuean*
18:00 - 20:30	Banquet @ Poolside Le Casa Hotel

Presentation Program, Day 1: August 20 — Caribbean III

Time	Caribbean III
08:45 - 09:20	Opening Ceremony
09:20 - 10:10	Keynote: Dr. Ayu Purwarianti (ITB)
10:10 - 10:30	Coffee break
10:30 - 10:50	Lattice Boltzmann Method for Two-dimensional Shallow Water Equations with CUDA *Jittavat Suksumlarn, Worasait Suwannik and Montri Maleewong*
10:50 - 11:10	Energy Efficient Routing in Cluster based Wireless Sensor Network *Sounak Paul and Tapan Kumar Dey*
11:10 - 11:30	Impact of End System Scheduling Policies on AFDX Performance in Avionic On-Board Data Network *Chakkaphong Suthaputchakun, Kin-Man-Benjamin Lee and Zhili Sun*
11:30 - 11:50	TippyDB: Geographically-Aware Distributed NoSQL Key-Value Store *Iskandar Setiadi and Achmad Imam Kistijantoro*
11:50 - 12:10	IFIX: A New Information Exchange Framework for Financial organizations *Pheerasak Tongkamonwat and Pattarasinee Bhattarakosol*
12:10 - 13:30	Luncheon
13:30 - 14:10	Quantum communications & computing *KeattisakSripimanwat, ECTI Association*
14:10 - 14:50	(NBTC Project: Technology Transfer and Human Resource Development of Perfectly Secure Quantum Communications *SuwitKiravittaya, Naresuan University*
14:50 - 15:10	Coffee break
15:10 - 16:00	Keynote: Dr. Keiji Yamada (NEC Asia Pacific)
18:00 - 20:30	Banquet @ Poolside Le Casa Hotel

Presentation Program, Day 2: August 21 — Caribbean I

Time	Caribbean I
09:00 - 09:20	The Bilateral Denoising Performance Influence of Window, Spatial and Radiometric Variance *Vorapoj Patanavijit*
09:20 - 09:40	System Development of Commercial Logo Analysis on Online Social Media *Dwi H. Widyantoro and Tino E.K. Sambora*
09:40 - 10:00	Objective Assessment and Quantification of Pearl Quality by Spectral-Spatial Features *Yuki Ota, Hiroshi Higashi and Shigeki Nakauchi*
10:00 - 10:20	Spectral-Difference Enhancing Illuminant for Improving Visual Detection of Blood Vessels

	Kazuya Ito, Yuki Ota, Hiroshi Higashi and Shigeki Nakauchi
10:20 - 10:40	Coffee break
10:40 - 11:00	Comparison of Skype VoIP Quality over 3G with Mobility: A Case Study of Fair Usage Policy Effects *Phisitsak Phusamchot, Pongpisit Wuttidittachotti,Vajirasak Vanijja and Therdpong Daengsi*
11:00 - 11:20	Equity in distributive justice to virtual characters *Mariko Hyodo, Shoji Itakura and Michiteru Kitazaki*
11:30 - 12:00	Closing
12:00 - 13:30	Luncheon
13:30 - 20:00	Social Study

Presentation Program, Day 2: August 21 Caribbean II

Time	Caribbean II
09:00 - 09:20	Handling Arbitrary Polygon Query based on the Boolean Overlay on a Geographical Information System *Iping Supriana Suwardi, Dessi Puji Lestari and Dicky Prima Satya*
09:20 - 09:40	Interactive Fluid Simulation Based on Material Point Method for Mobile Devices *Dody Dharma and Afwarman Manaf*
09:40 - 10:00	Smart Men Cloth-Dressing Robot Using DC Stepper Motor with RFID and Fuzzy-PID Control System *Songkran Kantawong*
10:00 - 10:20	Automated detection of spiculated masses using integrated method based on Active Contour *Piyatragoon Boonthong, Annupan Rodtook, Suwanna Rasmequan and Krisana Chinnasarn*
10:20 - 10:40	Coffee break
10:40 - 11:00	Autonomous Glaucoma Detection from Fundus Image using Cup to Disc Ratio and Hybrid Features *Anum Abdul Salam, M. Usman Akram, Kamran Wazir, Syed Muhammad Anwar and Muhammad Majid*
11:00 - 11:20	Color restoration of lighting scenes with locally adapted HDR images *Yuto Kubo, Takao Jinno and Shigeru Kuriyama*
11:30 - 12:00	Closing
12:00 - 13:30	Luncheon
13:30 - 20:00	Social Study

Presentation Program, Day 2: August 21 — Caribbean III

Time	Caribbean III
09:00 - 09:20	Frame-based Expert System for Monitoring Diesel Engine *Riyandika Andhi Saputra, Masayu Leylia Khodra and Nur Ulfa Maulidevi*
9:20 - 09:40	English to Japanese Spoken Lecture Translation System by Using DNN-HMM and Phrase-based SMT *Norioki Goto, Kazumasa Yamamoto and Seiichi Nakagawa*
09:40 - 10:00	Analyzing Yelp Reviews of Restaurants for Categorizing Health Related Cues to Action *Shanza Abbas, M. Usman Akram, Arslan Shaukat and Muazzam A Khan*
10:00 - 10:20	A Study on the Efficiency of Creating Stories by the use of Templates *Seiya Kawagoe, Miki Ueno and Hitoshi Isahara*
11:30 - 12:00	Closing
12:00 - 13:30	Luncheon
13:30 - 20:00	Social Study

Table of Content

Graphics, Image Processing and Intelligent Systems

Color restoration of lighting scenes with locally adapted HDR images
Yuto Kubo, Takao Jinno and Shigeru Kuriyama

Frame-based Expert System for Monitoring Diesel Engine
Riyandika Andhi Saputra, Masayu Leylia Khodra and Nur Ulfa Maulidevi

English to Japanese Spoken Lecture Translation System by Using DNN-HMM and Phrase-based SMT
Norioki Goto, Kazumasa Yamamoto and Seiichi Nakagawa

A Study on the Efficiency of Creating Stories by the use of Templates
Seiya Kawagoe, Miki Ueno and Hitoshi Isahara

Information Systems, Audit and Governance

Pushing Regularity Constraint on High Utility Itemsets Mining
Komate Amphawan and Athasit Surarerks

Event Extraction on Indonesian News Article Using Multiclass Categorization
Masayu Leylia Khodra

Improving Emotion Classification in Imbalanced YouTube Dataset Using SMOTE Algorithm
Phakhawat Sarakit, Thanaruk Theeramunkong and Choochart Haruechaiyasak

A Framework for Laptop Review Analysis
Thanapat Chatchaithanawat and Pakawan Pugsee

A New Feature Selection Based on Class Dependency and Feature Dissimilarity
Niphat Claypo and Saichon Jaiyen

Suggestion Analysis for Food Recipe Improvement
Pakawan Pugsee and Monsinee Niyomvanich

SAE Syntactic-based Aspect and Opinion Extraction from Product Reviews
Warih Maharani, Dwi H. Widyantoro and Masayu L. Khodra

Effectiveness of Social Media Text Classification by Utilizing the Online News Category
Phat Jotikabukkana, Virach Sornlertlamvanich, Okumura Manabu and Choochart Haruechaiyasak

Automatically Relation Modeling On Spatial Relationship As Self-Adaptation Ability
Iping Supriana Suwardi and Aradea

The Monitoring Management System for Tangible Cultural Heritage Surveillance
Chaowalit Nalad, Thatsanee Charoenporn and Nattapong Tongtep

Developing a Part-Time Lecturer Appointment Notification System using UCD
Suchanon Sengchuan, Wittawas Pantumchinda and Athita Onuean

High Performance Computing and Distributed Systems

Lattice Boltzmann Method for Two-dimensional Shallow Water Equations with CUDA
Jittavat Suksumlarn, Worasait Suwannik and Montri Maleewong

Energy Efficient Routing in Cluster based Wireless Sensor Network
Sounak Paul and Tapan Kumar Dey

Impact of End System Scheduling Policies on AFDX Performance in Avionic On-Board Data Network

Computational Science and Engineering

Pushing regularity constraint on high utility itemsets mining

Komate Amphawan
Computational Innovation Laboratory, Informatics,
Burapha university, Chonburi, 20131, Thailand
Email: komate@gmail.com

Athasit Surarerks
ELITE, Computer Engineering, Faculty of Engineering,
Chulalongkorn University Bangkok, 10330, Thailand
Email: athasit.s@chula.ac.th

***Abstract*—High utility itemsets mining (HUIM) is an interesting topic in data mining which can be applied in a wide range of applications, for example, on retail marketing–finding sets of sold products giving high profit, low cost, etc. However, HUIM only considers utility values of items/itemsets which may be insufficient to observe buying behavior of customers. To address this issue, we here introduce an approach to add regularity constraint into high utility itemsets mining. Based on this approach, sets of co-occurrence items with high utility values and regular occurrence, called *high utility-regular itemsets* (HURIs), are regarded as interesting itemsets. To mine HURIs, an efficient single-pass algorithm, called *HURI-UL*, is proposed. *HURI-UL* applies concept of remaining and overestimated utilities of itemsets to early prune search space (uninteresting itemsets) and also utilizes utility list structure to efficiently maintain utility values and occurrence information of itemsets. Experimental results on real datasets show that our proposed *HURI-UL* is efficient to discover high utility itemsets with regular occurrence.**

***Index Terms*—Data mining; Itemsets mining; High utility itemsets; Regularity constraint**

I. Introduction

Association rule mining (ARM) [1] is a fundamental research topic in data mining. It aims to discover a set of interesting rules from transactional database. ARM consists of 2 main steps: (*i*) discovering sets of items with frequency of occurrence no less than a user-specified support threshold, called *frequent itemsets*, and (*ii*) generating rules by frequent itemsets generated from step (*i*) that meets a user-given confidence threshold, respectively. Since the first proposal of Agrawal et al. [1], there are several extensions of ARM such as improving efficiency of frequent itemsets mining [2], [3], sequential patterns mining [4], mining quantitative association rules [5], closed and maximal itemsets mining [6], [7], n-most interesting and top-*k* frequent itemsets mining [8], [9], weighted-frequent itemsets mining [10], emerged itemsets mining [11], frequent-regular itemsets mining [12], [13], etc.

Unfortunately, most of above approaches only considers frequency and/or regularity of occurrences which cannot reflect the utility of itemsets. Thus, Chan et al. [14] proposed to consider the importances of items (*e.g.* profit, cost, and other user defined) including items' quantity of occurrence in transactions and then introduced the task of high-utility itemsets mining (HUIM). Based on HUIM, an itemset is called a *high-utility itemset* (*HUI*) if its utility (*i.e.* unit profit × quantity of occurrence) is no less than a user-specified utility threshold; otherwise, it is called a *low utility itemset* (*LUI*). High-utility itemsets mining has a wide range of applications such as cross-marketing in retail, website click stream, biomedical applications and mobile commerce.

Since the first attempt [14], HUIM is developed in various aspects such as improving efficiency of high utility itemsets mining [15], [16], mining high utility itemsets from incremental database/data streams [17], [18], top-*k* high-utility itemsets mining [19], [20], mining high utility sequential pattern [21], [22], mining high utility itemsets based on positive and/or negative unit profit [23], [24], etc. However, the consideration only utility may not be sufficient to observe interesting buying behavior of customer. For example, from database of Fig. 1, item 'h' tends to be a high utility itemsets due to its utility is 315 (*i.e.* (20 × 15)+ (1 × 15)). However, 'h' occurs only twice in transactional database and one of them has a heavy quantity of occurrence (as in the second transaction). On the other hands, the utility of item 'd' per piece (*i.e.* 30) is highest among all items, then 'd' and its superset might be HUIs even if 'd' has only few quantity of occurrence.

To address this issue, we introduce a new approach to discover itemsets based on their utility and regularity of occurrence, called *high utility-regular itemsets* (HURIs). These itemsets can tell us–"how is about regularity of purchases on products giving high profit"–and can help us to know customers' demand, to manage inventory, to make new promotion and so on. To mine HURIs, an efficient algorithm named *HURI-UL* (*High Utility-Regular Itemsets mining based on Utility List*) is proposed. *HURI-UL* can avoid repeatedly scan of database by scanning database only once. In addition, *HURI-UL* applies the concept of remaining utility calculation to prune search space and also utilizes utility list structure [25] to efficiently maintain utilities and occurrence information of itemsets. Experiments were conducted on both real and synthetic datasets in order to evaluate the performance of *HURI-UL*. Experimental results show that our proposed algorithm is efficient to discover high utility itemsets with regular occurrence.

The rest of this paper is organized as follows. Section II describes the basic concepts and notations for discovering HURIs. Section III introduces details of utility list structure and *HURI-UL* algorithm. Experiments are shown in Section IV. Finally, conclusions are given in Section V.

978-1-4673-8143-7/15/$31.00

II. PROBLEM STATEMENT

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items. Each item $i_j \in I$ has its own *external utility*, denoted as $eu(i_j)$, such as profit, cost, etc. A set $X \subseteq I$ is called a k-itemset, if X contains k items. A transactional database $D = \{t_1, t_2, \ldots, t_m\}$ is a set of m transactions such that each transaction $t_p \in D$ is a 2-tuple containing: (*i*) a unique identifier p (also called *tid*) and (*ii*) an itemset $Y \subseteq I$ where each $i_j \in Y$ is associated with *internal utility* (*i.e.* quantity of its own occurrence in t_p), denoted as $iu(i_j, t_p)$). If $X \subseteq Y$, it can be said that t_p contains X or X occurs in transaction t_p, denoted as t_p^X. Then, the set $T^X = \{t_p^X, \ldots, t_q^X\}$ is the ordered set w.r.t. *tid*s of transactions that contain X. The support value of X (*i.e.* frequency of occurrence) can be defined as $s^X = |T^X|$.

Definition 1: *The utility of item i_j occurs in transaction t_p, **defined as** $u(i_j, t_p) = iu(i_j, t_p) \times eu(i_j)$, is the product between internal utility of i_j in transaction t_p and external utility of i_j.*

Definition 2: *The utility of itemset X contained in transaction t_p, **defined as** $u(X, t_p) = \sum_{i_j \in X, X \in t_p} u(i_j, t_p)$, is the summation of the utilities of all items in X that occurs in t_p.*

Definition 3: *The utility of itemset X in database D, **denfinded as** $u(X) = \sum_{t_p \in D, X \in t_p} u(X, t_p)$, is the summation of the utilities of X in all transactions containing X.*

Based on three definitions above, the problem of high utility itemsets mining is the task of finding all itemsets whose utilities are not less than a user-specified utility threshold (σ_u). The main challenge of high utility itemsets mining is that *downward closure property* cannot be hold, since a superset of a low-utility itemset may be a high-utility one. Hence, the concept of *Transaction-Weighted Utility (TWU)* [26], an overestimated utility of an itemset that meets *downward closure property*, is proposed based on following definitions.

Definition 4: *The utility of transaction t_p, **defined as** $u(t_p) = \sum_{i_j \in t_p} u(i_j, t_p)$, is the summation of the utilities of all items occur in t_p.*

Definition 5: *The transaction-weighted utility of itemset X, **defined as** $TWU(X) = \sum_{t_p \in D, X \in t_p} tu(t_p)$, is the upper bound of utility value of X that refers to the summation of transaction utilities of all transactions containing X.*

Based on the concept of *TWU*, the *downward closure property* can be hold and it can be said that an itemset X and all supersets of X are low-utility itemsets, if $TWU(X) < \sigma_u$. However, *TWU* of an itemset is a loose upper bound of utility. Therefore, Liu et al. [25] proposed a new concept of tight upper bound of utility with less overestimated than *TWU*. Then, the following definitions are introduced for this purpose.

Definition 6: *Let $\succ$ be order on items from I. The remaining utility of X in a transaction t_p containing X, defined as $ru(X, t_p) = \sum_{i_j \in t_p, X \prec i_j} u(i_j, t_p)$, is the summation of utilities of all items ordered after X in t_p.*

Definition 7: *The remaining utility of an itemset X in a database D, **defined as** $ru(X) = \sum_{t_p \in D, X \in t_p} ru(X, t_p)$, is the summation of all remaining utilities of X in all transactions containing X.*

Definition 8: *The overestimated utility of an itemset X in a database D based on $\succ$, **defined as** $ou(X) = u(X) + ru(X)$, is the summation of the utility and the remaining utilities of X in database D.*

Based on the notions of overestimated utility, it can be said that an itemset X and all of its supersets are low-utility itemsets if $ou(X) < \sigma_u$. Thus, we can apply this concept to cut-down search space.

As proposed by Tanbeer et al. [12], notations on regularity of occurrence can be defined as follows.

Definition 9: *The regularity of itemset X from two consecutive transactions containing X (i.e. t_p^X and t_q^X where $p < q$ and there is no transaction between both transactions containing X), **defined as** $r(t_p^X, t_q^X) = q - p$, is the gap of occurrence of X between t_p^X and t_q^X.*

Definition 10: *The regularity of X in a database D, **defined as** $r(X) = max(r(t_p^X, t_q^X), \ldots, r(t_y^X, t_z^X))$ where $1 \leq p \leq q \leq y \leq z \leq |D|$, is the maximal gap of occurrence of X in database D.*

Problem statement. Given a database D, a user-given minimum utility threshold σ_u and a maximum regularity threshold σ_r, the problem of mining high utility-regular itemsets is to discover a complete set of itemsets whose (*i*) utility values are not less than σ_u, and (*ii*) regularity values are not greater than σ_r, respectively.

III. PROPOSED METHOD

In this section, we first introduces the main concept of utility list structure as proposed in [25] and then describes details of *HURI-UL* for mining high utility-regular itemsets.

A. Utility list

Based on the concept of the order on items ($\succ$), a utility list of an itemset X can be defined as an ordered set of 3-tuples: $\langle p, u(X, t_p), ru(X, t_p)\rangle$ where p is a tid of transaction containing X, $u(X, t_p)$ is the utility of X in the transaction t_p, and $ru(X, t_p)$ is the remaining utility of all items ordered after X ($\succ X$) in transaction t_p.

For example, consider item 'a' in database of Fig. 1 with external utility $eu(a) = 3$ and the occurrence of 'a' in transactions $\{t_1, t_3, t_5, t_7, t_8\}$. If the order of item is $a \prec b \prec \ldots \prec h$, then the utility list of a can be expressed as $UL^a = \{\langle 1, 9, 73\rangle, \langle 3, 6, 34\rangle, \langle 5, 6, 19\rangle, \langle 7, 15, 42\rangle, \langle 8, 9, 41\rangle\}$. Each element of UL^a contains 3 information. For example, the first element $\langle 1, 9, 73\rangle$ lets us know that (*i*) a occurs in t_1, (*ii*) utility of a in t_1, $u(a, t_1)$, is equal to 9, and (*iii*) remaining utility of a in t_1, $ru(a, t_1)$, is equal to 73, respectively.

B. HURI-UL algorithm

As mentioned above, *HURI-UL* not only applies the concept of remaining and overestimated utilities to cut-down search space but also utilizes utility list structure to efficiently maintain utility values and occurrence information of itemsets. Moreover, to avoid repeatedly scan database, a simple list, named *tu-List*, used for storing transaction utilities for all transactions is created. *HURI-UL* consists of 2 main steps:

a	b	c	d	e	f	g	h
3	2	1	30	5	3	4	15

(a) External utility of items

tid	items(internal utility)
1	a(3), c(8), d(2), e(1)
2	b(5), f(3), g(5), h(20)
3	a(2), c(4), d(1)
4	c(5), e(1), f(1)
5	a(2), b(3), c(1), f(4)
6	d(1), g(5), h(1)
7	a(5), b(1), d(1), e(2)
8	a(3), b(1), c(4), d(1), e(1)

(b) Transactional database

Fig. 1. Transactional database with internal utilities and external utility

1) **Identifing 1-HURIs** contained in *i-List*–a simple two-dimension list that separately stores itemsets by its own size, for example, *1-List* for 1-itemsets, *2-List* for 2-itemsets, and *k-List* for *k*-itemsets, respectively. Each entry in the *k-List* contains 5 pieces of information: itemset name I, its regularity r^I, its utility value $u(I)$, its remaining utility $ru(I)$, and its utility list UL^I,respectively.
2) **Mining a complete set of HURIs** from *i-List*.

1) Identifying 1-HURIs (algorithm 1): HURI-UL first initials *tu-List*, *i-List* and then creates entries for all items in *1-List*. Next, each transaction t in database is read and its utility is calculated and collected in *tu-List*. For each item i in transaction t, its $u(i,t)$ is computed and then collected into its utility list in *1-List* with *tid* of t. After scanning all of transactions, items with regularity value greater than the user-given threshold are identified as non-regular items (*i.e.* items that are not regularly appear in database) and then removed from *1-List*. Transaction utilities in *tu-List* corresponding to utilities of non-regular items are updated. Next, items in *1-List* are ordered by $\succ$ and remaining utility of each item on each transaction is calculated and collected in utility list. Actual and remaining utilities of each item are also calculated to know the overestimated utility of each items. Lastly, items with overestimated utilities less than the user-given utility threshold are removed from *1-List*. At the end of identifying 1-HURIs step, we gain 1-HURIs and non 1-HURIs (items with overestimated utilities no less than utility threshold) contained in *1-List* and ready to mine larger itemsets.

2) Mining a complete set of HURIs (algorithm 2): Each item i in *1-List* is considered. Then, item i is merged with other items in *1-List* (one by one) based on $\succ$ in order to generate 2-itemsets. For each merging, utility list of both items are intersected (in the same manner as [25]) in order to calculate regularity, utility, overestimated utility and to collect utility list of the new generated 2-itemset (itemset Z). If regularity of Z is not greater than regularity threshold and overestimated utility of Z is not less than utility threshold, an entry of Z is created in *2-List* with its information. After generating all of 2-itemsets corresponding to i, the merging process is repeatedly performed until there is only one or no more *k*-itemset contained in *k-List*. Then, *HURI-UL* move the consideration to another item j in the *1-List* (based on $\succ$) and then do the same manner as above. After consider all items in *1-List* based on merging step, we gain a complete set of HURIs that meet regularity and utility constraints.

Algorithm 1 1-HURIs identification

Input: D : transactional database, σ_u : a minimum utility threshold, σ_r : a maximum regularity threshold
Output: *i-List* : a two-dimension list containing 1-HURIs and non 1-HURIs (with potentail to be HURIs with other items) in *1-List*

• create *tu-List*, *i-List* and then initial *1-List* with all single items
for each transaction t in database D **do**
  • compute $tu(t)$ and then collect $tu(t)$ in *tu-List*(t)
  for each item i in t **do**
    • update utility list of i in *1-List* on utiltiy and regularity value
for each item i in *1-List* **do**
  if $r^i > \sigma_r$ **then**
    for each entry e in utility list of i **do**
      • decrease utility value in *tu-List* by e's tid and e's $u(i, t_{tid})$
    • remove i and all of its information from *1-List*
• sort *1-List* based on order of items $\succ$
for each item i in *1-List* **do**
  • set $u(i)$ and $ru(i)$ to be zero
  for each entry $e = \langle tid_e, u(i)_e, ru(i)_e \rangle$ in utility list of item i **do**
    • update *tu-List*(tid_e) by *tu-List*(tid_e) $- u(i)_e$
    • set $ru(i)_e$ to be equal to *tu-List*(tid_e) $- u(i)_e$
    • increase $ru(i)$ by $ru(i)_e$
    • increase $u(i)$ by $u(i, t_{tid})$
  if $u(i) \geq \sigma_u$ **then**
    • HURIs = HURIs $\cup i$

C. Example of HURI-UL

An example is given to illustrate how *HURI-UL* works for discovering a complete set of HURIs. Assume that regularity and utility threshold are set to be 3 and 50 (*i.e.* $\sigma_r = 3$ and $\sigma_u = 50$), respectively. A database with internal utility and external utilities of all items are shown in Fig. 1. From all of above parameters, itemsets with utility not less than 50 and regularly occur in database (*i.e.* itemsets can be disappear from

Algorithm 2 Mining HURIs

Input: *i-List* : a list of itemsets
Output: HURIs : a complete set of HURIs

for each item i in *1-List* **do**
- initial *2-List* to be empty

for each item j in *1-List* ($i \prec j$) **do**
- merge i and j to be itemset Z
- intersect utility lists of u and v to compute r^Z, $u(Z)$, $ru(Z)$, $ou(Z)$ and then to collect occurrence information and utilities in UL^Z

if $r^Z \leq \sigma_r$ or $ou(Z) \geq \sigma_u$ **then**
- create an entry of itemset Z in *2-List with* r^Z, $u(Z)$, $ru(Z)$, $ou(Z)$ and UL^Z

if $u(Z) \geq \sigma_u$ **then**
- HURIs = HURIs $\cup Z$

if |2-List| > 1 **then**
- Gen-Longer-Itemsets(2, *i-List*)

Procedure Gen-Longer-Itemsets(k, *i-List*)
- initial *(k+1)-List* to be empty

for each entry u in *k-List* **do**
for each entry v in *k-List* ($u \prec v$) **do**
- merge itemsets in entry u and v to create itemset Z
- intersect utility lists of u and v to compute r^Z, $u(Z)$, $ru(Z)$, $ou(Z)$ and then to collect occurrence information and utilities in UL^Z

if $r^Z \leq \sigma_r$ or $ou(Z) \geq \sigma_u$ **then**
- create an entry for itemset Z in *(k+1)-List* with r^Z, $u(Z)$, $ru(Z)$, $ou(Z)$ and UL^Z

if $u(Z) \geq \sigma_u$ **then**
- HURIs = HURIs $\cup Z$

if |*(k+1)-List*| > 1 **then**
- Gen-Long-Itemsets(k+1, *i-List*)

database at most 3 consecutive transactions) are regarded as interesting.

First, *tu-List* and *1-List* are first initialed. Next, each transaction in database is scanned, For the first transaction $t_1 = \{a(3), c(8), d(2), e(1)\}$, its transaction utility $tu(t_1)$ is computed (*i.e.* $(3 \times 3) + (8 \times 1) + (2 \times 30) + (1 \times 5) = 82$) and then collected in *tu-List*. In addition, entries for item a, c, d, and e are updated as illustrated in Fig. 2(a) (Notice that to save space, each entry of *1-List* only shows item-name, regularity value and utility list). For the second transaction $t_2 = \{b(5), f(3), g(5), h(20)\}$, *tu-List* and the entries of items b, f, g, and h are updated as shown in Fig. 2(b). For the third until the last transaction, *HURI-UL* acts in the same manner as above and then we gain *1-List* as in Fig. 2(c). The entries of items g and h are eliminated from *1-List*, since their regularity is greater than σ_r. Lastly, the remaining utility on transactions of each item is calculated and updated into its utility list (see Fig. 2(d)).

Next, an item a is considered and then merged together with items b, c, d, e and f, respectively. For each merging like a merged with b, both items are mered to be itemset $Z\langle a, b\rangle$. Then, utility lists UL^a and UL^b are intersected together in order to compute $r^{\langle a,b\rangle} = 5$, $u(\langle a, b\rangle) = 40$, $ou(\langle a, b\rangle) = 132$ and to collect $UL^{\langle a,b\rangle} = \{\langle 5, 12, 13\rangle, \langle 7, 17, 40\rangle, \langle 8, 11, 39\rangle\}$, respectively. Since $r^{\langle a,b\rangle}$ is greater than σ_r, then $\langle a, b\rangle$ does not regularly occur in database and it can be removed out of the consideration (based on the downward-closure property). The merging process is repeated for all pairs of item a and the others. Next, *HURI-UL* tries to generate longer itemsets such as 3-itemsets, 4-itemsets, ..., k-itemsets by recursively considering pairs of itemsets stored in *i-List*. Then, *HURI-UL* moves the consideration to other items such as b, c, d, e and applies the merging, intersection, and recursively generate long-itemsets to identify a complete set HURIs.

IV. Expermental evaluation

In this section, we show some experiments done to investigate performance of our proposed *HURI-UL* algorithm. However, from the best of our knowledge, this is the first attempt to add regularity constraint to high utility itemsets. Thus, there is no comparative study provided. *HURI-UL* is implemented on C and run on Xeon® 2.4 GHz with 64 GB of RAM. The regularity and utility thresholds are varied between $1-10\%$ and $0.1-1\%$ (similar parameter as previous works [14], [27], [12], [13]) to observe computational time and number of itemsets generated from *HURI-UL*. Four datasets downloaded from [28] are used as shown in Table I.

TABLE I
Database characteristics

Database	#items	Avg.length	#Transactions	type
chess	75	37	3,196	dense
connect	129	43	67,557	dense
mushroom	119	23	8,124	dense
retail	16,469	10.3	88,162	sparse

Figure 3 shows the computational time of *HURI-UL* on the variation of regularity threshold and a fix value of utility threshold. From the figure, higher value of regularity threshold causes higher computational time. The reason is that on high regularity threshold, there are more and more items/itemsets that meet the threshold. Then, *HURI-UL* has to spend more time to consider these itemsets. In contrast, Fig. 4 indicates the runtime of *HURI-UL* on the variation of utility threshold and a fix value of regularity threshold. From Fig. 4, it can be observed that runtime decreases as the increasing of utility threshold. This is because with the high utility threshold, there are less items/itemsets that meet the threshold. Then, this make a reduction on merging and on intersection of pairs of utility lists.

To observe number of itemsets discovered from *HURI-UL*, experiments were conducted in the same way as runtime investigation. Figure 5 shows number of discovered itemsets on variation of regularity threshold and a fix value of utility

tu-List

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
82	0	0	0	0	0	0	0

1-List

a,1, {[1,9,-]}	b,0, { }	c,1, {[1,8,-]}	d,1, {[1,60,-]}	e,1, {[1,5,-]}	f,0, { }	g,0, { }	h,0, { }

(a) After scanning t_1

tu-List

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
82	339	0	0	0	0	0	0

1-List

a,1, {[1,9,-]}	b,2, {[2,10,-]}	c,1, {[1,8,-]}	d,1, {[1,60,-]}	e,1, {[1,5,-]}	f,2, {[2,9,-]}	g,2, {[2,20,-]}	h,2, {[2,300,-]}

(b) After scanning t_2

tu-List

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
82	339	40	13	25	65	57	50

1-List

a,2, {[1,9,-], [3,6,-], [5,6,-] [7,15,-] [8,9,-]}	b,3, {[2,10,-], [5,6,-], [7,2,-], [8,2,-]}	c,3, {[1,8,-], [3,4,-], [4,5,-], [5,1,-], [8,4,-]}	d,3, {[1,60,-], [3,30,-], [6,30,-], [7,30,-], [8,30,-]}	e,3, {[1,5,-], [4,5,-], [7,10,-], [8,5,-]}	f,3, {[2,9,-], [4,3,-], [5,12,-]}	~~g,4,~~ {[2,20,-], [6,20,-]}	~~h,4,~~ {[2,300,-], [6,15,-]}

(c) After scanning all transactions

tu-List

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
82	10	40	10	25	30	57	50

1-List

a,2, {[1,9,73], [3,6,34], [5,6,19] [7,15,42] [8,9,41]}	b,3, {[2,10,0], [5,6,13], [7,2,40], [8,2,39]}	c,3, {[1,8,65], [3,4,30], [4,5,5], [5,1,12], [8,4,35}	d,3, {[1,60,5], [3,30,0], [6,30,0], [7,30,10], [8,30,5]}	e,3, {[1,5,0], [4,5,0], [7,10,0], [8,5,0]}	f,3, {[2,9,0], [4,3,0], [5,12,0]}

(d) 1-List containing all 1-HURIs

Fig. 2. *tu-List* and *i-List* gained from 1-HURIs identification process

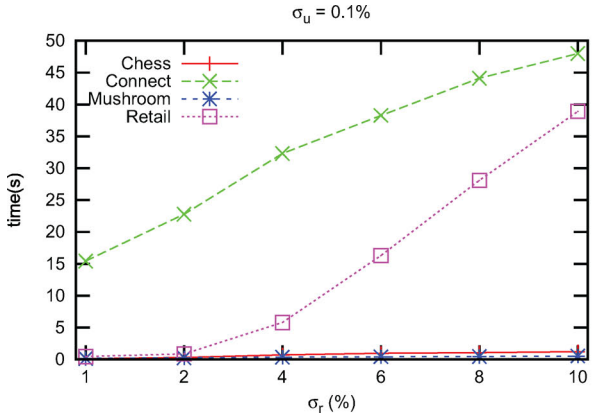


Fig. 3. Runtime with the variation of regularity threshold

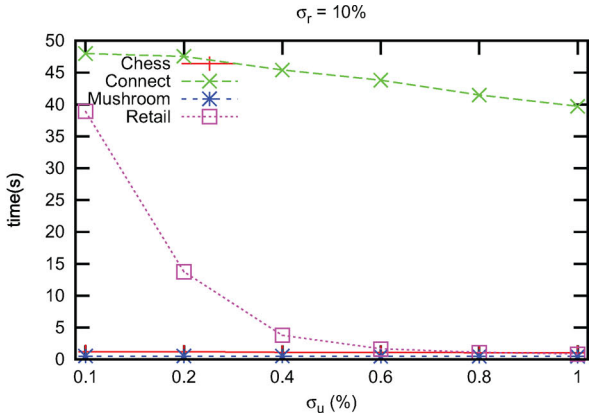


Fig. 4. Runtime with the variation of utility threshold

threshold. Meanwhile, Fig. 6 indicates number of discovered itemsets on a different way. From both figures, it can be seen that high regularity threshold causes *HURI-UL* can generate more results than the low one. With high regularity threshold, there are thousands of itemsets that can meets the thresholds. Meanwhile, high utility threshold results in *HURI-UL* generate less itemsets than the low one (the reason is in contrast with the value of regularity threshold).

V. Conclusion

In this paper, we have introduced a new approach to add regularity constraint on high utility itemsets mining. This can help to observe buying behavior of customer based on sold products giving high profit, low cost, etc. In this approach, any set of items that regularly appears in database and gives high utility is interesting. To mine such itemsets, an efficient single-pass named *HURI-UL* is proposed. *HURI-UL* applies the concept of remaining utility, overestimated utility and also utilizes utility list to efficiently prune search space. Experiments were conducted based on four real datasets and results show that our

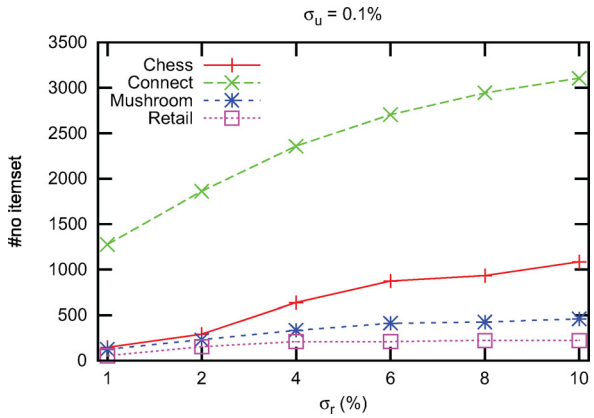


Fig. 5. Number of discovered itemsets with the variation of regularity threshold

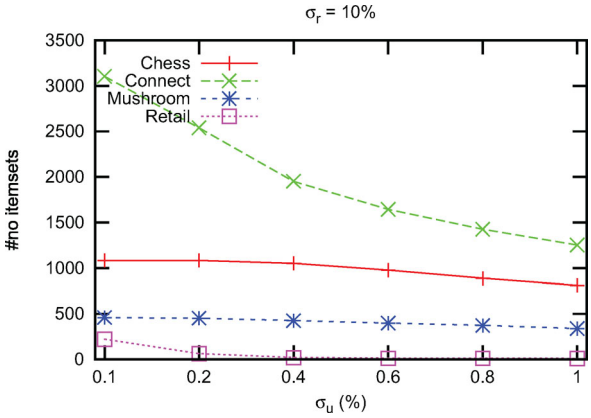


Fig. 6. Number of discovered itemsets with the variation of utility threshold

proposed *HURI-UL* is efficient to discover valuable itemsets.

Acknowledgment

The authors would like to thank National Research Council of Thailand (NRCT) for supports.

References

[1] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the ACM SIGMOD*, 1993, pp. 207–216.

[2] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *VLDB*, 1994, pp. 487–499.

[3] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *Proceedings of the ACM SIGMOD*, 2000, pp. 1–12.

[4] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Data Engineering, Proceedings of the Eleventh International Conference on*, 1995, pp. 3–14.

[5] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables," in *Proceedings of the ACM SIGMOD*, 1996, pp. 1–12.

[6] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering frequent closed itemsets for association rules," in *Proceedings of the 7th International Conference on Database Theory*, vol. 1540, 1999, pp. 398–416.

[7] D. Burdick, M. Calimlim, and J. Gehrke, "MAFIA: a maximal frequent itemset algorithm for transactional databases," in *Proceedings of the 17th International Conference on Data Engineering, April 2-6, 2001, Heidelberg, Germany*. IEEE Computer Society, 2001, pp. 443–452.

[8] A. W.-C. Fu, R. W. w. Kwong, and J. Tang, "Mining n-most interesting itemsets," in *Proceedings of the 12th International Symposium on Foundations of Intelligent Systems*, 2000, pp. 59–67.

[9] J. Han, J. Wang, Y. Lu, and P. Tzvetkov, "Mining top-k frequent closed patterns without minimum support," in *Proceedings of 2002 IEEE International Conference on Data Mining (ICDM 2002)*, 2002, pp. 211–218.

[10] C. Cai, A. Fu, C. Cheng, and W. Kwong, "Mining association rules with weighted items," in *Database Engineering and Applications Symposium, 1998. Proceedings. IDEAS'98. International*, 1998, pp. 68–77.

[11] G. Dong and J. Li, "Efficient mining of emerging patterns: Discovering trends and differences," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '99, 1999, pp. 43–52.

[12] S. K. Tanbeer, C. F. Ahmed, B.-S. Jeong, and Y.-K. Lee, "Discovering periodic-frequent patterns in transactional databases," in *Proceedings of 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 2009, pp. 242–253.

[13] K. Amphawan, P. Lenca, and A. Surarerks, "Mining top-k periodic-frequent patterns without support threshold," in *Proceedings of the 3rd International Conference on Advances in Information Technology*, vol. 55, 2009, pp. 18–29.

[14] R. Chan, Q. Yang, and Y.-D. Shen, "Mining high utility itemsets," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, 2003, pp. 19–26.

[15] V. S. Tseng, C.-W. Wu, B.-E. Shie, and P. S. Yu, "Up-growth: An efficient algorithm for high utility itemset mining," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 253–262.

[16] U. Yun, H. Ryang, and K. H. Ryu, "High utility itemset mining with techniques for reducing overestimated utilities and pruning candidates," *Expert Systems with Applications*, vol. 41, no. 8, pp. 3861–3878, 2014.

[17] C.-W. Lin, G.-C. Lan, and T.-P. Hong, "An incremental mining algorithm for high utility itemsets," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7173 – 7180, 2012.

[18] U. Yun and H. Ryang, "Incremental high utility pattern mining with static and dynamic databases," *Applied Intelligence*, vol. 42, no. 2, pp. 323–352, 2015.

[19] M. Zihayat and A. An, "Mining top-k high utility patterns over data streams," *Information Sciences*, vol. 285, no. 0, pp. 138 – 161, 2014.

[20] H. Ryang and U. Yun, "Top-k high utility pattern mining with effective threshold raising strategies," *Knowledge-Based Systems*, vol. 76, no. 0, pp. 109 – 126, 2015.

[21] J. Yin, Z. Zheng, and L. Cao, "Uspan: An efficient algorithm for mining high utility sequential patterns," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12, 2012, pp. 660–668.

[22] G.-C. Lan, T.-P. Hong, V. S. Tseng, and S.-L. Wang, "Applying the maximum utility measure in high utility sequential pattern mining," *Expert Systems with Applications*, vol. 41, no. 11, pp. 5071 – 5081, 2014.

[23] H.-F. Li, H.-Y. Huang, and S.-Y. Lee, "Fast and memory efficient mining of high-utility itemsets from data streams: with and without negative item profits," *Knowledge and Information Systems*, vol. 28, no. 3, pp. 495–522, 2011.

[24] P. Fournier-Viger, "FHN: Efficient mining of high-utility itemsets with negative unit profits," in *Advanced Data Mining and Applications*, 2014, pp. 16–29.

[25] M. Liu and J. Qu, "Mining high utility itemsets without candidate generation," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 2012, pp. 55–64.

[26] Y. Liu, W.-k. Liao, and A. Choudhary, "A two-phase algorithm for fast discovery of high utility itemsets," in *Advances in Knowledge Discovery and Data Mining*, 2005, vol. 3518, pp. 689–695.

[27] C. Ahmed, S. Tanbeer, B.-S. Jeong, and Y.-K. Lee, "Efficient tree structures for high utility pattern mining in incremental databases," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 21, no. 12, pp. 1708–1721, 2009.

[28] P. F. Viger, "SPMF: An Open-Source Data Mining Library," http://www.philippe-fournier-viger.com/spmf/, 2015.

ประวัติคณะผู้วิจัย

1. หัวหน้าโครงการวิจัย

1. ชื่อ - นามสกุล (ภาษาไทย) นายโกเมศ อัมพวัน
   ชื่อ - นามสกุล (ภาษาอังกฤษ) Mr. Komate Amphawan
2. ตำแหน่งปัจจุบัน ผู้ช่วยศาสตราจารย์และอาจารย์ประจำคณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา
3. หน่วยงานและสถานที่อยู่ที่ติดต่อได้สะดวก พร้อมหมายเลขโทรศัพท์ โทรสาร และไปรษณีย์อิเล็กทรอนิกส์ (e-mail)

   คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา 169 ถ. ลงหาดบางแสน ต. แสนสุข อ. เมือง จ. ชลบุรี 20131 หมายเลขโทรศัพท์ 038-103-097, 081-862-7819 หมายเลขโทรสาร 038-393-245 ไปรษณีย์อิเล็กทรอนิกส์ komate@buu.ac.th, komate@gmail.com
4. ประวัติการศึกษา

   วท.บ. (วิทยาการคอมพิวเตอร์) มหาวิทยาลัยบูรพา

   วท.ม. (วิทยาการคอมพิวเตอร์) จุฬาลงกรณ์มหาวิทยาลัย

   วศ.ด. (วิศวกรรมคอมพิวเตอร์) จุฬาลงกรณ์มหาวิทยาลัย
5. สาขาวิชาการที่มีความชำนาญพิเศษ (แตกต่างจากวุฒิการศึกษา) ระบุสาขาวิชาการ

   ปัญญาประดิษฐ์ การทำเหมืองข้อมูล การวิเคราะห์ข้อมูล โครงสร้างข้อมูล ขั้นตอนวิธี
6. ประสบการณ์ที่เกี่ยวข้องกับการบริหารงานวิจัยทั้งภายในและภายนอกประเทศ โดยระบุสถานภาพในการทำการวิจัยว่าเป็นผู้อำนวยการแผนงานวิจัย หัวหน้าโครงการวิจัย หรือผู้ร่วมวิจัยในแต่ละผลงานวิจัย

   หัวหน้าโครงการวิจัย (ในรอบ 5 ปีที่ผ่านมา) :

   1. การค้นหาความเปลี่ยนแปลงของรูปแบบภายใต้ความถี่และความสม่ำเสมอของการปรากฏเพื่อค้นหาแนวโน้มและความเปลี่ยนแปลงของพฤติกรรมผู้บริโภค (ทุนอุดหนุนการวิจัย สำนักงานคณะกรรมการวิจัยแห่งชาติ ประจำปีงบประมาณ 2558)
   2. การเพิ่มประสิทธิภาพผลลัพธ์สำหรับการค้นหารูปแบบที่ปรากฏบ่อยสุดเคอันดับแรกและปรากฏอย่างสม่ำเสมอ (ทุนอุดหนุนการวิจัย ประเภทเงินรายได้ คณะวิทยาการสารสนเทศ ประจำปีงบประมาณ 2557)
   3. การสร้างพจนานุกรมสำหรับการจำแนกคำประสมและการค้นหาคำ (ทุนอุดหนุนการวิจัย ประเภทเงินรายได้ คณะวิทยาการสารสนเทศ ประจำปีงบประมาณ 2556)

4. การค้นหารูปแบบที่เกิดขึ้นบ่อยและสม่ำเสมอเคอันดับแรกจากฐานข้อมูลรายการ (ทุนอุดหนุนการวิจัย ประเภทเงินรายได้ คณะวิทยาการสารสนเทศ ประจำปีงบประมาณ 2555)

งานวิจัยที่ทำเสร็จแล้ว (ในรอบ 5 ปีที่ผ่านมา) :

1. K. Amphawan, A. Surarerks, "Pushing Regularity Constraint on High Utility Itemsets Mining", The 2015 International Conference on Advanced Informatics: Concepts, Theory And Application (ICAICIT2015) [Best paper award]
2. K. Amphawan, P. Lenca, "Mining top-k frequent-regular closed patterns", Expert Systems with Applications, Elsevier, vol. 42(21), pp. 7882-7894, 2015.
3. K. Amphawan, P.Sittichaitaweekul, "Mining top-k frequent-regular patterns based on user-given length constraints", The 19th International Annual Symposium on Computational Science and Engineering, 2015.
4. S. Chompaisal, K. Amphawan, A. Surarerks, "Mining N-most Interesting Multi-level Frequent Itemsets without Support Threshold", Proceedings of Recent Advances in Information and Communication Technology, 2014.
5. P. Sittichaitaweekul, K. Amphawan, "Enhancing quality of results on Top-k Frequent-Regular Pattern mining", Proceedings of International Conference on Engineering Science and Innovative Technology, 2014.
6. K. Amphawan and P. Lenca, "Mining top-k frequent-regular patterns based on user-given trade-off between frequency and regularity", Proceedings of the 6th International Conference on Advances in Information Technology: IAIT-2013, 2013.
7. K. Amphawan, "SST: An efficient suffix-sharing trie structure for dictionary lookup", Proceedings of 7th Asia international conference on mathematical modeling and computer simulation, 2013.
8. K. Amphawan and A. Surarerks, "An efficient method for constructing dictionary based on decompounding words

technique", The 17th International Annual Symposium on Computational Science and Engineering, 2013.

9. K. Amphawan, P. Lenca, and A. Surarerks, "Mining top-k regular-frequent itemsets using database partitioning and support estimation", Expert Systems with Applications, Volume 39, Issue 2, February 1, Pages 1924-1936, 2012.
10. K. Amphawan K, P. Lenca , and A. Surarerks, "Efficient mining top-k regular-frequent itemset using compressed tidsets", Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7104 LNAI , pp. 124-135, 2011.

งานวิจัยที่กำลังทำ :

1. การค้นหาความเปลี่ยนแปลงของรูปแบบภายใต้ความถี่และความสม่ำเสมอของการปรากฏเพื่อค้นหาแนวโน้มและความเปลี่ยนแปลงของพฤติกรรมผู้บริโภค (ทุนอุดหนุนการวิจัย สำนักงานคณะกรรมการวิจัยแห่งชาติ ประจำปีงบประมาณ 2558)

2. ผู้ร่วมวิจัยคนที่ 1

1. ชื่อ - นามสกุล (ภาษาไทย) นายอรรถสิทธิ์ สุรฤกษ์
   ชื่อ - นามสกุล (ภาษาอังกฤษ) Mr. Athasit Surarerks
2. ตำแหน่งปัจจุบัน ผู้ช่วยศาสตราจารย์และอาจารย์ประจำภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
3. หน่วยงานและสถานที่อยู่ที่ติดต่อได้สะดวก พร้อมหมายเลขโทรศัพท์ โทรสาร และไปรษณีย์อิเล็กทรอนิกส์ (e-mail)

   ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย หมายเลขโทรศัพท์ 02-218-6996, 081-626-6116 หมายเลขโทรสาร 02-218-6955 ไปรษณีย์อิเล็กทรอนิกส์ athasit.s@chula.ac.th
4. ประวัติการศึกษา

   ปริญญาตรี B. Science (Mathematics)
   ปริญญาโท DEA (A.I.)
   ปริญญาเอก Ph.D. (Algorithms)

5. สาขาวิชาการที่มีความชำนาญพิเศษ (แตกต่างจากวุฒิการศึกษา) ระบุสาขาวิชาการ

ปัญญาประดิษฐ์ การทำเหมืองข้อมูล การวิเคราะห์ข้อมูล โครงสร้างข้อมูล ขั้นตอนวิธี ทฤษฎีการคำนวณ การคำนวณทางคณิตศาสตร์

6. ประสบการณ์ที่เกี่ยวข้องกับการบริหารงานวิจัยทั้งภายในและภายนอกประเทศ โดยระบุสถานภาพในการทำการวิจัยว่าเป็นผู้อำนวยการแผนงานวิจัย หัวหน้าโครงการวิจัย หรือผู้ร่วมวิจัยในแต่ละผลงานวิจัย

หัวหน้าโครงการวิจัย (ในรอบ 5 ปีที่ผ่านมา) : -
งานวิจัยที่ทำเสร็จแล้ว (ในรอบ 5 ปีที่ผ่านมา) :

1. K. Amphawan, A. Surarerks, "Pushing Regularity Constraint on High Utility Itemsets Mining", The 2015 International Conference on Advanced Informatics: Concepts, Theory And Application (ICAICIT2015) [Best paper award]
2. S. Chompaisal, K. Amphawan, A. Surarerks, "Mining N-most Interesting Multi-level Frequent Itemsets without Support Threshold", Proceedings of Recent Advances in Information and Communication Technology, 2014.
3. K. Amphawan and A. Surarerks, "An efficient method for constructing dictionary based on decompounding words technique", The 17th International Annual Symposium on Computational Science and Engineering, 2013.
4. A. Jitpattanakul, and A. Surarerks, "The study of learnability of the class of k-acceptable languages on Gold's learning model". Chiang Mai Journal of Science, Vol. 40(2) 2013, pp. 248-260.
5. K. Amphawan, P. Lenca, and A. Surarerks, "Mining top-k regular-frequent itemsets using database partitioning and support estimation", Expert Systems with Applications, Volume 39, Issue 2, February 1, Pages 1924-1936, 2012.
6. K. Amphawan K, P. Lenca , and A. Surarerks, "Efficient mining top-k regular-frequent itemset using compressed tidsets", Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7104 LNAI , pp. 124-135, 2011.
7. A. Jitpattanakul, and A., Surarerks, "Characteristic Sets for Learning k-Acceptable Languages ". ECTI Transactions on

Computer and Information Technology, Vol.5, No.1 May, 2011, pp 38-44.

งานวิจัยที่กำลังทำ :

1. การค้นหาความเปลี่ยนแปลงของรูปแบบภายใต้ความถี่และความสม่ำเสมอของการปรากฏเพื่อค้นหาแนวโน้มและความเปลี่ยนแปลงของพฤติกรรมผู้บริโภค (ทุนอุดหนุนการวิจัย สำนักงานคณะกรรมการวิจัยแห่งชาติ ประจำปีงบประมาณ 2558)

3. ผู้ร่วมวิจัยคนที่ 2

1. ชื่อ - นามสกุล (ภาษาไทย) นายอนุชิต จิตพัฒนกุล
   ชื่อ - นามสกุล (ภาษาอังกฤษ) Mr. Anuchit Jitpattanakul
2. ตำแหน่งปัจจุบัน ผู้ช่วยศาสตราจารย์และอาจารย์ประจำภาควิชาคณิตศาสตร์ประยุกต์ คณะวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
3. หน่วยงานและสถานที่อยู่ที่ติดต่อได้สะดวก พร้อมหมายเลขโทรศัพท์ โทรสาร และไปรษณีย์อิเล็กทรอนิกส์ (e-mail)

   ภาควิชาคณิตศาสตร์ประยุกต์ คณะวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ 1518 ถ. ประชาราษฎร์ เขตบางซื่อ กรุงเทพ 10800 หมายเลขโทรศัพท์ 02-587-8258, 089-773-9027 หมายเลขโทรสาร 02-587-8258 ไปรษณีย์อิเล็กทรอนิกส์ athasit.s@chula.ac.th
4. ประวัติการศึกษา

   ปริญญาตรี วท.บ.(คณิตศาสตร์ประยุกต์)
   ปริญญาโท วท.ม.(วิทยาการคณนา)
   ปริญญาเอก วศ.ด.(วิศวกรรมคอมพิวเตอร์)
5. สาขาวิชาการที่มีความชำนาญพิเศษ (แตกต่างจากวุฒิการศึกษา) ระบุสาขาวิชาการ

   ปัญญาประดิษฐ์ การทำเหมืองข้อมูล การวิเคราะห์ข้อมูล ทฤษฎีการคำนวณ ระบบอัจฉริยะ ทฤษฎีการคณนา ทฤษฎีภาษาธรรมชาติ
6. ประสบการณ์ที่เกี่ยวข้องกับการบริหารงานวิจัยทั้งภายในและภายนอกประเทศ โดยระบุสถานภาพในการทำการวิจัยว่าเป็นผู้อำนวยการแผนงานวิจัย หัวหน้าโครงการวิจัย หรือผู้ร่วมวิจัยในแต่ละผลงานวิจัย

   หัวหน้าโครงการวิจัย (ในรอบ 5 ปีที่ผ่านมา) : -

   1. คุณสมบัติความสามารถการตัดสินใจของระดับภาษารูปนัยที่รู้จำได้โดยออโตมาตาสถานะจำกัดขอบเค (ทุนงบประมาณแผ่นดิน ประจำปี 2558)

2. การเรียนรู้ออโตมาตาจำกัดเชิงสลับจากข้อซักถามความเป็นสมาชิกและข้อซักถามความสมมูล (ทุนสนับสุนนนักวิจัยทั่วไป มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ประจำปี 2558)
3. ออโตมาตาจำกัดสำหรับการแยกแยะสัญญาณคลื่นไฟฟ้าหัวใจโดยวิธีการอนุมานเชิงไวยากรณ์ (ทุนวิจัยคณะวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ปี 2558)
4. การเรียนรู้ออโตมาตาสถานะจำกัดเชิงกำหนดขอบเคในกรอบการเรียนรู้แบบโต้ตอบได้ (ทุนสนับสุนนนักวิจัยทั่วไป มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ประจำปี 2557)
5. คุณสมบัติปิดของระดับภาษาที่รู้จำได้โดยออโตมาตาสถานะจำกัดขอบเค (ทุนวิจัยคระวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ปี 2557)
6. คุณสมบัติปิดของระดับภาษายอมรับได้-เคแบบเข้ม (ทุนสนับสนุนนักวิจัยรุ่นใหม่ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ประจำปี 2556)
7. ความสามารถในการเรียนรู้ของระดับภาษายอมรับได้-เคแบบเข้ม (ทุนวิจัยคณะวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ปี 2556)
8. การอนุมานเชิงไวยากรณ์สำหรับการรู้จำตัวอักษรภาษาไทยที่เขียนด้วยลายมือแบบออฟไลน์ (ทุนวิจัยคณะวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ปี 2555)
9. การควบคุมแอตทิจูดแบบเหมาะสมที่สุดสำหรับยานอวกาศแบบแข็งเกร็งโดยใช้วิธีการซัคเซสซีฟ (ทุนสนับสนุนนักวิจัยรุ่นใหม่ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ประจำปี 2554)

งานวิจัยที่ทำเสร็จแล้ว (ในรอบ 5 ปีที่ผ่านมา) :

1. A. Jitpattanakul, “Closure properties of the classes of languages recognized by k-edge finite state automata”, Journal of Information and Communication Technology University of Prayao, Vol. 1(2), 2014.
2. A. Jitpattanakul and C. Pukdeboon, “Optimal Attitude Control for Rigid Spacecraft using Successive Approximation Approach”, Far East Journal of Mathematical Sciences, Vol. 74(1), 2013, pp.37-52.

3. C. Pukdeboon and A. Jitpattanakul, “Finite-Time Anti-Disturbance Inverse Optimal Attitude Tracking Control of Flexible Spacecraft”, Mathematical Problems in Engineering, Article Numbers: 967574, 2013.
4. A. Jitpattanakul, and A. Surarerks, "The study of learnability of the class of k-acceptable languages on Gold’s learning model". Chiang Mai Journal of Science, Vol. 40(2) 2013, pp. 248-260.
5. A. Jitpattanakul, "Learnability of the class of strictly k-acceptable languages". Far East Journal of Mathematical Sciences, Vol. 71(1), 2012, pp. 169-184.
6. A. Jitpattanakul, and A., Surarerks, "Characteristic Sets for Learning k-Acceptable Languages ". ECTI Transactions on Computer and Information Technology, Vol.5, No.1 May, 2011, pp 38-44.

งานวิจัยที่กำลังทำ :

1. การค้นหาความเปลี่ยนแปลงของรูปแบบภายใต้ความถี่และความสม่ำเสมอของการปรากฏเพื่อค้นหาแนวโน้มและความเปลี่ยนแปลงของพฤติกรรมผู้บริโภค (ทุนอุดหนุนการวิจัย สำนักงานคณะกรรมการวิจัยแห่งชาติ ประจำปีงบประมาณ 2558)

รายงานสรุปการเงิน

เลขที่โครงการระบบบริหารงานวิจัย สัญญาเลขที่ ...19/2558...

โครงการวิจัยประเภทงบประมาณรายได้จากเงินอุดหนุนรัฐบาล (งบประมาณแผ่นดิน)

ประจำปีงบประมาณ พ.ศ. ...2558...

ชื่อโครงการ ...การค้นหารูปแบบที่มีค่าคุณประโยชน์สูงที่ปรากฏอย่างสม่ำเสมอเพื่อการวิเคราะห์พฤติกรรม.....
ผู้บริโภค.....

ชื่อหัวหน้าโครงการวิจัยผู้รับทุน ...ผศ. ดร. โกเมศ อัมพวัน.....

รายงานในช่วงตั้งแต่วันที่ ...16 ตุลาคม พ.ศ. 2557... ถึงวันที่ ...18 สิงหาคม พ.ศ. 2559...

ระยะเวลาดำเนินการ ...1... ปี ...10... เดือน ตั้งแต่วันที่ ...16 ตุลาคม พ.ศ. 2557...

จำนวนเงินที่ได้รับ

งวดที่ 1 (50%) ...166,050... บาท เมื่อวันที่ ...1 ธันวาคม พ.ศ. 2557...

งวดที่ 2 (40%) ...132,840... บาท เมื่อวันที่ ...9 มิถุนายน พ.ศ. 2558...

งวดที่ 3 (40%) ...33,210..... บาท

รวม ...332,100... บาท

รายการ	งบประมาณที่ตั้งไว้	งบประมาณที่ใช้จริง	จำนวนเงินคงเหลือ/เกิน
1. ค่าตอบแทน	88,000	136,000	-48,000
2. ค่าจ้าง	0	0	0
3. ค่าวัสดุ	60,000	72,193	-12,193
4. ค่าใช้สอย	104,100	35,163	68,937
5. ค่าครุภัณฑ์	80,000	89,895	-9,895
6. ค่าใช้จ่ายอื่นๆ	0	0	0
รวม	332,100	333,251	-1,151

(.....)

ลงนามหัวหน้าโครงงานวิจัยผู้รับทุน

บทสรุปผู้บริหาร
(Executive Summary)

ข้าพเจ้า ผศ. ดร. โกเมศ อัมพวัน ได้รับทุนอุดหนุนโครงการวิจัยจากมหาวิทยาลัยบูรพา ประเภทงบประมาณเงินรายได้ จากเงินอุดหนุนรัฐบาล (งบประมาณแผ่นดิน) มหาวิทยาลัยบูรพา โครงการวิจัยเรื่อง (ภาษาไทย) การค้นหารูปแบบที่มีค่าคุณประโยชน์สูงที่ปรากฏอย่างสม่ำเสมอเพื่อการวิเคราะห์พฤติกรรมผู้บริโภค (ภาษาอังกฤษ)Mining high utility patterns with regular occurrence for customers' behavior analysis รหัสโครงการ 172596 / สัญญาที่ 19/2558 ได้รับงบประมาณรวมทั้งสิ้น 369,000 บาท (สามแสนหกหมื่นเก้าพันบาทถ้วน) ระยะเวลาเดินการ 1 ปี (ระหว่างวันที่ 1 ตุลาคม 2557 ถึง 30 กันยายน 2558)

บทคัดย่อ

การค้นหารูปแบบที่มีค่าคุณประโยชน์สูงเป็นหัวข้องานวิจัยหนึ่งภายใต้การทำเมืองข้อมูลที่น่าสนใจ การค้นหารูปแบบดังกล่าวสามารถประยุกต์ใช้ในแอพพลิเคชันต่าง ๆ อย่างแพร่หลาย ตัวอย่างเช่น การประยุกต์ใช้ในธุรกิจค้าปลีกเพื่อทำการค้นหาเซตของสินค้าที่ถูกซื้อจากลูกค้า โดยเซตของสินค้าดังกล่าวจะเป็นรายการสินค้าต่าง ๆ ที่ถูกซื้อร่วมกันที่จะให้ผลกำไรสูงหรือต้นทุนที่ต่ำเป็นต้น แต่อย่างไรก็ตาม การค้นหารูปแบบที่มีค่าคุณประโยชน์สูงจะทำการพิจารณาเพียงแค่ค่าคุณประโยชน์ของรายการต่าง ๆ เท่านั้นที่ซึ่งการดำเนินการดังกล่าวอาจไม่เพียงพอต่อการสังเกตุ/วิเคราะห์พฤติกรรมการซื้อสินค้าของผู้บริโภค ด้วยเหตุนี้ งานวิจัยนี้จึงมุ่งเน้นที่จะทำการเพิ่มเติมเงื่อนไขการพิจารณารูปแบบโดยจะทำการเพิ่มเติมเงื่อนไขของการปรากฏอย่างสม่ำเสมอร่วมกับการพิจารณาค่าคุณประโยชน์ของรายการต่าง ๆ ภายใต้แนวคิดข้างต้น รูปแบบที่น่าสนใจจะเป็นรูปแบบที่มีค่าคุณประโยชน์สูงและปรากฏขึ้นในชุดข้อมูลอย่างสม่ำเสมอ

ในการค้นหารูปแบบใหม่ที่นำเสนอ ผู้วิจัยได้เสนอขั้นตอนวิธีที่มีประสิทธิภาพที่ชื่อว่า HURI-UL ที่ซึ่งจะทำการอ่านข้อมูลจากฐานข้อมูลเพียงครั้งเดียว และทำการประยุกต์ใช้แนวความคิดเกี่ยวกับค่าคุณประโยชน์ที่หลงเหลือและค่าคุณประโยชน์แบบประมาณเพื่อช่วยลดทอนปริภูมิสถานะของการค้นหารูปแบบ นอกจากนั้นยังประยุกต์ใช้โครงสร้างลิสต์คุณประโยชน์เพื่อใช้ในการจัดเก็บค่าคุณประโยชน์และข้อมูลการปรากฏขึ้นของรูปแบบหนึ่ง ๆ โดยในการทดสอบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอ เราจะสังเกตุได้ว่าขั้นตอนวิธีที่นำเสนอสามารถค้นหารูปแบบที่มีค่าคุณประโยชน์สูงและปรากฏอย่างสม่ำเสมอได้อย่างมีประสิทธิภาพ

Output/outcome

1. ได้แนวทางในการวิเคราะห์รูปแบบที่มีประโยชน์สูงและปรากฏขึ้นอย่างสม่ำเสมอ ที่สามารถนำไปวิเคราะห์หาสาเหตุของการเกิดขึ้นของรูปแบบ ที่ซึ่งจะช่วยให้ผู้บริหารกิจการ บริษัท หรือเจ้าของธุรกิจ จะสามารถทำการปรับเปลี่ยนวิธีหรือกลยุทธ์ในการดำเนินธุรกิจเพื่อให้ธุรกิจที่ทำอยู่สามารถดำเนินไปได้ด้วยดี
2. ได้ขั้นตอนวิธีต้นแบบในการค้นหารูปแบบที่มีประโยชน์สูงและปรากฏขึ้นอย่างสม่ำเสมอ
3. สามารถนำขั้นตอนวิธีข้างต้นไปพัฒนาระบบซอฟต์แวร์ เพื่อใช้ในการวิเคราะห์พฤติกรรมของลูกค้าหรือผู้บริโภค ที่จะทำให้กิจการ บริษัท ห้างร้านต่าง ๆ สามารถปรับตัวตามพฤติกรรมของผู้บริโภคได้
4. ได้ขั้นตอนวิธีที่นำเสนอสามารถถูกใช้เป็นต้นแบบในการศึกษาและวิธีขั้นสูงต่อไป

ข้อเสนอแนะ

ผลงานวิจัยนี้สามารถนำไปใช้ได้กับภาคธุรกิจที่หลากหลาย แต่ผู้ที่จะประยุกต์ใช้จะต้องมีความรู้เบื้องต้นเกี่ยวกับข้อมูล และการกำหนดค่าพารามิเตอร์ให้มีความเหมาะสม