การสร้างกฎความสัมพันธ์จากเซตรายการความถี่และเซตรายการความถี่แบบปิด

Generating Association Rules from Frequent Itemsets and Frequent Closed Itemsets

พนิดา ทรงรัมย์ (Panida Songram)*

บทคัดย่อ

การสืบค้นกฎความสัมพันธ์ (Association Rule Mining) เป็นการค้นหาความสัมพันธ์ระหว่างเซตรายการ (Itemsets) ในฐานข้อมูลขนาดใหญ่ ซึ่งกฎความสัมพันธ์สามารถสร้างขึ้นจากเซตรายการความถี่ (Frequent Itemsets) และเซตรายการความถี่แบบปิด (Frequent Closed Itemsets) ในปัจจุบันมีการประยุกต์ใช้กฎความสัมพันธ์ในงานด้านต่างๆ อย่างแพร่หลาย บทความนี้จึงนำเสนอการสร้างกฎความสัมพันธ์จากเซตรายการความถี่และการสร้างกฎความสัมพันธ์จากเซตรายการความถี่แบบปิด และชี้ให้เห็นว่าการสร้างกฎความสัมพันธ์จากเซตรายการความถี่แบบปิดเป็นวิธีการที่ดีกว่าการสร้างกฎความสัมพันธ์จากเซตรายการความถี่

คำสำคัญ: กฎความสัมพันธ์ เซตรายการความถี่แบบปิด เหมืองข้อมูล

Abstract

Association rule mining is proposed to determine relationships between itemsets in very large databases. Association rules can be generated from two different itemsets, frequent itemsets and frequent closed itemsets. Since the association rule mining is extensively exploited in many applications, this article is proposed to present the generation of association rules from frequent itemsets and frequent closed itemsets. Moreover, it shows that the association rules should be generated from closed itemsets.

Keywords: Association rule, Closed Itemsets, Data Mining.

1. บทนำ

กฎความสัมพันธ์ถูกนำเสนอขึ้นครั้งแรกโดย Agrawal และคณะ [17], [18] เพื่อหาความสัมพันธ์ของข้อมูลจากฐานข้อมูลขนาดใหญ่ กฎความสัมพันธ์ถูกนำไปใช้อย่างแพร่หลายในด้านต่างๆ เช่น การวิเคราะห์พฤติกรรมของลูกค้า การทำนายรูปแบบการเข้าใช้เว็บไซต์ การทดลองทางวิทยาศาสตร์ การรักษาโรค และการทำนายปรากฏการณ์ธรรมชาติ เป็นต้น ขั้นตอนการหากฎความสัมพันธ์แบ่งออกเป็น 2 ขั้นตอนใหญ่ [15] ขั้นตอนแรก คือ การสืบค้นเซตรายการความถี่ (Frequent Itemset Mining) ซึ่งเป็นการสืบค้นเซตรายการที่มีความถี่หรือค่าสนับสนุน (Support) มากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ (Minimum Support) จากนั้นเซตรายการความถี่จะถูกนำไปสร้างกฎความสัมพันธ์ในขั้นตอนที่สอง ซึ่งจะเห็นได้ว่ากฎความสัมพันธ์ที่ได้จะขึ้นอยู่กับเซตรายการความถี่ ถ้าเซตรายการความถี่มีจำนวนมาก กฎความสัมพันธ์ที่ได้ก็จะมีจำนวนมากตาม ดังนั้นจึงได้มีการเสนอการสืบค้นเซตรายการความถี่แบบปิด (Frequent Closed Itemset Mining) เพื่อใช้ในการลดจำนวนเซตรายการความถี่โดยที่ข้อมูลไม่สูญหาย การสืบค้นเซตรายการความถี่แบบปิดจะสร้างเฉพาะเซตรายการความถี่ที่ไม่มีซุปเปอร์เซตที่มีค่าสนับสนุนเท่ากัน

บทความนี้นำเสนอการสร้างกฎความสัมพันธ์จากเซตรายการความถี่ และการสร้างกฎความสัมพันธ์จากเซตรายการความถี่แบบปิด เพื่อให้เข้าใจและเห็นถึงข้อได้เปรียบของการสร้างกฎความสัมพันธ์จากเซตรายการความถี่แบบปิด และสามารถเลือกใช้วิธีการสร้างกฎความสัมพันธ์เพื่อนำไปใช้ประโยชน์ต่อไปได้

2. เซตรายการความถี่กับเซตรายการความถี่แบบปิด

กำหนดให้ เป็นเซตของทรานแซคชัน (Transaction) ในฐานข้อมูล และ คือรายการ (Item) ทั้งหมดที่ปรากฏใน

* คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม

ฐานข้อมูล และ แทนความสัมพันธ์แบบไบนารี่ (Binary Relation) ระหว่าง ทรานแซคชัน $d \in D$ กับรายการ $i \in I$ ถ้าทรานแซคชัน มีรายการ สามารถเขียนแทนได้ด้วย $(d, i) \in R$

เซตรายการ (Itemset) $X = \{i_1, i_2, ..., i_l\}$ คือ เซตของรายการที่มีความยาวหรือจำนวนรายการเท่ากับ l แทนด้วย $|X| = l$ ซึ่ง $X \subseteq l$ และค่าสนับสนุนของเซตรายการ X คือจำนวนของทรานแซคชันใน D ที่บรรจุเซตรายการ X เขียนแทนด้วย $Supp(X)$ และเซตรายการ X เป็นเซตรายการความถี่ต่อเมื่อ $Supp(X) \geq \min_supp$ โดยที่ $\min_supp$ คือ ค่าสนับสนุนขั้นต่ำที่ถูกกำหนดขึ้นโดยผู้เชี่ยวชาญด้านข้อมูลนั้นๆ เซตรายการความถี่ทั้งหมดสามารถเขียนแทนด้วย $FI = \{X \mid Supp(X) \geq \min_supp\}$

ตารางที่ 1 เป็นฐานข้อมูลตัวอย่างที่ประกอบไปด้วยรายการทั้งหมด 4 รายการ คือ a, b, c, d และทรานแซคชันมีทั้งหมด 5 ทรานแซคชัน

ตารางที่ 1 *ฐานข้อมูลตัวอย่าง*

ลำดับ	รายการ
1.	a, b, c, d
2.	a, c
3.	c, d
4.	a, b, d
5.	a, b, c, d

จากตารางที่ 1 จะเห็นได้ว่าเซตรายการ *ab* มีความยาวเท่ากับ 2 และค่าสนับสนุนเท่ากับ 3 เนื่องมาจาก *ab* ปรากฏอยู่ใน 3 ทรานแซคชัน คือ 1, 4 และ 5 ถ้ากำหนดค่าสนับสนุนขั้นต่ำให้มีค่าเท่ากับ 3 แสดงว่า *ab* เป็นเซตรายการความถี่ เพราะ $Supp(ab) > \min_supp$ และจากตารางที่ 1 สามารถสืบค้นเซตรายการความถี่ได้ทั้งหมด 9 เซตรายการ ดังแสดงในตารางที่ 2

ตารางที่ 2 *เซตรายการความถี่ทั้งหมดเมื่อค่าสนับสนุนขั้นต่ำเท่ากับ 3*

ค่าสนับสนุน	เซตรายการความถี่
4	a, c, d
3	b, ab, ac, ad, bd, abd

เนื่องจากการสืบค้นเซตรายการความถี่อาจจะนำไปสู่การสร้างเซตรายการความถี่จำนวนมหาศาล [9], [10] ซึ่งเซตรายการความถี่จำนวนมหาศาลจะนำไปสู่การสร้างกฎจำนวนมหาศาลเช่นกัน ทำให้ประสิทธิภาพและประสิทธิผลลดลงเพราะจะต้องค้นหากฎที่จะนำไปใช้ประโยชน์จากกฎที่มีจำนวนมหาศาล [9], [10] เพื่อแก้ปัญหาดังกล่าวได้มีการนำเสนอการสืบค้นเซตรายการแบบปิดเพื่อลดจำนวนเซตรายการที่ได้จากการสืบค้นเซตรายการความถี่

การสืบค้นเซตรายการความถี่แบบปิดเป็นการสืบค้นเซตรายการความถี่บนพื้นฐานกาลัวคอนเน็ตชั่น (Galois Connection) เพื่อขจัดเซตรายการที่ซ้ำซ้อนและลดจำนวนการสร้างเซตรายการโดยที่ข้อมูลไม่สูญหาย [21] โดยจะทำการสืบค้นเฉพาะเซตรายการความถี่ที่ไม่มีซุปเปอร์เซตที่มีค่าสนับสนุนเท่ากัน มีหลายอัลกอริธึมที่ถูกนำเสนอเพื่อสืบค้นเซตรายการความถี่แบบเปิด เช่น CLOSET [9], CLOSET+ [10], FPCLOSE [6], TFP [8], [11], CLOSE [15], A-CLOSE [16], CHARM [14], DCI-CLOSED [3], [4], LCM [19], [20] และTopKMiner [2] เป็นต้น

การสืบค้นเซตรายการแบบปิดมีพื้นฐานมาจาก $\varsigma = fog$ ฟังก์ชัน เรียกว่ากาลัวโอเปอเรเตอร์ (Galois Operator) หรือโคลส์เซอร์โอเปอเรเตอร์ (Closure Operator) โดยที่ฟังก์ชัน f และ g แสดงได้ดังสมการที่ 1 และ 2 [3], [4]

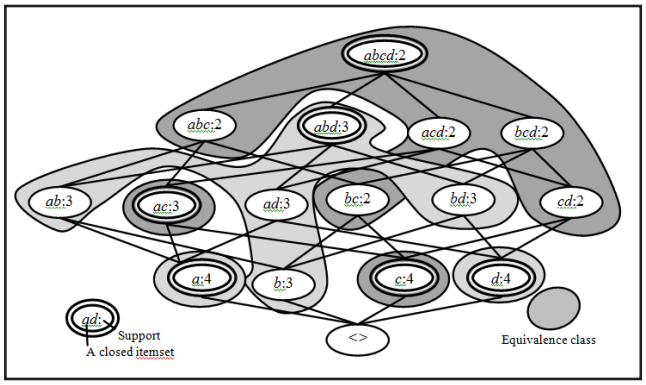
$$f(T) = \{i \in I \mid \forall d \in T, (d, i) \in R\} \quad (1)$$

$$g(X) = \{d \in D \mid \forall i \in X, (d, i) \in R\} \quad (2)$$

ฟังก์ชัน f คืนค่าเซตรายการที่มีขนาดยาวที่สุดที่อยู่ในทุกทรานแซคชันที่อยู่ใน T และฟังก์ชัน g คืนค่าเซตของทรานแซคชันที่มีเซตรายการ X โดยที่ $T \subseteq D$ และ $X \subseteq I$ เซตรายการ X ถูกเรียกว่าเซตรายการแบบปิดก็ต่อเมื่อ $\varsigma(X) = f(g(X)) = fog(X) = X$ ยกตัวอย่างเช่น จากตารางที่ 1 พิจารณา $\varsigma(ac) = f(g(ac)) = f(\{1, 2, 5\}) = ac$ แสดงว่า *ac* เป็นเซตรายการแบบปิด ในขณะที่ $\varsigma(bc) = f(g(bc)) = f(\{1, 5\}) = abcd \neq bc$ แสดงว่า *bc* ไม่ใช่เซตรายการแบบปิด

โคลส์เซอร์โอเปอเรเตอร์เป็นตัวกำหนดคลาสสมมูล (Equivalence Class) บนแลตทิซของเซตรายการ (Lattices of Itemset) ดังแสดงในรูปที่ 1 ซึ่งแสดงแลตทิซของเซตรายการในตารางที่ 1 เซตรายการอยู่ในคลาสเดียวกันก็ต่อเมื่อผลลัพธ์ที่ได้จากโคลส์เซอร์โอเปอเรเตอร์เป็นตัวเดียวกัน ยกตัวอย่าง เช่น *abc* และ *cd* อยู่ในคลาสสมมูลเดียวกันเพราะ $\varsigma(abc) = \varsigma(cd) = abcd$ ดังนั้นเซตรายการที่ได้ผลลัพธ์จาก

โคลส์เซอร์โอเปอเรเตอร์เท่ากันจะถูกจัดอยู่ในคลาสเดียวกัน ซึ่งแต่ละคลาสประกอบไปด้วยเซตรายการที่มีค่าสนับสนุนเดียวกัน เซตรายการที่ยาวที่สุดจะคลอบคลุมทุกเซตรายการที่อยู่ในคลาสเดียวกัน ซึ่งเซตรายการที่ยาวที่สุดก็คือเซตรายการแบบปิดนั้นเอง ถ้าเซตรายการแบบปิดมีค่าสนับสนุนมากกว่าค่าสนับสนุนขั้นต่ำ จะถูกเรียกว่า เซตรายการความถี่แบบปิด เซตรายการความถี่แบบปิดทั้งหมดแทนด้วย $FC = \{X \mid \varsigma(X) = X \wedge supp(X) \geq \min_supp\}$ ส่วนเซตรายการที่สั้นที่สุดในแต่ละคลาสสมมูลเรียกว่าตัวสร้าง (Generator) [3], [4] ซึ่งอัลกอริธึมส่วนใหญ่สร้างเซตรายการความถี่แบบปิดจากตัวสร้าง ตัวสร้างของเซตรายการความถี่แบบปิดทั้งหมดแทนด้วย FG เซตรายการความถี่ แบบปิดและตัวสร้างของเซตรายการความถี่แบบปิดสามารถแสดงได้ดังตารางที่ 3



ภาพที่ 1 *แลตทิซของเซตรายการ*

ตารางที่ 3 *เซตรายการความถี่แบบปิดและตัวสร้างเมื่อค่าสนับสนุนขั้นต่ำเท่ากับ 3*

ค่าสนับสนุน	ตัวสร้าง	เซตรายการความถี่แบบปิด
4	a c d	a c d
3	ac b	ac abd

จากตารางที่ 2 และ 3 จะเห็นได้ว่าเซตรายการความถี่แบบปิดมีจำนวน 5 เซตรายการ ซึ่งน้อยกว่าเซตรายการความถี่เมื่อสืบค้นจากฐานข้อมูลเดียวกันในตารางที่ 1 ดังนั้นการสร้างกฎความสัมพันธ์จากเซตรายการความถี่แบบปิดจะให้จำนวนกฎน้อยกว่าการสร้างกฎความสัมพันธ์จากเซตรายการความถี่

3. การสร้างกฎความสัมพันธ์จากเซตรายการความถี่

การสร้างกฎความสัมพันธ์จากเซตรายการความถี่สามารถสรุปเป็นขั้นตอนดังต่อไปนี้

ขั้นตอนที่ 1 กฎความสัมพันธ์ถูกสร้างจากเซตรายการความถี่ทั้งหมดที่มีความยาว 2 ขึ้นไป ซึ่งสามารถแทนด้วย $FR = \{X \mid \in FI \wedge |X| \geq 2\}$ จากตารางที่ 2 $FR = \{ab, ac, ad, bd, abd\}$

ขั้นตอนที่ 2 พิจารณาเซตรายการความถี่ที่เป็นซับเซตของเซตรายการความถี่ X ทั้งหมด ซึ่งสามารถแทนด้วย $FI_X = \{Y \mid Y \in FI \wedge |Y \subset X\}$ ตัวอย่างเช่น ต้องการหากฎความสัมพันธ์จากเซตรายการความถี่ abd ดังนั้นเซตรายการความถี่ที่เป็นซับเซตของ abd คือ $FI_{abd} = \{a, d, b, ab, ad, bd\}$

ขั้นตอนที่ 3 เซตรายการความถี่ X สามารถสร้างกฎความสัมพันธ์ทั้งหมดได้จาก $r : Y \rightarrow (X - Y)$ สำหรับ $\forall Y \in FI_X$ และ $Y \neq \varnothing$ [12], [13], [21], [22] โดยที่ Y คือ เหตุการณ์ที่เกิดก่อน (Antecedent) และ $(X - Y)$ คือ เหตุการณ์ที่เกิดขึ้นตามหลัง (Consequent) ตัวอย่างเช่น เซตรายการความถี่ abd สามารถสร้างกฎความสัมพันธ์ได้ทั้งหมด 6 กฎ คือ $a \rightarrow bd$, $d \rightarrow ab$, $b \rightarrow ad$, $ab \rightarrow d$, $ad \rightarrow b$, $bd \rightarrow a$ จะเห็นได้ว่า 1 เซตรายการความถี่ สามารถสร้างกฎความสัมพันธ์ได้หลายกฎ เซตรายการที่มีความยาว l สามารถสร้างกฎทั้งหมด 2^l-2 กฎ [12], [13]

ขั้นตอนที่ 4 หาค่าความเชื่อมั่นของกฎความสัมพันธ์ r ซึ่งหาได้จาก $conf(r) = supp(X) / supp(Y)$ ยกตัวอย่างเช่น ค่าความเชื่อมั่นของกฎ $a \rightarrow bd$ มีค่าเท่ากับ $supp(abd) / supp(a) = 3/4$

เมื่อได้กฎความสัมพันธ์และค่าความเชื่อมั่นแล้ว กฎความสัมพันธ์จะถูกนำมาพิจารณาเพื่อนำไปใช้ โดยกฎที่ยอมรับได้ที่จะนำไปใช้ คือ กฎความสัมพันธ์ที่มีค่าความเชื่อมั่นมากกว่าหรือเท่ากับค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence) ซึ่งถูกกำหนดโดยผู้เชี่ยวชาญด้านข้อมูล ถ้าค่าความเชื่อมั่นขั้นต่ำมีค่าเท่ากับ 75% กฎความสัมพันธ์ทั้งหมดที่สร้างจากเซตรายการความถี่ในตารางที่ 2 แสดงได้ในตารางที่ 4

ตารางที่ 4 *กฎความสัมพันธ์ที่สร้างจากเซตรายการความถี่*

เซตรายการความถี่	กฎความสัมพันธ์	ค่าความเชื่อมั่น	ค่าความเชื่อมั่น (%)
ab	$a \rightarrow b$	3/4	75%
	$b \rightarrow a$	3/3	100%
ac	$a \rightarrow c$	3/4	75%
	$c \rightarrow a$	3/3	75%
ad	$a \rightarrow d$	3/4	75%
	$d \rightarrow a$	3/4	75%
bd	$b \rightarrow d$	3/3	100%
	$d \rightarrow b$	3/4	75%
abd	$a \rightarrow bd$	3/4	75%
	$b \rightarrow ad$	3/3	100%
	$d \rightarrow ab$	3/4	75%
	$ab \rightarrow d$	3/3	100%
	$ad \rightarrow b$	3/3	100%
	$bd \rightarrow a$	3/3	100%

4. การสร้างกฎความสัมพันธ์จากเซตรายการความถี่แบบปิด

การสร้างกฎความสัมพันธ์จากเซตรายการความถี่แบบปิด พยายามหากฎความสัมพันธ์ที่ไม่ซ้ำซ้อน เป็นกฎที่มีประโยชน์และสำคัญที่สุด [22] ซึ่ง Bastide และคณะได้เสนอว่ากฎความสัมพันธ์ที่ไม่ซ้ำซ้อน คือ กฎความสัมพันธ์ $r : X_1 \rightarrow X_2$ ที่ไม่มีกฎความสัมพันธ์ $r' : X'_1 \rightarrow X'_2$ ซึ่ง $supp(r) = supp(r')$, $conf(r)$, $conf(r')$, $X'_1 \subseteq X_1$ และ $X_2 \subseteq X'_2$ โดยที่ $supp(r) = supp(X_1 \cup X_2)$ [12], [22] ยกตัวอย่างเช่น $ab \rightarrow d$ (3/3) เป็นกฎความสัมพันธ์ที่ซ้ำซ้อนเนื่องมาจากมีกฎความสัมพันธ์ $b \; r' \; X' \; ad$ (3/3) ซึ่ง $supp(abd) = supp(bad)$, $conf(ab \rightarrow d) = conf(b \rightarrow ad)$, $b \subseteq ab$ และ $d \subseteq ad$ ส่วนกฎความสัมพันธ์ $b \rightarrow ad$ ถือว่าเป็นกฎความสัมพันธ์ที่ไม่ซ้ำซ้อน จะเห็นได้ว่ากฎความสัมพันธ์ที่ไม่ซ้ำซ้อนที่สร้างจากเซตรายการ X คือ กฎความสัมพันธ์ที่มีซับเซตรายการของ X ที่สั้นที่สุดอยู่ฝั่งซ้ายและตามด้วยซับเซตรายการของ X ที่ยาวที่สุดฝั่งขวา [22]

Batide และคณะได้นำเสนอ 2 วิธีการเพื่อได้กฎดังกล่าว คือ วิธีการที่เรียกว่า Generic Basis ซึ่งวิธีการที่ใช้ในการสร้างกฎความสัมพันธ์ที่มีค่าความเชื่อมั่น 100% (Exact Association Rule) และวิธีการ Informative Basis ซึ่งเป็นวิธีการที่ใช้ในการสร้างกฎความสัมพันธ์ที่มีความเชื่อมั่นน้อยกว่า 100% (Approximate Association Rule) ทั้งสองวิธีการสร้างกฎความสัมพันธ์จากเซตรายการความถี่แบบปิดและตัวสร้าง (Generator) ขั้นตอนในการสร้างกฎความสัมพันธ์จะคล้ายกับการสร้างกฎความสัมพันธ์จากเซตรายการความถี่ ซึ่งสามารถสรุปได้ดังนี้

ขั้นตอนที่ 1 พิจารณาเซตรายการความถี่แบบปิดทั้งหมดที่มีความยาว 2 ขึ้นไป ซึ่งสามารถเขียนแทนด้วย

$$FCR = \{X \mid X \in FC \; \wedge |X| \geq 2\}$$

ขั้นตอนที่ 2 พิจารณาตัวสร้างของ X เพื่อสร้างกฎความสัมพันธ์ ซึ่งแบ่งออกเป็น 2 วิธี คือ

วิธีการ *Generic Basis* เป็นการสร้างกฎความสัมพันธ์จากตัวสร้างที่มีโคลส์เซอร์เท่ากับ X แทนด้วย $GE_X = \{Y \mid Y \in FG \; \wedge \varsigma(Y) = X)\}$ ตัวอย่างเช่น $GE_{abd} = \{b\}$ เนื่องมาจาก $\varsigma(Y) = abd$ จากนั้นสร้างกฎความสัมพันธ์จาก $Y \rightarrow (X - Y)$ สำหรับ $\forall Y \in GE_X$ และ $Y \neq \varnothing$ [12], [13, [22] จากวิธีการนี้เซตรายการความถี่แบบปิด abd สามารถสร้างกฎความสมพันธ์ได้ 1 กฎ คือ $b \rightarrow ad$ ซึ่งเป็นกฎความสัมพันธ์ที่มีความเชื่อมั่นเท่ากับ 100% เนื่องมาจาก $supp(b) = supp(\varsigma(b))$ = $supp(abd)$ ดังนั้น $conf(b \rightarrow ad) = supp(abd) / supp(b) = 1$ กฎที่สร้างจาก GE ทุกกฎจะได้ค่าความเชื่อมั่น 100% เนื่องมาจาก $supp(Y) = supp(\varsigma(Y))$ [12], [13]

วิธีการ Informative Basis พิจารณาตัวสร้างที่มีโคลส์เซอร์เป็นซับเซตของ X ตัวสร้างดังกล่าวสามารถแทนด้วย $GA_X = \{Y \mid Y \in FG \; \wedge \varsigma(Y) \subset X)\}$ จากนั้นสร้างกฎความสัมพันธ์ $Y \rightarrow (X - Y)$ สำหรับ $\forall Y \in GA_X$ และ $Y \neq \varnothing$ และหาค่าความเชื่อมั่นได้จาก $conf(r) = supp(X) / supp(Y)$ ตัวอย่างเช่น พิจารณาเซตรายการความถี่แบบปิด ab ตัวสร้าง $GA_{ac} = \{a, c\}$ เนื่องมาจาก $\varsigma(a) = a \subset ac$ และ $\varsigma(c) = c \subset ac$ ดังนั้นเซตรายการ ac สร้างกฎความสัมพันธ์ได้ 2 กฎ คือ $a \rightarrow c$ (3/4) และ $c \rightarrow a$ (3/4) ซึ่งกฎความสัมพันธ์ที่สร้างด้วยวิธี Informative Basic จะมีค่าความเชื่อมั่นน้อยกว่า 100%

กฎความสัมพันธ์ทั้งหมดที่ได้จากเซตรายการแบบปิดที่มีค่าความเชื่อมั่นมากกว่าหรือเท่ากับ 75% แสดงได้ดังตารางที่ 5

ตารางที่ 5 *กฎความสัมพันธ์ที่สร้างจากเซตรายการความถี่*

เซตรายการความถี่แบบปิด	กฎความสัมพันธ์	ค่าความเชื่อมั่น	ค่าความเชื่อมั่น (%)
ac	a → b	3/4	75%
	b → a	3/3	100%
abd	a → bd	3/4	75%
	b → ad	3/3	100%
	d → ab	3/4	75%

5. การเปรียบเทียบกฎความสัมพันธ์ที่สร้างจากเซตรายการความถี่และเซตรายการความถี่แบบปิด

จากการสร้างกฎความสัมพันธ์จากเซตรายการความถี่และการสร้างกฎความสัมพันธ์จากเซตรายการความถี่แบบปิดสามารถเปรียบเทียบและสรุปให้เห็นเป็นข้อๆ ดังนี้

1. จำนวนเซตรายการความถี่แบบปิดน้อยกว่าหรือเท่ากับจำนวนเซตรายการความถี่ เมื่อสืบค้นจากฐานข้อมูลเดียวกัน เนื่องมากจากเซตรายการแบบปิดจะทำการสืบค้นเฉพาะเซตรายการความถี่ที่ไม่มีซุปเปอร์เซตที่มีค่าสนับสนุนเท่ากัน ดังนั้น $FC \subseteq FI$ แสดงว่า $|FC| \leq |FI|$ ถ้าเซตรายการในฐานข้อมูลมีความยาวมาก ผลลัพธ์ที่ได้จากการสืบค้นเซตรายการความถี่แบบปิดจะมีจำนวนน้อยกว่าเซตรายการความถี่มาก ยกตัวอย่างเช่น สมมติฐานข้อมูลมีแค่ 1 ทรานแซคชัน คือ (i_1), (i_2) ..., (i_{100}) และค่าสนับสนุนขั้นต่ำมีค่าเท่ากับ 1 ผลลัพธ์ที่ได้จากการสืบค้นเซตรายการความถี่จะได้ทั้งหมด 2^{100} -1 เซตรายการ ซึ่งก็คือ (i_1), ..., (i_{100}), (i_1, i_2), ..., (i_{99}, i_{100}), ..., $(i_1, i_2, ..., i_{100})$ แต่เซตรายการความถี่แบบปิดถูกสร้างขึ้นมาแค่ 1 เซตรายการ คือ $(i_1, i_2, ..., i_{100})$

2. เนื่องมากจากจำนวนเซตรายการความถี่มีจำนวนมากกว่าหรือเท่ากับเซตรายการความถี่แบบปิด ดังนั้นกฎความสัมพันธ์ที่สร้างจากเซตรายการความถี่ย่อมมีจำนวนมากกว่ากฎความสัมพันธ์ที่สร้างจากเซตรายการความถี่แบบปิด และนอกจากนี้กฎความสัมพันธ์ที่สร้างจากเซตรายการความถี่แบบปิดยังสร้างเฉพาะกฎที่ไม่ซ้ำซ้อน ทำให้จำนวนกฎที่ได้ยิ่งลดลง สมมติฐานข้อมูลประกอบไปด้วยเซตรายการที่มีความยาวสูงสุด l เซตรายการความถี่ 1 เซตรายการ สามารถสร้างกฎความสัมพันธ์ $2^l - 2$ และกฎความสัมพันธ์ทั้งหมดที่สร้างจากเซตรายการความถี่จะเท่า $|FI| .(2^l - 2)$ แต่สำหรับการสร้างกฎความสัมพันธ์จากเซตรายการแบบปิด 1 เซตรายการ ในกรณีที่แย่ที่สุด คือ ซับเซตของเซตรายการแบบปิดทุกตัวเป็นเซตรายการแบบปิด ดังนั้นมีตัวสร้างทั้งหมด $2^l - 2$ ตัว และสามารถสร้างกฎความสัมพันธ์สูงสุดได้ $|FC| .(2^l - 2)$ กฎ เนื่องจาก $FC \leq |FI|$ ดังนั้นกฎความสัมพันธ์ที่สร้างจากเซตรายการความถี่แบบปิดน้อยกว่าหรือเท่ากับกฎความสัมพันธ์ที่สร้างจากเซตรายการความถี่ จากตารางที่ 4 และ 5 จะเห็นว่ากฎความสัมพันธ์ที่ได้จากเซตรายการความถี่มีจำนวน 14 กฎ แต่กฎความสัมพันธ์ที่ได้จากเซตรายการความถี่มีแค่ 5 กฎแค่นั้น ซึ่งการค้นหากฎที่ต้องการนำไปจากกฎความสัมพันธ์น้อยกว่าย่อมดีกว่าการค้นหากฎความสัมพันธ์ที่ต้องการนำไปใช้จากกฎความสัมพันธ์ที่มีจำนวนมากมายมหาศาล

3. จะต้องเสียเวลาในการสร้างกฎความสัมพันธ์จากเซตรายการความถี่มากกว่าการสร้างกฎความสัมพันธ์จากเซตรายการความถี่แบบปิด สมมติใช้เวลาในการสร้างกฎความสัมพันธ์เท่ากัน คือ c และกฎความสัมพันธ์ที่สร้างจากเซตรายการความถี่ทั้งหมด $|FI| .(2^l - 2)$ ดังนั้นความซับซ้อน (Time Complexity) ในการสร้างกฎความสัมพันธ์จากเซตรายการความถี่ทั้งหมดเท่ากับ $O(c. |FI| .2^l)$ โดย ค่าความซับซ้อนของการสร้างกฎความสัมพันธ์จากเซตรายการแบบปิดเท่ากับ $O(c. |FC| .2^l)$ เนื่องจาก $FC \leq |FI|$ ดังนั้นการสร้างกฎความสัมพันธ์จากเซตรายการความถี่แบบปิดจะเสียเวลาน้อยกว่าการสร้างกฎความสัมพันธ์จากเซตรายการความถี่

4. เมื่อเซตรายการความถี่มีจำนวนมากกว่าหรือเท่ากับเซตรายการความถี่แบบปิด ดังนั้นการคำนวณหาค่าสนับสนุนและการเปรียบหาค่าสนับสนุนกับค่าสนับสนุนขั้นต่ำย่อมเสียเวลากว่าการสืบค้นเซตรายการความถี่แบบปิด

5.การสร้างกฎความสัมพันธ์จากเซตรายการความถี่จะต้องเสียเวลาในการคำนวณหาค่าความเชื่อมั่นของกฎมากกว่าการสร้างกฎความสัมพันธ์จากเซตรายการความถี่แบบปิด ซึ่งการคำนวณหาค่าความเชื่อมั่นสามารถคำนวณได้จากค่าสนับสนุน สมมติใช้วิธีเดียวกัน คือ Sequential Searchในการค้นหาค่าสนับสนุนจากเซตรายการความถี่และค้นหาค่าสนับสนุนจากเซตรายการความถี่แบบปิด ค่าความซับซ้อนที่ใช้ในการหาค่าสนับสนุนของเซตรายการความถี่ 1 เซตรายการจะเท่ากับ $O(|FI|)$ เนื่องจากกฎความสัมพันธ์สร้างจากสองซับเซตของเซตรายการความถี่ ดังนั้นความซับซ้อนในการหาค่าสนับสนุนเพื่อคำนวณค่าความเชื่อมั่นของ

1 กฎความสัมพันธ์ เท่ากับ $O(2.|FI|)$ กฎความสัมพันธ์ที่ได้จากเซตรายการความถี่มีทั้งหมด $|FI|.(2^l - 2)$ กฎ ดังนั้นความซับซ้อนในการหาค่าสนับสนุนเพื่อใช้ในการหาค่าความเชื่อมั่นทั้งหมดเท่ากับ $O(|FI|^2.2^{l+1})$ แต่สำหรับการหาค่าความเชื่อมั่นของกฎความสัมพันธ์ที่สร้างจากเซตรายการความถี่แบบปิด กฎความสัมพันธ์ 1 กฎใช้เวลา $O(2.|FI|)$ ในกรณีที่แย่ที่สุดสร้างกฎความสัมพันธ์ได้ $|FI|^2.2^l - 2)$ กฎ ดังนั้นความซับซ้อนในการหาค่าสนับสนุนเพื่อใช้ในการหาค่าความเชื่อมั่นทั้งหมดเท่ากับ $O(|FI|^2.2^{l+1})$ แต่ $|FC| \leq |FI|$ ดังนั้นการคำนวณค่าความเชื่อมั่นสำหรับกฎความสัมพันธ์ที่สร้างจากเซตรายการความถี่เสียเวลากว่าการหาค่าความเชื่อมั่นสำหรับกฎความสัมพันธ์ที่สร้างจากเซตรายการความถี่แบบปิด

6. บทสรุป

กฎความสัมพันธ์สามารถสร้างได้จากสองเซตรายการคือ เซตรายการความถี่และเซตรายการความถี่แบบปิด เนื่องจากการสืบค้นเซตรายการความถี่จะสร้างเซตรายการจำนวนมหาศาลจากฐานข้อมูลขนาดใหญ่ นำไปสู่การสร้างกฎความสัมพันธ์จำนวนมหาศาลและเป็นกฎที่ซ้ำซ้อน ทำให้เกิดความยุ่งยากในการค้นหากฎความสัมพันธ์ที่ต้องการนำไปใช้ประโยชน์ นอกจากนี้ยังต้องเสียเวลาในการหาเซตรายการความถี่และกฎความสัมพันธ์จำนวนมาก รวมทั้งจะต้องเสียเวลาคำนวณหาค่าความเชื่อมั่นแต่ละกฎเพื่อพิจารณากฎที่จะนำไปใช้ ดังนั้นการสร้างกฎความสัมพันธ์จากเซตรายการแบบปิดจึงเป็นวิธีการแก้ปัญหาวิธีหนึ่ง ที่สามารถลดจำนวนเซตรายการความถี่โดยขจัดเซตรายการที่ซ้ำซ้อนโดยที่ข้อมูลไม่สูญหาย ทำให้จำนวนกฎความสัมพันธ์ลดลงและเป็นกฎความสัมพันธ์ที่ไม่ซ้ำซ้อน ส่งผลให้ง่ายในการพิจารณากฎที่จะไปใช้ประโยชน์ นอกจากนี้ยังทำให้ประหยัดเรื่องเวลาในการหากฎความสัมพันธ์อีกด้วย

7. เอกสารอ้างอิง

[1] A. Ceglar and J. F. Roddick, "Association mining," *ACM Computing Surveys*, Vol. 38, pp. 5, 2006.

[2] A. Pietracaprina and F.Vandin, "Efficient Incremental Mining of Top-k Frequent Closed Itemsets," *In Proc. of 10th International Conference on Discovery Science*, Sendai, Japan, pp. 275-280, 1-4 October, 2007.

[3] C. Lucchesse, S. Orlando, and R. Perego, "DCI-Closed: A Fast and Memory Efficient Algorithm to Mine Frequent Closed Itemsets," *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations* (FIMI 2004), volume 126 of CEUR Workshop Proceedings, Brighton, UK, 1 November, 2004.

[4] C. Lucchese, S. Orlando, and R. Perego, "Fast and Memory Efficient Mining of Frequent Closed Itemsets," *In IEEE Journal Transactions on Knowledge and Data Engineering (TKDE)*, January, Vol. 18(1), pp. 21–36, 2006.

[5] G. Li and H.J. Hamilton, "Basic association rules," *In Proc. Of the 4th SIAM International conference on Data Mining* (*SDM'04*), Orlando, FL. 2004

[6] G. Grahne and J. Zhu, "Efficiently Using Prefix-trees in Mining Frequent Itemsets," *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations* (*FIMI* 2003), volume 90 of CEUR Workshop Proceedings, Melbourne, Florida, USA, Vol. 8, Iss. 1, pp. 103, 19 November, 2003.

[7] J, Han, J. Pei, and Y. Yin, "Mining Frequent Itemsets without Candidate Generation," *Proceeding 2000 ACM SIGMOD International conference Management of Data*, pp. 1-12, May, 2000.

[8] J. Han, J. Wang, Y. Lu and P. Tzvetkov, "Mining Top-k Frequent Closed Patterns without Minimum Support," *In Proc. of the 2002 IEEE International Conference on Data Mining*, Washington, DC, USA, pp. 211-218, 2002.

[9] J. Pei, J. Han, and R. Mao, "Closet: An Efficient Algorithm for Mining Frequent Closed Itemsets," *Proceedings of the ACM-SIGMOD International Workshop on Data Mining and Knowledge Discovery* (*DMKD* 2000), Dallas, Texas, USA, pp. 21–30, 2000.

[10] J.Wang, J. Han, and J. Pei, "Closet+: Searching for the Best Strategies for Mining Frequent Closed Itemsets," Proceedings of the 9th ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2003), Washington D.C., USA, pp. 236-245,

24-27 August 2003.

[11] J. Wang, J. Han, Y. Lu and P. Tzvetkov, "TFP: An Efficient Algorithm for Mining Top-K Frequent Closed Itemsets," *Journal of IEEE Transaction on knowledge and data mining*, Vol. 17, No. 5, pp. 652-664, May, 2005.

[12] M. J. Zaki, "Generating non-redundant association rules," *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '00), ACM, New York, NY, USA, pp. 34-43, 2000.

[13] M. J. Zaki, "Mining Non-Redundant Association Rules," *Data Mining and Knowledge Discovery*, Vol. 9, pp. 223-248, 2004.

[14] M. J. Zaki and C. J. Hsiao, "CHARM: An Efficient Algorithm for Closed Itemset Mining," *Proceedings of the 2nd SIAM International Conference on Data Mining, Arlington*, Virginia, USA, pp. 34–43, April, 2002.

[15] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Efficient Mining of Association Rules Using Closed Itemset Lattices," *Journal of Information Systems*, Vol.24, No.1, pp. 25-46, 1999.

[16] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering Frequent Closed Itemsets for Association Rules," *In Proc. of 7th International Conference on Database Theory*, LNCS, Vol. 1540, Springer, Verlag, Jerusalem, Israel, pp. 398–416, January, 1999.

[17] R. Agrawal and R. Srikant, "Fast algorithm for Mining Association Rules," *In Pro. 20th Very Large Database*, pp. 487-499, 1994.

[18] R. Agrawal, T. Imielinski, and A. Swami, " Mining association rules between sets of items in large databases," *In Proc. of the ACM SIGMOD International Conference on Management of Data*, pp. 207-216, May, 1993.

[19] T. Uno, T. Asai, Y. Uchida, and H. Arimura, "An Efficient Algorithm for Enumerating Closed Itemsets in Transaction Databases," *Proceedings of the 7th International Conference on Discovery Science*, Padova, Italy, pp. 16-31, 2-5 October, 2004.

[20] T. Uno, M. Kiyomi, and H. Arimura, "LCM ver. 2: Efficient Mining Algorithms for Frequent/Closed/ Maximal Itemsets," *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations* (FIMI 2004), volume 126 of CEUR Workshop Proceedings, Brighton, UK, 1 November, 2004.

[21] V. Pudi and J. R. Haritsa, "Generalized Closed Itemsets for Association Rule Mining," *19th International Conference on Data Engineering* (ICDE'03), pp. 714, 2003.

[22] Y.Bastide, N. Pasquier, R. Taouil, G. Stumme, " Mining Minimal Non-Redundant Association Rules using Frequent Closed Itemsets," *Proceedings of the First International Conference on Computational Logic* (*CL* '00), London, UK, pp. 972-986, 2000.