

การทำเหมืองข้อมูลที่มีประสิทธิภาพของรูปแบบลำดับเหตุการณ์เชิงกลุ่ม

Efficient Mining of Grouping Sequential Pattern

พรรณี ลิทธิเดช¹ นิภาพร ชนะมาร² และ พิสิษฐ์ นาคใจ³

บทคัดย่อ

งานสืบค้นหารูปแบบลำดับเหตุการณ์ย่อยในฐานข้อมูลลำดับเหตุการณ์ขนาดใหญ่มีความสำคัญมาก งานวิจัยนี้ได้พัฒนากระบวนการสืบค้นรูปแบบลำดับเหตุการณ์เชิงกลุ่มชื่อไบเดก โดยดัดแปลงจากวิธีการของไบด์ที่มีประสิทธิภาพมากสำหรับลำดับเหตุการณ์เชิงเดี่ยว อัลกอริทึมไบเดกสามารถค้นหารูปแบบลำดับเหตุการณ์ที่เกิดขึ้นในเวลาที่แตกต่างกันหรือพร้อม ๆ กันได้ จากการทดสอบเปรียบเทียบการทำงานระหว่างอัลกอริทึมไบเดกกับอัลกอริทึมโคลด์สะแพน พบว่าผลลัพธ์มีความถูกต้องตรงกันและอัลกอริทึมไบเดกใช้เวลาน้อยกว่า ด้วยระดับนัยสำคัญ 0.05 ผู้วิจัยได้นำอัลกอริทึมไบเดกไปประยุกต์ใช้กับงานด้านการพยากรณ์อากาศ โดยการใช้ข้อมูลสภาพอากาศที่มีความสัมพันธ์กันย้อนหลัง 3 วัน ทำให้สามารถพยากรณ์อุณหภูมิในวันถัดไปได้ ด้วยค่าความถูกต้อง 82% และ 79% ในชุดข้อมูลฝึกฝนและชุดทดสอบ ตามลำดับ

คำสำคัญ: รูปแบบลำดับเหตุการณ์, อัลกอริทึมโคลด์สะแพน, อัลกอริทึมไบด์, การพยากรณ์

Abstract

Discovering frequent subsequences in large sequence database is a very important task. This work develops an algorithm for mining group sequential patterns named BIDE; BI-Direction Extension algorithm on Grouping sequences. The proposed algorithm is based on the BIDE algorithm which is very efficient algorithm for single-item sequences. BIDE algorithm can be applied for subsequences that may occur at different time intervals or appear at the same time. The proposed method was then applied to same datasets as a verifying result to the CloSpan algorithm. Our algorithm performs correct results as well as significantly less processing time ($\alpha = 0.05$). BIDE was applied for weather forecasting. By using related attributes measured on 3 previous days, the temperature value of the next day can be forecasted with 82% and 79% accuracies on training and testing data sets, respectively.

Keywords: Sequential pattern, CloSpan algorithm, BIDE algorithm, Forecasting

¹ อาจารย์, ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ, คณะวิทยาศาสตร์, มหาวิทยาลัยนเรศวร

² อาจารย์, สาขาวิชาวิทยาการคอมพิวเตอร์, คณะวิทยาศาสตร์และเทคโนโลยี, มหาวิทยาลัยราชภัฏสกลนคร

³ นิสิตปริญญาโท, ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ, คณะวิทยาศาสตร์, มหาวิทยาลัยนเรศวร

1. บทนำ

การสืบค้นรูปแบบลำดับเหตุการณ์ (Sequential Pattern Discovery) เป็นการหาเหมือนข้อมูลอีกวิธีหนึ่งในการสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ มีเป้าหมายในการหาความสัมพันธ์ของชุดข้อมูลที่เกี่ยวข้องกันทางเวลาหรือลำดับการเกิดของชุดข้อมูล ซึ่งผลที่ได้จากการสืบค้นรูปแบบลำดับเหตุการณ์นั้นจะได้รูปแบบลำดับเหตุการณ์ที่น่าสนใจหรือเกิดขึ้นบ่อย (Frequent Sequence Pattern) [1-2] ถ้ารูปแบบลำดับเหตุการณ์ดังกล่าวสามารถครอบคลุมรูปแบบลำดับเหตุการณ์ย่อย (Subsequence) ที่มีค่าสนับสนุนเท่ากันได้ เรียกว่ารูปแบบลำดับเหตุการณ์แบบปิด (Closed Sequence Pattern) โดยจะต้องพิจารณาถึงลำดับการเกิดของข้อมูลและการเกิดพร้อมกันของข้อมูล [3-4] มีการทำวิจัยหาวิธีการสืบค้นรูปแบบลำดับเหตุการณ์แบบปิดกันอย่างกว้างขวาง โดย Jianyong Wang และ Jiawei Han [3-4] ได้นำเทคนิคอัลกอริทึมโคลด์สแปน (CloSpan: Mining Closed Sequential Pattern) และอัลกอริทึมไบด์ (BIDE: Bi-Direction Extension) มาทำการสืบค้นรูปแบบลำดับเหตุการณ์แบบปิด ผลการทำงานของอัลกอริทึมไบด์มีประสิทธิภาพดีกว่าและใช้เวลาน้อยกว่าการทำงานของอัลกอริทึมโคลด์สแปน [3] ซึ่งอัลกอริทึมไบด์นี้สามารถทำการสืบค้นกับข้อมูลที่เป็นข้อมูลแบบลำดับเหตุการณ์เชิงเดี่ยวเท่านั้น งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อปรับปรุงอัลกอริทึมไบด์ให้สามารถทำการสืบค้นรูปแบบลำดับเหตุการณ์เชิงกลุ่มได้ เรียกอัลกอริทึมนี้ว่าอัลกอริทึมไบเดก (BIDEG: BI-Direction Extension algorithm on Grouping sequences) ผู้วิจัยได้ทดสอบการทำงานของอัลกอริทึมไบเดกด้วยข้อมูลชุดเดียวกันกับการทำงานของอัลกอริทึมโคลด์สแปน เพื่อตรวจสอบความถูกต้องของผลลัพธ์และทำการทดลองหาค่าข้อมูลแบบลำดับเหตุการณ์เชิงกลุ่ม เพื่อเปรียบเทียบประสิทธิภาพของเวลาที่ใช้ในการสืบค้นของทั้งสองอัลกอริทึม นอกจากนี้ เพื่อแสดงว่าอัลกอริทึมไบเดกสามารถใช้ประโยชน์สำหรับการสืบค้นกว้างขวางขึ้น ผู้วิจัยนี้ได้ประยุกต์ใช้อัลกอริทึมไบเดกกับการพยากรณ์ค่าอุณหภูมิจากข้อมูลอากาศด้วยกฎความสัมพันธ์ที่ได้จากอัลกอริทึมไบเดก ได้ผลค่าความถูกต้องเป็นที่ยอมรับได้

2. กฎความสัมพันธ์

การสืบค้นกฎความสัมพันธ์ เป็นการหาความสัมพันธ์ของข้อมูลภายในทรานเซกชันเดียวกัน กฎความสัมพันธ์ที่ได้จะแสดงว่า ถ้าพบข้อมูลชุดใดแล้ว มีแนวโน้มที่จะพบข้อมูลชุดอื่นด้วย [5] ความเป็นไปได้ที่จะเกิดข้อมูลทั้งสองชุดในทรานเซกชันเดียวกันมากกว่าหรือเท่ากับค่าสนับสนุน การค้นหากฎความสัมพันธ์ จึงมักเป็นการค้นหากฎความสัมพันธ์ของข้อมูลจากฐานข้อมูลขนาดใหญ่ที่มีอยู่ เพื่อนำไปใช้ในการวิเคราะห์หรือทำนายปรากฏการณ์ต่างๆ โดยเทคนิคนี้ใช้กันอย่างแพร่หลายในการขายสินค้าหรือการวิเคราะห์ข้อมูลที่เป็นทรานเซกชันประโยชน์ของเทคนิคการสืบค้นกฎความสัมพันธ์ ได้แก่ แสดงความสัมพันธ์ของการเกิดขึ้นของข้อมูล อธิบายความสัมพันธ์ที่ซ่อนอยู่ในฐานข้อมูลที่ไม่สามารถพิจารณาได้ด้วยตาเปล่า ใช้เป็นข้อมูลประกอบการตัดสินใจ เป็นต้น โดยปกติแล้วกฎความสัมพันธ์ไม่ได้บอกเราเฉพาะข้อมูลที่เกิดขึ้นพร้อมกันเท่านั้น กฎที่ได้ยังอธิบายลำดับเหตุการณ์ที่ต่อเนื่องกันตามลำดับของเวลาที่เกิดขึ้นได้ด้วย

ข้อมูลลำดับเหตุการณ์ สามารถแบ่งออกเป็น 2 แบบ คือ

1) ข้อมูลแบบลำดับเหตุการณ์เชิงเดี่ยว คือ ลำดับเหตุการณ์โดยไม่มีลำดับเหตุการณ์ที่เกิดลำดับเหตุการณ์ย่อยพร้อมกัน เช่น ลำดับเหตุการณ์ $\langle (A) (B) (C) (D) \rangle$ หมายถึง ลำดับเหตุการณ์ A เกิดขึ้นก่อนลำดับเหตุการณ์ B ลำดับเหตุการณ์ B เกิดขึ้นก่อนลำดับเหตุการณ์ C และลำดับเหตุการณ์ D จะเกิดขึ้นลำดับสุดท้าย

2) ข้อมูลลำดับเหตุการณ์เชิงกลุ่ม คือ ลำดับเหตุการณ์โดยอาจมีลำดับเหตุการณ์ย่อยเกิดขึ้นพร้อมกันได้ เช่น ลำดับเหตุการณ์ $\langle (AB) (D) (E) (F) \rangle$ หมายถึง ลำดับเหตุการณ์ A และ B เกิดขึ้นพร้อมกัน จากนั้นเกิดลำดับเหตุการณ์ D E และ F ตามลำดับ

2.1 การสืบค้นรูปแบบลำดับเหตุการณ์แบบปิด

คือ รูปแบบลำดับเหตุการณ์ที่เกิดบ่อยที่ครอบคลุม รูปแบบลำดับเหตุการณ์ที่มีค่าสนับสนุนเท่ากัน ถูกเสนอขึ้นมาเพื่อแก้ปัญหาการสืบค้นรูปแบบลำดับเหตุการณ์ที่สร้างลำดับเหตุการณ์จำนวนมากเมื่อฐานข้อมูลมีขนาดใหญ่ เพื่อคัด

รูปแบบลำดับเหตุการณ์ที่ซ้ำซ้อนหรือรูปแบบที่ถูกครอบคลุมโดยรูปแบบลำดับเหตุการณ์อื่นออกไป ซึ่งทำให้ได้จำนวนรูปแบบน้อยกว่าจำนวนรูปแบบที่ได้จากการสืบค้นรูปแบบลำดับเหตุการณ์ที่เกิดบ่อย นิยามพื้นฐานของรูปแบบลำดับเหตุการณ์แบบปิด [3] ดังนี้

นิยามที่ 1 $\alpha = \langle a_1 a_2 \dots a_n \rangle$ คือ ลำดับเหตุการณ์ ซึ่ง $a_i \subseteq I$ โดยที่ I คือ เซตของชิ้นข้อมูล (Items) และ n คือ จำนวนชิ้นข้อมูลหรือความยาวของลำดับเหตุการณ์ α แทนด้วย n - sequence

นิยามที่ 2 ลำดับเหตุการณ์ $\alpha = \langle a_1 a_2 \dots a_n \rangle$ เป็นลำดับเหตุการณ์ย่อย (Subsequence) ของลำดับเหตุการณ์ $\beta = \langle b_1 b_2 \dots b_m \rangle$ ก็ต่อเมื่อมีจำนวนเต็มบางตัว $i_1, i_2, \dots, i_n$ ซึ่ง $1 \leq i_1 < i_2 < \dots < i_n \leq m$ และ $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, \dots, a_n \subseteq b_{i_n}$ เรียกลำดับเหตุการณ์ β ว่า ซุปเปอร์ลำดับเหตุการณ์ (Super sequence) ของลำดับเหตุการณ์ α เขียนแทนด้วย $\alpha \subseteq \beta$

ข้อสังเกต ถ้าค่าสนับสนุน (Support) ของลำดับเหตุการณ์ α และลำดับเหตุการณ์ β มีค่าเท่ากัน แสดงว่าลำดับเหตุการณ์ α ถูกครอบคลุมโดยลำดับเหตุการณ์ β

นิยามที่ 3 เซตของลำดับเหตุการณ์ที่เกิดขึ้นบ่อย (Frequent Sequence : FS) ประกอบด้วยลำดับเหตุการณ์ที่มีความถี่หรือค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ เขียนแทนด้วย $FS = \{\alpha \mid \text{support}(\alpha) \geq \text{min_sup}\}$

นิยามที่ 4 เซตของรูปแบบลำดับเหตุการณ์แบบปิด (Closed Sequence : CS) ประกอบด้วยรูปแบบลำดับเหตุการณ์ที่ไม่มีซูปเปอร์ลำดับเหตุการณ์ที่มีค่าสนับสนุนเท่ากัน เขียนแทนด้วย $CS = \{\alpha \mid \alpha \in FS \wedge (\exists \beta \in FS \mid \alpha \subseteq \beta \wedge \text{support}(\alpha) = \text{support}(\beta))\}$

นิยามที่ 5 ค่าสนับสนุนขั้นต่ำ (Minimum Support) คือ ค่าที่กำหนดขึ้นเพื่อใช้เป็นเกณฑ์ขั้นต่ำในการตัดสินใจว่ากลุ่มข้อมูลใดเป็นกลุ่มข้อมูลที่เกิดขึ้นพร้อมกันบ่อย กลุ่มข้อมูลใดๆ ที่มีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำที่กำหนดจะเรียกว่า กลุ่มข้อมูลนั้นเกิดขึ้นบ่อยและน่าสนใจ

2.2 อัลกอริทึมโคลด์สแพน

Xifeng Yan และคณะ [4] ได้เสนออัลกอริทึมโคลด์สแพน โดยสร้างเซตของรูปแบบลำดับเหตุการณ์แบบปิดคู่แข่ง (Candidate Closed Sequential Pattern) จากนั้นจะคัดลำดับที่ไม่ใช่ลำดับเหตุการณ์แบบปิดออก เพราะว่าอัลกอริทึมโคลด์สแพน จำเป็นที่จะต้องมีการเก็บเซตของลำดับเหตุการณ์แบบปิดคู่แข่งไว้ ซึ่งจะทำให้สิ้นเปลืองหน่วยความจำมาก และทำให้เสียเวลาในการค้นหาส่วนปิดคลุมหรือ Closure เมื่อมีลำดับเหตุการณ์แบบปิดคู่แข่งจำนวนมาก

ตารางที่ 1 ฐานข้อมูลลำดับเหตุการณ์เชิงกลุ่ม (3 ระเบียบ)

ระเบียบ	ลำดับเหตุการณ์
1	$\langle (AF) D E A \rangle$
2	$\langle E A B \rangle$
3	$\langle E (ABF) (BDE) \rangle$

จากตารางที่ 1 กำหนดตัวอักษร A, B, D, E และ F แทนชิ้นข้อมูล (Items) ในลำดับเหตุการณ์ และกำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 2 ลำดับเหตุการณ์ $\langle (AF) D E A \rangle$ หมายความว่า ลำดับเหตุการณ์ A และ F เกิดขึ้นพร้อมกัน จากนั้นเกิดลำดับเหตุการณ์ D E และ A ตามลำดับ เมื่อทำการสืบค้นจากฐานข้อมูลพบลำดับเหตุการณ์ $\langle (AF) D \rangle$ ในระเบียบ 1 และระเบียบ 3 โดยในระเบียบ 3 รูปแบบลำดับเหตุการณ์ $\langle E (ABF) (BDE) \rangle$ พบเหตุการณ์ $\langle (AF) \rangle$ ในเหตุการณ์ $\langle (ABF) \rangle$ และพบเหตุการณ์ $\langle (D) \rangle$ ในเหตุการณ์ $\langle (BDE) \rangle$ สามารถนับลำดับเหตุการณ์ $\langle (AF) D \rangle$ ในระเบียบ 3 ได้ เมื่อค่าสนับสนุนมีค่าเท่ากับ 2 ซึ่งมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ แสดงว่าลำดับเหตุการณ์ $\langle (AF) D \rangle$ เป็นรูปแบบลำดับเหตุการณ์ และพบลำดับเหตุการณ์ $\langle (AF) \rangle$ ในระเบียบ 1 และ 3 ดังนั้น ค่าสนับสนุน มีค่าเท่ากับ 2 แสดงว่าลำดับเหตุการณ์ $\langle (AF) \rangle$ ก็เป็นรูปแบบลำดับเหตุการณ์เช่นกัน แต่เนื่องจากรูปแบบลำดับเหตุการณ์ $\langle (AF) \rangle$ เป็นลำดับเหตุการณ์ย่อยของรูปแบบลำดับเหตุการณ์ $\langle (AF) D \rangle$ และมีค่าสนับสนุนเท่ากัน แสดงว่ารูปแบบลำดับเหตุการณ์ $\langle (AF) \rangle$ ถูกครอบคลุมโดยรูปแบบลำดับเหตุการณ์ $\langle (AF) D \rangle$ ซึ่งจะสามารถพิจารณา

รูปแบบลำดับเหตุการณ์ <(AF)>จากรูปแบบลำดับเหตุการณ์ <(AF) D> ได้

การสืบค้นรูปแบบลำดับเหตุการณ์แบบปิดจึงทำการสืบค้นเฉพาะรูปแบบลำดับเหตุการณ์ที่มีขนาดใหญ่และครอบคลุมรูปแบบลำดับเหตุการณ์ย่อยได้ ซึ่งผลลัพธ์ที่ได้จากการสืบค้นรูปแบบลำดับเหตุการณ์แบบปิดจากฐานข้อมูลในตารางที่ 1 จะสืบค้นได้ทั้งหมด 4 รูปแบบลำดับเหตุการณ์ คือ { <(AF) D>: 2, <(AF) E>: 2, <E A>: 3, <E A B>: 2 }

2.3 อัลกอริทึมไบต์

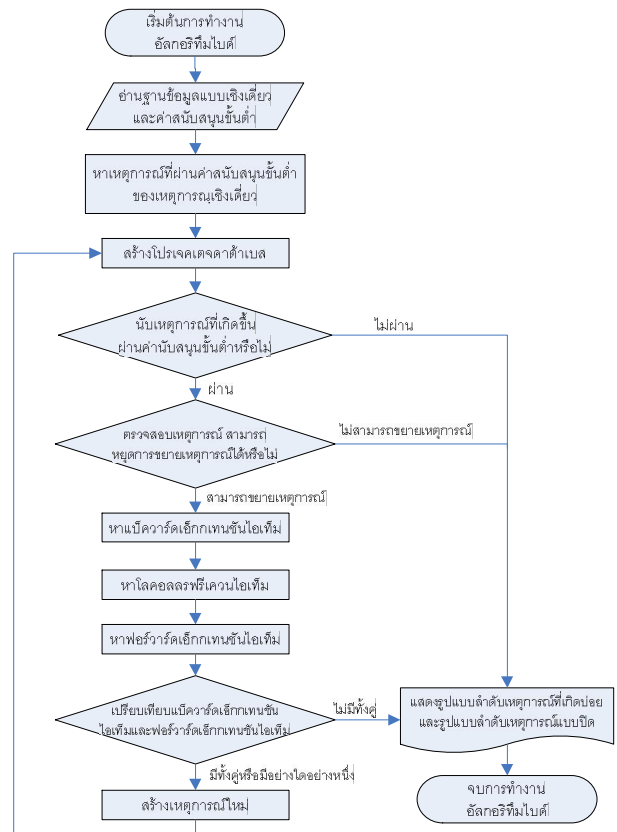
อัลกอริทึมไบต์ เป็นเทคนิคหนึ่งในการสืบค้นรูปแบบลำดับเหตุการณ์แบบปิด ซึ่งอัลกอริทึมนี้ไม่มีการเก็บรูปแบบลำดับเหตุการณ์คู่แข่ง และจะใช้วิธีการตรวจสอบโคลสเซอร์แบบสองทิศทางเพื่อหาลำดับเหตุการณ์แบบปิด และทำการขจัดเส้นทางที่ไม่สามารถนำไปสู่ลำดับเหตุการณ์แบบปิด โดยการหยุดขยายลำดับเหตุการณ์เดิมนั้น และใช้วิธีการถดถอยเพื่อให้การคำนวณเร็วขึ้น อัลกอริทึมนี้จะทำการสืบค้นรูปแบบลำดับเหตุการณ์กับข้อมูลลำดับเหตุการณ์เชิงเดียว [3] แสดงดังตารางที่ 2 ดังนี้

ตารางที่ 2 ฐานข้อมูลลำดับเหตุการณ์เชิงเดียว

ระเบียบ	ลำดับเหตุการณ์
1	<C A A B C>
2	<A B C B>
3	<C A B C>
4	<A B B C A>

ข้อมูลลำดับเหตุการณ์ในตารางที่ 2 เป็นข้อมูลที่ใช้ในการพิจารณา จะเห็นว่าแต่ละระเบียบของฐานข้อมูลลำดับเหตุการณ์ประกอบด้วยชุดของลำดับเหตุการณ์ที่เรียงตามเวลาการเกิดก่อนหลังของชุดข้อมูล โดยมีลำดับเป็นสำคัญ มีทั้งหมด 4 ระเบียบ ซึ่งแต่ละระเบียบประกอบด้วย ระเบียบและชุดของลำดับเหตุการณ์ ตัวอย่าง เช่น จากระเบียบที่ 1 หมายความว่า เหตุการณ์ที่เกิดเหตุการณ์ย่อย C ตามด้วยเหตุการณ์ A, A, B และ C

ตามลำดับซึ่งการทำงานของอัลกอริทึมไบต์ แสดงดังรูปที่ 1 ดังนี้



รูปที่ 1 ขั้นตอนการทำงานของอัลกอริทึมไบต์

จากรูปที่ 1 อธิบายการทำงานของอัลกอริทึมไบต์ได้ดังนี้

- 1) อ่านฐานข้อมูลลำดับเชิงเดียวและค่าสนับสนุนขั้นต่ำ
- 2) หาเหตุการณ์ที่ผ่านค่าสนับสนุนขั้นต่ำ ของเหตุการณ์เชิงเดียวขนาด 1 ขึ้นข้อมูล
- 3) นำเหตุการณ์ที่ได้มาทำการสร้างโปรเจกต์เดตาเบส (Projected Database) หรือฐานข้อมูลใหม่
- 4) นับเหตุการณ์ในโปรเจกต์เดตาเบสว่าผ่านค่าสนับสนุนขั้นต่ำหรือไม่ โดยพิจารณาจากการปรากฏของเหตุการณ์ในโปรเจกต์เดตาเบส
 - ถ้าไม่ผ่านค่าสนับสนุนขั้นต่ำให้ขจัดเหตุการณ์ โดยการแสดงผลเหตุการณ์และหยุดทำการขยายเหตุการณ์นั้น

- ถ้าผ่านค่าสนับสนุนขั้นต่ำให้นำวิธีการ แบ็คสแกนพรุณนิง (Back Scan Pruning) มาตรวจสอบเหตุการณ์ว่าสามารถจัดออกได้อีกหรือไม่

5) วิธีการแบ็คสแกนพรุณนิง ใช้ตรวจสอบว่าสามารถที่จะหยุดขยายเหตุการณ์ได้อีกหรือไม่ ในส่วนนี้จะตรวจสอบตามลำดับเหตุการณ์ที่สืบค้นทีละเหตุการณ์จนกว่าจะไม่สามารถขยายเหตุการณ์ได้อีก

6) วิธีการตรวจสอบแบ็ควาร์ดเอ็กเทนชันไอเท็ม (Backward-extension Items) ใช้ตรวจสอบความเป็นไปได้ของเหตุการณ์นั้น สามารถเป็นรูปแบบลำดับเหตุการณ์แบบปิดได้หรือไม่

7) วิธีการตรวจสอบโลคอลลีฟรเควนไอเท็ม (Locally Frequent Item) เป็นการนับไอเท็มในโปรเจกต์เดคดาต้าเบสของเหตุการณ์ที่สามารถขยายเหตุการณ์ได้ ว่ามีไอเท็มใดบ้างที่ผ่านค่าสนับสนุนขั้นต่ำ เพื่อนำไปเพิ่มที่ได้ไปตรวจสอบหาฟอว์วาร์ดเอ็กเทนชันไอเท็ม (Forward-extension Items) เมื่อได้ไอเท็มที่ผ่านค่าสนับสนุนแล้ว สามารถนำไอเท็มไปขยายเป็นเหตุการณ์ย่อยได้

8) วิธีการตรวจสอบฟอว์วาร์ดเอ็กเทนชันไอเท็ม เป็นการตรวจสอบว่าโลคอลลีฟรเควนไอเท็มที่ได้นั้นมีเหตุการณ์ใดที่มีค่าสนับสนุนเท่ากับค่าสนับสนุนของเหตุการณ์นั้น เพื่อให้ทราบว่ามีฟอว์วาร์ดเอ็กเทนชันไอเท็มหรือไม่

9) ทำการเปรียบเทียบแบ็ควาร์ดเอ็กเทนชันไอเท็มฟอว์วาร์ดเอ็กเทนชันไอเท็ม คือ ถ้าไม่มีแบ็ควาร์ดเอ็กเทนชันไอเท็มและฟอว์วาร์ดเอ็กเทนชันไอเท็มทั้งคู่ แสดงว่าเป็นเหตุการณ์แบบปิดหรือรูปแบบลำดับเหตุการณ์แบบปิด

10) แสดงรูปแบบลำดับเหตุการณ์ที่เกิดขึ้นบ่อยและรูปแบบลำดับเหตุการณ์แบบปิด

11) ทำการสร้างเหตุการณ์ใหม่ โดยการนำเอาเหตุการณ์เก่ามารวมกับโลคอลลีฟรเควนไอเท็ม เมื่อได้เหตุการณ์ใหม่ แล้วให้ทำซ้ำขั้นตอนที่ 3 ถึง 9 จนไม่สามารถหาเหตุการณ์ได้อีก

โดยการสืบค้นรูปแบบลำดับเหตุการณ์แบบปิดของอัลกอริทึมไบเค แสดงสืบค้นรูปแบบลำดับเหตุการณ์บนข้อมูลแบบลำดับเหตุการณ์เชิงเดี่ยว ซึ่งข้อมูลลักษณะนี้มีข้อจำกัดว่า เหตุการณ์ย่อยแต่ละเหตุการณ์เกิดขึ้น ณ เวลาที่แตกต่างกัน ในงานวิจัยนี้สนใจประยุกต์อัลกอริทึมไบเค ให้

สามารถทำการสืบค้นรูปแบบลำดับเหตุการณ์กับข้อมูลแบบลำดับเหตุการณ์เชิงกลุ่ม โดยที่เหตุการณ์ย่อยแต่ละเหตุการณ์อาจเกิดขึ้นพร้อมกัน ซึ่งทำให้มีประโยชน์สำหรับการสืบค้นกว้างขวางขึ้น ดังแสดงวิธีการดำเนินการวิจัยในหัวข้อต่อไป

3. วิธีดำเนินการวิจัย

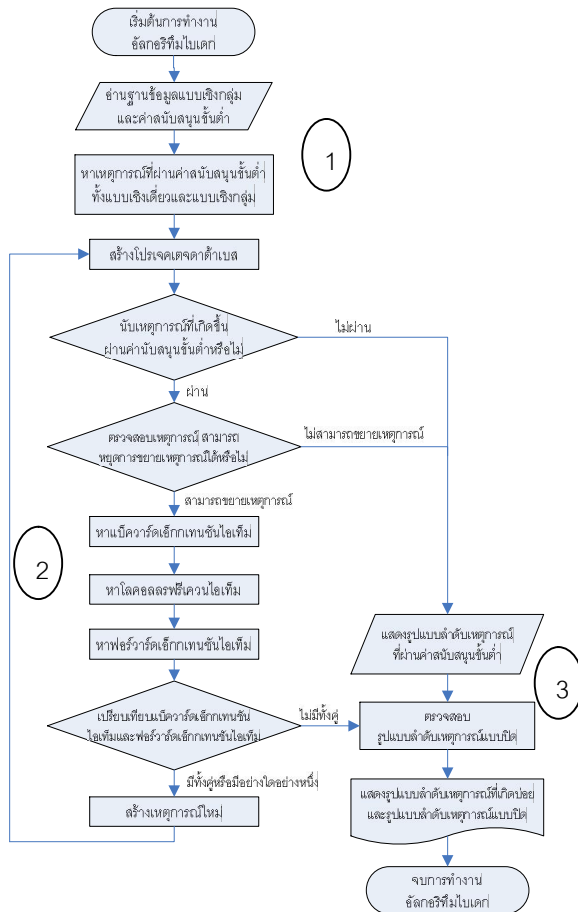
การทำงานของอัลกอริทึมไบเค แสดงดังรูปที่ 2 โดยจะทำการปรับปรุงเพิ่มเติมจากอัลกอริทึมไบเค เพื่อให้อัลกอริทึมนี้สามารถทำการสืบค้นรูปแบบลำดับเหตุการณ์กับข้อมูลลำดับเหตุการณ์เชิงกลุ่ม จากนั้นจะใช้ฐานข้อมูลลำดับเหตุการณ์เชิงกลุ่ม จากตารางที่ 1 ในการทดสอบความถูกต้องของอัลกอริทึมไบเค ว่าสามารถทำการสืบค้นรูปแบบลำดับเหตุการณ์ถูกต้องตรงกับรูปแบบลำดับเหตุการณ์ที่สืบค้นด้วยอัลกอริทึมโคลัสสะแพนหรือไม่ ในลำดับต่อไปได้ทำการสุ่มข้อมูลลำดับเหตุการณ์เชิงกลุ่ม ขนาดข้อมูล 100 ระเบียบจำนวน 10 ชุดข้อมูล เพื่อทดสอบประสิทธิภาพเวลาในการสืบค้นรูปแบบลำดับเหตุการณ์กับอัลกอริทึมโคลัสสะแพน

จากรูปที่ 2 เป็นขั้นตอนการทำงานของอัลกอริทึมไบเค โดยปรับปรุงเพิ่มเติมในอัลกอริทึมไบเค ดังแสดงตามหมายเลข ดังนี้

หมายเลข ① เพิ่มเดิมการสืบค้นเหตุการณ์ข้อมูลแบบเชิงกลุ่ม โดยปรับปรุงจากอัลกอริทึมไบเคเดิมจะสืบค้นเหตุการณ์แบบเชิงเดี่ยวเท่านั้น เมื่ออ่านฐานข้อมูลลำดับเหตุการณ์และค่าสนับสนุนขั้นต่ำเข้าในโปรแกรมแล้วมีวิธีการสืบค้น ดังนี้

1) ตรวจสอบเหตุการณ์เชิงเดี่ยวและเชิงกลุ่ม โดยเก็บค่าในตัวแปรหนึ่ง เช่น เหตุการณ์ <(D)> เป็นเหตุการณ์เชิงเดี่ยว เหตุการณ์ <(ABF)> เป็นเหตุการณ์เชิงกลุ่ม

2) นำเหตุการณ์เชิงเดี่ยวและเชิงกลุ่มมานับค่าสนับสนุนว่าเหตุการณ์ใดผ่านค่าสนับสนุนขั้นต่ำ โดยเหตุการณ์เชิงกลุ่มจะต้องแยกเป็นเหตุการณ์ย่อยก่อน เช่น เหตุการณ์ <(ABF)> ได้เหตุการณ์ย่อยเป็น <(ABF) (AB) (AF) (BF) (A) (B) (F)> สามารถตรวจสอบได้ดังนี้



รูปที่ 2 ขั้นตอนการทำงานของอัลกอริทึมไบนารี

- ถ้าเหตุการณ์ซ้ำและไม่เคยปรากฏมาก่อน ให้เพิ่มค่าลำดับชั้นขึ้น 1
- ถ้าเหตุการณ์ซ้ำและเคยปรากฏมาก่อนให้ข้ามไป
- ถ้าเหตุการณ์ไม่ซ้ำ ให้เพิ่มเหตุการณ์นั้นและเพิ่มค่าลำดับชั้นให้เหตุการณ์นั้น 1

หมายเลข ② เพิ่มเติมในระหว่างการสืบค้นแต่ละขั้นตอน ของการทำงานจะพิจารณาข้อมูลเป็นแบบเชิงกลุ่มด้วย ซึ่งอัลกอริทึมไบนารีระหว่างการสืบค้นเหตุการณ์จะพิจารณาข้อมูลแบบเชิงเดี่ยวเท่านั้น

หมายเลข ③ เพิ่มเติมการตรวจสอบรูปแบบลำดับเหตุการณ์แบบปิดของข้อมูลแบบเชิงกลุ่ม ซึ่งอัลกอริทึมไบนารีเดิมจะไม่จำเป็นต้องตรวจสอบ เนื่องจากข้อมูลเป็นแบบเชิงเดี่ยวเท่านั้น โดยมีวิธีการตรวจสอบ ดังนี้

1) เหตุการณ์ที่จะเป็นรูปแบบลำดับเหตุการณ์แบบปิดได้ เหตุการณ์นั้นต้องอยู่ในระดับที่มากกว่าหนึ่งระดับ

2) ความยาวโหนดลูกมากกว่าความยาวโหนดแม่ และค่าสนับสนุนมากกว่าหรือเท่ากับความยาวโหนดแม่

3) โหนดลูกที่มีค่าสนับสนุนน้อยกว่าความยาวโหนดแม่ บางเหตุการณ์ สามารถเป็นรูปแบบลำดับเหตุการณ์แบบปิดได้ เนื่องจากไอเท็มในโหนดลูกมีไอเท็มครบทุกตัวในโหนดแม่และโหนดข้างเคียง

ผู้วิจัยได้พัฒนาอัลกอริทึมไบนารีและทดสอบการทำงานของอัลกอริทึมได้ใช้ฐานข้อมูลลำดับเหตุการณ์เชิงกลุ่ม (3 ระดับ) ดังตารางที่ 1 และฐานข้อมูลลำดับเหตุการณ์เชิงกลุ่ม (4 ระดับ) ดังตารางที่ 3 ผลของการสืบค้นรูปแบบลำดับเหตุการณ์ด้วยอัลกอริทึมไบนารี โดยกำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 2 สามารถสืบค้นรูปแบบลำดับเหตุการณ์ที่เกิดขึ้นบ่อย และรูปแบบลำดับเหตุการณ์แบบปิด ได้ถูกต้องตรงกับการสืบค้นด้วยอัลกอริทึมโคลัสสะแพน ทุกประการ

ตารางที่ 3 ฐานข้อมูลลำดับเหตุการณ์เชิงกลุ่ม (4 ระดับ)

ระดับ	ลำดับเหตุการณ์
1	<(AC) (D) (E)>
2	<(E) (ABCF) (E)>
3	<(A) (E) (B)>
4	<(D) (AC) (E)>

4. การทดสอบประสิทธิภาพ

ผู้วิจัยได้ทำการทดสอบเปรียบเทียบประสิทธิภาพการสืบค้นรูปแบบลำดับเหตุการณ์ของอัลกอริทึมโคลัสสะแพนและอัลกอริทึมไบนารี โดยสุ่มฐานข้อมูลขนาด 100 ระดับ จำนวน 10 ชุดข้อมูล โดยกำหนดค่าสนับสนุนขั้นต่ำเท่ากับ 8 ทุกชุดข้อมูล จากการทดสอบด้วยโปรแกรม 3 ครั้ง แล้วนำเวลาที่ใช้ในการสืบค้นมาหาค่าเฉลี่ย ได้เวลาเฉลี่ยในการประมวลผลของอัลกอริทึม ดังตารางที่ 4

ตารางที่ 4 เวลาเฉลี่ยการสืบค้นรูปแบบลำดับเหตุการณ์ด้วยอัลกอริทึมโคลัสสะแพนและอัลกอริทึมไบเดก

ขนาดข้อมูล 100 ระเบียบ	เวลาเฉลี่ยในการประมวลผล (มิลลิวินาที)	
	อัลกอริทึมโคลัสสะแพน	อัลกอริทึมไบเดก
ชุดที่ 1	279	173
ชุดที่ 2	280	172
ชุดที่ 3	288	224
ชุดที่ 4	318	183
ชุดที่ 5	347	189
ชุดที่ 6	302	161
ชุดที่ 7	286	142
ชุดที่ 8	261	162
ชุดที่ 9	282	173
ชุดที่ 10	318	168
เวลาเฉลี่ย	296.1	174.7

ข้อมูลตารางที่ 4 ถูกนำมาตรวจสอบค่าความแปรปรวนของข้อมูลทั้ง 2 ชุด พบว่า เวลาในการสืบค้นรูปแบบลำดับเหตุการณ์ด้วยอัลกอริทึมโคลัสสะแพนกับการสืบค้นรูปแบบลำดับเหตุการณ์ด้วยอัลกอริทึมไบเดกมีความแปรปรวนพอกัน และเวลาเฉลี่ยของการสืบค้นรูปแบบลำดับเหตุการณ์ด้วยอัลกอริทึมไบเดกใช้เวลาในการสืบค้นน้อยกว่าอัลกอริทึมโคลัสสะแพนที่ระดับนัยสำคัญ 0.05

5. การประยุกต์ใช้อัลกอริทึมไบเดกเพื่อการพยากรณ์

5.1 การพยากรณ์อากาศ

ในการพยากรณ์อากาศโดยทั่วไปแล้ว จะทำโดยการบันทึกผลการตรวจอากาศลงบนแผนที่อากาศ โดยมีกระบวนการที่ซับซ้อน มีงานวิจัยที่เกี่ยวข้องกับการพยากรณ์อากาศมากมาย ทั้งที่ใช้วิธีทางด้านสถิติสำหรับข้อมูลอนุกรมเวลา และ ใช้วิธีทางด้านการทำเหมืองข้อมูลด้วยเทคนิคอื่น [7-8] งานวิจัยนี้ได้นำอัลกอริทึมไบเดกมาประยุกต์ใช้ในการพยากรณ์อุณหภูมิ โดยใช้ข้อมูลจากกรมอุตุนิยมวิทยาประเทศไทย ปี 2005 และปี 2006 ซึ่งจะต้องมีการเตรียมข้อมูลให้อยู่ในรูปแบบลำดับ

เหตุการณ์เพื่อนำมาใช้ในการสร้างกฎด้วยอัลกอริทึมไบเดก และการพยากรณ์ ค่าอุณหภูมิด้วยกฎความสัมพันธ์ ดังอธิบายในหัวข้อต่อไป

5.2 การเตรียมข้อมูล

ข้อมูลอากาศ ซึ่งประกอบด้วย ข้อมูลอุณหภูมิ ความชื้น ความเร็วลม ความกดอากาศ และปริมาณน้ำฝน จำเป็นจะต้องจัดเตรียมข้อมูลโดยมีขั้นตอนดังนี้ (1) คัดกรองข้อมูลเพื่อทำการตัดข้อมูลบางส่วนที่ไม่เกี่ยวข้องกับการพยากรณ์ทิ้ง โดยมีหลักการพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์ ประกอบความเห็นจากผู้รู้ รวมทั้งคัดระเบียบที่มีข้อมูลสูญหายออกไป (2) คำนวณข้อมูลรายวัน และแบ่งกลุ่มข้อมูลของแต่ละตัวแปร (Clustering) (3) กำหนดสัญลักษณ์เพื่อกำหนดช่วงของข้อมูลแต่ละกลุ่ม และ (4) แปลงข้อมูลให้เป็นลำดับเหตุการณ์ โดยที่ข้อมูลอากาศในวันเดียวกันจะถือว่าเป็นข้อมูลเชิงกลุ่มเดียวกัน ดังนั้นใน 1 ระเบียบจะมีลำดับเหตุการณ์ทั้งหมด 4 เหตุการณ์ ประกอบด้วย เหตุการณ์ที่ 1 (วันที่ 1) เหตุการณ์ที่ 2 (วันที่ 2) เหตุการณ์ที่ 3 (วันที่ 3) และข้อมูลกลุ่มเป้าหมายของการพยากรณ์ (วันที่ 4) ซึ่งเป็นข้อมูลของอุณหภูมิเพียงอย่างเดียว ดังตัวอย่าง

$$\langle (BJTdm)(EMWgq)(HPZju)(y) \rangle$$

โดยที่ (BJTdm) เป็นข้อมูลอากาศ 5 ตัวแปร ของวันที่ 1
(EMWgq) เป็นข้อมูลอากาศ 5 ตัวแปร ของวันที่ 2
(HPZju) เป็นข้อมูลอากาศ 5 ตัวแปร ของวันที่ 3
(y) เป็นข้อมูลอุณหภูมิของวันที่ต้องการพยากรณ์

5.3 การสร้างกฎและการพยากรณ์

หลังจากเตรียมข้อมูลแล้ว จึงแบ่งชุดข้อมูลออกเป็น 2 ส่วน เพื่อใช้เป็นชุดข้อมูลฝึกฝน และชุดทดสอบ โดยเลือกใช้ข้อมูลในปี 2005 เป็นข้อมูลฝึกฝน เพื่อสร้างกฎการพยากรณ์ ด้วยระดับค่าสนับสนุนที่เพิ่มขึ้นเรื่อยๆ และจะหยุดเมื่อมีค่าความถูกต้อง (Accuracy) ลดลง โดยบันทึกกฎชุดนี้ไว้เพื่อใช้ในการพยากรณ์ข้อมูลชุดทดสอบในปี 2006

6. สรุปผล

งานวิจัยนี้ได้พัฒนาอัลกอริทึมไบเดคที่มีประสิทธิภาพทางด้านความเร็วในการสร้างกฎความสัมพันธ์จากการสืบค้นรูปแบบลำดับเหตุการณ์ในข้อมูลลำดับเหตุการณ์เชิงกลุ่ม โดยดัดแปลงวิธีการของอัลกอริทึมไบด์ ที่มีความสามารถสูงในการสืบค้นหาข้อมูลลำดับเหตุการณ์เชิงเดี่ยว ผลการทดสอบประสิทธิภาพการสืบค้นรูปแบบลำดับเหตุการณ์ในข้อมูลลำดับเหตุการณ์เชิงกลุ่ม ด้วยอัลกอริทึมไบเดค เปรียบเทียบกับอัลกอริทึมโคลด์สะแพน ด้วยข้อมูลชุดเดียวกัน พบว่าได้ผลลัพธ์กฎความสัมพันธ์ถูกต้องตรงกันในทุกกลุ่มข้อมูลที่ทดสอบ นอกจากนั้นอัลกอริทึมไบเดคยังใช้เวลาในการทำงานน้อยกว่าอัลกอริทึมโคลด์สะแพน ด้วยการทดสอบสถิติเปรียบเทียบค่าเฉลี่ยของเวลาในการสืบค้นที่ระดับนัยสำคัญ 0.05

นอกจากนี้ผู้วิจัยได้นำอัลกอริทึมไบเดคไปประยุกต์ใช้กับงานด้านการพยากรณ์อากาศ โดยการใช้ข้อมูลสภาพอากาศที่มีความสัมพันธ์กันย้อนหลัง 3 วัน ทำให้สามารถพยากรณ์อุณหภูมิในวันถัดไปได้ ด้วยค่าความถูกต้อง 82% และ 79% ในชุดข้อมูลฝึกฝนและชุดทดสอบ ตามลำดับ

7. ข้อเสนอแนะ

งานวิจัยนี้ได้ทำการทดลองเปรียบเทียบประสิทธิภาพกับการสืบค้นด้วยอัลกอริทึมโคลด์สะแพนเพียงอย่างเดียว หากได้ทำการเปรียบเทียบกับวิธีการอื่น ๆ จะทำให้เกิดความมั่นใจมากขึ้น และ ข้อมูลอากาศที่นำมาประยุกต์ใช้เป็นข้อมูลเฉพาะถิ่น และอาจเปลี่ยนแปลงตามสภาพของภูมิประเทศ ซึ่งสามารถนำเทคโนโลยีสารสนเทศทางภูมิศาสตร์ มาใช้ให้เกิดประโยชน์มากขึ้นได้ นอกจากนั้นการพยากรณ์ที่ทดสอบครั้งนี้ เป็นการพยากรณ์ระยะสั้น ควรศึกษาวิเคราะห์ปัจจัยสำคัญที่มีอิทธิพลในภาพรวมและอาจขยายผลเพื่อการพยากรณ์ระยะปานกลางได้

8. กิตติกรรมประกาศ

ขอขอบคุณรองคณบดีมหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี ที่อนุญาตให้ข้อมูลในการศึกษาครั้งนี้ เป็นอย่างสูง

เอกสารอ้างอิง

- 1 J. Han and M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, Second Edition, 2006.
- 2 ดวงทอง วัตรจิกรุด และ กฤษณะ ไวยมัย, “การค้าไมน์นิ่ง: เทคนิคการสืบค้นรูปแบบลำดับเหตุการณ์บนฐานข้อมูลขนาดใหญ่” การประชุมวิชาการเรื่องวิทยาการคอมพิวเตอร์และวิศวกรรม สาาสตร์ นานาชาติ ครั้งที่ 6, มหาวิทยาลัยเชียงใหม่, เชียงใหม่, 2544, หน้า 203-211.
- 3 J. Wang and J. Han “BIDE:Efficient Mining of Frequent Closed Sequences” Proceedings of the 20th International Conference of Data Engineering (ICDE'04), Boston, USA., 2004. pp. 79-90.
- 4 X. Yan, J. Han and R. Afshar, “CloSpan: Mining Closed Sequential Patterns in Large Datasets” In Proceedings of the 3rd SIAM international Conference on Data Mining (SDM'03) San Francisco,CA., USA, 2003, pp.166-177,
- 5 R. Agrawal and R. Srikant, “Mining Sequential Patterns” Proceeding of the International Conference on Data Engineering (ICDE'95), Taipei, Taiwan.1995, pp. 101-112.
- 6 P. Tzvetkov, X. Yan and J. Han “TSP: Mining Top-K Closed Sequential Patterns” In Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM'03), Florida,USA, 2003, pp. 347-354.
- 7 M. Behrouz and B. L. Seyed, “Correlation Mining between Time Series Stream and Event Stream” In Proceedings of the 4th International Conference on Networked Computing and Advanced Information Management, IEEE Xplore, 2008, pp. 333-338.
- 8 P. Siripitayananon, H.C. Chen, and K.R. Jin, “A Modified Nearest Neighbors Approach for Estimating Missing Wind Data” International Journal of Smart Engineering System Design, Vol. 5, 2003, pp. 149-160.