

1.1 บทสรุปสำหรับผู้บริหาร (Executive Summary)

ข้อมูลที่อยู่รอบ ๆ ตัวและเกี่ยวข้องกับชีวิตประจำวันของเรานั้น มีอยู่ในรูปของอนุกรมเวลา (Time Series) เป็นจำนวนมาก ซึ่งได้นำไปสู่ความสนใจที่เพิ่มขึ้นในแวดวงของ ฐานข้อมูล (Database) การค้นคืนข้อมูลอนุกรมเวลา (Time Series Retrieval) และการทำเหมืองข้อมูล (Data Mining) แต่โดยทั่วไปแล้ว เมื่อกล่าวถึงข้อมูลอนุกรมเวลา คนส่วนใหญ่มักจะนึกถึงข้อมูลทางการเงิน เช่น ข้อมูลตลาดหุ้นเป็นอันดับต้น ๆ และข้อมูลเหล่านั้นมักจะถูกนำมาใช้ในการทำนาย (Prediction) การหาแนวโน้มของข้อมูล (Trend) การวิเคราะห์การถดถอย (Regression Analysis) สหสัมพันธ์อัตโนมัติ (Autocorrelation) หรือการวิเคราะห์ทางสถิติต่าง ๆ

อย่างไรก็ตาม จริง ๆ แล้วข้อมูลอนุกรมเวลานั้นมีความเกี่ยวข้องกับชีวิตเรามากกว่านั้น เนื่องจากข้อมูลอนุกรมเวลานั้น ได้จากการเก็บข้อมูลใด ๆ ที่ทำต่อเนื่องกันเป็นลำดับเวลาก่อนหลัง และทราบได้ที่มนุษย์เรามีการตรวจวัดสิ่งต่าง ๆ ซึ่งสิ่งเหล่านี้มักมีการเปลี่ยนแปลงไปตามกาลเวลา ข้อมูลเหล่านี้ก็สามารถแสดงได้ในรูปของอนุกรมเวลาเช่นเดียวกัน ตัวอย่างเช่น ข้อมูลอัตราการเต้นของหัวใจ/คลื่นไฟฟ้าหัวใจ ข้อมูลปริมาณน้ำตาในเลือด ความดันโลหิต อุณหภูมิของร่างกาย รายได้จากการขายบัตรชมภาพยนตร์ในแต่ละสัปดาห์ คะแนนการหยั่งเสียงความนิยมของนักการเมือง ปริมาณน้ำฝนต่อปี ค่าความกดอากาศ ข้อมูลค่าความเป็นกรดต่างของน้ำทะเล อัตราการเติบโตทางการตลาด ฯลฯ และในปัจจุบันข้อมูลอนุกรมเวลา ก็ได้เข้ามามีบทบาทที่สำคัญกับงานวิจัยในศาสตร์แขนงต่าง ๆ เช่น ชีวสารสนเทศศาสตร์ (Bioinformatics) วิทยาการหุ่นยนต์ (Robotics) การแพทย์ (Medicine) เคมี (Chemistry) การรู้จำท่าทาง (Gesture Recognition) การสะกดรอย (Tracking) การเงิน (Finance) ชีวมาตร (Biometrics) ดาราศาสตร์ (Astronomy) อุตสาหกรรม (Manufacturing) และอื่น ๆ อีกมากมาย ข้อมูลเหล่านี้วันยังมีการเก็บข้อมูลที่เพิ่มขึ้นอย่างมหาศาล ทำให้การค้นหาข้อมูล หรือนำข้อมูลเหล่านั้นมาหาความสัมพันธ์กันอย่างมีประสิทธิภาพแทบจะไม่มีทางเป็นไปได้

ปัจจุบันงานวิจัยด้านการทำเหมืองข้อมูลในข้อมูลอนุกรมเวลา มีความสำคัญเพิ่มมากขึ้นเนื่องจากความแพร่หลายของการนำไปใช้งานในด้านต่าง ๆ เช่น การตรวจคลื่นไฟฟ้าหัวใจในงานทางด้าน การแพทย์ [3] การลดมิติของข้อมูลอนุกรมเวลา [11] การตรวจจับสิ่งผิดปกติที่เกิดขึ้นในข้อมูลอนุกรมเวลา [2] และการพยากรณ์ข้อมูลตลาดหุ้น [4] เป็นต้น การค้นพบโมทีฟของข้อมูลอนุกรมเวลา เป็นอีกสาขาหนึ่งของงานวิจัย ซึ่งเน้นการค้นหารูปแบบเริ่มต้นที่เกิดขึ้นซ้ำกันในข้อมูลอนุกรมเวลา เพื่อนำรูปแบบที่ค้นพบนั้น ๆ เข้าไปสู่กระบวนการถัดไปของการวิจัย เช่น การจำแนกประเภท [10] และการจัดกลุ่มของข้อมูลอนุกรมเวลา [9] เป็นต้น

การจัดกลุ่มข้อมูล (Clustering) นับเป็นภารกิจที่สำคัญยิ่งภารกิจหนึ่งในการทำเหมืองข้อมูล ซึ่งเป็นการจัดข้อมูลออกเป็นกลุ่มย่อย ๆ โดยที่ข้อมูลที่ถูกจัดอยู่ในกลุ่มเดียวกันจะมีคุณลักษณะที่ คล้ายคลึงกัน ในด้านการทำเหมืองข้อมูลอนุกรมเวลานั้น การบ่งชี้ความคล้ายกันของข้อมูลมักใช้วิธีการวัดระยะทางของข้อมูล (Distance Measure) เช่นมาตรวัดระยะทางแบบ ไดนามิกไทม์วอร์ปิง (Dynamic Time Warping Distance Measure) ซึ่งมีความแม่นยำสูง หากแต่วิธีนี้ใช้เวลาในการคำนวณที่สูงมาก การนำเอาไดนามิกไทม์วอร์ปิงมาใช้ในการจัดกลุ่มข้อมูลแบบกระแสด จึงมีความ

ท้าทายในตัวเองเป็นอย่างมาก และจำเป็นต้องมีการพัฒนาต่อ หากแต่เมื่อไม่นานมานี้ ได้มีงานวิจัยที่พิสูจน์ให้เห็นว่าการจัดกลุ่มลำดับย่อย (Subsequence Clustering) สำหรับข้อมูลอนุกรมเวลาในปัจจุบันนั้นมีความผิดพลาด (หรืออีกนัยหนึ่ง คือ ไร้ความหมาย) บางงานวิจัยได้พิจารณาว่าความผิดพลาดนั้นเกิดจากการใช้มาตรวัดที่ไม่เหมาะสม จึงได้มีกลุ่มนักวิจัยมากมายที่พยายามเสนอมาตรวัดใหม่ ที่พยายามแก้ไขปัญหาคือผิดพลาดดังกล่าว แต่จนถึงปัจจุบัน ก็ยังไม่มีงานวิจัยใดที่สามารถแก้ไขปัญหานี้ได้อย่างถูกต้องและสมบูรณ์ ซึ่งนับได้ว่าปัญหานี้เป็นปัญหาที่มีผู้ให้ความสำคัญและมีความท้าทายเป็นอย่างมาก

นอกจากนี้ข้อมูลแบบกระแส (Streaming Data) นับเป็นข้อมูลที่มีอยู่โดยทั่วไป และยังมีเพิ่มมากขึ้นตลอดเวลา หากแต่เครื่องมือที่ใช้ในการทำเหมืองข้อมูลสำหรับข้อมูลแบบกระแสในปัจจุบัน ยังมีอยู่อย่างจำกัด ทั้งในแง่ของประสิทธิภาพ ที่ไม่สามารถรองรับข้อมูลขนาดใหญ่ได้ดั่งนัก ความเร็ว และความถูกต้องแม่นยำ งานวิจัยนี้จึงมีจุดประสงค์ที่จะต่อยอดงานวิจัยไปสู่การคิดค้นวิธีการจัดกลุ่มลำดับย่อยสำหรับข้อมูลอนุกรมเวลาแบบกระแสอย่างมีความหมายและแม่นยำ สืบเนื่องจากงานวิจัยที่ได้ทำการพิสูจน์ให้นักวิจัยในวงการเห็นแล้วว่า การจัดกลุ่มลำดับย่อยสำหรับข้อมูลอนุกรมเวลาแบบกระแสที่ใช้กันอยู่ในปัจจุบันนั้นมีความผิดพลาด ซึ่งได้ส่งผลโดยตรง กับอัลกอริทึมที่ใช้การจัดกลุ่มลำดับย่อยสำหรับกระแสข้อมูลอนุกรมเวลาเป็นส่วนหนึ่งนั้น เกิดข้อผิดพลาดตามไปด้วย ดังนั้น องค์ความรู้ที่ได้จากงานวิจัยนี้ จะสามารถนำไปใช้ทดแทนการจัดกลุ่มแบบเดิม และนักวิจัยอื่น ๆ ยังสามารถใช้อีกความรู้ใหม่นี้ เพื่อต่อยอดในการวิจัยด้านการทำเหมืองข้อมูลอนุกรมเวลาต่อไปได้อย่างต่อเนื่อง

การดำเนินโครงการวิจัยโดยรวม เริ่มจากการปรับปรุงการจัดกลุ่มข้อมูลอนุกรมเวลาโดยใช้วิธีการจัดกลุ่มตามรูปร่าง (Shape-based Time Series Clustering) ซึ่งนำอัลกอริทึมสำหรับเฉลี่ยข้อมูลอนุกรมเวลาแบบ Cubic-Spline Interpolation มาใช้ในการหาตัวแทนของกลุ่มข้อมูล และนำระยะไดนามิกไทม์วอร์ปิงมาใช้แทนระยะยุคลิด เพื่อเพิ่มความแม่นยำในการจัดกลุ่มข้อมูลแบบ K-Means จากนั้นเป็นส่วนของการหาโมทีฟความยาวแปรผันของข้อมูลอนุกรมเวลา (Variable-length Time Series Motif Discovery) ซึ่งเป็นวิธีการหาโมทีฟหรือรูปแบบของข้อมูลอนุกรมเวลา โดยที่ไม่ต้องกำหนดความยาวของโมทีฟล่วงหน้า ดังเช่นงานวิจัยที่ผ่านมา ทั้งสองส่วนนี้ มีความสำคัญและจะช่วยให้ต่อยอดไปสู่การหาวิธีจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาทั้งแบบข้อมูลคงที่และแบบกระแสอย่างมีความหมายและแม่นยำ ที่สามารถหารูปแบบซ้ำๆ กันของข้อมูลอนุกรมเวลาใดๆ และจัดกลุ่มให้ผู้ใช้ โดยเฉพาะนักวิเคราะห์ทางข้อมูล นำไปใช้ในการวิเคราะห์เฉพาะทางได้ ซึ่งในงานวิจัยนี้ได้นำเสนอการใช้เทคนิค ที่ผสมผสานการวัดระยะทางแบบไดนามิกไทม์วอร์ปิง และการหาค่าเฉลี่ยตามรูปเพื่อให้ได้กลุ่มข้อมูลและรูปแบบที่มีความถูกต้องมากที่สุด

1.2 วัตถุประสงค์ของการวิจัย

งานวิจัยนี้ มีจุดประสงค์หลัก เพื่อคิดค้นวิธีการจัดกลุ่มลำดับย่อยสำหรับข้อมูลอนุกรมเวลาแบบกระแสอย่างมีความหมายและแม่นยำ โดยเป็นการเริ่มจากงานวิจัย สองส่วนย่อย ๆ คือ การจัดกลุ่มข้อมูลอนุกรมเวลาตามรูปร่าง (Shape-based Time Series Clustering) และ การหาโมทีฟความยาวแปรผันของข้อมูลอนุกรมเวลา (Variable-length Time Series Motif Discovery) ซึ่งเป็นสองส่วนสำคัญที่จะนำไปสู่การคิดค้นวิธีการจัดกลุ่มลำดับย่อยสำหรับข้อมูลอนุกรมเวลาอย่างมีความหมายและแม่นยำในที่สุด

1.3 การนำผลงานวิจัยไปใช้ประโยชน์

แม้ว่างานวิจัยนี้ ยังไม่ได้มีการนำผลงานไปใช้ประโยชน์อย่างเป็นทางการเป็นรูปธรรมในเชิงพาณิชย์หรือเชิงนโยบาย แต่ได้เกิดผลประโยชน์ในเชิงสาธารณะและเชิงวิชาการ กล่าวคือ เนื่องจากงานวิจัยนี้ได้เสนอการจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาแบบกระแส ซึ่งถือได้ว่าเป็นงานวิจัยแรกของโลกที่สามารถแก้ปัญหานี้ได้ นับตั้งแต่การค้นพบโดยนักวิจัยเมื่อ 8 ปีที่ผ่านมาว่างานวิจัยเกี่ยวกับเรื่องนี้มีการใช้งานที่ผิดมาตลอด และได้ผลลัพธ์ที่ไม่ถูกต้อง หลังจากการค้นพบครั้งใหญ่นั้น มีนักวิจัยหลายคนจากทั่วโลกพยายามแก้ปัญหานี้มาอย่างต่อเนื่อง แต่ก็ไม่มีใครทำสำเร็จและถูกต้อง ซึ่งงานวิจัยนี้ได้เสนอวิธีการแก้ปัญหาแบบใหม่ทำให้ได้วิธีการจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาแบบกระแสวิธีแรกของโลก จึงเป็นการจุดประกายและสร้างกระแสความสนใจในวงกว้าง นอกจากนี้ยังได้มีการวิจัยต่อยอดกับนักวิจัยและนิสิตอีกหลากหลายผลงาน

1.4 ระเบียบวิธีดำเนินการวิจัย

- (1) ปรับปรุงอัลกอริทึมในการหาตัวแทนกลุ่ม ตามรูปร่างของลำดับย่อยสำหรับข้อมูลอนุกรมเวลา
- (2) คิดค้นวิธีการจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาอย่างมีความหมาย โดยพิจารณาที่มาตรวัดต่าง ๆ ซึ่งน่าจะเป็นสาเหตุหลักที่ทำให้อัลกอริทึมในปัจจุบันมีปัญหาและข้อผิดพลาด จากนั้นพยายามคิดค้นขั้นตอนวิธีที่สามารถจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาได้อย่างมีความหมายและถูกต้องแม่นยำ
- (3) การหาตัวแทนกลุ่มข้อมูล ตามรูปร่างของข้อมูลลำดับย่อยสำหรับข้อมูลอนุกรมเวลาแบบกระแส และหาวิธีการจัดกลุ่มลำดับย่อยสำหรับข้อมูลอนุกรมเวลาแบบกระแสอย่างมีความหมาย
- (4) ทำการทดลอง ประเมินผล และปรับปรุงแก้ไขเทคนิคและอัลกอริทึม
- (5) ตีพิมพ์ผลงานวิจัย

1.5 ผลงานตีพิมพ์ในวารสารวิชาการนานาชาติ

- Sura Rodpongpun, Vit Niennattrakul, Chotirat Ann Ratanamahatana: Selective Subsequence Time Series clustering. Knowledge-Based Systems. 35: 361-368 (2012)
- Vit Niennattrakul, Dararat Srisai, Chotirat Ann Ratanamahatana: Shape-based template matching for time series data. Knowledge-Based Systems. 26: 1-8 (2012)
- Sura Rodpongpun, Vit Niennattrakul, and Chotirat Ann Ratanamahatana, Efficient Subsequence Search on Streaming Data Based on Time Warping Distance. ECTI Transactions on Computer and Information Technology Vol.5 (1) May 2011, pp. 1-7.

1.6 ผลงานตีพิมพ์และการเสนอผลงานในที่ประชุมวิชาการระดับนานาชาติ

- Warissara Meesrikamulkul, Vit Niennattrakul, and Chotirat Ann Ratanamahatana, Shape-based Clustering for Time Series Data. In Proceedings of 16th Pacific-Asia Conference, PAKDD 2012, Kuala Lumpur, Malaysia, May 29-June 1, 2012.

- Sorrachai Yingchareonthawornchai, Haemwaan Sivaraks, Sura Rodpongpun, and Chotirat Ann Ratanamahatana. 2012. The Proper Length Motif Discovery Algorithm. In Proceedings of the 16th International Computer Science and Engineering Conference (ICSEC'12), October 17-19, 2012, Thailand.
- Warissara Meesrikamolkul, Vit Niennattrakul and Chotirat Ann Ratanamahatana, Multiple Shape-based Templates Matching for Time Series Data. In Proceedings of 8th ECTI-CON 2011, May 17-19, 2011, Thailand.
- Pawan Nunthanid, Vit Niennattrakul, and Chotirat Ann Ratanamahatana, Discovery of Variable Length Time Series Motif. In Proceedings of 8th ECTI-CON 2011, May 17-19, 2011, Thailand.