บทคัดย่อ

รหัสโครงการ: MRG5380130

ชื่อโครงการ: การจัดกลุ่มกระแสข้อมูลอนุกรมเวลาอย่างมีความหมายและแม่นยำ

ผู้วิจัย: ผู้ช่วยศาสตราจารย์ โชติรัตน์ รัตนามหัทธนะ และคณะ
ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

อีเมล์: chotirat@gmail.com, chotirat.r@chula.ac.th

ระยะเวลาโครงการ: 2 ปี

บทคัดย่อ:

การจัดกลุ่มลำดับย่อยสำหรับข้อมูลอนุกรมเวลาแบบกระแส เป็นหนึ่งในปัญหาที่ท้าทายมากที่สุดของการทำเหมืองข้อมูลอนุกรมเวลา ตั้งแต่การจัดกลุ่มลำดับย่อยได้ถูกแสดงให้เห็นว่า การจัดกลุ่มจะให้คำตอบที่ไร้ความหมายในเชิงการทดลอง และทฤษฎี การจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาที่ถูกใช้ในหลายร้อยงานวิจัยนั้นจะให้คลื่นไซน์เป็นตัวแทนกลุ่มเสมอ ถ้าให้ข้อมูลอนุกรมเวลาหนึ่ง ๆ การจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาควรคืนค่าตัวแทนกลุ่มที่เป็นลักษณะของทุกลำดับย่อยในข้อมูลอนุกรมเวลา ส่วนสาเหตุที่ทำให้เกิดความไร้ความหมาย ถูกระบุไว้เป็นสองสาเหตุได้แก่ การใช้ระยะทางยุคลิดเป็นตัววัดระยะทางที่ไม่เหมาะสม และการใช้การเฉลี่ยค่าตามแอมพลิจูดเป็นฟังก์ชันการเฉลี่ยที่ไม่เหมาะสม เพื่อที่จะได้มาซึ่งคำตอบของการจัดกลุ่มที่มีความหมาย ในงานวิจัยนี้ได้เสนอการจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาตามรูป โดยใช้ระยะทางไดนามิกไทม์วอร์ปปิงและการเฉลี่ยค่าตามรูปแทนระยะทางยุคลิด และการเฉลี่ยค่าตามแอมพลิจูดตามลำดับ ดังนั้นการจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาตามรูปจะคืนผลลัพธ์ที่มีความหมายที่มากกว่าการจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาแบบเดิม แต่อย่างไรก็ตาม การจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาตามรูปไม่สามารถประยุกต์ใช้กับข้อมูลแบบกระแสได้โดยตรง เนื่องจากการจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาตามรูปใช้เวลาในการประมวลผลนาน โดยคำนวณลำดับย่อยที่ผ่านมาทั้งหมดเมื่อมีจุดข้อมูลใหม่เข้ามา งานวิจัยนี้จึงได้เสนอการจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาแบบกระแสตามรูป ให้รองรับกรณีข้อมูลแบบกระแส โดยคำนวณบนชุดข้อมูลขนาดเล็กของลำดับย่อยที่เก็บไว้ แทนที่จะคำนวณจากลำดับย่อยทั้งหมด ซึ่งชุดข้อมูลของลำดับย่อยที่เก็บไว้ถูกปรับปรุงสำหรับทุก ๆ จุดข้อมูล เพื่อรักษาจำนวนลำดับย่อยในชุดข้อมูล ไม่ให้เกินกว่าจำนวนมากสุดที่อนุญาต ดังนั้นการจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาแบบกระแสตามรูป จึงเร็วกว่าการจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาตามรูปอย่างมาก

คำหลัก : การทำเหมืองข้อมูล, การจัดกลุ่มลำดับย่อย, อนุกรมเวลา, ข้อมูลแบบกระแส

Abstract

Project Code : MRG5380130

Project Title : Meaningful and Accurate Subsequence Clustering for Time Series Data Stream

Investigator : Assistant Professor Chotirat Ratanamahatana, Ph.D.
Dept. of Computer Engineering, Faculty of Engineering, Chulalongkorn University

E-mail Address : chotirat@gmail.com, chotirat.r@chula.ac.th

Project Period : 2 years

Abstract:

Subsequence clustering for time series data streams is one of the most challenging issues of time series data mining since subsequence clustering has been proven both theoretically and empirically that it produces meaningless clustering results, where hundreds of research works that utilize STSC as a preprocessing step and a subroutine are all affected. Given a time series s equence, subsequence clustering should return cluster representatives which represent characteristics of all subsequences in time series. Therefore, if cluster representatives are always sine waves regardless of inputs, clustering results are meaningless since they do not reflect characteristics of the subsequences. The causes of meaninglessness are identified in twofold, i.e., inappropriate uses of Euclidean distance as a distance measure and Amplitude Averaging as an averaging function. To achieve meaningful clustering results, in this research, Shape-based Subsequence Time Series Clustering (2STSC) is proposed to use Dynamic Time Warping (DTW) distance measure and Shape-based Averaging function. Therefore, 2STSC returns more meaningful results than those from STSC. However, 2STSC cannot directly apply to data streams since 2STSC consumes large computational complexity by considering all previous subsequences for every new incoming data point. Shape-based Streaming Subsequence Time Series Clustering (3STSC) is then proposed to handle the streaming case by calculating a clustering result on a small set of stored subsequences instead of calculating from all previous subsequences. The small set of stored subsequences is updated for every new incoming data point to maintain the number of stored subsequences not to exceed the maximum allowance number. 3STSC, therefore, is much faster than 2STSC, while 3STSC returns small distortions of clustering results.

Keywords : Data Mining, Subsequence Clustering, Time Series, Data Stream