การจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาแบบกระแสอย่างมีความหมาย

นายวิชญ์ เนียรนาทตระกูล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2553

การจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาแบบกระแสอย่างมีความหมาย

นายวิชญ์ เนียรนาทตระกูล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2553

MEANINGFUL SUBSEQUENCE CLUSTERING FOR TIME SERIES DATA STREAM

Mr. Vit Niennattrakul

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Computer Engineering
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2010

Thesis Title	MEANINGFUL SUBSEQUENCE CLUSTERING FOR TIME SERIES DATA STREAM
By	Mr. Vit Niennattrakul
Field of Study	Computer Engineering
Thesis Advisor	Assistant Professor Chotirat Ann Ratanamahatana, Ph.D.

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctoral Degree

.... Dean of the Faculty of Engineering

(Associate Professor Boonsom Lerdhirunwong, Dr.Ing.)

THESIS COMMITTEE

.... Chairman

(Professor Boonserm Kijsirikul, Ph.D.)

.... Thesis Advisor

(Assistant Professor Chotirat Ann Ratanamahatana, Ph.D.)

.... Examiner

(Professor Prabhas Chongstitvattana, Ph.D.)

.... Examiner

(Assistant Professor Sukree Sinthupinyo, Ph.D.)

Charnyote Pluempitiwiriyawej

.... External Examiner

(Assistant Professor Charnyote Pluempitiwiriyawej, Ph.D.)

วิชญ์ เนียรนาทตระกูล: การจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาแบบกระแสอย่างมีความหมาย. (MEANINGFUL SUBSEQUENCE CLUSTERING FOR TIME SERIES DATA STREAM) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ผู้ช่วยศาสตราจารย์ ดร. โชติรัตน์ รัตนามหัทธนะ, 192 หน้า.

การจัดกลุ่มลำดับย่อยสำหรับข้อมูลอนุกรมเวลาแบบกระแสเป็นหนึ่งในปัญหาที่ท้าทายมากที่สุดของการทำเหมืองข้อมูลอนุกรมเวลาตั้งแต่การจัดกลุ่มลำดับย่อยได้ถูกแสดงให้เห็นว่าการจัดกลุ่มจะให้คำตอบที่ไร้ความหมายในเชิงการทดลองและทฤษฎี การจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาที่ถูกใช้ในหลายร้อยงานวิจัยนั้นจะให้คลื่นไซน์เป็นตัวแทนกลุ่มเสมอ ถ้าให้ข้อมูลอนุกรมเวลาหนึ่ง ๆ การจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาควรคืนค่าตัวแทนกลุ่มที่เป็นลักษณะของทุกลำดับย่อยในข้อมูลอนุกรมเวลา สาเหตุที่ทำให้เกิดความไร้ความหมายถูกระบุไว้มาจากสองสาเหตุได้แก่ การใช้ระยะทางยุคลิดเป็นตัววัดระยะทางที่ไม่เหมาะสมและการใช้การเฉลี่ยค่าตามแอมพลิจูดเป็นฟังก์ชันการเฉลี่ยที่ไม่เหมาะสม เพื่อที่จะได้มาซึ่งคำตอบของการจัดกลุ่มที่มีความหมาย ในวิทยานิพนธ์นี้การจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาตามรูปได้ถูกเสนอโดยใช้ระยะทางไดนามิกไทม์วอร์ปปิงและการเฉลี่ยค่าตามรูปแทนระยะทางยุคลิดและการเฉลี่ยค่าตามแอมพลิจูดตามลำดับ ดังนั้นการจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาตามรูปจะคืนผลลัพธ์ที่มีความหมายที่มากกว่าการจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาแบบเดิม แต่อย่างไรก็ตามการจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาตามรูปไม่สามารถประยุกต์ใช้กับข้อมูลแบบกระแสได้ เนื่องจากการจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาตามรูปใช้เวลาในการประมวลผลนานโดยคำนวณลำดับย่อยที่ผ่านมาทั้งหมดเมื่อมีจุดข้อมูลใหม่เข้ามา การจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาแบบกระแสตามรูปจึงถูกเสนอให้รองรับกรณีข้อมูลแบบกระแสโดยคำนวณบนชุดข้อมูลขนาดเล็กของลำดับย่อยที่เก็บไว้แทนที่จะคำนวณจากลำดับย่อยทั้งหมด ซึ่งชุดข้อมูลของลำดับย่อยที่เก็บไว้ถูกปรับปรุงสำหรับทุกๆจุดข้อมูลเพื่อรักษาจำนวนลำดับย่อยในชุดข้อมูลไม่ให้เกินกว่าจำนวนมากสุดที่อนุญาต ดังนั้นการจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาแบบกระแสตามรูปจึงเร็วกว่าการจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาตามรูปอย่างมาก

ภาควิชา ....วิศวกรรมคอมพิวเตอร์....... ลายมือชื่อนิสิต ....

สาขาวิชา.......วิศวกรรมคอมพิวเตอร์....... ลายมือชื่ออ.ที่ปรึกษาวิทยานิพนธ์หลัก ....

ปีการศึกษา ....2553 ....

4971828021: MAJOR COMPUTER ENGINEERING

KEYWORDS: DATA MINING / SUBSEQUENCE CLUSTERING / TIME SERIES / DATA STREAM

VIT NIENNATTRAKUL : MEANINGFUL SUBSEQUENCE CLUSTERING FOR TIME SERIES DATA STREAM. ADVISOR : ASSISTANT PROFESSOR CHOTIRAT RATANAMAHATANA, PH.D., 192 pp.

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 Subsequence Time Series Clustering (STSC) as a preprocessing step and a subroutine are all affected. Given a time series sequence, 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 thesis, 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. 3STSC, therefore, is much faster than 2STSC, while 3STSC returns small distortions of clustering results.

Department: ..... Computer Engineering ..... Student's Signature ....

Field of Study: .... Computer Engineering .... Advisor's Signature ....

Academic Year: .... 2010 ....

Acknowledgments

I would like to express my sincere gratitude to my thesis advisor, Dr. Chotirat Ann Ratanamahatana for her invaluable guidance and support during my graduate studies at Chulalongkorn University. She always encourages my progress, new ideas, as well as motivates my research work. She is always full of energy and untiringly available for giving me insightful advices. In addition, I have learned precious lessons through her past and present remarkable research work, as well as her exceptional presentation. I truly consider it a great privilege in having the opportunity to work with her as my graduate advisor.

I am grateful to Prof. Eamonn Keogh who supported me when I was in the United States. I also express my thankfulness to my dissertation committee: Dr. Prabhas Chongstitvatana, Dr. Boonserm Kitsirikul, Dr. Sukree Sinthupinyo, and Dr. Charnyote Pluempitiwiriyawej. I am indebted to every teacher, especially, Dr. Proadpran, Dr. Atiwong, Dr. Athasit, Dr. Pizzanu, Dr. Vishnu, Dr. Somchai, Ajarn Mandhana for introducing me a rabbit hole of computer engineering for nine years since my undergraduate years. I would like to thank my lovely friends, Soung, Nart, Aim, Ji, Ping, Heng, Ton, Nui, Yong, Jen, Rote, Pick, Pong, Kwang, Guk, N'Pam, N'Bird, N'Bim, N'Pao, N'Au, N'Rong, N'Pop, N'Pun, Kook, Tohn, Poo, P'Lin, P'Komate, P'Woon, P'Petch, P'Jung, P'Yui, P'Nan, P'Noot, P'Ae, P'O, P'Woot, and P'Ko for their wonderful friendships, encouragement, and many valuable discussions. I also would like to thank Took, Moo, Nakorn, P'Tookta, P'Lek, P'Fad, P'Pang, P'Art, P'Dao, P'Hao, P'May, P'Pui, P'Tee-Guay, P'Ae, P'O, N'Parn, N'Neoy, N'Ping, N'Nanah, and N'Giffy for making my stay in the U.S. wonderful; without them, my visit would be colorless. Additionally, I am thankful to the administrative staffs for always being helpful during my whole time attending the Department of Computer Engineering.

I greatly appreciate the financial support from the Thailand Research Fund given through the Royal Golden Jubilee Ph.D. Program (PHD/0141/2549) and the Chulalongkorn University Graduate Scholarship to Commemorate the 72^{nd} Anniversary of His Majesty King Bhumibol Adulyadej for giving me the invaluable opportunity and providing financial support during my precious a half decade of years studying in Ph.D. program and going abroad to the United States. I also greatly appreciate the research fund from the 90^{th} Anniversary of Chulalongkorn University Fund (Ratchadaphiseksomphot Endowment Fund) to make my research significantly forward and my idea becomes reality.

Finally, with my utmost gratitude, this dissertation is dedicated to my beloved parents for shaping my life for what it is today, and to my sisters for always being there. Without their love, encouragement, understanding, and support, this research could not have been completed.

Contents

Page

List of Tables

List of Figures

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page