การหารูปแบบความสัมพันธ์ของลำดับมรดกจากฐานข้อมูลราชการ

นายโกเมศ อัมพวัน

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2553
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

การหารูปแบบความถี่สม่ำเสมอเคอันดับแรกจากฐานข้อมูลรายการ

นายโกเมศ อัมพวัน

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2553

MINING TOP-K REGULAR-FREQUENT ITEMSETS FROM TRANSACTIONAL DATABASE

Mr. Komate Amphawan

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	MINING TOP-K REGULAR-FREQUENT ITEMSETS FROM TRANSACTIONAL DATABASE
By	Mr. Komate Amphawan
Field of Study	Computer Engineering
Thesis Advisor	Assistant Professor Athasit Surarerks, 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
(Associate Professor Wanchai Rivepiboon, Ph.D.)

..... Thesis Advisor
(Assistant Professor Athasit Surarerks, Ph.D.)

..... Examiner
(Assistant Professor Pizzanu Kanongchaiyos, Ph.D.)

..... External Examiner
(Assistant Professor Krisana Chinnasarn, Ph.D.)

..... External Examiner
(Assistant Professor Arnon Rungsawang, Ph.D.)

โกเมศ อัมพวัน: การหารูปแบบความถี่สม่ำเสมอเคอันดับแรกจากฐานข้อมูลรายการ. (MINING TOP-K REGULAR-FREQUENT ITEMSETS FROM TRANSACTIONAL DATABASE) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ผู้ช่วยศาสตราจารย์ อรรถสิทธิ์ สุรฤกษ์, 197 หน้า.

การหากฎความสัมพันธ์ภายใต้ค่าสนับสนุนและค่าความเชื่อมั่นที่ผู้ใช้ระบุเป็นงานสำคัญของกลุ่มวิจัยด้านการทำเหมืองข้อมูล อย่างไรก็ตามความถี่ของการเกิดข้อมูลในรูปแบบต่างๆอาจไม่เพียงพอที่จะเป็นเงื่อนไขของการหารูปแบบที่มีความหมาย ดังนั้นลักษณะการเกิดของรูปแบบที่มีความสัมพันธ์กันจึงเป็นประเด็นสำคัญของงานในหลายๆด้าน รูปแบบหนึ่งของการเกิดความสัมพันธ์ของข้อมูลก็คือการเกิดขึ้นอย่างสม่ำเสมอ นั่นคือข้อมูลจะปรากฏขึ้นห่างกันเป็นช่วงๆในระยะห่างที่ผู้ใช้ให้ความสำคัญ ซึ่งในการหาข้อมูลที่มีการเกิดขึ้นในลักษณะดังกล่าวจะต้องมีการกำหนดค่าสนับสนุนเพื่อใช้ในการเลือกกรองความสม่ำเสมอของข้อมูลที่เกิดขึ้น แต่ในทางปฏิบัติแล้วการกำหนดค่าขีดแบ่งที่เหมาะสมนี้ทำได้ยาก เนื่องจากถ้ากำหนดค่าขีดแบ่งนี้น้อยเกินไปจำนวนผลลัพธ์ที่ได้ก็จะมีปริมาณมากและในทางกลับกันถ้ากำหนดค่าขีดแบ่งมากผลลัพธ์ที่ได้ก็จะมีจำนวนน้อยหรืออาจจะไม่พบคำตอบที่สอดคล้องกับค่าขีดแบ่งนี้ ดังนั้นแทนที่จะกำหนดค่าขีดแบ่งนี้การกำหนดจำนวนผลลัพธ์ที่ต้องการจะเป็นการเหมาะสมกว่าจากการสำรวจงานวิจัยที่มีอยู่ในปัจจุบันพบว่าไม่มีวิธีใดที่ให้ผู้ใช้ระบุจำนวนการเกิดข้อมูลแบบสม่ำเสมอที่ผู้ใช้ต้องการ ดังนั้นวิทยานิพนธ์นี้จึงนำเสนอการใช้จำนวนผลลัพธ์เป็นเป้าหมายในการค้นหาข้อมูลที่มีการเกิดอย่างสม่ำเสมอ นั่นคือการหาผลลัพธ์ที่มีค่าสนับสนุนสูงที่สุดเคอันดับแรก ซึ่งวิธีที่วิทยานิพนธ์นี้นำเสนอมีดังนี้ (1) วิธีการหาคำตอบโดยอ่านข้อมูลจากฐานข้อมูลเพียงครั้งเดียว (2) การใช้การค้นหาแบบเลือกทางดีที่สุดเพื่อลดปริภูมิการค้นหา (3) การแบ่งข้อมูลและการประมาณค่าสนับสนุนเพื่อช่วยลดการคำนวณที่ไม่จำเป็นลง (4) การใช้วิธีการแทนข้อมูลแบบใหม่เพื่อลดเวลาและหน่วยความจำที่ใช้ในการประมวลผล ซึ่งจากการวิเคราะห์ประสิทธิภาพในเชิงเวลาและหน่วยความจำของวิธีการที่นำเสนอพบว่า ไม่ว่าจะกำหนดจำนวนของผลลัพธ์มากหรือน้อย ไม่ว่าข้อมูลจะมีการกระจุกหรือกระจายตัว ขั้นตอนวิธีที่นำเสนอนี้ก็สามารถให้ผลลัพธ์ได้อย่างมีประสิทธิภาพ

ภาควิชา .....วิศวกรรมคอมพิวเตอร์..... ลายมือชื่อนิสิต โกเมศ อัมพวัน

สาขาวิชา.....วิศวกรรมคอมพิวเตอร์..... ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา ..... 2553 .....

4871857021: MAJOR COMPUTER ENGINEERING

KEYWORDS: DATA MINING / ASSOCIATION RULES / TOP-*K* SIGNIFICANT ITEMSETS / REGULAR-FREQUENT ITEMSETS / TOP-*K* REGULAR-FREQUENT ITEMSETS

KOMATE AMPHAWAN : MINING TOP-K REGULAR-FREQUENT ITEMSETS.FROM TRANSACTIONAL DATABASE. THESIS ADVISOR : Assistant Professor Athasit Surarerks, Ph.D., 53 pp.

Association rule based on support-confident framework is an important task in data mining community. However, the occurrence frequency of a pattern may not be sufficient criterion for mining meaningful patterns. The occurrence behavior can be revealed as an important key in several applications. A pattern is a regular pattern if it regularly occurs in a user-given period (regularity threshold). To mine regular itemsets, a support threshold is used to filter some regular itemsets. However, in practice, it is often difficult for users to provide an appropriate support threshold. Indeed, a too small support threshold could yield a number of regular-frequent itemsets impractically large while a too large threshold could yield very few or no regular-frequent itemsets. Therefore, the use of a support threshold tends to produce a large number of regular-frequent itemsets and it could be better to ask for the number of desired results.

Currently, from the deep survey, there is no existing approach permitting users to specify the number of regular-frequent itemsets to be mined. Therefore, a new approach allowing the users to control the number of results (*i.e.* k regular itemsets with the highest supports) is presented. There are several techniques proposed in this dissertation to mine this kind of itemsets: (*i*) a single-scan approach, (*ii*) a best-first search strategy used to prune the search space, (*iii*) the partitioning and estimation techniques assisting in reducing unnecessary computational costs, and (*iv*) a new concise representation helping to save runtime and memory. From the performance study, the proposed approaches are efficient and scalable in terms of time and memory for small and large values of desired results on sparse and dense datasets.

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

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

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

Acknowledgements

First and foremost, I wish to thank the Higher Education Commission of Thailand who gave me a research scholarship. I would like to express my deep gratitude to my supervisor, Athasit Surarerks, Ph.D., to whom with his guidance, advices and helps me to overcome the difficulties of the process of conducting this dissertation. He patiently and thoroughly guided me for an almost 5-year period from a starting point to reading the chapters of this thesis. I also would like to express my deepest gratitude to my co-supervisor, Philippe Lenca, Ph.D. at Telecom Bretagne, France. His invaluable advice and patience allowed me to undergo and finish the challenging process of my doctoral degree program. I also express my thankfulness to my dissertation committees: Wanchai Rivepiboon, Ph.D., Pizzanu Kanongchaiyos, Ph.D., Krisana Chinnasan, Ph.D. and Arnon Rungsawang, Ph.D., for their advices and guidance to help me focus on my research activities. I am also grateful to Sunisa Rimchareon, Ph.D., and Miss Warisa Sritriratanarak whom always encourage me to do research, give me moral support and help me everything as they can do.

I also would like to thank all of my lovely friends (e.g. Pai, Phae, Te, Yring, Rin, Chan, Aui, Krit, Tuk, Kai, Kik, Camp, Oat, Zeng, Wat, Tar, Trung, Kwan, Pum, etc.) and the graduate students in Computer Engineering department (e.g. P' Chai, Koh, Pook, Pla, Woot, Wut, P' Tom, P' Chang, P' Jim, P' Phueng, P' Chit, P'Vic, Tei, Pae, Pun, Dong, Kai, Puh, Pair, Na, Pong, Dear, Oat(Elite), Oat(CG), Petch(Van), Petch(Dek), P'Keng, Tair, Chris, Tuy, Jumbo, Noot, P' Ae, P' Aoe, Matt, P' Mao,P'Aui P' Sakorn, P' Yui, Petch(Mhee), P' Dae, P' Jung, J' Nhan, P'Fu, June, Pia, Vit, etc.) whose advice and friendship gave me a lot of encouragement for my studying at Chulalongkorn University.

I would like to thank the Department of Computer Engineering of Chulalongkorn University for giving me the opportunity to pursue a Ph.D. degree in computer engineering. I am also thankful to the support staffs in Computer Engineering Department for always being helpful.

Last but not least, I am very grateful to my family: Father, Mother and my Sister for their love, patience, continuous moral support and encouragement. Without all of these I could not have accomplished my doctoral degree.

Contents

Chapter Page

Chapter Page

List of Tables

List of Figures

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page