รายงานวิจัยฉบับสมบูรณ์

โครงการ

การค้นพบโดยประมาณของรูปแบบที่ปรากฏบ่อยในข้อมูลสตรีม

Approximate Frequent Pattern Discovery over Data Stream

โดย รองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ

20 สิงหาคม 2553

สัญญ

รายงานวิจัยฉบับสมบูรณ์

โครงการ

การค้นพบโดยประมาณของรูปแบบที่ปรากฏบ่อยในข้อมูลสตรีม

Approximate Frequent Pattern Discovery over Data Stream

ผู้วิจัย

รองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ

สาขาวิชาวิศวกรรมคอมพิวเตอร์

สำนักวิชาวิศวกรรมศาสตร์

มหาวิทยาลัยเทคโนโลยีสุรนารี

สนับสนุนโดยสำนักงานคณะกรรมการการอุดมศึกษา
และสำนักงานกองทุนสนับสนุนการวิจัย

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณสำนักงานคณะกรรมการการอุดมศึกษา (สกอ.) และสำนักงานกองทุนสนับสนุนการวิจัย (สกว.) ที่ได้เล็งเห็นถึงความสำคัญของการเสริมสร้างนักวิจัยและพัฒนางานวิจัยของชาติ โดยทั้งสองหน่วยงานได้ร่วมมือกันจัดสรรทุนวิจัยให้กับนักวิจัยที่ทำงานอยู่ในมหาวิทยาลัยต่างๆ ทั่วประเทศมาอย่างต่อเนื่อง ความสำเร็จของงานวิจัยนี้จะเกิดขึ้นมิได้หากไม่ได้รับการสนับสนุนจากผู้บริหารของหน่วยงานต้นสังกัดได้แก่รองศาสตราจารย์ นาวาอากาศเอก ดร.วรพจน์ ขำพิศ คณบดีสำนักวิชาวิศวกรรมศาสตร์ และศาสตราจารย์ ดร.ประสาท สืบค้า อธิการบดีมหาวิทยาลัยเทคโนโลยีสุรนารี ที่ตระหนักถึงความสำคัญของการสร้างนักวิจัย และสนับสนุนคณาจารย์ในมหาวิทยาลัยให้ได้ทำงานวิจัยอย่างต่อเนื่อง โดยได้เตรียมสภาพแวดล้อมในการทำงานทั้งสถานที่และอุปกรณ์การวิจัยที่เหมาะสมสำหรับนักวิจัย รวมถึงการสนับสนุนให้คณาจารย์ได้มีโอกาสเดินทางไปเผยแพร่ผลงานวิจัยต่อที่ประชุมวิชาการทั้งในและต่างประเทศ ผู้วิจัยขอขอบคุณทุกท่านไว้ ณ โอกาสนี้

บทคัดย่อ

รหัสโครงการ: RMU5080026

ชื่อโครงการ: การค้นพบโดยประมาณของรูปแบบที่ปรากฏบ่อยในข้อมูลสตรีม

ชื่อนักวิจัย: รศ.ดร.กิตติศักดิ์ เกิดประสพ
สาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

E-mail address: kerdpras@sut.ac.th, KittisakThailand@gmail.com

ระยะเวลาโครงการ: 25 มิถุนายน 2550 – 24 มิถุนายน 2553

การค้นพบรูปแบบที่ปรากฏบ่อย เป็นปฏิบัติการที่สำคัญในงานวิเคราะห์ความสัมพันธ์ กระบวนการค้นพบนี้เป็นการคัดแยกโดยอัตโนมัติ เพื่อค้นหารูปแบบที่น่าสนใจและความสัมพันธ์ที่เกิดร่วมกันในฐานข้อมูลขนาดใหญ่ รูปแบบที่น่าสนใจเหล่านี้สามารถนำไปใช้เพื่อการสร้างกฎความสัมพันธ์สำหรับสนับสนุนการตัดสินใจในงานด้านต่างๆ เช่น การพยากรณ์ด้านการเงิน การวินิจฉัยทางการแพทย์ งานวิจัยปัจจุบันด้านการทำเหมืองข้อมูลเพื่อค้นหาความสัมพันธ์ จะมุ่งความสนใจไปที่การพัฒนาวิธีการที่มีประสิทธิภาพสูงเพื่อค้นหาวัตถุหรือไอเท็มที่ปรากฏร่วมกัน ทั้งนี้เนื่องจากถ้าในฐานข้อมูลประกอบด้วยไอเท็มที่แตกต่างกันจำนวน m รายการ ไอเท็มเหล่านี้สามารถทำให้เกิดรูปแบบร่วมได้มากถึง 2^m รูปแบบ ซึ่งการค้นหารูปแบบร่วมจำนวนมากเช่นนี้จะเป็นงานประมวลผลที่ใช้เวลานานมาก การค้นหานี้จะยิ่งยากขึ้นถ้าข้อมูลเป็นลักษณะสตรีม เนื่องจากลักษณะของสตรีมจะผลิตข้อมูลอย่างต่อเนื่องในปริมาณมาก ทำให้การทำงานกับข้อมูลต้องกระทำในรอบเดียวเพื่อให้ได้ผลการวิเคราะห์ความสัมพันธ์ที่ทันต่อการใช้งาน จากการศึกษาถึงลักษณะของปัญหาและข้อจำกัดต่างๆเหล่านี้ ผู้วิจัยจึงได้พัฒนาวิธีการโดยประมาณเพื่อค้นหารูปแบบที่ปรากฏบ่อยในข้อมูลสตรีม วิธีการโดยประมาณจะถูกนำมาใช้กับข้อมูลสตรีมก่อนที่จะนำข้อมูลตัวแทนไปประมวลผลต่อเพื่อค้นหารูปแบบที่ปรากฏบ่อย วิธีการที่เสนอขึ้นนี้ได้รับการพัฒนาเป็นโปรแกรมต้นแบบด้วยภาษาเชิงฟังก์ชัน ผลการทดสอบโปรแกรมกับข้อมูลจริงได้ผลลัพธ์ที่น่าพอใจในด้านประสิทธิภาพและความถูกต้องของรูปแบบที่เป็นผลลัพธ์ของโปรแกรม แนวทางการพัฒนาต่อยอดของงานวิจัยนี้ จะเป็นการใช้วิธีการประมวลผลโปรแกรมแบบคู่ขนานที่คาดว่าจะช่วยให้สามารถทำงานกับข้อมูลขนาดใหญ่มากได้

คำหลัก: การค้นพบรูปแบบที่ปรากฏบ่อย, ข้อมูลสตรีม, วิธีการโดยประมาณ

Abstract

Project Code: RMU5080026

Project Title: Approximate frequent pattern discovery over data stream

Investigator: Dr. Kittisak Kerdprasop, Asscociate Professor
School of Computer Engineering, Suranaree University of Technology

E-mail Address: kerdpras@sut.ac.th, KittisakThailand@gmail.com

Project Period: 25 June 2007 – 24 June 2010

Frequent pattern discovery is an essential operation for association analysis. The discovery process concerns an automatic extraction of interesting patterns and correlations from a large database. These patterns can reveal implicit relationships among set of objects (or items) that lead to the generation of association rules to be used for decision support, financial forecast, medical diagnosis and many other applications. Current studies in association rule mining concentrate on how to effectively find all objects frequently co-occurring. Given m objects, there are as much as 2^m frequent patterns to consider. Frequent pattern discovery is thus a computationally expensive problem. It is even harder over data stream because a continuously generated nature of stream does not allow a revisit on each data element. Furthermore, pattern discovery process must be fast to produce timely results. Based on these requirements, we devise an approximate approach to tackle the problem of discovering frequent patterns over continuous stream. Our approximation algorithm is intended to be applied to process a stream prior to the pattern discovery process. We propose a stochastic method to get a good guess of the stream characteristics, and then draw a set of representatives from the incoming stream. These representatives are subsequently used in the process of frequent pattern mining. Our design had been implemented with the functional programming paradigm and the experimental results confirm the efficiency and reliability of our method. For a massive database, parallel method is a solution for the scalability problem. That is the main direction of our future research.

Keywords: Frequent pattern discovery, Data stream, Approximation method

บทสรุปสำหรับผู้บริหาร
Executive Summary

รหัสโครงการ: RMU5080026

ชื่อโครงการ: การค้นพบโดยประมาณของรูปแบบที่ปรากฏบ่อยในข้อมูลสตรีม
Approximate frequent pattern discovery over data stream

ชื่อนักวิจัย: รศ.ดร.กิตติศักดิ์ เกิดประสพ
สาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

E-mail address: kerdpras@sut.ac.th, KittisakThailand@gmail.com

ระยะเวลาโครงการ: 25 มิถุนายน 2550 – 24 มิถุนายน 2553

ปัญหาที่ทำการวิจัย และความสำคัญของปัญหา

การค้นพบความสัมพันธ์หรือความเชื่อมโยงของสินค้าที่ถูกซื้อร่วมกันบ่อยเช่น ลูกค้าที่ซื้อเบียร์มักจะซื้อผ้าอ้อมเด็กด้วย เรียกว่า *การทำเหมืองเพื่อค้นหาความสัมพันธ์* (association mining) ซึ่งเป็นวิธีการทำเหมืองข้อมูลประเภทหนึ่ง การค้นพบความสัมพันธ์อาศัยขั้นตอนพื้นฐานที่สำคัญคือ *การค้นพบรูปแบบที่ปรากฏบ่อย* (frequent pattern discovery) รูปแบบนี้มีได้หลากหลายประเภท เช่น รูปแบบการซื้อสินค้าของลูกค้าส่วนใหญ่ในห้างสรรพสินค้า, รูปแบบการเกิดแผ่นดินไหวในรอบสิบปีที่ผ่านมาของภูมิภาคเอเชียตะวันออกเฉียงใต้, รูปแบบการค้นหาข้อมูลจากเว็บเพจต่างๆของผู้ใช้อินทอร์เน็ต, รูปแบบการจัดโครงสร้างโมเลกุลในสารประกอบโปรตีน เป็นต้น

นับตั้งแต่ปีค.ศ.1993 ที่ได้มีการเริ่มเสนอแนวคิดและเทคนิคของการทำเหมืองข้อมูลประเภทการค้นพบรูปแบบที่ปรากฏบ่อย งานวิจัยด้านนี้ก็ได้รับความนิยมอย่างกว้างขวางเนื่องจากใช้ประโยชน์ได้กับการวิเคราะห์ข้อมูลในหลากหลายสาขา มีการพัฒนาเทคนิคต่างๆของการค้นหารูปแบบที่ปรากฏบ่อยให้มีประสิทธิภาพสูงขึ้น ในระยะหลังความก้าวหน้าของเทคโนโลยีอินเทอร์เนตทำให้รูปแบบของข้อมูลเปลี่ยนแปลงไปจากข้อมูลที่มีขนาดใหญ่แต่คงตัว (static) กลายเป็นข้อมูลที่มีลักษณะพลวัต (dynamic) เปลี่ยนแปลงได้ทั้งการเพิ่มปริมาณอย่างต่อเนื่องและการเปลี่ยนรูปแบบการกระจายของข้อมูล ข้อมูลลักษณะเช่นนี้เรียกว่า *ข้อมูลสตรีม* (data stream)

ในช่วงเวลาห้าปีที่ผ่านมานักวิจัยจำนวนหนึ่งในสาขาการทำเหมืองข้อมูล เริ่มให้ความสนใจกับข้อมูลสตรีม พยายามปรับปรุงเทคนิคการค้นหารูปแบบที่ปรากฏบ่อยที่ใช้งานได้ดีกับข้อมูลคงตัวให้ทำงานกับข้อมูลสตรีมได้ แต่เทคนิคเหล่านั้นยังมีข้อจำกัดและใช้ได้กับเพียงบางแอพพลิเคชัน ผู้วิจัยจึงได้เสนอโครงการวิจัยนี้ขึ้นเพื่อพัฒนาเทคนิคการค้นหารูปแบบที่ปรากฏบ่อยจากข้อมูลสตรีม ที่ใช้งานได้กับหลากหลายแอพพลิเคชัน เช่น web-page click stream, DNA stream, market basket data stream และเพื่อให้ได้ผลลัพธ์ของการค้นหารูปแบบที่ปรากฏบ่อยในเวลาที่รวดเร็วเพื่อทันต่อความต้องการใช้งาน

การประมวลผลข้อมูลสตรีมจะใช้การประมาณค่าด้วยกระบวนการ stochastic ในขั้นตอนของการพัฒนาโปรแกรมต้นแบบจะใช้แนวทางการโปรแกรมเชิงฟังก์ชัน ที่มีโครงสร้างของภาษาสนับสนุนการตรวจจับแพทเทิร์น การพัฒนาเทคนิคการค้นพบรูปแบบที่ปรากฏบ่อยในข้อมูลสตรีมโดยใช้ stochastic process วิธีการค้นพบโดยประมาณและการทำ rapid prototyping ด้วยการโปรแกรมเชิงฟังก์ชัน รวมถึงแนวทางการทดสอบประสิทธิภาพของโปรแกรม จะเป็นองค์ความรู้ใหม่ที่ช่วยพัฒนาความก้าวหน้าของงานวิจัยในสาขาการทำเหมืองข้อมูลประเภทการค้นพบรูปแบบที่ปรากฏบ่อย และการค้นพบความสัมพันธ์

วัตถุประสงค์

- เพื่อพัฒนาเทคนิคและอัลกอริทึมในการค้นพบโดยประมาณของรูปแบบที่ปรากฏบ่อยในข้อมูลสตรีม
- พัฒนากรอบงาน stochastic process ที่เหมาะสมกับการทำ association mining กับข้อมูลสตรีม
- ออกแบบและพัฒนาโปรแกรมต้นแบบในการทำ approximate association mining ด้วยเทคนิคการโปรแกรมเชิงฟังก์ชัน

ผลสำเร็จของโครงการ

1. ผลงานตีพิมพ์ในหนังสือวิชาการ (Book chapters)

    1) K. Kerdprasop and N. Kerdprasop (2009). Knowledge mining with a higher-order logic approach. In K. Nakamatsu, G. Phillips-Wrens, L.C. Jain, R.J. Howlett (Eds.), *New Advances in Intelligent Decision Technologies*, pp. 151-159, Springer. (ISBN: 978-3-642-00908-2, DOI: 10.1007/978-3-642-00909-9)

    2) N. Kerdprasop, S. Pilabutr, and K. Kerdprasop (2009). Improving medical database consistency with induced trigger rules. In K. Nakamatsu, G. Phillips-Wrens, L.C. Jain, R.J. Howlett (Eds.), *New Advances in Intelligent Decision Technologies*, pp. 265-274, Springer. (ISBN: 978-3-642-00908-2, DOI: 10.1007/978-3-642-00909-9)

2. ผลงานตีพิมพ์ในวารสารวิชาการนานาชาติ (Journal articles)

    1) N. Pannurat, N Kerdprasop, and K. Kerdprasop (2010). Database reverse engineering based on association rule mining. *International Journal of Computer Science Issues (IJCSI)*, Volume 7, Issue 2, March 2010, pp. 10-15.

    2) N. Kerdprasop and K. Kerdprasop (2009). Knowledge induction from medical databases with higher-order programming. *WSEAS Transactions on Information Science and Applications*, Issue 10, Volume 6, October 2009, pp. 1719-1728.

    3) K. Kerdprasop and N. Kerdprasop (2008). Approximate frequent pattern discovery over data stream. *International Journal of Computer Science and Engineering (IJCSE)*, Volume 2, Number 1, pp. 28-32.

    4) K. Kerdprasop and N. Kerdprasop (2007). On pattern-based programming towards the discovery of frequent patterns. *International Journal of Computer Science (IJCS)*, Volume 2, Number 4, pp. 268-273.

5) N. Kerdprasop and K. Kerdprasop (2007). Mining frequent patterns with functional programming. *International Journal of Computer and Information Science and Engineering (IJCISE)*, Volume 1, Number 2, pp. 66-71.

6) N. Kerdprasop and K. Kerdprasop (2007). Moving data mining tools toward a business intelligence system. *International Journal of Intelligent Technology (IJIT)*, Volume 2, Number 2, pp. 99-104.

Cited by:

Z. Zhu, J. Gu, L. Zhang, W. Song, and R. Gao. (2009). Research on domain-driven actionable knowledge discovery. *Proc. 20th Int. Conf. on Cutting-Edge Research Topics on Multiple Criteria Decision Making (MCDM 2009)*, 21-26 June 2009, China, pp.176-183.

M. Al-Noukari and W. Al-Hussan. (2007). Using data mining techniques for predicting future car market demand: dcx case study. *Proc. 3rd Int. Conf. on Information and Communication Technology (ICTTA 2008)*, 7-11 April 2008, pp.1-5.

3. การเสนอผลงานในที่ประชุมวิชาการและตีพิมพ์ผลงานใน Conference Proceedings

1) K. Kerdprasop and N. Kerdprasop (2009). Automated induction of frequent patterns with knowledge-based software engineering. *Proceedings of the Joint Conference of ASCM 2009 (Asian Symposium on Computer Mathematics) and MACIS 2009 (Mathematical Aspects of Computer and Information Sciences)*, 14-17 December 2009, Fukuoka, Japan, pp. 431-434.

2) N. Kerdprasop and K. Kerdprasop (2008). A declarative programming paradigm and the development of knowledge mining agents. *Proceedings of IADIS Multi Conference on Computer Science and Information Systems*, Amsterdam, Netherlands, 22-27 July 2008, pp. 45-52.

4. การนำผลงานวิจัยไปใช้ประโยชน์เชิงวิชาการ

1) ใช้ประกอบการเรียนการสอนในรายวิชา 423681 Selected Topics in Computer Engineering I ซึ่งเป็นรายวิชาเลือกในหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์

2) สร้างนักวิจัยใหม่ด้วยการนำส่วนหนึ่งของงานวิจัยไปเป็นหัวข้อวิทยานิพนธ์ "การปรับรูปแบบบรรทัดฐานในฐานข้อมูลเชิงสัมพันธ์ด้วยเทคนิคการวิเคราะห์ความสัมพันธ์" (Normalization in relational database with association analysis technique) โดยนายณัฐพล พันนุรัตน์ นักศึกษาในหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์

สารบัญ

สารบัญภาพ

สารบัญตาราง

หน้า