

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 เพื่อค้นหารูปแบบที่น่าสนใจและความสัมพันธ์ที่เกิดขึ้นร่วมกันในฐานข้อมูลขนาดใหญ่ รูปแบบที่น่าสนใจเหล่านี้สามารถนำไปใช้เพื่อการสร้างกฎความสัมพันธ์สำหรับสนับสนุนการตัดสินใจในงานด้านต่างๆ เช่น การพยากรณ์ด้านการเงิน การวินิจฉัยทางการแพทย์ งานวิจัยปัจจุบันด้านการทำเหมืองข้อมูลเพื่อค้นหาความสัมพันธ์ จะมุ่งความสนใจไปที่การพัฒนาวิธีการที่มีประสิทธิภาพสูงเพื่อค้นหาวัตถุหรือไอเท็มที่ปรากฏร่วมกัน ทั้งนี้เนื่องจากถ้าในฐานข้อมูลประกอบด้วยไอเท็มที่แตกต่างกันจำนวน m รายการ ไอเท็มเหล่านี้สามารถทำให้เกิดรูปแบบรวมได้มากถึง 2^m รูปแบบ ซึ่งการค้นหารูปแบบรวมจำนวนมากเช่นนี้จะเป็งานประมวลผลที่ใช้เวลานานมาก การค้นหานี้จะยิ่งยากขึ้นถ้าข้อมูลเป็นลักษณะสตรีม เนื่องจากลักษณะของสตรีมจะผลิตข้อมูลอย่างต่อเนื่องในปริมาณมาก ทำให้การทำงานกับข้อมูลต้องกระทำในรอบเดียวเพื่อให้ได้ผลการวิเคราะห์ความสัมพันธ์ที่ทันต่อการใช้งาน จากการศึกษาถึงลักษณะของปัญหาและข้อจำกัดต่างๆเหล่านี้ ผู้วิจัยจึงได้พัฒนาวิธีการโดยประมาณเพื่อค้นหารูปแบบที่ปรากฏบ่อยในข้อมูลสตรีม วิธีการโดยประมาณจะถูกนำมาใช้กับข้อมูลสตรีมก่อนที่จะนำข้อมูลตัวแทนไปประมวลผลต่อเพื่อค้นหารูปแบบที่ปรากฏบ่อย วิธีการที่เสนอขึ้นนี้ได้รับการพัฒนาเป็นโปรแกรมต้นแบบด้วยภาษาเชิงฟังก์ชัน ผลการทดสอบโปรแกรมกับข้อมูลจริงได้ผลลัพธ์ที่น่าพอใจในด้านประสิทธิภาพและความถูกต้องของรูปแบบที่เป็นผลลัพธ์ของโปรแกรม แนวทางการพัฒนาต่อยอดของงานวิจัยนี้ จะเป็นการใช้วิธีการประมวลผลโปรแกรมแบบคู่ขนานที่คาดว่าจะช่วยให้สามารถทำงานกับข้อมูลขนาดใหญ่ได้

คำหลัก: การค้นพบรูปแบบที่ปรากฏบ่อย, ข้อมูลสตรีม, วิธีการโดยประมาณ

Abstract

Project Code: RMU5080026

Project Title: Approximate frequent pattern discovery over data stream

Investigator: Dr. Kittisak Kerdprasop, Associate 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