ธรรมศักดิ์ เธียรนิเวศน์ : การลดขนาดข้อมูลด้วยน้ำหนักความหนาแน่นเพื่อการจัดกลุ่มข้อมูลขนาดใหญ่ (A DENSITY-BASED DATA REDUCTION FOR CLUSTERING ON LARGE DATA SETS) อาจารย์ที่ปรึกษา : ผศ. ดร.กิตติศักดิ์ เกิดประสพ, 120 หน้า.
ISBN 974-533-521-5

กระบวนการจัดกลุ่มข้อมูลอัตโนมัติบนชุดข้อมูลที่มีขนาดใหญ่มากๆ เป็นกระบวนการที่ต้องใช้เวลาและสิ้นเปลืองหน่วยความจำเป็นจำนวนมาก การลดขนาดข้อมูลเป็นแนวทางหนึ่งที่จะช่วยแก้ปัญหานี้ได้ งานวิจัยนี้จึงมุ่งที่จะศึกษาค้นคว้า ตลอดจนพัฒนาวิธีการลดขนาดข้อมูลด้วยเทคนิคการสุ่มข้อมูลที่มีความทนทานต่อข้อมูลรบกวน สำหรับงานการจัดกลุ่มข้อมูลขนาดใหญ่ที่มีการกระจายแบบไม่ปกติ และสามารถสร้างชุดข้อมูลสุ่มแบบต่อเนื่องได้ภายในการอ่านข้อมูลเพียงรอบเดียว จากการวิเคราะห์หาลักษณะเด่นของอัลกอริธึมสุ่มข้อมูล RVS, DBS, และ DBRVS ที่เคยมีนักวิจัยท่านอื่นเสนอไว้ พบว่าอัลกอริธึม DBS เป็นอัลกอริธึมที่มีความแม่นตรงสูง การสุ่มข้อมูลด้วยอัลกอริธึม DBS เพียง 2% สามารถให้ผลลัพธ์ของกลุ่มข้อมูลได้เทียบเท่ากับข้อมูลทั้งหมด อีกทั้งยังสามารถลดเวลาในการจัดกลุ่มข้อมูลได้มากกว่า 95% แต่เมื่อข้อมูลที่ใช้มีข้อมูลรบกวนปะปนอยู่ คุณภาพของการสุ่มข้อมูลของอัลกอริธึม DBS กลับลดลงอย่างเห็นได้ชัด ซึ่งแสดงให้เห็นว่าอัลกอริธึม DBS มีความอ่อนไหวต่อข้อมูลรบกวนสูงเช่นกัน

งานวิจัยนี้จึงได้เสนออัลกอริธึมสุ่มข้อมูล DBSPACE ซึ่งพัฒนาขึ้นเพื่อเพิ่มศักยภาพในการทนทานต่อข้อมูลรบกวน โดยการพิจารณาค่าความเป็นปึกแผ่นของข้อมูล ซึ่งบริเวณที่มีความน่าจะเป็นที่จะเป็นกลุ่มข้อมูลที่สนใจจะมีค่าความเป็นปึกแผ่นของข้อมูลสูงกว่าบริเวณที่มีความน่าจะเป็นที่จะเป็นข้อมูลรบกวน

จากผลการวิจัยพบว่า อัลกอริธึม DBSPACE สามารถให้ผลลัพธ์ที่ดีเทียบเท่ากับอัลกอริธึม DBS ในกรณีที่ข้อมูลปราศจากข้อมูลรบกวน และสามารถให้ผลลัพธ์ที่ดีกว่า ในกรณีที่ข้อมูลมีข้อมูลรบกวนปะปนอยู่

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนักศึกษา _____

ปีการศึกษา 2548 ลายมือชื่ออาจารย์ที่ปรึกษา _____

ลายมือชื่ออาจารย์ที่ปรึกษาร่วม _____

THAMMASAK THIANNIWET : A DENSITY-BASED DATA REDUCTION FOR CLUSTERING ON LARGE DATA SETS. THESIS ADVISOR : ASST. PROF. KITTISAK KERDPRASOP, Ph.D., 120 PP. ISBN 974-533-521-5

DATA MINING/DATA REDUCTION/SAMPLING/CLUSTERING/ DENSITY-BIASED/COMPACTNESS

Determining clusters in large data sets takes a very long time and consumes many resources. Data reduction is an important step to increase the efficiency of determining clusters in large data sets. Our work is intended to examine and develop the appropriate sampling technique as a data reduction scheme for clustering that requires only a single data set scan. From the experimental results performed on three sampling algorithms (RVS, DBS, and DBRVS), we found that DBS is the most accurate sampling algorithm. A 2% DBS sample of the original data set can produce the same result as the whole original data set and also help reduce time to find the clusters by over 95%. However, it shows sensitivity on a noisy data set.

Our research is intended to propose the DBSPACE algorithm which is developed to gain the potential of noise tolerance by considering the compactness within the region of data. The region with more compactness will be a good area from which representative sample should be drawn.

The results of this research showed that, DBSPACE sample can produce the result as accurate as DBS sample drawn from clean data sets. It can produce the better result while the original data set is surrounded by many noises.

School of Computer Engineering

Academic Year 2005

Student's Signature

Advisor's Signature

Co-advisor's Signature