บทความวิจัย (Research Article)

ข้อมูลก่อนประมวลผลสำหรับการจัดกลุ่มเคมีนส์แบบขนาน

Data pre-processing for parallel k-means clustering

ไพชยนต์ คงไชย[1*], ภาสพิชญ์ ชูใจ[1], วิภาสิทธิ์ หิรัญรัตน์[2], นิตยา เกิดประสพ[1] และกิตติศักดิ์ เกิดประสพ[1]

Phaichayon Kongchai[1*], Pasapitch Chujai[1], Wiphasith Hiranrat[2], Nittaya Kerdprasop[1] and Kittisak Kerdprasop[1]

บทคัดย่อ

ข้อมูลก่อนประมวลผลเป็นกลุ่มเป็นขั้นตอนสำคัญในการทำเหมือนแยกแยะข้อมูล การเตรียมข้อมูลที่ดีนำไปสู่กลุ่มสมรรถนะที่ดี การเตรียมข้อมูลมีหลากหลายวิธีการ การวิจัยนี้เสนอแนวคิดของการเตรียมความสมดุลของข้อมูลอย่างไรก่อนประมวลผลในระบบการปฏิบัติงานแบบขนาน วัตถุประสงค์เพื่อหารูปแบบการแบ่งจำนวนข้อมูลที่เหมาะสม สำหรับขั้นตอนวิธีการจัดกลุ่มเคมีนส์แบบขนาน รวมถึงพัฒนาขั้นตอนวิธีด้วยการใช้ภาษาเออร์แลงอันเป็นภาษาเชิงหน้าที่ร่วมกันภาษาหนึ่ง จากนั้นทดลองกับชุดข้อมูลสังเคราะห์ลักษณะหลายมิติ และมีจำนวนกลุ่มแปรผันตั้งแต่ 2 ถึง 10 กลุ่ม ผลลัพธ์การทดลองแสดงให้เห็นว่า สมรรถนะเวลาของรูปแบบการแยกเท่าเทียมกันดีกว่ารูปแบบอื่น

คำสำคัญ: การจัดกลุ่มเคมีนส์แบบขนาน, ข้อมูลก่อนประมวลผล, การจัดกลุ่ม

Abstract

Data pre-processing in clustering is an important step in data mining. A good data preparation can lead to a good clustering performance. Preparation of data has variety of methods. This research proposed the concept of how to prepare data balancing before processing in a parallel machine. The objective is to find the pattern of splitting the amount of data that is appropriate for parallel k-means clustering algorithm. In addition, the algorithm was developed using Erlang language, which is a concurrent functional language. Then we experiment with synthetic data sets that are multi-dimensional, and have a number of clusters varying from 2 to 10. The experimental results show that the time performance of equal pattern decomposition is better than other patterns.

Keywords: Parallel k-means clustering, data pre-processing, clustering

---

[1] สาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี จ.นครราชสีมา 30000

[2] สาขาวิชาเทคโนโลยีสารสนเทศ สำนักวิชาเทคโนโลยีสังคม มหาวิทยาลัยเทคโนโลยีสุรนารี จ.นครราชสีมา 30000

[1] School of Computer Engineering, Institute of Engineering, Suranaree University of Technology, Nakhon Ratchasima Province 30000

[2] School of Information Technology, Institute of Social Technology, Suranaree University of Technology, Nakhon Ratchasima Province 30000

[*] Corresponding author, e-Mail: zaguraba_ii@hotmail.com

Received: 12 September 2012; Accepted: 30 November 2013

บทนำ

การจัดกลุ่ม (clustering) [1] หมายถึง การจำแนกสิ่งเหมือนกันหรือคล้ายกันเป็นกลุ่มเดียวกัน การจัดกลุ่มข้อมูลแบ่งเป็น 2 ประเภท ได้แก่ การเข้าถึงข้อมูลเชิงลำดับชั้น (hierarchical approach) และการเข้าถึงเชิงแบ่งพวก (partition approach) [10] ทั้งสองประเภทแตกต่างกันในรายละเอียดและจุดประสงค์การใช้งาน การเข้าถึงเชิงลำดับชั้นอาศัยแผนภาพลักษณะคล้ายต้นไม้ (dendrogram) จัดกลุ่มข้อมูลแบบแบ่งพวกด้วย การแบ่งแยกข้อมูลหมวดหมู่ข้อมูลหนึ่งเป็นกลุ่มย่อย [10] วิธีการแบ่งพวกที่ใช้แพร่หลาย ได้แก่ ขั้นตอนวิธีเคมีนส์ (k-means) และขั้นตอนวิธีฟัซซี่ซีมีน (fuzzy c-means) เป็นต้น [4]

ขั้นตอนวิธีเคมีนส์นิยมใช้มากที่สด [13] ด้วยสามารถทำงานกับข้อมูลปริมาณมาก รวดเร็ว และเข้าใจง่ายอาศัยการแบ่งข้อมูลเป็นกลุ่มย่อยตามความคล้ายกันหรือเหมือนและทำงานวนซ้ำจนได้กลุ่มข้อมูลตามเกณฑ์กำหนดหรือมีเสถียรภาพ โดยปราศจากการเปลี่ยนกลุ่มหรือมีข้อผิด พลาดต่ำที่สุด [6, 12] ขั้นตอนวิธีเคมีนส์ใช้กับข้อมูลได้หลายลักษณะงาน แต่ไม่ทุกประเภท ด้วยมีข้อจำกัดในการจัดกลุ่มข้อมูลขนาดใหญ่ เป็นผลให้สูญเสียเวลาการทำงานและหน่วยความจำที่ใช้ประมวลผล [5] หากมีข้อมูลจำนวน n ข้อมูล การจัดการกลุ่มข้อมูลของขั้นตอนวิธีเคมีนส์เป็นการทำงานกับทั้ง n ข้อมูลในทุกรอบการทำงาน ดังนั้น เพื่อเพิ่มประสิทธิภาพการทำงานของขั้นตอนวิธีเคมีนส์จำเป็นต้องจัดกลุ่มข้อมูลให้ทำงานพร้อมกันหรือแบ่งงานในการทำงานแต่ละรอบ

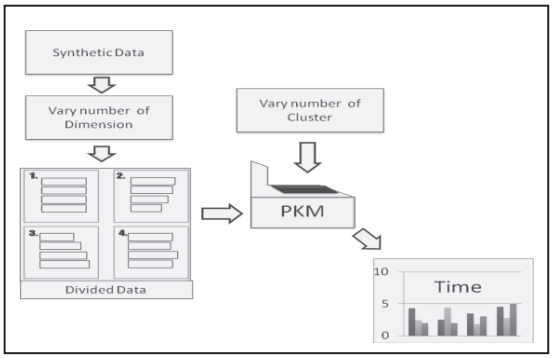
การประมวลผลแบบพร้อมกันหรือการทำงานแบบขนาน (parallel) เป็นวิธีหนึ่งอันมุ่งเน้นลดเวลาการประมวลผล [7] ยิ่งนำเอาหลักการทำงานแบบขนานมาประยุกต์ใช้กับเทคนิคเคมีนส์ ทำให้ลดจำนวนรอบการทำงานและเพิ่มความรวดเร็วของการจัดกลุ่มข้อมูล [8] สอดคล้องกับงานวิจัยของ Kerdprasop K [7] พบว่าเทคนิคเคมีนส์แบบขนานช่วยลดเวลาการประมวลผล และการจัดกลุ่มข้อมูลทำงานแบบขนานทำให้การประมวลผลเร็วขึ้นเช่นเดียวกับ นเรศ ผ่องสวัสดิ์กุล [3] พบว่า การจัดกลุ่มข้อมูลแบบขนานลดเวลาการประมวล และข้อมูลมากขึ้น

กระนั้นก็ตาม ก่อนประมวลผลการเตรียมข้อมูลเป็นขั้นตอนสำคัญหนึ่ง ด้วยการบรรจุข้อมูลอย่างสมดุล (load balancing) และเป็นปัจจัยเพิ่มประสิทธิภาพการประมวลผล [2, 11] สอดคล้องกับการศึกษาของ Liang H, [9] พบว่าการบรรจุข้อมูลแบบสมดุลด้วยพลวัตลดเวลาประมวลผลคล้ายคลึงกัน Youssef M Marzouk [11] สรุปว่าการประมวลผลแบบขนานมีความแม่นยำร้อยละ 98 ในการประมวลผลด้วย 1024 การประมวลส่วน Kijsipongse E [8] เสนอการบรรจุข้อมูลแบบสมดุลบนหน่วยประมวลผลกราฟิก (graphic processing unit - GPU) และระบบความจำร่วมแบบกระจาย พบว่า การจัดกลุ่มเคมีนส์แบบขนานร่วมกับการบรรจุข้อมูลแบบสมดุลด้วยพลวัตลดเวลาในการประมวลผล

คณะผู้วิจัยมีจุดมุ่งหมายระบุหารูปแบบของการแบ่งข้อมูลที่เหมาะสมเพื่อเพิ่มประสิทธิภาพและลดเวลาการประมวลผลแบบขนานด้วยขั้นตอนวิธีเคมีนส์

วิธีดำเนินการวิจัย

เป็นการทดลองตามกระบวนการในการดำเนินงาน ดังภาพที่ 1



ภาพที่ 1 กระบวนการหารูปแบบของการแบ่งจำนวนข้อมูลที่เหมาะสมกับการประมวลผลแบบขนานด้วยขั้นตอนวิธีเคมีนส์

ภาพที่ 1 อธิบายรายละเอียดของแต่ละขั้นตอนดังต่อไปนี้

1. สร้างข้อมูลสังเคราะห์ (synthetic data) จำนวนมิติตั้งแต่ 2 มิติ ถึง 5 มิติ โดยข้อมูลมีลักษณะ

เป็นตัวเลข แสดงตัวอย่างข้อมูล 2 มิติ ได้ดังภาพที่ 2 ข้อมูลแต่ละจุดจะเขียนอยู่ในโครงสร้างลิสต์ เช่น [3521, 7420]

```
[3521,7420]. [3417,3828]. [9992,1700]. [1011,9796]. [9328,1023].
[8982,3404]. [7882,1696]. [2021,7202]. [8344,793].  [8734,2380].
[7712,1412]. [2216,8080]. [7899,905].  [7082,3819]. [841,4085].
[703,6993].  [5821,5423]. [4257,7034]. [2784,3626]. [3967,7018].
[2887,4190]. [9070,4442]. [3240,4436]. [1838,2991]. [3708,6518].
[9066,3437]. [675,1285].  [1092,4291]. [8318,3015]. [1749,8098].
[1500,8705]. [9619,4013]. [3765,3504]. [6139,1849]. [2641,6981].
[8781,1137]. [1319,9857]. [4544,8316]. [3234,1316]. [8287,6993].
[951,2520].  [2117,7592]. [8676,8938]. [7633,7598]. [7472,9660].
[6749,2009]. [962,4102].  [6487,7184]. [8873,6993]. [6813,2651].
```

ภาพที่ 2 ตัวอย่างข้อมูลสังเคราะห์ลักษณะเป็น 2 มิติ

2. นำข้อมูลได้จากการสังเคราะห์แบ่งเป็น 4 รูปแบบ ดังต่อไปนี้

1) แบ่งจำนวนข้อมูลแบบเท่ากัน
2) แบ่งจำนวนข้อมูลจากมากไปน้อย
3) แบ่งจำนวนข้อมูลจากน้อยไปมาก
4) แบ่งจำนวนข้อมูลแบบสุ่ม

การแบ่งข้อมูลรูปแบบต่างๆ แสดงดังตารางที่ 1

ตารางที่ 1 รูปแบบการแบ่งข้อมูล

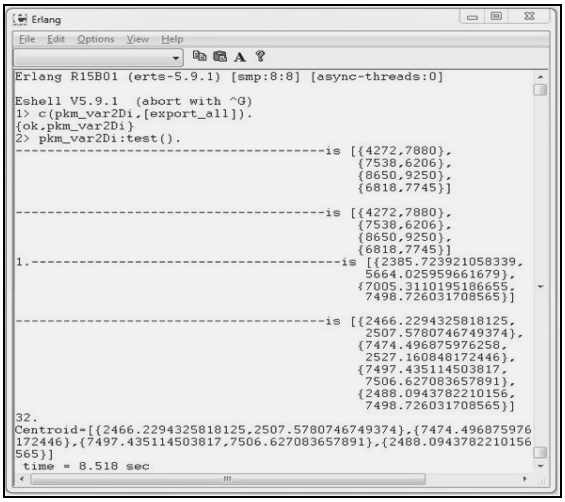
รูปแบบ	การประมวลผล (ร้อยละ)			
	1	2	3	4
เท่ากัน	25	25	25	25
มากไปน้อย	40	30	20	10
น้อยไปมาก	10	20	30	40
สุ่ม	NA	NA	NA	NA

NA (not applicable) = ไม่เกี่ยว

ตารางที่ 1 รูปแบบเท่ากันแบ่งข้อมูลแต่ละการประมวลผลเท่ากับร้อยละ 25 รูปแบบมากไปน้อยกำหนดการประมวลผลแรกได้รับข้อมูลมากสุดร้อยละ 40 ประมวลผลหลังลดหลั่นร้อยละ 10 ตามลำดับ รูปแบบน้อยไปมากกำหนดการประมวลผลแรกได้รับข้อมูลน้อยสุดร้อยละ 10 ประมวลผลหลังเพิ่มร้อยละ 10 ตามลำดับ ส่วนรูปแบบสุ่มกำหนดแต่ละการประมวลผลได้รับข้อมูลคละกันในสัดส่วนร้อยละ 10, 20, 30 และ 40

3. ประมวลผลข้อมูลตามรูปแบบกำหนด ด้วยขั้นตอนวิธีเคมีนส์แบบขนาน ทดลองกลุ่มข้อมูลแปรผันตั้งแต่ 2 ถึง 10 กลุ่ม โดยเพิ่มขึ้นทีละ 2

4. เปรียบเทียบประสิทธิภาพด้านเวลาของแต่ละรูปแบบ ตัวอย่างผลการจัดกลุ่มเคมีนส์แบบขนาน ในรูปแบบของการแบ่งจำนวนของข้อมูลที่เท่ากันด้วยโปรแกรมเออร์แลง (Erlang) ดังภาพที่ 3



ภาพที่ 3 ตัวอย่างผลการจัดกลุ่มด้วยขั้นตอนวิธีเคมีนส์แบบขนานในรูปแบบของการแบ่งจำนวนข้อมูลที่เท่ากัน

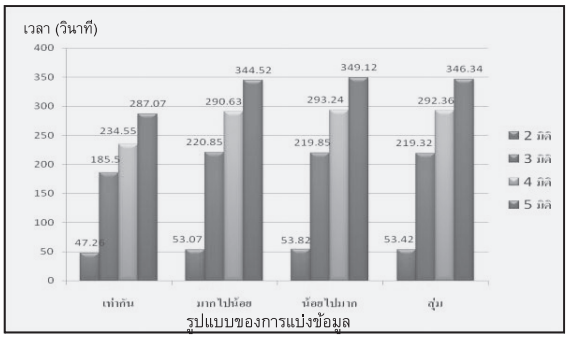
ผลการวิจัย

งานวิจัยนี้ศึกษาและทดลองเกี่ยวกับกระบวนการเตรียมข้อมูลสำหรับการจัดกลุ่มข้อมูลด้วยขั้นตอนเคมีนส์แบบขนาน ใช้ชุดข้อมูลจากการสังเคราะห์มีขนาด 500,000 จุด และเป็นตัวเลขจำนวนเต็มที่มีช่วงข้อมูลตั้งแต่ 1 ถึง 10,000 ใช้เครื่องคอมพิวเตอร์ที่มีหน่วยประมวลผลกลางแบบ 4 แกนหลัก และทำการพัฒนาด้วยภาษาเออร์แลง อันเป็นภาษาเชิงหน้าที่ (functional language) ที่เหมาะกับงานประมวลผลแบบขนาน

การทดลองเพื่อหาประสิทธิภาพด้านเวลาการศึกษานี้กำหนดจำนวนการประมวลผล 4 การ

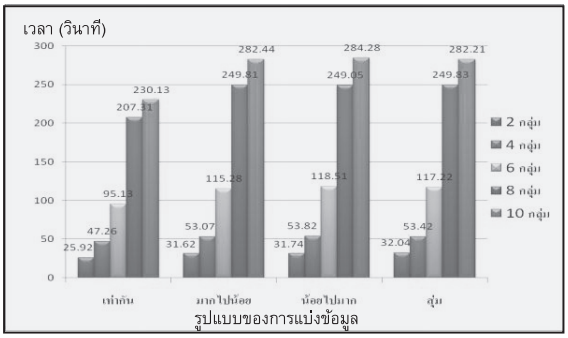
ประมวลผล เพราะการแบ่งข้อมูลของแต่ละการประมวลผลชัดเจนมากกว่าแบ่งเป็น 8 การประมวลผล

กรณีของข้อมูลมีหลายมิติตั้งแต่ 2 มิติ ถึง 5 มิติ และกำหนด 4 จำนวนกลุ่มแสดงดังภาพที่ 4 พบว่า เมื่อจำนวนมิติเพิ่มขึ้นเป็นผลให้เวลาประมวลผลมากขึ้น สำหรับการแบ่งจำนวนข้อมูลรูปแบบเท่ากัน ทุกมิติใช้เวลาประมวลผลน้อยสุด ส่วน 3 รูปแบบที่เหลือใช้เวลาใกล้เคียงกัน



ภาพที่ 4 การเปรียบเทียบประสิทธิภาพด้านเวลาของแต่ละมิติที่มีรูปแบบการแบ่งข้อมูลต่างกัน

กรณีของการกำหนดจำนวนกลุ่มข้อมูล ทดลองกับข้อมูลลักษณะ 2 มิติ และกำหนดจำนวนกลุ่มข้อมูลตั้งแต่ 2 ถึง 10 กลุ่ม โดยเพิ่มขึ้นทีละ 2 เป็น 2, 4, 6,...,0 ตามลำดับ ผลลัพธ์แสดงดังภาพที่ 5 เมื่อจำนวนกลุ่มเพิ่มขึ้น เวลาประมวลผลมากขึ้น ส่วนการแบ่งจำนวนข้อมูลในรูปแบบเท่ากันใช้เวลาประมวลผลน้อยสุด ส่วน 3 รูปแบบที่เหลือใช้เวลาใกล้เคียงกัน



ภาพที่ 5 การเปรียบเทียบประสิทธิภาพด้านเวลาของแต่ละกลุ่มที่มีรูปแบบการแบ่งข้อมูลต่างกัน

สรุปและอภิปรายผล

งานวิจัยนี้เป็นการเตรียมข้อมูลก่อนการประมวลผลเพื่อหารูปแบบที่ดีที่สุดในการแบ่งจำนวนข้อมูล สำหรับใช้ในการจัดกลุ่มข้อมูลที่มีการประมวลผลแบบขนานด้วยขั้นตอนเคมีนส์ โดยทดลองกับข้อมูลสังเคราะห์มิติ ตั้งแต่ 2 มิติ ถึง 5 มิติ และแบ่งจำนวนกลุ่มตั้งแต่ 2 กลุ่มถึง 10 กลุ่มโดยเพิ่มทีละ 2 ผลลัพธ์ทั้งสองวิธีแสดงว่า รูปแบบการแบ่งจำนวนข้อมูลปริมาณข้อมูลมากไปน้อย น้อยไปมาก และแบบสุ่ม ใช้เวลาในการประมวลผลมากกว่าการแบ่งจำนวนข้อมูลรูปแบบเท่ากัน

เอกสารอ้างอิง

1. จุฬาลักษณ์ วัฒนานนท์, อนิราช มิ่งขวัญ. การเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มของข้อมูลด้วยวิธี Correlation Plot. วิชาการพระจอมเกล้าพระนครเหนือ. 2555; 22(1):77-88.
2. ณัฏฐกฤตย์ สงวนดีกุล, ณัฐวุฒิ หนูไพโรจน์. วิธีการแบ่งงานเชิงประสิทธิภาพในขณะทำงานจริงสำหรับระบบประมวลผลที่ประกอบไปด้วยหลายๆ คลัสเตอร์. วิชาการพระจอมเกล้าพระนครเหนือ. 2552;19(3) : 358-366.
3. นเรศ ผ่องสวัสดิ์กุล, จีรพร วีระพันธุ์. อัลกอริทึมเคมีนคลัสเตอร์ริงแบบขนานที่มีประสิทธิภาพบนระบบคลัสเตอร์. ใน: เอกสารการประชุมวิชาการ Knowledge and Smart Technologies สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยบูรพา. ชลบุรี 2551. หน้า 142-147.
4. ประหยัด เลวัน, อัศนีย์ ก่อตระกูล. การรู้จำตัวพิมพ์อักษรภาษาไทยโดยเทคนิคผสม: ฟัซซี่ซีมีนและเคมีน. ใน: การประชุมทางวิชาการของมหาวิทยาลัยเกษตรศาสตร์ ครั้งที่ 41 สาขาวิศวกรรมศาสตร์และสาขาสถาปัตยกรรมศาสตร์. มหาวิทยาลัยเกษตรศาสตร์. กรุงเทพฯ 2551. หน้า 269-276.

5. เมธี โขนงนุช, ชุลีรัตน์ จรัสกุลชัย. การจัดกลุ่มเอกสารแบบขนานโดยขั้นตอนวิธี spherical k-means แบบปรับจุดเริ่มต้น. ใน: เอกสารการประชุมวิชาการมหาวิทยาลัยเกษตรศาสตร์ ครั้งที่ 41 สาขาวิทยาศาสตร์ สาขาการจัดการทรัพยากรและสิ่งแวดล้อม. มหาวิทยาลัยเกษตรศาสตร์. กรุงเทพฯ 2546. หน้า 68-75.
6. Bradley PS, Bennett KP, Demiriz A. Constrained k-means clustering. [serial online] 2000 May; [9 Screens] Available From: URL:http:// machinelearning 102.pbworks.com/f/ConstrainedKMeanstr-2000-65. pdf August 12, 2012.
7. Kerdprasop K, Kerdprasop N. A lightweight method to parallel k-means clustering. mathematics and computers in simulation. 2003;4(4):144-153.
8. Kijsipongse E, U-ruekolan S. Dynamic load balancing on GPU clusters for large-scale k-means clustering. In: 9th International joint conference on computer science and software engineering; 2010. p.346-350.
9. Liang H, Faner M, Ming H. A dynamic load balancing system based on data migration In: 8th International conference on computer supported cooperative work in design proceedings; 2003. p. 493-499.
10. Vesanto J, Alhoniemi E. Clustering of the self-organizing map. IEEE Trans Neural Network 2000;(11)3:586–600.
11. Youssef M. Marzouk, Ahmed F. Ghoniem. k-means clustering for optimal domain decomposition of parallel hierarchical N-body simulations. Computational Physics. 2003;207:493–528.
12. Zhang B, Hsu M, Dayal U. K-harmonic means - a data clustering algorithm. Hewlett-Packard company; 1999.
13. Zhang Y, Xiong Z, Mao J, Ou L. The study of parallel k-means algorithm, In: proceedings of the 6th world congress on intelligent control and automation; 2006. p. 5868-5871.