การจัดกลุ่มเอกสารแบบขนานโดยขั้นตอนวิธี Spherical K-Means แบบปรับจุดเริ่มต้น

Parallel Refining Initial Point Spherical K-Means for Document Clustering

เมธี โขนงนุช[1] **และ ชุลีรัตน์ จรัสกุลชัย**[1]

Metee Khanongnuch[1] and Chuleerat Jaruskulchai[1]

บทคัดย่อ

ถึงแม้ว่าการจัดกลุ่มเอกสารด้วยขั้นตอนวิธี Spherical k-means มีความซับซ้อนของเวลาในการประมวลผลเป็นแบบเชิงเส้น แต่เวลาที่ใช้ในการประมวลผลไม่เหมาะสำหรับข้อมูลขนาดใหญ่ ดังนั้นขั้นตอนวิธี Spherical k-means ได้รับการออกแบบด้วยขั้นตอนวิธีการประมวลผลแบบขนาน เพื่อมุ่งเน้นความสามารถในการรองรับข้อมูลขนาดใหญ่ และลดเวลาในการประมวลผล ซึ่งผลลัพธ์ที่ได้สามารถลดเวลาในการจัดกลุ่มเอกสารได้อย่างมาก อย่างไรก็ตามการจัดกลุ่มเอกสารด้วยขั้นตอนวิธี Spherical k-means เป็นการประมวลผลแบบ iterative ผลลัพธ์ที่ได้มีความถูกต้องมากหรือน้อยขึ้นอยู่กับการกำหนดค่าจุดเริ่มต้น เนื่องจากขั้นตอนวิธี Spherical k-means ใช้กรรมวิธีแบบสุ่มในการหาจุดเริ่มต้น ซึ่งเป็นจุดที่เราสามารถปรับปรุงให้ดีขึ้นได้อีก ในงานวิจัยนี้นำเสนอขั้นตอนวิธีการปรับจุดเริ่มต้นการจัดกลุ่มเอกสารบนพื้นฐานขั้นตอนวิธี Spherical k-means โดยได้รับการออกแบบเป็นขั้นตอนวิธีการประมวลผลแบบขนาน จากผลการทดลองเราได้ความถูกต้องจากการจัดกลุ่มเอกสารเพิ่มขึ้นอีก 7 เปอร์เซ็นต์ แสดงให้เห็นว่าขั้นตอนวิธีการปรับจุดเริ่มต้นการจัดกลุ่มเอกสารบนพื้นฐานขั้นตอนวิธี Spherical k-means สามารถออกแบบให้ประมวลผลแบบขนาน และเพิ่มความถูกต้องให้กับการจัดกลุ่มเอกสารได้

ABSTRACT

Although the Spherical k-means algorithm gave a linear time complexity, it is not appropriate for large scale data size. Hence, the parallel algorithm has been applied to the Spherical k-means for large scale data size, and to reduce the computation time. However, the Spherical k-means use an iterative procedure, it is known that, the accuracy of results are especially sensitive to initial staring point. Due to the Spherical k-means used random algorithm to setup the initial starting point, there is a room for improving the initial starting point. In this paper presents a procedure for refined starting point based on Spherical k-means. Furthermore, the Refining Initial Point implement in parallel algorithm. Experimental result shows that the accuracy of documents clustering is increased up to 7%. This means the improvement of the accuracy of documents clustering can be obtained by applying Parallel Refining Initial Point Spherical k-means algorithm.

---

1. ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ Department of Computer Science, Faculty of Science, Kasetsart University

คำนำ

ในปัจจุบันปริมาณเอกสารมีแนวโน้มที่เพิ่มขึ้นอย่างรวดเร็ว เช่นการรวบรวมหน้าเว็บจากเซิร์ฟเวอร์ต่าง ๆ ทั่วโลกของเว็บเสิร์จเอ็นจิน ส่งผลให้ไม่สามารถแยกและจัดกลุ่มเอกสารต่าง ๆ ได้ด้วยกำลังของมนุษย์และถึงแม้ว่าจะทำได้ก็ต้องใช้ทรัพยากรมนุษย์จำนวนมากแต่ที่สำคัญอาจไม่ทันต่อการใช้งาน ดังนั้นการจัดกลุ่มเอกสารด้วยคอมพิวเตอร์จึงเข้ามามีบทบาทที่สำคัญและได้รับการพัฒนาให้มีประสิทธิภาพมากขึ้นเพื่อให้สามารถประมวลผลเอกสารจำนวนมาก ๆ โดยแนวทางหลักในการปรับปรุงคือ การใช้วิธีการประมวลแบบขนานเพื่อเพิ่มความเร็วในการประมวลผล การบีบลดขนาดของข้อมูลเพื่อประหยัดเนื้อที่ในการจัดเก็บและนำขึ้นมาประมวลผล และการเพิ่มประสิทธิภาพในการค้นคืนข้อมูล ซึ่งสามารถวัดได้จากค่าประสิทธิภาพ ค่าประสิทธิผล และค่าความถูกต้อง

การจัดกลุ่มเอกสารซึ่งเป็นหนึ่งในศาสตร์การค้นคืนเอกสาร (Information Retrieval) เป็นการแยกกลุ่มเอกสารตามความสัมพันธ์ของเนื้อหาหรือข้อความภายในเอกสาร ซึ่งสามารถแบ่งวิธีการประมวลผลได้ 2 วิธีใหญ่ ๆ คือ แบบลำดับชั้น (hierarchical) และแบบไม่เป็นลำดับชั้น (nonhierarchical) หรือการแบ่งกลุ่ม (partitioning) (Rasmussen, 1992) จากผลการทดลองประมวลผลด้วยขั้นตอนวิธีแบบลำดับชั้น และแบบไม่เป็นลำดับชั้น เปรียบเทียบกันพบว่า ขั้นตอนวิธีการประมวลผลแบบลำดับชั้นมีความถูกต้องสูงกว่าขั้นตอนวิธีการประมวลผลแบบไม่เป็นลำดับชั้น (Steinbach et al, 2000) อย่างไรก็ตามข้อได้เปรียบของการประมวลผลแบบไม่เป็นลำดับชั้น คือมีค่าความซับซ้อนของเวลาในการประมวลผลต่ำ เมื่อนำมาประมวลผลข้อมูลขนาดใหญ่จะใช้เวลาในการประมวลผลน้อยกว่าขั้นตอนวิธีการประมวลผลแบบลำดับชั้นมาก เนื่องจากเวลาในการประมวลผลต่อจำนวนเอกสารเป็นค่าเชิงเส้น (linear time) สำหรับขั้นตอนวิธีที่นิยมใช้คือ k-means (Cutting et al, 1992) เป็นต้น

ถึงแม้ว่าการจัดกลุ่มโดยใช้การประมวลผลแบบไม่เป็นลำดับชั้นจะใช้เวลาในการประมวลผลแบบเชิงเส้น แต่เมื่อนำไปใช้กับข้อมูลขนาดใหญ่มาก ๆ เวลาที่ใช้ในการประมวลผลอาจจะไม่ทันกับการใช้งานเช่นกัน ดังนั้นจึงมีผู้วิจัยเกี่ยวกับการจัดกลุ่มเอกสารโดยใช้การประมวลผลแบบขนานเพื่อมุ่งเน้นการลดเวลาในการประมวลผลเพื่อให้ได้ผลลัพธ์ในเวลาที่ไม่มากเกินไป ซึ่งในการจัดกลุ่มข้อมูลแบบขนานนี้ได้มีงานวิจัยต่าง ๆ เช่น ในปี 1997 Ruocco และ Frieder (Ruocco and Frieder, 1997) นำเสนอขั้นตอนวิธี single-pass และ single-link มาทำการประมวลผลแบบขนาน ซึ่งใช้โมเดล multiple instruction, multiple data stream (MIMD) โดยประมวลผลบนเครื่อง Intel Paragon และในปี 1999 Dhillon และ Modha (Dhillon and Modha,1999) นำเสนอขั้นตอนวิธี k-means มาประมวลผลแบบขนานโดยใช้โมเดล single instruction, multiple data stream (SIMD) ซึ่งใช้ message-passing interface (MPI) นอกจากนี้ยังมีผู้วิจัยที่นำเอาขั้นตอนวิธี k-means แบบ SIMD มาประยุกต์ใช้อีก เช่น งานวิจัยของ Kantabutra และ Couch (Kantabutra and Couch, 2000) ในปี 2000 ได้ทำการวิจัยบนเครือข่ายของ workstations (network of workstations: NOWs) และการใช้ Spherical k-means (Kruengkrai and Jaruskulchai, 2001) กับการทดลองบนเครือข่ายพิรุณคลัสเตอร์ ของมหาวิทยาลัยเกษตรศาสตร์ ซึ่งผลลัพธ์ที่ได้สามารถเพิ่มความเร็วในการจัดกลุ่มได้อย่างมาก จากงานวิจัยดังกล่าวพบว่าเราสามารถปรับปรุงการจัดกลุ่มเอกสารให้มีความถูกต้องเพิ่มขึ้นได้อีก

งานวิจัยฉบับนี้นำเสนอขั้นตอนวิธีการจัดกลุ่มเอกสารแบบขนาน ซึ่งมีจุดมุ่งหมายเพื่อเพิ่มความถูกต้องในการประมวลผล โดยอยู่บนพื้นฐานขั้นตอนวิธี Spherical k-means โดยใช้โมเดล SIMD และ message-passing interface (Dhillon and Modha,1999) ผสมผสานกับขั้นตอนวิธีการปรับจุดเริ่มต้นการจัดกลุ่ม (Refining Initial Point) ประยุกต์จากขั้นตอนวิธีปรับจุดเริ่มต้นของ Bradley และ Fayyad (Bradley and Fayyad, 1998) โดยเปรียบเทียบผลลัพธ์กับการจัดกลุ่มเอกสารแบบขนานโดยขั้นตอนวิธี Spherical k-means ของ Kruengkrai และ Jaruskulchai (Kruengkrai and Jaruskulchai, 2001)

อุปกรณ์และวิธีการ

อุปกรณ์การทดลอง

ในการทดลองจะเขียนโปรแกรมเป็นภาษา C โดยใช้ message passing interface (MPI) ไลบรารี และประมวลผลบนเครื่องคลัสเตอร์ AMATA (Athlon and Myrinet Advanced Testbed Architecture) (Parallel Research Group, 2002) ของคณะวิศวกรรมศาสตร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์ ซึ่งเป็นการนำเอาเครื่องพีซี (personal computer) มาเชื่อมต่อกันด้วยเครือข่ายความเร็วสูง และให้ทำงานร่วมกันแบบขนาน โดยประมวลผลแบบคลัสเตอร์บนระบบปฏิบัติการลีนุกซ์

การจัดเตรียมเอกสาร

เอกสารที่ใช้ในการทดลองเป็นเอกสารภาษาไทย ซึ่งเป็นข่าวที่นำมาจากหนังสือพิมพ์เดลินิวส์ตั้งแต่วันที่ 1 ก.ย. 2538 ถึง 22 พ.ย. 2538 จำนวน 4800 ข่าว แบ่งเป็นหัวข้อข่าว 4 ชนิดคือ ข่าวเศรษฐกิจ 1146 ข่าว ข่าวต่างประเทศ 1654 ข่าว ข่าวการเมือง 827 ข่าว ข่าวสังคม 1126 ข่าว และข่าวในพระราชสำนัก 47 ข่าว ซึ่ง 1 ข่าว แทน 1 เอกสาร

การนำระบบคอมพิวเตอร์เข้ามาช่วยในการจัดกลุ่มเอกสารนั้น เราต้องทำการจัดเตรียมเอกสารต่าง ๆ ให้อยู่ในรูปแบบที่สามารถประมวลผลได้ง่าย ในขั้นตอนแรกต้องแยกเอกสารออกเป็นคำ ๆ ซึ่งเอกสารภาษาอังกฤษมีการเว้นวรรคอยู่แล้ว แต่สำหรับเอกสารภาษาไทยมีลักษณะการเขียนที่ต่อกันไม่มีการแบ่งวรรคตอนชัดเจน ดังนั้นการตัดคำภาษาไทยจึงเป็นเรื่องที่สำคัญเรื่องหนึ่ง จากการทดลองของ Kruengkrai และ Jaruskulchai พบว่าการตัดคำด้วยขั้นตอนวิธี Longest Matching (LM) ให้ผลการจัดกลุ่มเอกสารที่ดีกว่าการตัดคำด้วยขั้นตอนวิธีกฎทางภาษาศาสตร์ (Kruengkrai and Jaruskulchai, 2001) ดังนั้นในการทดลองเราใช้โปรแกรมตัดคำของ Kruengkrai และ Jaruskulchai ซึ่งใช้ขั้นตอนวิธี LM (Sornlertlamvanich, 1993) โดยใช้รายการคำไทยจำนวน 32,675 คำ จากนั้นเราจะแทนเอกสารที่ไม่มีโครงสร้างแต่ละฉบับด้วย Vector Space Model (VSM) (Salton, Wong and Yang, 1975) ซึ่งเป็นเวคเตอร์ของคำที่ไม่ซ้ำกันที่ปรากฏอยู่ในเอกสารทั้งหมด จากนั้นจะดึงคำ stop word ออกจากเอกสาร เช่น “a” “and” “the” “และ” “หรือ” และพิจารณาตัดคำที่ปรากฏในเอกสารต่าง ๆ ที่มีความถี่มากเกินไปหรือน้อยเกินไปออกจากเอกสาร เพื่อไม่ให้เอกสารมีมิติมากเกินไป จากนั้นให้น้ำหนักกับคำแต่ละคำ (Salton and Allan. 1993) ในแต่ละเอกสาร โดยวิธีที่นิยมใช้คือ normalized term frequency inverse document frequency

การจัดกลุ่มเอกสาร

ในการจัดกลุ่มเอกสารเราแบ่งเป็น 2 ขั้นตอน คือ ขั้นตอนการหาจุดเริ่มต้น (Initial) และขั้นตอนการจัดกลุ่ม (Clustering) โดยขั้นตอนวิธี Spherical k-means ในขั้นตอนการหาจุดเริ่มต้น เป็นการแบ่งเอกสารทั้งหมดออกเป็นกลุ่ม ๆ โดยการสุ่ม จากนั้นคำนวณหาจุดศูนย์กลางของเอกสารแต่ละกลุ่ม เพื่อใช้เป็นตัวแทนเริ่มต้นในการแบ่งเอกสาร (Dhillon and Modha,1999) สำหรับการเพิ่มความถูกต้องในการประมวลผลเราจะนำขั้นตอนวิธีการปรับจุดเริ่มต้นการจัดกลุ่ม ซึ่งประยุกต์มาจากขั้นตอนวิธีปรับจุดเริ่มต้นของ Bradley และ Fayyad (Bradley and Fayyad, 1998) เพื่อหาจุดเริ่มต้นการประมวลผล โดยในขั้นตอนวิธีการปรับจุดเริ่มต้นการจัดกลุ่มเป็นการสุ่มเอกสารบางส่วนขึ้นมาแบ่งเป็นกลุ่ม ๆ และใช้ขั้นตอนวิธี Spherical k-means คำนวณหาจุดศูนย์กลางของเอกสารแต่ละกลุ่ม ทำเช่นนี้หลาย ๆ รอบ ทำให้ได้จุดศูนย์กลางหลาย ๆ จุด จากนั้นนำจุดศูนย์กลางที่ได้ทั้งหมดมาผ่านขั้นตอนวิธี Spherical k-means อีกครั้งเพื่อเฉลี่ยหาจุดศูนย์กลางที่ดีที่สุด และใช้เป็นตัวแทนเริ่มต้นในการแบ่งเอกสาร ซึ่งในขั้นตอนวิธีการปรับจุดเริ่มต้นของการจัดกลุ่มและขั้นตอนการจัดกลุ่มเอกสาร เราใช้การประมวลผลแบบขนานเพื่อลดเวลาในการประมวลผล โดยการแบ่งเอกสารออกเป็นชุดย่อย ๆ เท่า ๆ กันให้แต่ละโพรเซสเซอร์ประมวลผล และเรียกใช้ฟังก์ชัน MPI_Allreduce (Dhillon and Modha,1999) เพื่อรวมผลลัพธ์ที่ได้และส่งกลับคืนไปยังทุกโพรเซสเซอร์

การวัดประสิทธิภาพของการจัดกลุ่ม

การวัดประสิทธิภาพของการจัดกลุ่มเอกสาร เราวัดด้วย F-measure metric (Larsen and Anoe, 1999) โดยวัดจากค่าความถูกต้อง (Precision, P) ค่าระลึกได้ (Recall, R) และนำมาคำนวณออกมาเป็นค่า F-measure มีสมการดังนี้

$$P_{k,t} = \frac{\text{จำนวนเอกสารหัวเรื่อง t ในกลุ่ม k}}{\text{จำนวนเอกสารทั้งหมดในกลุ่ม k}} \qquad \text{สมการที่ (1)}$$

$$R_{k,t} = \frac{\text{จำนวนเอกสารหัวเรื่อง t ในกลุ่ม k}}{\text{จำนวนเอกสารหัวเรื่อง t ทั้งหมด}} \qquad \text{สมการที่ (2)}$$

$$F_{k,t} = \frac{2\left(P_{k,t} R_{k,t}\right)}{P_{k,t} + R_{k,t}} \qquad \text{สมการที่ (3)}$$

$$F_{Total} = \frac{\sum_{t \in T} \left(|t| \max_k F_{k,t} \right)}{\sum_{t \in T} |t|} \qquad \text{สมการที่ (4)}$$

โดยที่ | t | เป็นจำนวนเอกสารในหัวเรื่อง t และ $F_{k,t}$ เป็นค่า F ของหัวเรื่อง t ในกลุ่ม k หากผลลัพธ์ในการจัดกลุ่มเอกสารมีความถูกต้องสูง ค่า F ที่ได้จะมีค่าสูง

สำหรับการวัดเวลาในการประมวลผลในการเขียนโปรแกรมขนานแบบ MPI จะใช้ฟังก์ชั่น MPI_Wtime() เป็นตัวจับเวลาในการประมวลผล (Dhillon and Modha,1999) ซึ่งในการทดลองนี้เวลาที่วัดจะไม่รวมเวลาของ I/O หรือเวลาในการอ่านข้อมูลจากฮาร์ดดิสก์

ผลการทดลองและวิจารณ์

ในการวัดประสิทธิภาพและวัดเวลาในการจัดกลุ่มซึ่งเป็นการเปรียบเทียบความถูกต้องและเวลาที่ใช้ในการประมวลผลแบบขนานระหว่างขั้นตอนวิธี Spherical k-means และขั้นตอนวิธี Spherical k-means แบบปรับจุดเริ่มต้นของการจัดกลุ่ม โดยแต่ละเงื่อนไขจะทำการทดลองแบบขนาน 2 โพรเซสเซอร์เป็นจำนวน 10 ครั้ง แล้วนำค่าที่ได้มาหาค่าเฉลี่ย ซึ่งในการทดลองนี้เวลาที่วัดจะไม่รวมเวลาของ I/O หรือเวลาในการอ่านข้อมูลจากฮาร์ดดิสก์

เงื่อนไขแรกเป็นการทดสอบผลตอบสนองของขั้นตอนวิธี Spherical k-means แบบปรับจุดเริ่มต้นการจัดกลุ่ม ต่อจำนวนตัวอย่างเอกสารที่ต่างกัน โดยกำหนดให้จำนวนรอบในการสุ่มตัวอย่างเป็น 10 รอบคงที่ และเปลี่ยนจำนวนตัวอย่างที่สุ่มเป็น 5% 10% 15% และ 20% ตามลำดับ ซึ่งการเปรียบเทียบความถูกต้องในการจัดกลุ่มเอกสารแสดงในตารางที่ 1 และการเปรียบเทียบเวลาในการประมวลผลแสดงในตารางที่ 2

Table 1 Comparison of F-Measure by variation of Number of Documents

F-Measure	Spherical k-means	Refining initial point Spherical k-means			
		s = 5%	s = 10%	s = 15%	s = 20%
Mean	0.7138123	**0.7650168**	**0.7654902**	**0.7660425**	**0.781296**
Max	0.7519461	0.8131706	0.8342100	0.8003366	0.8243215
Min	0.6130865	0.7450032	0.7437408	0.74332	0.7479487

Table 2 Comparison of Processing Time by Variation Number of Documents

Time (Sec)	Spherical k-means	Refining initial point Spherical k-means			
		s = 5%	s = 10%	s = 15%	s = 20%
Initial	1.6041	25.7067	69.6603	99.7445	137.0711
Clustering	126.1369	**70.2335**	**67.6328**	**65.4918**	**53.2102**
Total	127.7410	**95.9402**	**109.1841**	118.1496	190.6513

เงื่อนไขที่สองเป็นการทดสอบผลตอบสนองของขั้นตอนวิธี Spherical k-means แบบปรับจุดเริ่มต้นการจัดกลุ่ม ต่อจำนวนรอบการสุ่มตัวอย่างที่ต่างกัน โดยกำหนดให้จำนวนตัวอย่างที่สุ่มเป็น 10% คงที่และเปลี่ยนจำนวนรอบในการสุ่มเป็น 5 รอบ 10 รอบ 15 รอบ และ 20 รอบตามลำดับ ซึ่งการเปรียบเทียบความถูกต้องในการจัดกลุ่มเอกสารแสดงในตารางที่ 3 และการเปรียบเทียบเวลาในการประมวลผลแสดงในตารางที่ 4

Table 3 Comparison of F-Measure by Variation of Number of Iterations

F-Measure	Spherical k-means	Refining initial point Spherical k-means			
		I = 5	I = 10	I = 15	I = 20
Mean	0.7138123	**0.7637019**	**0.7654902**	**0.7766148**	**0.7834789**
Max	0.7519461	0.8119083	0.8342100	0.8091732	0.8386282
Min	0.6130865	0.7332211	0.7437408	0.7534189	0.7519461

Table 4 Comparison of Processing Time by Variation of Number of Iterations

Time (Sec)	Spherical k-means	Refining initial point Spherical k-means			
		I = 5	I = 10	I = 15	I = 20
Initial	1.6041	30.8270	69.6603	94.8359	130.5458
Clustering	126.1369	**68.9700**	**67.6328**	**61.2517**	**59.1563**
Total	127.7410	**99.7950**	**109.1841**	156.0877	189.7022

การจัดกลุ่มเอกสารโดยการใช้ขั้นตอนวิธี Spherical k-means แบบปรับจุดเริ่มต้นการจัดกลุ่ม ได้ผลลัพธ์ความถูกต้องดีกว่าการจัดกลุ่มเอกสารด้วยขั้นตอนวิธี Spherical k-means จากการทดลองความถูกต้องเพิ่มขึ้น 5 ถึง 7 เปอร์เซ็นต์ โดยพิจารณาจากค่า F-Measure ทั้งนี้ผลของความถูกต้องจะเพิ่มขึ้นถ้าเพิ่มจำนวนตัวอย่างเอกสาร ดังแสดงในตารางที่ 1 หรือเพิ่มจำนวนรอบในการสุ่มตัวอย่าง ดังแสดงในตารางที่ 3 ซึ่งสรุปได้ว่าความถูกต้องของการจัดกลุ่มเอกสารโดยการใช้ขั้นตอนวิธี Spherical k-means แบบปรับจุดเริ่มต้นการจัดกลุ่มจะแปรผันตามจำนวนตัวอย่างเอกสาร

สำหรับเวลาที่ใช้ในการประมวลผล เมื่อพิจารณาเฉพาะเวลาในการจัดกลุ่มเอกสาร (Clustering) พบว่าการจัดกลุ่มเอกสารโดยการใช้ขั้นตอนวิธี Spherical k-means แบบปรับจุดเริ่มต้นการจัดกลุ่ม ใช้เวลาในการจัดกลุ่มเอกสารน้อยกว่าการจัดกลุ่มเอกสารแบบขั้นตอนวิธี Spherical k-means ซึ่งแสดงให้เห็นว่าการใช้จุดเริ่มต้นของการประมวลผลที่ดีมีผลให้การประมวลผลลู่เข้าสู่คำตอบได้เร็วขึ้น ดังแสดงในตารางที่ 2 และตารางที่ 4 เมื่อพิจารณาเวลารวม (Total) ในการจัดกลุ่ม พบว่าในกรณีการใช้จำนวนตัวอย่างเอกสารหรือจำนวนรอบของการสุ่มตัวอย่างน้อย ๆ ขั้นตอนวิธี Spherical k-means แบบปรับจุดเริ่มต้นการจัดกลุ่ม ใช้เวลารวมในการจัดกลุ่มน้อยกว่าขั้นตอนวิธี Spherical k-means สำหรับกรณีที่ใช้จำนวนตัวอย่างเอกสารหรือจำนวนรอบของการสุ่มตัวอย่างที่มากขึ้น การจัดกลุ่มเอกสารโดยการใช้ขั้นตอนวิธี Spherical k-means แบบปรับจุดเริ่มต้นการจัดกลุ่ม ต้องเสียเวลาในการคำนวณหาจุดเริ่มต้น ซึ่งเวลาที่ใช้ขึ้นอยู่กับปริมาณตัวอย่างเอกสาร ในขณะที่ขั้นตอนวิธี Spherical k-means ใช้เวลาคำนวณหาจุดเริ่มต้นน้อยมาก เมื่อรวมเวลาทั้งสองส่วนคือในส่วนคำนวณหาจุดเริ่มต้น และเวลาในการจัดกลุ่มเอกสารแล้วพบว่า การจัดกลุ่มเอกสารโดยการใช้ขั้นตอนวิธี Spherical k-means แบบปรับจุดเริ่มต้นการจัดกลุ่ม ใช้เวลารวมทั้งหมดมากกว่าการจัดกลุ่มแบบ Spherical k-means อย่างไรก็ตามเวลาที่เสียไปนั้นเพื่อแลกกับความถูกต้องในการจัดกลุ่มเอกสารนั่นเอง

ในการวัดเวลาเพื่อหาค่า speedup ของขั้นตอนวิธี Spherical k-means แบบปรับจุดเริ่มต้นการจัดกลุ่ม เรากำหนดให้จำนวนตัวอย่างเป็น 10% และทำการสุ่ม 10 รอบ ซึ่งทดลองประมวลผลโดยใช้จำนวนโพรเซสเซอร์ 1 โพรเซสเซอร์ 2 โพรเซสเซอร์ 4 โพรเซสเซอร์ และ 8 โพรเซสเซอร์ตามลำดับ และคำนวณหาค่า speedup จากเวลาประมวลผลใน 1 โพรเซสเซอร์ หารด้วยเวลาในการประมวลผลด้วยจำนวนโพรเซสเซอร์ต่าง ๆ เวลาที่ใช้ในการประมวลผลแสดงดังรูปที่ 1 สำหรับค่า speedup แสดงดังรูปที่ 2

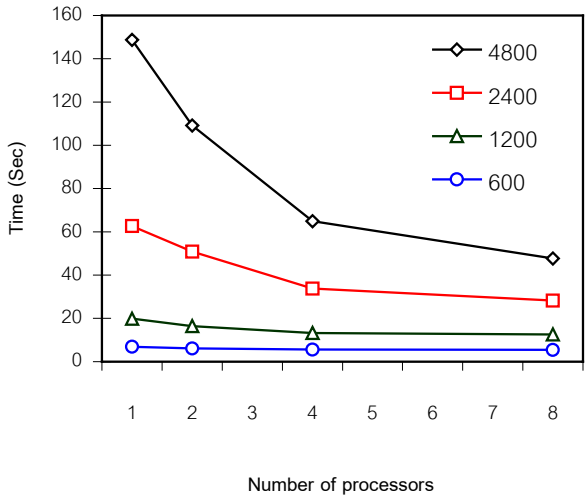

Figure 1 Comparison of Time and Number of Documents

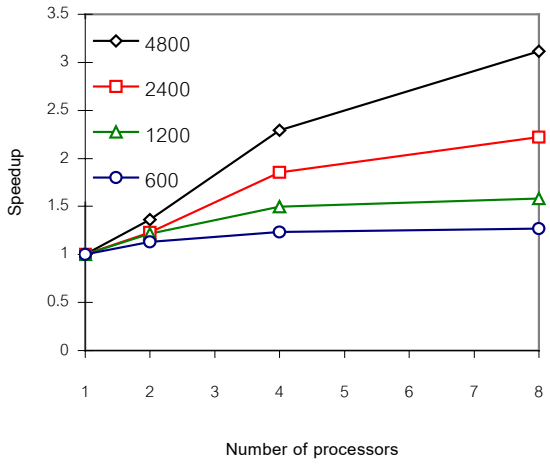

Figure 2 Parallel Speedup to execute on different Document Size, 600,1200,2400 and 4800

สำหรับประสิทธิภาพในการประมวลผลแบบขนานตามรูปที่ 2 จากจำนวนเอกสาร 600 เอกสาร พบว่าการเพิ่มจำนวนโพรเซสเซอร์จนถึง 8 โพรเซสเซอร์ ค่า speedup เกือบจะไม่มีการเปลี่ยนแปลง แต่เมื่อเพิ่มขนาดของปัญหาให้มีขนาดใหญ่ขึ้นเป็น 1200 เอกสาร 2400 เอกสาร และ 4800 เอกสาร ตามลำดับพบว่าค่า speedup เพิ่มขึ้นเรื่อย ๆ ดังนั้นสรุปได้ว่าหากเอกสารมีปริมาณน้อยการเพิ่มจำนวนโพรเซสเซอร์ไม่มีผลต่อความเร็วในการจัดกลุ่มเอกสาร แต่เมื่อเอกสารมีปริมาณมาก ๆ เราสามารถลดเวลาของการประมวลผลได้โดยการเพิ่มจำนวนโพรเซสเซอร์

สรุป

งานวิจัยนี้ได้นำเสนอขั้นตอนวิธี Spherical k-means แบบปรับจุดเริ่มต้นของการจัดกลุ่ม ในการจัดกลุ่มเอกสารแบบขนาน เพื่อเพิ่มถูกต้องในการจัดกลุ่มเอกสาร โดยทำการประมวลผลบน AMATA คลัสเตอร์ จากผลการทดลองจัดกลุ่มเอกสารภาษาไทย ซึ่งเป็นข่าวที่นำมาจากหนังสือพิมพ์เดลินิวส์จำนวน 4800 ข่าว พบว่าเราสามารถเพิ่มความถูกต้องการจัดกลุ่มเอกสารจากขั้นตอนวิธี Spherical k-means ได้อีก 5 ถึง 7 เปอร์เซ็นต์

เอกสารอ้างอิง

Bradley, P.S. and U.M. Fayyad. 1998. Refining Initial Points for K-Means Clustering, Proceedings of the Fifteenth International Conference on Machine Learning ICML98. pp. 91-99.

Cutting, D.R., D.R. Karger, J.O. Pedersen, and J.W. Tukey. 1992. Scatter/Gather: Acluster-based Approach to Browsing Large Document Collections. Proceedings of SIGIR'92.

Dhillon, I.S., and D. Modha. 1999. A Data-Clustering Algorithm on Distributed Memory Multiprocessors. Large-Scale Parallel Data Mining. pp. 245-260.

Gerard, W.A. and C.S. Yang. 1975. A Vector Space Model for Automatic Indexing, CACM 18(11). pp. 613-620

Kantabutra, S. and A.L. Couch. 2000. Parallel k-means Clustering Algorithm on NOWs. NECTEC Technical Journal Vol.1 No. 6

Kruengkrai, C.and C. Jaruskulchai. 2001. Thai Text Document Clustering Using Parallel Spherical K-Means Algorithm on PIRUN Linux Cluster, The Fifth National Computer Science and Engineering Conference.

Larsen, B. and C. Anoe. 1999. Fast and Effective Text Mining Using Linear-time Document Clustering. KDD-99, San Diego, California.

Parallel Research Group. 2002. .Athlon and Myrinet Advanced Testbed Architecture (AMATA). Computer Engineering, Kasetsart University, Thailand. Available Source: http://amata.cpe.ku.ac.th/

Rasmussen, E.M. 1992. Clustering Algorithms. *In* W.B. Frakes and R. Baeza-Yates. Information Retrieval: Data Structures and Algorithms, Prentice Hall. pp. 419-442

Ruocco, A. and O. Frieder. 1997. Clustering and classification of large document bases in a parallel environment. Journal of the American Society for Information Science pp. 932-943.

Salton, G. and J. Allan. 1993. Selective Text Utilization and Text Traversal. Proceedings of Hypertext'93. pp 131-144.

Sornlertlamvanich, V. 1993. Word Segmentation for Thai in Machine Translation System. Machine Translation, Nation Electronics and Computer Technology Center. Bangkok. pp. 50-56.

Steinbach, M., G. Karypis, and V. Kumar. 2000. A Comparison of Document Clustering Techniques. *In* KDD Workshop on Text Mining