บทที่ 1
บทนำ

การค้นคว้าแบบอิสระเรื่องการจัดกลุ่มบทคัดย่อทางวิชาการด้วยอัลกอริทึมเคมีเดียนส์ ศึกษาและค้นคว้าขึ้นมาโดยมีมูลเหตุจูงใจ และประเด็นที่สนใจดังนี้

- ที่มาและความสำคัญของปัญหาที่นำไปสู่การค้นคว้าวิจัย
- สรุปสาระสำคัญจากเอกสารที่เกี่ยวข้อง
- หลักการ ทฤษฎี เหตุผล และ/หรือสมมุติฐาน
- วัตถุประสงค์ของการศึกษา
- ขอบเขตของการศึกษา

1.1 ที่มาและความสำคัญของปัญหาที่นำไปสู่การค้นคว้าวิจัย

ปัจจุบันบทความทางวิชาการมีจำนวนมากมาย และมีแนวโน้มเพิ่มขึ้นเรื่อยๆ ในการศึกษาค้นคว้าข้อมูล บางครั้งต้องใช้บทความทางวิชาการจำนวนมาก เวลาในการค้นหาและคัดเลือกบทความทางวิชาการที่ตรงกับความต้องการก็จะมากตามจำนวนบทความที่มีอยู่ ทำให้เสียเวลาและเกิดความล่าช้าในการทำงาน จึงได้คิดหาวิธีการที่จะใช้เพื่อลดเวลาในส่วนนี้ลง การศึกษาค้นคว้าข้อมูลก็จะได้มีความรวดเร็วยิ่งขึ้น โดยจะทำการจัดกลุ่มบทคัดย่อของบทความออกเป็นกลุ่ม ๆ ตามความคล้ายคลึงกันของบทคัดย่อ หากต้องการค้นหาบทความเกี่ยวกับเรื่องใด ก็สามารถหาได้จากกลุ่มที่ได้ทำการแบ่งไว้ ซึ่งจะเห็นว่าสามารถลดเวลาในการค้นหาบทความที่ต้องการได้มาก เพราะไม่ต้องทำการค้นหาบทความที่ต้องการจากบทความที่มีอยู่ทั้งหมด แต่สามารถหาได้จากกลุ่มที่แบ่งไว้ซึ่งมีจำนวนบทความน้อยกว่า

1.2 สรุปสาระสำคัญจากเอกสารที่เกี่ยวข้อง

1.2.1 อัลกอริทึมเคมีเดียนส์

อัลกอริทึมเคมีเดียนส์ (K-Medians Algorithm) เป็นอัลกอริทึมในการจัดกลุ่มข้อมูลตามความคล้ายคลึงกันของข้อมูล โดยที่ความคล้ายคลึงกันของข้อมูลจะพิจารณาจากจุดศูนย์กลางของกลุ่มข้อมูลหรือที่เรียกว่าจุดมีเดียน (Median) เพื่อให้การอธิบายอัลกอริทึมเคมีเดียนส์ได้สะดวกขึ้น ขอนิยามคำศัพท์ที่เกี่ยวข้องดังต่อไปนี้

นิยามของจุดมีเดียน : ให้ S เป็นเซตของจุดจำนวน N จุด ในมิติขนาด D และจุด x ที่เป็นสมาชิกของ S จะเป็นจุดศูนย์กลาง หรือที่เรียกว่าจุดมีเดียนของ S ก็ต่อเมื่อผลรวมของระยะทางระหว่าง x กับจุดอื่นๆ ใน S มีค่าน้อยที่สุด

นิยามของค่าผิดพลาดยกกำลังสอง (Squared-Error Function) : ให้ S เป็นเซตของจุดจำนวน N จุด ในมิติขนาด D และ k เป็นจำนวนเต็ม จะได้ค่าผิดพลาดยกกำลังสอง $E = \sum_{i=1}^{k} \sum_{x \in Ci} d(C_i^m, x)$ โดยที่ C_i^m คือจุดมีเดียนของกลุ่มข้อมูลที่ i

นิยามของปัญหาการจัดกลุ่มด้วยอัลกอริทึมเคมีเดียนส์ (K-Medians Clustering Algorithm) : ให้ S เป็นเซตของจุดจำนวน N จุด ในมิติขนาด D และ k เป็นจำนวนเต็ม ปัญหาการจัดกลุ่มด้วยอัลกอริทึมเคมีเดียนส์เป็นการแบ่งเซต S เป็นกลุ่มข้อมูล k กลุ่ม C_1, C_2, C_3,..., C_k โดยให้มีค่าผิดพลาดยกกำลังสองน้อยที่สุด

อัลกอริทึมเคมีเดียนส์มีหลักการคือ แบ่งข้อมูลออกเป็น k กลุ่ม โดยการแบ่งกลุ่มจะพิจารณาความคล้ายคลึงกันของข้อมูลด้วยระยะทางระหว่างจุดข้อมูลกับจุดมีเดียนของกลุ่มข้อมูลนั้น โดยจุดมีเดียนของกลุ่มข้อมูลจะเป็นสมาชิกในกลุ่มข้อมูลนั้นที่มีผลรวมของระยะทางระหว่างจุดข้อมูลในกลุ่มกับจุดมีเดียนน้อยที่สุด โดยจะมีขั้นตอนการทำงานคือ

(1) ทำการรับค่า k จากผู้ใช้แล้วสุ่มแบ่งกลุ่มเป็น k กลุ่ม

(2) ตรวจสอบแต่ละจุดข้อมูลว่าระยะทางระหว่างจุดข้อมูลกับจุดมีเดียนของกลุ่มข้อมูลใดน้อยที่สุด แสดงว่ามีความคล้ายคลึงกันมากที่สุด แล้วจึงกำหนดให้จุดข้อมูลนั้นเป็นสมาชิกของกลุ่มข้อมูลที่คล้ายคลึงกันมากที่สุด

(3) ถ้ามีการย้ายกลุ่มของจุดข้อมูล จะต้องมีการคำนวณหาจุดมีเดียนของแต่ละกลุ่มข้อมูลใหม่

(4) วนทำตามขั้นตอนที่ 2 และ 3 ไปจนกระทั่งไม่มีการย้ายกลุ่มของจุดข้อมูล ก็จะได้กลุ่มของข้อมูลที่มีความคล้ายคลึงกัน จำนวน k กลุ่ม

(Sanpawat Kantabutra, 2001)

1.2.2 อัลกอริทึมเคมีนส์

อัลกอริทึมเคมีนส์ (K-Means Algorithm) เป็นอัลกอริทึมในการจัดกลุ่มข้อมูลตามความคล้ายคลึงกันของข้อมูล โดยที่ความคล้ายคลึงกันของข้อมูลจะพิจารณาจากจุดศูนย์กลางของกลุ่ม หรือที่เรียกว่าจุดมีน (Mean) เพื่อให้การอธิบายอัลกอริทึมเคมีนส์ได้สะดวกขึ้น ขอนิยามคำศัพท์ที่เกี่ยวข้องดังต่อไปนี้

นิยามของจุดมีน : ให้ S เป็นเซตของจุดจำนวน N จุด ในมิติขนาด D และจุด x เป็นสมาชิกของ S จะได้ว่า m เป็นจุดศูนย์กลางของ S โดยเรียกจุดศูนย์กลางนี้ว่าจุดมีน ก็ต่อเมื่อ $m = (1/N)\sum^{N}_{i=1} x_i$

นิยามของปัญหาการจัดกลุ่มด้วยอัลกอริทึมเคมีนส์ (K-Means Clustering Algorithm) : ให้ S เป็นเซตของจุดจำนวน N จุด ในมิติขนาด D และ k เป็นจำนวนเต็ม ปัญหาการจัดกลุ่มด้วยอัลกอริทึมเคมีนส์เป็นการแบ่งเซต S เป็นกลุ่มข้อมูล k กลุ่ม $C_1, C_2, C_3, \ldots, C_k$ โดยให้มีค่าผิดพลาดยกกำลังสองน้อยที่สุด

อัลกอริทึมเคมีนส์มีหลักการคือ แบ่งข้อมูลออกเป็น k กลุ่ม โดยการแบ่งกลุ่มจะพิจารณาความคล้ายคลึงกันของข้อมูลด้วยระยะทางระหว่างจุดข้อมูลกับจุดมีนของกลุ่มข้อมูลนั้น โดยจะมีขั้นตอนการทำงานคือ

(1) ทำการรับค่า k จากผู้ใช้แล้วสุ่มแบ่งกลุ่มเป็น k กลุ่ม

(2) ตรวจสอบแต่ละจุดข้อมูลว่าระยะทางระหว่างจุดข้อมูลกับจุดมีนของกลุ่มข้อมูลใดน้อยที่สุด แสดงว่ามีความคล้ายคลึงกันมากที่สุด แล้วจึงกำหนดให้จุดข้อมูลนั้นเป็นสมาชิกของกลุ่มข้อมูลที่คล้ายคลึงกันมากที่สุด

(3) ถ้ามีการย้ายกลุ่มของจุดข้อมูล จะต้องมีการคำนวณหาจุดมีนของแต่ละกลุ่มข้อมูลใหม่

(4) วนทำตามขั้นตอนที่ 2 และ 3 ไปจนกระทั่งไม่มีการย้ายกลุ่มของจุดข้อมูล ก็จะได้กลุ่มของข้อมูลที่มีความคล้ายคลึงกัน จำนวน k กลุ่ม

(Jiawei Han and Micheline Kamber, 2001)

1.3 หลักการ ทฤษฎี เหตุผล และ/หรือ สมมติฐาน

การจัดกลุ่มบทคัดย่อออกเป็นกลุ่ม ๆ ตามความคล้ายคลึงกันของบทคัดย่อจะอาศัยวิธีการจัดกลุ่มข้อมูล (Clustering) เข้ามาช่วย โดยการจัดกลุ่มข้อมูล คือการแบ่งข้อมูลออกเป็นกลุ่มๆ โดยที่ข้อมูลในกลุ่มเดียวกันจะมีความคล้ายคลึงกัน (Similarity) มากกว่าข้อมูลที่อยู่คนละกลุ่ม ซึ่งความคล้ายคลึงกันของข้อมูลสามารถวัดได้จากระยะทางระหว่างข้อมูล (Distance) โดยมีอัลกอริทึมที่นิยมใช้ในการจัดกลุ่มข้อมูลอยู่หลายอัลกอริทึม อาทิเช่น อัลกอริทึมเคมีนส์ และอัลกอริทึมเคมีเดียนส์

ในการจัดกลุ่มบทคัดย่อทางวิชาการ ต้องการข้อมูลศูนย์กลางที่เป็นข้อมูลภายในกลุ่มนั้นๆ ซึ่งเป็นข้อมูลที่มีอยู่จริง และอัลกอริทึมเคมีเดียนส์ เป็นอัลกอริทึมที่จุดศูนย์กลางของแต่ละกลุ่มข้อมูล (Cluster) เป็นข้อมูลภายในกลุ่มนั้นๆ และข้อมูลที่เป็นจุดศูนย์กลางจะมีความคล้ายคลึงกับข้อมูล

อื่นๆ ภายในกลุ่มเดียวกันมากที่สุด จึงทำให้ความคล้ายคลึงกันของข้อมูลในกลุ่มที่แบ่งได้มีความถูกต้องและน่าเชื่อถือ

ดังนั้นในการศึกษาครั้งนี้จึงได้นำเอาอัลกอริทึมเคมีเดียนส์มาใช้ในการจัดบทคัดย่อทางวิชาการออกเป็นกลุ่มตามจำนวนกลุ่มที่ต้องการ เพื่อเป็นประโยชน์ในการศึกษาค้นคว้าข้อมูลให้มีความรวดเร็วยิ่งขึ้น

1.4 วัตถุประสงค์ของการศึกษา

เพื่อพัฒนาโปรแกรมจัดกลุ่มบทคัดย่อทางวิชาการตามจำนวนกลุ่มที่กำหนดให้ โดยใช้วิธีการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีเดียนส์

1.5 ขอบเขตของการศึกษา

(1) ศึกษาวิธีการทำงานของอัลกอริทึมเคมีเดียนส์

(2) ศึกษาแนวทางประยุกต์ใช้อัลกอริทึมเคมีเดียนส์กับการจัดกลุ่มบทคัดย่อทางวิชาการ

(3) พัฒนาโปรแกรมที่มีความสามารถในการจัดกลุ่มบทคัดย่อทางวิชาการตามจำนวนกลุ่มที่กำหนดให้ โดยใช้วิธีการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีเดียนส์ ซึ่งบทคัดย่อที่ใช้ในการทดสอบโปรแกรม เป็นบทคัดย่อทางวิชาการที่เกี่ยวกับฐานข้อมูล (Database) มีตั้งแต่ 1 ถึง 1,000 บทคัดย่อ และเป็นแฟ้มข้อมูลเอกสาร (Text File) เท่านั้น

การค้นคว้าแบบอิสระนี้มีวัตถุประสงค์เพื่อพัฒนาโปรแกรมสำหรับช่วยจัดกลุ่มบทคัดย่อทางวิชาการที่มีจำนวนมากออกเป็นกลุ่ม ๆ เพื่อเป็นประโยชน์ในการศึกษาค้นคว้าข้อมูลให้มีความรวดเร็วยิ่งขึ้น โดยได้นำอัลกอริทึมเคมีเดียนส์มาประยุกต์ใช้ในการศึกษาครั้งนี้