การเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มของข้อมูลด้วยวิธี Correlation Plot
A Comparative Efficiency of Correlation Plot Data Classification

จุฬาลักษณ์ วัฒนานนท์[1]* และ อนิราช มิ่งขวัญ[2]
Julaluk Watthananon[1]* and Anirach Mingkhwan[2]

บทคัดย่อ

ปัจจุบันการจำแนกข้อมูลเป็นส่วนหนึ่งของงานหลายชนิด ที่นำมาช่วยจัดการกับข้อมูลที่มีขนาดใหญ่ และคล้ายคลึงกันได้อย่างรวดเร็วและมีประสิทธิภาพ บทความนี้ได้ทำการศึกษาผลการเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มของข้อมูลด้วยวิธี Correlation Plot กับวิธีการอื่นในการใช้ข้อมูลแบบสหความสัมพันธ์ โดยทำการทดลองกับข้อมูลความรู้จำนวน 2 ชุด เกี่ยวกับบทความวิชาการทางคอมพิวเตอร์ และหนังสือที่จัดเก็บในห้องสมุด ซึ่งมีความแตกต่างกันหลากหลายศาสตร์ โดยได้ทำการวิเคราะห์สหความสัมพันธ์ด้วยเทคนิค DDC-MR ผลการทดลองกับข้อมูลความรู้ทั้ง 2 ชุด พบว่าค่าความถูกต้อง (Accuracy) ในการจำแนกกลุ่มของข้อมูลด้วยวิธีการนี้มีค่าเท่ากับ 96.04% และ 90.66% ตามลำดับ ซึ่งใช้เวลาตอบสนองดีกว่าวิธีการอื่นที่ใช้ทดสอบ

คำสำคัญ: การจำแนกกลุ่ม จุดสหความสัมพันธ์ ระบบทศนิยมดิวอี้แบบสหความสัมพันธ์ ระนาบสหความสัมพันธ์ ซัพพอร์ตเวกเตอร์แมชชีน

Abstract

Nowadays, data classification helping many kinds of work that manages large and similar datasets in a rapid and effective manner. This article studies results of comparison of Correlation Plot effectiveness; a sample group of data was classified using correlation plot versus other well known methods (SVM, K-Mean and Hierarchical). The experimental testing two sets of data: academic articles from the conferences and a sample books from the library. The experimental results on both sets of data showed that accuracy of Correlation Plot are 96.04% and 90.66%, respectively which is higher than the other methods. And response time of this method was found better than other methods as well.

Keywords: Classification, Correlation Plot, Dewey Decimal Classification-Multiple Relations, Correlation Plane, Support Vector Machines

[1] นักศึกษา ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

[2] ผู้ช่วยศาสตราจารย์ ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีและการจัดการอุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

* Corresponding Author, Tel.08-1351-1236, E-mail: watthananon@hotmail.com

รับเมื่อ 30 พฤศจิกายน 2553 ตอบรับเมื่อ 15 สิงหาคม 2554

1. บทนำ

การจำแนกกลุ่มของข้อมูลสามารถทำได้หลากหลายวิธีกับความรู้ในรูปแบบต่างๆ ที่มีอยู่มากมายในปัจจุบัน เพื่อจำแนกสิ่งที่แตกต่างกันออกจากกัน หรือทำการจัดกลุ่มสิ่งที่คล้ายกันรวมเข้าไว้ด้วยกัน ยกตัวอย่างเช่น วิธีการวัดความคล้ายคลึงกันของข้อมูล หรือการเลือกใช้วิธีการวิเคราะห์ทางสถิติคำนวณความน่าจะเป็นเข้ามาช่วยในการศึกษาเปรียบเทียบว่ามีข้อมูลใดบ้างที่คล้ายคลึงกันหรือแตกต่างกัน โดยอาศัยเทคนิคการจำแนกประเภทของข้อมูล (Data Classification) ซึ่งเป็นเทคนิคหนึ่งที่สำคัญสำหรับการสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from Very Large Database: KDD) หรือที่เรียกว่า วิธีการดาต้าไมน์นิง (Data Mining) จุดประสงค์ของการจำแนกประเภทข้อมูล คือ การพัฒนาแบบจำลองการแยกคุณลักษณะหนึ่ง โดยขึ้นอยู่กับคุณลักษณะอื่น [1] ซึ่งแบบจำลองที่ได้จากการจำแนกประเภทข้อมูลจะทำให้สามารถพิจารณาประเภทของข้อมูลที่ยังไม่ได้แบ่งกลุ่มในอนาคตได้ และได้ถูกนำไปประยุกต์ใช้กับงานหลายด้านเพื่อเพิ่มประสิทธิภาพในการทำงาน ซึ่งมีอยู่หลากหลายวิธีการ และเป็นประโยชน์ ยกตัวอย่างเช่น การจัดกลุ่มลูกค้าทางการตลาดด้วยวิธีการ Support Vector Machine (SVM) การวิเคราะห์พฤติกรรมของลูกค้า ต่อการชื่นชอบสินค้าด้วยอัลกอริธึมแบบเพื่อนบ้านใกล้เคียง (K-Nearest Neighbor: KNN) การวิเคราะห์กลุ่มงานภายในห้องสมุด [2] หรือการค้นหารูปภาพโดยใช้วิธีการ Hierarchical Clustering [3] เป็นต้น

จากการศึกษาพบว่า ปัจจุบันนักวิจัยจำนวนมากได้ให้ความสำคัญกับปัญหาที่ต้องการวิธีการจำแนกข้อมูลที่ถูกต้อง แม่นยำ และรวดเร็ว โดยนักวิจัยได้พยายามพัฒนาและนำเสนอวิธีการต่างๆ ให้แบบจำลองให้มีประสิทธิภาพสูงขึ้น อาทิเช่น ในงานวิจัยของ Huanga [4] ได้ทำการเปรียบเทียบผลลัพธ์ระหว่างวิธีการ Support Vector Machine (SVM) ซึ่งให้ผลลัพธ์ที่แม่นยำว่าวิธีการโครงข่ายประสาทเทียม และได้มีการนำมาประยุกต์ใช้ในการจำแนกข้อมูล [5], [6] ซึ่งในงานวิจัยของ Yongchen et al. [7] ได้นำเสนอขั้นตอนวิธีเชิงพันธุกรรม (Genetic Algorithm: GA) เพื่อปรับปรุงประสิทธิภาพการจำแนกกลุ่มให้สูงขึ้น และงานวิจัยของ Watthananon et al. [8] ได้นำเสนอวิธีการ Correlation Plot ซึ่งเป็นเทคนิคที่ใช้หลักการทางคณิตศาสตร์ เข้ามาช่วยในการจำแนกกลุ่มข้อมูลด้วยวิธีการคำนวณหาความสัมพันธ์ระหว่างพิกัดฉาก (Cartesian) และพิกัดเชิงขั้ว (Polar Coordinate) จากข้อมูลที่ปรากฏบน Radar Chart ภายใต้แนวคิดที่ว่าตัวแปรหนึ่งตัว สามารถมีสมาชิกภายในเท่ากับ $1, \ldots, n$ จากนั้นทำการคำนวณตำแหน่งพิกัดของกลุ่มที่เหมาะสม ซึ่งได้มาจากทิศทางและระยะทางของค่าความสัมพันธ์เหล่านั้น โดยจุดแต่ละจุดที่คำนวณได้นี้เรียกว่า จุดสหความสัมพันธ์ (Correlation Plot) ซึ่งปรากฏอยู่ภายในพื้นที่ของแนวเส้นเชื่อมที่ต่อเนื่องกัน เพื่อปิดล้อมพื้นที่ทำให้เกิดเป็นแกน n แกน โดยจะทำการตัดแบ่งระนาบออกเป็น n ส่วน โดยแต่ละส่วนถูกเรียกว่า ระนาบสหความสัมพันธ์ (Correlation Plane) [9]

ดังนั้น บทความนี้ผู้วิจัยได้นำวิธีการ Correlation Plot มาประยุกต์ใช้ในการจำแนกกลุ่มของข้อมูลที่มีความคล้ายคลึงของความรู้จำนวน 2 ชุดข้อมูลที่ใช้ในการทดลอง จากนั้นทำการประเมินประสิทธิภาพของค่าความถูกต้อง หรือ Accuracy [10] และการประเมินประสิทธิภาพของระยะเวลาการตอบสนองของการแสดงผล หรือ Response Times [11], [12] โดยทำการเปรียบเทียบกับวิธีการ Support Vector Machine (SVM), K-means Clustering และ Hierarchical Clustering สำหรับจำแนกกลุ่มของข้อมูลที่นำเสนอ เนื่องจากเนื้อหาของความรู้ที่ได้จากการวิเคราะห์ เป็นค่าความสัมพันธ์แบบสหความสัมพันธ์ (MR: Multiple Relation) ซึ่งในหัวข้อต่อไปของบทความ ผู้วิจัยจะกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง หัวข้อที่ 3 วิธีการจำแนกกลุ่มของข้อมูลด้วย Correlation Plot ตามด้วยหัวข้อที่ 4 การทดลองและการอภิปรายผลกับข้อมูลความรู้จำนวน 2 ชุด และหัวข้อสุดท้ายเป็นบทสรุป

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การจำแนกกลุ่มของข้อมูลด้วย Correlation Plot

เป็นการประยุกต์ใช้ตำแหน่งพิกัดความสัมพันธ์ของความรู้ โดยการเปลี่ยนลักษณะไปจากรูปร่างเดิมของความสัมพันธ์ที่เชื่อมโยงกันหลายตัวแปรสู่การแสดงผลด้วยระบบพิกัดเชิงขั้ว เพื่อแสดงภาพตัวแทนความสัมพันธ์ของความรู้แบบสหความสัมพันธ์ และนำเสนอการจำแนกกลุ่มของข้อมูลที่มีเนื้อหาคล้ายคลึงกัน โดยได้มีการแบ่งทฤษฎีและงานวิจัยที่ทำการศึกษาเป็น 6 หัวข้อ ดังต่อไปนี้

2.1 การจัดรูปแบบตามกรอบความรู้ระบบทศนิยมดิวอี้

การจัดรูปแบบตามกรอบความรู้ระบบทศนิยมดิวอี้ เป็นการจำแนกหมวดหมู่ของความรู้อย่างเป็นระบบ ด้วยการเลือกใช้กรอบความรู้มาตรฐานสากลที่นิยมใช้ในห้องสมุด เพื่อกำหนดขอบเขตความสัมพันธ์ของความรู้ และช่วยในการอธิบายความสัมพันธ์ของความรู้ที่คล้ายคลึงกันในฐานความรู้ขนาดใหญ่ได้อย่างชัดเจนเป็นลำดับ เนื่องจากกรอบความรู้ DDC มีการแบ่งกลุ่มความรู้แบบลำดับชั้นอย่างชัดเจน จากหมวดหลัก 10 หมวด จำแนกสู่หมวดย่อย 100 หมวด และจากหมวดย่อยจำแนกเป็นหมวดหมู่ 1,000 หมวด ด้วยการใช้ตัวเลขสัญลักษณ์ (Pure Notation) [13] สำหรับงานวิจัยนี้ ได้ประยุกต์ใช้ผลจากการวิจัยที่ได้ทำการศึกษาถึงข้อดีและข้อจำกัดเข้ามาช่วยในการนำเสนอข้อมูลของบทความวิชาการทางคอมพิวเตอร์และหนังสือที่จัดเก็บในห้องสมุด ร่วมกับกรอบมาตรฐาน DDC เพื่ออธิบายทิศทางการขยายตัวของกลุ่มข้อมูลในแต่ละหมวดหมู่ โดยเลือกใช้การแสดงผลที่ลำดับชั้นที่ 1 ของการจัดหมู่ระบบทศนิยมดิวอี้จำนวน 10 หมวดหลัก เป็นเกณฑ์ในการแสดงผลตำแหน่งพิกัดของความรู้ เนื่องจากการนำเสนอข้อมูลทุกลำดับชั้นในคราวเดียวไม่เป็นผลดีต่อข้อมูล และผู้ใช้ [14] โดยที่แต่ละตำแหน่งอ้างถึงเนื้อหาของแต่ละบทความที่ใช้ในการทดสอบ ซึ่งได้จากการวิเคราะห์แบบสหความสัมพันธ์

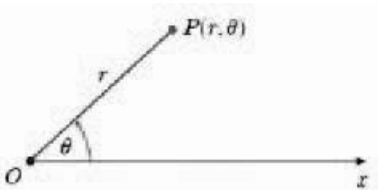
2.2 การวิเคราะห์หมวดหมู่แบบสหความสัมพันธ์ด้วยระบบทศนิยมดิวอี้

กระบวนการ DDC-MR เป็นเทคนิคการวิเคราะห์หมวดหมู่แบบสหความสัมพันธ์ที่พัฒนามาจากกระบวนการค้นคืนสารสนเทศของเครื่องมือสืบค้น (Search Engine) และกรอบความรู้การจัดหมวดหมู่หนังสือระบบทศนิยมดิวอี้ (Dewey Decimals Classification) ซึ่งเป็นการเปลี่ยนวิธีการวิเคราะห์จากหมวดหมู่เดียว สู่การวิเคราะห์หมวดหมู่แบบหลากหลาย ซึ่งจัดเก็บเนื้อหาสารสนเทศในทุกข้อมูลที่เกี่ยวข้อง เพื่อวิเคราะห์ให้เห็นว่าข้อมูลที่จัดเก็บมีเนื้อหาเกี่ยวข้องกับหมวดหมู่ใด เท่าใด และมีรูปแบบความสัมพันธ์กันอย่างไร โดยมุ่งเน้นให้มีการวิเคราะห์สัดส่วนและกำหนดสัญลักษณ์ ให้กับเนื้อหาสำคัญทุกประเด็นที่ปรากฏในวัสดุสารสนเทศ [15], [16] เทคนิคนี้เป็นการวิเคราะห์อย่างละเอียดด้วยการสกัดเนื้อหาสารสนเทศที่มีอยู่ออกมา [8], [17] ทำให้สามารถแสดงผลสารสนเทศที่มองไม่เห็นได้อย่างชัดเจน ส่งผลให้ผู้ใช้เป็นผู้ตัดสินใจในการเลือกใช้สารสนเทศที่ตรงความต้องการได้มากยิ่งขึ้น ซึ่งในงานวิจัยที่ผ่านมา [9], [15] ได้ทำการทดสอบแล้วว่าเทคนิคการวิเคราะห์แบบ DDC-MR สามารถนำมาใช้ในการจัดกลุ่มเนื้อหาที่อยู่ภายในหนังสือได้อย่างแม่นยำและรวดเร็วมีประสิทธิภาพ

2.3 ระบบพิกัดเชิงขั้ว (Polar coordinates)

ระบบพิกัดเชิงขั้ว เป็นระบบพิกัดที่แสดงตำแหน่งของจุดจุดหนึ่ง ซึ่งจะถูกอธิบายด้วยการวัดจากทิศทางและระยะห่างจากลักษณะสำคัญที่ถูกกำหนดไว้ ซึ่งมีการกางออกของมุมหนึ่งมุมหรือมากกว่าที่มีความสัมพันธ์กัน โดยบทความนี้เป็นการประยุกต์ใช้ตัวแปรแบบสหความสัมพันธ์ (Multiple Relations) [8], [15]-[17] นั่นหมายความว่า ตัวแปรหนึ่งตัวมีค่าสมาชิกระยะทางที่มีความสัมพันธ์กันมากกว่า 1 ค่า เช่น $x = x_0, x_1, x_2, \ldots, x_n$ ซึ่งในระบบพิกัดเชิงขั้ว เราเลือกจุด O ในระนาบซึ่งเรียกว่า ขั้ว (Pole) หรือจุดกำเนิด (Origin) และลากรังสี (Ray) จากจุด O เรียกรังสีดังกล่าวว่า แกนเชิงขั้ว (Polar Axis) ซึ่งมักจะลากในแนวราบไปทางขวามือ ดังแสดงในรูปที่ 1

กำหนดให้ P เป็นจุดใดๆ ในระนาบ, r เป็นระยะจากจุด O ถึงจุด P และ θ เป็นมุมระหว่างแกนเชิงขั้ว



รูปที่ 1 ส่วนของเส้นตรง OP

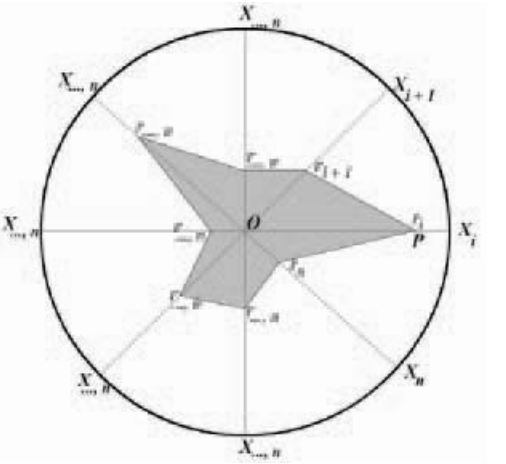
กับส่วนของเส้นตรง *OP* ซึ่งมุม θ มีค่าบวก $(\theta \geq 0)$ เมื่อวัดในทิศทวนเข็มนาฬิกา และมีค่าลบ $(\theta < 0)$ เมื่อวัดในทิศตามเข็มนาฬิกา ดังนั้น จุด P จึงแทนได้ด้วยคู่อันดับ $(r;\theta)$ และเรียก r กับ θ ว่าพิกัดเชิงขั้ว (Polar Coordinates) ของจุด P ในกรณีที่จุด P คือจุด O เราจะพบว่า $r=0$ และตกลงให้ $(0,\theta)$ แทนขั้วสำหรับทุกค่าใดๆ ของมุม θ ดังนั้น หลักสำคัญของการบอกตำแหน่งพิกัดที่เหมาะสม หรือค่าคู่อันดับ $(r; \theta)$ จากความรู้ที่มีหลายตัวแปร จะต้องบอกให้ได้ว่าทิศทางของเวกเตอร์นั้นมีความละเอียดและชัดเจนเพียงทิศเดียว ด้วยวิธีการคำนวณหาขนาดและทิศทาง

2.4 การคำนวณหาขนาดและทิศทาง

จากรูปที่ 2 แสดงตัวอย่างข้อมูลของตัวแปรหนึ่งตัวมีค่าสมาชิกระยะทาง ที่มีความสัมพันธ์กันมากกว่า 1 ค่า โดยกำหนดให้ r_{ij} แทนระยะทางของจุด P จากจุดกำเนิด โดยที่มุม θ คือองศาของแนวเส้นรัศมีจากจุด P ถึง O และมีค่า r เป็นระยะทางของแต่ละค่าสมาชิกจากจุด P ถึงจุดกำเนิด ดังนั้น ในการคำนวณหาขนาดและทิศทางความสัมพันธ์บนระบบพิกัดเชิงขั้ว เลือกใช้กฎของ *Cosine & Sine* เพื่อเชื่อมต่อจุด n จุด ที่ตำแหน่ง P_ij สำหรับ $i = 1,\ldots, n$ สามารถคำนวณดังแสดงในสมการที่ (1) ดังนี้

$$P_{ij} = \begin{cases} x_{ij} = r\cos\theta_i \\ y_{ij} = r\sin\theta_i \end{cases} \qquad (1)$$

โดยที่ $r \geq 0, 0 \leq \theta < 2\pi$, ซึ่งทุกจุดสามารถเขียนแทนด้วย $P(x_{ij}, y_{ij})$ บนแกนระนาบ xy แทนได้ด้วยคู่อันดับ (r,θ) ซึ่งจากตัวแปรแบบสหความสัมพันธ์ดังกล่าว



รูปที่ 2 ตัวอย่างข้อมูลของตัวแปรหนึ่งตัวที่มีค่าความสัมพันธ์กันมากกว่า 1 ค่า โดยที่ n คือจำนวนของสมาชิก, r_i คือความสัมพันธ์ระหว่างพิกัดฉากและพิกัดเชิงขั้ว (r, θ)

ถูกเรียกว่าจุดสหสัมพันธ์บนระบบพิกัดเชิงขั้ว ซึ่งจุด P สามารถอ้างถึงความสัมพันธ์ $x_{ij}^2 + y_{ij}^2 = r_{ij}^2 \; (\cos^2\theta_i + \sin^2\theta_i) \rightarrow x_{ij}^2 + y_{ij}^2 = r_{ij}^2$ (เพื่อใช้ในการอ้างถึง $P(x_{ij},y_{ij})$ และ $(\cos^2\theta_i + \sin^2\theta_i) = 1$) ที่ต่อเนื่องกันซึ่งอยู่บนวงกลมรัศมี r จากจุดศูนย์กลางที่ตำแหน่ง O ดังนั้น สามารถคำนวณมุม θ ได้ดังแสดงในสมการที่ (2) ดังนี้

$$\tan\theta_i = \frac{y_{ij}}{x_{ij}} \rightarrow \theta_i = \arctan\left(\frac{y_{ij}}{x_{ij}}\right), \qquad (2)$$

โดยที่ θ อยู่ในช่วง $0 \leq \theta < 2\pi$ กำหนดให้ การหาค่า Arctangent ของ x และ y แทนด้วยการใช้สมการที่ (3) ดังนี้

$$\theta_i = \arctan\left(\frac{y_{ij}}{x_{ij}}\right) = \begin{cases} Arctan\left(\frac{y}{x}\right) & if\ -\frac{\pi}{2} < \theta < \frac{\pi}{2}, \\ Arctan\left(\frac{y}{x}\right) + \pi & if\ \frac{\pi}{2} < \theta \leq \frac{3\pi}{2}, \end{cases} \qquad (3)$$

2.5 การคำนวณหาพื้นที่ในระบบพิกัดเชิงขั้ว

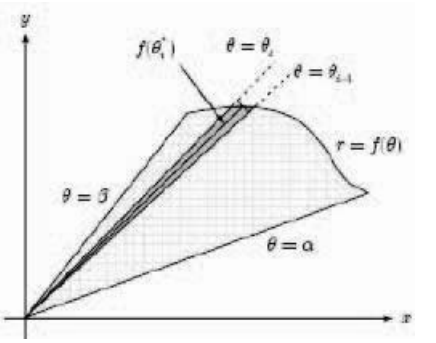
จากข้อมูลที่ปรากฏบน Radar Chart (รูปที่ 2) เพื่อใช้คำนวณจุดสหความสัมพันธ์ ที่ได้จากการคำนวณ

ทิศทางและระยะทางของสมาชิกในตัวแปร การคำนวณหาพื้นที่ในระบบพิกัดเชิงขั้ว เป็นการกำหนดบริเวณของขอบเขตในระบบพิกัดเชิงขั้ว บนระนาบสหความสัมพันธ์ของแนวเส้นเชื่อมต่อเนื่องกัน ซึ่งปิดล้อมด้วยฟังก์ชันเส้นโค้งที่ต่อเนื่องกัน $r=f(\theta)$ และโดยรังสีเส้นตรง $\theta=\alpha$ และ $\theta=\beta$ เมื่อ $r\geq 0$ ดังแสดงในรูปที่ 3 (ก) ซึ่งจุดที่ตัดกันของ n แกน เรียกว่า จุดเริ่มต้นหรือจุดกำเนิด (Origin) และแกน n แกน ตัดกันจะแบ่งระนาบสหความสัมพันธ์ออกเป็น n ส่วน โดยที่ f เป็นฟังก์ชันค่าบวกที่ต่อเนื่องในช่วง $\alpha \leq\theta\leq\beta$ โดยเราจะแบ่งช่วง $[\alpha,\beta]$ ออกเป็นช่วงย่อยๆ ด้วยจุด $\alpha= \theta_0,\theta_1,\theta_2,\ldots,\theta_{n-1},\theta_n=\beta$ โดยแต่ละช่วงมีความกว้าง $\Delta\theta$ เท่ากัน ทำให้แบ่งบริเวณดังกล่าวออกเป็น n บริเวณย่อยๆ ดังแสดงในรูปที่ 3 (ข) เกี่ยวกับการแสดงส่วนของพื้นที่ n ส่วน ของวงกลมรัศมี r ที่รองรับมุม θ

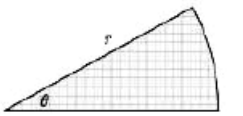
เมื่อพิจารณาสัดส่วนของวงกลมรัศมี r ที่รองรับมุม θ จะพบว่าพื้นที่ของสัดส่วนดังกล่าวคือ $=\frac{1}{2}r^2\theta$ ดังนั้น เมื่อนำหลักการคำนวณหาพื้นที่ไปประยุกต์ใช้กับพื้นที่ของแต่ละบริเวณย่อยๆ ที่แบ่งไว้ จะพบว่าในช่วง $[\theta_i=1,\theta_i]$ จะมี $\theta_i^*\epsilon(\theta_{i-1},\theta_i)$ ซึ่งทำให้ $\Delta A_i=\frac{1}{2}\left(f(\theta_i)\right)^2\Delta$ ดังนั้น เมื่อรวมพื้นที่ n ส่วน เข้าด้วยกันเป็นระนาบที่สัมพันธ์กันจะได้ฟังก์ชันที่ต่อเนื่อง $A=\sum_{i=1}^{n}\frac{1}{2}\left(f(\theta_i^*)\right)^2\Delta\theta$ ในลิมิตที่ $n\rightarrow\infty$ ดังแสดงในสมการ (4)

$$A=\int_{\alpha}^{\beta}\frac{1}{2}\left(f(\theta_i^*)\right)^2 d\theta=\int_{\alpha}^{\beta}\frac{1}{2}r^2 d\theta \qquad (4)$$

ดังนั้น การคำนวณหาพื้นที่ในระบบพิกัดเชิงขั้นเป็นการประยุกต์ใช้ตำแหน่งพิกัดที่เหมาะสม สำหรับความสัมพันธ์ของความรู้ มีวัตถุประสงค์เพื่อแสดงภาพตัวแทนความสัมพันธ์ของความรู้โดยแปลงความสัมพันธ์แบบสหความสัมพันธ์ให้เป็นภาพพิกัดกราฟที่ง่ายต่อความเข้าใจ [8] และเพื่อให้ทราบถึงทิศทางความรู้ที่มีอยู่ในระบบ [9], [17] เนื่องจากความรู้ที่คำนวณได้ปรากฏค่าสัดส่วนความสัมพันธ์ที่หลากหลายและเกี่ยวข้องกัน ทำให้ไม่สามารถแสดงผลความสัมพันธ์ทั้งหมดออก



(ก)



(ข)

รูปที่ 3 การหาพื้นที่ในระบบพิกัดเชิงขั้ว

มาได้ ดังนั้น จึงต้องทำการคำนวณตัวแทนที่เหมาะสมเพื่อกำหนดความสัมพันธ์เหล่านั้นอีกครั้ง โดยการใช้ผลรวมจากทุกโหนดความสัมพันธ์ของทิศทางและระยะทาง กำหนดเป็นตำแหน่งพิกัดที่เหมาะสมของความรู้ ซึ่งตำแหน่งพิกัดที่คำนวณได้นั้นเป็นตัวแทนสหความสัมพันธ์ของทุกค่าความสัมพันธ์ที่ซ่อนอยู่ภายในความรู้ไม่ได้ระบุจากค่าใดค่าหนึ่งที่สูงสุดเท่านั้น นอกจากนี้พื้นที่และขอบเขตในการแสดงความสัมพันธ์ของความรู้ไม่สามารถใช้พื้นที่สำหรับความรู้ใดองค์ความหนึ่งเป็นเกณฑ์ในการแสดงผลได้ เนื่องจากพื้นที่และขอบเขตที่เกิดขึ้นเป็นการซ้อนทับกันของความรู้ จึงต้องคำนวณขอบเขตและพื้นที่สหความสัมพันธ์ เพื่อใช้แสดงตำแหน่งสหความสัมพันธ์ และทิศทางของความรู้

2.6 การจำแนกข้อมูลด้วย Data Mining

การจำแนกข้อมูลด้วย Data Mining [1] เป็นกระบวนการหนึ่งที่สำคัญ เพื่อกลั่นกรองและจำแนกกลุ่มของข้อมูลบนฐานความรู้ขนาดใหญ่ (KDD) ซึ่งวัตถุประสงค์ของการจำแนกข้อมูล คือการสร้างโมเดลการแยกคุณลักษณะหนึ่งของข้อมูลโดยขึ้นกับคุณลักษณะอื่น

ซึ่งทำให้สามารถพิจารณากลุ่มของข้อมูลที่ยังไม่ได้แบ่งกลุ่มในอนาคตได้ บทความนี้ผู้วิจัยได้นำวิธีการ Correlation Plot มาทำการเปรียบเทียบกับการจำแนกข้อมูลกับอีก 3 วิธีการ ซึ่งมีความแตกต่างกันระหว่างการจัดหมู่แบบ Classification ด้วยเทคนิค Support Vector Machine เนื่องจากวิธีการที่พัฒนาขึ้นเมื่อทำการวิเคราะห์ด้วย DDC-MR ข้อมูลที่ได้จะมีการให้คุณลักษณะตามหมวดหมู่ เพื่อนำไปใช้ในการตัดสินใจ และทำการเปรียบเทียบการจัดหมู่แบบ Clustering ด้วยเทคนิค K-means Clustering และเทคนิคการจัดกลุ่มแบบ Agglomerative Hierarchical Cluster เนื่องจากวิธีการที่พัฒนาเป็นการใช้กรอบการจัดหมู่ระบบทศนิยมดิวอี้ ซึ่งมีพื้นฐานโครงสร้างเป็นลำดับชั้นโดยดูจากลักษณะพื้นฐานที่คล้ายคึงกัน ดังนั้น บทความนี้ผู้วิจัยจึงเลือกใช้วิธีการดังกล่ามาเปรียบเทียบการทดลองในครั้งนี้

2.6.1 Support Vector Machines (SVM) เป็นวิธีการที่สามารถนำมาใช้ในการจำแนกรูปแบบหรือกลุ่มของข้อมูลได้ โดยใช้การจำแนกกลุ่มของข้อมูลด้วยหลักการทางสถิติ [18]-[20] โดยหลักการสำคัญจะอาศัยระนาบหรือหาเส้นแบ่งที่เหมาะสมที่สุดมาใช้ในการแบ่งเขตของข้อมูลสองชุดออกจากกัน (Hyper Plane) และ Support Vector Machines นี้ สามารถแบ่งออกได้เป็น 2 ประเภท คือ แบบเชิงเส้น (Linear) และแบบไม่เป็นเชิงเส้น (Non-linear) ข้อจำกัดของวิธีการนี้คือ การกำหนดค่าพารามิเตอร์ หากกำหนดค่าไม่เหมาะสมจะมีผลทำให้โมเดลที่ได้มีประสิทธิภาพไม่ดีนัก อีกทั้งวิธีซัพพอร์ตเวกเตอร์แมชชีนไม่คงทนต่อข้อมูลรบกวน (Noise) และข้อมูลที่มีลักษณะผิดแยกออกมาจากกลุ่ม (Outlier) [6]

2.6.2 K-means Clustering เป็นวิธีการที่นิยมใช้กันมากในหลายงานวิจัย [21]-[23] เนื่องจากมีขั้นตอนวิธีในการปฏิบัติที่ง่ายและได้ผลดี โดยการใช้วิธีการวัดระยะห่างของข้อมูลเป็นข้อกำหนด แต่วิธีการนี้มีข้อจำกัดในเรื่องของเวลากับจำนวนของ Vectors และจำนวนกลุ่มต่อการวนลูป ซึ่งต้องใช้ทรัพยากรสิ้นเปลืองมากเมื่อต้องทำงานกับชุดข้อมูลขนาดใหญ่

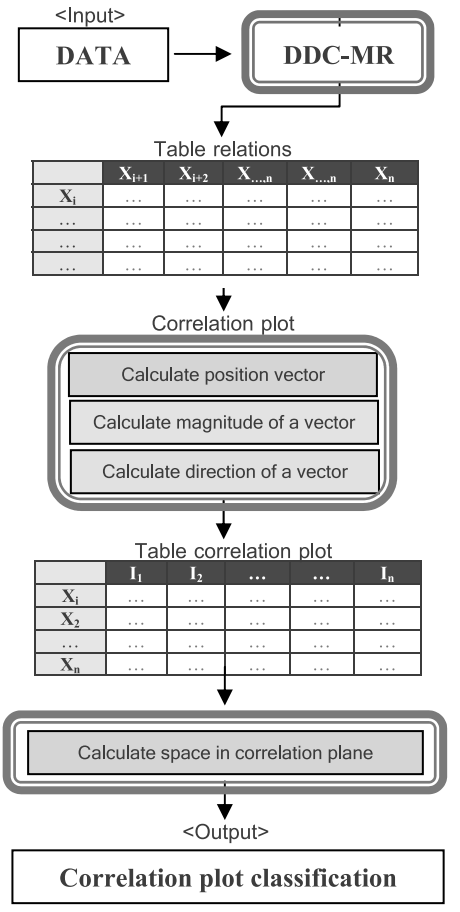
2.6.3 Agglomerative Hierarchical Clustering หรือวิธีการการแบ่งกลุ่มแบบลำดับชั้น เป็นวิธีการที่นิยมกันมาก ในการแบ่งกลุ่มตัวแปร [24] โดยจะทำการจำแนกข้อมูลที่คล้ายกันให้รวมเป็นกลุ่ม แล้วดูว่ากลุ่มใดคล้ายกันอีกก็ทำการรวมเป็นกลุ่มที่ใหญ่ขึ้น ซึ่งจะกระทำจนกว่าจะเหลือเพียง 1 กลุ่มใหญ่เท่านั้น เช่น การจัดกลุ่มลูกค้าทางการตลาด การวิเคราะห์กลุ่มงานภายในห้องสมุด [2] หรือการค้นหารูปภาพ [3] เป็นต้น ด้วยการจัดกลุ่มที่คล้ายกันให้อยู่ในกลุ่มเดียวกัน เพื่อช่วยให้การบริการมีประสิทธิภาพ และเป็นประโยชน์ต่อลูกค้ามากยิ่งขึ้น

3. วิธีการจำแนกกลุ่มของข้อมูลด้วย Correlation Plot

การเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มของข้อมูลด้วยวิธีการ Correlation Plot เป็นวิธีการการเปลี่ยนลักษณะ (Deformation) ไปจากรูปร่างเดิมของความสัมพันธ์ที่เชื่อมโยงกันด้วยการแสดงผลแบบ Radar Graph (รูปที่ 2) ที่มีความสัมพันธ์กันของสมาชิกภายในตัวแปร ซึ่งสิ่งสำคัญที่นำมาพิจารณา คือรูปร่าง ปริมาณความสัมพันธ์ของเนื้อหา ระยะ ตำแหน่งสัมพันธ์ และทิศทางของพิกัดที่ถูกกำหนด [8] เมื่อนำมาคำนวณเพื่อหาตำแหน่ง Plot ในการเป็นตัวแทนที่ดีของความรู้แสดงว่าจะต้องถูกทำให้เปลี่ยนลักษณะการ Plot ด้วยความสัมพันธ์ที่ซ่อนอยู่ภายใน บนระนาบและขอบเขตที่สอดคล้องกับความสัมพันธ์นั้น ดังนั้น เราจึงไม่สามารถใช้ระนาบปกติอธิบายความสัมพันธ์ของตำแหน่งที่คำนวณนั้นได้ เนื่องจากจุดสหความสัมพันธ์ที่เกิดขึ้นมาจากตัวแปรที่มีสมาชิกภายในหลากหลายความสัมพันธ์ ซึ่งโครงสร้างของวิธีการ Correlation Plot มีการทำงานดังรูปที่ 4 ดังนี้

ขั้นตอนที่ 1 การทำงานเริ่มจากการนำข้อมูลเข้ามาซึ่งจะเป็นข้อมูลจำนวน 4 ส่วน คือ ชื่อเรื่อง บทคัดย่อ คำสำคัญ และวัน/เดือน/ปี โดยต้องทำการแทนบทความให้อยู่ในรูปแบบของ Cartesian Product ดังนี้

$$Doc = Title_{w1} \times Abstract_{w2} \times Keywords_{w3} \times Date \quad (5)$$



รูปที่ 4 โครงสร้างวิธีการ Correlation Plot

โดยที่ w_i คือ จำนวนสัดส่วนที่เหมาะสมของการคำนวณน้ำหนักในแต่ละส่วน

ขั้นตอนที่ 2 นำข้อมูลที่ได้ในขั้นตอนที่ 1 ไปวิเคราะห์เนื้อหาการจัดหมู่แบบสหความสัมพันธ์ด้วย DDC-MR และนำค่าที่ได้จากการวิเคราะห์จัดเก็บไว้ในฐานข้อมูล ผลลัพธ์ที่ได้เป็นตารางค่าความสัมพันธ์ในลำดับชั้นที่ 3 จำนวน 1,000 หมวดหมู่

ขั้นตอนที่ 3 การคำนวณจุดสหความสัมพันธ์ ซึ่งประกอบด้วยวิธีการคำนวณตำแหน่ง วิธีการคำนวณระยะทาง และการคำนวณทิศทาง ด้วยการประยุกต์ใช้จากทฤษฎีในส่วนที่ 2 ซึ่งเป็นการคำนวณหาตำแหน่งพิกัดของบทความจากผลรวมความสัมพันธ์ที่เชื่อมระหว่างตัวแปร ผลลัพธ์ที่ได้เป็นตารางค่าความสัมพันธ์ของพิกัดที่คำนวณได้ในรูปแบบของค่าคู่อันดับ

จากรูปที่ 4 โครงสร้างวิธีการ Correlation Plot สามารถอธิบายได้ดังนี้

ขั้นตอนที่ 4 การคำนวณขอบเขตและพื้นที่ความสัมพันธ์ เพื่อใช้กำหนดขอบเขตและระนาบสมมติที่เรียกว่า ขอบเขตสหความสัมพันธ์ และระนาบสหความสัมพันธ์ ด้วยการคำนวณให้เกิดระนาบของแนวเส้นที่เชื่อมต่อเนื่องกัน และซ้อนทับกันหลายบทความ ซึ่งปิดล้อมด้วยแกนเชิงขั้วที่มีการจัดแบ่งหมวดหมู่ตามองศาของแกนด้วยการคำนวณองศาของมุมจำนวน 10 ส่วน สำหรับการกำหนดตำแหน่งของจุดสหความสัมพันธ์ในแต่ละหมวดหมู่

4. การทดลองและการอภิปรายผล

การทดลองประสิทธิภาพ การจำแนกกลุ่มของข้อมูลด้วยวิธีการ Correlation Plot บทความนี้ผู้วิจัยมุ่งเน้นให้ความสนใจใน 2 ประเด็นหลัก คือ ความถูกต้องในการจำแนกกลุ่ม และระยะเวลาที่ใช้ในการประมวลผลข้อมูล โดยทำการทดลองกับข้อมูลความรู้จำนวน 2 ชุด เกี่ยวกับบทความวิชาการทางคอมพิวเตอร์ และหนังสือที่จัดเก็บในห้องสมุด ซึ่งมีความแตกต่างกันหลากหลายศาสตร์

4.1 ข้อมูลที่ใช้ในการทดลอง

ข้อมูลที่ใช้ในการทดลองสำหรับบทความนี้ เป็นข้อมูลความรู้ที่จัดเก็บความสัมพันธ์ที่สอดคล้องกัน ด้วยศาสตร์หลากหลายศาสตร์ไว้ภายใน โดยความรู้ชุดแรกที่ใช้ในการทดลอง คือ บทความวิชาการทางคอมพิวเตอร์ จำนวน 700 บทความ ซึ่งได้รับการตีพิมพ์เผยแพร่ระหว่างปี 2005 – 2010 โดยเลือกใช้ข้อมูล 4 ส่วน เป็นตัวแทนของความรู้ในการวิเคราะห์ความสัมพันธ์เพื่อจำแนกกลุ่มของข้อมูล ได้แก่ ชื่อเรื่อง (Title) บทคัดย่อ (Abstract) คำสำคัญ (Keyword) และปี (Year) ของบทความที่เผยแพร่ และความรู้ชุดที่สองที่ใช้ในการทดลอง คือ หนังสือที่จัดเก็บในห้องสมุด จำนวน 150 เล่ม โดยตัวแทนของความรู้ เพื่อใช้ในการวิเคราะห์ข้อมูลชุดนี้

เลือกใช้ 3 ส่วน คือ ชื่อหนังสือ (Title) สารบัญ (Table of Content) และดัชนี (Index) เนื่องจากในงานวิจัยของ Fuller [25] ได้ให้เหตุผลเกี่ยวกับการเลือกตัวแทนของข้อมูลบางส่วน โดยไม่สนใจเนื้อหาของเอกสารทั้งหมด ก็สามารถอ้างถึงเอกสารที่ต้องการได้เช่นเดียวกัน ในขณะที่งานวิจัยของ Peery และคณะ [26] ได้ทำการศึกษาและทดลองความแม่นยำในการค้นคืนเอกสาร 3 ส่วน คือ Content, Metadata และ Structure ผลการทดลองพบว่าสามารถค้นค้นแม่นยำเพิ่มขึ้น 50% ดังนั้น ในงานวิจัยนี้ผู้วิจัยจึงเลือกใช้ตัวแทนในแต่ละส่วนของชุดข้อมูล เป็นตัวแทนข้อมูลเพื่อทำการทดลอง

4.2 การคำนวณประสิทธิภาพ

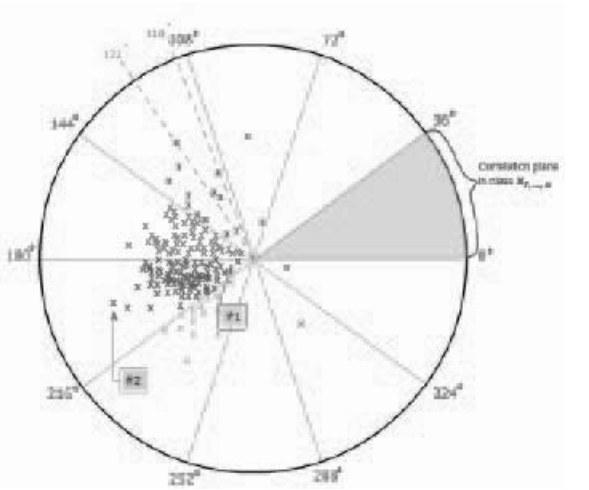
ในการทดลองครั้งนี้ เลือกใช้วิธีการวัดค่าความถูกต้อง (Accuracy) [10] เพื่อวัดประสิทธิภาพมาตรฐานสำหรับการประเมินผลการจำแนกกลุ่มข้อมูลที่ใช้ในการทดลองเปรียบเทียบ โดยค่าข้อมูลของตัวชี้วัดเหล่านี้ขึ้นอยู่กับจำนวนของระเบียนข้อมูลที่สามารถทำนายถูกต้อง และทำนายไม่ถูกต้องด้วยตาราง Confusion Matrix (ดังแสดงในตารางที่ 1) ซึ่งวิธีการนี้จะทำการวัดได้ทั้งค่าความถูกต้อง และค่าความผิดพลาดของการจำแนกกลุ่ม

ตารางที่ 1 ตาราง Confusion Matrix

		ค่าพยากรณ์			
		1	2	…	M
ค่าจริง	1	$C_{1,1}$	$C_{1,2}$	…	$C_{1,M}$
	2	$C_{2,1}$	$C_{2,2}$	…	$C_{2,M}$
	…	…	…	…	…
	M	$C_{M,1}$	$C_{M,2}$	$C_{M,3}$	$C_{M,M}$

กำหนดให้ $C_{i,j}$ แทนค่าที่ได้จากการทำนาย โดยที่ i=1,...,M และ j=1,...,M ถ้า i=j แสดงว่าทำนายถูก (CorrectPrediction) และถ้าแสดงว่าทำนายไม่ถูก (Incorrect Prediction) โดยสามารถคำนวณได้จากสมการที่ (6)

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (6)$$



รูปที่ 5 ตัวอย่างจุดสหความสัมพันธ์จำแนกตามกลุ่มบนระนาบสหความสัมพันธ์

4.3 การอภิปรายผล

จากการทดลองจำแนกกลุ่มข้อมูลด้วยวิธี Correlation Plot ตำแหน่งของจุดสหความสัมพันธ์ที่เกิดขึ้นหรือเรียกว่า Correlation Plot สามารถนำมาใช้อ้างถึงตัวแปรที่มีจำนวน n ความสัมพันธ์ ซึ่งแต่ละความสัมพันธ์มีปริมาณ และทิศทางที่แตกต่างกัน ส่งผลให้ตำแหน่งพิกัดที่คำนวณได้อยู่บนระนาบสหความสัมพันธ์ มีระยะห่างที่แตกต่างกันจากจุดศูนย์กลาง ดังนั้น ถ้าเราแทนแต่ละจุดเป็นองค์ความรู้หรือหนังสือ Correlation Plot ที่คำนวณได้จะแทนเนื้อหาของหนังสือที่มีองค์ความรู้สัมพันธ์กัน Correlation Plane จะแทนด้วยพื้นที่ความสัมพันธ์ของโครงสร้างเนื้อหาในองค์ความรู้ ซึ่งจุดที่รวมตัวหนาแน่นเป็นกลุ่มเส้น จะปรากฏให้เห็นเป็นทิศทางที่มีความสัมพันธ์กันภายในระนาบสหความสัมพันธ์ โดยที่จุดที่มีระยะห่างจากจุดศูนย์กลางมาก (#2 ในรูปที่ 5) หมายถึง หมวดหนังสือที่มีเนื้อหาเฉพาะเจาะจงลงไปในระดับลึกขององค์ความรู้นั้นมาก เนื่องจากน้ำหนักของแรง และทิศทางของตัวแปรที่สัมพันธ์กันมุ่งไปสู่ทิศทางเดียวกันสูง ในขณะที่จุดที่มีระยะเข้าใกล้จุดศูนย์กลางมาก (#1 ในรูปที่ 5) หมายถึง หมวดหนังสือเฉพาะทางในองค์ความรู้นั้นเช่นเดียวกัน แต่มีเนื้อหาสอดคล้องกับหลากหลายองค์ความรู้ ดังแสดงในรูปที่ 5

จากรูปที่ 5 แสดงตัวอย่างของจุดสหความสัมพันธ์ที่ปรากฏอยู่บนระนาบสหความสัมพันธ์ ซึ่งได้ทดลองกับข้อมูลชุดที่ 1 โดยที่ Correlation Plane ใน Class $X_1,\ldots,$ n หมายถึง ช่วงขอบเขตระนาบสหความสัมพันธ์ในแต่ละความรู้ ที่ได้จากการคำนวณ จากตัวอย่างข้างต้น *n* หมายถึง จำนวนความรู้ 10 หมวดหลัก ซึ่งจำแนกกลุ่มตามการวิเคราะห์ความสัมพันธ์ด้วยเทคนิค DDC-MR โดยความรู้แรกหมายถึง Class General มีระนาบอยู่ในช่วง 0^{O}–35^{O} ความรู้ที่สองหมายถึง Class Philosophy มีอยู่ในช่วง 36^{O}–71^{O} ความรู้ถัดไปได้แก่ Class Religion มีระนาบอยู่ในช่วง 72^{O}–107^{O} Class Social Sciences มีระนาบอยู่ในช่วง 108^{O}–143^{O} Class Languages มีระนาบอยู่ในช่วง 144^{O}–179^{O} Class Pure Science and Mathematics มีระนาบอยู่ในช่วง 180^{O}–215^{O} Class Technology and Applied Science มีระนาบอยู่ในช่วง 216^{O}–251^{O} Class Arts and Recreation มีระนาบอยู่ในช่วง 252^{O}–287^{O} Class Literature มีระนาบอยู่ในช่วง 288^{O}–323^{O} และ Class History and Geography มีระนาบอยู่ในช่วง 324^{O}–360^{O} ตามลำดับ ด้วยวิธีการนี้สามารถนำมาใช้ทดสอบกับข้อมูลที่มีความสัมพันธ์กันได้ ซึ่งประเด็นสำคัญที่นำมาพิจารณาถึงประสิทธิภาพ มีดังนี้

• **Accuracy:** ผลการทดสอบความถูกต้อง หลังจากแทนข้อมูลด้วยวิธีจุดสหความสัมพันธ์เรียบร้อยแล้ว ผลจะถูกส่งไปทดสอบการจำแนกโดยใช้ WEKA และ SPSS เพื่อวิเคราะห์หาประสิทธิภาพในการจำแนก จากนั้นทำการเปรียบเทียบประสิทธิภาพการแทนข้อมูลด้วยจุดสหความสัมพันธ์กับการวิธีการ Support Vector Machine (SVM) โดยกำหนด Random Seed = 1,000 และเลือกฟังก์ชันการทำงานเป็น Normalized Poly Kernel วิธีการ K-Mean Clustering กำหนดจำนวนกลุ่ม = 10 Clusters เลือกวิธีการคำนวณเป็น Euclidean Distance และวิธีการ Hierarchical Clustering กำหนด Random = 999 รอบ โดยผลของการเปรียบเทียบประสิทธิภาพการจำแนก ดังแสดงในตารางที่ 2

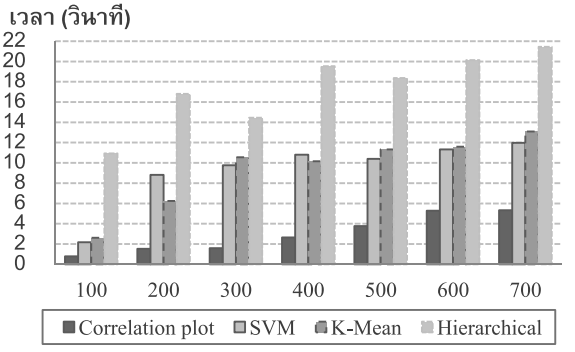
จากตารางที่ 2 แสดงผลการเปรียบเทียบประสิทธิภาพการจำแนกด้วยค่าความถูกต้องของทั้ง 4 วิธี คือวิธีการจำแนกด้วยจุดสหความสัมพันธ์ที่นำเสนอ วิธีการ Support Vector Machine (SVM) วิธีการ K-Mean Clustering และวิธีการ Hierarchical Clustering จากตารางพบว่า ผลการทดสอบที่ได้ทำการเปรียบเทียบค่าความถูกต้องในการจำแนกกลุ่มข้อมูลของวิธีกาจุดสหความสัมพันธ์ที่นำเสนอ ให้ค่าความถูกต้องสูงกว่าวิธีการอื่น ซึ่งมีค่าเท่ากับ 96.04%และ 90.60% สำหรับบทความวิชาการทางคอมพิวเตอร์จำนวน 700 บทความ และหนังสือที่จัดเก็บในห้องสมุดจำนวน 150 เล่ม ตามลำดับ

ตารางที่ 2 ผลการเปรียบเทียบประสิทธิภาพการจำแนกด้วยค่าความถูกต้อง (Accuracy)

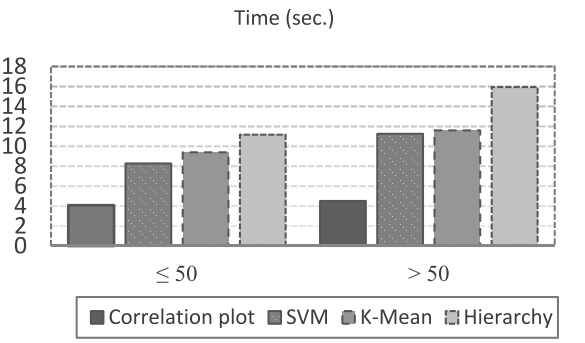
Compare	Accuracy (%)	
	Articles (700)	Books (150)
Correlation plot	96.04	90.60
SVM	90.40	89.93
K-Mean	90.16	88.67
Hierarchical	87.34	85.21

• **Response Times:** ผลการทดสอบประสิทธิภาพเกี่ยวกับเวลาตอบสนองของการแสดงผล ด้วยการเปรียบเทียบการทำงานของวิธีการทั้ง 4 กับเกณฑ์ประสิทธิภาพของ Miller [11] โดยได้ทำการแบ่งจำนวนของข้อมูลแต่ละชุดเพื่อทดสอบ ซึ่งข้อมูลชุดแรก คือ บทความวิชาการได้แบ่งจำนวนบทความที่ใช้ทดสอบไว้ที่ 100, 200, 300, 400, 500, 600 และ 700 บทความ ข้อมูลชุดที่สอง คือหนังสือได้ถูกแบ่งจำนวนในการทดสอบไว้ 2 ช่วง คือ 1 – 50 เล่ม และ มากกว่า 50 เล่ม ซึ่งผลของการทดสอบสามารถแสดงได้ดังตารางที่ 3 และ 4 ตามลำดับ

• จากตารางที่ 3 แสดงผลของการทดสอบประสิทธิภาพเกี่ยวกับเวลาตอบสนองของการแสดงผลสำหรับข้อมูลชุดที่ 1 คือ บทความวิชาการทางคอมพิวเตอร์ จากตารางพบว่า วิธีการที่นำเสนอให้เวลาตอบสนองของการแสดงผลต่ำกว่าวิธีการอื่นที่นำมาใช้ทดสอบเพื่อเปรียบเทียบ นอกจากนี้เวลาตอบสนองที่ได้จากวิธีการจุดสหความสัมพันธ์ยังให้ประสิทธิภาพตามเกณฑ์ [11] ประสิทธิภาพระดับดีของการแสดงผล ซึ่งอยู่ในช่วงการแสดงผล 1 – 6 วินาที ดังแสดงในรูปที่ 6



รูปที่ 6 ผลการเปรียบเทียบความแตกต่างของเวลาตอบสนองสำหรับข้อมูลชุดที่ 1



รูปที่ 7 ผลการเปรียบเทียบความแตกต่างของเวลาตอบสนองสำหรับข้อมูลชุดที่ 2

ตารางที่ 3 ผลการทดสอบประสิทธิภาพของเวลาตอบสนองการแสดงผลข้อมูลชุดที่ 1 จำนวน 700 บทความ

No. articles	Response time (second)			
	Correlation plot	SVM	K-Mean	Hierarchical
100	00.78	2.17	2.62	10.94
200	01.52	8.82	6.25	16.80
300	01.59	9.76	10.55	14.45
400	02.64	10.80	10.16	19.53
500	03.77	10.40	11.33	18.36
600	05.29	11.33	11.59	20.13
700	05.33	11.97	13.09	21.45

ตารางที่ 4 ผลการทดสอบประสิทธิภาพของเวลาตอบสนองการแสดงผลสำหรับข้อมูลชุดที่ 2 จำนวน 150 เล่ม

No. Books	Response time (second)			
	Correlation plot	SVM	K-Mean	Hierarchical
≤ 50	04.10	08.29	09.40	11.16
> 50	04.52	11.26	11.59	15.94

จากตารางที่ 4 แสดงผลของการทดสอบประสิทธิภาพเกี่ยวกับเวลาตอบสนองของการแสดงผล สำหรับข้อมูลชุดที่ 2 คือ หนังสือที่จัดเก็บในห้องสมุด จากตารางพบว่าวิธีการจุดสหความสัมพันธ์ที่นำเสนอให้เวลาตอบสนองของการแสดงผลต่ำกว่าวิธีการอื่นที่นำมาใช้ทดสอบเพื่อเปรียบเทียบ แม้ว่าขนาดของข้อมูลมีปริมาณมากกว่าบทความวิชาการ นอกจากนี้สามารถอธิบายได้ว่า การเพิ่มหรือลดจำนวนข้อมูลที่ใช้ทดสอบไม่ส่งผลกระทบต่อเวลาตอบสนองของการแสดงผลสำหรับวิธีจุดสหความสัมพันธ์ เนื่องจากวิธีการนี้ไม่จำเป็นต้องคำนวณตำแหน่งของข้อมูลใหม่ทุกครั้งที่มีการเพิ่มหรือเปลี่ยนแปลง

ดังนั้น เมื่อมีข้อมูลเพิ่มหรือลดจำนวนข้อมูลเดิมจะยังคงอยู่ที่ตำแหน่งและกลุ่มเดิม โดยที่ไม่ทำให้ข้อมูลเกิดความสูญเสียนัยสำคัญ จึงแตกต่างจากวิธีการอื่นที่ใช้เปรียบเทียบซึ่งต้องทำการจำแนกกลุ่มใหม่ทุกครั้งที่มีการเปลี่ยนแปลง ซึ่งไม่เป็นผลดีต่อการใช้งานต่อปริมาณที่มากขึ้นอย่างชัดเจนเนื่องจากตัวแปรทุกตัวมีผลต่อการจัดกลุ่มข้อมูล ดังแสดงในรูปที่ 7

5. สรุป

บทความนี้ เป็นการศึกษาเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มของข้อมูลด้วยวิธีการ Correlation Plot อธิบายได้ว่า มีประสิทธิภาพด้วยค่าความถูกต้องที่สูงกว่าวิธีการอื่น และให้ประสิทธิภาพของระยะเวลาตอบสนองในการแสดงผลเร็วกว่าวิธีการอื่น และสามารถระบุได้ชัดเจนว่า บทความวิชาการทางคอมพิวเตอร์ที่ใช้ในการทดสอบมีความสัมพันธ์ค่อนข้างใกล้ชิดกัน หมายความว่าบทความที่ใช้ในการทดสอบเป็นบทความที่มีเนื้อหา

คล้ายคลึงกัน ดังนั้น ทิศทางการกระจายตัวและความหนาแน่นของบทความ โดยส่วนใหญ่จึงเกาะกลุ่มกัน แต่วิธีการนี้มีข้อจำกัดเกี่ยวกับการระบุหมวดหมู่ที่ชัดเจนของเนื้อหาที่คำนวณได้ เนื่องจากจุดสหความสัมพันธ์ที่เกิดขึ้น แม้ว่าปรากฏบนหมวดภาษาศาสตร์ก็ไม่ได้หมายความว่าเป็นเนื้อหาของหมวดภาษาศาสตร์เพียงอย่างเดียว แต่เป็นเนื้อหาที่เกิดจากแรงดึงระหว่างหมวดสังคมศาสตร์และหมวดเทคโนโลยี จึงไม่สามารถระบุได้ว่าจุดที่เกิดขึ้นมาจากเนื้อหาของหมวดใดหมวดหนึ่งได้ชัดเจน โดยที่เส้นระบุขอบเขตสหความสัมพันธ์ที่ปรากฏเป็นเส้นสมมติตามค่าองศาสำหรับการระบุพิกัดตำแหน่งของความรู้บนแผนที่ความรู้เท่านั้น ขณะเดียวกัน ผู้วิจัยเชื่อว่า การประยุกต์ใช้รูปแบบ Correlation สามารถนำมาช่วยในการอธิบาย ด้านการวิเคราะห์และการประยุกต์ใช้งาน ซึ่งจะมีประโยชน์ต่อผู้ใช้ ดังนี้

1. ประโยชน์ต่อผู้ใช้ เกี่ยวกับการใช้งานในด้านของความชัดเจน เข้าใจง่าย และมีความน่าเชื่อถือต่อการใช้ข้อมูล ซึ่งเป็นการนำเสนอในรูปแบบ Bar Graph และมีเส้นแบ่งขอบเขตของกลุ่มอย่างชัดเจน อีกทั้งยังสามารถประยุกต์ใช้เป็นเครื่องมือช่วยในการแนะนำข้อมูลในกรณีที่ผู้ใช้กำลังสืบค้นสารสนเทศ และต้องการเข้าถึงข้อมูลที่สนใจเพิ่มเติมได้ ซึ่งทำให้ผู้ใช้ได้ข้อมูลที่ตรงกับความต้องการเพิ่มขึ้น

2. ประโยชน์ต่อการจัดการองค์ความรู้ในองค์กร เกี่ยวกับการเพิ่มขีดความสามารถของระบบจัดเก็บความรู้ ซึ่งโดยส่วนใหญ่มุ่งเน้นในเรื่องของการจัดเก็บ รวบรวม แลกเปลี่ยน ถ่ายทอด และเผยแพร่ แต่ไม่ได้สนใจว่าข้อมูลเหล่านั้นมีความเกี่ยวข้องกันมากหรือน้อยเพียงไร ดังนั้นเพื่อให้เป็นประโยชน์มากที่สุด ข้อมูลนั้นควรจะบอกได้ด้วยว่ามีความเกี่ยวข้องและสัมพันธ์กับข้อมูลใดบ้าง ซึ่งวิธีการ Correlation สามารถนำมาช่วยในการวิเคราะห์และนำเสนอถึงลำดับชั้นของข้อมูลที่มีความใกล้เคียงกัน ทำให้รู้ทิศทางการใช้งานองค์ความรู้ที่เป็นประโยชน์ เพราะถ้าเรามีการอธิบายข้อมูลอย่างมีประสิทธิภาพ ข้อมูลนั้นก็จะสามารถสนับสนุนการทำงานให้ประสบผลสำเร็จ และทำให้เกิดการเปลี่ยนแปลงอย่างต่อเนื่อง

3. ประโยชน์ทางด้าน Data Mining เกี่ยวกับการนำข้อมูลไปใช้ ซึ่งในงานวิจัยที่ผ่านมาโดยทั่วไปมุ่งเน้นนำเสนอที่ความสามารถเกี่ยวกับการจัดกลุ่มข้อมูลเท่านั้น หรือการให้ความสำคัญกับสารสนเทศที่ใกล้เคียงกันให้จัดอยู่ในกลุ่มเดียวกัน ซึ่งไม่ได้สนใจถึงความเกี่ยวข้องของเนื้อหาภายใน ซึ่งวิธีการ Correlation นอกจากจะมีความสามารถในการจัดกลุ่มข้อมูลถูกต้อง และลดระยะเวลาในการประมวลผล (อ้างอิงจากผลการทดลองในตารางที่ 2, 3 และ 4) ยังสามารถอธิบายความสัมพันธ์ที่ปรากฏได้ในระดับลึกถึงเนื้อหาของแต่ละสารสนเทศพร้อมทั้งแบ่งลำดับชั้นความสัมพันธ์ภายในกลุ่มได้อย่างชัดเจน

บทความนี้ เป็นการศึกษาเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มของข้อมูลด้วยวิธีการ Correlation Plot ซึ่งสำหรับงานวิจัยในอนาคตเพื่ออำนวยประโยชน์ให้มีการจำแนกความรู้แบบสหความสัมพันธ์ได้ตรงประเด็นยิ่งขึ้น อาจมีการนำไปประยุกต์ใช้กับข้อมูลประเภทอื่นที่มีความหลายหลายมากยิ่งขึ้น และเพิ่มจำนวนของข้อมูลที่ใช้ในการทดสอบ เช่น ความรู้ทางศิลปะศาสตร์ ความรู้ทางประวัติศาสตร์ หรือความรู้ทางวิทยาศาสตร์ เป็นต้น

เอกสารอ้างอิง

[1] J. Han and M. Kamber, *Data mining concepts and techniques*. United States of America: Morgan Kaufman Publishers, 2006.

[2] JianWei Li and PingHua Chen. "The application of Cluster Analysis in Library system," in *Proceedings IEEE of the International Symposium* on 21-22 Dec, Page(s) 907-910, 2008.

[3] S. Björk, "Hierarchical flip zooming: enabling parallel exploration of hierarchical visualizations," in *Proceedings of the working conference on Advanced visual interfaces Palermo*, *Italy: ACM,* 2000.

[4] Z. Huanga, Hsinchun Chena, C. J. Hsua, W. H. Chenb, and S. Wuc, "Credit rating analysis with support vector machines and neural networks,"

a market comparative study, 2004.

[5] A. Fan, and M. Palaniswami, "Selecting Bankruptcy predictors using a support vector machine Approach,"in *Proceedings of the International Joint Conference on Neural Networks*, 2000.

[6] K.S. Shin, T.K. Lee, and H.J. Kim, "An application of Support vector machines in bankruptcy prediction Model," *Expert Systems with Applications*, vol. 28, pp. 127-135, 2005.

[7] L. Yongchen and X. Honge, "Credit Rating Analysis with Support Vector Machines Optimized by Genetic Algorithm," in *Wireless Communications, Networking and Mobile Computing, WiCOM'08, 4th International Conference*, 2008.

[8] J. Watthananon and A. Mingkwan, "Multiple Relation Knowledge Mapping using Rectangular and Polar Coordinates." In *Proceedings PG Net 2009, the 10th Annual Postgraduate Symposium on The Convergence of Telecommunications, Networking and Broadcasting*, Liverpool UK. Liverpool John Moores University. [n.p. : n.p.], 2009.

[9] J. Watthananon and A. Mingkwan, "The Innovative Application of Multiple Correlation Plane," *Published in IJCSIS International Journal of Computer Science and Information Security*, vol. 8, no.9, pp. 61-69. 2010.

[10] B. Liu, "Web Data Mining: Exploring Hyperlinks Contents and Usage Data," [n.p. : n.p.], 2006.

[11] R. B. Miller, "Response time in man-computer conversational transactions." In *Proceedings of American Federation of Information Processing Societies: AFIPS Conference on Spring Joint Computer Conference*, pp.267-277, 1968.

[12] S. K Card, G. G. Robertson, and J. D. Mackinlay, "The information visualizer: An information workspace," in *Proceedings of the ACM CHI'91 Human Factors in Computing Systems Conference*, pp.181-188, 1991.

[13] M. Dewey, *Dewey Decimal Classification and Relative Index*, 21st ed. Albany, New York: OCLC Forest Press, 2003.

[14] M. Meyer, T. Girba, and M. Lungu, "Modrian: An Agile Information Visualization Framework," In *Proceeding s of ACM Symposium on Software Visualization (SofeVis 2006)*, [n.p. : n.p.], 2006.

[15] W. Lertmahakiat, "Information Retrieval using DDC Multiple Relations," Ph.D. Thesis in *the Faculty of Information Technology at King Mongkut's University of Technology North Bangkok (KMUTNB)*, 2010.

[16] W. Lertmahakiat and A. Mingkhwan, "The Innovation of Multiple Relations Information Retrieval," *The Journal of KMUTNB*, vol. 20, no. 3, pp. 514 – 523, Sep. – Dec. 2010 (in Thai).

[17] J. Watthananon and A. Mingkhwan, "A Magnify Classification technique for group of Knowledge using DDC-MR," in *The 5th National Conference on Computing and Information Technology (NCCIT 2009)*, pp. 164 – 169, 2009 (in Thai). [18] S. Aixin, L. Ee-Peng, and N. Wee-Keong, "Web Classification Using Support Vector Machine," in *Proceeding WIDM'02, the 4th International Workshop on Web Information and Data Management*, New York, USA, 2002.

[19] A. Manuel, P. Alberto, R. Juan, B. Fernando, and C. Fidel "Extracting lists of data records from semi-structured web pages," *Published in Elsevier Science Journal Data & Knowledge Engineering*, vol. 64, Issue 2, February, 2008.

[20] O. Tadachika, and S. Toramatsu, "Paper Classification for Recommendation on Research Support System Papits." *Published in IJCSNS International Journal*

of Computer Science and Network Security, vol. 6, no. 5A, May 2006.

[21] L. Yan, H. Edward, C. Korris, and H. Joshua, "Building a Decision Cluster Classification Model for High Dimensional Data by a Variable Weighting k-Means Method," in *Proceeding AI'08, the 21st Australasian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence*, Springer-Verlag Berlin, Heidelberg, 2008.

[22] L. Yan and H. Edward, "Building a Decision Cluster Forest Model to Classify High Dimensional Data with Multi-classes," in *Proceeding ACML'09, the 1st Asian Conference on Machine Learning: Advances in Machine Learning, Springer-Verlag Berlin, Heidelberg*, 2009.

[23] O. J. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance," *(IJCSIS) International Journal of Computer Science and Information Security*, vol. 7, no. 1, pp. 292 – 295, 2010.

[24] W. H. E. Day and H. Edelsbrunner, "Efficient Algorithms for Agglomerative Hierarchical Clustering Methods," *Journal of Classification*, vol. 1, pp. 7-24, 1984.

[25] M. C. Fuller, P. D. Biros, and D. Delen, "Exploration of Feature Selection and Advanced Classification Models for High-Stakes Deception Detection," in *Proceedings of the 41st Hawaii International Conference on System Sciences*, IEEE, 2008.

[26] C. Peery, A. M. Wei Wang, and Thu D. Nguyen, "Multi-Dimensional Search for Personal Information Management Systems," in *EDBT'08. France*, March 25-30, 2008.