วิทยานิพนธ์

การจัดกลุ่มข้อมูลด้านความปลอดภัยของอาหารด้วยวิธีการแบบสองขั้นตอน โดย SOM และขั้นตอนวิธีเชิงพันธุกรรม

CLUSTERING OF FOOD SAFETY DOCUMENTS BY USING TWO STEP CLUSTERING: SOM AND GENETIC ALGORITHMS

นายวงกต พจน์พงศ์สรรค์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

พ.ศ. 2551

ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

ปริญญา

วิทยาการคอมพิวเตอร์	วิทยาการคอมพิวเตอร์
สาขา	**ภาควิชา**

เรื่อง การจัดกลุ่มข้อมูลด้านความปลอดภัยของอาหารด้วยวิธีการแบบสองขั้นตอน โดย SOM และขั้นตอนวิธีเชิงพันธุกรรม

Clustering of Food Safety Documents by Using Two Step Clustering: SOM and Genetic Algorithms

นามผู้วิจัย นายวงกต พจน์พงศ์สรรค์

ได้พิจารณาเห็นชอบโดย

ประธานกรรมการ .....

(รองศาสตราจารย์อนงค์นาฏ ศรีวิหค, Ph.D.)

กรรมการ .....

(ผู้ช่วยศาสตราจารย์นวลวรรณ สุนทรภิษัช, วศ.ด.)

กรรมการ .....

(รองศาสตราจารย์กฤษณะ ไวยมัย, D.U.)

หัวหน้าภาควิชา .....

(ผู้ช่วยศาสตราจารย์ศิริกร จันทร์นวล, M.S.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่ 29 **เดือน** พฤษภาคม **พ.ศ.** 2551

วิทยานิพนธ์

เรื่อง

การจัดกลุ่มข้อมูลด้านความปลอดภัยของอาหารด้วยวิธีการแบบสองขั้นตอน โดย SOM และขั้นตอนวิธีเชิงพันธุกรรม

Clustering of Food Safety Documents by Using Two Step Clustering: SOM and Genetic Algorithms

โดย

นายวงกต พจน์พงศ์สรรค์

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
พ.ศ. 2551

วงกต พจน์พงศ์สรรค์ 2551: การจัดกลุ่มข้อมูลด้านความปลอดภัยของอาหารด้วยวิธีการแบบสองขั้นตอน โดย SOM และขั้นตอนวิธีเชิงพันธุกรรม ปริญญาวิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์) สาขาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์ ประธานกรรมการที่ปรึกษา: รองศาสตราจารย์อนงค์นาฏ ศรีวิหค, Ph.D. 75 หน้า

งานวิจัยนี้นำเสนอหลักการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหารออกเป็นกลุ่ม ตามความคล้ายคลึงกันของข้อมูล ซึ่งพิจารณาจากคำสำคัญของเอกสารที่กำหนดขึ้นโดยผู้เชี่ยวชาญหรือเจ้าของเอกสารและผ่านการตัดคำและกรองคำหยุด จากนั้น ใช้ระบบน้ำหนัก Inverse Document Frequency (IDF) เพื่อปรับค่า น้ำหนักของสำคัญ เพื่อทดลองเปรียบเทียบระหว่างการจัดกลุ่มข้อมูลที่ไม่มีการปรับค่าน้ำหนัก (CN) กับการจัดกลุ่มข้อมูลที่มีการนำระบบน้ำหนักมาใช้ (CW)

โดยการทดลองจะนำเสนอการใช้ขั้นตอนวิธีการจัดกลุ่มข้อมูลตามคำสำคัญของเอกสารแบบ 2 ขั้นตอน ระหว่างขั้นตอนวิธี Kohonen's Self-Organizing Maps (SOM) ซึ่งเป็นขั้นตอนวิธีทางโครงข่ายประสาทเทียม (Neural Network) มาใช้ในการหาจำนวนกลุ่มที่เหมาะสมกับข้อมูล จากนั้น นำจำนวนกลุ่มที่ได้มาทำการจัดกลุ่มเปรียบเทียบกันระหว่างขั้นตอนวิธีเชิงพันธุกรรม (Genetic Algorithms: GA) กับขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีน (K-Means) โดยเปรียบเทียบประสิทธิภาพการจัดกลุ่มจากค่าความแปรปรวน (Variance) ความคล้ายคลึงของเอกสารภายในกลุ่มโดยใช้ค่า RMSSTD และความแตกต่างระหว่างแต่ละกลุ่มโดยใช้ค่า (RS) เพื่อให้ทราบถึงขั้นตอนวิธีที่เหมาะสมที่ใช้ในการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร

จากการวิจัยพบว่า การจัดกลุ่มที่ให้ผลการจัดกลุ่มที่ดีที่สุดคือ จำนวนกลุ่ม 8 กลุ่มและขั้นตอนวิธีที่ดีที่สุดคือ ขั้นตอนวิธีการจัดกลุ่มแบบ SOM และ Genetic Algorithms เมื่อเปรียบเทียบกับการนำระบบน้ำหนักมาใช้พบว่าการใช้ระบบน้ำหนักจะให้ผลการจัดกลุ่มที่ดีกว่า

วงกต พจน์พงศ์สรรค์ 14 / 5 / 2551

ลายมือชื่อนิสิต ลายมือชื่อประธานกรรมการ

Wongkot Pojpongsan 2008: Clustering of Food Safety Documents by Using Two Step Clustering: SOM and Genetic Algorithms. Master of Science (Computer Science), Major Field: Computer Science, Department of Computer Science. Thesis Advisor: Associate Professor Anongnart Srivihok, Ph.D. 75 pages.

This research proposes an approach to improve document clustering technique to segment Food Safety documents by the similarity of documents. The keywords that provided by the specialist were used to represent the documents. Keywords were extracted from documents by using Stemming and Stopping Word Removal technique and then Inverse Document Frequency (IDF) was used to calculate keywords' weight. The comparison of clustering documents by using weighted keywords (CW) and non -weighted keywords (CN) was conducted.

Two step clustering technique was improved in this study to cluster Food Safety documents. These techniques included (1) Self-Organizing Maps (SOM) and (2) Genetic Algorithms (GA) and (3) K-Means algorithm. SOM was used find the number of clusters. Then both GA and K-Means were used to cluster data and evaluate their performances. Evaluation methods included Variance analysis, Root-mean-square Standard Deviation (RMSSTD) and R-Square (RS). RMSSTD was used to find the cohesion of data or the difference between data in the same cluster while RS was used to find the differences between clusters.

Results revealed that the appropriate cluster numbers was equal to 8 while having the best of value variance analysis, RMSTD and RS. As well, Genetic Algorithms was the best performance algorithm in clustering. When we use document weighting technique. The results of clustering with weighted keyword Documents (CW) were better than clustering with non-weighted keywords Documents (CN).

Wongkot Pojpongsan — Student's signature

Anongnart Srivihok — Thesis Advisor's signature

14 / 5 / 2008

กิตติกรรมประกาศ

ข้าพเจ้าขอกราบขอบพระคุณ รองศาสตราจารย์ ดร. อนงค์นาฏ ศรีวิหค ประธานกรรมการที่ปรึกษา และผู้ช่วยศาสตราจารย์ ดร. นวลวรรณ สุนทรภิษัช กรรมการวิชาเอก สำหรับความช่วยเหลือ และคำแนะนำที่มีให้ตลอดมา ขอกราบขอบพระคุณรองศาสตราจารย์ ดร. กฤษณะ ไวยมัย กรรมการวิชารอง ที่กรุณาให้ความช่วยเหลือในการสอบวิทยานิพนธ์ให้สำเร็จลุล่วงไปด้วยดี ขอกราบขอบพระคุณคณาจารย์ทุกท่าน ที่แนะนำ สั่งสอน และให้ความรู้แก่ข้าพเจ้าตลอดระยะเวลาการศึกษา

ขอขอบพระคุณ คุณพ่อ คุณแม่ ของข้าพเจ้า ที่ให้ความรัก ความห่วงใย คอยให้กำลังใจ พร้อมทั้งให้ความช่วยเหลือ และสนับสนุนในด้านค่าใช้จ่ายในการศึกษาจนสำเร็จ ลุล่วงได้ ขอขอบคุณเพื่อน ๆ พี่ ๆ ภาควิชาวิทยาการคอมพิวเตอร์ ทั้งรุ่นเดียวกัน รุ่นพี่ และรุ่นน้อง ที่ช่วยเหลือ และให้คำแนะนำในการทำวิทยานิพนธ์ตลอดมา ขอขอบพระคุณโครงการระบบสารสนเทศวิชาการด้านความปลอดภัยอาหาร รวมทั้งขอขอบคุณบุคลากรในโครงการทุกท่าน และสำนักงานมาตรฐานสินค้าเกษตรและอาหารแห่งชาติ ที่ให้การเอื้อเฟื้อข้อมูลสำหรับงานวิจัยนี้

สุดท้ายขอขอบคุณภาควิชาวิทยาการคอมพิวเตอร์ที่ให้โอกาสข้าพเจ้าได้เรียนรู้จนสำเร็จการศึกษา งานวิทยานิพนธ์นี้ ข้าพเจ้าหวังเป็นอย่างยิ่งว่าจะเป็นประโยชน์ต่อผู้ศึกษาค้นคว้าและสนใจ หากข้อผิดพลาดประการใด ข้าพเจ้าขอน้อมรับไว้เพื่อนำไปใช้ในการปรับปรุงให้วิทยานิพนธ์นี้มีความสมบูรณ์ยิ่งขึ้น สำหรับความดีที่ได้รับจากวิทยานิพนธ์ นี้ข้าพเจ้ามอบให้แก่ผู้มีพระคุณทุกท่าน

วงกต พจน์พงศ์สรรค์

พฤษภาคม 2551

สารบัญ

หน้า

สารบัญ	(1)
สารบัญตาราง	(2)
สารบัญภาพ	(4)
คำนำ	1
วัตถุประสงค์	3
การตรวจเอกสาร	5
อุปกรณ์และวิธีการ	41
อุปกรณ์	41
วิธีการ	41
ผลและวิจารณ์	50
ผล	50
วิจารณ์	66
สรุปและข้อเสนอแนะ	69
สรุป	69
ข้อเสนอแนะ	70
เอกสารและสิ่งอ้างอิง	71
ประวัติการศึกษา และการทำงาน	75

สารบัญตาราง

ตารางที่		หน้า
1	ตัวอย่างข้อมูลเพื่อนำมาจัดกลุ่มโดยขั้นตอนวิธี SOM	15
2	แสดงผลการจัดกลุ่มข้อมูลโดยขั้นตอนวิธี SOM	19
3	ข้อมูลเพื่อนำมาจัดกลุ่มโดยขั้นตอนวิธีแบบเค-มีน	30
4	ตัวอย่างการคำนวณค่าจุดกึ่งกลางของกลุ่ม	30
5	แสดงค่าเฉลี่ยใหม่ของแต่ละกลุ่ม	31
6	แสดงการเปรียบเทียบการทำงานของขั้นตอนวิธีในการจัดกลุ่ม	32
7	ตารางสรุปผลงานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มข้อมูลขนาดใหญ่	38
8	ตารางสรุปผลงานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มข้อมูลด้วยขั้นตอนวิธีเชิงพันธุกรรม	38
9	ตารางสรุปผลงานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มข้อมูลโดยใช้วิธีการหลายขั้นตอน	39
10	ตารางตัวอย่างคุณลักษณะของเอกสาร	46
11	ค่าพารามิเตอร์ต่างๆ ของขั้นตอนวิธีเชิงพันธุกรรมที่ใช้ในการทดลองจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร	48
12	ตัวอย่างตารางข้อมูลการจัดกลุ่มที่ไม่มีการปรับระบบน้ำหนัก (CN)	51
13	ตัวอย่างตารางข้อมูลการจัดกลุ่มที่มีการปรับระบบน้ำหนัก (CW)	52
14	ผลการวัดประสิทธิภาพการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร โดยขั้นตอนวิธี SOM ตั้งแต่ 2 - 10 กลุ่ม ในข้อมูล CN	53
15	ผลการวัดประสิทธิภาพการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร เมื่อกำหนดจำนวนกลุ่มเท่ากับ 8 กลุ่ม ในข้อมูลชุด CN	54
16	ผลการเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีโดยนำค่าของ RMSSTD มาทดสอบ Two Paired Sample Test วิธี t-Test ในข้อมูล CN โดยการวัดค่า RMSSTD	55
17	แสดงจำนวนสมาชิกและร้อยละของการจัดกลุ่มข้อมูลทั้ง 8 กลุ่มในข้อมูล CN	56
18	ผลการวัดประสิทธิภาพการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร โดยขั้นตอนวิธี SOM ตั้งแต่ 2 - 10 กลุ่ม ในข้อมูล CW	58

สารบัญตาราง (ต่อ)

ตารางที่		หน้า
19	ผลการวัดประสิทธิภาพการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร เมื่อกำหนดจำนวนกลุ่มเท่ากับ 8 กลุ่ม ในข้อมูล CW โดยการวัดค่า RMSSTD	59
20	ผลการเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีโดยนำค่าของ RMSSTD มาทดสอบ Two Paired Sample Test วิธี t-Test ในข้อมูล CW	60
21	แสดงจำนวนสมาชิกและร้อยละของการจัดกลุ่มข้อมูลทั้ง 8 กลุ่มในข้อมูล CW	61
22	แสดงการเปรียบเทียบผลการทดลองการจัดกลุ่มระหว่างข้อมูล CN และข้อมูล CW	63
23	แสดงการเปรียบเทียบประสิทธิภาพของการจัดกลุ่มที่เพิ่มขึ้นหลังปรับระบบน้ำหนักของคำ	63
24	แสดงการเปรียบเทียบเวลาและจำนวนรอบการทำงานที่ใช้ในแต่ละขั้นตอนวิธี	64
25	เปรียบเทียบเวลาและผลของจำนวนกลุ่มที่ได้ในการจัดกลุ่มของการใช้ขั้นตอนวิธี SOM ร่วมกับขั้นตอนวิธีเชิงพันธุกรรมกับการใช้ขั้นตอนวิธีเชิงพันธุกรรมเพียงวิธีเดียว ที่จำนวนกลุ่มตั้งแต่ 2-10 กลุ่ม	65
26	เปรียบเทียบผลการจัดกลุ่มกับงานวิจัยก่อนหน้าที่จำนวนกลุ่มเท่ากับ 8 ในการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร	66

สารบัญภาพ

ภาพที่		หน้า
1	แสดงขั้นตอนการทำเหมืองข้อมูล	6
2	แสดงการจัดกลุ่มข้อมูลแบบลำดับชั้น (Hierarchical Clustering)	8
3	แสดงการจัดกลุ่มแบบแบ่งส่วน	9
4	แสดงการทำงานของการจัดกลุ่มเอกสารโดยวิธีคลัสเตอริ่ง	10
5	แสดงลักษณะการทำงานของขั้นตอนวิธี SOM	11
6	ขั้นตอนการจัดกลุ่มโดยขั้นตอนวิธี SOM	12
7	แสดงขั้นตอนการทำงานของเทคนิคการจัดกลุ่มแบบ SOM	14
8	แสดงการทำงานของขั้นตอนวิธีเชิงพันธุกรรม	20
9	ภาพแสดงขั้นตอนวิธีทางคอมพิวเตอร์ของขั้นตอนวิธีเชิงพันธุกรรม	23
10	แสดงการแทนที่สายโครโมโซม	24
11	แสดงตัวอย่างการกำหนดประชากรเริ่มต้น	24
12	แสดงตัวอย่างการจัดกลุ่มข้อมูลโดยขั้นตอนวิธีแบบเค-มีน	27
13	แสดงการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีแบบเค-มีน	28
14	ขั้นตอนวิธีทางคอมพิวเตอร์ของวิธีการจัดกลุ่มข้อมูลแบบเค-มีน	29
15	หน้าจอเว็บไซต์โครงการระบบสารสนเทศวิชาการด้านความปลอดภัยของอาหาร	42
16	ตัวอย่างเอกสารด้านความปลอดภัยของอาหาร	43
17	แสดงโครงสร้างขั้นตอนการดำเนินการทดลองการจัดกลุ่มเอกสาร	44
18	ตัวอย่างคำสำคัญของเอกสารด้านความปลอดภัยของอาหาร	45
19	ตัวอย่างการตัดคำสำคัญโดยใช้โปรแกรม SWATH	45
20	ตัวอย่างการกรองคำหยุดออกจากคำสำคัญ	46
21	ตัวอย่างระเบียนข้อมูลในตาราง Metadataในฐานข้อมูล MySQL	50
22	ตัวอย่างระเบียนข้อมูลในตาราง Keyword ในฐานข้อมูล MySQL	51
23	แผนภาพวงกลมแสดงจำนวนข้อมูล (ร้อยละ) ของการจัดกลุ่ม ในข้อมูล CN	56
24	แผนภาพวงกลมแสดงจำนวนข้อมูล (ร้อยละ) ของการจัดกลุ่มในข้อมูล CW	61

คำอธิบายสัญลักษณ์และคำย่อ

CW	=	Clusters with weighted keywords document
CN	=	Clusters with non-weighted keywords document
GA	=	Genetic Algorithms
KM	=	K-Means Algorithms
RMSSTD	=	Root Mean Square Standard Deviation
RS	=	R-Square
SOM	=	Kohonen’s Self-Organizing Maps
SWATH	=	Smart Word Analysis for THai

การจัดกลุ่มข้อมูลด้านความปลอดภัยของอาหารด้วยวิธีการแบบสองขั้นตอน โดย SOM และขั้นตอนวิธีเชิงพันธุกรรม

Clustering of Food Safety Documents by Using Two Step Clustering: SOM and Genetic Algorithms

คำนำ

ปริมาณข้อมูลในปัจจุบันมีการเพิ่มขึ้นอยู่ตลอดเวลา การคัดแยกข้อมูลเหล่านี้ออกจากกันให้เป็นกลุ่ม หากกระทำโดยมนุษย์ด้วยการพิจารณาทีละชุดข้อมูล จะต้องใช้เวลานานและความถูกต้องแม่นยำในการจัดกลุ่มจะลดลง และหากไม่ทราบจำนวนกลุ่มที่แน่นอน จะทำได้ลำบากมากขึ้น การเพิ่มหรือลดจำนวนกลุ่มหมายถึงการนำเอกสารทั้งหมดมาจัดกลุ่มใหม่อีกครั้ง ทำให้เกิดความยุ่งยากในการจัดกลุ่ม จึงทำการศึกษาการจัดกลุ่มข้อมูลซึ่งเป็นส่วนหนึ่งของการทำเหมืองข้อมูล (Data Mining) เพื่อช่วยแก้ปัญหาการคัดแยกข้อมูลออกจากกันให้เป็นกลุ่มตามความคล้ายคลึงกันของข้อมูลหรือตามความสัมพันธ์บางอย่างของข้อมูล ด้วยขั้นตอนวิธีต่างๆที่หลากหลาย ซึ่งให้ผลในการจัดกลุ่มที่มีประสิทธิภาพในระดับหนึ่ง

ข้อมูลด้านความปลอดภัยอาหารเป็นข้อมูลของโครงการระบบสารสนเทศวิชาการด้านความปลอดภัยอาหาร จัดทำขึ้นเพื่อบูรณาการข้อมูลที่ได้จากแหล่งต่างๆ ในประเทศที่เกี่ยวกับความปลอดภัยด้านสินค้าเกษตรและอาหารทุกกลุ่มสินค้าให้เป็นระบบข้อมูลกลาง โดยมีการนำเสนอข้อมูลผ่านทางเว็บไซต์ จัดทำโดย สำนักงานมาตรฐานสินค้าเกษตรและอาหารแห่งชาติ (มกอช.) เพื่อสนับสนุนด้านข้อมูลให้แก่นักวิจัย เพื่อใช้สืบค้นข้อมูล และเป็นแหล่งเก็บรวบรวมเอกสาร แต่เนื่องจากข้อมูลภายในระบบมีปริมาณมาก การคัดแยกข้อมูลออกจากกันให้เป็นกลุ่ม จะพบปัญหาในการจัดกลุ่มเช่นเดียวกับการจัดกลุ่มโดยมนุษย์ กล่าวคือ ไม่ทราบจำนวนกลุ่มของข้อมูลที่แน่นอน และเกิดความสับสนในการกำหนดลักษณะเฉพาะของแต่ละกลุ่ม จึงได้ทำการค้นคว้าหาวิธีการที่เหมาะสมเพื่อนำมาใช้ปรับปรุงและแก้ปัญหาที่เกิดขึ้น เพื่อเพิ่มประสิทธิภาพของระบบและเป็นประโยชน์แก่ผู้ใช้งาน

งานวิจัยนี้จึงนำเสนอหลักการทำเหมืองข้อมูลมาใช้ในการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร ตามความคล้ายคลึงกันของข้อมูล ซึ่งพิจารณาจากคำสำคัญที่ปรากฏอยู่ภายในเอกสารที่ได้

วัตถุประสงค์

1. เพื่อนำเสนอวิธีการจัดกลุ่มข้อมูลด้วยวิธีผสมผสานกันระหว่างขั้นตอนวิธี Kohonen's Self-Organizing Maps (SOM) กับ ขั้นตอนวิธีเชิงพันธุกรรม (Genetic Algorithm - GA)

2. เพื่อจัดกลุ่มข้อมูลด้านความปลอดภัยอาหารของโครงการระบบสารสนเทศวิชาการด้านความปลอดภัยอาหาร

3. เพื่อวิเคราะห์และเปรียบเทียบประสิทธิภาพของการจัดกลุ่มของขั้นตอนวิธี SOM ร่วมกับ ขั้นตอนวิธีเชิงพันธุกรรม และการจัดกลุ่มโดยใช้ขั้นตอนวิธี SOM ร่วมกับขั้นตอนวิธีเค-มีน

ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถนำไปใช้ในการจัดกลุ่มเอกสารด้านความปลอดภัยอาหารของโครงการระบบสารสนเทศวิชาการด้านความปลอดภัยอาหาร

2. ได้ทราบถึงข้อดีและข้อเสียของการจัดกลุ่มด้วยขั้นตอนวิธี SOM ขั้นตอนวิธีเชิงพันธุกรรม และขั้นตอนวิธีเค-มีน กับข้อมูลด้านความปลอดภัยของอาหาร

ขอบเขตของงานวิจัย

1. ข้อมูลด้านความปลอดภัยอาหาร เก็บรวบรวมจากแหล่งข้อมูลภายในประเทศที่มีการนำเสนอและเผยแพร่ ตั้งแต่ปี พ.ศ. 2543 – 2548 ซึ่งเป็นเอกสารเกี่ยวกับ งานวิจัย บทความทางวิชาการ วิทยานิพนธ์ และการประชุมวิชาการ ที่ผ่านการกลั่นกรองจากผู้เชี่ยวชาญแล้ว จำนวน 4,806 เอกสาร

2. จำนวนคุณลักษณะที่นำมาใช้กำหนดเป็นตัวแทนเอกสารได้แก่คำสำคัญที่ปรากฏในแต่ละเอกสาร จำนวนทั้งหมด 4,395 คำ

3. จำนวนกลุ่มที่กำหนดในการทดลองตั้งแต่ 2 ถึง 10 กลุ่ม

4. งานวิจัยมุ่งเน้นที่ประสิทธิภาพโดยดูที่ความใกล้กันของเอกสารภายในกลุ่มและความแตกต่างระหว่างกลุ่มเป็นหลัก

การตรวจเอกสาร

ความรู้เบื้องต้นเกี่ยวกับงานวิจัย

1. การทำเหมืองข้อมูล

การทำเหมืองข้อมูล (Data mining) เป็นวิธีการค้นหาความรู้ในฐานข้อมูล (Knowledge Discovery in Databases – KDD) (Frawley, 1992) หรือการสกัดข้อมูล (Data Extraction) เพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่โดยอัตโนมัติ เพื่อให้ได้สารสนเทศ (Information) ที่มีเหตุผล และสามารถนำไปใช้ได้จริง ความรู้ที่ได้จากการทำเหมืองข้อมูลมีหลายรูปแบบ ได้แก่

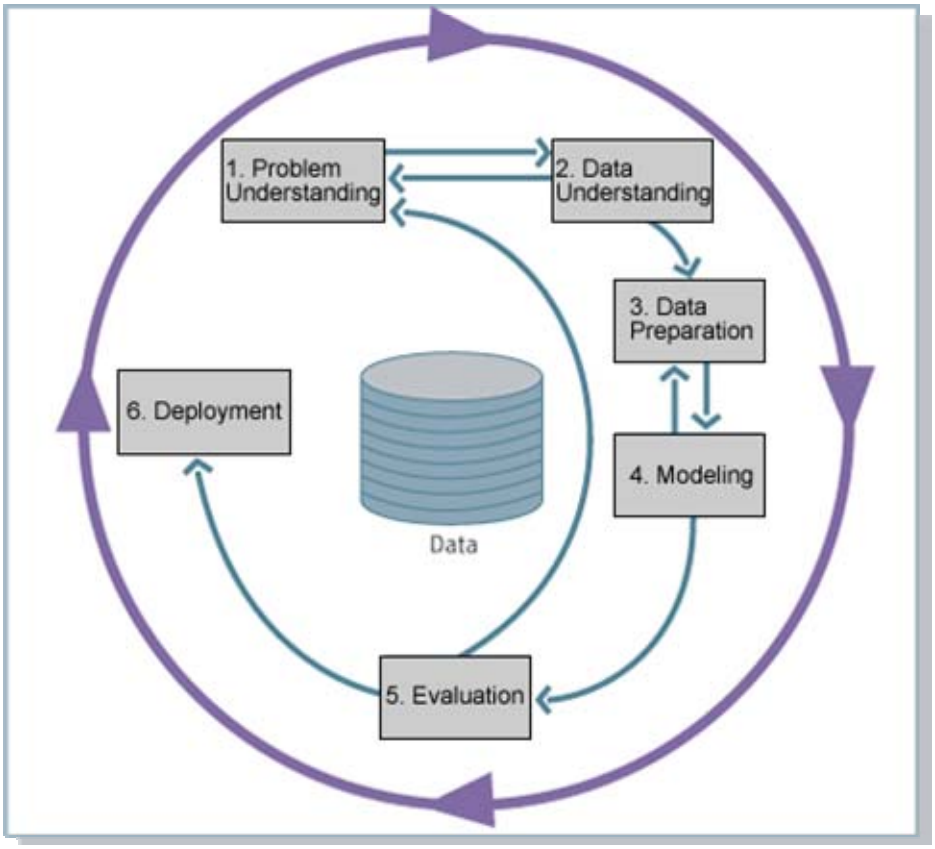
กฎความสัมพันธ์ (Association rule) แสดงความสัมพันธ์ของเหตุการณ์หรือวัตถุ ที่เกิดขึ้นพร้อมกัน เช่น การวิเคราะห์ข้อมูลการขายสินค้า โดยเก็บข้อมูลจากร้านค้าออนไลน์ แล้วพิจารณาสินค้าที่ผู้ซื้อมักจะซื้อพร้อมกัน เช่น ถ้าพบว่าคนที่ซื้อเทปวิดีโอมักจะซื้อเทปกาวด้วย ร้านค้าอาจจะจัดร้านให้สินค้าสองอย่างอยู่ใกล้กัน เพื่อเพิ่มยอดขาย หรืออาจจะพบว่าหลังจากคนซื้อหนังสือ ก แล้ว มักจะซื้อหนังสือ ข ด้วย กฎความสัมพันธ์ที่ได้สามารถนำไปใช้แนะนำผู้ที่สนใจซื้อหนังสือ ก ได้

การจำแนกข้อมูล (Data classification) เป็นการระบุประเภทของวัตถุจากคุณสมบัติของวัตถุ เช่น หาความสัมพันธ์ระหว่างผลการตรวจร่างกายต่าง ๆ กับการเกิดโรคโดยใช้ข้อมูลผู้ป่วยและการวินิจฉัยของแพทย์ที่เก็บไว้ เพื่อนำมาช่วยวินิจฉัยโรคของผู้ป่วย หรือการวิจัยทางการแพทย์ ในทางธุรกิจจะใช้การจำแนกข้อมูล เพื่อประกอบการพิจารณาการอนุมัติเงินกู้ ว่าผู้ที่มาขอกู้เงินจะก่อหนี้ดีหรือหนี้เสีย

การจัดกลุ่มข้อมูล (Data clustering) แบ่งข้อมูลที่มีลักษณะคล้ายกันออกเป็นกลุ่ม เช่น การจัดกลุ่มผู้ป่วยที่เป็นโรคเดียวกันตามลักษณะอาการ เพื่อนำไปใช้ประโยชน์ในการวิเคราะห์หาสาเหตุของโรค โดยพิจารณาจากผู้ป่วยที่มีอาการคล้ายคลึงกัน

จิตทัศน์ (Visualization) สร้างภาพคอมพิวเตอร์กราฟิกที่สามารถนำเสนอข้อมูลอย่างครบถ้วนแทนการใช้ข้อความนำเสนอข้อมูลที่มากมาย โดยอาจพบข้อมูลที่ซ้อนเร้นเมื่อดูข้อมูลชุดนั้นด้วยจิตทัศน์

ขั้นตอนการทำเหมืองข้อมูลตามโมเดลของ CRISP (CRISP, 2008) มีการดำเนินการดังภาพ



ภาพที่ 1 แสดงขั้นตอนการทำเหมืองข้อมูล
ที่มา: Adapted from CRISP (2008)

1. ทำความเข้าใจปัญหา (Problem Understanding) ว่าการทำเหมืองข้อมูลต้องการแก้ปัญหาใด เช่น เป้าหมายการทำเหมืองข้อมูลเพื่อเพิ่มจำนวนลูกค้า หรือเพื่อจัดกลุ่มข้อมูลให้เป็นหมวดหมู่ เป็นต้น

2. ทำความเข้าใจข้อมูล (Data Understanding) ทำการเก็บรวบรวมข้อมูลจากแหล่งต่างๆ แล้วทำความเข้าใจกับข้อมูล เพราะหากไม่เข้าใจจะทำให้ไม่สามารถแก้ปัญหาที่ตรงกับข้อมูลได้ อีกทั้งยังไม่สามารถได้ข้อมูลที่สมบูรณ์และถูกต้อง

3. การเตรียมข้อมูล (Data Preparation) การจัดการข้อมูลให้อยู่ในรูปแบบที่สามารถนำเข้าไปวิเคราะห์โดยขั้นตอนวิธีการทำเหมืองข้อมูล เช่น เลือกข้อมูลที่จะนำมาใช้ หรือแบ่งข้อมูลที่สนใจจากฐานข้อมูล จากนั้นจะเป็นการปรับเปลี่ยนรูปแบบข้อมูลและทำข้อมูลให้สะอาดโดยการแยกข้อมูลที่ไม่เกี่ยวข้องหรือไม่สอดคล้องกันออกไป

4. การสร้างแบบจำลอง (Modeling) เป็นการเลือกอัลกอริทึมที่เหมาะสมในการทำเหมืองข้อมูล โดยมีเทคนิคและวิธีการหลายรูปแบบ เช่น Database Segmentation Descriptive Modeling หรือ Link Analysis โดยแต่ละรูปแบบจะมีขั้นตอนวิธีต่างๆ ให้เลือกใช้แตกต่างกัน เช่น การจัดกลุ่มข้อมูลใช้ขั้นตอนวิธีเค-มีนหรือ Kohonen's Neural Net หรือการทำโมเดลทำนาย (Predictive Modeling) ใช้อัลกอริทึม CART (Classification And Regression Tree) หรือ Back Propagation Neural Net หรือการทำ Link Analysis ใช้ Apriori Algorithms เป็นต้น เมื่อเลือกขั้นตอนวิธีที่ต้องการเรียบร้อย จะมีการกำหนดรูปแบบการทดสอบและผลลัพธ์ พร้อมทั้งสร้างแบบจำลองตามขั้นตอนวิธีที่เลือกต่อไป

5. การประเมินผล (Evaluation) ทำการประเมินแบบจำลองที่สร้างขึ้นด้วยการทดลองกับสถานการณ์จริง เพื่อพิจารณาแบบจำลองนี้ได้ผลเพียงใด

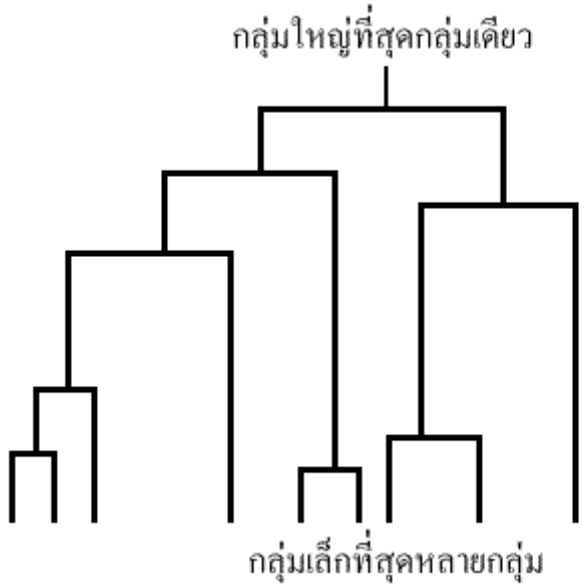
6. การนำไปใช้งาน (Deployment) เริ่มใช้งานจริง เพื่อตรวจผลการทำเหมืองข้อมูลว่าบรรลุตามเป้าหมายที่ตั้งไว้มากน้อยเพียงใด

2. ความรู้เบื้องต้นเกี่ยวกับการจัดกลุ่มเอกสาร

การจัดกลุ่มเอกสาร (Jain *et al*., 1999) เป็นกระบวนการที่แบ่งเอกสารออกเป็นกลุ่มๆ ตามความสัมพันธ์ของเนื้อหา ซึ่งมีการนำไปใช้ในสาขาวิชาหลายแขนง เช่น การเรียนรู้ของเครื่อง (Machine Learning) การทำเหมืองข้อมูล (Data Mining) การรู้จำรูปแบบ (Pattern Recognition) การวิเคราะห์รูปภาพ (Image Analysis) หรือสาขาชีวสารสนเทศ (Bioinformatics) เป็นต้น ซึ่งหลักการ

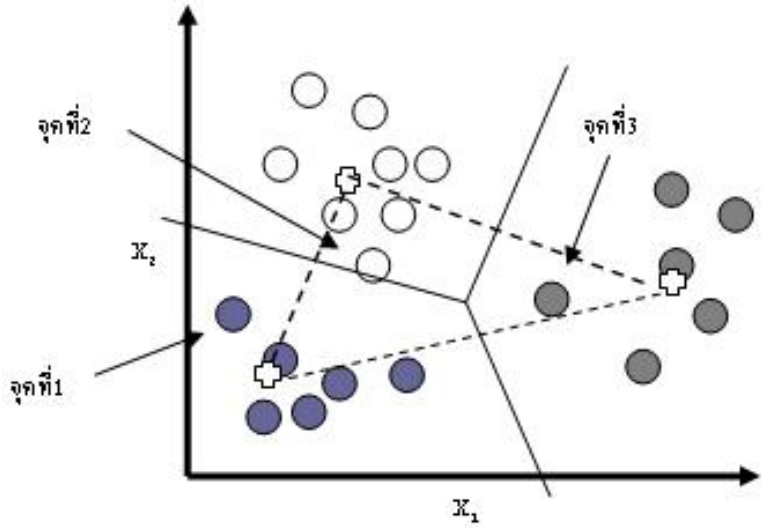
การจัดกลุ่มเอกสารสามารถแบ่งออกเป็นหลายวิธีตามขั้นตอนการทำงานดังนี้

1. การจัดกลุ่มแบบเป็นลำดับชั้น (Hierarchical Clustering) เป็นการจัดกลุ่มโดยอาศัยมุมมองการเชื่อมโยงเอกสารแบบโครงสร้างต้นไม้ ซึ่งแบ่งเป็นลำดับชั้นไป จนกว่าจะได้ผลการจัดกลุ่มที่ต้องการ ซึ่งสามารถแบ่งออกเป็น 2 ประเภทคือ Agglomerative hierarchical clustering เป็นการจัดกลุ่มที่เริ่มจากการแบ่งข้อมูลออกเป็นกลุ่มย่อยก่อน แล้วจึงรวมข้อมูลที่มีความใกล้เคียงกันเข้าด้วยกันเป็นกลุ่ม และ Divisive hierarchical clustering เป็นการจัดกลุ่มที่เริ่มจากกลุ่มข้อมูลเพียงกลุ่มเดียวแล้วแบ่งแยกออกเป็นหลายกลุ่ม



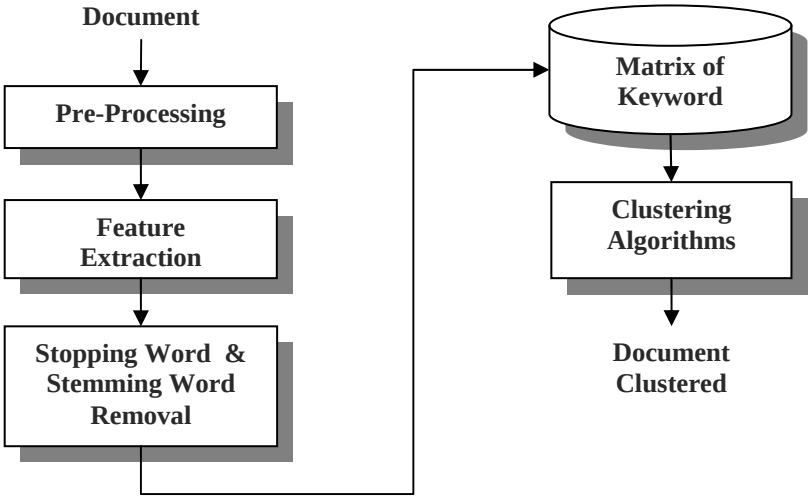
ภาพที่ 2 แสดงการจัดกลุ่มข้อมูลแบบลำดับชั้น (Hierarchical Clustering)
ที่มา: Berson (1999)

2. การจัดกลุ่มแบบไม่เป็นลำดับชั้นหรือการจัดกลุ่มแบบแบ่งส่วน (non-Hierarchical Clustering หรือ Partitional Clustering) เป็นการจัดกลุ่มที่ทำการแบ่งข้อมูลออกเป็นกลุ่มตามจำนวนกลุ่มที่กำหนด มีวิธีการจัดกลุ่มหลายแบบ ที่นิยมใช้กัน เช่น ขั้นตอนวิธี K-Means ขั้นตอนวิธี Self-Organizing Maps ขั้นตอนวิธีเชิงพันธุกรรมและ ขั้นตอนวิธี Fuzzy c-means เป็นต้น การจัดกลุ่มประเภทนี้จำเป็นต้องทราบถึงจำนวนกลุ่มที่ใช้ในการจัดกลุ่ม



ภาพที่ 3 แสดงการจัดกลุ่มแบบแบ่งส่วน

ตัวอย่างขั้นตอนการจัดกลุ่มด้วยวิธีคลัสเตอริ่ง (Grossman, 2006) เป็นวิธีที่นิยมใช้กันและสามารถให้ผลของการจัดกลุ่มที่ดี โดยทั่วไปมีขั้นตอนการทำงานดังภาพ และมีรายละเอียดในแต่ละขั้นตอนดังนี้



ภาพที่ 4 แสดงการทำงานของการจัดกลุ่มเอกสารโดยวิธีคลัสเตอริ่ง
ที่มา: Grossman (2006)

การปรุงแต่งข้อมูลให้สมบูรณ์ก่อนการวิเคราะห์ (Pre-Processing) เป็นขั้นตอนเริ่มแรก หลังจากรับเอกสารเข้ามา โดยเอกสารที่เข้ามาจะประกอบไปด้วยหัวข้อของเอกสาร เนื้อหาของเอกสาร หรือเอกสารบางประเภทเช่น ข่าวจะมีที่มาของข่าว หรือเอกสารประเภทเว็บเพจ จะมีแท็กต่างๆ ของ HTML จึงต้องทำการเอาแท็กเหล่านี้ออก

การสกัดตัวแทนเอกสาร (Feature Extraction) สกัดตัวแทนเอกสารออกจากเอกสารมีวิธีการอยู่หลายแบบ เช่น การนำคำที่ปรากฏอยู่มาใช้เป็นตัวแทนเอกสาร หรือการนำคำสำคัญ (Keywords) ของเอกสารมาใช้

การสกัดคำออกมาจากเอกสารเพื่อนำไปใช้ในการจัดกลุ่มโดยทุกๆเอกสารจะมีขั้นตอนการดำเนินการคือ สกัดคำทุกคำออกมาจากเอกสาร และจัดเก็บคำและความถี่ที่คำนั้นๆ ปรากฏในเอกสาร

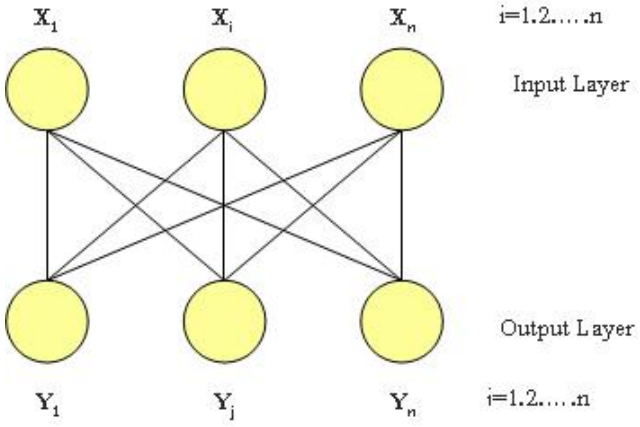
เมื่อทำการสกัดคำออกมาจากเอกสารครบทุกเอกสาร จะได้เมตริกซ์ของเอกสารโดยข้อมูลในแถวคือเอกสาร คอลัมน์คือคำที่ปรากฏในเอกสาร เพื่อใช้เป็นตัวแทนของเอกสารในขั้นตอนการจัดกลุ่มต่อไป

ขั้นตอนการตัดคำหยุด (Stop Word Removal) ขั้นตอนในการตัดคำที่ไม่เกี่ยวข้อง (stopping word) ออก เช่น คำว่า is am are if then a an the ในภาษาอังกฤษ หรือคำว่า ที่ ซึ่ง อัน แล้ว ก็ คือ ในภาษาไทย เป็นต้น และ รวมคำที่มีความหมายคล้ายกัน (word stemming) เข้าไว้ด้วยกัน เช่น คำที่มีการเติม –ed –ing ในภาษาอังกฤษ เพื่อลดคำที่ไม่ใช่คำที่เป็นตัวแทนของเอกสารออก

ขั้นตอนวิธีการจัดกลุ่ม (Clustering Algorithms) ขั้นตอนที่สำคัญที่สุดในการจัดกลุ่มเอกสาร โดยนำเอกสารต่างๆที่อยู่ในเมตริกซ์มาสร้างเป็นเวคเตอร์ตัวแทนของเอกสาร จากนั้นนำเข้าสู่ขั้นตอนวิธีการจัดหมวดหมู่ เช่น ขั้นตอนวิธีเค-มีน K-Nearest Neighbor โครงข่ายประสาทเทียม Support Vector Machine หรือขั้นตอนวิธีเชิงพันธุกรรม เป็นต้น เพื่อให้ได้ผลลัพธ์คือเอกสารที่มีการจัดกลุ่มเสร็จเรียบร้อย

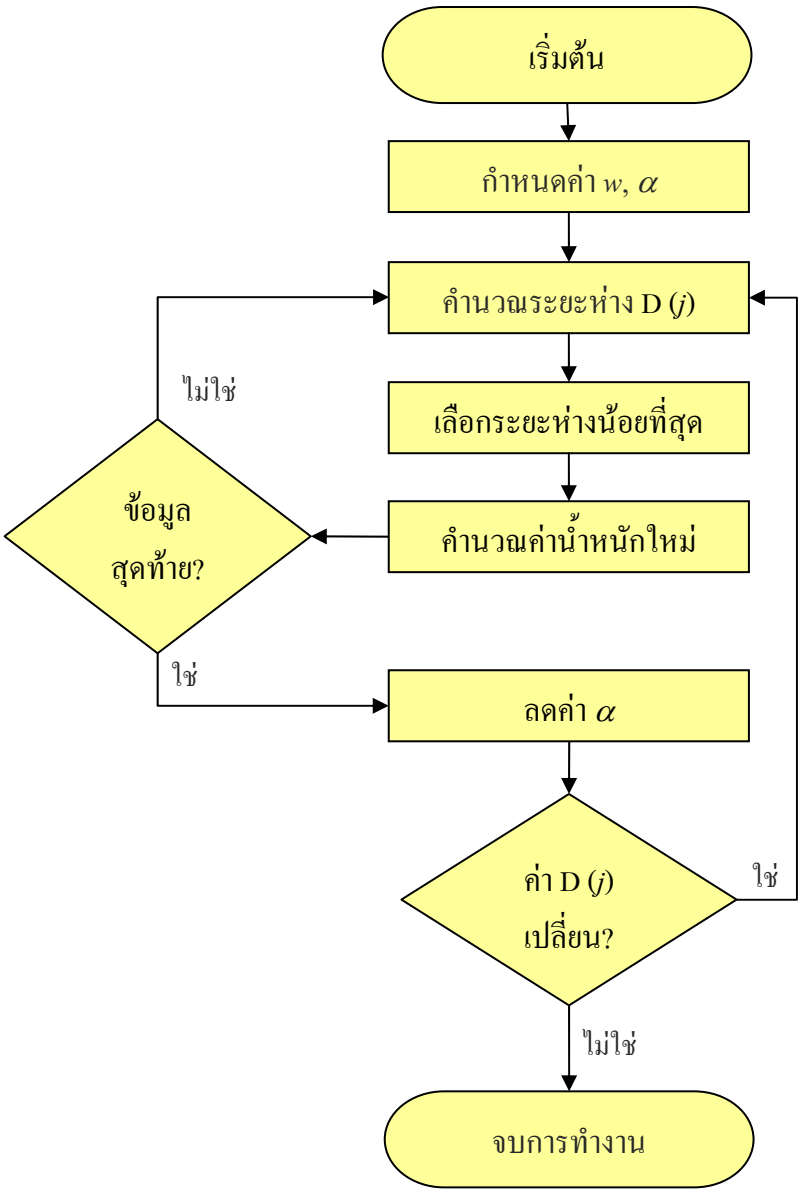
3. ความรู้เบื้องต้นเกี่ยวกับขั้นตอนวิธี Kohonen's Self-Organizing Maps

Kohonen's Self-Organizing Maps (SOM) เป็นวิธีหนึ่งของขั้นตอนวิธีแบบโครงข่ายประสาทเทียม (Kohonen, 1995) เป็นลักษณะโครงข่ายประสาทเทียมแบบไม่มีชั้นซ่อน (Hidden Layer) นำมาใช้ในการจัดกลุ่มโดยจับคู่ข้อมูลจำนวนหลายมิติในชั้นอินพุต (Input Layer) มาสู่เอ้าท์พุต (Output Layer) โดยจำนวนเอ้าท์พุตขึ้นอยู่กับจำนวนกลุ่มที่ต้องการ ดังภาพที่ 5



ภาพที่ 5 แสดงลักษณะการทำงานของขั้นตอนวิธี SOM
ที่มา: Kohonen (1995)

ขั้นตอนการทำงานจะวนรอบทำซ้ำการคำนวณปรับค่าน้ำหนักตามจำนวนกลุ่มและมิติของข้อมูล เพื่อหาค่าน้ำหนักที่เหมาะสมที่สุด โดยขั้นตอนการทำงานเป็นไปตามภาพที่ 6



ภาพที่ 6 ขั้นตอนการจัดกลุ่มโดยขั้นตอนวิธี SOM

ที่มา: Kohonen (1995)

ขั้นตอนวิธีของ SOM มีดังนี้

1. เริ่มต้นโดยการกำหนดค่าน้ำหนัก (w) และค่าความผิดพลาดที่ยอมรับได้ (α) ซึ่งทั้ง 2 ค่ามีค่าอยู่ระหว่าง 0 ถึง 1

2. หาระยะห่างของข้อมูลด้วยการคำนวณระยะห่างจากสูตร

$$D(j) = \sum_{i=1}^{n} \left(w_{i,j} - x_i \right)^2 \quad (1)$$

โดยที่ D คือระยะห่างของข้อมูล
j คือจำนวนกลุ่มที่ต้องการจัดกลุ่ม
i คือจำนวนคอลัมน์หรือมิติของข้อมูล
n คือจำนวนมิติสูงสุด
w คือค่าน้ำหนัก
x คือค่าของข้อมูล

3. นำค่าระยะห่างของแต่ละกลุ่มมาเปรียบเทียบกันเพื่อหาค่าน้อยที่สุด

4. นำน้ำหนักของกลุ่มที่มีค่าน้อยที่สุดมาคำนวณน้ำหนักใหม่จากสูตร

$$w_{i,j}(new) = w_{i,j}(old) + \alpha\left(x_i - w_{i,j}(old) \right) \quad (2)$$

โดยที่ $w\ (new)$ คือ ค่าน้ำหนักใหม่
$w\ (old)$ คือ ค่าน้ำหนักเก่า
j คือจำนวนกลุ่มที่ต้องการจัดกลุ่ม
i คือจำนวนคอลัมน์หรือมิติของข้อมูล
x คือค่าของข้อมูล

5. ทำซ้ำขั้นตอนที่ 2 ถึง 4 กับข้อมูลแถวอื่นๆ จนกระทั่งหมดทุกข้อมูล

6. ลดค่า α ลงทีละครึ่งหนึ่งของค่าเดิม

7. หากค่า *D (j)* ในแต่ละกลุ่มมีการเปลี่ยนแปลงให้ทำซ้ำตั้งแต่ขั้นตอนที่ 2 จนกว่าจะไม่มีการเปลี่ยนแปลงค่า จึงเสร็จสิ้นการทำงาน

Algorithm Basic SOM Algorithm.

1: Initialize the centroids.
2: **repeat**
3: Select the next object.
4: Determine the closest centroid to the object.
5: Update this centroid and the centroids that are close, i.e., in a specified neighborhood.
6: **until** The centroids don't change much or a threshold is exceeded.
7: Assign each object to its closest centroid and return the centroids and clusters.

ภาพที่ 7 แสดงขั้นตอนการทำงานของเทคนิคการจัดกลุ่มแบบ SOM
ที่มา: Tan *et al*. (2006)

เวลาในการประมวลผลโดยทั่วไปของขั้นตอนวิธี SOM คือ O(*n*) ในกรณีที่ข้อมูลมี 2 มิติ และมีเวลาในการประมวลผลเป็น O(*dn*) เมื่อ *d* คือ จำนวนมิติ (Dimension) ของข้อมูล (Amarasiri *et al*., 2004)

ตัวอย่างการคำนวณสำหรับขั้นตอนวิธี SOM

สมมุติมีข้อมูลอยู่ 4 รายการ โดยแต่ละชุดข้อมูลมี 4 คอลัมน์คือ X_1 X_2 X_3 และ X_4 กำหนดให้มีการจัดกลุ่มข้อมูลเป็น 2 กลุ่ม ดังตัวอย่างตารางที่ 1

ตารางที่ 1 ตัวอย่างข้อมูลเพื่อนำมาจัดกลุ่มโดยขั้นตอนวิธี SOM

X_1	X_2	X_3	X_4
1	1	0	0
0	0	0	1
1	0	0	0
0	0	1	1

กำหนดค่า α โดยใช้สูตรดังต่อไปนี้

$$\alpha(t+1) = \alpha(t)/2 \tag{3}$$

โดยกำหนดค่า α รอบแรกเป็น 0.6 เขียนในรูปสูตรจะเป็นดังนี้

$$\alpha\,(t=0) = 0.6$$

สมมุติกำหนดให้ค่าน้ำหนัก (w) แก่โครงข่ายประสาทเทียม ต่าง ๆ ที่ต้องการทำการจัดกลุ่มดังต่อไปนี้

$$\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \\ w_{41} & w_{42} \end{bmatrix} = \begin{bmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \\ 0.5 & 0.7 \\ 0.9 & 0.3 \end{bmatrix}$$

นำข้อมูลแถวแรกจากตาราง คือ 1100 มาทำการคำนวณ โดยการหาค่าด้วย สูตรที่ 1 เพื่อหาระยะห่างจากสูตรที่ 1 เมื่อแทนค่าจะได้ดังต่อไปนี้

$$\begin{aligned} D\,(1) &= (0.2-1)^2 + (0.6-1)^2 + (0.5-0)^2 + (0.9-0)^2 \\ &= 1.86 \\ D\,(2) &= (0.8-1)^2 + (0.4-1)^2 + (0.7-0)^2 + (0.3-0)^2 \\ &= 0.98 \end{aligned}$$

หลังจากได้ค่า $D(1)$ และ $D(2)$ แล้วให้เลือกตัวที่มีค่าน้อยที่สุดเพื่อมาคำนวณหาน้ำหนักใหม่ จากตัวอย่างที่คำนวณจะต้องคำนวณน้ำหนักใหม่ของ $D(2)$ (เนื่องจาก $j = 2$) ซึ่งจะต้องใช้สูตรการปรับค่าน้ำหนักจากสูตรที่ 2 ซึ่งจะแทนค่าได้ดังต่อไปนี้

$$
\begin{aligned}
w_{12} &= 0.8 + 0.6\,(1 - 0.8) \\
&= 0.92 \\
w_{22} &= 0.4 + 0.6\,(1 - 0.4) \\
&= 0.76 \\
w_{32} &= 0.7 + 0.6\,(1 - 0.7) \\
&= 0.28 \\
w_{42} &= 0.3 + 0.6\,(1 - 0.3) \\
&= 0.12
\end{aligned}
$$

ซึ่งเขียนในรูปเมตริกซ์ได้ดังต่อไปนี้

$$
\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \\ w_{41} & w_{42} \end{bmatrix} = \begin{bmatrix} 0.2 & 0.92 \\ 0.6 & 0.76 \\ 0.5 & 0.28 \\ 0.9 & 0.12 \end{bmatrix}
$$

นำข้อมูลแถวที่ 2 จากตารางมาคำนวณซ้ำ โดยใช้วิธีเดียวกับข้อมูลในแถวแรก ซึ่งจะได้ดังผลต่อไปนี้ สำหรับเวคเตอร์ 0001

$$
\begin{aligned}
D(1) &= 0.66 \\
D(2) &= 2.2768
\end{aligned}
$$

เลือก $D(1)$ (เนื่องจาก $j = 1$) หลังจากนั้นทำการคำนวณหาค่าน้ำหนักใหม่ซึ่งจะได้ดังต่อไปนี้

$$\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \\ w_{41} & w_{42} \end{bmatrix} = \begin{bmatrix} 0.08 & 0.92 \\ 0.24 & 0.76 \\ 0.20 & 0.28 \\ 0.96 & 0.12 \end{bmatrix}$$

นำค่าจากข้อมูลในแถวที่ 3 ตามตาราง คือ 1000 มาทำการคำนวณซ้ำ โดยวิธีเดียวกับข้อมูลในแถวแรก ซึ่งจะได้ดังผลต่อไปนี้ สำหรับเวคเตอร์ 1000

$$D(1) = 1.8656$$
$$D(2) = 0.6768$$

เลือก $D(2)$ (เนื่องจาก $j = 2$) หลังจากนั้นทำการคำนวณหาค่าน้ำหนักใหม่ซึ่งจะได้ดังต่อไปนี้

$$\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \\ w_{41} & w_{42} \end{bmatrix} = \begin{bmatrix} 0.08 & 0.968 \\ 0.24 & 0.304 \\ 0.20 & 0.112 \\ 0.96 & 0.948 \end{bmatrix}$$

นำค่าจากข้อมูลในแถวต่อไปคือ 0011 มาทำการคำนวณซ้ำในเหมือนกับข้อมูลในแถวแรก ซึ่งจะได้ดังผลต่อไปนี้ สำหรับเวคเตอร์ 0011

$$D(1) = 0.7056$$
$$D(2) = 2.724$$

เลือก $D(1)$ (เพราะว่า $j = 1$) หลังจากนั้นทำการคำนวณหาค่าน้ำหนักใหม่ซึ่งจะได้ดังต่อไปนี้

$$\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \\ w_{41} & w_{42} \end{bmatrix} = \begin{bmatrix} 0.032 & 0.968 \\ 0.096 & 0.304 \\ 0.680 & 0.112 \\ 0.984 & 0.948 \end{bmatrix}$$

ให้ลดค่า α ลงครึ่งหนึ่งดังต่อไปนี้

$$\alpha(1) = \alpha(0) / 2$$

$$= 0.6 / 2$$

$$= 0.3$$

คำนวณซ้ำใหม่จนกว่าไม่มีการเปลี่ยนแปลงค่า ซึ่งหลังจากคำนวณ 100 รอบผลลัพธ์ต่อไปนี้

$$\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \\ w_{41} & w_{42} \end{bmatrix} = \begin{bmatrix} 6.7\times10^{-17} & 1 \\ 2.0\times10^{-16} & 0.49 \\ 0.51 & 2.3\times10^{-16} \\ 1 & 1.0\times10^{-16} \end{bmatrix}$$

เมื่อปรับค่าในเมตริกซ์ให้อยู่ในรูปแบบที่ง่ายขึ้น ได้ผลดังนี้

$$\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \\ w_{41} & w_{42} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0.5 \\ 0.5 & 0 \\ 1 & 0 \end{bmatrix}$$

จากเมตริกซ์นี้ สามารถจัดข้อมูลได้เป็น 2 กลุ่มอย่างชัดเจน

หลังจากได้ข้อมูลน้ำหนักสุดท้ายแล้วนั้น ถ้าต้องการที่จะจัดกลุ่มข้อมูลให้นำน้ำหนักที่ได้มาคำนวณกับข้อมูลที่ต้องการจะจัดกลุ่ม ซึ่งการคำนวณตามสูตรที่ 1

ตัวอย่าง การคำนวณข้อมูลในแถวแรกคือ 1100 อยู่ในกลุ่มข้อมูลใด ซึ่งจะต้องแทนค่าได้ดังต่อไปนี้

$$D(1) = (0-1)^2 + (0-1)^2 + (0.5-0)^2 + (1-0)^2 = 3.25$$

$$D(2) = (1-1)^2 + (0.5-1)^2 + (0-0)^2 + (0-0)^2 = 0.25$$

หลังจากนั้นเลือกข้อมูลในแถวแรกอยู่ที่กลุ่มที่ 2 เพราะว่าค่า D(2) มีค่าน้อยกว่า D(1) ซึ่งจากการคำนวณครบทั้ง 4 แถวข้อมูลแถวใดอยู่ในกลุ่มข้อมูลใด ซึ่งได้ผลลัพธ์ดังตารางที่ 2

ตารางที่ 2 แสดงผลการจัดกลุ่มข้อมูลโดยขั้นตอนวิธี SOM

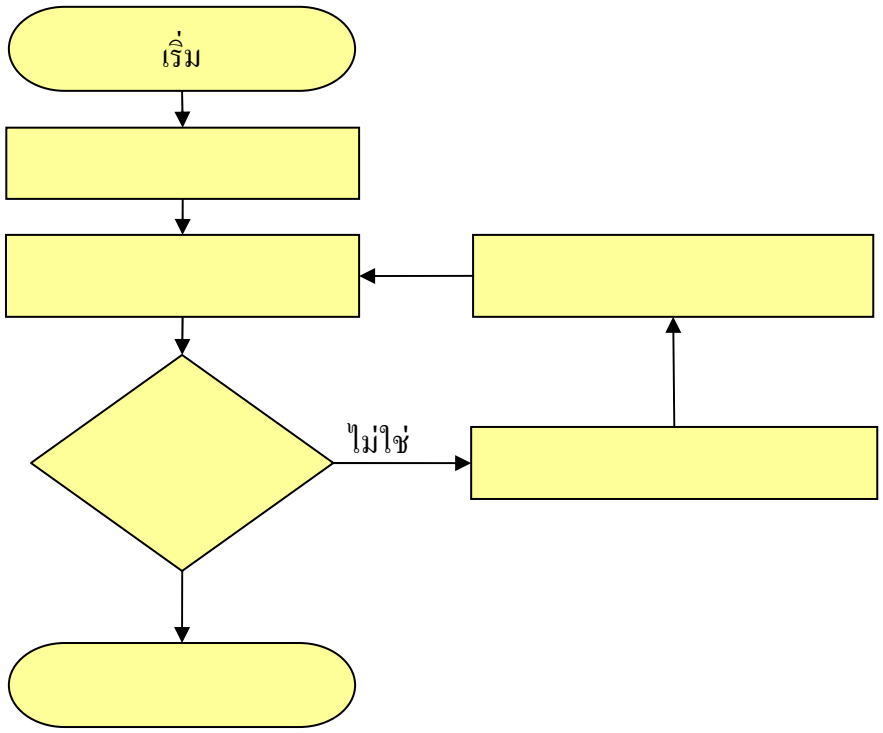
x_1	x_2	x_3	x_4	Cluster
1	1	0	0	2
0	0	0	1	1
1	0	0	0	2
0	0	1	1	1

จุดเด่นของขั้นตอนวิธีการจัดกลุ่มแบบ SOM คือ สามารถจัดการกับข้อมูลจำนวนมากได้ ส่วนข้อเสียคือ ใช้เวลาในการคำนวณสูง และ การใช้ตัวแปรต่างกันก็จะให้ผลแตกต่างกัน

4. ความรู้เบื้องต้นเกี่ยวกับขั้นตอนวิธีเชิงพันธุกรรม

ขั้นตอนวิธีเชิงพันธุกรรม (Genetic Algorithms - GA) เป็นวิธีการค้นหาคำตอบโดยมีพื้นฐานมาจากกระบวนการคัดเลือกทางธรรมชาติ (Natural Selection) และกระบวนการคัดเลือกทางพันธุศาสตร์ (Natural Genetic Selection) (Goldberg, 1989, 2002) โดยการพิจารณาและดำเนินการจากกลุ่มคำตอบของปัญหาที่ถูกสร้างขึ้นโดยการเข้ารหัส (Coding) คือการแปลงค่าตัวแปรหรือพารามิเตอร์ต่างๆ ของปัญหาให้อยู่ในรูปโครงสร้างของโครโมโซม (Chromosomes) หรือสายอักษร (String) ตามที่กำหนด โดยจะทำการคัดเลือกโครโมโซมคำตอบที่มีความเหมาะสมจากกลุ่มของโครโมโซมทั้งหมดด้วยวิธีการสุ่ม และนำโครโมโซมเหล่านี้ไปผ่านกระบวนการคัดเลือกที่เลียนแบบกระบวนการคัดเลือกทางธรรมชาติเพื่อหาโครโมโซมที่มีความเหมาะสมในการอยู่รอด โดยใช้ค่าฟังก์ชั่นความเหมาะสม (Fitness Function) ที่สอดคล้องกับฟังก์ชั่นเป้าหมาย (Objective Function) ซึ่งโครโมโซมที่มีความเหมาะสมนี้คือคำตอบที่ดีที่สุดหรือใกล้เคียงคำตอบที่ดีที่สุดของปัญหา (Optimal Solution) รูปแบบโครโมโซม ในการแก้ปัญหานั้น ได้กำหนดปัญหาเท่ากับโครโมโซมหนึ่ง ซึ่งประกอบไปด้วยยีนลักษณะต่างๆ เปรียบเหมือนกับตัวแสดงค่าคำตอบของปัญหาที่แปรผันไปตามการประยุกต์ใช้งานซึ่งได้แก่ ตัวแปร พารามิเตอร์ เงื่อนไขหรือข้อกำหนด

ลักษณะการทำงานของขั้นตอนวิธีเชิงพันธุกรรมจะเป็นการทำงานในลักษณะของการวนรอบ (Cycle) คือทำงานวนซ้ำไปเรื่อยๆตามขั้นตอนวิธีที่กำหนดโดยมีกระบวนการในการสร้างประชากรรุ่นใหม่และการคัดเลือกประชากร ทำงานไปเรื่อยๆจนกว่าจะได้ผลที่ต้องการหรือจนกว่าจะถึงจำนวนรอบที่กำหนด (ดูภาพที่ 8)



ภาพที่ 8 แสดงการทำงานของขั้นตอนวิธีเชิงพันธุกรรม

ที่มา: Goldberg (1989)

การสร้างประชากรต้นกำเนิด (Initial Population) จากการสุ่มสร้างค่าแต่ละยีนของแต่ละโครโมโซม

การวิเคราะห์ค่าความเหมาะสม ทำการวิเคราะห์ค่าความเหมาะสมของแต่ละโครโมโซม โดยถอดรหัสค่าตัวแปร พารามิเตอร์ต่างๆ ของแต่ละยีนในโครโมโซมและคำนวณค่าความเหมาะสมจากฟังก์ชั่นความเหมาะสมที่กำหนดไว้

การเลือกชุดโครโมโซมต้นแบบ ทำการสร้างชุดโครโมโซมต้นแบบ (Mating Pool) หรือชุดโครโมโซมพ่อแม่ที่สามารถอยู่รอดเป็นต้นแบบโดยใช้วิธีการคัดเลือกทาง ธรรมชาติที่พิจารณาจากค่าความเหมาะสมของแต่ละโครโมโซม ถ้าโครโมโซมใดมีค่าความเหมาะสมมากก็จะมีโอกาสถูกคัดเลือกเป็นต้นแบบมาก โดยวิธีการเลือกโครโมโซมมีวิธีการที่หลากหลาย เช่น

1. การเลือกตามลำดับที่ (Ranking Selection) เป็นวิธีการคัดเลือกโดยทำการเลือกโครโมโซมที่มีค่าความเหมาะสมสูงที่สุดในแต่ละรอบการทำงาน

2. การเลือกแบบกงล้อรูเล็ต (Roulette Wheel Selection) เป็นวิธีการคัดเลือกที่มีลักษณะคล้ายกับวงล้อรูเล็ตต์ ซึ่งโครโมโซมแต่ละตัวจะมีความน่าจะเป็นในการถูกเลือกขึ้นอยู่กับค่าความเหมาะสมของโครโมโซมตัวนั้น

3. การเลือกแบบทัวร์นาเม้น (Tournament Selection) เป็นวิธีการคัดเลือกโครโมโซมโดยใช้วิธี Roulette Wheel Selection ก่อนให้ครบตามจำนวนที่กำหนด จากนั้นทำการเลือกโครโมโซมที่ดีที่สุดจากขั้นตอนแรก

4. การเลือกแบบเฟ้นสุ่มกลุ่มตัวอย่างสากล (Stochastic Universal Sampling (Baker, 1987) เป็นวิธีการคัดเลือกที่คล้ายกับวิธี Roulette Wheel Selection แต่จะลดรอบในการหมุนวงล้อลงเหลือเพียงแค่ครั้งเดียว แล้วใช้การกำหนดระยะห่างในการเลือกแต่ละครั้งเท่าเดิม

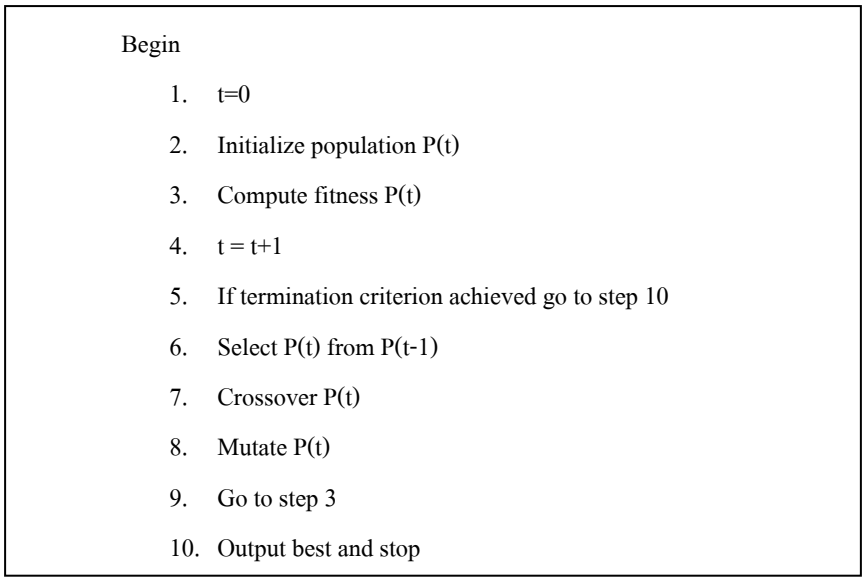
การสร้างประชากรรุ่นใหม่ เป็นการดำเนินการทางพันธุศาสตร์โดยการสุ่มจับคู่โครโมโซมต้นแบบเพื่อสร้างประชากรรุ่นใหม่ โดยอาศัยการดำเนินการทางพันธุศาสตร์ (Genetic Operation) ประกอบไปด้วย 3 กระบวนการหลักๆ ดังต่อไปนี้

1. อิลิทิสซึม (Elitism) หรือเรียกว่า การรีโปรดักชั่น (Reproduction) หรือการสืบพันธุ์ คือ กระบวนการคัดเลือกโครโมโซมที่มีความเหมาะสมสูงที่สุดในรุ่นปัจจุบันเพื่อให้อยู่รอดในประชากรรุ่นต่อไป

2. การไขว้เปลี่ยน (Crossover) เป็นขั้นตอนที่ทำภายหลังการรีโปรดักชั่น โดยการแลกเปลี่ยนส่วนของโครโมโซมพ่อแม่ (Parent) ตามอัตราความน่าจะเป็นในการไขว้เปลี่ยน (Probability of Crossover: Pc) เพื่อสร้างชุดโครโมโซมรุ่นใหม่หรือโครโมโซมรุ่นลูก (Offspring) อัตราการไขว้เปลี่ยน (Pc) นั้นเป็นพารามิเตอร์ที่สำคัญสำหรับการหาคำตอบของขั้นตอนวิธีเชิงพันธุกรรม ซึ่งก็คืออัตราส่วนของจำนวนโครโมโซมลูกที่ถูกสร้างขึ้นในแต่ละรุ่นต่อขนาดของประชากร (population size)

3. การกลายพันธุ์ (Mutation) เป็นขั้นตอนที่อาจช่วยให้โครโมโซมมีค่าความเหมาะสมดีขึ้นหลังจากการไขว้เปลี่ยน โดยการกลับค่าบางส่วนของโครโมโซมเป็นค่าใหม่ในตำแหน่งที่สุ่มได้ตามอัตราความน่าจะเป็นในการกลายพันธุ์ (Probability of Mutation: Pm) ที่กำหนด อัตราการกลายพันธุ์ (Pm) หมายถึงเปอร์เซ็นต์ของจำนวนยีนทั้งหมดในประชากรที่จะเกิดการกลายพันธุ์ขึ้น

ขั้นตอนวิธีเชิงพันธุกรรมสามารถเขียนเป็นขั้นตอนวิธีทางคอมพิวเตอร์ได้ดังภาพที่ 9

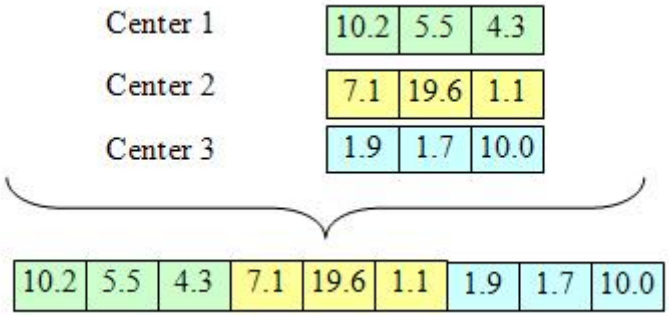


ภาพที่ 9 ภาพแสดงขั้นตอนวิธีทางคอมพิวเตอร์ของขั้นตอนวิธีเชิงพันธุกรรม
ที่มา: Goldberg (2002)

5. ความรู้เบื้องต้นเกี่ยวกับการใช้ขั้นตอนวิธีเชิงพันธุกรรมในการจัดกลุ่มข้อมูล

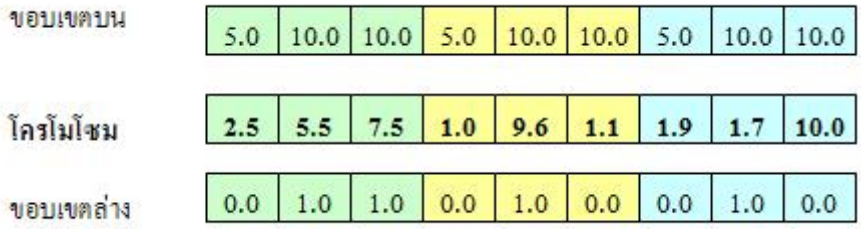
ขั้นตอนวิธีเชิงพันธุกรรม สามารถนำมาประยุกต์ใช้กับการทำงานได้หลายด้าน ซึ่งสามารถนำมาใช้ในการจัดกลุ่มข้อมูลเพื่อเพิ่มประสิทธิภาพในการทำงานได้เช่นกัน ที่สำคัญในการนำขั้นตอนวิธีเชิงพันธุกรรมมาใช้ในการจัดกลุ่มข้อมูลคือ การแทนที่ค่าในแต่ละโครโมโซม ซึ่งค่าที่นำมาใช้แทนที่โครโมโซมคือ จุดศูนย์กลางของกลุ่มข้อมูล จะแทนด้วยชุดของตัวเลขทศนิยม (Floating Point Representation) (Maulik and Bandyopadhyay, 2000) รายละเอียดในแต่ละขั้นตอนเป็นดังนี้

1. การแทนที่สายโครโมโซม จากแต่ละเอกสารจะแทนด้วยเวคเตอร์ จะนำจุดศูนย์กลางของกลุ่ม มาสร้างเป็นโครโมโซม ตัวอย่างเช่น มิติของเวคเตอร์ (N) = 3 และจำนวน cluster (K) = 3 จะแทนจุดศูนย์กลาง (C_{11}, C_{12}, C_{13}) (C_{21}, C_{22}, C_{23}) และ (C_{31}, C_{32}, C_{33}) ด้วยสายโครโมโซม (Maulik and Bandyopadhyay, 2000) ดังภาพที่ 10



ภาพที่ 10 แสดงการแทนที่สายโครโมโซม
ที่มา: Maulik and Bandyopadhyay (2000)

2. การกำหนดประชากรเริ่มต้น จะทำโดยการสุ่มค่าของแต่ละยีนขึ้นมาโดยค่าที่สุ่มจะอยู่ภายในขอบเขตบนและขอบเขตล่างของข้อมูลในแต่ละยีน ทำการสุ่มให้ครบทั้งโครโมโซม จากนั้นจะทำการสุ่มไปจนกระทั่งได้จำนวนโครโมโซมเท่ากับที่กำหนดไว้ในแต่ละรอบการทำงาน ตัวอย่างการสุ่มประชากรเริ่มต้น แสดงดังภาพที่ 11



ภาพที่ 11 แสดงตัวอย่างการกำหนดประชากรเริ่มต้น
ที่มา: Maulik and Bandyopadhyay (2000)

3. การคำนวณค่าความเหมาะสม จากการจัดกลุ่มข้อมูลที่มีการคัดแยกเอกสารที่อยู่ใกล้กันให้อยู่ในกลุ่มเดียวกัน ดังนั้นการวัดค่าความเหมาะสมของแต่ละโครโมโซมจะวัดจากการที่ชุดของโครโมโซม กล่าวคือ ชุดของจุดศูนย์กลางของกลุ่มนั้น ทำให้ข้อมูลมีค่าความใกล้กันภายในกลุ่มน้อยที่สุด ซึ่งก็คือการหาผลรวมของระยะห่างระหว่างแต่ละเอกสาร (M) ภายในกลุ่มกับจุดศูนย์กลางของกลุ่มนั้น ดังสมการที่ 4 (Maulik and Bandyopadhyay, 2000)

$$M = \sum_{i=1}^{k} \sum_{X_j \in C_i} \left\| X_j - Z_i \right\| \quad (4)$$

โดยที่ X คือเอกสารภายในกลุ่ม

Z คือจุดศูนย์กลางของกลุ่ม

k เป็นจำนวนกลุ่ม

$C_1, C_2, ..., C_k$ คือกลุ่มที่ 1 ถึง k

จะได้ค่า M ซึ่งเป็นผลรวมออกมา ซึ่งในกรณีที่ M มีค่าน้อยแสดงว่า เอกสารภายในกลุ่มมีความใกล้เคียงกันมาก จากนั้นจะหาค่าความเหมาะสม (f) โดย

$$f = 1/M \quad (5)$$

4. การเลือกประชากรต้นแบบ เป็นการเลือกประชากรเพื่อใช้ในขั้นตอนการสร้างประชากรรุ่นใหม่ โดยใช้วิธีการคัดเลือกแบบ Roulette Wheel Selection ซึ่งเป็นการเลือกที่อาศัยค่าความเหมาะสมของแต่ละโครโมโซมนำมาเป็นค่าความน่าจะเป็นในการเลือกโครโมโซม โดยจะทำการเลือกจนกว่าจะได้จำนวนโครโมโซมครบตามจำนวนสูงสุดในแต่ละรุ่น

5. การสร้างประชากรรุ่นใหม่ ประกอบด้วย 3 ขั้นตอน คือ

อิลิทิสซึม (Elitism) การเลือกโครโมโซมที่มีค่าความเหมาะสมสูงที่สุดเก็บไว้ โดยหากกำหนดจำนวนประชากรแต่ละรุ่นเป็นเลขคี่จะทำการเลือกโครโมโซมที่ดีที่สุดเก็บไว้ 1 ตัว หากกำหนดจำนวนประชากรแต่ละรุ่นเป็นเลขคู่ จะทำการเลือกโครโมโซมที่ดีที่สุดเก็บไว้ 2 ตัว ตามลำดับ

การไขว้เปลี่ยน (Crossover) จะทำแบบไขว้เปลี่ยนจุดเดียว จุดที่ใช้ในการแบ่งจะสุ่มขึ้นมาโดยมีค่าอยู่ในช่วง [1, length-1] เมื่อ length คือความยาวของสายโครโมโซม โดยมีความน่าจะเป็นที่จะเกิดการไขว้เปลี่ยนเท่ากับค่าคงที่ μ_c นั่นคือถ้าไม่เกิดการไขว้เปลี่ยนขึ้น โครโมโซมที่ได้จะเหมือนกับรุ่นพ่อแม่ทุกประการ ผลที่ได้จากการไขว้เปลี่ยนจะเกิดประชากรรุ่นลูกขึ้น 2 โครโมโซม

การกลายพันธุ์ (Mutation) ความน่าจะเป็นที่จะเกิดการกลายพันธุ์เท่ากับค่าคงที่ μ_m โดยจะเกิดขึ้นในทุกๆ ยีนในแต่ละโครโมโซมรุ่นลูกหลังจากการไขว้เปลี่ยน และเนื่องจากค่าในแต่ละยีนจะเป็นเลขทศนิยม ดังนั้นจึงใช้การกลายพันธุ์แบบไม่กำหนดค่าแน่นอน (Non-uniform mutation) (Elliott, 2002) โดยขึ้นกับค่าของจำนวนรอบการทำงานและค่าสุ่ม จากสมการที่ 6

$$v_k' = \begin{cases} v_k + \delta(i, u_k - v_k) & \text{เมื่อสุ่มได้ค่า +} \\ v_k - \delta(i, v_k - l_k) & \text{เมื่อสุ่มได้ค่า -} \end{cases} \tag{6}$$

โดยที่

$$\delta(i, y) = \delta(i, I, y, r) = y(1 - r^{(1-\frac{i}{I})^b}) \tag{7}$$

เมื่อ

	v_k	คือค่าของแต่ละยีน
	$[l_k, u_k]$	คือขอบเขตต่ำสุดและสูงสุดของข้อมูล
	i	คือรอบที่ทำงานขณะนั้น
	I	คือจำนวนรอบการทำงานสูงสุดที่กำหนด
	R	คือ ค่าสุ่มอยู่ระหว่าง 0 ถึง 1
	b	คือ ค่าดีกรีของ non-uniformity = 5

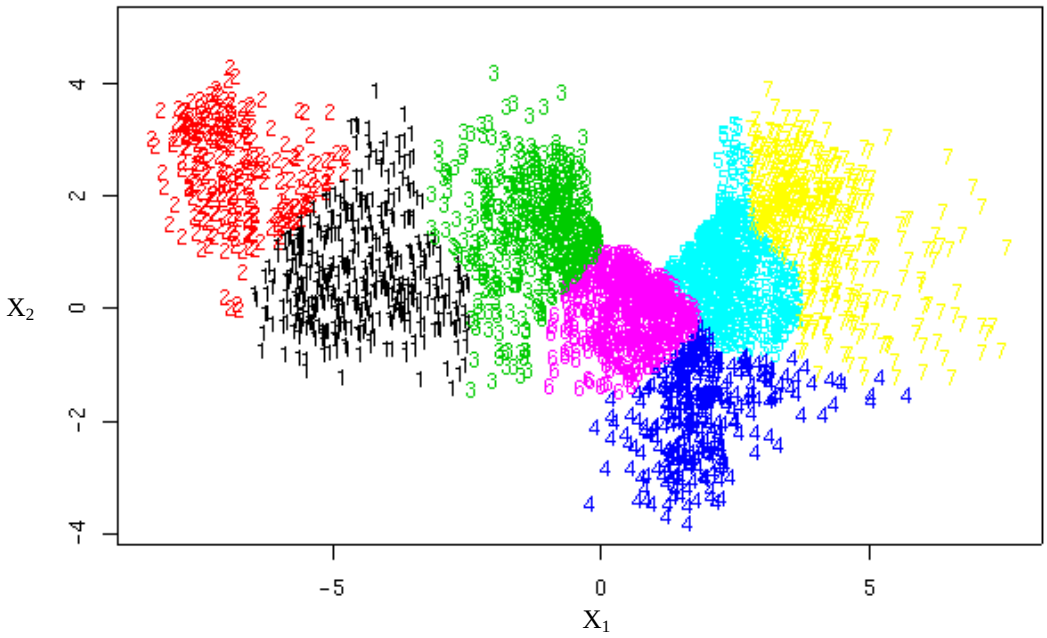
6. การสิ้นสุดการทำงาน เมื่อครบจำนวนรอบการทำงาน (Iteration)

เวลาในการประมวลผลโดยทั่วไปของขั้นตอนวิธีเชิงพันธุกรรมคือ O ($I*K*c*d*p*n^2$) (Rebelsky, 1997) โดยที่ I คือจำนวนรอบการทำงาน K คือจำนวนกลุ่มของข้อมูล c คือจำนวนโครโมโซมในแต่ละรอบการทำงาน d คือ จำนวนมิติของข้อมูล p คือจำนวนประชากรในแต่ละรุ่น และ n คือจำนวนข้อมูล

ข้อดีของการใช้การจัดกลุ่มแบบโดยขั้นตอนวิธีเชิงพันธุกรรมคือ สามารถจัดการกับข้อมูลจำนวนมากได้ และให้ผลการจัดกลุ่มที่ดีเนื่องจากมีการวนรอบการทำงานทำเรื่อยๆ และมีการเก็บผลการทำงานที่ดีที่สุดไว้เสมอ ส่วนข้อเสียคือ ใช้เวลาในการคำนวณสูง

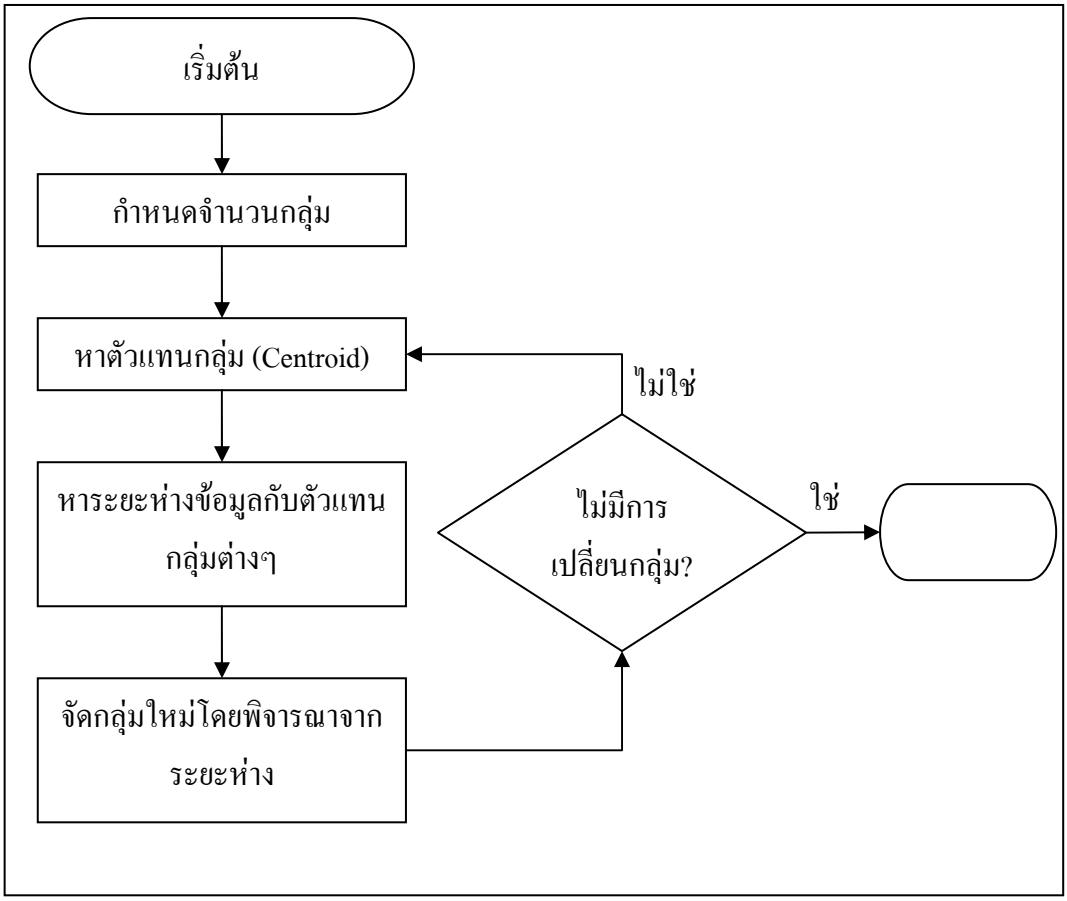
6. ความรู้เบื้องต้นเกี่ยวกับการจัดกลุ่มแบบเค-มีน (K-Means Clustering)

เทคนิคการจัดกลุ่มข้อมูลโดยขั้นตอนวิธีแบบเค-มีน (K-means Algorithm) มีการพัฒนาและนำเสนอโดย MacQueen ซึ่งแสดงให้เห็นถึงขั้นตอนวิธีในการจัดกลุ่มที่สมาชิกภายในแต่ละกลุ่ม จะมีระยะใกล้จุดศูนย์กลางหรือตัวแทนของกลุ่ม (Mean) (MacQueen, 1967)



ภาพที่ 12 แสดงตัวอย่างการจัดกลุ่มข้อมูลโดยขั้นตอนวิธีแบบเค-มีน
ที่มา: Cassidy (2002)

ขั้นตอนการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีแบบเค-มีนประกอบด้วย การกำหนดจำนวนกลุ่มเริ่มต้น กำหนดตัวแทนกลุ่ม การจัดข้อมูลแต่ละตัวเข้ากลุ่ม และการปรับปรุงตัวแทนกลุ่มในแต่ละกลุ่ม ปัญหาที่พบคือ การกำหนดจำนวนจุดเริ่มต้น ที่ส่งผลต่อประสิทธิภาพของการจัดกลุ่ม หรือการที่กลุ่มบางกลุ่มมีจำนวนสมาชิกน้อยหรืออาจจะไม่มีสมาชิกในกลุ่ม จึงมีการกำหนดกฎเกณฑ์ว่ากลุ่มที่จะอยู่ในรอบถัดไป จะต้องมีสมาชิกอยู่ไม่น้อยกว่าค่าคงที่ค่าหนึ่งที่กำหนดขึ้นมา (Bradley, 1998) มีขั้นตอนการทำงานดังภาพที่ 13



ภาพที่ 13 แสดงการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีแบบเค-มีน
ที่มา: Tan *et al*. (2006)

ขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีน มีขั้นตอนดังนี้

1. กำหนดจำนวนกลุ่มที่ต้องการ โดยกำหนดให้ K เท่ากับจำนวนกลุ่มที่ต้องการจัดกลุ่ม และเลือกจุดศูนย์กลางของกลุ่มเริ่มต้น

2. คำนวณหาตัวแทนกลุ่ม หรือจุดศูนย์กลางของกลุ่ม (Centroid) ในแต่ละกลุ่ม

3. จัดกลุ่มข้อมูลใหม่โดยพิจารณาจากค่าความใกล้ชิดหรือระยะห่างของข้อมูลในกลุ่มกับตัวแทนของกลุ่มต่างๆ ว่าข้อมูลนั้นสมควรที่จะอยู่กลุ่มใด

4. ทำการปรับปรุงการจัดกลุ่มโดยย้อนกลับไปทำข้อ 2 และจะหยุดเมื่อข้อมูลสมาชิกในกลุ่มแต่ละกลุ่มไม่มีการเปลี่ยนแปลง

ขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีนต้องการพื้นที่ (space) ที่ใช้ในการทำงานน้อย (modest) เนื่องจากเก็บเฉพาะข้อมูล (data points) และตัวแทนของกลุ่ม (centroids) เท่านั้น ดังนั้นจึงต้องการพื้นที่ (space complexity) เป็น O((*m* + *K*)*n*) เมื่อ *n* คือจำนวนข้อมูลทั้งหมด และ *m* คือจำนวนของตัวแปรที่ใช้ (attributes) ส่วนเวลาที่ใช้ในการประมวลผลจะเป็นสมการเชิงเส้น (linear) ในรูปของจำนวนข้อมูล ดังนั้นเวลา (time complexity) ที่ต้องการใช้เป็น O(*I* * *K* * *m* * *n*) เมื่อ *I* เป็นจำนวนรอบที่ใช้เมื่อตัวแทนของกลุ่มมีการเปลี่ยนแปลง และ *K* คือ จำนวนกลุ่มที่ต้องการจัดกลุ่ม เมื่อเรากำหนดค่าตัวแปรให้คงที่ เวลาที่ใช้จะเป็น O(*n*) (Tan *et al*., 2006)

Algorithm Basic K-Means algorithm.

1: Select *K* points as initial centroids.

2: **repeat**

3: From *K* clusters by assigning each point to its closest centroid.

4: Recompute the centroid of each cluster.

5: **until** Centroids do not change.

ภาพที่ 14 ขั้นตอนวิธีทางคอมพิวเตอร์ของวิธีการจัดกลุ่มข้อมูลแบบเค-มีน

ตัวอย่างการจัดกลุ่มโดยใช้ขั้นตอนวิธีแบบเค-มีน

สมมุติมีข้อมูลอยู่ 4 รายการคือ A B C และ D โดยแต่ละชุดข้อมูลมี 2 คอลัมน์ คือ X และ Y จากนั้นกำหนดให้มีการจัดกลุ่มข้อมูลเป็น 2 กลุ่ม (K=2) ดังตัวอย่างข้อมูลตารางที่ 3

ตารางที่ 3 ข้อมูลเพื่อนำมาจัดกลุ่มโดยขั้นตอนวิธีแบบเค-มีน

ข้อมูล	X	Y
A	5	3
B	-1	1
C	1	-2
D	-3	-2

การดำเนินงานมีดังนี้

1. ให้สุ่มเลือกกลุ่มโดยสุ่มจับคู่ ดังตัวอย่างจะจับคู่ A กับ B และ C กับ D หลังจากนั้น ทำการหาค่าเฉลี่ยจุดกึ่งกลาง จะได้ค่าดังตัวอย่างตารางที่ 4

ตารางที่ 4 ตัวอย่างการคำนวณค่าจุดกึ่งกลางของกลุ่ม

กลุ่ม	ค่าเฉลี่ยจุดกึ่งกลาง	
	X'	*X'*
AB	(5 + (-1)) / 2 = 2	(5 + (-1)) / 2 = 2
CD	(1 + (-3)) / 2 = -1	(1 + (-3)) / 2 = -1

2. หาค่าระยะห่างระหว่างข้อมูลกับกลุ่มโดยใช้สูตรการหาระยะห่างแบบ Euclidean Distance ดังสูตรที่ 8

$$d(A,B) = (x_1 - x_2)^2 + (y_1 - y_2)^2 \tag{8}$$

เมื่อ $A = (x_1, y_1)$

$B = (x_2, y_2)$

3. นำสูตรมาคำนวณค่าระยะห่างระหว่างจุดข้อมูลกับกลุ่ม จากนั้นจึงจัดข้อมูลให้อยู่ในกลุ่มที่ซึ่งคำนวณค่าได้น้อยกว่า

$$d\,(A,(AB)) = (5 - 2)^2 + (3 - 2)^2 = 10$$
$$d\,(A,(CD)) = (5 + 1)^2 + (3 + 2)^2 = 61$$

4. คำนวณสูตรกับข้อมูลที่เหลือให้ครบทั้งหมด

$$d\,(B,(AB)) = (-1 - 2)^2 + (1 - 2)^2 = 10$$
$$d\,(B,(CD)) = (-1 + 1)^2 + (1 + 2)^2 = 9$$

5. คำนวณหาค่าเฉลี่ยของแต่ละกลุ่มใหม่ดังตาราง

ตารางที่ 5 แสดงค่าเฉลี่ยใหม่ของแต่ละกลุ่ม

กลุ่ม	ค่าเฉลี่ยจุดกึ่งกลาง	
	X'	*Y'*
A	5	3
BCD	-1	-1

6. ทำซ้ำขั้นตอนที่กล่าวมาข้างต้นจนกว่าข้อมูลจะไม่มีการเปลี่ยนแปลงสุดท้ายแล้วจะได้กลุ่มข้อมูล 2 กลุ่มโดยแบ่งเป็นกลุ่มที่ 1 คือ A และกลุ่มที่ 2 คือ BCD

การจัดกลุ่มโดยใช้ขั้นตอนวิธีแบบเค-มีน มีข้อเด่นในการทำงานคือ ง่ายและเป็นที่นิยมในการนำไปใช้งาน อีกทั้งยังมีการทำงานที่รวดเร็ว ส่วนข้อเสียของการจัดกลุ่มโดยขั้นตอนวิธีแบบเค-มีนคือ การกำหนดจำนวนกลุ่มและตัวแทนกลุ่มเริ่มต้น มีผลต่อประสิทธิภาพการจัดกลุ่มทั้งในเชิงเวลาและความถูกต้อง และ ขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีนจะมีปัญหากับข้อมูลที่มีค่าผิดปรกติ (outlier) คือ มีข้อมูลบางตัวที่มีค่าสูงมากหรือต่ำมาก การใช้ค่าเฉลี่ยจะได้รับผลของข้อมูลผิดปรกติ ทำให้ค่าเฉลี่ยไม่เป็นตัวแทนของกลุ่มที่ดี

ตารางที่ 6 แสดงการเปรียบเทียบการทำงานของขั้นตอนวิธีในการจัดกลุ่ม

ขั้นตอนวิธี (Algorithms)	ชนิดของข้อมูล (Type of data)	เวลาที่ใช้ (Complexity) *	ข้อมูลนำเข้า (Input parameters)	ผลลัพธ์ (Result)
SOM	ตัวเลข	O(n)	จำนวนกลุ่ม, น้ำหนัก	จุดศูนย์กลางของกลุ่ม
GA	ตัวเลข	$O(n^2)$	จำนวนกลุ่ม	จุดศูนย์กลางของกลุ่ม
เค-มีน	ตัวเลข	O(n)	จำนวนกลุ่ม	จุดศูนย์กลางของกลุ่ม

หมายเหตุ * n คือ จำนวนข้อมูลทั้งหมด

ที่มา: Amarasiri *et al.* (2004), Rebelsky (1997), Tan *et al.* (2006)

7. การวัดประสิทธิภาพ

การวัดประสิทธิภาพการจัดกลุ่มข้อมูล ประกอบด้วยวิธีที่นิยมใช้ในปัจจุบัน 3 วิธี คือ (1) ค่าความแปรปรวน (2) Root Mean Square Standard Deviation และ (3) R-Square

1. ค่าความแปรปรวน (Variance)

ค่าความแปรปรวน เป็นค่าสถิติที่ใช้บ่งบอกถึงความแปรผันของข้อมูลที่เบี่ยงเบนออกจากค่าเฉลี่ย ซึ่งคำนวณได้โดยการนำค่าเบี่ยงเบน (Standard Deviation: S.D.) ยกกำลังสอง (Ronald, 1995) ซึ่งค่าที่น้อยหมายถึงได้ผลการจัดกลุ่มที่ดี

$$Variance = \frac{\sum_{i=1}^{N}(x_i - \overline{X})^2}{N} \tag{9}$$

โดยที่ N คือ จำนวนข้อมูล

x คือ ข้อมูล

$\overline{X}$ คือ จุดศูนย์กลางของกลุ่มที่ข้อมูลนั้นๆ อยู่

2. Root-mean-square Standard Deviation: RMSSTD

เป็นการวัดค่าความแตกต่างของข้อมูลภายในกลุ่ม กรณีที่ RMSSTD มีค่าน้อยแสดงว่าข้อมูลการจัดกลุ่มดีเพราะข้อมูลภายในกลุ่มมีระยะห่างกันน้อย (Kovacs *et al*., 2005)

$$RMSSTD = \sqrt{\frac{\sum_{j=1..d}^{i=1..nc} \sum_{k=1}^{n_{ij}} (x_k - x'_j)^2}{\sum_{j=1..d}^{i=1..nc} (n_{ij} - 1)}} \tag{10}$$

โดยที่ x คือข้อมูลในแต่ละตัวแปร

x'_j คือค่าเฉลี่ยของข้อมูลในกลุ่มของตัวแปรที่ j

nc คือจำนวนกลุ่ม

d คือจำนวนตัวแปร

n_{ij} คือจำนวนข้อมูลภายในกลุ่มที่ i

3. R-Square: RS

เป็นการวัดความแตกต่างของข้อมูลระหว่างกลุ่ม ถ้าค่าความแตกต่างระหว่างกลุ่มมีค่ามากแสดงว่าการจัดกลุ่มได้ผลดีเนื่องจากแต่ละกลุ่มอยู่ห่างกันมาก โดยค่า R-Square จะมีค่าระหว่าง 0 กับ 1 (Kovacs *et al*., 2005) โดยค่า R-Square มีค่าเป็น 1 หมายถึงข้อมูลมีความแตกต่างมากที่สุด ส่วนกรณีที่ R-Square มีค่าเป็น 0 นั้น แสดงว่า ข้อมูลระหว่างกลุ่มไม่มีความแตกต่าง

$$RS = \frac{SS_t - SS_w}{SS_t} \tag{11}$$

โดยที่

$$SS_t = \sum_{j=1}^{d} \sum_{k=1}^{n_j} \left(x_k - x'_j\right)^2 \tag{12}$$

และ

$$SS_w = \sum_{i=1}^{nc}\sum_{j=1}^{d}\sum_{k=1}^{n_{ij}}\left(x_k - x'_j\right)^2 \quad (13)$$

งานวิจัยที่เกี่ยวข้อง

1. งานวิจัยเกี่ยวกับการจัดกลุ่มข้อมูลขนาดใหญ่

งานวิจัยของ Sinka and Corne (2002) นำเสนอการจัดกลุ่มข้อมูลที่มีจำนวนข้อมูลมาก โดยใช้ขั้นตอนวิธีเค-มีนทำการจัดกลุ่มข้อมูลที่เป็นเอกสารเว็บจำนวน 11 กลุ่ม กลุ่มละ 1000 เอกสาร แต่การทดลองจะเปรียบเทียบครั้งละ 2 กลุ่ม โดยเลือกทั้งกลุ่มที่มีความแตกต่างกันมาก เช่น กลุ่มของธุรกิจกับกีฬา และ กลุ่มที่มีความแตกต่างกันน้อยเช่น กลุ่มของกีฬาฟุตบอลกับกีฬาเทนนิส และทำการวัดผลโดยดูจากความถูกต้องของผลของการจัดกลุ่มเทียบกับผลเฉลย ซึ่งผลที่ได้พบว่าการจัดกลุ่มโดยขั้นตอนวิธีเค-มีนให้ผลของการจัดกลุ่มมีค่าความถูกต้องสูงสุดที่ประมาณ 90% แต่ค่าเฉลี่ยของผลการทดลองจัดกลุ่มทั้งหมดอยู่ที่ประมาณ 50%

2. งานวิจัยเกี่ยวกับการจัดกลุ่มข้อมูลด้วยขั้นตอนวิธีเชิงพันธุกรรม

ในงานวิจัยของ Maulik and Bandyopadhyay (2000) ได้นำเสนอการทดลองที่ใช้ขั้นตอนวิธีเชิงพันธุกรรมมาจัดกลุ่มข้อมูลโดยเปรียบเทียบกับวิธีการจัดกลุ่มแบบขั้นตอนวิธีเค-มีน โดยใช้ชุดข้อมูลจะใช้ Artificial data sets ซึ่งเป็นชุดข้อมูลที่จำลองขึ้นมา 3 ชุด ได้แก่ Vowel data ซึ่งมีขนาด 3 มิติ จำนวน 871 แถว สำหรับ Iris data มีขนาด 4 มิติ จำนวน 50 แถว และ Crude oil data มีขนาด 5 มิติ จำนวน 56 แถว แล้ววัดผลโดยการเปรียบเทียบความใกล้กันของเอกสาร ซึ่งผลการทดลองจะพบว่าขั้นตอนวิธีเชิงพันธุกรรมจะทำงานได้ผลของการจัดกลุ่มที่ดีกว่าขั้นตอนวิธีเค-มีน

ต่อมา Painho and Bacao (2000) นำเสนองานวิจัยที่ใช้ขั้นตอนวิธีเชิงพันธุกรรมเปรียบเทียบกับขั้นตอนวิธี SOM และขั้นตอนวิธีเค-มีนในการจัดกลุ่มข้อมูลทางด้านภูมิศาสตร์ และใช้ชุดข้อมูลที่จำลองขึ้นมา การวัดผลดูที่ความถูกต้องของการจัดกลุ่มโดยใช้ค่า Square Error (SE) ซึ่ง SE ที่มีค่าน้อยหมายถึงค่าความถูกต้องที่ดี ผลการทดลองที่ได้คือขั้นตอนวิธีเชิงพันธุกรรมให้ผลที่ดีที่สุด

3. งานวิจัยเกี่ยวกับการจัดกลุ่มข้อมูลโดยใช้วิธีการหลายขั้นตอน

Lu *et al*. (2004) นำเสนอการใช้ขั้นตอนวิธีเชิงพันธุกรรมมาผสานกับขั้นตอนวิธีเค-มีน ซึ่งเป็นการรวมการทำงานตั้งแต่ 2 ขั้นตอนวิธีขึ้นไป นำมาใช้ในการวิเคราะห์รูปแบบของยีน โดยการนำเอาขั้นตอนวิธีเชิงพันธุกรรมมาใช้ในการจัดกลุ่มข้อมูลและนำขั้นตอนวิธีเค-มีนมาเพิ่มประสิทธิภาพเข้าไปภายในขั้นตอนวิธีเชิงพันธุกรรม อีกทั้งยังปรับปรุงส่วนของในกรณีที่ค่าความน่าจะเป็นในการเกิดการกลายพันธุ์มีค่าน้อยแล้วจะทำให้การทำงานช้าลง แต่ทำให้เกิดข้อด้อยในกรณีที่ค่าความน่าจะเป็นในการเกิดการกลายพันธุ์มีค่ามากขึ้นแทน สุดท้ายจึงใช้วิธีผสานงานวิจัยทั้ง 2 แบบเข้าด้วยกันเรียกว่า Hybrid Genetic K-Means Algorithm โดยการทดสอบ จะใช้ชุดข้อมูล 2 ชุดคือ ชุดข้อมูลของเซรุ่ม ที่มีจำนวนยีน 517 ยีน จำนวนลักษณะข้อมูลเท่ากับ 19 แบ่งออกเป็น 10 กลุ่ม และ ชุดข้อมูลของยีสต์จำนวน 2907 ยีน แบ่งเป็น 30 กลุ่ม ผลที่ได้คือการใช้ขั้นตอนวิธีแบบ Hybrid Genetic K-Means Algorithm ใช้เวลาที่ดีกว่า มีความพ้องกัน (convergence) ของข้อมูลภายในกลุ่มที่ดี และมีความเสถียร (stability) ที่ดีกว่าแบบที่ไม่ใช้ขั้นตอนวิธีร่วม

จากนั้น Weng and Liu (2004) ได้ใช้วิธีการผสมผสานขั้นตอนวิธีระหว่างขั้นตอนวิธี SOM และขั้นตอนวิธีเค-มีนเพื่อนำมาประยุกต์ใช้ในเรื่องของการทำระบบการแนะนำลูกค้าของการตลาดแบบหนึ่งต่อหนึ่ง โดยข้อมูลที่ใช้ทำงานวิจัย คือ ข้อมูลของลูกค้าที่ซื้อสินค้าและข้อมูลของสินค้าที่ถูกสั่งซื้อ หรือประเภทสินค้าและบริการที่ลูกค้าสนใจ โดยวิธีการทดลอง คือ ผู้วิจัยใช้คุณลักษณะ (Feature) ของสินค้า และลูกค้าในการจัดกลุ่ม โดยในขั้นตอนที่ 1 ใช้ขั้นตอนวิธี SOM จะได้ข้อมูลจำนวนของกลุ่มที่ดีที่สุด จากนั้นนำข้อมูลที่ได้ คือจำนวนกลุ่มไปเป็นข้อมูลนำเข้าของ ขั้นตอนวิธีในขั้นตอนที่ 2 คือขั้นตอนวิธีเค-มีนสุดท้ายก็จะได้ข้อมูลที่ถูกจัดเป็นกลุ่มแล้ว ซึ่งจากการทดลองพบว่าจำนวนกลุ่มที่ดีที่สุด คือ 9 กลุ่ม นั่นคือให้ค่า CV (Coefficient of Variance) ที่ดีที่สุดจากจำนวนกลุ่มทั้งหมด

ส่วนงานวิจัยของ Chaimeun and Srivihok (2005) ได้นำเสนอการใช้ขั้นตอนการจัดกลุ่มด้วยขั้นตอนวิธี SOM ร่วมกับขั้นตอนวิธีเค-มีนแบบ 2 ขั้นตอน (Two Step Clustering) โดยใช้ขั้นตอนวิธี SOM หาจำนวนกลุ่มและจัดกลุ่มโดยขั้นตอนวิธีเค-มีนต่อไป ใช้ในการศึกษาพฤติกรรมของลูกค้า และนำมาใช้ปรับปรุงและวางแผนกลยุทธ์ทางการตลาด โดยข้อมูลที่นำมาใช้คือ ข้อมูลของลูกค้าที่ซื้อสินค้าสินค้าหัตถกรรมของบริษัทแห่งหนึ่ง ผลการวิจัยสามารถจัดกลุ่มข้อมูลออกได้เป็น 9 กลุ่มที่มีลักษณะเฉพาะแตกต่างกัน

งานวิจัยของ Phattaraworamet and Rattanapongporn (2006) มีการใช้ขั้นตอนวิธีเชิงพันธุกรรมร่วมกับขั้นตอนวิธีเค-มีนคือนำขั้นตอนวิธีเค-มีนมาช่วยในการคำนวณค่าความเหมาะสม (fitness) ภายในขั้นตอนวิธีเชิงพันธุกรรม ซึ่งได้ทดสอบกับข้อมูลทั้งหมด 638 จำนวน และมีลักษณะเฉพาะเท่ากับ 2 ผลการทดลองได้เปรียบเทียบความเร็วในการทำงานและจำนวนรุ่นของประชากร เทียบกับการใช้ขั้นตอนวิธีเชิงพันธุกรรมเพียงอย่างเดียวในการจัดกลุ่ม ผลที่ได้คือการใช้ขั้นตอนวิธีเค-มีนมาช่วยจะทำให้เวลาและจำนวนรอบในการทำงานลดลงอย่างมาก

4. งานวิจัยเกี่ยวกับการวัดประสิทธิภาพการจัดกลุ่มข้อมูล

งานวิจัยที่นำเสนอเกี่ยวกับวิธีในการวัดประสิทธิภาพประกอบด้วย งานวิจัยของ Halkidi และคณะ (2001) ศึกษากระบวนการในการจัดกลุ่ม หลักการพื้นฐานของการจัดกลุ่ม รวมถึงวิธีในการวัดประสิทธิภาพ โดยทำการเปรียบเทียบวิธีในการวัดประสิทธิภาพเพื่อหาจำนวนกลุ่มที่เหมาะสม ประกอบด้วย 4 วิธี คือ (1) Davies-Bouldin (2) RMSSTD (3) RS และ (4) SD Validity index โดยใช้ข้อมูลทดสอบ 5 ชุด ประกอบด้วย ข้อมูลมาตรฐานจำนวน 4 ชุด ข้อมูลชุดที่ 5 เป็นข้อมูลจริง จากผลการทดลองพบว่า SD validity index สามารถให้ผลถูกต้องถึง 3 ใน 5 ของชุดข้อมูลทดสอบการทดลองกับข้อมูลจริง ทั้ง Davies-Bouldin RMSSTD RS และ SD Validity index ให้จำนวนกลุ่มที่เท่ากัน

จากการศึกษางานวิจัยที่ผ่านมาพบว่า ขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีนต้องกำหนดค่าเริ่มต้นและการจัดกลุ่มที่ไม่มีสมาชิกข้อมูลหรือมีจำนวนสมาชิกจำนวนน้อย ส่งผลกระทบต่อประสิทธิภาพของขั้นตอนวิธี ค่าเริ่มต้นเหล่านี้ที่สำคัญได้แก่ จำนวนกลุ่มของข้อมูล ตัวแทนของกลุ่ม และความเป็นสมาชิก เป็นต้น งานวิจัยที่ผ่านมาพบว่าขั้นตอนวิธีเชิงพันธุกรรมให้ผลของการจัดกลุ่มที่ดีที่สุด แต่จะพบปัญหาที่ต้องใช้เวลาในการจัดกลุ่มนาน ถ้านำมาใช้คำนวณจำนวนกลุ่มของข้อมูลจะใช้เวลาในการทำงานมาก และงานวิจัยที่ผ่านมาได้ทำการทดลองกับชุดข้อมูลจริงมีปริมาณน้อยมาก จึงเป็นที่น่าสนใจว่าถ้านำขั้นตอนวิธีเชิงพันธุกรรมมาใช้จัดกลุ่มข้อมูลที่เป็นข้อมูลจริง จะให้ผลของการทดลองเป็นเช่นไร ปัจจุบันมีการนำขั้นตอนวิธี SOM มาใช้ในการจัดกลุ่มเพื่อกำหนดจำนวนกลุ่มเริ่มต้น เนื่องจากขั้นตอนวิธี SOM สามารถหาจำนวนกลุ่มที่ดีที่สุดได้ ก่อนที่จะนำผลลัพธ์ที่ได้ไปเป็นข้อมูลนำเข้าของขั้นตอนวิธีอื่น ๆ เช่น ขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีนหรือขั้นตอนวิธีเชิงพันธุกรรมต่อไป

งานวิจัยนี้จึงเลือกใช้การทำงานแบบ 2 ขั้นตอน คือ ขั้นตอนที่ 1 ขั้นตอนวิธี SOM มาใช้ในการจัดกลุ่มเพื่อให้ได้จำนวนกลุ่มที่ดีที่สุด ขั้นตอนที่ 2 นำผลลัพธ์ที่ได้จากขั้นตอนที่ 1 มาใช้เป็นข้อมูลนำเข้าของขั้นตอนที่ 2 คือ ขั้นตอนวิธีเชิงพันธุกรรม และขั้นตอนวิธีเค-มีนเพื่อเปรียบเทียบการทำงานของแต่ละขั้นตอนวิธี เพื่อนำขั้นตอนวิธีที่ให้ผลของการจัดกลุ่มที่ดีที่สุดมาใช้ในการจัดกลุ่มข้อมูลด้านความปลอดภัยของอาหารต่อไป

ตารางที่ 7 เป็นตารางสรุปงานวิจัยที่เกี่ยวข้อง ซึ่งเป็นการวิจัยที่ทำการเปรียบเทียบขั้นตอนวิธีในการจัดกลุ่ม

ตารางที่ 7 ตารางสรุปผลงานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มข้อมูลขนาดใหญ่

งานวิจัย	เทคนิค	ข้อมูล	ผล
Sinka and Corne (2002)	นำเสนอการจัดกลุ่มข้อมูลที่มีจำนวนข้อมูลมาก โดยใช้ขั้นตอนวิธีเค-มีนทำการจัดกลุ่มข้อมูลที่เป็นเอกสารเว็บทดลองเปรียบเทียบทีละ 2 กลุ่ม	เอกสารเว็บจำนวน 11 กลุ่ม กลุ่มละ 1000 เอกสาร	ขั้นตอนวิธีเค-มีนให้ผลการจัดกลุ่มมีค่าความถูกต้องสูงสุดที่ประมาณ 90% แต่ค่าเฉลี่ยของผลการทดลองจัดกลุ่มทั้งหมดอยู่ที่ประมาณ 50%

ตารางที่ 8 ตารางสรุปผลงานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มข้อมูลด้วยขั้นตอนวิธีเชิงพันธุกรรม

งานวิจัย	เทคนิค	ข้อมูล	ผล
Maulik and Bandyopadhyay (2000)	ใช้ขั้นตอนวิธีเชิงพันธุกรรมมาจัดกลุ่มข้อมูลโดยเปรียบเทียบกับขั้นตอนวิธีเค-มีน	Vowel data มีขนาด 3 มิติ จำนวน 871 แถว Iris data มีขนาด 4 มิติ จำนวน 50 แถว และ Crude oil data มีขนาด 5 มิติ จำนวน 56 แถว	ขั้นตอนวิธีเชิงพันธุกรรมจะทำงานได้ผลของการจัดกลุ่มที่ดีกว่าขั้นตอนวิธีเค-มีน
Painho and Bacao (2000)	ใช้ขั้นตอนวิธีเชิงพันธุกรรมเปรียบเทียบกับขั้นตอนวิธี SOM และขั้นตอนวิธีเค-มีนในการจัดกลุ่มข้อมูล	ข้อมูลทางด้านภูมิศาสตร์ และใช้ชุดข้อมูลที่จำลองขึ้นมา	ขั้นตอนวิธีเชิงพันธุกรรมให้ผลที่ดีที่สุด

ตารางที่ 9 ตารางสรุปผลงานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มข้อมูลโดยใช้วิธีการหลายขั้นตอน

งานวิจัย	เทคนิค	ข้อมูล	ผล
Yi Lu *et al.* (2004)	ใช้ขั้นตอนวิธีเชิงพันธุกรรมกับขั้นตอนวิธีเค-มีนในการวิเคราะห์รูปแบบของยีน โดยนำขั้นตอนวิธีเชิงพันธุกรรมมาใช้ในการจัดกลุ่มข้อมูลและนำขั้นตอนวิธีเค-มีนมาเพิ่มประสิทธิภาพเข้าไปภายในขั้นตอนวิธีเชิงพันธุกรรม	ใช้ชุดข้อมูล 2 ชุดคือ ชุดข้อมูลของเซรุ่ม มีจำนวนยีน 517 ยีน มิติเท่ากับ 19 แบ่ง 10 กลุ่ม และ ชุดข้อมูลของยีสต์จำนวน 2907 ยีน 30 กลุ่ม	ผลที่ได้คือการใช้ขั้นตอนวิธีแบบ Hybrid Genetic K-Means Algorithm ใช้เวลาที่ดีกว่า มีความพ้องกันของข้อมูลภายในกลุ่มที่ดีและมีความเสถียรที่ดีกว่าแบบที่ไม่ใช้ขั้นตอนวิธีร่วม
Weng and Liu (2004)	ผสมผสานขั้นตอนวิธีระหว่างขั้นตอนวิธี SOM และขั้นตอนวิธีเค-มีน เพื่อนำมาประยุกต์ใช้ในเรื่องของการทำระบบการแนะนำลูกค้าของการตลาดแบบหนึ่งต่อหนึ่ง	ข้อมูลของลูกค้าที่ซื้อสินค้าและสินค้าที่ถูกสั่งซื้อ หรือประเภทสินค้าและบริการที่ลูกค้าสนใจ	ผลการวิจัยสามารถพบว่าจำนวนกลุ่มที่ดีที่สุด คือ 9 กลุ่ม
Chaimeun and Srivihok (2005)	ใช้ขั้นตอนการจัดกลุ่มด้วยขั้นตอนวิธี SOM ร่วมกับขั้นตอนวิธีเค-มีนโดยใช้ขั้นตอนวิธี SOM หาจำนวนกลุ่มและจัดกลุ่มโดยขั้นตอนวิธีเค-มีน	ข้อมูลของลูกค้าที่ซื้อสินค้าหัตถกรรมของบริษัทแห่งหนึ่ง	ผลการวิจัยสามารถแบ่งกลุ่มลูกค้าได้ 9 กลุ่ม

ตารางที่ 9 (ต่อ)

งานวิจัย	เทคนิค	ข้อมูล	ผล
Phattaraworamet and Rattanapongporn (2005)	ใช้ขั้นตอนวิธีเชิงพันธุกรรมร่วมกับขั้นตอนวิธีเค-มีนโดยนำขั้นตอนเค-มีนมาช่วยในการคำนวณค่าความเหมาะสม (fitness)	ข้อมูลจำลอง 638 จำนวน และมีลักษณะเฉพาะเท่ากับ 2	ผลที่ได้คือการใช้ขั้นตอนเค-มีนมาช่วยจะทำให้เวลาและจำนวนรอบในการทำงานลดลงอย่างมาก

อุปกรณ์และวิธีการ

อุปกรณ์

1. Hardware

คอมพิวเตอร์ตั้งโต๊ะ หน่วยประมวลผล AMD Athlon64 X2 Dual Core 3600+ ความเร็ว 1.90 กิกะเฮิร์ต หน่วยความจำหลัก DDR2 667 ขนาด 2048 เมกะไบต์ หน่วยความจำสำรอง 160 กิกะไบต์ แบบ SATA II

2. Software

คอมไพลเลอร์ Java Development Kits (JDK) 6 Update 4 พร้อมกับ NetBeans 5.5.1

3. Database

MySql 5.0 สำหรับจัดเก็บข้อมูลลงฐานข้อมูล

4. Operating System

Microsoft Windows XP Service Pack 2

วิธีการ

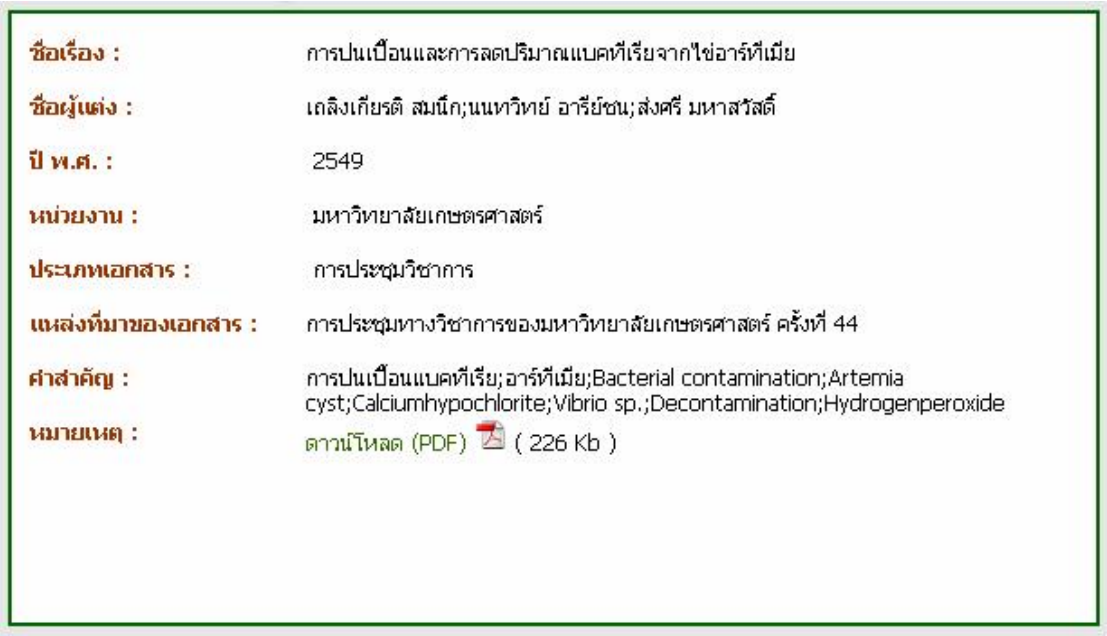
1. ข้อมูลสำหรับทำการทดลอง (Data set)

ข้อมูลที่ใช้ในการจัดกลุ่มเป็นข้อมูล เป็นข้อมูลด้านความปลอดภัยของอาหาร ซึ่งมีที่มาจากโครงการระบบสารสนเทศวิชาการด้านความปลอดภัยอาหาร (มกอช., 2550) ซึ่งจัดทำขึ้นเพื่อบูรณาการข้อมูลที่ได้จากแหล่งต่างๆ ในประเทศ ที่เกี่ยวกับความปลอดภัยด้านสินค้าเกษตรและอาหารทุก

ภาพที่ 15 หน้าจอเว็บไซต์โครงการระบบสารสนเทศวิชาการด้านความปลอดภัยของอาหาร
ที่มา: มกอช. (2007)

โดยโครงการระบบสารสนเทศวิชาการด้านความปลอดภัยอาหาร ได้จัดทำเว็บไซต์เพื่อให้สมาชิกโครงการได้ใช้เป็นแหล่งสืบค้นข้อมูล และแลกเปลี่ยนความรู้ระหว่างสมาชิก รวมถึงแจ้งข่าวสารให้แก่สมาชิกได้ทราบ

เอกสารที่โครงการระบบสารสนเทศวิชาการด้านความปลอดภัยของอาหารได้ทำการรวบรวมจากแหล่งต่างๆ ภายในประเทศไทย โดยเป็นการรวบรวมข้อมูลที่มีการนำเสนอและเผยแพร่ ตั้งแต่ปี พ.ศ. 2543 – พ.ศ. 2548 ซึ่งเป็นเอกสารเกี่ยวกับงานวิจัย บทความทางวิชาการ

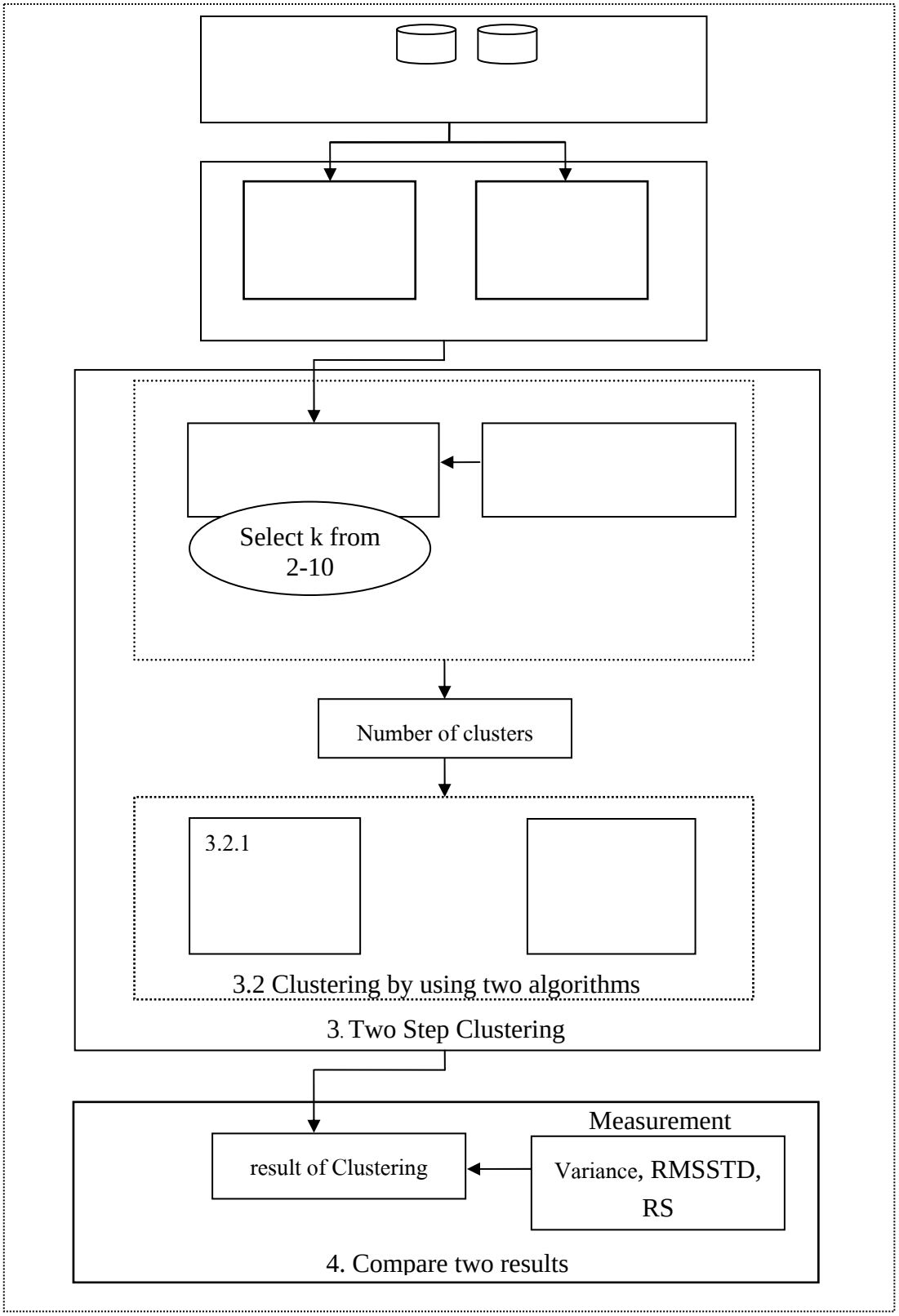


ภาพที่ 16 ตัวอย่างเอกสารด้านความปลอดภัยของอาหาร
ที่มา: มกอช. (2007)

โดยในงานวิจัยนี้จะใช้คำสำคัญที่ปรากฏในเอกสารมาใช้เป็นตัวแทนของเอกสาร

2. ขั้นตอนการทดลอง

ในงานวิจัยนี้สามารถเขียนขั้นตอนการทดลอง โดยใช้เทคนิคการทำเหมืองข้อมูล แสดงการทำงานดังภาพที่ 17



ภาพที่ 17 แสดงโครงสร้างขั้นตอนการดำเนินการทดลองการจัดกลุ่มเอกสาร

ขั้นตอนการทดลอง

1. ฐานข้อมูลด้านความปลอดภัยของอาหาร

ข้อมูลด้านความปลอดภัยของอาหาร ได้ทำการรวบรวมและถูกจัดเก็บอยู่ในฐานข้อมูลชนิด MySQL โดยมีการเก็บเป็นลักษณะระเบียน (record) ซึ่งข้อมูลที่เก็บอยู่ได้แก่ ชื่อเอกสาร ชื่อเรื่อง ผู้แต่ง ปี พ.ศ.ที่แต่ง ชนิดเอกสาร คำสำคัญ หน่วยงาน และที่มาของเอกสาร ซึ่งที่นำมาใช้ในงานวิจัยนี้คือคำสำคัญของเอกสาร

2. การเตรียมข้อมูล (Data Preprocessing)

2.1 การดึงคำสำคัญจากเอกสารมาเป็นตัวแทนเอกสาร

ในงานวิจัยนี้ ค่าคุณลักษณะที่นำมาใช้มาจากคำสำคัญที่ปรากฏอยู่ภายในเอกสาร ซึ่งกำหนดโดยผู้เชี่ยวชาญ ซึ่งคำสำคัญมีทั้งภาษาไทยและภาษาอังกฤษ ตัวอย่างของคำสำคัญที่ปรากฏในเอกสาร จะเป็นไปดังภาพที่ 18

ภาพที่ 18 ตัวอย่างคำสำคัญของเอกสารด้านความปลอดภัยของอาหาร

ซึ่งจะเห็นว่าคำดังกล่าวมีลักษณะเป็นวลีและมีการรวมคำหลายคำเอาไว้ในคำเดียว จึงมีการตัดคำโดยใช้ซอฟต์แวร์ชื่อ SWATH (เนคเทค, 2551) หลังจากตัดคำจะได้คำดังภาพที่ 19

ภาพที่ 19 ตัวอย่างการตัดคำสำคัญโดยใช้โปรแกรม SWATH

จากภาพ เมื่อผ่านการตัดคำแล้วพบว่า มีคำอยู่หลายคำที่มีลักษณะเป็นคำหยุด คือ เป็นคำที่ไม่ได้บ่งบอกถึงเอกสาร หรือไม่อาจนำมาเป็นตัวแทนของเอกสารได้ เช่น การ ความ และ หรือ ดังนั้นจึงมีการกรองคำหยุดเหล่านี้ออก ซึ่งจะเหลือคำดังภาพที่ 20

ภาพที่ 20 ตัวอย่างการกรองคำหยุดออกจากคำสำคัญ

จากนั้นนำคำที่เหลือมานับจำนวนคำและกำหนดเป็นคุณลักษณะของเอกสาร ซึ่งสามารถเขียนตัวอย่างแสดงคำสำคัญที่ใช้เป็นตัวแทนของแต่ละเอกสารได้ดังตารางที่ 10

ตารางที่ 10 ตารางตัวอย่างคุณลักษณะของเอกสาร

เอกสารที่	คุณลักษณะของเอกสาร	ความถี่ของคำ
1	ข้าว	5
1	กล้อง	1
1	เก่า	1
1	ใหม่	1
1	แช่	1
1	ร้อน	1
1	แห้ง	1

ซึ่งเอกสารทั้งหมดที่นำมาใช้ในงานวิจัย จะได้จำนวนคุณลักษณะของเอกสารเท่ากับ 4,395 คำ และจำนวนเอกสารเท่ากับ 4,806 ชุดเอกสาร

2.2 การใช้ระบบน้ำหนักเพื่อปรับค่าน้ำหนักของคำสำคัญ

ในงานวิจัยนี้ ได้มีการทดลองโดยใช้คำสำคัญเพียงอย่างเดียว เปรียบเทียบกับการนำคำสำคัญนั้นมาเข้าระบบน้ำหนักคำ โดยใช้ระบบน้ำหนัก Inverse Document Frequency (IDF) (Salton, 1983) ซึ่งมีวิธีการคำนวณดังนี้

$$IDF = 1 - \log(\left|\frac{N}{DF}\right|) \tag{14}$$

โดยที่ N คือ จำนวนของเอกสารทั้งหมด

DF คือจำนวนเอกสารที่มีคำนั้นๆปรากฏอยู่

ดังนั้น ค่าของ Term Frequency ใหม่ของคำสำคัญจะหาได้จาก

$$TFIDF = TF \times IDF \tag{15}$$

โดยที่ TF คือความถี่ของคำนั้นๆในแต่ละเอกสาร

จากนั้นจะได้ค่าของ TFIDF หรือค่าของความถี่ของคำคูณกับน้ำหนักของคำนั้นๆ ออกมา ซึ่งจะได้ข้อมูลที่ใช้ในการทดลอง 2 ชุด คือ ข้อมูลการจัดกลุ่มที่ไม่มีการปรับระบบน้ำหนัก (CN) และข้อมูลการจัดกลุ่มที่มีการปรับระบบน้ำหนัก (CW) โดยใช้วิธี TFIDF เพื่อนำไปใช้ในการทดลองต่อไป

3. การจัดกลุ่มข้อมูลโดยวิธีการแบบ 2 ขั้นตอน (Two Step Clustering)

การดำเนินการทดลองจะแบ่งเป็น 2 ส่วนคือ ส่วนแรกทำการหาจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่ม ต่อมาจึงทำการจัดกลุ่มเพื่อเลือกขั้นตอนวิธีที่ให้ผลของการจัดกลุ่มที่ดีที่สุด

3.1 การหาจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่ม (Find appropriate number of clusters) ขั้นตอนนี้จะทำการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธี SOM จัดกลุ่มข้อมูลตั้งแต่ 2 -10 กลุ่ม

3.2 การจัดกลุ่มข้อมูล (Clustering by using two algorithms) นำจำนวนกลุ่มที่ได้จากขั้นตอนแรกเป็นข้อมูลนำเข้าเพื่อใช้ในการเปรียบเทียบประสิทธิภาพของการจัดกลุ่ม ซึ่งในที่นี้ใช้ขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบขั้นตอนวิธีเชิงพันธุกรรม และขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีน เพื่อหาขั้นตอนวิธีที่มีประสิทธิภาพดีที่สุดและเหมาะสมกับข้อมูล

3.2.1 การจัดกลุ่มข้อมูลโดยขั้นตอนวิธีเชิงพันธุกรรม (Genetic Algorithms) เป็นวิธีการจัดกลุ่มที่มีการทำงานในลักษณะวนรอบเพื่อปรับค่าความเหมาะสมในทางพันธุกรรม โดยทำการกำหนดค่ารายละเอียดต่างๆ ในขั้นตอนวิธีเชิงพันธุกรรม ดังตารางที่ 11

ตารางที่ 11 ค่าพารามิเตอร์ต่างๆ ของขั้นตอนวิธีเชิงพันธุกรรมที่ใช้ในการทดลองจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร

ค่าพารามิเตอร์	การกำหนดค่า
จำนวนรอบการทำงาน	100 รอบ
การเลือกประชากรต้นแบบ	Roulette Wheel Selection
การไขว้เปลี่ยน	ใช้การไขว้เปลี่ยนจุดเดียว
อัตราการไขว้เปลี่ยน	0.8
การกลายพันธุ์	แบบ Non-uniform mutation
อัตราการกลายพันธุ์	0.04
อิลิทิสซึม	ทุกรอบการทำงาน
จำนวนประชากรสูงสุดในแต่ละรอบ	6 โครโมโซม

ส่วนสำคัญในขั้นตอนวิธีเชิงพันธุกรรมคือการวัดค่าความเหมาะสมของโครโมโซม ในงานวิจัยนี้ ได้ทำการปรับปรุงการคิดค่าความเหมาะสมใหม่ โดยนำค่าของ RMSSTD และค่า RS มาเพื่อใช้ในการคำนวณค่าความเหมาะสม เนื่องจากค่า RMSSTD และค่า RS เป็นการวัดประสิทธิภาพการจัดกลุ่มข้อมูลโดยพิจารณาถึงระยะห่างระหว่างข้อมูลภายในกลุ่ม และระยะห่าง

$$f = \left(w_{RMSSTD} \times \frac{1}{RMSSTD} \right) + \left(w_{RS} \times RS \right) \quad (16)$$

เมื่อ w_{RMSSTD} คือค่าน้ำหนักของค่า RMSSTD

w_{RS} คือ ค่าน้ำหนักของค่า RS โดยที่

$$w_{RMSSTD} + w_{RS} = 1 \quad (17)$$

โดยในการทดลองนี้ จะใช้ค่า $w_{RMSSTD} = 0.5$ และ ค่า $w_{RS} = 0.5$ เนื่องจากให้ความสำคัญของทั้งสองค่าเท่ากัน

3.2.2 การจัดกลุ่มข้อมูลโดยขั้นตอนวิธีเค-มีน เป็นขั้นตอนวิธีที่มีการใช้งานแพร่หลาย และมีขั้นตอนการทำงานที่ไม่ซับซ้อน จึงเลือกขั้นตอนวิธีนี้มาใช้ในการเปรียบเทียบ โดยการทำงานจะเป็นแบบวนรอบจนกระทั่งค่าเฉลี่ย (mean) ของกลุ่มไม่เปลี่ยนแปลง

4. การเปรียบเทียบผลการจัดกลุ่มแต่ละขั้นตอนวิธี

ทำการเปรียบเทียบผลของการจัดกลุ่มโดยใช้ค่าความแปรปรวน RMSSTD และ RS เป็นตัววัดประสิทธิภาพ สรุปผลขั้นตอนวิธีที่เหมาะสม เพื่อนำผลของการจัดกลุ่มจากขั้นตอนวิธีนั้นไปวิเคราะห์ต่อไป

ผลและวิจารณ์

ผล

1. ผลการเตรียมฐานข้อมูลด้านความปลอดภัยอาหารสำหรับการทดลอง

การเตรียมฐานข้อมูลด้านความปลอดภัยอาหารสำหรับการทดลอง จะจัดเก็บอยู่ในฐานข้อมูล MySQL ในลักษณะระเบียน (Record) ดังภาพที่ 21 และ ภาพที่ 22

Document_ID	Keyword_ID	frequency
0	0	2
0	1	1
0	2	1
0	3	1
1	0	1
1	1	1
1	4	1
1	5	1
1	6	1
1	7	1
2	1	1
2	8	1
2	9	1
2	10	1

ภาพที่ 21 ตัวอย่างระเบียนข้อมูลในตาราง Metadataในฐานข้อมูล MySQL

keyword_index	keyword_name
2967	หมอนทอง
3443	หมอไทย
498	หมัก
3860	หมัน
619	หมั่น
3588	หมีก
2434	หมุน
1232	หมุนเวียน
1539	หมู
1540	หมูยอ

ภาพที่ 22 ตัวอย่างระเบียนข้อมูลในตาราง Keyword ในฐานข้อมูล MySQL

ซึ่งเมื่อรวบรวมข้อมูลที่ใช้ในการทดลองทั้งหมด จะมีจำนวนระเบียนทั้งหมด 44,913 ระเบียน

2. ผลการเตรียมข้อมูลสำหรับการจัดกลุ่มข้อมูล

ในงานวิจัยนี้ ใช้ข้อมูลในการทดลอง 2 ชุด คือ ข้อมูลการจัดกลุ่มที่ไม่มีการปรับระบบน้ำหนัก (CN) มีจำนวนคุณลักษณะของเอกสารเท่ากับ 4,395 คำ และจำนวนเอกสารเท่ากับ 4,806 ชุดเอกสาร ซึ่งเก็บอยู่ในรูปแบบเมตริกซ์ ดังตารางที่ 12

ตารางที่ 12 ตัวอย่างตารางข้อมูลการจัดกลุ่มที่ไม่มีการปรับระบบน้ำหนัก (CN)

เอกสารที่	คำสำคัญ	ความถี่ของคำ
1	Food	2
1	Rice	1
1	Product	1
1	Improve	1
2	Food	1
2	Rice	1
2	Product	0

ข้อมูลอีกชุดคือ ข้อมูลการจัดกลุ่มที่มีการปรับระบบน้ำหนัก (CW) มีจำนวนคุณลักษณะของเอกสารเท่ากับ 4,395 คำ และจำนวนเอกสารเท่ากับ 4,806 ชุดเอกสาร เช่นกัน ซึ่งเก็บอยู่ในรูปแบบเมตริกซ์ ดังตารางที่ 13

ตารางที่ 13 ตัวอย่างตารางข้อมูลการจัดกลุ่มที่มีการปรับระบบน้ำหนัก (CW)

เอกสารที่	คำสำคัญ	ความถี่ของคำ
1	Food	14.46
1	Rice	6.35
1	Product	9.23
1	Improve	9.42
2	Food	7.23
2	Rice	6.35
2	Product	0

จากตารางที่ 12 และ ตารางที่ 13 เมื่อพิจารณาถึงความแตกต่าง จะพบว่า แต่ละคำสำคัญจะมีค่าน้ำหนักที่ต่างกัน ดังเช่น คำสำคัญที่ 2 และ 3 ในเอกสารที่ 1 ทั้ง 2 คำมีความถี่เท่ากับ 1 เหมือนกัน แต่เมื่อนำมาคิดระบบน้ำหนักของคำ จะพบว่าทั้ง 2 คำมีค่าที่แตกต่างกัน เป็นต้น ซึ่งข้อมูลทั้งสองกลุ่ม จะนำไปใช้ในการจัดกลุ่มข้อมูลโดยวิธีการแบบ 2 ขั้นตอน ต่อไป

3. ผลการทดลองการจัดกลุ่มข้อมูลโดยวิธีการแบบ 2 ขั้นตอน

การทดลอง เมื่อทำการเตรียมข้อมูลที่ต้องการและจัดให้อยู่ในรูปแบบที่เหมาะสมในการทำเหมืองข้อมูลแล้ว จึงนำข้อมูลที่จัดเตรียมมาทำการจัดกลุ่มข้อมูล โดยเทคนิคการจัดกลุ่มข้อมูลแบบ 2 ขั้นตอน ซึ่งขั้นตอนแรกจะใช้ขั้นตอนวิธี SOM เพื่อหาจำนวนกลุ่มที่เหมาะสม เป็นข้อมูลนำเข้าในขั้นตอนที่ 2 คือ การเปรียบเทียบวิธีในการจัดกลุ่มแบบขั้นตอนวิธีเชิงพันธุกรรมเปรียบเทียบกับขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน และเลือกวิธีที่ดีที่สุดเพื่อนำไปวิเคราะห์ผลการทดลองต่อไป โดยสรุปผลการจัดกลุ่มข้อมูลมีดังต่อไปนี้

วิธีที่ 1 ข้อมูลด้านความปลอดภัยอาหารที่ไม่มีการปรับระบบน้ำหนัก (Clustering with Non-Weighted Keyword: CN)

ขั้นตอนที่ 1 การหาจำนวนกลุ่มที่เหมาะสมในการแบ่งกลุ่มโดยใช้ขั้นตอนวิธี SOM ที่จำนวนกลุ่มตั้งแต่ 2-10 กลุ่ม

ขั้นตอนนี้ นำขั้นตอนวิธี SOM มาจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร ซึ่งมีข้อมูลจำนวน 4,806 ชุดเอกสาร มาจัดกลุ่มตั้งแต่ 2-10 กลุ่มเพื่อหาจำนวนกลุ่มที่เหมาะสม ได้ผลการทดลองดังต่อไปนี้

ตารางที่ 14 ผลการวัดประสิทธิภาพการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหารโดยขั้นตอนวิธี SOM ตั้งแต่ 2 - 10 กลุ่ม ในข้อมูล CN

จำนวนกลุ่ม (Cluster)	Variance	RMSSTD	RS
2	11.31373	3.36429	0.00090
3	11.28870	3.36091	0.00311
4	11.29006	3.36147	0.00299
5	11.27786	3.36000	0.00407
6	11.26916	3.35905	0.00484
7	11.26241	3.35840	0.00543
8	**11.24805**	**3.35660**	**0.00670**
9	11.25334	3.35774	0.00623
10	11.26383	3.35966	0.00531

จากตารางที่ 14 แสดงให้เห็นถึงค่าความแปรปรวน ค่า RMSSTD และค่า RS เมื่อทำการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธี SOM ระบุจำนวนกลุ่มในการจัดกลุ่มตั้งแต่ 2 - 10 กลุ่ม โดยจะสังเกตได้ว่าเมื่อจำนวนกลุ่มมีค่าเท่ากับ 8 จะมีค่าความแปรปรวน ค่า RMSSTD และค่า RS ที่ดีที่สุด คือมีค่าเท่ากับ 11.24805 3.35660 และ 0.00670 ตามลำดับ เนื่องจากค่าความแปรปรวน เมื่อค่าที่ได้ยิ่งน้อยหมายถึงผลการจัดกลุ่มที่ดี ค่า RMSSTD ที่น้อยหมายถึงผลการจัดกลุ่มที่ดี และค่า RS ที่มากจะ

จากผลลัพธ์ดังกล่าวนี้ ทำให้สามารถสรุปได้ว่าจำนวนกลุ่มเท่ากับ 8 กลุ่มนั้น เป็นผลลัพธ์ที่ดีที่สุดที่ได้จากขั้นตอนที่ 1 ซึ่งจะนำไปใช้ในขั้นตอนที่ 2 คือการเปรียบเทียบการจัดกลุ่มโดยขั้นตอนวิธีเชิงพันธุกรรม และ ขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีน ต่อไป

ขั้นตอนที่ 2 การแบ่งกลุ่มข้อมูลโดยเปรียบเทียบการจัดกลุ่มโดยขั้นตอนวิธีเชิงพันธุกรรม และ ขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน

จากทดลองในขั้นตอนที่ 1 โดยจัดกลุ่มด้วยขั้นตอนวิธี SOM เพื่อหาจำนวนกลุ่มที่เหมาะสม แล้วพบว่าจำนวนกลุ่มเท่ากับ 8 เหมาะสมที่สุด จากนั้น จึงทดลองจัดกลุ่มโดยขั้นตอนวิธีเชิงพันธุกรรมเปรียบเทียบกับขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีน ซึ่งผลของการจัดกลุ่มแสดงดังนี้

ตารางที่ 15 ผลการวัดประสิทธิภาพการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร เมื่อกำหนดจำนวนกลุ่มเท่ากับ 8 กลุ่ม ในข้อมูลชุด CN

วิธีการจัดกลุ่ม	จำนวนกลุ่ม (Cluster)	Variance	RMSSTD	RS
Genetic Algorithms	8	**10.62939**	**3.26299**	**0.06133**
K-Means	8	11.12090	3.33758	0.01793

จากตารางที่ 15 พบว่า ขั้นตอนวิธีเชิงพันธุกรรม ให้ผลของการจัดกลุ่มที่ดีกว่าขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน ในทุกวิธีการวัดประสิทธิภาพที่นำมาใช้วัดผล คือ ให้ค่าความแปรปรวนในขั้นตอนวิธีเชิงพันธุกรรม (Variance =10.62939) ให้ค่าที่น้อยกว่าขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน (Variance =11.12090) ค่า RMSSTD ในขั้นตอนวิธีเชิงพันธุกรรม (RMSSTD=3.26299) ให้ค่าที่น้อยกว่าขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน (RMSSTD=3.33758) ซึ่งหมายความว่าขั้นตอนวิธีเชิงพันธุกรรมจะให้ผลการจัดกลุ่มที่คุณลักษณะข้อมูลสมาชิกภายในกลุ่มมีความใกล้เคียงกันมากกว่าขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน ส่วนค่า RS ในขั้นตอนวิธีเชิงพันธุกรรม (RS=0.06133) ให้ค่าที่มากกว่าขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน (RS=0.01793) ซึ่งหมายความว่าขั้นตอนวิธีเชิง

จากผลการทดลองข้างต้น มีข้อสังเกตคือ RMSSTD และ RS ของแต่ละขั้นตอนวิธีมีค่าที่แตกต่างกันไม่มากนัก มีข้อสงสัยว่าค่าที่ได้มีความแตกต่างกันหรือไม่ จึงนำค่า RMSSTD และค่า RS ทำการวัดค่า t-Test ทางสถิติ (Ronald, 1995) โดยใช้การวัดค่า t-Test แบบ Two-Paired Sample Test เปรียบเทียบระหว่างขั้นตอนวิธี 3 คู่ คือ ระหว่างขั้นตอนวิธีเชิงพันธุกรรมกับขั้นตอนวิธีเค-มีน ระหว่างขั้นตอนวิธีเชิงพันธุกรรมกับขั้นตอนวิธี SOM และระหว่างขั้นตอนวิธี SOMกับขั้นตอนวิธีเค-มีน ซึ่งผลการทดลองแสดงดังตารางที่ 16

ตารางที่ 16 ผลการเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีโดยนำค่าของ RMSSTD มาทดสอบ Two Paired Sample Test วิธี t-Test ในข้อมูล CN โดยการวัดค่า RMSSTD

ขั้นตอนวิธี	GAs	SOM	K-Means
GAs	-	0.000 *	0.000 *
SOM	-	-	0.058
K-Means	-	-	-

หมายเหตุ: * หมายถึง significant ที่ระดับนัยสำคัญ 0.05

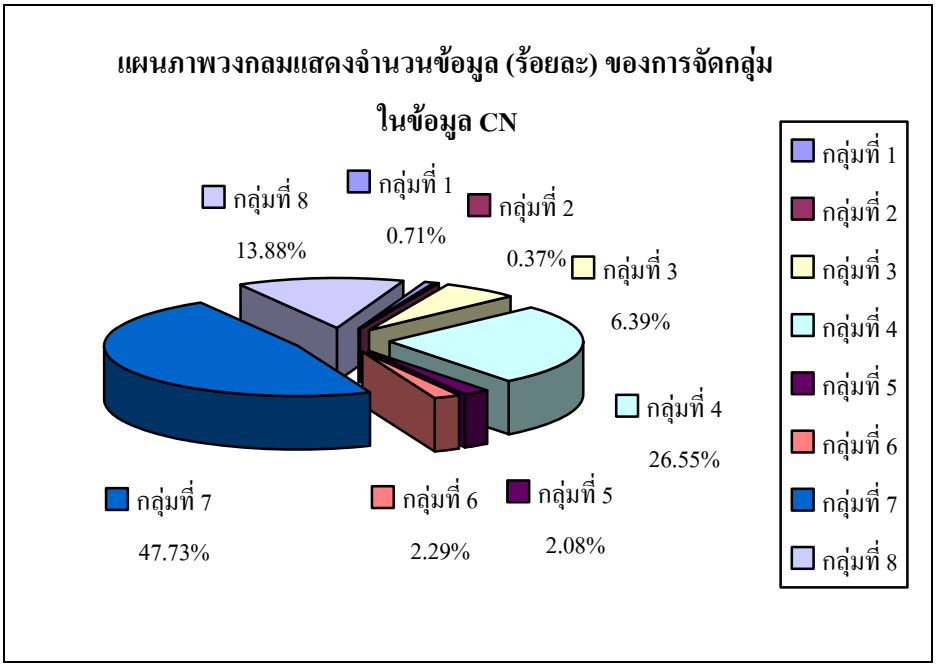
จากตารางที่ 16 จึงสรุปได้ว่า ขั้นตอนวิธีเชิงพันธุกรรมมีค่า RMSSTD ที่แตกต่างกับค่า RMSSTD จากขั้นตอนวิธี SOM และ ขั้นตอนวิธีเค-มีน ในขณะที่การเปรียบเทียบระหว่างขั้นตอนวิธี SOM กับขั้นตอนวิธีเค-มีน มีผลการทำงานที่ให้ค่าที่ไม่แตกต่างกัน

จากผลการเปรียบเทียบประสิทธิภาพของการจัดกลุ่มข้อมูลข้างต้นนั้น สามารถสรุปได้ว่าในข้อมูล CN ซึ่งใช้ข้อมูลด้านความปลอดภัยอาหาร โดยไม่ใช้ระบบน้ำหนัก เมื่อนำข้อมูลมาจัดกลุ่มแบบ 2 ขั้นตอน จะได้จำนวนกลุ่มเท่ากับ 8 กลุ่ม และใช้ขั้นตอนวิธีเชิงพันธุกรรมในการจัดกลุ่ม จะให้ผลการจัดกลุ่มที่ดีที่สุด

เมื่อนำผลที่ได้มาวิเคราะห์ผลการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร พบว่า จำนวนสมาชิกในแต่ละกลุ่มเป็นไปดังนี้

ตารางที่ 17 แสดงจำนวนสมาชิกและร้อยละของการจัดกลุ่มข้อมูลทั้ง 8 กลุ่มในข้อมูล CN

กลุ่มที่	จำนวนสมาชิก (รายการ)	จำนวนสมาชิก (ร้อยละ)
1	34	0.71
2	18	0.37
3	307	6.39
4	1,276	26.55
5	100	2.08
6	110	2.29
7	2,294	47.73
8	667	13.88
รวม	4,806	100



ภาพที่ 23 แผนภาพวงกลมแสดงจำนวนข้อมูล (ร้อยละ) ของการจัดกลุ่ม ในข้อมูล CN

ภาพที่ 23 แสดงถึงผลการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร ซึ่งในแต่ละกลุ่มนั้นมีจำนวนสมาชิกแตกต่างกันไป โดยกลุ่มที่มีสมาชิกมากที่สุดนั้น คือ กลุ่มที่ 7 ซึ่งมีจำนวนสมาชิกมาก

เมื่อพิจารณาถึงสมาชิกภายในกลุ่ม เพื่อดูลักษณะเฉพาะของแต่ละกลุ่ม โดยดูที่คำสำคัญที่ปรากฏอยู่ภายในกลุ่ม พบว่าแต่ละกลุ่มมีลักษณะดังนี้

กลุ่มที่ 1 คำสำคัญที่ปรากฏอยู่มากที่สุดในกลุ่มนี้ ได้แก่ ผลไม้ การแปรรูปผลไม้ ลำไย

กลุ่มที่ 2 คำสำคัญที่ปรากฏอยู่มากที่สุดในกลุ่มนี้ ได้แก่ ดิน จุลินทรีย์

กลุ่มที่ 3 คำสำคัญที่ปรากฏอยู่มากที่สุดในกลุ่มนี้ ได้แก่ น้ำ สัตว์น้ำ กุ้ง ปลา พันธุกรรม

กลุ่มที่ 4 คำสำคัญที่ปรากฏอยู่มากที่สุดในกลุ่มนี้ ได้แก่ ข้าว การพัฒนาคุณภาพ โรค การเจริญเติบโต

กลุ่มที่ 5 คำสำคัญที่ปรากฏอยู่มากที่สุดในกลุ่มนี้ ได้แก่ ความเป็นกรด แบคทีเรีย สารพิษ สารปนเปื้อน

กลุ่มที่ 6 คำสำคัญที่ปรากฏอยู่มากที่สุดในกลุ่มนี้ ได้แก่ การเลี้ยงสัตว์ ประมง ทะเล อาหารสัตว์

กลุ่มที่ 7 คำสำคัญที่ปรากฏอยู่มากที่สุดในกลุ่มนี้ ได้แก่ อาหาร (Food Science)

กลุ่มที่ 8 คำสำคัญที่ปรากฏอยู่มากที่สุดในกลุ่มนี้ ได้แก่ การเก็บรักษาผลผลิต คุณภาพผลผลิต

วิธีที่ 2 ข้อมูลด้านความปลอดภัยอาหารที่มีการปรับระบบน้ำหนัก (Clustering with Weighted Keyword: CW)

ขั้นตอนที่ 1 การหาจำนวนกลุ่มที่เหมาะสมในการแบ่งกลุ่มโดยใช้ขั้นตอนวิธี SOM ที่จำนวนกลุ่มตั้งแต่ 2-10 กลุ่ม

ขั้นตอนนี้ นำขั้นตอนวิธี SOM มาจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร ซึ่งมีข้อมูลจำนวน 4,806 ชุดเอกสาร มาจัดกลุ่มตั้งแต่ 2-10 กลุ่มเพื่อหาจำนวนกลุ่มที่เหมาะสม ได้ผลการทดลองดังต่อไปนี้

ตารางที่ 18 ผลการวัดประสิทธิภาพการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร โดยขั้นตอนวิธี SOM ตั้งแต่ 2 - 10 กลุ่ม ในข้อมูล CW

จำนวนกลุ่ม (Cluster)	Variance	RMSSTD	RS
2	11.28420	3.35989	0.00351
3	11.22112	3.35084	0.00908
4	10.81862	3.29054	0.04462
5	10.80729	3.28915	0.04562
6	10.80205	3.28870	0.04609
7	10.78500	3.28645	0.04759
8	**10.66008**	**3.26770**	**0.05862**
9	10.73855	3.28005	0.05169
10	11.25879	3.35891	0.00575

จากตารางที่ 18 แสดงให้เห็นถึงค่าความแปรปรวน ค่า RMSSTD และค่า RS เมื่อทำการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธี SOM ระบุจำนวนกลุ่มในการจัดกลุ่มตั้งแต่ 2 - 10 กลุ่ม โดยสังเกตได้ว่าเมื่อจำนวนกลุ่มมีค่าเท่ากับ 8 จะมีค่าความแปรปรวน ค่า RMSSTD และค่า RS ที่ดีที่สุด คือมีค่าความแปรปรวนเท่ากับ 10.66008 ค่า RMSSTD เท่ากับ 3.26770 และค่า RS เท่ากับ 0.05862 เนื่องจากค่าความแปรปรวน เมื่อค่าที่ได้น้อยหมายถึงผลการจัดกลุ่มที่ดี ในทำนองเดียวกัน ค่า RMSSTD ที่น้อยหมายถึงผลการจัดกลุ่มที่ดี และค่า RS ที่มากจะหมายถึงให้ผลการจัดกลุ่มที่ดี ดังนั้นจึงเลือกจำนวนกลุ่มที่เหมาะสมเท่ากับ 8 กลุ่ม เนื่องจากมีตัววัดประสิทธิภาพที่ให้ผลดีที่สุดทั้ง 3 ค่าการวัดผล ณ จำนวนกลุ่มเท่ากับ 8

จากผลลัพธ์ดังกล่าว ทำให้สามารถสรุปได้ว่าจำนวนกลุ่มเท่ากับ 8 กลุ่มนั้น เป็นผลลัพธ์ที่ดีที่สุดที่ได้จากขั้นตอนที่ 1 ซึ่งจะนำไปใช้ในขั้นตอนที่ 2 คือการเปรียบเทียบการจัดกลุ่มโดยขั้นตอนวิธีเชิงพันธุกรรม และ ขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีน ต่อไป

ขั้นตอนที่ 2 การแบ่งกลุ่มข้อมูลโดยเปรียบเทียบการจัดกลุ่มโดยขั้นตอนวิธีเชิงพันธุกรรม และ ขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน

จากทดลองในขั้นตอนที่ 1 โดยจัดกลุ่มด้วยขั้นตอนวิธี SOM เพื่อหาจำนวนกลุ่มที่เหมาะสม แล้วพบว่าจำนวนกลุ่มเท่ากับ 8 เหมาะสมที่สุด จากนั้น จึงทดลองจัดกลุ่มโดยขั้นตอนวิธีเชิงพันธุกรรมเปรียบเทียบกับขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีน ซึ่งผลของการจัดกลุ่มแสดงดังนี้

ตารางที่ 19 ผลการวัดประสิทธิภาพการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร เมื่อกำหนดจำนวนกลุ่มเท่ากับ 8 กลุ่ม ในข้อมูล CW โดยการวัดค่า RMSSTD

วิธีการแบ่งกลุ่ม	จำนวนกลุ่ม (Cluster)	Variance	RMSSTD	RS
Genetic Algorithms	8	**10.60892**	**3.25985**	**0.06314**
K-means	8	10.73131	3.27860	0.05233

จากตารางที่ 19 พบว่า ขั้นตอนวิธีเชิงพันธุกรรม ให้ผลของการจัดกลุ่มที่ดีกว่าขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน ในทุกวิธีการวัดประสิทธิภาพที่นำมาใช้วัดผล คือ ให้ค่าความแปรปรวนในขั้นตอนวิธีเชิงพันธุกรรม (Variance =10.60892) ให้ค่าที่น้อยกว่าขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน (Variance =10.73131) ค่า RMSSTD ในขั้นตอนวิธีเชิงพันธุกรรม (RMSSTD=3.25985) ให้ค่าที่น้อยกว่าขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน (RMSSTD=3.27860) ซึ่งหมายความว่าขั้นตอนวิธีเชิงพันธุกรรมจะให้ผลการจัดกลุ่มที่สมาชิกภายในกลุ่มอยู่ใกล้กันกว่าขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน ส่วนค่า RS ในขั้นตอนวิธีเชิงพันธุกรรม (RS=0.06314) ให้ค่าที่มากกว่าขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน (RS=0.05233) ซึ่งหมายความว่าขั้นตอนวิธีเชิงพันธุกรรมจะให้ผลการจัดกลุ่มที่แต่ละกลุ่มจะอยู่ห่างกันมากกว่าขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน

เช่นเดียวกับการทดลองในข้อมูล CN เมื่อนำผลการทดลองข้างต้นมาทำการวัดค่า t-Test แบบ Two-Paired Sample Test เปรียบเทียบระหว่างขั้นตอนวิธี 3 คู่ คือ ระหว่างขั้นตอนวิธีเชิงพันธุกรรมกับขั้นตอนวิธีเค-มีน ระหว่างขั้นตอนวิธีเชิงพันธุกรรมกับขั้นตอนวิธี SOM และระหว่างขั้นตอนวิธี SOMกับขั้นตอนวิธีเค-มีน ได้ผลการทดลองแสดงดังตารางที่ 20

ตารางที่ 20 ผลการเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีโดยนำค่าของ RMSSTD มาทดสอบ Two Paired Sample Test วิธี t-Test ในข้อมูล CW

ขั้นตอนวิธี	GAs	SOM	K-Means
GAs	-	0.736	0.170
SOM	-	-	0.345
K-Means	-	-	-

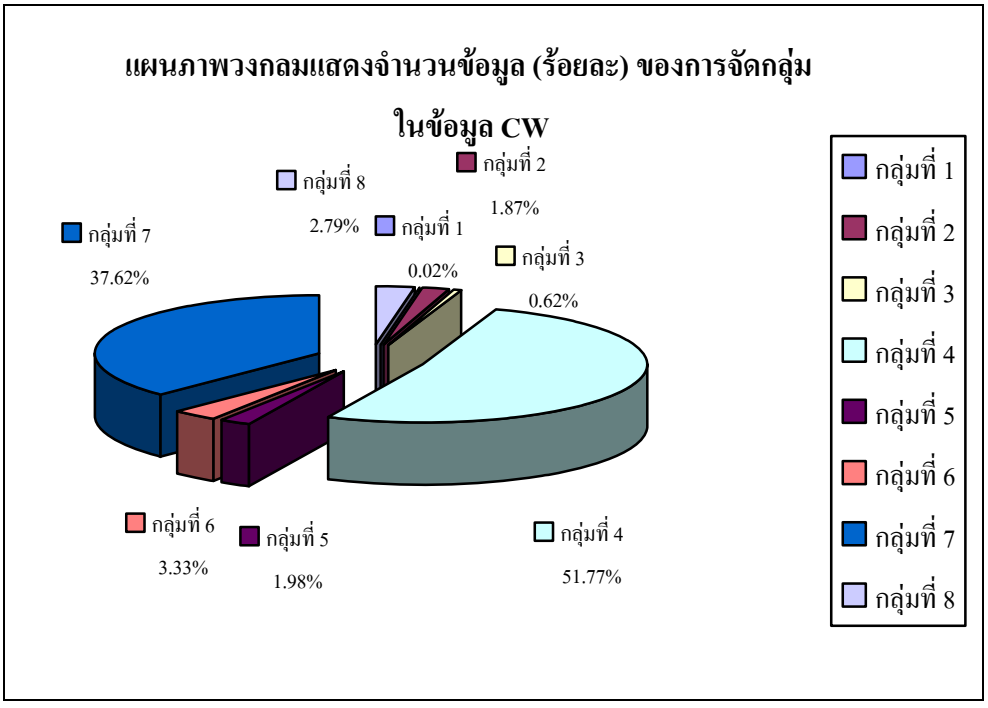
จากตารางที่ 20 สรุปได้ว่า เมื่อนำข้อมูลมาปรับค่าน้ำหนักของคำสำคัญ จะทำให้ค่า RMSSTD ของทั้ง 3 ขั้นตอนวิธี มีค่าไม่แตกต่างกันมาก เนื่องจากการปรับค่าน้ำหนักจะทำให้ผลการจัดกลุ่มของขั้นตอนวิธี SOM และขั้นตอนวิธีเค-มีน มีผลการจัดกลุ่มที่ดีขึ้นมาก ในขณะที่ผลของขั้นตอนวิธีเชิงพันธุกรรมดีขึ้นเพียงเล็กน้อย

จากผลการเปรียบเทียบประสิทธิภาพของการจัดกลุ่มข้อมูลข้างต้นนั้น สามารถสรุปได้ว่า ในข้อมูล CW ซึ่งใช้ข้อมูลด้านความปลอดภัยอาหาร โดยนำระบบน้ำหนักมาใช้ เมื่อนำข้อมูลมาจัดกลุ่มแบบ 2 ขั้นตอน จะได้จำนวนกลุ่มเท่ากับ 8 กลุ่ม และใช้ขั้นตอนวิธีเชิงพันธุกรรมในการจัดกลุ่มจะให้ผลการจัดกลุ่มที่ดีที่สุด

เมื่อนำผลที่ได้มาวิเคราะห์ผลการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร พบว่า จำนวนสมาชิกในแต่ละกลุ่มเป็นไปดังนี้

ตารางที่ 21 แสดงจำนวนสมาชิกและร้อยละของการจัดกลุ่มข้อมูลทั้ง 8 กลุ่มในข้อมูล CW

กลุ่มที่	จำนวนสมาชิก (รายการ)	จำนวนสมาชิก (ร้อยละ)
1	1	0.02
2	90	1.87
3	30	0.62
4	2,488	51.77
5	95	1.98
6	160	3.33
7	1808	37.62
8	134	2.79
รวม	4,806	100



ภาพที่ 24 แผนภาพวงกลมแสดงจำนวนข้อมูล (ร้อยละ) ของการจัดกลุ่มในข้อมูล CW

ภาพที่ 24 แสดงถึงผลการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร ซึ่งในแต่ละกลุ่มนั้นมีจำนวนสมาชิกแตกต่างกันไป โดยกลุ่มที่มีสมาชิกมากที่สุดนั้น คือ กลุ่มที่ 4 ซึ่งมีจำนวนสมาชิกมากที่สุด ร้อยละ 51.77 รองลงมาคือกลุ่มที่ 7 และ6 ตามลำดับ โดยมีจำนวนสมาชิกร้อยละ 37.62 และ 3.33 ตามลำดับ ส่วนกลุ่มที่มีจำนวนสมาชิกน้อยที่สุดคือกลุ่มที่ 1 โดยมีสมาชิกร้อยละ 0.02 ของจำนวนสมาชิกทั้งหมด

เมื่อพิจารณาถึงสมาชิกภายในกลุ่ม เพื่อดูลักษณะเฉพาะของแต่ละกลุ่ม โดยดูที่คำสำคัญที่ปรากฏอยู่ภายในกลุ่ม พบว่าแต่ละกลุ่มมีลักษณะดังนี้

กลุ่มที่ 1 คำสำคัญที่ปรากฏอยู่มากที่สุดในกลุ่มนี้ ได้แก่ แกมมา อไรซานอล ไขสบู่ น้ำมันรำข้าว

กลุ่มที่ 2 คำสำคัญที่ปรากฏอยู่มากที่สุดในกลุ่มนี้ ได้แก่ ถั่ว ถั่วลิสง พืช พันธุกรรม

กลุ่มที่ 3 คำสำคัญที่ปรากฏอยู่มากที่สุดในกลุ่มนี้ ได้แก่ ขอนแก่น พฤติกรรม การบริโภค

กลุ่มที่ 4 คำสำคัญที่ปรากฏอยู่มากที่สุดในกลุ่มนี้ ได้แก่ อาหาร (Food Science) ปลา ข้าว โรค

กลุ่มที่ 5 คำสำคัญที่ปรากฏอยู่มากที่สุดในกลุ่มนี้ ได้แก่ น้ำมัน ปาล์ม ไขมัน น้ำมันหอมระเหย

กลุ่มที่ 6 คำสำคัญที่ปรากฏอยู่มากที่สุดในกลุ่มนี้ ได้แก่ การเลี้ยงกุ้ง กุลาดำ ก้ามกราม

กลุ่มที่ 7 คำสำคัญที่ปรากฏอยู่มากที่สุดในกลุ่มนี้ ได้แก่ การเจริญเติบโต การเลี้ยง การรักษาคุณภาพ

กลุ่มที่ 8 คำสำคัญที่ปรากฏอยู่มากที่สุดในกลุ่มนี้ ได้แก่ ไก่ เนื้อไก่ ไข่

จากการทดลองการจัดกลุ่มกับข้อมูลทั้ง 2 วิธี สามารถเปรียบเทียบผลการทดลองการจัดกลุ่มแบบ 2 ขั้นตอนได้ดังนี้

ตารางที่ 22 แสดงการเปรียบเทียบผลการทดลองการจัดกลุ่มระหว่างข้อมูล CN และข้อมูล CW

การเปรียบเทียบ	CN	CW
จำนวนกลุ่มที่แบ่งได้	8	8
ค่าความแปรปรวน	10.62939	10.60892
ค่า RMSSTD	3.26299	3.25985
ค่า RS	0.06133	0.06314
ขั้นตอนวิธีที่ให้ผลการทดลองที่ดีที่สุด	GAs	GAs

จากตารางที่ 22 พบว่าข้อมูลทั้ง 2 วิธี ให้ผลการจัดกลุ่มที่ดีที่สุดที่ 8 กลุ่มเท่ากัน และขั้นตอนวิธีที่ดีที่สุดในการทดลองคือ การจัดกลุ่มโดยใช้ขั้นตอนวิธีเชิงพันธุกรรม แต่ผลการทดลองในค่าความแปรปรวน ค่า RMSSTD และ ค่า RS ในข้อมูล CW ให้ผลที่ดีกว่าข้อมูลในCN ทั้ง 3 ค่า คือ ค่าความแปรปรวนใน CW (Variance=10.60892) น้อยกว่า ใน CN (Variance=10.62939) ค่า RMSSTD ใน CW (RMSSTD=3.25985) น้อยกว่า ใน CN (RMSSTD=3.26299) และค่า RS ใน CW (RS=0.06314) มากกว่า ใน CN (RS=0.06133)

เมื่อพิจารณาเปรียบเทียบถึงผลของการจัดกลุ่มของข้อมูลที่นำระบบน้ำหนักของคำมาใช้กับข้อมูลที่ไม่ใช้ระบบน้ำหนักของแต่ละขั้นตอนวิธีพบว่า การใช้ระบบน้ำหนักจะทำให้การจัดกลุ่มของทั้ง 3 ขั้นตอนวิธี ได้ผลที่ดีขึ้น ดังตารางที่ 23

ตารางที่ 23 แสดงการเปรียบเทียบประสิทธิภาพของการจัดกลุ่มที่เพิ่มขึ้นหลังปรับระบบน้ำหนักของคำ

ขั้นตอนวิธี	Variance			RMSSTD			RS		
	CN	CW	ค่าลดลง	CN	CW	ค่าลดลง	CN	CW	ค่าเพิ่มขึ้น
GA	10.62939	10.60892	0.19 %	3.26299	3.25985	0.09 %	0.06133	0.06314	2.95 %
KM	11.12090	10.73131	3.50 %	3.33758	3.27860	1.76 %	0.01793	0.05233	191.18 %
SOM	11.24805	10.66008	5.22 %	3.35660	3.26770	2.64 %	0.00670	0.05862	774.92 %

จากตารางที่ 23 พบว่า เมื่อนำระบบปรับน้ำหนักมาใช้ในการทดลอง ขั้นตอนวิธี SOM ประสิทธิภาพเพิ่มขึ้นมากที่สุด คือค่าความแปรปรวนมีค่าลดลง 5.22 % ค่า RMSSTD มีค่าลดลง 2.64 % และค่า RS มีค่าเพิ่มขึ้น 774.92 % รองลงมาคือขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน ค่าความแปรปรวนมีค่าลดลง 3.50 % ค่า RMSSTD มีค่าลดลง 1.76 % และค่า RS มีค่าเพิ่มขึ้น 191.18 % และสุดท้ายคือขั้นตอนวิธีเชิงพันธุกรรมให้ประสิทธิภาพเพิ่มขึ้นน้อยที่สุดคือ ค่าความแปรปรวนมีค่าลดลง 0.19 % ค่า RMSSTD มีค่าลดลง 0.09 % และค่า RS มีค่าเพิ่มขึ้น 2.95 % โดยที่ค่าความแปรปรวนที่น้อยหมายถึงประสิทธิภาพที่ดีขึ้น ในทำนองเดียวกัน ค่า RMSSTD ที่น้อยลงหมายถึงประสิทธิภาพที่ดีขึ้น และค่า RS ที่มากขึ้นหมายถึงประสิทธิภาพที่ดีขึ้น

เมื่อทำการจับเวลาที่ใช้ในการทดลอง และจำนวนรอบการทำงานของทั้ง 3 ขั้นตอนวิธี แล้วนำมาคำนวณหาค่าเฉลี่ย ได้ผลดังนี้

ตารางที่ 24 แสดงการเปรียบเทียบเวลาและจำนวนรอบการทำงานที่ใช้ในแต่ละขั้นตอนวิธี

ขั้นตอนวิธี	จำนวนรอบการทำงาน	เวลาที่ใช้	เฉลี่ยเวลาที่ใช้ต่อรอบ (วินาที)	ความซับซ้อนด้านเวลา *
GAs	100	153 นาที 12 วินาที	91.93	$O(n^2)$
K-Means	23	4 นาที 17 วินาที	11.19	O(n)
SOM	66	12 นาที 26 วินาที	11.30	O(n)

หมายเหตุ * n คือ จำนวนข้อมูลทั้งหมด

ซึ่งขั้นตอนวิธีเชิงพันธุกรรมจะใช้เวลาและจำนวนรอบมากที่สุดคือ 91.93 วินาทีต่อรอบ เนื่องจากกำหนดจำนวนรอบการทำงานที่ 100 รอบเสมอ ส่วนขั้นตอนวิธีเค-มีน จะใช้เวลาและจำนวนรอบน้อยที่สุด และใช้เวลาเฉลี่ยต่อรอบน้อยที่สุดคือ 11.19 วินาทีต่อรอบ ซึ่งเป็นไปตามค่า Time Complexity ของขั้นตอนวิธี

เมื่อพิจารณาถึงเวลาและผลของจำนวนกลุ่มที่ได้เมื่อทดลองจัดกลุ่มโดยขั้นตอนวิธี SOM ร่วมกับขั้นตอนวิธีเชิงพันธุกรรม ในการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหารที่จำนวนกลุ่มตั้งแต่

ตารางที่ 25 เปรียบเทียบเวลาและผลของจำนวนกลุ่มที่ได้ในการจัดกลุ่มของการใช้ขั้นตอนวิธี SOM ร่วมกับขั้นตอนวิธีเชิงพันธุกรรมกับการใช้ขั้นตอนวิธีเชิงพันธุกรรมเพียงวิธีเดียว ที่จำนวนกลุ่มตั้งแต่ 2-10 กลุ่ม

ขั้นตอนวิธี	จำนวนกลุ่มที่ได้	เวลาที่ใช้ในการทดลอง
SOM กับ GA	8	4 ชั่วโมง 15 นาที 29 วินาที
GA	8	20 ชั่วโมง 28 นาที 4 วินาที

จากตารางที่ 25 เมื่อเปรียบเทียบจำนวนกลุ่มที่ได้จากการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร พบว่าการจัดกลุ่มโดยขั้นตอนวิธี SOM ร่วมกับขั้นตอนวิธีเชิงพันธุกรรมและการจัดกลุ่มโดยใช้ขั้นตอนวิธีเชิงพันธุกรรมเพียงวิธีเดียว ให้ผลของจำนวนกลุ่มที่เท่ากันคือ 8 กลุ่ม และเมื่อเปรียบเทียบเวลาที่ใช้ในการทำงาน พบว่า การจัดกลุ่มโดยขั้นตอนวิธี SOM ร่วมกับขั้นตอนวิธีเชิงพันธุกรรมจะใช้เวลาในการจัดกลุ่มเท่ากับ 4 ชั่วโมง 15 นาที 29 วินาที ซึ่งน้อยกว่าการจัดกลุ่มโดยใช้ขั้นตอนวิธีเชิงพันธุกรรมเพียงวิธีเดียวที่ใช้เวลาเท่ากับ 20 ชั่วโมง 28 นาที 4 วินาที

สุดท้ายเมื่อเปรียบเทียบผลการทดลองกับงานวิจัยก่อนหน้า โดยทำการทดลองเปรียบเทียบกับการจัดกลุ่มด้านความปลอดภัยอาหารที่จำนวนกลุ่มเท่ากับ 8 และวัดประสิทธิภาพโดยใช้ค่าความแปรปรวน ค่า RMSSTD และค่า RS ได้ผลดังตารางที่ 26

ตารางที่ 26 เปรียบเทียบผลการจัดกลุ่มกับงานวิจัยก่อนหน้าที่จำนวนกลุ่มเท่ากับ 8 ในการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร

ผู้ทดลอง	ขั้นตอนวิธี	Variance		RMSSTD		RS	
		CN	CW	CN	CW	CN	CW
Wongkot (2008)	SOM กับ GA	10.62940	10.60892	3.26299	3.25985	0.06133	0.06314
Chaimeun (2005)	SOM กับ K-Means	11.12090	10.73131	3.33758	3.27860	0.01793	0.05233
Maulik (2000)	GA	10.83193	10.63944	3.29393	3.26453	0.04345	0.06045

จากตารางที่ 26 เมื่อเปรียบเทียบผลการทดลองนี้ กับการใช้ขั้นตอนวิธีของงานวิจัยก่อนหน้าโดยใช้ข้อมูลชุดเดียวกัน พบว่าขั้นตอนวิธีที่นำเสนอคือการใช้ SOM ร่วมกับขั้นตอนวิธีเชิงพันธุกรรมที่มีการคำนวณค่าความเหมาะสมใหม่ ให้ผลการจัดกลุ่มที่ดีที่สุด โดยให้ค่าความแปรปรวนในข้อมูล CN เท่ากับ 10.62940 ในข้อมูล CW เท่ากับ 10.60892 ค่า RMSSTD ในข้อมูล CN เท่ากับ 3.26299 ในข้อมูล CW เท่ากับ 3.25985 และค่า RS ในข้อมูล CN เท่ากับ 0.06133 ในข้อมูล CW เท่ากับ 0.06314

วิจารณ์

งานวิจัยที่นำเสนอนี้เป็นการนำเทคนิคการจัดกลุ่มแบบ 2 ขั้นตอนมาปรับปรุงประสิทธิภาพในการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหาร โดยใช้คำสำคัญเป็นตัวแทนของเอกสาร และหาลักษณะเฉพาะข้อมูลในแต่ละกลุ่มโดยใช้ชุดข้อมูลแบบที่มีการใช้ระบบน้ำหนักของคำและไม่ใช้ระบบน้ำหนักของคำ

สำหรับงานวิจัยนี้ขั้นตอนแรก ได้นำข้อมูลด้านความปลอดภัยอาหารมาทำการจัดกลุ่มโดยใช้ขั้นตอนวิธี SOM เพื่อหาจำนวนกลุ่มที่เหมาะสม พบว่า ในชุดข้อมูลที่ไม่มีการนำระบบน้ำหนักของคำมาใช้ จากตารางที่ 14 จะได้จำนวนกลุ่มที่เหมาะสมที่สุดเท่ากับ 8 เช่นเดียวกับในชุดข้อมูลที่

ในขั้นตอนต่อมาเป็นการเปรียบเทียบการจัดกลุ่มโดยใช้ขั้นตอนวิธีเชิงพันธุกรรมเปรียบเทียบกับขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน เพื่อหาขั้นตอนวิธีที่เหมาะสมที่ให้ผลของการจัดกลุ่มที่ดีที่สุด เพื่อนำมาใช้ในการวิเคราะห์การจัดกลุ่มต่อไป พบว่า ในชุดข้อมูลที่ไม่มีการนำระบบน้ำหนักของคำมาใช้ จากตารางที่ 15 ผลการจัดกลุ่มโดยขั้นตอนวิธีเชิงพันธุกรรมให้ผลการจัดกลุ่มที่ดีกว่าการจัดกลุ่มโดยใช้ขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน และในชุดข้อมูลที่นำระบบน้ำหนักของคำมาใช้ จากตารางที่ 19 พบว่า ผลการจัดกลุ่มโดยขั้นตอนวิธีเชิงพันธุกรรมให้ผลการจัดกลุ่มที่ดีกว่าการจัดกลุ่มโดยใช้ขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน เช่นเดียวกัน

เมื่อเปรียบเทียบถึงผลของการจัดกลุ่มของข้อมูลที่นำระบบน้ำหนักของคำมาใช้กับข้อมูลที่ไม่ใช้ระบบน้ำหนักจากตารางที่ 23 พบว่า ผลของการจัดกลุ่มข้อมูลที่นำระบบน้ำหนักของคำมาใช้ให้ผลที่ดีกว่าชุดข้อมูลที่ไม่มีการนำระบบน้ำหนักของคำมาใช้ทั้ง 3 ขั้นตอนวิธี โดยขั้นตอนวิธี SOM ประสิทธิภาพเพิ่มขึ้นมากที่สุด รองลงมาคือขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน และสุดท้ายคือขั้นตอนวิธีเชิงพันธุกรรมให้ประสิทธิภาพเพิ่มขึ้นน้อยที่สุด ซึ่งส่งผลให้การเปรียบเทียบทั้ง 3 ขั้นตอนวิธี ในตารางที่ 20 พบว่าทั้ง 3 ขั้นตอนวิธี ไม่มีความแตกต่างกันมากเมื่อนำข้อมูลมาปรับน้ำหนักของคำ

เมื่อพิจารณาถึงเวลาที่ใช้และจำนวนรอบการทำงานของทั้ง 3 ขั้นตอนวิธี คือ ขั้นตอนวิธีการจัดกลุ่มแบบ SOM ขั้นตอนวิธีเชิงพันธุกรรม และขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน พบว่าขั้นตอนวิธีการจัดกลุ่มแบบเค-มีนใช้จำนวนรอบและเวลาน้อยที่สุด โดยที่ขั้นตอนวิธีเชิงพันธุกรรมใช้เวลามากที่สุดแต่มีการกำหนดจำนวนรอบการทำงานที่ 100 รอบเสมอ และขั้นตอนวิธีการจัดกลุ่มแบบเค-มีนมีเวลาเฉลี่ยต่อรอบการทำงานน้อยที่สุด ส่วนขั้นตอนวิธีเชิงพันธุกรรมมีเวลาเฉลี่ยต่อรอบมากที่สุด

จากการเปรียบเทียบเรื่องเวลาและผลของจำนวนกลุ่มที่ได้เมื่อทดลองจัดกลุ่มโดยขั้นตอนวิธี SOM ร่วมกับขั้นตอนวิธีเชิงพันธุกรรม ในการจัดกลุ่มข้อมูลด้านความปลอดภัยอาหารที่จำนวนกลุ่มตั้งแต่ 2-10 กลุ่ม เปรียบเทียบกับการใช้ขั้นตอนวิธีเชิงพันธุกรรมเพียงวิธีเดียว ดังตารางที่ 25 พบว่าการจัดกลุ่มโดยขั้นตอนวิธี SOM ร่วมกับขั้นตอนวิธีเชิงพันธุกรรมและการจัดกลุ่มโดยใช้

สุดท้ายเมื่อเปรียบเทียบผลการทดลองจัดกลุ่มกับงานวิจัยก่อนหน้า พบว่าการใช้ขั้นตอนวิธี SOM ร่วมกับขั้นตอนวิธีเชิงพันธุกรรมที่มีการคำนวณค่าความเหมาะสมโดยใช้ค่า RMSSTD และค่า RS ให้ผลการทดลองที่ดีกว่าการใช้ขั้นตอนวิธีเชิงพันธุกรรมที่ใช้ผลรวมของระยะห่างของข้อมูลกับจุดศูนย์กลางกลุ่ม และขั้นตอนวิธี SOM ร่วมกับขั้นตอนวิธีเค-มีน

สรุปและข้อเสนอแนะ

สรุป

การจัดกลุ่มข้อมูลด้านความปลอดภัยอาหารโดยใช้คำสำคัญเป็นตัวแทนของเอกสารด้วยวิธีการแบบ 2 ขั้นตอน คือใช้ขั้นตอนวิธี SOM เพื่อหาจำนวนกลุ่มมาใช้ในขั้นตอนต่อไปคือเปรียบเทียบขั้นตอนวิธีเชิงพันธุกรรมกับขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน และนำคำสำคัญมาใช้เป็นตัวแทนของเอกสาร ทดลองเปรียบเทียบกับการนำระบบน้ำหนักของคำมาใช้ในการปรับน้ำหนักของคำสำคัญ พบว่าการจัดกลุ่มข้อมูลด้วยการใช้ระบบน้ำหนักและไม่ใช้ระบบน้ำหนักให้ผลการจัดกลุ่มที่ 8 กลุ่ม และขั้นตอนวิธีที่ดีที่สุดคือขั้นตอนวิธีเชิงพันธุกรรม

เมื่อเปรียบเทียบผลการจัดกลุ่มระหว่างชุดข้อมูลที่มีการใช้ระบบน้ำหนักและไม่ใช้ระบบน้ำหนักพบว่า ชุดข้อมูลที่มีการปรับระบบน้ำหนักให้ผลการจัดกลุ่มที่ดีกว่าชุดข้อมูลที่ไม่มีการปรับระบบน้ำหนักในการทดสอบกับวิธีการแบบ 2 ขั้นตอน คือ SOM กับเค-มีน และ SOM กับ ขั้นตอนวิธีเชิงพันธุกรรม

เมื่อเปรียบเทียบผลการทดลองในทางสถิติโดยการทดสอบ Two Paired Sample Test วิธี t-Test โดยใช้ค่าของ RMSSTD เพื่อทดสอบความแตกต่างของแต่ละขั้นตอนวิธี พบว่า ชุดข้อมูลที่ไม่มีการใช้ระบบน้ำหนัก ขั้นตอนวิธีเชิงพันธุกรรม มีผลการทดสอบที่แตกต่างกับขั้นตอนวิธี SOM และขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน อย่างมีนัยสำคัญ สำหรับชุดข้อมูลที่มีการใช้ระบบน้ำหนัก พบว่า ทั้ง 3 ขั้นตอนวิธี มีผลการทดสอบที่ไม่แตกต่างกัน เนื่องจาก ชุดข้อมูลที่มีการปรับระบบน้ำหนัก จะทำให้ผลการจัดกลุ่มของขั้นตอนวิธี SOM และขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน มีประสิทธิภาพที่ดีขึ้นมากกว่าประสิทธิภาพที่ดีขึ้นของขั้นตอนวิธีเชิงพันธุกรรม

การเปรียบเทียบเวลาที่ใช้ในการทดลองและจำนวนรอบการทำงานของการจัดกลุ่มพบว่าขั้นตอนวิธีการจัดกลุ่มแบบเค-มีนมีเวลาเฉลี่ยต่อรอบการทำงานน้อยที่สุด ส่วนขั้นตอนวิธีเชิงพันธุกรรมมีเวลาเฉลี่ยต่อรอบมากที่สุด

การจัดกลุ่มข้อมูลด้านความปลอดภัยของอาหารโดยวิธีการแบบ 2 ขั้นตอนคือ ขั้นตอนวิธี SOM ร่วมกับขั้นตอนวิธีเชิงพันธุกรรม ใช้เวลาในการจัดกลุ่มลดลงเมื่อเทียบกับการใช้ขั้นตอนวิธีเชิงพันธุกรรมเพียงวิธีเดียว

ข้อเสนอแนะ

ในงานวิจัยนี้ มุ่งเน้นในการปรับปรุงประสิทธิภาพการจัดกลุ่มเอกสาร ซึ่งประเมินผลโดยวัดความใกล้เคียงกันของเอกสารที่ได้รับการจัดกลุ่ม ซึ่งไม่ได้มุ่งเน้นที่เวลาที่ใช้ในการทำงาน ผลการทดลองพบว่าขั้นตอนวิธีเชิงพันธุกรรมเป็นวิธีที่มีประสิทธิภาพดีที่สุด แต่ใช้เวลาที่นานมากกว่าขั้นตอนวิธีอื่นๆ แนวทางในการศึกษาต่อไปคือ ปรับปรุงเรื่องเวลาการทำงานในขั้นตอนวิธีเชิงพันธุกรรม เช่นการใช้ขั้นตอนวิธีเชิงพันธุกรรมแบบบีบอัดแทนขั้นตอนวิธีเชิงพันธุกรรมแบบเดิมเป็นต้น

ข้อสังเกตที่ได้จากการทดลองคือเมื่อพิจารณาถึงลักษณะเฉพาะของการจัดกลุ่มของชุดข้อมูลทั้ง 2 ชุด โดยพิจารณาคำสำคัญที่ปรากฏอยู่ภายในกลุ่ม พบว่า ชุดข้อมูลที่ไม่มีการนำระบบน้ำหนักของคำมาใช้ ลักษณะของกลุ่มที่ได้จะเป็นลักษณะเชิงกว้าง เช่น การเลี้ยงสัตว์ ผลไม้ คุณภาพผลผลิต เป็นต้น ส่วนชุดข้อมูลที่นำระบบน้ำหนักของคำมาใช้จะทำให้ผลการจัดกลุ่มมีลักษณะในเชิงลึก หรือกลุ่มเฉพาะ เช่นกลุ่มของ ไก่ กุ้ง น้ำมัน เป็นต้น ซึ่งทำให้เราสามารถเลือกได้ว่าจะนำระบบน้ำหนักของคำมาใช้หรือไม่ ขึ้นอยู่กับความต้องการของผลในการจัดกลุ่ม ถ้าต้องการผลการจัดกลุ่มในเชิงลึกควรนำระบบน้ำหนักของคำมาใช้ และถ้าต้องการผลการจัดกลุ่มในลักษณะเชิงกว้าง ไม่จำเป็นต้องใช้ระบบน้ำหนักของคำ จึงเป็นจุดที่น่าสนใจในการวิจัยถึงความสัมพันธ์ของเอกสารภายในกลุ่มในกรณีที่ใช้ระบบน้ำหนักกับไม่ใช้ระบบน้ำหนักในการปรับปรุงน้ำหนักของคำสำคัญของเอกสารสำหรับการจัดกลุ่มต่อไป

เอกสารและสิ่งอ้างอิง

เนคเทค. 2551. **SWATH: โปรแกรมตัดคำและกำกับหน้าที่คำภาษาไทยอัตโนมัติ.** ฝ่ายวิจัยและพัฒนาสารสนเทศ. แหล่งที่มา: http://www.links.nectec.or.th/download_thai.php, 12 กุมภาพันธ์ 2551.

มกอช. สำนักงานมาตรฐานเกษตรและอาหารแห่งชาติ. 2550. **โครงการระบบสารสนเทศวิชาการด้านความปลอดภัยอาหาร.** แหล่งที่มา: http://fsci.acfs.go.th, 23 ธันวาคม 2550.

Amarasiri, R., D. Alahakoon and K.A. Smith. 2004. HDGSOM: a modified growing self-organizing map for high dimensional data clustering, pp. 216- 221. *In* **Proceedings of the Fourth International Conference on Hybrid Intelligent Systems**. Kitakyushu, Japan.

Baker, J. 1987. Reducing Bias and Inefficiency in the Selection Algorithm, pp. 14-21. *In* **the Second International Conference on Genetic Algorithms and their Application**. Hillsdale, New Jersey, USA.

Berson, A., S. Smith and K. Thearling. 1999. **Building Data Mining Applications for CRM.** McGraw-Hill, USA.

Bradley, P. S and U. M. Fayyad. 1998. Refining Initial Points for K-Means Clustering, pp. 91-99. *In* **the Fifteenth International Conference on Machine Learning (ICML98)**. San Francisco, USA.

Cassidy, S. 2002. **COMP449: Speech Recognition.** Available Source: http://www.ics.mq.edu.au/~cassidy/comp449/html/comp449.html, February 23, 2008.

Chaimeun, O. and A. Srivihok. 2005. Clustering of Thai handcraft Customers by using Combined SOM and K-means Algorithm, *In* **The 8th IASTED International Conference on Database and Application Innsbruck**. Innsbruck, Austria.

CRISP. **CRoss Industry Standard Process for Data Mining.** CRISP-DM. Available Source: http://www.crisp-dm.org/, February 23, 2008.

Elliott, L., D.B. Ingham, A.G. Kyne, N.S. Mera, M. Pourkashanian and C.W. Wilson. 2002. A real coded genetic algorithm for the optimization of reaction rate parameters or chemical kinetic modelling in a perfectly stirred reactor, pp. 138-144. *In* **Genetic and Evolutionary Computation Conference (GECCO)**. New York, USA.

Frawley, W. J., G. Piatetsky-Shapiro and C. J. Matheus. 1992. Knowledge discovery in databases: An overview. **AI Magazine.** 13 (3): 57-70.

Goldberg, D.E. 1989. **Genetic Algorithms in Search, Optimization and Machine Learning.** Addison-Wesley Professional, New York, USA.

_____. 2002. **The Design of Innovation.** Kluwer Academic Publishers, USA.

Grossman, D. A. and O. Frieder. 2006. *In* **Information Retrieval: Algorithms and Heuristics.** 2ed. Springer, Germany.

Jain, A. K., M. N. Murty and P. J. Flynn. 1999. Data Clustering: A Review. **ACM Computing Surveys.** 31 (3): 264-323.

Kohonen, T. 2001. **Self-Organizing Maps.** 3 Extended ed. Springer, Berlin, Heidelberg.

Kovács, F., C. Legány and A. Babos. 2005. Cluster validity measurement techniques, *In* **International Symposium of Hungarian Researches on Computational Intelligenc Hungarian Researches on Computational Intelligenc**. Hungary.

Lu, Y., S. Lu, Y. Deng and S. Brown. 2004. Incremental Genetic K-means Algorithm and its Application in Gene Expression Data Analysis. **BMC bioinformatics.** 5 (1): 172.

MacQueen, J. B. 1967. Some Methods for classification and Analysis of Multivariate Observations, pp. 281-297. *In* **5th Berkeley Symposium on Mathematical Statistics and Probability**. Berkeley, University of California, USA.

Maulik, U. and S. Bandyopadhyay. 2000. Genetic algorithm-based clustering technique. **Pattern Recognition.** 33 (9): 1455-1465.

Painho, M. and F. Bacao. 2000. Using genetic Algorithms in Clustering Problems, *In* **the 5th International Conference on GeoComputation University of Greenwich**. United Kingdom.

Phattaraworamet, T. and P. Rattanapongporn. 2006. Genetic Algorithm with K-Means, *In* **the third annual international conference organized by Electrical Engineering/Electronics, Computer, Telecomunications and Information Technology Association (ECTI-CON 2006)**. Ubon Ratchathani, Thailand.

Rebelsky, S. A. 1997. **Genetic Algorithms and Simulated Annealing**. Available Source: http://www.cs.grinnell.edu/~rebelsky/Courses/CS223/97F/Outlines/outline.38.html, February 23, 2008.

Ronald, E. 1995. **Probability and Statistics for Engineers and Scientists.** Walpole, Macmillan, United Kingdom.

Salton, G. and M. J. McGill. 1983. **Introduction to modern information retrieval.** McGraw-Hill, USA.

Sinka, M.P. and D.W. Corne. 2002. A large benchmark dataset for web document clustering. **Soft Computing Systems: Design, Management and Applications.** 87: 881-890.

Tan, P.N., M. Steinbach and V. Kumar. 2006. **Introduction to Data Mining.** Pearson Education, Inc., USA.

Weng, S.S. and M.J. Liu. 2004. Feature-based recommendations for one-to-one marketing. **Expert Systems with Applications.** 26 (4): 493-508.

Wikipedia. **Cluster Analysis**. Available Source: http://en.wikipedia.org/wiki/Data_clustering, February 23, 2008.

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	นายวงกต พจน์พงศ์สรรค์
วัน เดือน ปี ที่เกิด	วันที่ 26 เมษายน 2523
สถานที่เกิด	กรุงเทพมหานคร
ประวัติการศึกษา	วท.บ. (วิทยาการคอมพิวเตอร์) เกียรตินิยมอันดับ 2 มหาวิทยาลัยเทคโนโลยีราชมงคล ธัญบุรี (พ.ศ. 2545)
ตำแหน่งหน้าที่การงานปัจจุบัน	-
สถานที่ทำงานปัจจุบัน	-
ผลงานดีเด่นและรางวัลทางวิชาการ	1. นำเสนอ และตีพิมพ์บทความวิชาการเรื่อง “Unsupervised Data Clustering using Hybrid Genetic Algorithms” ในงานประชุมวิชาการ The 5th International Conference on e-Business 2006 (NCEB 2006) กรุงเทพฯ วันที่ 2 - 3 พฤศจิกายน 2549 2. นำเสนอ และตีพิมพ์บทความวิชาการเรื่อง “การจัดกลุ่มแบบ 2 ขั้นตอน โดย Self-Organizing Maps และ ขั้นตอนวิธีเชิงพันธุกรรมสำหรับเอกสารด้านความปลอดภัยของอาหาร” ในงานประชุมวิชาการ The 11th National Computer Science and Engineering Conference (NCSEC 2007) กรุงเทพฯ วันที่ 19 - 21 พฤศจิกายน 2550
ทุนการศึกษาที่ได้รับ	ทุนงานวิจัยจากโครงการปริญญาโท