ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

ปริญญา

วิทยาการคอมพิวเตอร์	วิทยาการคอมพิวเตอร์
สาขา	**ภาควิชา**

เรื่อง การแบ่งกลุ่มผู้เชี่ยวชาญในมหาวิทยาลัยของรัฐบาล โดยใช้เทคนิคการจัดกลุ่มแบบอัลกอริทึม 2 ขั้นตอนและการหาน้ำหนักของคุณลักษณะ

Enterprise Expert Clustering of Public University by Two Stage Algorithms and Feature Weighting Techniques

นามผู้วิจัย นางสาวจารุวรรณ กาญจนศุภวรรณ

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก .....

(รองศาสตราจารย์อนงค์นาฎ ศรีวิหค, Ph.D.)

หัวหน้าภาควิชา .....

(ผู้ช่วยศาสตราจารย์ศิริกร จันทร์นวล, M.S.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่..... **เดือน**..... **พ.ศ.**.....

วิทยานิพนธ์

เรื่อง

การแบ่งกลุ่มผู้เชี่ยวชาญในมหาวิทยาลัยของรัฐบาล โดยใช้เทคนิคการจัดกลุ่มแบบอัลกอริทึม 2 ขั้นตอนและการหาน้ำหนักของคุณลักษณะ

Enterprise Expert Clustering of Public University by Two Stage Algorithms and Feature Weighting Techniques

โดย

นางสาวจารุวรรณ กาญจนศุภวรรณ

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

เพื่อความสมบูรณ์แห่งปริญญาวิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

พ.ศ. 2554

จารุวรรณ กาญจนศุภวรรณ 2554: การแบ่งกลุ่มผู้เชี่ยวชาญในมหาวิทยาลัยของรัฐบาล โดยใช้เทคนิคการจัดกลุ่มแบบอัลกอริทึม 2 ขั้นตอนและการหาน้ำหนักของคุณลักษณะ ปริญญาวิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์) สาขาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: รองศาสตราจารย์อนงค์นาฏ ศรีวิหค, Ph.D. 104 หน้า

มหาวิทยาลัยเป็นองค์กรการศึกษาที่มีขนาดใหญ่ที่ประกอบด้วยผู้เชี่ยวชาญมากมายหลายสาขา ข้อมูลผู้เชี่ยวชาญมักถูกเก็บไว้โดยไม่ได้นำมาใช้ประโยชน์อย่างเต็มที่ การจะค้นหาผู้เชี่ยวชาญเป็นเรื่องยาก เนื่องจากในองค์กรขนาดใหญ่มีผู้เชี่ยวชาญจำนวนมาก ดังนั้นการศึกษานี้จึงนำเทคนิคการทำเหมืองข้อมูลมาช่วยในการจัดการข้อมูลผู้เชี่ยวชาญที่มีอยู่ (ซึ่งผู้เชี่ยวชาญต่างสาขากันอาจมีความเชี่ยวชาญที่เหมือนกันหรือคาบเกี่ยวกัน) โดยพิจารณาจากคำสำคัญซึ่งได้มาจากการกำหนดไว้แล้วจากผู้เชี่ยวชาญ ข้อมูลในเบื้องต้นได้จาก (1) ข้อมูลบุคลากร (2) ข้อมูลผลงานวิจัยและ (3) ข้อมูลวิทยานิพนธ์ เมื่อรวมข้อมูลดังกล่าวพบว่ามีจำนวนคุณลักษณะขนาดใหญ่ (จำนวน 971 คำสำคัญ) จึงจำเป็นต้องนำเทคนิคการคัดเลือกคุณลักษณะมาช่วยลดขนาดคุณลักษณะลง ซึ่งเทคนิคที่นำมาใช้ คือ การผสมผสานระหว่างวิธีการ CFS (Correlation-based Feature Subset Selection) และ วิธีการเชิงพันธุกรรม (Genetic Search) เทคนิคการจัดกลุ่มที่ใช้ คือ เทคนิคการจัดกลุ่มแบบอัลกอริทึม 2 ขั้นตอน ในขั้นตอนแรกหาจำนวนกลุ่มที่เหมาะสมโดยใช้อัลกอริทึม SOM (Self-Organizing Maps) และในขั้นตอนที่สองเป็นการจัดกลุ่มโดยใช้อัลกอริทึม K-Means และอัลกอริทึม Fuzzy C-Means ร่วมกับการให้น้ำหนักคุณลักษณะด้วยวิธีการ TF-IDF (Term Frequency-Inverse Document Frequency) วิธีการ Logarithm Weight และ วิธีการ Augmented Weight

จากผลการศึกษาพบว่าการลดขนาดคุณลักษณะให้ผลการจัดกลุ่มที่ดีขึ้น (เดิม 971 คุณลักษณะเหลือ 258 คุณลักษณะ) สำหรับตัววัดประสิทธิภาพการจัดกลุ่มที่นำมาใช้ ได้แก่ F-Statistic, R-Squared และ Silhouetteจำนวนกลุ่มที่ดีที่สุด คือ 7 กลุ่มและจัดกลุ่มโดยใช้อัลกอริทึม K-Means ร่วมกับวิธีการ Logarithm Weight ให้ผลดีที่สุด เมื่อพิจารณาผลการจัดกลุ่มและคำสำคัญที่มีความถี่สูงสุดต่อพบว่ากลุ่มที่ (1) ผู้เชี่ยวชาญ 50 คน คำสำคัญ “เศรษฐศาสตร์” (2) ผู้เชี่ยวชาญ 249 คน คำสำคัญ “ป่า” (3) ผู้เชี่ยวชาญ 63 คน คำสำคัญ “Rice” (4) ผู้เชี่ยวชาญ 11 คน คำสำคัญ “Corn” (5) ผู้เชี่ยวชาญ 43 คน คำสำคัญ “สอน” (6) ผู้เชี่ยวชาญ 2,671 คน เป็นกลุ่มใหญ่ มีคำสำคัญที่ปะปนกันสูง (7) ผู้เชี่ยวชาญ 107 คน คำสำคัญ “ไฟฟ้า” สำหรับในกลุ่มที่ 6 มีผู้เชี่ยวชาญปะปนมาก จึงนำกลุ่มนี้มาจัดกลุ่มต่อโดยทดลองกับอัลกอริทึม K-Means พบว่าผลการจัดกลุ่มรอบนี้ให้แนวโน้มที่ไม่ดีนัก จึงหยุดการจัดกลุ่มไว้เท่านี้ จากการศึกษานี้จะเป็นแนวทางในการต่อยอดพัฒนาวิธีจัดกลุ่มอื่นที่มีประสิทธิภาพมากขึ้นและสามารถนำมาพัฒนาเป็นระบบสารสนเทศในมหาวิทยาลัยได้

_____ _____ ___ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Jaruwan Kanjanasupawan 2011: Enterprise Expert Clustering of Public University by Two Stage Algorithms and Feature Weighting Techniques. Master of Science (Computer Science), Major Field: Computer Science, Department of Computer Science. Thesis Advisor: Associate Professor Anongnart Srivihok, Ph.D. 104 pages.

In a large university there are many experts in several domains such as agriculture, economics, science, and engineering. Therefore finding who is an expert in which domain is time consuming. The experts in different domains may have the same domain of expertise. This research proposed data mining techniques to segment expert's data by using keywords. Datasets were collected from 3 databases included (1) personal database (2) research projects and (3) advisors and thesis. Since, dataset included too many attributes (971 attributes), so hybrid attribute selection techniques using CFS (Correlation-based Feature subset Selection) and Genetic Search were conducted.

Clustering techniques used were two stage algorithms to determine the appropriate number of clusters by using SOM (Self-Organizing Maps) and to cluster data by using K-Means and Fuzzy C-Means. Clustering techniques were joined feature weighting techniques including (1) TF-IDF (Term Frequency-Inverse Document Frequency), (2) Logarithm Weight (Logarithm Term Frequency-Inverse Document Frequency) and (3) Augmented Weight.

Evaluators were F-Statistic, R-Squared and Silhouette. Results showed that the performance of dataset with attribute selection (from 971 to 258 attributes) were better than original dataset (non attribute selection). The appropriate number of clusters were 7 clusters by using SOM. The most efficient clustering was K-Means joined with Logarithm Weighting techniques. Next, expert extraction in each cluster was: (1) 50 experts, keyword: economic (2) 249 experts, keyword: agriculture (3) 63 experts, keyword: agriculture (4) 11 experts, keyword: agriculture (5) 43 experts, keyword: education (6) 2,671 experts and (7) 107 experts, keyword: engineer. Therefore, cluster 6 was too large. So this cluster was re-clustered. However, results showed that the re-clustered should be stopped in this process. The results of this study should bring about the efficient experiment clustering methods. In other ways, practitioners can apply this knowledge to develop information systems in the university such as expert searching system in the future.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

สำหรับความสำเร็จของวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอกราบขอบพระคุณรองศาสตราจารย์ ดร.อนงค์นาฎ ศรีวิหค อาจารย์ที่ปรึกษาที่ได้กรุณาสนับสนุน แนะนำแนวทางในการแก้ไขปัญหาให้ข้าพเจ้าโดยตลอดมา นอกจากนี้ข้าพเจ้าขอกราบขอบพระคุณผู้ช่วยศาสตราจารย์ ดร. วรเศรษฐ สุวรรณิก ประธานการสอบและรองศาสตราจารย์ ดร. ประสงค์ ปราณีตพลกรัง ผู้ทรงคุณวุฒิจากบัณฑิต ที่ได้กรุณาให้คำแนะนำเพิ่มเติมเพื่อให้วิทยานิพนธ์ฉบับนี้สมบูรณ์มากยิ่งขึ้น และขอกราบขอบพระคุณคณาจารย์ทุกท่านที่ได้แนะนำ สั่งสอนวิชาความรู้ให้แก่ข้าพเจ้าเพื่อเป็นแนวทางในการทำงานวิจัยชิ้นนี้อย่างสูงยิ่ง

ขอกราบขอบพระคุณบิดา มารดา พี่และน้องของข้าพเจ้าที่ได้ให้กำลังใจและสนับสนุนให้ข้าพเจ้าได้ศึกษามาจนกระทั่งดำเนินการวิทยานิพนธ์ฉบับนี้ได้สำเร็จสมบูรณ์

ขอขอบพระคุณรุ่นพี่ โดยเฉพาะพี่วงกต พจน์พงศ์สรรค์และพี่ศศิธร มงคลศรีพัฒนา ที่ได้กรุณาแนะนำและช่วยเหลือการเตรียมข้อมูล การใช้งานโปรแกรม MATLAB รวมทั้งรุ่นพี่คนอื่น ๆ ด้วย ขอขอบคุณเพื่อน ๆ และรุ่นน้องที่ให้คำแนะนำ ช่วยเหลือและแลกเปลี่ยนความรู้เพื่อเป็นแนวทางในการดำเนินงานวิจัยชิ้นนี้ได้สำเร็จ

ท้ายสุดนี้ขอขอบพระคุณ โครงการปริญญาโท ภาควิชาวิทยาการคอมพิวเตอร์ที่ได้สนับสนุนให้ข้าพเจ้าได้ศึกษาจนสำเร็จ ประโยชน์จากวิทยานิพนธ์ฉบับนี้ขอน้อมมอบให้แก่ผู้มีพระคุณทุกท่านด้วยความเคารพอย่างสูงยิ่ง

จารุวรรณ กาญจนศุภวรรณ

เมษายน 2554

สารบัญ

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญภาพ

คำอธิบายสัญลักษณ์และคำย่อ

SOM	=	Kohonen's Self-Organizing Maps
TF-IDF	=	Term Frequency-Inverse Document Frequency
CFS	=	Correlation-based Feature subset Selection
GA	=	Genetic Algorithm Search Method
MSTr	=	Mean Square for Treatments
MSE	=	Mean Square Error
RS	=	R-Squared

การแบ่งกลุ่มผู้เชี่ยวชาญในมหาวิทยาลัยของรัฐบาล โดยใช้เทคนิคการจัดกลุ่มแบบอัลกอริทึม 2 ขั้นตอนและการหาน้ำหนักของคุณลักษณะ

Enterprise Expert Clustering of Public University by Two Stage Algorithms and Feature Weighting Techniques

คำนำ

ในยุคแห่งการพัฒนาด้านไอที องค์กรต่าง ๆ รวมทั้งองค์กรทางการศึกษาได้มีการนำเอาเทคโนโลยีด้านสารสนเทศมาประกอบการทำงานเพื่อเพิ่มความสะดวกรวดเร็วมากขึ้น ฐานข้อมูลจึงเป็นเทคโนโลยีที่ช่วยอำนวยความสะดวกในการเก็บข้อมูลซึ่งเป็นสิ่งที่สำคัญยิ่งในองค์กร เมื่อมีการเก็บข้อมูลไปนานวัน ปัญหาที่เกิดขึ้น คือ ข้อมูลที่จัดเก็บในฐานข้อมูลจะเพิ่มปริมาณมากขึ้นอย่างมหาศาลและสิ่งที่องค์กรคาดหวังจากข้อมูลในฐานข้อมูล คือ การใช้ประโยชน์จากข้อมูลที่เก็บนั้นแต่จะใช้ประโยชน์จากข้อมูลมหาศาลเหล่านี้ได้อย่างไร

เทคนิคที่จะช่วยให้องค์กรทางการศึกษาสามารถใช้ประโยชน์จากข้อมูลที่เก็บไว้ คือ การทำเหมืองข้อมูล หรือการขุดค้นสิ่งที่เป็นแก่นสารของข้อมูลหรือความรู้ที่คาดว่าจะเป็นประโยชน์ต่อองค์กร ซึ่งมีวิธีการทำงาน คือ การจำแนกข้อมูล โดยข้อมูลที่จะใช้วิธีการนี้จะต้องมีผลเฉลย เพื่อตรวจสอบความถูกต้องของแบบจำลองที่ดีที่สุดเพื่อใช้จำแนกข้อมูล และการจัดกลุ่มข้อมูล วิธีการนี้ไม่จำเป็นต้องอาศัยผลเฉลยในการจัดกลุ่มแต่จะใช้หลักการของความสัมพันธ์ของข้อมูลมาใช้ในการจัดกลุ่ม เช่น ความสัมพันธ์ที่มีความคล้ายคลึงกัน ความสัมพันธ์ที่มีความเกี่ยวข้องกัน เป็นต้น

ในองค์กรทางการศึกษา โดยทั่วไปจะมีข้อมูลของผู้เชี่ยวชาญเก็บไว้โดยที่ไม่ได้นำมาใช้ประโยชน์ จะสามารถใช้ประโยชน์จากข้อมูลดังกล่าวได้อย่างไร ข้อมูลของผู้เชี่ยวชาญนี้จะให้ประโยชน์อะไรกับองค์กร จากปัญหาดังกล่าวจึงทำให้มีแนวคิดในการจัดกลุ่มผู้เชี่ยวชาญ เนื่องจากผู้เชี่ยวชาญคนหนึ่ง ๆ อาจมีความเชี่ยวชาญมากกว่า 1 ความเชี่ยวชาญ หรือผู้เชี่ยวชาญต่างสายงานต่างสาขากัน อาจมีความเชี่ยวชาญที่เหมือนกันหรือคาบเกี่ยวกัน โดยลักษณะข้อมูลสำหรับงานวิจัยนี้เป็นข้อมูลที่ไม่มีผลเฉลย วิธีการที่ทำให้สามารถใช้ประโยชน์จากข้อมูลที่มีอยู่ในเบื้องต้น คือ การ

จัดกลุ่มหรือการแบ่งกลุ่มผู้เชี่ยวชาญ ดังนั้นในงานวิจัยฉบับนี้จึงมุ่งไปที่การทำเหมืองข้อมูลผู้เชี่ยวชาญโดยอาศัยเทคนิคการจัดกลุ่มแบบอัลกอริทึม 2 ขั้นตอนร่วมกับการให้น้ำหนักของคุณลักษณะ

วัตถุประสงค์

1. เพื่อศึกษาและวัดประสิทธิภาพในการจัดกลุ่มผู้เชี่ยวชาญแบบ 2 ขั้นตอน คือ (1) อัลกอริทึม SOM (Self-Organizing Maps) กับ อัลกอริทึม K-Means และ (2) อัลกอริทึม SOM กับ อัลกอริทึม Fuzzy C-Means

2. เพื่อศึกษาและเปรียบเทียบประสิทธิภาพจากเทคนิคในการจัดกลุ่มข้อมูลโดยให้น้ำหนักของคุณลักษณะโดยใช้วิธีการ Term Frequency-Inverse Document Frequency (TF-IDF) วิธีการ Logarithm Weight วิธีการ Augmented Weight และวิธีการไม่ให้น้ำหนักของคุณลักษณะ

3. เพื่อจัดกลุ่มผู้เชี่ยวชาญในมหาวิทยาลัยโดยใช้เทคนิคการจัดกลุ่มร่วมกับการให้น้ำหนักของคุณลักษณะ

ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถนำเอาผลที่ได้จากการทดลองนี้ไปใช้ในการพัฒนาต้นแบบของระบบค้นหาผู้เชี่ยวชาญในองค์กรได้ เช่น ระบบค้นหาผู้เชี่ยวชาญภายในมหาวิทยาลัย

2.สามารถแบ่งกลุ่มผู้เชี่ยวชาญตามลักษณะของงานวิจัยและงานวิทยานิพนธ์โดยใช้คำสำคัญของงาน

ขอบเขตของงานวิจัย

ข้อมูลที่นำมาใช้ในการทดลองเป็นเพียงข้อมูลในสถาบันการศึกษาของรัฐบาลแห่งหนึ่งในประเทศไทยที่มีการศึกษาในระดับปริญญาตรี ปริญญาโทและปริญญาเอก นอกจากนี้ข้อมูลที่ใช้ในการทดลองของงานวิจัยนี้เป็นข้อมูลผลงานวิจัยของสถาบันการศึกษาดังกล่าวในช่วงเวลาปี ค.ศ. 2002-2007 (พ.ศ. 2545-2550) ซึ่งสามารถเปลี่ยนแปลงความเชี่ยวชาญได้อีกหากทดลองกับข้อมูลในช่วงปีที่มากกว่าปี ค.ศ. 2007 ผู้เชี่ยวชาญในมหาวิทยาลัยนี้เป็นบุคคลากรของมหาวิทยาลัยที่มีผลงานวิจัยหรือเป็นอาจารย์ที่ปรึกษาวิทยานิพนธ์ของนิสิตระดับปริญญาโทหรือระดับปริญญาเอก

ข้อจำกัดของงานวิจัย

เนื่องจากผู้เชี่ยวชาญคนหนึ่งอาจจะมีความเชี่ยวชาญหลายด้านจึงสามารถมีคำสำคัญได้มากกว่าหนึ่งคำสำคัญและผู้เชี่ยวชาญต่างกลุ่มสายงานกันสามารถมีคำสำคัญที่เหมือนกันได้ ดังนั้นในการจัดกลุ่มจึงมีโอกาสสูงที่ผู้เชี่ยวชาญจากกลุ่มสายงานเดียวกันกระจายไปยังกลุ่มต่าง ๆ ได้ทุกกลุ่มและผู้เชี่ยวชาญจากกลุ่มสายงานที่ต่างกันอาจมีคำสำคัญที่เหมือนกันได้ งานวิจัยนี้ดำเนินงานกับสถาบันการศึกษาของรัฐบาลแห่งหนึ่งในประเทศไทยที่มีการศึกษาในระดับปริญญาตรี ปริญญาโทและปริญญาเอกเท่านั้น

การตรวจเอกสาร

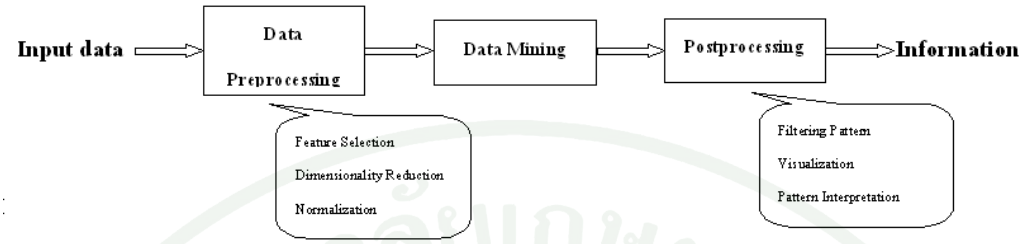
1. เทคนิคการทำเหมืองข้อมูล

การทำเหมืองข้อมูล (Data Mining) เป็นส่วนหนึ่งของการค้นพบความรู้ (Knowledge Discovery) ที่ใช้ระบุประเภทของรูปแบบข้อมูลที่สามารถค้นพบได้ คือ ข้อมูลที่ใช้ประโยชน์เพื่อการทำนาย (Predictive) และข้อมูลที่ใช้ประโยชน์เพื่อเชิงอธิบายหรือพรรณนา (Descriptive) ที่สามารถนิยามการทำเหมืองข้อมูลได้แตกต่างออกไปอีก เช่น การสกัดความรู้ (Knowledge Extraction) การค้นพบสารสนเทศ (Information Discovery) การประมวลรูปแบบข้อมูล (Data Pattern Processing) เป็นต้น (Cios *et al*., 2007) ในภาพที่ 1 เป็นการอธิบายกระบวนการในการค้นพบความรู้ที่เกี่ยวข้องกับการทำเหมืองข้อมูล

1. การเตรียมข้อมูลก่อนการประมวลผล (Data Preprocessing) ในขั้นตอนนี้จะเป็นการเตรียมข้อมูลนำเข้าก่อนที่จะนำข้อมูลไปประมวลผล เช่น การเปลี่ยนแปลงรูปแบบข้อมูลและสกัดข้อมูลที่ซ้ำซ้อน (Normalization) การสกัดสิ่งรบกวน (Outlier) การคัดเลือกคุณลักษณะ (Features Selection) เป็นต้น

2. การทำเหมืองข้อมูล เป็นขั้นตอนการสกัดความรู้จากข้อมูลที่มีอยู่ ทั้งนี้วิธีการที่จะใช้ขึ้นอยู่กับลักษณะของข้อมูล เช่น ข้อมูลมีผลเฉลยหรือข้อมูลไม่มีผลเฉย

3. การประเมินผล (Postprocessing) ในขั้นตอนนี้เป็นการประเมินผลที่ได้จากการทำเหมืองข้อมูลและนำผลดังกล่าวมาสรุปผล วิเคราะห์สิ่งที่ได้จากผลการทำเหมืองข้อมูลว่าส่วนใดสามารถนำมาใช้ประโยชน์ได้ จนท้ายสุดจะได้สารสนเทศ (Information) มาใช้ประโยชน์ในองค์กรต่อไป



ภาพที่ 1 แสดงกระบวนการค้นพบความรู้ (Knowledge Discovery in database-KDD)

ที่มา: Tan *et al*. (2006)

การทำเหมืองข้อมูลสามารถแบ่งการวิเคราะห์ออกเป็น 2 ชนิด (Cios *et al*., 2007; Han and Kamber, 2001) ได้แก่

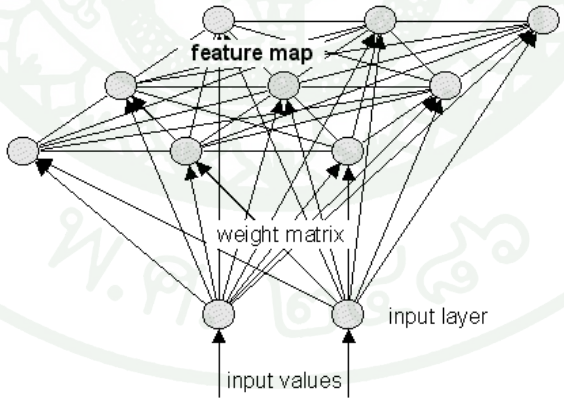
1.1 การจำแนกและการทำนาย (Classification and Predictive) เป็นวิธีการเรียนรู้หนึ่งในวิธีการเรียนรู้ข้อมูลแบบมีผลเฉลย (Supervised Learning) วิธีการเรียนรู้ข้อมูลแบบมีผลเฉลยใช้หลักการค้นหาแบบจำลอง (Model) เพื่อใช้ในการอธิบายหรือทำนายกลุ่มหรือคลาส (Classes) ของข้อมูลและนำไปสู่การวิเคราะห์ความถูกต้องของแบบจำลองที่ได้มา การวิเคราะห์แบบจำลองดังกล่าวจะอยู่บนพื้นฐานของเซตข้อมูลฝึก (Training Data Set) และผลเฉลย (Class Label Prediction) แบบจำลองเพื่อใช้ในการจำแนกนี้สามารถแทนอยู่ในรูปแบบ กฎ (Rules) ต้นไม้การตัดสินใจ (Decision Tree) และโครงข่ายประสาทเทียม (Neural Network) เป็นต้น

1.2 การจัดกลุ่มข้อมูล (Cluster Analysis) เป็นวิธีการเรียนรู้หนึ่งในวิธีการเรียนรู้ข้อมูลแบบไม่มีผลเฉลย (Unsupervised Learning) ซึ่งวิเคราะห์ข้อมูลโดยปราศจากผลเฉลย แต่อาศัยการจัดกลุ่มข้อมูลโดยพิจารณาจากความคล้ายคลึงภายในกลุ่ม (Intragroup) และความแตกต่างกันระหว่างกลุ่ม (Intergroup) ข้อมูลจะถูกจัดให้อยู่ในกลุ่มเดียวกันได้จะต้องมีความคล้ายคลึงภายในกลุ่มและความแตกต่างระหว่างกลุ่มมาก การจัดกลุ่มสามารถช่วยในการแบ่งประเภท (Taxonomy Formation) ได้ง่าย นั่นคือสามารถจัดกลุ่มที่มีความคล้ายคลึงกันให้อยู่ในรูปแบบลำดับขั้น (Hierarchy of Classes) ซึ่งสามารถบอกความสัมพันธ์ภายในกลุ่มเป็นลำดับขั้นได้ เช่น กลุ่มข้าว มีความสัมพันธ์ย่อยภายในกลุ่ม คือ ข้าวสาลี ข้าวเหนียว เป็นต้น

2. การหาจำนวนกลุ่ม

ปัญหาหลักในการจัดกลุ่ม คือ ข้อมูลไม่มีผลเฉลยและไม่อาจทราบได้ว่าข้อมูลที่จะนำมาจัดกลุ่มควรมีจำนวนกลุ่มที่เหมาะสมเท่าใด (วงกต, 2551) โดยทั่วไปจะใช้การเลือกมาหนึ่งอัลกอริทึมแล้วทดลองประมวลผลข้อมูลตั้งแต่ 2 กลุ่มไปจนถึงจำนวนกลุ่มที่ต้องการ เลือกจำนวนกลุ่มที่ให้ผลการทดลองที่ดีที่สุด ในที่นี้จะกล่าวถึงอัลกอริทึม Self-Organizing Maps (SOM) ซึ่งเป็นอัลกอริทึมหนึ่งที่นิยมนำมาใช้ในการหาจำนวนกลุ่ม สำหรับหลักการทำงานมีดังนี้

2.1 อัลกอริทึม SOM หรือเรียกอีกแบบหนึ่งว่า โครงข่าย Kohonen (Kohonen's map) พัฒนาโดย Kohonen ในปี 1982 และปี 1984 อัลกอริทึม SOM เป็นอัลกอริทึมโครงข่ายประสาทเทียมที่มีประสิทธิภาพถึงแม้ข้อมูลจะมีขนาดใหญ่ ลักษณะของอัลกอริทึมนี้จะเป็นโครงข่ายประสาทเทียมที่ไม่มีชั้นกลาง (Hidden Layer) มีแต่เพียงชั้นอินพุต (Input Layer) และชั้นเอ้าพุต (Output Layer) เท่านั้น เชื่อมด้วยเส้นประสาทเทียมที่มีการคำนวณค่าน้ำหนัก เพื่อใช้ประกอบการพิจารณาข้อมูลที่อยู่ในเซลล์ประสาทเทียมชั้นอินพุตขณะนี้ว่าควรจัดให้อยู่ในกลุ่มใด (Tan *et al*., 2006; Cios *et al*., 2007)



ภาพที่ 2 ตัวอย่างลักษณะโครงข่าย SOM

ที่มา: Koprinska (2001)

สูตรในการคำนวณน้ำหนัก

$$w_{i,j}(t+1) = w_{i,j}(t) + \eta(t)(x_i - w_{i,j}(t)) \quad (1)$$

ค่า $w_{i,j}(t+1)$ คือ ค่าน้ำหนักของรอบใหม่
ค่า $w_{i,j}(t)$ คือ ค่าน้ำหนักของรอบที่แล้ว
ค่า j คือจำนวนกลุ่มที่ต้องการจัดกลุ่ม
ค่า i คือจำนวนคอลัมน์หรือมิติของข้อมูล
ค่า η คือ อัตราการเรียนรู้ (Learning Rate) มีค่า $0 < \eta(t) < 1$
ค่า x_i คือ ข้อมูลที่เข้ามาอยู่ในชั้นอินพุตในขณะนั้น

สำหรับประสิทธิภาพเชิงเวลา (Time Complexity) คือ O(dn) เมื่อ d คือ จำนวนมิติของข้อมูลและ n คือ จำนวนข้อมูลทั้งหมด (ศศิธร, 2550) วิธีการทำงานของ SOM พื้นฐานได้แสดงดังนี้

ขั้นตอนการทำงานของอัลกอริทึม SOM (Tan *et al*., 2006)

1. เลือกจุดศูนย์กลางเริ่มต้น ซึ่งอาจได้มาจากการสุ่ม
2. ทำซ้ำข้อ 3 ถึง 5 เป็นต้นไป
3. นำข้อมูลใส่เข้าไปในเซลล์ประสาทเทียมในชั้นอินพุต
4. คำนวณระยะทางโดย

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

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

5. นำข้อมูลตัวที่ให้ระยะทางสั้นที่สุดจากการคำนวณข้อ 4 มาหาน้ำหนักในรอบใหม่และลดค่าอัตราการเรียนรู้ทีละครึ่ง

6. อัพเดตจุดศูนย์กลางตัวใหม่

7. หยุดการทำงานเมื่อจุดศูนย์กลางและสมาชิกไม่มีการเปลี่ยนแปลง

จุดแข็งของอัลกอริทึม SOM

1. มีประสิทธิภาพการทำงานดีกับข้อมูลที่มีขนาดใหญ่และยุ่งยากซับซ้อน หรือข้อมูลที่มีกระจายของจุดข้อมูลเป็นรูปร่างที่ยากต่อการจัดกลุ่ม เช่น ข้อมูลที่มีการกระจายจุดเป็นรูปวงแหวน

จุดด้อยของอัลกอริทึม SOM

1. จำเป็นต้องกำหนดพารามิเตอร์บางค่าเอง เช่น ค่าอัตราการเรียนรู้ ถ้ากำหนดค่าไม่เหมือนกันในแต่ละรอบ ผลการจัดกลุ่มจะได้ไม่เหมือนกัน, ค่าจำนวนโครงข่ายประสาทเทียม เป็นต้น

2. อาจจำเป็นต้องใช้ทรัพยากรบ้าง เช่น เวลาในการประมวลผลนาน ทั้งนี้ขึ้นอยู่กับจำนวนข้อมูลและการกำหนดจำนวนโครงข่ายเทียม

3. ทฤษฎีการจัดกลุ่ม (Clustering)

การจัดกลุ่มเป็นวิธีการเรียนรู้ข้อมูลที่ไม่มีผลเฉลย (Unsupervised Learning) ดังนั้นการจัดกลุ่มข้อมูลจะอาศัยหลักความสัมพันธ์ของคุณลักษณะ (Features or Attributes) และความคล้ายคลึงกันของข้อมูล ข้อมูลที่มีความคล้ายคลึงภายในกลุ่มมากและมีความแตกต่างระหว่างกลุ่มมาก จะเป็นการจัดกลุ่มที่ดี วิธีการในการจัดกลุ่มโดยทั่วไปมีหลายประเภทด้วยกัน ดังต่อไปนี้

3.1 วิธีการแบบแบ่งส่วน (Partitioning Method) วิธีการนี้จะทำการแบ่งข้อมูลออกเป็นส่วน k ส่วน และค่า k ต้องน้อยกว่าจำนวนแถวข้อมูล โดยค่า k แทนจำนวนกลุ่มที่ต้องการ วิธีการจัดกลุ่มประเภทนี้จะพิจารณาจากความใกล้และใช้หลักการทำงานแบบวนซ้ำ (Iterative Relocation Technique) จนกว่าข้อมูลที่แบ่งไว้ไม่มีการเปลี่ยนกลุ่ม ตัวอย่างอัลกอริทึมที่ใช้วิธีการแบบแบ่งส่วน คือ อัลกอริทึม K-Means และอัลกอริทึม Fuzzy C-Means

3.1.1 อัลกอริทึม K-Means

เป็นอัลกอริทึมการจัดกลุ่มพื้นฐาน ใช้รอบในการทำงานน้อย ซึ่งอาศัยหลักความใกล้กันของคุณลักษณะข้อมูล โดยเริ่มต้นจะทำการเลือกจุดศูนย์กลาง (Centroids) มาจำนวน k ตัว โดยแทนค่า k เท่ากับจำนวนกลุ่ม การพิจารณาเข้ากลุ่มจะพิจารณาจากระยะของจุดใด ๆ หรือข้อมูลใด ๆ เทียบกับจุดศูนย์กลาง ถ้าข้อมูลดังกล่าวมีความใกล้กับจุดศูนย์กลางใดมากกว่าจะถูกจัดให้เป็นสมาชิกกลุ่มนั้น ๆ ในรอบต่อไปจะมีการอัพเดตจุดศูนย์กลางแต่ละกลุ่มใหม่ ทำงานวนซ้ำไปเรื่อย ๆ จนสมาชิกในกลุ่มไม่มีการเปลี่ยนแปลงอีก สำหรับประสิทธิภาพเชิงเวลา คือ O(i*k*m*n) โดยที่ i ถูกกำหนดให้เป็นจำนวนรอบการทำงานเมื่อตัวแทนของกลุ่มเปลี่ยนแปลง ค่า k กำหนดให้เป็นจำนวนกลุ่ม ค่า m เป็นจำนวนคุณลักษณะและ n กำหนดเป็นจำนวนข้อมูลทั้งหมด หากรอบการทำงานน้อยจะทำให้ประสิทธิภาพเชิงเวลาเป็น O(n) ในภาพที่ 3 ได้แสดงการทำงานของอัลกอริทึม K-Means (Tan *et al*., 2006) และในภาพที่ 4 เป็นการแสดงตัวอย่างการจัดกลุ่มของอัลกอริทึม K-Means

---

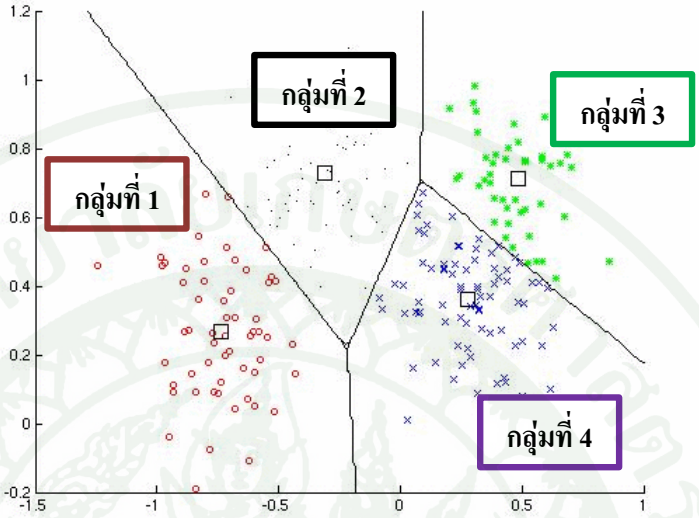
Algorithm Basic K-Means Algorithm

---

1. Select *k* points (may random) as initial centroids.
2. **repeat:**
3. From *k* clusters by assign as point which closest centroid.
4. Recompute the centroid of each cluster.
5. **until:** The centroids do not change much.

ภาพที่ 3 แสดงขั้นตอนการทำงานของอัลกอริทึม K-Means

ที่มา: Tan *et al*. (2006)



ภาพที่ 4 ตัวอย่างการจัดกลุ่มของอัลกอริทึม K-Means

ที่มา: Center for Machine Perception (2004)

จุดแข็งของอัลกอริทึม K-Means

1. เป็นอัลกอริทึมที่ง่ายและรวดเร็ว
2. สามารถใช้กับข้อมูลได้หลากหลายประเภท
3. ข้อมูลหนึ่งจุดสามารถถูกจัดให้อยู่ในกลุ่มเดียวเท่านั้น

จุดด้อยของอัลกอริทึม K-Means

1. เกิดข้อผิดพลาดในการจัดกลุ่มหากมีสิ่งรบกวนหรือข้อมูลมีลักษณะที่ยากต่อการจัดกลุ่ม กล่าวคือ พิจารณาจากการพล็อตจุดแล้ว รูปร่างข้อมูลไม่สามารถแยกกลุ่มให้เห็นชัดเจน
2. ผู้ใช้จำเป็นต้องกำหนดจำนวนกลุ่มหรือค่า k

3.1.2 อัลกอริทึม Fuzzy C-Means

เป็นอัลกอริทึมที่พัฒนามาจากอัลกอริทึม K-Means ดั้งเดิมจึงอาจกล่าวได้ว่าเป็นอัลกอริทึม K-Means ในรูปแบบ Fuzzy เนื่องจากมีลักษณะการทำงานที่คล้ายคลึงกับอัลกอริทึม K-Means มาก เป็นขั้นตอนการทำงานโดยใช้ทฤษฎี Fuzzy Set คือ เซตของกลุ่มที่มีการยอมรับสมาชิกใหม่เข้ามาอยู่ในกลุ่มโดยมีการกำหนดเงื่อนไขในการพิจารณาการรับเป็นสมาชิก อัลกอริทึม Fuzzy C-Means จะใช้ทำงานโดยใช้หลักความใกล้หรือความเป็นสมาชิกในการเข้ากลุ่มหลัก โดยในการทำงานเริ่มต้นจะทำการเลือกจุดศูนย์กลางและคำนวณค่า Fuzzy Pseudo Partition หรือค่าความเป็นสมาชิก ซึ่งก็คือ การกำหนดเงื่อนไขในการเข้ากลุ่มว่าควรอยู่ในกลุ่มใดจึงจะเหมาะสม กระบวนการทำงานโดยรวมมีดังนี้ ในขั้นแรกจะมีการกำหนดค่าความเป็นสมาชิก (w_{ij}) ในรอบแรกจะให้ข้อมูลทุกตัวมีค่า w_{ij} เท่ากันและต้องมีค่ารวมกันทั้งหมดเท่ากับ 1 (Tan *et al*., 2006; Cios *et al*., 2007)

$$\sum_{j=1}^{k} w_{ij} = 1 \qquad (2)$$

โดยมีเงื่อนไขหลักในการทำงาน คือ $0 < \sum_{i=1}^{m} w_{ij} < m$ เสมอ

ค่า m คือจำนวนข้อมูล m แถวหรือ m จุด

ค่า i คือข้อมูลตัวที่ i

ค่า k คือจำนวนกลุ่มหรือคลัสเตอร์

ค่า j คือกลุ่มหรือคลัสเตอร์ที่ j

ในรอบต่อไปจะมีการอัพเดตค่า w_{ij} และค่าจุดศูนย์กลางใหม่เรื่อย ๆ จนสมาชิกในกลุ่มและจุดศูนย์กลางไม่มีการเปลี่ยนแปลง

การอัพเดตค่า w_{ij} ในรอบต่อไปจาก

$$w_{ij} = (1/\mathrm{dis}(x_i, c_j)^2)^{\frac{1}{p-1}} / \sum_{q=1}^{k} (1/\mathrm{dis}(x_i, c_q)^2)^{\frac{1}{p-1}} \qquad (3)$$

ค่า x_i คือข้อมูลที่ i ที่กำลังพิจารณาอยู่

ค่า c_j คือค่าจุดศูนย์กลางของกลุ่มที่ j

ค่า q คือกลุ่มหรือคลัสเตอร์ที่ q

ค่า p คือค่าความสำคัญในการตัดสินใจ ซึ่งมีค่าระหว่าง 0 ถึง ∞

โดยทั่วไปจะกำหนดที่ค่า p เป็น 2 ถ้าหากค่า p ยิ่งเข้าใกล้ 1 จะทำให้การทำงานคล้ายกับอัลกอริทึม K-Means แต่ถ้าหากกำหนดมากเกินไป จุดศูนย์กลางจะครอบคลุมครอบข้อมูลในวงกว้างเกินไป ดังนั้นในกรณีนี้จึงใช้ค่า $p = 2$

การคำนวณค่าจุดศูนย์กลาง (c_j) ในรอบแรกอาจมาจากการสุ่มแต่ในรอบต่อไปจะคำนวณจาก

$$c_j = \sum_{i=1}^{m} w_{ij}{}^{p} x_i / \sum_{i=1}^{m} w_{ij}{}^{p} \qquad (4)$$

สำหรับประสิทธิภาพเชิงเวลาเป็น O(n) เมื่อ n คือจำนวนข้อมูลทั้งหมด (ศศิธร, 2550) ในภาพที่ 5 แสดงขั้นตอนการทำงานของอัลกอริทึม Fuzzy C-Means

จุดแข็งของอัลกอริทึม Fuzzy C-Means

1. เป็นอัลกอริทึมที่ง่ายและรวดเร็ว
2. สามารถใช้กับข้อมูลได้หลากหลายประเภท
3. เพราะมีการคำนวณค่า Fuzzy Pseudo Partition ซึ่งเป็นตัวบ่งบอกความคล้ายกันในการรวมกลุ่มดังนั้นจึงมีโอกาสที่ข้อมูลหนึ่งสามารถจัดให้อยู่ในกลุ่มใดก็ได้ ซึ่งการคำนวณค่าดังกล่าวทำให้การจัดกลุ่มมีความถูกต้องมาก

จุดด้อยของอัลกอริทึม Fuzzy C-Means

1. ใช้ทรัพยากรในการทำงานมาก เช่น มีการคำนวณค่า Fuzzy Pseudo Partition และเก็บค่าไว้

2. การกำหนดพารามิเตอร์ เช่น ค่า p สามารถส่งผลต่อการคำนวณค่า Fuzzy Pseudo Partition และจุดศูนย์กลาง

Algorithm Fuzzy C-Means Algorithm

---

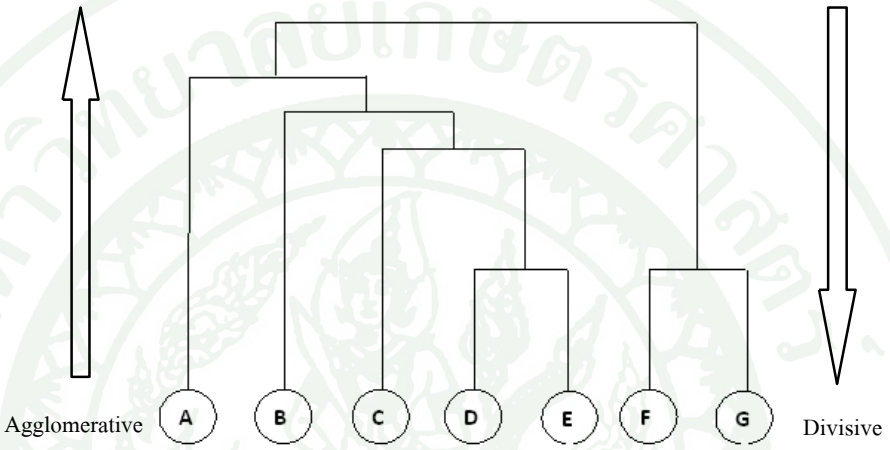
1. Select an initial fuzzy pseudo partition, i.e., assign values to all the w_{ij}
2. **repeat:**
3. Compute the centroid of each cluster using the fuzzy pseudo partition.
4. Recompute the fuzzy pseudo partition, i.e., the w_{ij}.
5. **until:** The centroids do not change or stopping conditions are "if the change in the error is below a speciified threshold" or "if the absolute change in any w_{ij} is below a given threshold".

ภาพที่ 5 แสดงขั้นตอนการทำงานของอัลกอริทึม Fuzzy C-Means

ที่มา: Tan *et al*. (2006)

3.2 วิธีการแบบลำดับขั้น (Hierarchical Method) เป็นวิธีการจัดกลุ่มที่พิจารณาการรวมกลุ่มหรือแยกกลุ่มจากเทคนิคความคล้ายคลึงกันหรือระยะทาง แบ่งออกเป็น 2 วิธีการหลัก ได้แก่ วิธีการจากล่างขึ้นบน (Bottom-up Mode) หรือวิธีการ Agglomerative Clustering ลักษณะการทำงานจะเริ่มจากกำหนดข้อมูลทุกตัวที่จะจัดกลุ่มแยกเป็นข้อมูลละหนึ่งกลุ่ม จากนั้นค่อย ๆ รวมกลุ่มไปเรื่อย ๆ จนได้ตามจำนวนกลุ่มที่ต้องการ จึงถือได้ว่าวิธีการจากล่างขึ้นบนนี้ใช้หลักการผสาน (Merge) และวิธีการจากบนลงล่าง (Top-Down Mode) หรือวิธีการ Divisive Clustering ซึ่งวิธีการทำงานจะกลับกันกับวิธีการจากล่างขึ้นบน โดยจะเริ่มจากให้ข้อมูลทุกตัวรวมตัวกันเป็นหนึ่งกลุ่ม จากนั้นค่อย ๆ แตกกลุ่มย่อยไปเรื่อย ๆ จนได้จำนวนกลุ่มตามที่ต้องการ สามารถกล่าวได้ว่าวิธีการจากบนลงล่าง

ใช้หลักการแตกย่อย (Split) ในการจัดกลุ่ม ในภาพที่ 6 เป็นตัวอย่างของวิธีการจัดกลุ่มแบบลำดับขั้น ทั้งวิธีการจากล่างขึ้นบนและบนลงล่างตามลำดับ (Han and Kamber, 2001; Tan *et al*., 2006)



ภาพที่ 6 ตัวอย่างการจัดกลุ่มแบบลำดับขั้น

3.3 วิธีการวิเคราะห์จากความหนาแน่น (Density-based Method) เป็นวิธีที่ใช้ที่พิจารณาการจัดกลุ่มจากความหนาแน่นหรือการเกาะกันของข้อมูลโดยพิจารณาเงื่อนไขหรือค่า Threshold บางค่าจากบริเวณที่ใกล้เคียงกับข้อมูลที่กำลังพิจารณาอยู่ เงื่อนไขดังกล่าวในการรวมกลุ่ม คือ การกำหนดค่ารัศมีในการกวาดสมาชิกที่อยู่ในรัศมีจากจุดศูนย์กลางให้เข้าร่วมกลุ่ม ทั้งนี้การรวมกลุ่มกันได้จะต้องขึ้นกับจำนวนจุดต่ำที่สุด (Minimum Number of Points) ที่กำหนดไว้เท่าใดจึงจะยอมให้จัดเป็นหนึ่งกลุ่มได้ ตัวอย่างอัลกอริทึมที่ใช้การวิเคราะห์จากความหนาแน่น ได้แก่ DBSCAN (A Density-Based Clustering Method Based on Connected Regions with Sufficiently High Desity) และ OPTICS (Ordering Points To Identify the Clustering Structure) (Han and Kamber, 2001)

3.4 วิธีการวิเคราะห์จากกริด (Grid-based Method) เป็นวิธีการจัดกลุ่มโดยแบ่งพื้นที่ข้อมูลทั้งหมดออกเป็นช่องหรือกริด (Grid) และใช้หลักช่องว่างในขอบเขตของเซลล์ในกริด วิธีการนี้ทำงานได้เร็วเพราะในการวิเคราะห์ช่องว่าง ข้อมูลแต่ละจุดเป็นอิสระต่อกันและขึ้นตรงกับจำนวนเซลล์เท่านั้น ตัวอย่างอัลกอริทึมที่ใช้วิธีการนี้ ได้แก่ STRING (Statistic Information Grid) (Han and Kamber, 2001)

3.5 วิธีการวิเคราะห์จากแบบจำลอง (Model-based Method) เป็นวิธีที่ใช้ข้อมูลในการทดสอบความเหมาะสมของแบบจำลองทีละแบบจำลองตามที่ได้มา วิธีการนี้ส่วนใหญ่มักใช้ในการทดสอบหาจำนวนกลุ่มที่เหมาะสมโดยตัดสินจากการทดสอบทางสถิติ สามารถแบ่งวิธีการในการวิเคราะห์จากแบบจำลองได้ 2 วิธีการ คือ การวิเคราะห์โดยใช้สถิติ (Statistical Approach) ซึ่งอัลกอริทึมที่ใช้วิธีการนี้ ได้แก่ COBWEB และวิธีการวิเคราะห์โดยใช้หลักโครงข่ายประสาทเทียม (Neural Network Approach) สำหรับอัลกอริทึมที่ใช้วิธีการแบบโครงข่ายประสาทเทียม ได้แก่ SOM ซึ่งได้อธิบายไว้ในข้อ 2 (Han and Kamber, 2001)

4. การให้น้ำหนักของคุณลักษณะ

โดยทั่วไปจะพบการให้น้ำหนักของคุณลักษณะ (Features Weighting) ในงานด้านการค้นคืนสารสนเทศและการทำเหมืองข้อมูลเพื่อเพิ่มประสิทธิภาพในการจัดกลุ่ม กล่าวคือ หากความถี่ของคำสำคัญที่ต่ำ แต่อาจเป็นคำเฉพาะที่สามารถแยกแยะเอกสารเพื่อใช้ในการจัดกลุ่มได้ดี ก็ควรเพิ่มความถี่ของคำสำคัญนั้น ๆ เพื่อเพิ่มความสำคัญของคำสำคัญ ตัวอย่างเทคนิคการให้น้ำหนักคุณลักษณะ ได้แก่ TF-IDF (Terms Frequency-Inverse Document Frequency) แบบจำลองของโรเบิร์ตสัน (Robertson, 1970, 1980) หรือ แบบจำลอง BM25 (Zhao and Lu, 2007) ซึ่งใช้หลักการค้นคืนโดยพิจารณาจากการเข้าคู่กันของเอกสารและใช้พื้นฐานความรู้ด้านความน่าจะเป็นซึ่งคิดค้นในปี 1970 และปี 1980 (http://en.wikipedia.org/wiki/BM25), วิธีการ Logarithm Weight, วิธีการ Augmented Weight เป็นต้น

4.1 TF-IDF (Term Frequency-Inverse Document Frequency) เป็นเทคนิคที่ใช้กันโดยทั่วไปและเป็นที่นิยม ซึ่งใช้หลักการนับความถี่ของคำสำคัญ (Terms Frequency) และจำนวนของเอกสาร (Document Frequency) ที่มีคำสำคัญที่สนใจปรากฏอยู่ เทคนิค TF-IDF มีสูตรในการคำนวณดังสมการต่อไปนี้ (Manning *et al*., 2008)

$$TF\text{-}IDF_{t,d} = TF_{t,d} * IDF_t \qquad (5)$$

ค่า $TF_{t,d}$ คือ จำนวนครั้งหรือความถี่ของคำสำคัญ t ปรากฏอยู่ในเอกสาร d

ค่า IDF_t คือ น้ำหนักของคำ t

โดยที่*ค่า* IDF_t สามารถคำนวณได้จาก

$$IDF_t = \log \frac{N}{df_t} \quad (6)$$

ค่า N คือ จำนวนเอกสารทั้งหมดที่ใช้พิจารณา

ค่า df_t คือ จำนวนเอกสารที่มีคำสำคัญ t ปรากฏอยู่

ข้อสังเกตหากค่า $TF\text{-}IDF_{t,d}$(Manning *et al*., 2008)

1. ค่าสูงมาก แสดงว่ามีคำสำคัญ t ปรากฏบ่อยครั้งในน้อยเอกสาร มีความสามารถในการจำแนกเอกสารสูง จึงถือว่าความสำคัญของคำ t ดังกล่าวมีมาก

2. ค่าต่ำ แสดงว่ามีคำ t ปรากฏเพียงเล็กน้อยในเอกสารทั้งหมด คำ t มีความสามารถในการจำแนกค่อนข้างต่ำ

3. ค่าต่ำมากหรือต่ำที่สุด แสดงว่ามีคำ t ปรากฏอยู่ในแทบทุกเอกสาร มีความสามารถในการจำแนกต่ำ จึงถือว่าความสำคัญของคำ t มีน้อย

ในการสังเกตค่า $TF\text{-}IDF_{t,d}$ เพื่อพิจารณาความสำคัญของคำ t จะพิจารณาเทียบจากเอกสารทั้งหมดในกรณีนั้น ๆ

4.2 Logarithm Weight หรือเรียกว่าวิธีการ Log TF-IDF (Logarithm Terms Frequency-Inverse Document Frequency) เป็นวิธีการที่ประยุกต์จาก TF-IDF โดยนำเอา Logarithm มาช่วยในการเพิ่มความถี่ของคำสำคัญ โดยทั่วไปมักใช้เพื่อการให้น้ำหนักคุณลักษณะสำหรับการค้นคืน แต่สามารถนำมาใช้เพิ่มประสิทธิภาพเพื่อการจัดกลุ่มได้ สำหรับ Log TF-IDF มีวิธีการคำนวณดังนี้ (Gong and Liu, 2001; Lan *et al*., 2005)

$$\text{Log } TF\text{-}IDF_{t,d} = \log(1+TF_{t,d}) * IDF_t \quad (7)$$

ค่า $TF_{t,d}$ คือ จำนวนครั้งหรือความถี่ของคำสำคัญ t ปรากฏอยู่ในเอกสาร d

สำหรับการคำนวณค่า IDF_t จะคำนวณตามสมการ (6)

4.3 Augmented Weight เป็นอีกวิธีการหนึ่งที่ประยุกต์มาจาก TF-IDF ซึ่งวิธีการนี้จะคำนึงถึงความยาวเอกสารหรือลักษณะของเอกสารในการเพิ่มความถี่ของคำสำคัญ โดยมีวิธีการคำนวณตามสมการต่อไปนี้ (Gong and Liu, 2001; Ginsparg, 2009)

$$Augmented = C + (\frac{C * TF_{t,d}}{\text{Max}_t (TF_{t,d})}) * IDF_t \quad (8)$$

ค่า $TF_{t,d}$ คือ จำนวนครั้งหรือความถี่ของคำสำคัญ t ปรากฏอยู่ในเอกสาร d

ค่า $\text{Max}(TF_{t,d})$ คือ ความถี่สูงสุดของคำใด ๆ ในเอกสาร d หรือเอกสารที่พิจารณาอยู่

ค่า C คือ ค่าคงที่ (Constant)

สำหรับการคำนวณค่า IDF_t จะคำนวณตามสมการ (6)

ข้อสังเกตในการกำหนดค่าคงที่ (Polettini, 2004)

1. หากกำหนดค่าคงที่ C เป็นค่า 0.5 หมายถึงเอกสารที่ใช้ในการทดลองเป็นเอกสารที่สั้น
2. แต่หากเป็นเอกสารที่ยาว (เอกสารมีความยาวเป็นหน้า) ค่าคงที่ C ควรกำหนดเป็นค่า 0.3
3. ดังนั้นค่า 0.3 และค่า 0.5 จึงเป็นค่ามาตรฐานขั้นต้น แต่สามารถเปลี่ยนแปลงเพิ่มหรือลดหลั่นตามความเหมาะสม
4. โดยทั่วไปไม่ควรต่ำกว่าค่า 0.5 และไม่ควรกำหนดเป็นค่า 1 สำหรับเอกสารที่สั้น
5. สำหรับเอกสารที่ยาวซึ่งโดยทั่วไปควรกำหนดเป็นค่า 0.3 แต่ไม่ควรเป็นค่า 0

สำหรับการกำหนดว่าเอกสารมีขนาดสั้นหรือขนาดยาวขึ้นอยู่กับผู้ทดลองกำหนดตามความเหมาะสม แต่โดยทั่วไปค่า 0.5 ถือเป็นค่ามาตรฐาน

5. การคัดเลือกคุณลักษณะ

ชุดข้อมูลขนาดใหญ่ที่มีคุณลักษณะจำนวนมากเป็นสาเหตุหนึ่งที่ทำให้สิ้นเปลืองทรัพยากรมากในการวิเคราะห์ข้อมูล เช่น เวลาในการทำงานและหน่วยความจำ เป็นต้น ดังนั้นเทคนิคการคัดเลือกคุณลักษณะ (Features Selection) จึงเป็นทางเลือกหนึ่งที่ช่วยลดการใช้ทรัพยากรลง เทคนิคการคัดเลือกคุณลักษณะ คือ กระบวนการในการจำแนกแยกแยะและลบข้อมูลที่ไม่จำเป็นออกไป ซึ่งช่วยลดปริมาณของคุณลักษณะ และเพิ่มประสิทธิภาพเรื่องเวลาในการทำงานกับอัลกอริทึมอื่นในขั้นตอนถัดไป การคัดเลือกคุณลักษณะจะใช้หลักการค้นหาคุณลักษณะผ่านทางช่องว่างหรือระยะห่างของเซตคุณลักษณะย่อย (Feature subset space) ซึ่งมีขั้นตอนการค้นหาหลัก ๆ 4 ขั้นตอน (Hall, 1999) คือ

1. ขั้นตอนเริ่มต้น ทำการเลือกจุดเริ่มต้นในระยะห่างของเซตคุณลักษณะย่อย หากเริ่มต้นในรอบแรกไม่มีคุณลักษณะใด ๆ เซตคุณลักษณะย่อยที่ดีที่สุดจะถูกเลือกก่อนซึ่งเป็นการประมวลผลแบบเดินหน้าทีละขั้น (Forward) ในทางกลับกัน หากเป็นการประมวลผลแบบย้อนหลังทีละขั้น (Backward) จะเริ่มรอบแรกจากเซตคุณลักษณะทั้งหมดแล้วคัดเซตคุณลักษณะย่อยที่ไม่ดีออกไปก่อน จะหยุดเมื่อไม่มีคุณลักษณะที่ผ่านตามเงื่อนไข ในกรณีนี้จะค้นหาช่องว่างในเซตคุณลักษณะโดยประมวลผลแบบย้อนหลัง

2. ขั้นตอนค้นหา แบบ Exhaustive (การค้นหาทั้งหมด) เป็นวิธีการค้นหาระยะห่างของคุณลักษณะ สมมติมีคุณลักษณะทั้งหมด N คุณลักษณะ ต้องใช้เวลาในการค้นหาระยะห่างของคุณลักษณะทั้งหมด 2^N แต่ในการค้นหาแบบฮิวริสติก (Heuristic Search) จะมีกลยุทธ์ในการค้นหาที่ดีกว่านั้นถึงแม้จะไม่สามารถการันตีได้ว่าดีที่สุดแต่ถือว่ายังดีกว่าวิธีการ Exhaustive เพราะการค้นหาแบบฮิวริสติกจะทำการค้นหาข้อมูลเพียงบางส่วน

3. ขั้นตอนประเมินการค้นหาคุณลักษณะ การประเมินคุณลักษณะที่ได้มาเป็นส่วนประกอบในการตัดสินใจว่าคุณลักษณะที่ค้นหามายอมรับได้หรือไม่ มี 2วิธีการหลัก คือ Filter ซึ่งเป็นวิธีการที่ทำงานโดยเป็นอิสระจากอัลกอริทึมที่เข้ามาทำงานด้วยฮิวริสติกจึงเป็นหนึ่งในวิธีการที่ใช้การประเมินแบบ Filter โดยทั่วไปจะประเมินผลจากค่า Merit ของเซตย่อยคุณลักษณะ อีกวิธีการหนึ่งคือ Wrapper ใช้หลักการสุ่มซ้ำ ๆ (Re-Sampling) เช่นการทำ Cross-Validation ในการประเมินค่าความถูกต้องของเซตย่อยคุณลักษณะ

4. หยุดการค้นหา ในขั้นตอนการหยุดค้นหาช่องว่างของเซตย่อยคุณลักษณะเมื่อพบว่าไม่มีเซตย่อยคุณลักษณะใดที่ทำให้ค่า Merit สูงขึ้นอีกแล้ว

โดยเทคนิคจะกล่าวถึงนี้ คือ การใช้ตัวคำนวณความสัมพันธ์ของคุณลักษณะ (Attribute Evaluator) โดย CFS (Correlation-based Feature Subset Selection) และใช้วิธีการเชิงพันธุกรรมพื้นฐาน (Genetic Search) ในการค้นหาคุณลักษณะ (Search Method) สำหรับรายละเอียดของทั้งสองวิธีการมีดังต่อไปนี้

5.1 CFS Subset Evaluator เป็นเทคนิคที่ใช้หลักการค้นหาสหสัมพันธ์เชิงฮิวริสติก (Correlation based Heuristic) โดยทั่วไปจะใช้วิธีการค้นหาแบบดีที่สุดก่อน (Best First Search) ประกอบการตัดสินใจในการคัดเลือกคุณลักษณะ โดยวิธีการค้นหาแบบดีที่สุดก่อนจะคล้ายการปีนเขาแบบเชิงละโมบ (Greedy Hill Climbing) ใช้ทฤษฎีกราฟในการเก็บข้อมูลตรวจสอบโหนดที่เชื่อมโยงกับโหนดที่พิจารณาจากการคำนวณโดยใช้ฟังก์ชันฮิวริสติกแล้วเลือกเส้นทางที่จะไปยังโหนดถัดไป (ขึ้นกับลักษณะของปัญหา) ซึ่งเป็นโหนดที่ดีกว่าโหนดเดิม CFS นี้จะใช้หลักการการคำนวณค่าความไม่เป็นระเบียบของข้อมูล (Entropy) ดังสมการ (9) และสมการ (10) หากค่าที่ได้ยิ่งน้อย จะสามารถบ่งบอกได้ว่าคุณลักษณะนั้นสามารถลดความไม่เป็นระเบียบของข้อมูลได้ แล้วนำค่าความไม่เป็นระเบียบนี้มาคำนวณค่าความไม่แน่นอน (Uncertainty Coefficient) ต่อตามสมการ (11) ในการพิจารณาความสัมพันธ์ระหว่างคุณลักษณะซึ่งสามารถคำนวณได้จาก

$$H(Y) = -\sum_{y \in R_y} p(y)\log(p(y)) \tag{9}$$

$$H(Y|X) = -\sum_{x \in R_x} p(x) \sum_{y \in R_y} p(y \mid x)\log(p(y \mid x)) \tag{10}$$

$$C(Y|X) = \frac{H(Y) - H(Y \mid X)}{H(Y)} \tag{11}$$

ค่า $p(y)$ คือ โอกาสที่เลือกคุณลักษณะ y (Probability Distribution of y)

ค่า p(x) คือ โอกาสที่เลือกคุณลักษณะ x (Probability Distribution of x)

ค่า p(y|x) คือ โอกาสที่เลือกคุณลักษณะทั้ง x และคุณลักษณะ y

ค่า $H(Y)$ คือค่าความไม่เป็นระเบียบของคุณลักษณะ Y

ค่า $H(Y|X)$ คือค่าความไม่เป็นระเบียบเมื่อมีคุณลักษณะ X แล้วจะมีคุณลักษณะ Y หรือคุณลักษณะ X และ Y ประกอบกัน

ค่า $C(Y|X)$ คือค่าความไม่แน่นอนเมื่อคุณลักษณะ X และ Y ประกอบกัน

โดยค่าความไม่แน่นอนนี้เป็นตัววัดความสัมพันธ์ระหว่างคุณลักษณะที่สนใจอยู่ มีค่าได้ระหว่าง 0 ถึง 1 หากค่ายิ่งเข้าใกล้ 1 แสดงว่าคุณลักษณะทั้งสองคุณลักษณะมีความสัมพันธ์กันมาก นำค่าความไม่แน่นอนนี้มาประกอบการตัดสินผลการคัดเลือกคุณลักษณะ โดยนำค่าความไม่แน่นอนนี้มาคิดค่าเฉลี่ยของความสัมพันธ์ของคุณลักษณะ ซึ่ง CFS จะตัดสินจากการคำนวณค่าความสามารถในการจัดการคุณลักษณะที่มีความสัมพันธ์กันต่ำ (Merit Value) ภายในของกลุ่มคุณลักษณะ S ใด ๆ ที่มี k คุณลักษณะสามารถคำนวณได้จาก

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (12)$$

ค่า $\overline{r_{cf}}$ คือ ค่าเฉลี่ยของความสัมพันธ์ระหว่างคุณลักษณะกับกลุ่มของคุณลักษณะที่ถูกเลือก (Class Attribute) โดยที่ $f \in S$

ค่า $\overline{r_{ff}}$ คือ ค่าเฉลี่ยของความสัมพันธ์ระหว่างคุณลักษณะที่ถูกเลือกภายในกลุ่มของคุณลักษณะที่ถูกเลือก

ค่า k คือ จำนวนคุณลักษณะในกลุ่ม S (เซตคุณลักษณะที่ถูกคัดเลือก)

ค่า S คือ กลุ่มคุณลักษณะที่มี k คุณลักษณะ

ค่าความสัมพันธ์ของคุณลักษณะที่สูงจะถือว่าเป็นการคัดเลือกคุณลักษณะที่ดี เพราะสามารถบ่งบอกได้ว่าคุณลักษณะของข้อมูลแต่ละกลุ่มมีความเกี่ยวข้องกันสูง (Hall and Smith, 1997)

5.2 Genetic Algorithm (GA) Search Method โดยทั่วไปการคัดเลือกคุณลักษณะโดยใช้วิธีการ CFS นี้จะทำการค้นหาพื้นฐานโดยวิธีการค้นหาแบบดีที่สุดก่อน แต่ในงานวิจัยนี้จะใช้การค้นหาโดยใช้วิธีการ GA Search Method ซึ่งเป็นการค้นหาโดยใช้เทคนิคทางพันธุกรรมพื้นฐาน เป็นวิธีการค้นหาคำตอบโดยอาศัยหลักทางพันธุศาสตร์ กล่าวคือ กำหนดข้อมูลที่ต้องการทดสอบให้อยู่ในรูปแบบโครโมโซม ซึ่งภายในโครโมโซมจะมียีนเป็นส่วนประกอบของแต่ละโครโมโซม จากนั้นทำการวัดความแข็งแรงของประชากรหรือโครโมโซมโดยอาศัยจากการคำนวณค่าความแข็งแรง (Fitness Function) ประชากรที่แข็งแรงเท่านั้นจึงจะอยู่รอดต่อไป ซึ่งกระบวนการทางพันธุกรรมหลัก ๆ (Fent, 1999) ได้แก่

1. เริ่มจากการสร้างประชากรต้นกำเนิด โดยกำหนดให้อยู่ในรูปแบบโครโมโซม (Chromosomes) ดังที่ได้แสดงในภาพที่ 7 (ก) ในกรณีศึกษานี้มีจำนวนยีน 971 ยีน

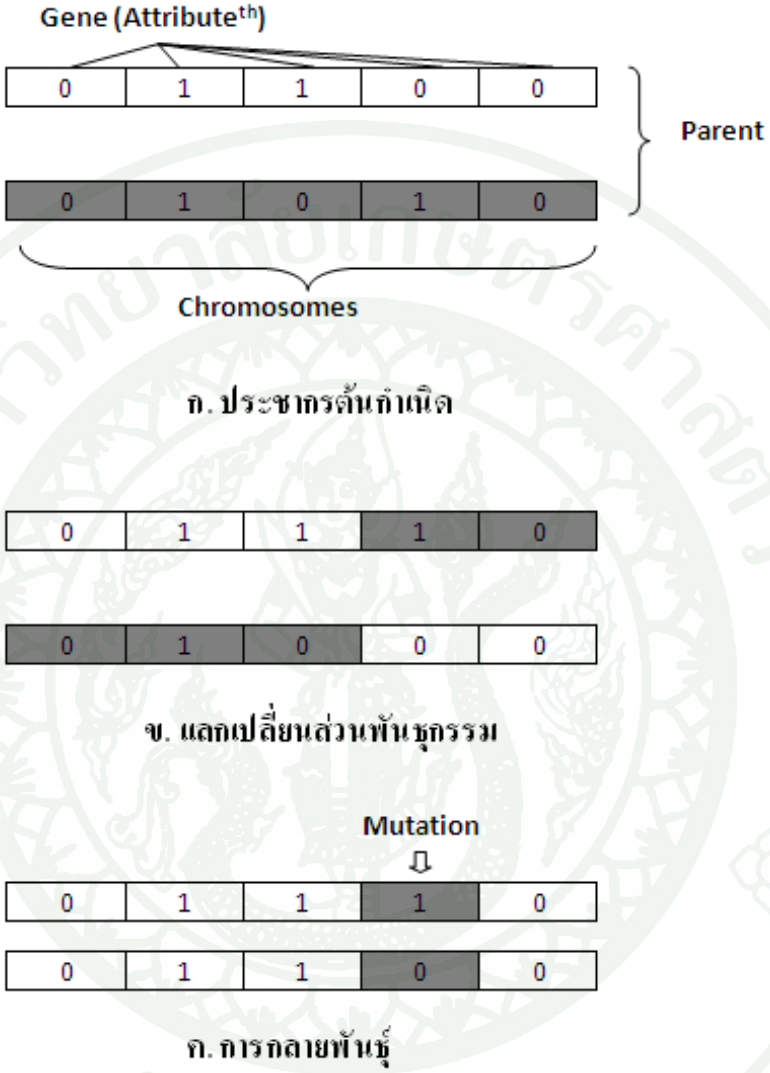
2. การแลกเปลี่ยนส่วนของพันธุกรรม (Crossover) อาจใช้วิธีการสุ่มคู่ของประชากรเพื่อนำมาแลกเปลี่ยนกันได้ ในขั้นตอนการแลกเปลี่ยนนี้เองอาจทำให้ค้นพบสายของประชากรใหม่หรือกฎใหม่ที่ไม่เคยพบมาก่อนก็ได้ ดังที่ได้แสดงในภาพที่ 7 (ข) ในกรณีศึกษานี้กำหนดอัตราการแลกเปลี่ยนส่วนพันธุกรรมเท่ากับ 0.6

3. การคำนวณค่าความแข็งแรง

4. คัดเลือกประชากร (Select Population) เพื่อใช้เป็นประชากรรุ่นต่อไปโดยคัดเลือกจากการคำนวณค่าความแข็งแรงมาประกอบการคัดเลือกประชากร

5. การกลายพันธุ์ (Mutation) ซึ่งในขั้นตอนนี้อาจไม่มีก็ได้ ขึ้นอยู่กับการกำหนดอัตราการกลายพันธุ์ ในการกลายพันธุ์อาจทำให้ค้นพบสายของประชากรใหม่หรือกฎใหม่ที่ไม่เคยพบมาก่อน ซึ่งตัวอย่างการกลายพันธุ์ได้แสดงในภาพที่ 7 (ค) ในกรณีศึกษานี้กำหนดอัตราการกลายพันธุ์เท่ากับ 0.033

ในภาพที่ 7 ได้นำเสนอวิธีการแลกเปลี่ยนส่วนพันธุกรรมและการกลายพันธุ์ใน Genetic Algorithm



ภาพที่ 7 แสดงขั้นตอนการแลกเปลี่ยนส่วนพันธุกรรมและการกลายพันธุ์

ในภาพที่ 8 แสดงการทำงานของ Genetic Algorithm Search Method (Hall, 1999)

Algorithm Simple genetic search strategy

1. Begin by randomly generating an initial population P.
2. Calculate e(x) for each member x ∈ P. //คำนวณค่าความแข็งแรง
3. Define a probability distribution p over the members of P where p(x) α e(x).
4. Select two population members x and y with respect to p. //เลือกประชากรที่ดีที่สุด
5. Apply crossover to x and y to produce new population members x′ and y′.
6. Apply mutation to x′ and y′.
7. Insert x′ and y′ into P′ (the next generation).
8. If | P′| < |P|, goto 4.
9. Let P ← P′.

//ทำต่อหากจำนวนรุ่นยังไม่ครบ ตามกำหนด

10. If there are more generations to process, goto 2.
11. Return x ∈ P for which e(x) is highest. //คืนค่าประชากรที่มีความแข็งแรงมากที่สุด

ภาพที่ 8 แสดงขั้นตอนการทำงานของ Genetic Search

ที่มา: Hall (1999)

หลักการทำงานของการคัดเลือกคุณลักษณะ (สุคนธ์ทิพย์, 2551; Hall, 1999)

1. กำหนดประชากรเริ่มต้น ซึ่งอาจได้มาจากการสุ่ม (ในที่นี้กำหนดรุ่นประชากรสูงสุด 20 รุ่น (Generation)) และกำหนดให้อยู่ในรูปแบบโครโมโซม ดังภาพที่ 7

2. ทำการแทนค่าคุณลักษณะแต่ละคุณลักษณะด้วย 0 และ 1 (0 หมายถึง คุณลักษณะถูกใช้, 1 หมายถึง คุณลักษณะไม่ถูกใช้) เพื่อแสดงว่าคุณลักษณะถูกใช้ โดยทำเป็นเซต Individual (เซตของคุณลักษณะที่ถูกคัดเลือกจากคุณลักษณะทั้งหมด)

3. วัดค่าความเหมาะสมหรือค่าความแข็งแรงของ Individual แต่ละเซต โดยตัดสินจากวิธีการ CFS ว่าเซตของคุณลักษณะใดเหมาะสมที่สุด คุณลักษณะเซตนั้นจะถูกคัดเลือก โดยพิจารณาจากการคำนวณค่า Merit

4. สลับบางส่วนของเซต Individual หรือใช้เทคนิคกลายพันธุ์เพื่อให้ได้เซต Individual เซตใหม่ ๆ

5. ทำซ้ำข้อ 3 ใหม่ หยุดการทำงานเมื่อเป็นไปตามเงื่อนไขที่กำหนด เช่น ประชากรทั้งหมดซ้ำกับรอบเดิม หรือคลอบคลุมประชากรทุกแบบที่เป็นไปได้

6. การวัดประสิทธิภาพการจัดกลุ่ม

ตัววัดประสิทธิภาพในการจัดกลุ่มเป็นสิ่งที่สามารถอ้างอิงผลการทำงานหรือผลการจัดกลุ่มได้ว่ามีประสิทธิภาพในการจัดกลุ่มดีหรือไม่ดี ตัววัดประสิทธิภาพโดยทั่วไปนั้นจะใช้หลักการทางสถิติเป็นตัวอ้างอิง

6.1 F-Statistic เป็นทดสอบอัตราระหว่างความแปรปรวนของข้อมูลระหว่างกลุ่ม (Mean Square for Treatments-MSTr) กับความแปรปรวนระหว่างข้อมูลภายในกลุ่ม (Mean Square Error-MSE) หากค่า F-Statistic มีค่ามากแสดงว่าข้อมูลระหว่างกลุ่มมีความแตกต่างกันมาก ถือว่าเป็นการจัดกลุ่มที่ดี ในทางกลับกันหากค่า F-Statistic มีค่าน้อยแสดงว่าความแปรปรวนระหว่างกลุ่มกับความแปรปรวนภายในกลุ่มมีค่าใกล้เคียงกันแต่ไม่อาจสรุปไปในแนวโน้มความแปรปรวนระหว่างกลุ่มได้ สามารถคำนวณค่า F-Statistic ได้จาก (Matignon, 2007)

$$F\text{-}Statistic = \frac{MSTr}{MSE} \tag{13}$$

$$MSTr = \frac{\sum_{i=1}^{n} \| x_i - \bar{x} \|^2}{k-1} \tag{14}$$

$$\text{MSE} = \frac{\sum_{i=1}^{k} \| x_i - \bar{x}_k \|^2}{n - k} \tag{15}$$

ค่า $MSTr$ คือ ความแปรปรวนของข้อมูลระหว่างกลุ่ม (Intergroup)

ค่า MSE คือ ความแปรปรวนของข้อมูลภายในกลุ่ม (Intragroup)

ค่า x_i คือ ข้อมูลตัวที่ i

ค่า $\bar{x}$ คือ ค่ากลางของข้อมูล

ค่า n คือ จำนวนข้อมูล

ค่า k คือ จำนวนกลุ่ม

ค่า $\bar{x}_k$ คือ ค่ากลางของข้อมูลภายในกลุ่มตัวที่ k

6.2 Silhouette เป็นตัววัดความคล้ายคลึงกันของข้อมูลภายในกลุ่ม พิจารณาค่า Silhouette (s_i) จากอัตราระหว่างค่าเฉลี่ยของระยะทางภายในกลุ่มกับค่าเฉลี่ยของระยะทางระหว่างกลุ่มโดยจะมีค่าอยู่ระหว่าง -1 ถึง 1 ถ้าหากค่า Silhouette ติดลบ แสดงว่าระยะทางภายในกลุ่มมีความห่างกันมากและหากค่า Silhouette ยิ่งเข้าใกล้ 1 แสดงว่าระยะทางเฉลี่ยภายในกลุ่มมีค่าน้อย (หรืออีกนัยหนึ่งระยะทางเฉลี่ยระหว่างกลุ่มมีค่ามาก) มีการแบ่งแยกกลุ่มได้ชัดเจน สามารถคำนวณค่า Silhouette ได้จาก (Tibshirani *et al*., 2001; Fanizzi *et al*., 2007)

$$a_i = \frac{1}{|C_j|} \sum_{r \in c_j} d_p(a_i, x) \tag{16}$$

$$b_i = \frac{1}{|C_j|} \sum_{r \in c_h}^{h \neq j} d_p(a_i, x) \tag{17}$$

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{18}$$

ค่า a_i คือ ระยะทางเฉลี่ยภายในกลุ่ม

ค่า b_i คือ ระยะทางเฉลี่ยระหว่างกลุ่ม

ค่า d_p คือ ระยะทางระหว่างระยะทางเฉลี่ยของข้อมูลภายในกลุ่มกับข้อมูลที่สนใจอยู่

ค่า x คือ ข้อมูลที่สนใจอยู่

ค่า C_j คือ กลุ่มที่ j โดยที่ $j \in \{1,..., k\}$ จากจำนวนกลุ่มทั้งหมด k กลุ่ม

ค่า C_h คือ กลุ่มที่ h โดยที่ $h \in \{1,..., k\}$ จากจำนวนกลุ่มทั้งหมด k กลุ่ม และ $h \neq j$

จากนั้นทำการหาค่าเฉลี่ยของค่า Silhouette ดังนี้

$$s_i = \frac{1}{k} \sum_{1=1}^{k} s_i \tag{19}$$

ค่า k คือ จำนวนกลุ่ม

หากค่า Silhouette ที่สูงถือว่าเป็นการจัดกลุ่มที่ดี เนื่องจากค่าเฉลี่ยของระยะทางข้อมูลระหว่างกลุ่มมีค่ามากกว่าค่าเฉลี่ยของระยะทางข้อมูลภายในกลุ่ม ซึ่งก็คือ กลุ่มที่แบ่งมานั้นถูกแบ่งแยกอย่างชัดเจน

6.3 R-Squared ในที่นี้เรียกย่อว่า RS เป็นตัววัดความแตกต่างกันระหว่างกลุ่มและอัตราการรวมเป็นกลุ่มเดียวกัน (Degree of Homogeneity) ค่า RS จะมีค่าได้ระหว่าง 0 ถึง 1 หากค่ายิ่งเข้าใกล้ 0 จะหมายถึงข้อมูลที่จัดกลุ่มแทบไม่มีความแตกต่างระหว่างกลุ่มเลย ในทางกลับกันหากค่ายิ่งเข้าใกล้ 1 แสดงว่าข้อมูลที่จัดกลุ่มมีความแตกต่างระหว่างกลุ่มมาก ซึ่งค่า RS สามารถใช้เป็นตัวสนับสนุนควบคู่กับค่า F-Statistic ได้เพื่อวิเคราะห์ว่าสิ่งที่คาดการณ์หรือผลการจัดกลุ่มที่ได้นั้นมีความเหมาะสมหรือไม่ (Kovacs *et al*., 2006; Matignon, 2007)

$$RS = 1 - \left(\frac{\sum_{i=1}^{k} || x_i - \bar{x}_k ||}{\sum_{i=1}^{n} || x_i - \bar{x} ||} \right) \qquad (20)$$

ค่า x_i คือ ข้อมูลตัวที่ i

ค่า $\bar{x}$ คือ ค่ากลางของข้อมูล

ค่า n คือ จำนวนข้อมูล

ค่า k คือ จำนวนกลุ่ม

ค่า $\bar{x}_k$ คือ ค่ากลางของข้อมูลภายในกลุ่มตัวที่ k

7. งานวิจัยที่เกี่ยวข้อง

งานของ Zheng และคณะ (Zheng *et al*., 2003) ได้ทำการทดลองจัดกลุ่มเอกสารโดยใช้การเยี่ยมชมโหนดในระยะทางที่สั้นที่สุด (Minimum Spanning Tree-MST) และใช้เทคนิค VSM ในการแทนคำสำคัญและใช้เทคนิค TF-IDF ในการจัดลำดับการแยกส่วนของเอกสารเพื่อแทนโหนดใน MST โดยทำการทดลองเทียบกับวิธีการจัดกลุ่มแบบธรรมดา จากผลการทดลองพบว่าวิธีการ MST ให้ประสิทธิภาพดีที่สุด

Wang และคณะ (Wang *et al*., 2004) ทำการศึกษาการคัดเลือกคุณลักษณะยีนของการเกิดโรคมะเร็ง ซึ่งโดยใช้วิธีการ Wrapper, filters และ CFS โดยกำหนดคุณลักษณะให้อยู่ในรูปแบบเวคเตอร์ การใช้ Wrapper ร่วมกับอัลกอริทึม J48, Naïve Bayes และ Support Vector Machines จากผลการทดลองพบว่าวิธีการ Wrapper สามารถเลือกยีนเพื่อใช้ในการสร้างโมเดลในการจำแนกโรคมะเร็งได้ดีที่สุดและวิธีการ CFS สามารถเลือกยีนที่มีความสัมพันธ์เชื่อมโยงกับการเกิดโรคมะเร็งได้ดีที่สุด นอกจากนี้วิธีการ CFS ยังเป็นวิธีการที่ใช้เวลาในการทำงานเร็วกว่าวิธีการ Wrapper

Aarabi และคณะ (Aarabi *et al*., 2005) ได้ทำการทดลองจำแนกอาการปัจจุบันของโรคในทารกแรกเกิดจากคลื่นสมอง (คลื่นสมองในแต่ละช่วงสามารถจับโรคได้ต่างโรคกัน) ด้วยวิธีโครงข่ายประสาทเทียม เปรียบเทียบระหว่างข้อมูลที่ผ่านคัดเลือกคุณลักษณะด้วย CFS, RelifF, FCBF (Fast correlation-based filter) และ ข้อมูลที่ไม่ผ่านการคัดเลือกคุณลักษณะ โดยกำหนดคุณลักษณะให้อยู่ในรูปแบบเวคเตอร์รวม 132 คุณลักษณะของโรงพยาบาล North Hospital of Amiens จากผลการทดลองพบว่าการจำแนกข้อมูลด้วยวิธีการโครงข่ายประสาทเทียมและคัดเลือกคุณลักษณะด้วยวิธีการทั้ง CFS และ ReliefF (Relevance of Features) ช่วยเพิ่มประสิทธิภาพในการจำแนกอาการปัจจุบันของโรคในทารกแรกเกิดดีที่สุดแต่คุณลักษณะที่คัดเลือกจาก CFS มีความสัมพันธ์ของคุณลักษณะดีที่สุด

การทดลองของ Lan และคณะ (Lan *et al*., 2005) ได้ทำการศึกษางานด้านการให้น้ำหนักคุณลักษณะสำหรับการจัดกลุ่มข้อความ โดยเทคนิคที่ใช้ในการให้น้ำหนักเอกสารหลายเทคนิค เช่น (1) เทคนิค Binary (2) เทคนิค TF (Term Frequency) และ (3) เทคนิค Logarithm Weight เป็นต้น ทำการเปรียบเทียบทุกวิธีโดยใช้กับข้อมูลข่าว ได้แก่ ข้อมูลของสำนักข่าวรอยเตอร์ (Reuters)

จำนวน 10 กลุ่มข่าวและข้อมูลข่าวจากแหล่งอื่นสุ่มมา 20 กลุ่มข่าว กลุ่มละ 300 ตัวอย่าง ทำการทดลองกับจำนวนคุณลักษณะที่มีขนาดต่าง ๆ กัน เทคนิคที่ให้ประสิทธิภาพที่ดีที่สุด คือ TF.RF (Term Frequency-Relevance Frequency) รองลงมา คือ TF วิธีการ Logarithm Weight และ ITF (Inverse Term Frequency) ตามลำดับ

สำหรับการวัดประสิทธิภาพการจัดกลุ่ม ได้มีงานวิจัยที่ได้นำเสนอวิธีการในการวัดประสิทธิภาพในการจัดกลุ่ม ในงานวิจัยของ Kovacs และคณะ (Kovacs *et al*., 2006) ได้นำเสนอเทคนิคการวัดประสิทธิภาพการจัดกลุ่ม ได้แก่ Davies Bouldin Index (DB-Index), Root Mean Squared Standard Deviation (RMSSD), Dunn and Dunn Like Indices, R-Squared, SD Validity Index (SD-Index) และ S_Dbw Validity Index นำมาทดลองกับชุดข้อมูลสามชุด โดยใช้อัลกอริทึม K-Means ในการจัดกลุ่ม ในชุดข้อมูลชุดแรกพบว่า Dunn และ SD-Index ให้ประสิทธิภาพดีที่สุด ข้อมูลชุดที่สองพบว่า DB-Index และ SDIndex ให้ประสิทธิภาพดีที่สุด และในชุดข้อมูลชุดสุดท้ายไม่สามารถสรุปได้เนื่องจากสภาพข้อมูลมีสิ่งรบกวนมากและลักษณะข้อมูลไม่เกาะกลุ่มกัน

งานวิจัยของศศิธร (ศศิธร, 2550) ได้ทำการทดลองจัดกลุ่มศูนย์บริการและถ่ายทอดเทคโนโลยีทางการเกษตรประจำตำบลตามการปฏิบัติงาน โดยใช้ (1) เทคนิคการทำเหมืองข้อมูลด้วยอัลกอริทึมแบบ 2 ขั้นตอน (ใช้ SOM ในการหาจำนวนกลุ่มและจัดกลุ่มโดยใช้อัลกอริทึม K-Means, อัลกอริทึม Bisecting K-Means และอัลกอริทึม Fuzzy C-Means และ (2) การจัดกลุ่มโดยใช้หลักการคำนวณค่าระดับคะแนนรวมประสิทธิภาพ (P Score) การทำงานของศูนย์ จากผลการทดลองพบว่าใน (1) อัลกอริทึม Fuzzy C-Means ให้ประสิทธิภาพในการจัดกลุ่มดีที่สุด สำหรับผลของ (2) ค่า P Score ให้ผลคะแนนส่วนใหญ่อยู่ที่ระดับพอใช้และปานกลาง และจำนวนกลุ่มที่เหมาะสมที่สุด คือ 6 กลุ่ม

งานวิจัยของวงกต (วงกต, 2551) ได้ทำการทดลองจัดกลุ่มข้อมูลเอกสารวิชาการความปลอดภัยของอาหารด้วยอัลกอริทึมแบบ 2 ขั้นตอน โดยใช้อัลกอริทึม SOM ในการหาจำนวนกลุ่มที่เหมาะสมและทำการจัดกลุ่มโดยใช้อัลกอริทึม K-Means และวิธีการจัดกลุ่มเชิงพันธุกรรม (Genetic Algorithm) ให้น้ำหนักคุณลักษณะโดยใช้เทคนิค TF-IDF จากการทดลองพบว่าวิธีการเชิงพันธุกรรมให้ผลการจัดกลุ่มดีกว่าอัลกอริทึม K-Means แต่ประสิทธิภาพเชิงเวลาอัลกอริทึม K-Means ดีกว่า และจำนวนกลุ่มที่เหมาะสมที่สุด คือ 8 กลุ่ม

งานวิจัยของวงศ์พันธ์และศรีวิหค (Wongpun และ Srivihok, 2008) ได้ศึกษาการหาเทคนิคการคัดเลือกคุณลักษณะที่ให้ประสิทธิภาพการจำแนกข้อมูลดีที่สุดนำมาหาปัจจัยที่มีผลต่อพฤติกรรมการกระทำความผิดของนักเรียนระดับอาชีวศึกษา โดยใช้เทคนิคการคัดเลือกคุณลักษณะเชิงพันธุกรรม (Genetic Search) ร่วมกับ Correlation-based Feature Selection (CFS), Consistency-based Subset Evaluation, Wrapper Subset Evaluation จากผลการทดลองพบว่า การคัดเลือกคุณลักษณะทั้ง 3 เทคนิค คือ Correlation-based Feature Selection, Consistency-based Subset Evaluation, Wrapper Subset Evaluation ก่อนนำไปประมวลผลต่อในตัวจำแนกประเภทเบย์อย่างง่าย, ข่ายงานความเชื่อแบบเบย์ และ C4.5 จะให้ประสิทธิภาพความถูกต้องในการจำแนกดีขึ้นกว่าการไม่คัดเลือกคุณลักษณะ โดยการคัดเลือกคุณลักษณะจากเทคนิค GA ร่วมกับ CFS ให้ประสิทธิภาพดีที่สุด

การทดลองของ Madahvi และคณะ (Madahvi *et al*., 2008) ได้ทำการทดลองจัดกลุ่มเอกสารเว็บ โดยการแทนเอกสารให้อยู่รูป VSM (Vector Space Model) ที่ผ่านการให้น้ำหนักด้วยเทคนิค TF-IDF ไว้แล้ว จัดกลุ่มด้วยอัลกอริทึม K-Means, อัลกอริทึม Harmony, อัลกอริทึม Hybrid Harmony K-Means และอัลกอริทึม Integrated K-Means in Harmony จากกการทดลองพบว่าอัลกอริทึม Integrated K-Means in Harmony ให้ประสิทธิภาพเชิงคุณภาพดีที่สุด และอัลกอริทึม K-Meansให้ประสิทธิภาพเชิงเวลาดีที่สุด

Zhang และคณะ (Zhang *et al.*, 2008) ได้นำเสนองานวิจัยโดยแบ่งการทดลองออกเป็นสองช่วง คือ การทดลองด้านการจำแนกข้อความ (Text Categorization) และงานด้านค้นคืนสารสนเทศ (Information Retrieval) โดยได้นำเอาเอกสารภาษาอังกฤษและภาษาจีนมาใช้ในการทดลองเปรียบเทียบสามเทคนิคได้แก่ TF-IDF เทคนิค Latent Semantic Indexing (LSI) และเทคนิค Multi-word ผลการทดลองด้านงานค้นคืนสารสนเทศสามารถสรุปได้ว่าเอกสารทั้งสองภาษาให้ผลดีที่สุดเมื่อใช้เทคนิค Multiword ส่วนผลการทดลองงานด้านการจำแนกข้อความ พบว่า TF-IDF ให้ผลดีที่สุดในเอกสารข้อความภาษาจีนและ Multiword ให้ผลดีที่สุดในเอกสารข้อความภาษาอังกฤษ จึงสามารถสรุปรวมได้ว่าการเลือกใช้เทคนิคต่าง ๆ อย่างเหมาะสมจะให้ผลดีในข้อมูลที่แตกต่างกัน

งานวิจัยที่เกี่ยวข้องดังที่นำเสนอไปแล้วข้างต้นได้สรุปในตารางที่ 1 สรุปงานวิจัยด้านการให้น้ำหนักคุณลักษณะ ในตารางที่ 2 เป็นการสรุปงานวิจัยด้านการจัดกลุ่ม ตารางที่ 3 สรุปงานวิจัย

ด้านการวัดประสิทธิภาพการจัดกลุ่มและในตารางที่ 4 สรุปงานวิจัยที่เกี่ยวข้องกับการคัดเลือกคุณลักษณะ

ตารางที่ 1 งานวิจัยด้านการให้น้ำหนักคุณลักษณะ

งานวิจัย	วิธีการ	ชุดข้อมูล	ผลการทดลอง
Zheng และคณะ 2003	จัดกลุ่มข่าวโดยใช้ MST (Minimum Spanning Tree) และให้น้ำหนักโดยใช้ TF-IDF เปรียบเทียบผลกับอัลกอริทึม K-Means	ชุดข้อมูลข่าวมาตรฐานหนังสือพิมพ์จีน CIRB010 ของ NTCIR-2	MST ที่ผ่านการให้น้ำหนักให้ประสิทธิภาพในการจัดกลุ่มดีที่สุด
Lan และคณะ 2005	ศึกษาเทคนิคการให้น้ำหนักคุณลักษณะสำหรับการจัดกลุ่มข้อความ เช่น เทคนิค Binary เทคนิค TF เทคนิค Logarithm Weight เป็นต้น	1. ข่าว 21,578 เรื่อง จากสำนักข่าว Reuters จำนวน 10 กลุ่ม 2. ข่าวจากแหล่งอื่นสุ่มมาจำนวน 20 กลุ่มข่าว กลุ่มละ 300 ตัวอย่าง	วิธีการที่ให้ประสิทธิภาพที่ดีที่สุด คือ TF.RF รองลงมา คือ TF วิธีการ Logarithm Weight และ Inverse Term Frequency ตามลำดับ
Zhang และคณะ 2008	ทำการทดลองงานด้านค้นคืน 2 งาน ได้แก่ งานการจำแนกข้อความและงานค้นคืนสารสนเทศในภาษาอังกฤษและภาษาจีน เปรียบเทียบการทำงาน 3 เทคนิค ได้แก่ TF-IDF, Latent Semantic Indexing (LSI) และ Multi-word	1. TanCorpV.1.0 สำหรับเอกสารข่าวภาษาจีนจำนวน 14,150 เอกสาร 2. Reuters สำหรับเอกสารข่าวภาษาอังกฤษจำนวน 21,578 เอกสาร	งานด้านการค้นคืนในภาษาอังกฤษและภาษาจีนให้ผลดีที่สุดเมื่อใช้เทคนิค Multiword สำหรับงานด้านจำแนกข้อความ TF-IDF ให้ผลดีที่สุดในเอกสารภาษาจีนและ Multi-word ให้ผลดีที่สุดในเอกสารภาษาอังกฤษ

ตารางที่ 2 งานวิจัยด้านการจัดกลุ่ม

งานวิจัย	วิธีการ	ชุดข้อมูล	ผลการทดลอง
ศศิธร 2550	จัดกลุ่มศูนย์บริการและถ่ายทอดเทคโนโลยีทางการเกษตรประจำตำบลตามการปฏิบัติงาน โดยใช้ (1) เทคนิคการทำเหมืองข้อมูลด้วยอัลกอริทึมแบบ 2 ขั้นตอน (ใช้ SOM ในการหาจำนวนกลุ่มและจัดกลุ่มโดยใช้อัลกอริทึม K-Means, อัลกอริทึม Bisecting K-Means และอัลกอริทึม Fuzzy C-Means และ (2) การจัดกลุ่มโดยใช้หลักการคำนวณค่าระดับคะแนนรวมประสิทธิภาพ (P Score) การทำงานของศูนย์	ข้อมูลของศูนย์บริการและถ่ายทอดเทคโนโลยีทางการเกษตรประจำตำบล ของประเทศไทย ประจำปี 2548	(1) อัลกอริทึม Fuzzy C-Means ที่จำนวนกลุ่ม 6 กลุ่ม ให้ประสิทธิภาพในการจัดกลุ่มดีที่สุด สำหรับผลของ (2) ค่า P Score ให้ผลคะแนนส่วนใหญ่อยู่ที่ระดับพอใช้และปานกลาง

ตารางที่ 2 (ต่อ)

งานวิจัย	วิธีการ	ชุดข้อมูล	ผลการทดลอง
Madahvi และคณะ 2008	จัดกลุ่มเอกสารเว็บด้วยอัลกอริทึม K-Means, อัลกอริทึม Harmony, อัลกอริทึม Hybrid Harmony K-Means และอัลกอริทึม Integrated K-Means in Harmony	1. ด้านการเมือง 176 เอกสาร 2. ข่าว 424 เอกสาร 3. ชุดเอกสารรวมหลายด้านจาก DMOZ จำนวน 697 เอกสาร	อัลกอริทึม Integrated K-Means in Harmony ให้ประสิทธิภาพเชิงคุณภาพดีที่สุด และอัลกอริทึม K-Means ให้ประสิทธิภาพเชิงเวลาดีที่สุด
วงกต 2551	นำเสนอวิธีการจัดกลุ่มด้านความปลอดภัยของอาหารโดยใช้วิธีการ SOM ในการหาจำนวนกลุ่มและเปรียบเทียบการจัดกลุ่มโดยใช้อัลกอริทึม K-Means และอัลกอริทึมเชิงพันธุกรรม (Genetic Algorithm) และให้น้ำหนักคุณลักษณะโดยใช้ TF-IDF	เอกสารด้านความปลอดภัยของอาหารที่มีการกำหนดคำสำคัญไว้แล้ว จำนวน 4,806 เอกสาร	จำนวนกลุ่มที่เหมาะสมจากวิธีการ SOM คือ 8 กลุ่ม และผลการจัดกลุ่มพบว่าอัลกอริทึมเชิงพันธุกรรมให้ผลในการจัดกลุ่มดีกว่าอัลกอริทึม K-Means

ตารางที่ 3 งานวิจัยด้านการวัดประสิทธิภาพการจัดกลุ่ม

งานวิจัย	วิธีการ	ชุดข้อมูล	ผลการทดลอง
Kovacs และคณะ 2006	นำเสนอเทคนิคการวัดประสิทธิภาพการจัดกลุ่ม ได้แก่ Davies Bouldin index, Root Mean Squared Standard Deviation (RMSSDT), Dunn and Dunn like indices, R-Squared, SD Validity index (SD-Index) และ S_Dbw Validity index ทดลองจัดกลุ่มโดยใช้อัลกอริทึม K-Means	เซตข้อมูลที่มีรูปร่างต่างกัน ได้แก่ (1) ข้อมูลที่มีการเกาะเป็นกลุ่มก้อนกันชัดเจน (2) ข้อมูลที่เกาะกลุ่มเป็นรูปวงแหวนและ (3) ข้อมูลที่เกาะกลุ่มกันแบบกระจัดกระจายมาก	ชุดข้อมูลชุดแรกพบว่า Dunn และ SD-Index ให้ประสิทธิภาพดีที่สุด ข้อมูลชุดที่สองพบว่า DB-Index และ SD-Index ให้ประสิทธิภาพดีที่สุด และในชุดข้อมูลชุดสุดท้ายไม่สามารถสรุปได้เนื่องจากสภาพข้อมูลมีสิ่งรบกวนมาก ไม่เกาะกลุ่มกัน

ตารางที่ 4 งานวิจัยด้านการคัดเลือกคุณลักษณะ

งานวิจัย	วิธีการ	ชุดข้อมูล	ผลการทดลอง
Wang และคณะ 2004	คัดเลือกคุณลักษณะยีนของการเกิดโรคมะเร็ง ซึ่งโดยใช้วิธีการ Wrapper, filters และ CFS ร่วมกับอัลกอริทึม J48, Naïve Bayes และ Suport Vector Machines	1. Acute leukemia 2. Diffuse large B-cell lymphoma	วิธีการ Wrapper สามารถเลือกยีนเพื่อใช้ในการสร้างโมเดลในการจำแนกดีที่สุดและวิธีการ CFS สามารถเลือกยีนที่มีความสัมพันธ์เชื่อมโยงกับการเกิดโรคได้ดีที่สุด นอกจากนี้วิธีการ CFS ยังเป็นวิธีการที่ใช้เวลาในการทำงานเร็วกว่าวิธีการ Wrapper
Aarabi และคณะ 2005	จำแนกคลื่นสมองเพื่อจับอาการปัจจุบันของโรคในทารกแรกเกิด โดยกำหนดคุณลักษณะให้อยู่ในรูปแบบเวคเตอร์ด้วยวิธีการโครงข่ายประสาทเทียมเปรียบเทียบระหว่างข้อมูลที่ผ่านคัดเลือกคุณลักษณะด้วย CFS, RelifF, FCBF และข้อมูลที่ไม่ผ่านการคัดเลือกคุณลักษณะ	คลื่นสมองทารกแรกเกิดที่มีอายุระหว่าง 32-42 สัปดาห์ จำนวน 11 คนของโรงพยาบาล North Hospital of Amiens สกัดเวคเตอร์เพื่อกำหนดเป็นคุณลักษณะจำนวน 132 คุณลักษณะ	การจำแนกข้อมูลด้วยวิธีการโครงข่ายประสาทเทียมและคัดเลือกคุณลักษณะด้วยวิธีการ CFS และ RelifF ช่วยเพิ่มประสิทธิภาพในการจำแนกคลื่นสมองดีที่สุดแต่คุณลักษณะที่คัดเลือกจาก CFS มีความสัมพันธ์ของคุณลักษณะดีที่สุด

ตารางที่ 4 (ต่อ)

งานวิจัย	วิธีการ	ชุดข้อมูล	ผลการทดลอง
Wongpun และ Srivihok 2008	หาเทคนิคการคัดเลือกคุณลักษณะที่ให้ประสิทธิภาพการจำแนกข้อมูลดีที่สุดนำมาหาปัจจัยที่มีผลต่อพฤติกรรมการกระทำความผิดของนักเรียนระดับอาชีวศึกษา โดยเทคนิคการคัดเลือกคุณลักษณะเชิงพันธุกรรม (Genetic Search) ร่วมกับ Correlation-based Feature Selection (CFS), Consistency-based Subset Evaluation, Wrapper Subset Evaluation และจำแนกผ่านวิธีการเบย์อย่างง่าย (Naive Bayes classifier), ข่ายงานความเชื่อแบบเบย์ (Baysian belief network) และ C4.5	ข้อมูลประวัตินักเรียนอาชีวศึกษา	ในการคัดเลือกคุณลักษณะที่เหมาะสมที่สุด คือวิธีการ CFS ร่วมกับ GA และจำแนกข้อมูลผ่านวิธีการ C4.5 ให้ประสิทธิภาพดีที่สุด

อุปกรณ์และวิธีการ

อุปกรณ์

ที่ใช้ในการศึกษามีดังนี้ คือ

1. ฮาร์ดแวร์ (Hardware) ประกอบด้วย

   คอมพิวเตอร์ Laptop Toshiba Satellite L310 CPU Pentium Dual Core 1.86 GHz, 3072 MB DDR2 SDRAM, ฮาร์ดดิสก์ 200 GB

2. ซอฟต์แวร์ (Software) ประกอบด้วย

   2.1 โปรแกรม Matlab 7.0.1 สำหรับการจัดกลุ่มและวัดประสิทธิภาพการจัดกลุ่ม
   2.2 โปรแกรม Weka 3.5.8 สำหรับการคัดเลือกคุณลักษณะ

3. ระบบปฏิบัติการ (Operating System) ที่นำมาใช้ คือ

   Microsoft Window XP Service Pack 2

วิธีการ

1. ชุดข้อมูลสำหรับการทดลอง ข้อมูลที่ใช้ในงานวิจัยนี้มี 3 แหล่ง ได้แก่

1.1 ฐานข้อมูลบุคลากรหรือผู้เชี่ยวชาญ (Personnel Database) ของมหาวิทยาลัยซึ่งได้จากกองการเจ้าหน้าที่จำนวน 3,194 ระเบียน ซึ่งประกอบด้วย รหัสนักวิจัย (ID), รหัสบัตรประชาชน, คำนำหน้านาม (Prefix), ชื่อ (Name), นามสกุล (Surname),วันเกิด, เดือนเกิด, ปีเกิด (พ.ศ.), วันที่เริ่มทำงาน, เดือนที่เริ่มทำงาน, ปีที่เริ่มทำงาน (พ.ศ.), ตำแหน่งทางวิชาการ (Position), ภาควิชา (Department) , คณะ(Faculty), วิทยาเขต (Campus) และประเภทพนักงาน (Officer Class) ดังตารางที่ 5

ตารางที่ 5 ตัวอย่างข้อมูลนักวิจัยจากฐานข้อมูลบุคลากร

รหัสนักวิจัย	**รหัสบัตรประชาชน**	**คำนำหน้า**	**ชื่อ**	**นามสกุล**	**ปีที่เริ่มทำงาน**	**ภาควิชา**	**...**
1	1234567891234	นาย	สมศักดิ์	ใจดี	2536	ฝ่ายโภชนาการและสุขภาพ	...
2	4567893453334	น.ส.	สมศรี	เชิดชู	2529	ฝ่ายจุลชีววิทยาประยุกต์	...
...	...	...	...	...	...	...	...

1.2 ฐานข้อมูลผลงานวิจัย (Research Database) จากสถาบันวิจัยและพัฒนามหาวิทยาลัย จำนวน 3,704 ระเบียน ซึ่งประกอบด้วย รหัสบัตรประชาชน, ชื่อ (ภาษาไทย), นามสกุล(ภาษาไทย), ชื่อ (ภาษาอังกฤษ), นามสกุล(ภาษาอังกฤษ), ชื่องานวิจัยภาษาไทย (Research Name), ชื่องานวิจัยภาษาอังกฤษ, ชื่อแหล่งทุน (Scholarship Name), ประเภทแหล่งทุน (Scholarship Class), จำนวนเงินทุน (Amount), ระยะเวลาในการทำงานและคำสำคัญ (Keywords) ดังตารางที่ 6

ตารางที่ 6 ตัวอย่างข้อมูลผลงานวิจัยของนักวิจัยจากฐานข้อมูลผลงานวิจัย

เลขประจำตัว	ชื่อ	นามสกุล	ชื่อโครงการ (ไทย)	ชื่อโครงการ (Eng)	คำสำคัญ	...
3456238900023	นายหมาย	ยิ่งยง	การพัฒนาคุณภาพและผลิตภัณฑ์ผ้าไหม	Quality Improvement of Silk Fabric	ผ้าไหม	...
9908765465431	น.ส. จริง	ใจงาม	การศึกษาความปลอดภัยด้านอาหารของประเทศไทยในภาพรวม	Thailand Food Safety Policy Study	ความปลอดภัย, safety, food	...
4535744464200	นายศักดิ์	อิ่มเอม	การใช้สารสกัดเบต้ากลูแคนจากยีสต์ในการกระตุ้นภูมิคุ้มกันของกุ้งขาวแวนนาไม	-	เบต้ากลูแคน, ยีสต์, แวนนาไม, กุ้ง	...
...	...	...	...	...		...

1.3 ฐานข้อมูลวิทยานิพนธ์ของบัณฑิตวิทยาลัย (Thesis Database) โดยส่วนที่นำมาใช้ คือ ข้อมูลอาจารย์ที่ปรึกษา จำนวน 194 ระเบียน ซึ่งประกอบด้วย ภาควิชา, รหัสอาจารย์, ชื่อและนามสกุล, รหัสสาขา, สาขาวิชา, เชี่ยวชาญ, สนใจ ดังตารางที่ 7

ตารางที่ 7 ตัวอย่างข้อมูลวิทยานิพนธ์ของบัณฑิตวิทยาลัย

ภาควิชา	รหัสอาจารย์	ชื่ออาจารย์	รหัสสาขา	สาขาวิชา	เชี่ยวชาญ	สนใจ
คณิตศาสตร์	D1020	รศ.สมศักดิ์ ใจดี	XD01	คณิตศาสตร์	คณิตศาสตร์	พีชคณิต
จุลชีววิทยา	D3113	รศ.ดร.สมศรี เชิด	XD06	จุลชีววิทยา	Enzyme Technology	Protein Engineer
...	...	...	...	...	...	...

ต่อมาได้มีการแบ่งหมวดหมู่ผู้เชี่ยวชาญตามหน่วยงานที่สังกัดตามสายงาน โดยพิจารณาจากสาขาที่เกี่ยวข้องกันได้ 9 กลุ่มสายงาน ได้แก่

1. กลุ่มสายงานวิทยาศาสตร์ ผู้เชี่ยวชาญจำนวน 432 คน คิดเป็น 13.53% ประกอบด้วย
   1.1 คณะวิทยาศาสตร์ วิทยาเขตบางเขน
   1.2 คณะศิลปะศาสตร์และวิทยาศาสตร์ วิทยาเขตกำแพงแสน
   1.3 คณะทรัพยากรและสิ่งแวดล้อม วิทยาเขตศรีราชา
   1.4 วิทยาลัยสิ่งแวดล้อม วิทยาเขตบางเขน

2. กลุ่มสายงานวิศวกรรม ผู้เชี่ยวชาญจำนวน 480 คน คิดเป็น 15.03% ประกอบด้วย
   2.1 คณะวิทยาศาสตร์และวิศวกรรมศาสตร์ วิทยาเขตสกลนคร
   2.2 คณะวิศวกรรมศาสตร์ วิทยาเขตกำแพงแสน
   2.3 คณะวิศวกรรมศาสตร์ วิทยาเขตบางเขน
   2.4 คณะวิศวกรรมศาสตร์ศรีราชา วิทยาเขตศรีราชา
   2.5 คณะสถาปัตยกรรมศาสตร์ วิทยาเขตบางเขน
   2.6 วิทยาลัยพาณิชยนาวีนานาชาติ วิทยาเขตศรีราชา

3. กลุ่มสายงานเกษตร อุตสาหกรรมเกษตร ผู้เชี่ยวชาญจำนวน 642 คน คิดเป็น 20.1% ประกอบด้วย

3.1 คณะเกษตรบางเขน วิทยาเขตบางเขน

3.2 คณะเกษตรกำแพงแสน วิทยาเขตกำแพงแสน

3.3 คณะทรัพยากรธรรมชาติและอุตสาหกรรมเกษตร วิทยาเขตสกลนคร

3.4 คณะประมง วิทยาเขตบางเขน

3.5 คณะวนศาสตร์ วิทยาเขตบางเขน

3.6 คณะอุตสาหกรรมเกษตร วิทยาเขตบางเขน

4. กลุ่มสายงานบริหาร เศรษฐศาสตร์ ผู้เชี่ยวชาญจำนวน 231 คน คิดเป็น 7.23% ประกอบด้วย

4.1 คณะวิทยาการจัดการ วิทยาเขตศรีราชา

4.2 คณะศิลปะศาสตร์และวิทยาการจัดการ วิทยาเขตสกลนคร

4.3 คณะบริหารธุรกิจ วิทยาเขตบางเขน

4.4 คณะเศรษฐศาสตร์ วิทยาเขตบางเขน

5. กลุ่มสายงานสังคมศาสตร์ ผู้เชี่ยวชาญจำนวน 220 คน คิดเป็น 6.89% ประกอบด้วย

5.1 คณะมนุษยศาสตร์ วิทยาเขตบางเขน

5.2 คณะสังคมศาสตร์ วิทยาเขตบางเขน

6. กลุ่มสายงานศึกษาศาสตร์ ผู้เชี่ยวชาญจำนวน 441 คน คิดเป็น 13.81% ประกอบด้วย

6.1 คณะวิทยาศาสตร์การกีฬา วิทยาเขตกำแพงแสน

6.2 คณะศึกษาศาสตร์ วิทยาเขตบางเขน

6.3 คณะศึกษาศาสตร์และพัฒนศาสตร์ วิทยาเขตกำแพงแสน

7. กลุ่มสายงานสัตว์แพทย์ศาสตร์ ผู้เชี่ยวชาญจำนวน 170 คน คิดเป็น 5.32% ประกอบด้วย

7.1 คณะเทคนิคการสัตวแพทย์ วิทยาเขตบางเขน

7.2 คณะสัตวแพทยศาสตร์ วิทยาเขตกำแพงแสน

7.3 คณะสัตวแพทยศาสตร์ วิทยาเขตบางเขน

8. กลุ่มสถาบันวิจัย ผู้เชี่ยวชาญจำนวน 258 คน คิดเป็น 8.08% ประกอบด้วย

8.1 ศูนย์นานาชาติสิรินธรเพื่อการวิจัยพัฒนาและถ่ายทอดเทคโนโลยี วิทยาเขตบางเขน

8.2 สถาบันค้นคว้าและพัฒนาผลิตผลทางการเกษตรและอุตสาหกรรมเกษตร วิทยาเขตบางเขน

8.3 สถาบันค้นคว้าและพัฒนาผลิตภัณฑ์อาหาร วิทยาเขตบางเขน

8.4 สถาบันค้นคว้าและพัฒนาระบบนิเวศเกษตร วิทยาเขตบางเขน

8.5 สถาบันภาษาศาสตร์และวัฒนธรรมศึกษาราชนครินทร์ วิทยาเขตบางเขน

8.6 สถาบันวิจัยและพัฒนากำแพงแสน วิทยาเขตกำแพงแสน

8.7 สถาบันวิจัยและพัฒนาวิทยาเขตเฉลิมพระเกียรติจังหวัดสกลนคร วิทยาเขตสกลนคร

8.8 สถาบันวิจัยและพัฒนาแห่งมหาวิทยาลัย วิทยาเขตกำแพงแสน

8.9 สถาบันวิจัยและพัฒนาแห่งมหาวิทยาลัย วิทยาเขตบางเขน

8.10 สถาบันวิชาการด้านสหกรณ์ วิทยาเขตบางเขน

8.11 สถาบันสุวรรณวาจกกสิกิจเพื่อการค้นคว้าและพัฒนาปศุสัตว์และผลิตภัณฑ์สัตว์ วิทยาเขตกำแพงแสน

8.12 สถาบันอินทรียจันทรสถิตย์เพื่อการค้นคว้าและพัฒนาพืชศาสตร์ วิทยาเขตบางเขน

9. กลุ่มสำนักงานภายในมหาวิทยาลัย ผู้เชี่ยวชาญจำนวน 320 คน คิดเป็น 10.02% ประกอบด้วย

9.1 บัณฑิตวิทยาลัย วิทยาเขตบางเขน

9.2 วิทยาลัยบัณฑิตศึกษาศรีราชา วิทยาเขตศรีราชา

9.3 สำนักงานโครงการจัดตั้งวิทยาเขตลพบุรี วิทยาเขตบางเขน

9.4 สำนักงานโครงการจัดตั้งวิทยาเขตสุพรรณบุรี วิทยาเขตบางเขน

9.5 สำนักงานวิทยาเขต วิทยาเขตกำแพงแสน

9.6 สำนักงานวิทยาเขตเฉลิมพระเกียรติจังหวัดสกลนคร วิทยาเขตสกลนคร

9.7 สำนักงานวิทยาเขตศรีราชา วิทยาเขตศรีราชา

9.8 สำนักงานอธิการบดี วิทยาเขตบางเขน

9.9 สำนักทะเบียนและประมวลผล วิทยาเขตบางเขน

9.10 สำนักบริการคอมพิวเตอร์ วิทยาเขตบางเขน

9.11 สำนักพิพิธภัณฑ์และวัฒนธรรมการเกษตร วิทยาเขตบางเขน

9.12 สำนักวิทยบริการ วิทยาเขตสกลนคร

9.13 สำนักส่งเสริมและฝึกอบรม วิทยาเขตกำแพงแสน

9.14 สำนักส่งเสริมและฝึกอบรม วิทยาเขตบางเขน

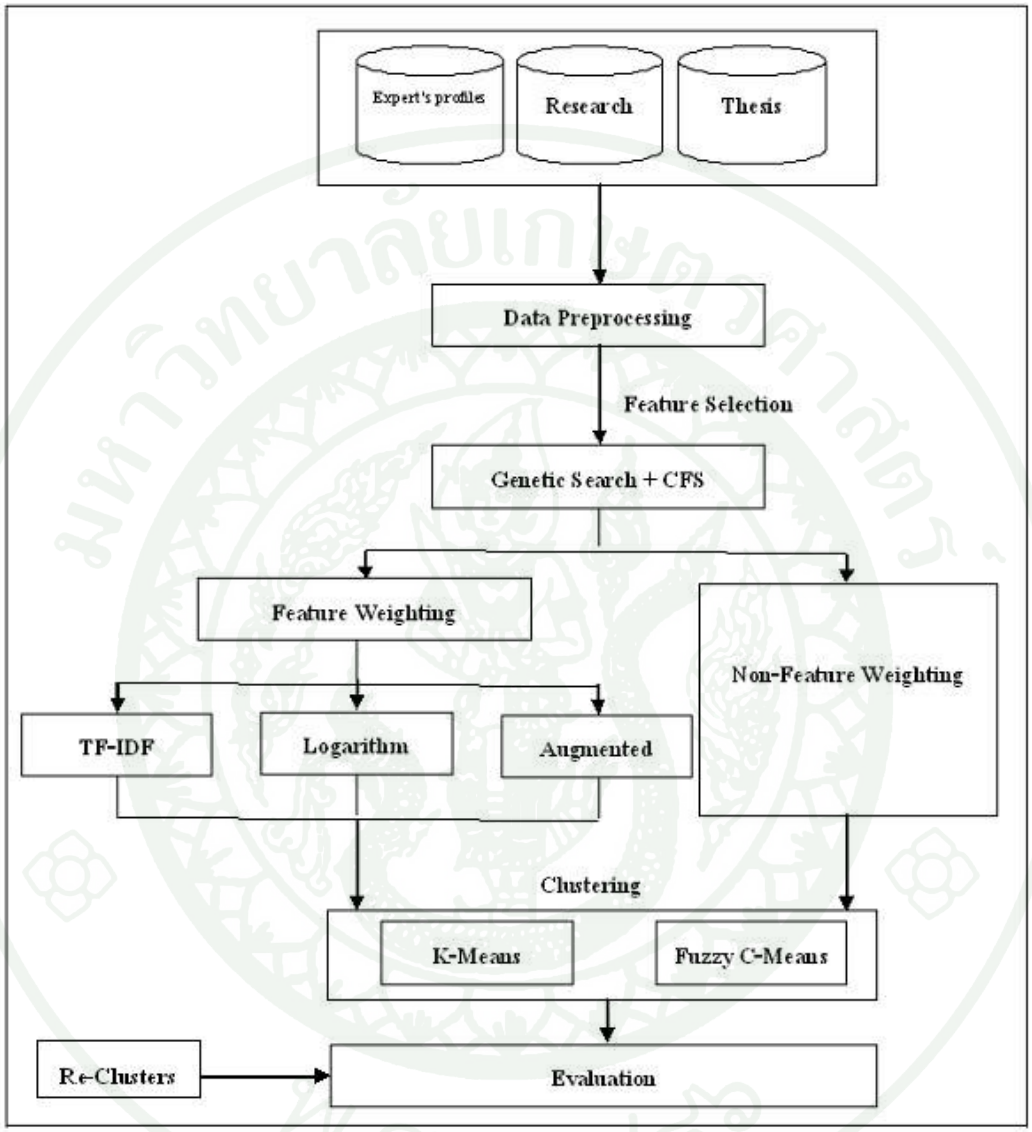
9.15 สำนักหอสมุด วิทยาเขตบางเขน

9.16 สำนักหอสมุดกำแพงแสน วิทยาเขตกำแพงแสน

รวมจำนวนผู้เชี่ยวชาญ 3,194 คน โดยข้อมูลเริ่มต้นประกอบด้วย รหัสประจำตัวของผู้เชี่ยวชาญ ชื่อผู้เชี่ยวชาญ สังกัดที่ผู้เชี่ยวชาญคนนั้น ๆ สังกัด ชื่องานวิจัยที่ผู้เชี่ยวชาญคนนั้น ๆ รับผิดชอบ และคำสำคัญของงานวิจัย ดังที่ได้แสดงไว้ในตารางที่ 6 ตารางที่ 7 และตารางที่ 8

2. ขั้นตอนการทดลอง

สำหรับขั้นตอนการทดลองในงานวิจัยนี้ได้สรุปดังภาพที่ 9



ภาพที่ 9 แสดงขั้นตอนการทำงานของงานวิจัยการจัดกลุ่มผู้เชี่ยวชาญ

2.1 การเตรียมข้อมูล (Data Preprocessing)

เป็นขั้นตอนการกลั่นกรองข้อมูลที่ใช้ในการทดลอง โดยชุดข้อมูลที่ได้ตามข้อ 1 มีการกำหนดคำสำคัญ (Keywords) ไว้แล้วโดยผู้เชี่ยวชาญของงานวิจัยนั้น ๆ นำคำสำคัญนี้ไปกำหนดให้

เป็นคุณลักษณะของข้อมูล ในงานวิจัยนี้สามารถสกัดคำสำคัญออกมาได้ 971 คำ ดังนั้นข้อมูลชุดนี้จึงมี 3,194 แถว 971 คุณลักษณะ โดยคุณลักษณะทั้ง 971 คุณลักษณะ ได้แสดงไว้ในตารางผนวกที่ 1

ในขั้นตอนต่อไปจะทำการนับความถี่ของคำที่เกิดขึ้นซ้ำ ๆ กันในข้อมูลแต่ละแถว เพื่อเข้าสู่กระบวนการให้ค่าน้ำหนักคุณลักษณะต่อไป ในตารางที่ 8 เป็นตัวอย่างคำสำคัญที่ผ่านการเข้ากระบวนการนับความถี่คำสำคัญแล้ว เพื่อให้ง่ายต่อการอ้างอิงจึงมีการกำหนดรหัสให้กับคำสำคัญ

ตารางที่ 8 ตัวอย่างคำสำคัญของนักวิจัยที่มีการนับความถี่และรหัสอ้างอิงคำสำคัญ

รหัสพนักงาน	รหัสคำสำคัญ	ความถี่
1	227	1
1	119	5
...	...	...

หมายเหตุ รหัสนักวิจัยซึ่งกำหนดขึ้นมาเองและสามารถอ้างอิงไปยังข้อมูลนักวิจัยได้
รหัสคำสำคัญซึ่งกำหนดขึ้นมาเองและสามารถอ้างอิงไปยังคำสำคัญได้

ในตารางที่ 9 เป็นตัวอย่างข้อมูลที่ผ่านการสกัดคำสำคัญและนับความถี่จนเป็นข้อมูลที่พร้อมใช้งาน คือ เป็นข้อมูลที่พร้อมเข้าสู่กระบวนการทำเหมืองข้อมูลต่อไป

2.2 การคัดเลือกคุณลักษณะ (Features Selection)

การคัดเลือกคุณลักษณะเป็นขั้นตอนหนึ่งที่นิยมใช้ในการทำเหมืองข้อมูล เนื่องจากข้อมูลที่ใช้ในการทดลองอาจมีขนาดของมิติที่ใหญ่หรือคุณลักษณะที่มีจำนวนมากจนเกินไป อันเป็นสาเหตุให้เกิดการใช้ทรัพยากร เช่น หน่วยความจำ มากเกินความจำเป็น ที่สำคัญการคัดเลือกคุณลักษณะอาจสามารถยกระดับประสิทธิภาพในการทำเหมืองข้อมูลให้ดีกว่าการทำเหมืองข้อมูลที่ไม่ผ่านกระบวนการคัดเลือกคุณลักษณะ กระบวนการคัดเลือกคุณลักษณะนี้จึงเป็นขั้นตอนสำคัญที่สามารถทำให้ประสิทธิภาพในการเหมืองข้อมูลดีขึ้นและช่วยลดการใช้ทรัพยากรลง

ตารางที่ 9 ตัวอย่างเลขประจำตัวนักวิจัยและความถี่ของคำสำคัญที่พร้อมเข้าสู่กระบวนการทำเหมืองข้อมูล

	ID_Keyword			
IDExp	**1**	**2**	**3**	**...**
1	0	0	1	...
2	4	0	3	...
3	2	1	7	...
...	...	...	...	...

หมายเหตุ IDExp คือ รหัสนักวิจัยซึ่งกำหนดขึ้นในการทดสอบและสามารถอ้างอิงไปยังข้อมูลนักวิจัยได้

ID_Keyword คือ รหัสคำสำคัญซึ่งกำหนดขึ้นในการทดสอบและสามารถอ้างอิงไปยังคำสำคัญได้

ในงานวิจัยนี้ได้นำเทคนิคในการคัดเลือกคุณลักษณะมาใช้เนื่องจากการคัดเลือกคุณลักษณะสามารถลดการใช้ทรัพยากรลงและคาดหวังได้ว่าจะสามารถให้ประสิทธิภาพในการจัดกลุ่มที่ดีขึ้น เพราะข้อมูลที่ใช้ในการทดลองนี้มีขนาดใหญ่ โดยมีจำนวน 3,194 แถวและ 971 คุณลักษณะ เครื่องมือที่ใช้ในการคัดเลือกคุณลักษณะ คือ Weka 3.5.8 ซึ่งเป็น Open Source ซอฟต์แวร์ที่พัฒนาโดย The University of Waikato ใช้งานง่าย เป็นที่นิยมและมีความน่าเชื่อถือ สำหรับอัลกอริทึมที่ใช้ในการประเมินคุณลักษณะ คือ Correlation-based Feature Subset Selection และวิธีการหาคุณลักษณะที่เหมาะสมโดยใช้วิธีการเชิงพันธุกรรม (Genetic Search) ซึ่งได้อธิบายวิธีการไว้ในข้อ 5 หน้า 19 เนื่องจากผลงานวิจัยที่ผ่านมา (วงศ์พันธ์ และ ศรีวิหค, 2008) ได้นำเสนอว่าวิธีการ CFS และ GA เป็นวิธีการที่มีประสิทธิภาพในการคัดเลือกชุดข้อมูลของนักศึกษาระดับอาชีวศึกษา สำหรับในชุดข้อมูลชุดนี้หากใช้วิธีการอื่นจะได้จำนวนคุณลักษณะที่น้อยเกินไปจนไม่สามารถพบความสัมพันธ์ของคุณลักษณะได้สมเหตุสมผล

2.3 การหาจำนวนกลุ่มที่เหมาะสม

เนื่องจากข้อมูลที่ใช้ในการจัดกลุ่มเป็นข้อมูลที่ไม่มีผลเฉลย สิ่งที่ได้จากข้อมูลในเบื้องต้นก่อนเริ่มต้นกระบวนการจัดกลุ่ม คือ ข้อมูลดิบที่ผ่านการกลั่นกรองรอการจัดกลุ่มต่อไป ดังนั้นปัญหาหลักในการจัดกลุ่ม คือ การหาจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่ม

วิธีการในการจัดกลุ่มแบบ 2 ขั้นตอนจึงเป็นทางเลือกหนึ่งในการหาจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่ม กล่าวคือ ในขั้นแรกจะทำการหาจำนวนกลุ่มที่เหมาะสมของนักวิจัย โดยการเลือกอัลกอริทึมการจัดกลุ่มมาหนึ่งอัลกอริทึมและทำการประมวลผลไปเรื่อย ๆ จาก 2 กลุ่มไปจนถึงจำนวนกลุ่มตามที่ต้องการแล้วทำการวัดประสิทธิภาพในการจัดกลุ่ม อัลกอริทึมดังกล่าวให้ประสิทธิภาพในการจัดกลุ่มดีที่สุดที่จำนวนกลุ่มเท่าใดจะยึดถือให้เป็นจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มในกรณีนั้น ๆ

เนื่องจากอัลกอริทึม K-Means เป็นอัลกอริทึมที่มีประสิทธิภาพในการจัดกลุ่มข้อมูลที่ดี แต่การทำงานของอัลกอริทึมนี้ต้องมีการกำหนดจำนวนกลุ่มที่เหมาะสมก่อนการทำงาน ดังนั้นงานวิจัยนี้จึงใช้อัลกอริทึม SOM ในการหาจำนวนกลุ่ม อัลกอริทึม SOM เป็นอัลกอริทึมที่สามารถทำงานได้ดีกับข้อมูลที่มีขนาดใหญ่ เนื่องจากอัลกอริทึม SOM เป็นหนึ่งในอัลกอริทึมประเภทวิเคราะห์โดยใช้แบบจำลอง ซึ่งจุดเด่นของอัลกอริทึมประเภทวิเคราะห์โดยใช้แบบจำลอง คือ นิยมใช้ในการหาจำนวนกลุ่มที่เหมาะสม *แต่ทำไมไม่ใช้อัลกอริทึม COWEB ทั้งที่เป็นอัลกอริทึมกลุ่มเดียวกับอัลกอริทึม SOM* ยังไม่ผลงานวิจัยใดยืนยันว่าอัลกอริทึม COWEB สามารถคำนวณหาจำนวนกลุ่มได้ดีกว่า SOM แต่มีงานวิจัยที่ผ่านมา (วงกต, 2551; ศศิธร, 2550) ได้มีการนำอัลกอริทึม SOM มาใช้ในการหาจำนวนกลุ่ม ดังนั้นในกรณีศึกษานี้จึงนำอัลกอริทึม SOM มาใช้ซึ่งสามารถเข้าใจรูปแบบการทำงานและกำหนดค่าต่าง ๆ ที่จำเป็นในการทำงานได้ง่ายกว่าอัลกอริทึม COWEB ในขั้นตอนการหาจำนวนกลุ่มนี้จะประมวลผลจากจำนวน 2 กลุ่มถึง 10 กลุ่ม แล้วประเมินหาผลการจัดกลุ่มที่ดีที่สุด สำหรับการประเมินผลจะอธิบายในข้อ 2.6

สิ่งที่น่าสังเกตในการหาจำนวนกลุ่ม คือ *ทำไมถึงจำเป็นต้องหาจำนวนกลุ่มที่เหมาะสมก่อน* เนื่องจาก SOM เป็นอัลกอริทึมที่ค่อนข้างเที่ยงตรงกับข้อมูลที่มีขนาดใหญ่ ไม่ว่าจะประมวลผลกี่ครั้งก็จะได้ผลเท่าเดิม ดังนั้นจึงควรเลือกหาอัลกอริทึมที่เที่ยงกับการประมวลผลมาหา

จำนวนกลุ่มก่อนที่จะศึกษาแนวทางในการพัฒนางานทดลองต่อ ในเมื่อเป็นเช่นนี้แล้ว *ทำไมไม่นำผลที่ได้จาก SOM มาเป็นผลในการจัดกลุ่มเลย* เนื่องจากได้มีงานวิจัยก่อนหน้านี้รายงานว่าการหาจำนวนกลุ่มโดย SOM ร่วมกับอัลกอริทึมด้านการจัดกลุ่มอื่น ๆ เช่น K-Means และ Fuzzy C-Means (วงกต, 2551; ศศิธร, 2550) ให้ผลที่ดีกว่าการไม่นำอัลกอริทึมมาใช้ร่วมกัน

2.4 การให้น้ำหนักของคุณลักษณะ (Features Weight)

การให้น้ำหนักคุณลักษณะเป็นเทคนิคหนึ่งที่นิยมใช้ในการทำเหมืองข้อมูลและค้นคืนสารสนเทศเพื่อเพิ่มประสิทธิภาพในการค้นคืนและทำเหมืองข้อมูล

สำหรับขั้นตอนนี้ได้นำเอาเทคนิค TF-IDF เทคนิค Logarithm Weight และเทคนิค Augmented Weight มาใช้ในการให้น้ำหนักคุณลักษณะ ซึ่งใช้หลักความถี่ของคำและเอกสารที่สนใจ เนื่องวิธีการ TF-IDF เป็นวิธีการที่ง่าย นิยมใช้กันมากและให้ประสิทธิภาพที่ดีวิธีการหนึ่ง จึงได้นำมาเป็นมาใช้ในการให้น้ำหนักคุณลักษณะในขั้นต้น และนำวิธีการ Logarithm Weight และ Augmented Weight ซึ่งเป็นวิธีการที่ต่อยอดจากวิธีการ TF-IDF มาเปรียบเทียบประสิทธิภาพเพื่อหาผลการจัดกลุ่มที่ดีที่สุด โดยวิธีการ Augmented Weight เป็นวิธีการที่ค่อนข้างน่าเชื่อถือมาก เนื่องจากมีการพิจารณาความยาวของเอกสาร สำหรับวิธีการให้น้ำหนักด้วยเทคนิคทั้งสามได้อธิบายไว้ในข้อ 4 หน้า 16

2.5 การจัดกลุ่ม (Clustering)

ขั้นตอนนี้ถือว่าเป็นขั้นตอนที่สำคัญมากขั้นตอนหนึ่ง เนื่องจากเป็นขั้นตอนในการขุดค้นความรู้จากข้อมูลที่มีอยู่เพื่อนำความรู้ที่ได้จากการทำเหมืองข้อมูลหรือขุดค้นข้อมูลนั้นไปใช้ประโยชน์ในองค์กร สำหรับอัลกอริทึมที่ใช้ในการจัดกลุ่มในปัจจุบันนั้นมีอยู่มาก ในงานวิจัยนี้จะเลือกใช้วิธีการแบบแบ่งส่วน คือ อัลกอริทึม K-Means และอัลกอริทึม Fuzzy C-Means เนื่องจากเป็นวิธีการที่ง่าย ไม่ซับซ้อนมากนัก และสามารถใช้กับข้อมูลที่หลากหลายประเภทได้ เช่น ข้อมูลที่เป็นตัวเลข ในการทดลองจะเปรียบเทียบกับผลการทดลองระหว่างการจัดกลุ่มแบบให้น้ำหนักและการจัดกลุ่มไม่ให้น้ำหนักว่าวิธีการใดให้ผลที่ดีที่สุด โดยอาศัยตัวประเมินผลการจัดกลุ่มที่จะอธิบายในข้อ 2.6 ประกอบการตัดสินใจผลการทดลอง

2.6 การประเมินผล

ในขั้นตอนนี้เป็นขั้นตอนที่ใช้ในการประกอบการตัดสินว่าจำนวนกลุ่มเท่าใดหรือผลการจัดกลุ่มวิธีการใดให้ประสิทธิภาพที่ดีที่สุด โดยตัวประเมินที่ใช้ในงานวิจัยนี้ได้แก่ F-Statistic Silhouette และ R-Squared (RS) ผลการประเมินที่มีประสิทธิภาพ ตัวประเมินทั้งสามตัวจะต้องมีค่ามากที่สุด เมื่อได้ผลการจัดกลุ่มที่มีประสิทธิภาพแล้ว ในขั้นตอนต่อไปจะสกัดคำสำคัญหรือความเชี่ยวชาญในกลุ่มโดยใช้ค่าความถี่คำสำคัญในการพิจารณา

สิ่งที่น่าสังเกตในขั้นตอนนี้ คือ *ทำไมไม่ใช้ค่า IDF ในการสกัดคำสำคัญ* เนื่องจากพิจารณาแล้วพบว่าค่า IDF มีค่าที่ซ้ำกันมากหลายคำตั้งแต่อันดับต้น ๆ ซึ่งเพิ่มโอกาสให้คำสำคัญที่สกัดได้มีความขัดแย้งกันภายในกลุ่มมากกว่าการพิจารณาจากความถี่คำ จึงใช้วิธีการพิจารณาจากความถี่สูงสุดแทน

2.7 การจัดกลุ่มครั้งที่ 2 (Re-Clusters)

ในขั้นตอนนี้เป็นขั้นตอนการลดขนาดสมาชิกของกลุ่มที่มีขนาดใหญ่โดยทั่วไปในการจัดกลุ่มข้อมูลมักจะปรากฏกลุ่มที่มีขนาดใหญ่อย่างน้อยหนึ่งกลุ่ม หากผลการทดลองที่ได้จากในขั้นตอนนี้มีแนวโน้มที่ไม่ดีก็จะยึดถือผลการจัดกลุ่มในครั้งแรก สำหรับอัลกอริทึมที่ทำการทดลองจะใช้กับอัลกอริทึม K-Means แบบไม่ให้น้ำหนักคุณลักษณะ เนื่องจากหากผลการทดลองมีแนวโน้มไม่ดีจะสามารถหยุดการประมวลได้โดยที่ไม่ต้องทดลองกับอัลกอริทึมอื่นอีก

ผลและวิจารณ์

ผล

1. ผลการทดลอง

1.1 การคัดเลือกคุณลักษณะ

ข้อมูลที่ใช้ในงานวิจัยนี้มีขนาดของคุณลักษณะที่ใหญ่ คือ มีจำนวน 3,194 แถว 971 คุณลักษณะ จึงนำเอาการคัดเลือกคุณลักษณะที่สามารถลดการใช้ทรัพยากรมาใช้ อัลกอริทึมที่ใช้ คือ CFS และ วิธีการเชิงพันธุกรรม สำหรับผลการทดลองได้แสดงดังตารางที่ 10 โดยทดลองประมวลผลจากขนาดประชากรตั้งแต่ 50 ไปเรื่อย ๆ

ตารางที่ 10 ผลการคัดเลือกคุณลักษณะโดยใช้อัลกอริทึม CFS และ GA

Population Size	Merit	จำนวนคุณลักษณะที่ถูกเลือก
50	0.28331	384
100	**0.30816**	**258**
150	0.29915	256
200	0.28239	407
250	0.28399	312
300	0.30094	310
350	0.29332	342

จากตารางที่ 10 ในการทดลองคัดเลือกคุณลักษณะโดยประมวลผลถึงขนาดประชากรสูงสุด 350 เนื่องจากคาดการณ์แนวโน้มว่าค่า Merit เมื่อไปถึงจุดใดจุดหนึ่งแล้วจะค่อย ๆ ลดต่ำลงเรื่อย ๆ และสามารถมีโอกาสอีกที่ค่า Merit จะกลับมาสูง แต่ไม่สามารถคาดการณ์ได้ว่าค่า Merit จะกลับมาสูงอีกครั้งเมื่อใด จึงทำการประมวลผลที่ขนาดประชากรสูงสุดเพียง 350

จากผลการคัดเลือกคุณลักษณะตามตารางที่ 10 สามารถตัดสินได้ว่าที่ขนาดประชากรเท่ากับ 100 จะให้ค่า Merit สูงสุด (0.30816) และเลือกจำนวนคุณลักษณะมาได้ 258 คุณลักษณะ ดังนั้นคุณลักษณะที่เลือกมาเบื้องต้นจึงตัดสินที่ 258 คุณลักษณะ โดยได้แสดงไว้ใน ตารางผนวกที่ 2 ปัญหาใหม่ คือ หลังจากที่ข้อมูลผ่านการคัดเลือกคุณลักษณะแล้ว การประเมินผลของการจัดกลุ่มจะให้ผลที่ดีขึ้นหรือไม่เมื่อเปรียบเทียบกับข้อมูลที่ไม่ผ่านกระบวนการคัดเลือกคุณลักษณะ ซึ่งผลเปรียบเทียบระหว่างข้อมูลที่ผ่านการคัดเลือกคุณลักษณะและข้อมูลที่ไม่ผ่านการคัดเลือกคุณลักษณะจะอธิบายต่อในข้อที่ 1.2 การหาจำนวนกลุ่มที่เหมาะสม

1.2 การหาจำนวนกลุ่มที่เหมาะสม

ปัญหาหลักในการจัดกลุ่ม คือ การหาจำนวนกลุ่มที่เหมาะสม เนื่องจากข้อมูลที่นำมาใช้ในการจัดกลุ่มเป็นข้อมูลที่ไม่มีผลเฉลย หนึ่งในวิธีการที่ใช้หาจำนวนกลุ่ม คือ วิธีการแบบ 2 ขั้นตอนดังที่ได้กล่าวไว้ในข้อ 2.3 ในส่วนวิธีการ โดยเลือกใช้อัลกอริทึม SOM ประกอบการหาจำนวนกลุ่ม เนื่องจากข้อมูลที่ใช้ในการทดลองมีขนาดใหญ่ จึงมีการคัดเลือกคุณลักษณะก่อนการหาจำนวนกลุ่ม โดยทำการประมวลผลตั้งแต่ 2 กลุ่มไปจนถึง 10 กลุ่ม ในตารางที่ 11 เป็นผลการทดลองในการหาจำนวนกลุ่มที่เหมาะสมโดยใช้อัลกอริทึม SOM

จากผลการทดลองในตารางที่ 11 ซึ่งใช้ข้อมูลที่ผ่านการคัดเลือกคุณลักษณะแล้วและวัดประสิทธิภาพด้วยค่า F-Statistic, ค่า Silhouette และค่า RS พบว่าค่า F-Statistic (35.08) และค่า RS (0.75513) ให้จำนวนกลุ่มที่เหมาะสมจำนวน 7 กลุ่ม ในขณะที่ค่า Silhouette (0.61552) ให้จำนวนกลุ่มที่เหมาะสมจำนวน 2 กลุ่ม

ตารางที่ 11 แสดงผลการหาจำนวนกลุ่มที่เหมาะสมโดยใช้อัลกอริทึม SOM กับข้อมูลที่ผ่านการคัดเลือกคุณลักษณะแล้ว

Number of cluster	F-Statistic	Silhouette	RS
2	22.692	**0.61552**	0.66608
3	29.86	0.49134	0.72413
4	26.982	0.3529	0.70343
5	33.668	0.33196	0.74745
6	30.551	0.31029	0.72867
7	**35.08**	0.35024	**0.75513**
8	29.616	0.30261	0.72248
9	29.612	0.29373	0.72246
10	33.714	0.33281	0.74771

เมื่อได้จำนวนกลุ่มที่เหมาะสมในเบื้องต้นแล้ว สิ่งที่ต้องคำนึงต่อ คือ ควรจะยอมรับจำนวนกลุ่มที่เหมาะสม 7 กลุ่มนี้หรือไม่ เนื่องจากข้อมูลที่ทดสอบในเบื้องต้นใช้ข้อมูลที่ผ่านการคัดเลือกคุณลักษณะ ดังนั้นจึงต้องเพิ่มความมั่นใจในความเหมาะสมของจำนวนกลุ่มต่อโดยการเปรียบเทียบกับข้อมูลที่ไม่ผ่านการคัดเลือกคุณลักษณะ

การเปรียบเทียบระหว่างข้อมูลที่ผ่านการคัดเลือกคุณลักษณะและข้อมูลที่ไม่ผ่านการคัดเลือกคุณลักษณะจะได้ผลการทดลองดังตารางที่ 12

ตารางที่ 12 แสดงการเปรียบเทียบผลการทดลองระหว่างข้อมูลที่ผ่านการคัดเลือกคุณลักษณะและข้อมูลที่ไม่ผ่านการคัดเลือกคุณลักษณะโดยใช้อัลกอริทึม SOM

Number of cluster	Features Selection	F-Statistic	Silhouette	RS
7	No	35.07	-0.079855	**0.93879**
7	Yes	**35.08**	**0.35024**	0.75513

จากการเปรียบเทียบระหว่างข้อมูลที่ผ่านการคัดเลือกคุณลักษณะกับข้อมูลที่ไม่ผ่านการคัดเลือกคุณลักษณะด้วยจำนวนกลุ่ม 7 กลุ่ม โดยใช้อัลกอริทึม SOM ในการทดสอบ พบว่า เมื่อข้อมูลผ่านการคัดเลือกคุณลักษณะจะให้ประสิทธิภาพดีขึ้นเนื่องจากตัววัดประสิทธิภาพ 2 ใน 3 ตัว ได้แก่ ค่า F-Statistic (35.08) และค่า Silhouette (0.35024) ให้ประสิทธิภาพดีขึ้นหลังจากข้อมูลผ่านขั้นตอนการคัดเลือกคุณลักษณะแล้ว ถึงแม้ว่าค่า F-Statistic จะไม่สามารถชี้ขาดกับข้อมูลที่ยังไม่ผ่านการคัดเลือกคุณลักษณะแล้วก็ตาม แต่เมื่อพิจารณาประกอบกับค่า Silhouette แล้วทำให้มีน้ำหนักในการตัดสินมากขึ้นอีก เพราะค่า Silhouette จากที่มีค่าเป็นจำนวนติดลบมาเป็นค่าที่เป็นจำนวนเต็มบวก จึงสามารถสรุปได้ว่าจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มของข้อมูลชุดนี้อยู่ที่จำนวน 7 กลุ่มและข้อมูลควรเป็นข้อมูลที่ผ่านการคัดเลือกคุณลักษณะ

1.3 การจัดกลุ่ม

ในส่วนนี้จะเป็นการอธิบายผลการจัดกลุ่ม ด้วยการเปรียบเทียบประสิทธิภาพในการจัดกลุ่มระหว่างอัลกอริทึม K-Means และอัลกอริทึม Fuzzy C-Means ที่จัดกลุ่มแบบมีการให้น้ำหนักและการจัดกลุ่มแบบไม่ให้น้ำหนัก โดยใช้ข้อมูลที่ผ่านการคัดเลือกคุณลักษณะแล้วและกำหนดจำนวนกลุ่มที่เหมาะสมจำนวน 7 กลุ่ม สำหรับผลการวัดประสิทธิภาพของอัลกอริทึม K-Means และอัลกอริทึม Fuzzy C-Means ได้นำเสนอในตารางที่ 13 และตารางที่ 14

ตารางที่ 13 แสดงการเปรียบเทียบการจัดกลุ่มอัลกอริทึม K-Means ที่ผ่านการให้น้ำหนักคุณลักษณะและไม่ให้น้ำหนักคุณลักษณะ

Algorithm	F-Statistic	Silhouette	RS
K-means	56.586	0.20191	0.83261
K-means(TF-IDF)	56.934	**0.68768**	0.83346
K-means(Log)	**79.234**	0.46166	**0.87445**
K-means(Aug)	50.236	0.30611	0.81536

ในการพิจารณาเฉพาะอัลกอริทึม K-Means ดังที่ได้แสดงไว้ในตารางที่ 13 วิธีการ TF-IDF และวิธีการ Logarithm Weight ให้ประสิทธิภาพในการจัดกลุ่มดีขึ้นเมื่อเทียบจากการไม่ให้น้ำหนักคุณลักษณะ โดยพิจารณาจากตัววัดประสิทธิภาพทั้ง 3 ตัว แต่วิธีการ Augmented Weight มีเพียงตัววัดประสิทธิภาพ Silhouette ตัวเดียวเท่านั้นที่ดีขึ้นกว่าการไม่ให้น้ำหนักคุณลักษณะ เมื่อพิจารณาโดยรวมจากทั้งตาราง วิธีการ Logarithm Weight ให้ประสิทธิภาพที่ดีที่สุด เนื่องจากตัววัดประสิทธิภาพ 2 ตัวของการจัดกลุ่มแบบให้น้ำหนักคุณลักษณะ ได้แก่ F-Statistic (79.234) และ RS (0.87445) ให้ผลการจัดกลุ่มที่ดีขึ้น แต่ตัววัดประสิทธิภาพ Silhouette (0.46166) ของวิธีการ Logarithm Weight ซึ่งเป็นตัววัดประสิทธิภาพภายในกลุ่มกลับสู้วิธีการ TF-IDF ไม่ได้ จึงคาดการณ์ในขั้นต้นได้ว่าวิธีการ Logarithm Weight ทำให้ข้อมูลระหว่างกลุ่มแบ่งกันชัดเจนมากขึ้นในขณะที่วิธีการ TF-IDF ทำให้ข้อมูลภายในกลุ่มเกาะกลุ่มกันดีขึ้น

ตารางที่ 14 แสดงการเปรียบเทียบการจัดกลุ่มอัลกอริทึม Fuzzy C-Means ที่ผ่านการให้น้ำหนักคุณลักษณะและไม่ให้น้ำหนักคุณลักษณะ

Algorithm	F-Statistic	Silhouette	RS
Fuzzy C-means	12.983	0.50732	0.53299
Fuzzy C-means(TF-IDF)	18.497	**0.53488**	0.61919
Fuzzy C-means(Log)	21.761	0.41225	0.6567
Fuzzy C-means(Aug)	**43.223**	0.24147	**0.79165**

การทดสอบของอัลกอริทึม Fuzzy C-Means จะพบว่า อัลกอริทึม Fuzzy C-Means แบบมีการให้น้ำหนักคุณลักษณะด้วยวิธีการ Augmented Weight ให้ผลการจัดกลุ่มดีที่สุด พิจารณาจากค่า F-Statistic (43.223) และค่า RS (0.79165) ที่มากที่สุด และพิจารณาผลการทดลองโดยรวมอัลกอริทึม Fuzzy C-Means ให้ผลการจัดกลุ่มแบบให้น้ำหนักคุณลักษณะทั้งสามวิธีดีกว่าการจัดกลุ่มแบบไม่ให้น้ำหนักคุณลักษณะ เมื่อพิจารณาจากตัววัดประสิทธิภาพระหว่างกลุ่ม คือ ค่า F-Statistic และค่า RS แต่สิ่งที่น่าสังเกต คือ ตัววัดประสิทธิภาพภายในกลุ่มอย่าง Silhouette ของข้อมูลที่ผ่านการให้น้ำหนักคุณลักษณะ มีเพียงวิธีการ TF-IDF เท่านั้นที่ดีกว่าการจัดกลุ่มแบบไม่ให้คุณลักษณะ

ส่วนวิธีการ Logarithm Weight และวิธีการ Augmented Weight กลับแย่ลงกว่าการจัดกลุ่มแบบไม่ให้คุณลักษณะ

ตารางที่ 15 แสดงการเปรียบเทียบประสิทธิภาพการจัดกลุ่มระหว่างอัลกอริทึม K-Means และอัลกอริทึม Fuzzy C-Means ที่ผ่านการให้น้ำหนักคุณลักษณะและไม่ให้น้ำหนักคุณลักษณะ

Algorithm	F-Statistic	Silhouette	RS
K-means	56.586	0.20191	0.83261
K-means(TF-IDF)	56.934	**0.68768**	0.83346
K-means(Log)	**79.234**	0.46166	**0.87445**
K-means(Aug)	50.236	0.30611	0.81536
Fuzzy C-means	12.983	0.50732	0.53299
Fuzzy C-means(TF-IDF)	18.497	0.53488	0.61919
Fuzzy C-means(Log)	21.761	0.41225	0.6567
Fuzzy C-means(Aug)	43.223	0.24147	0.79165

พิจารณาผลการทดลองจากตารางที่ 15 เปรียบเทียบระหว่างอัลกอริทึม K-Means และอัลกอริทึม Fuzzy C-Means ทั้งการจัดกลุ่มแบบให้น้ำหนักคุณลักษณะทั้งสามเทคนิค ได้แก่ วิธีการ TF-IDF วิธีการ Logarithm Weight (Log) และวิธีการ Augmented Weight (Aug) กับการจัดกลุ่มแบบไม่ให้น้ำหนักคุณลักษณะ จากการเปรียบเทียบโดยรวมแล้วพบว่าการจัดกลุ่มโดยใช้อัลกอริทึม K-Means แบบมีการให้น้ำหนักคุณลักษณะด้วยวิธีการ Logarithm Weight (K-means(Log)) ให้ประสิทธิภาพในการจัดกลุ่มดีที่สุดจากตัววัดประสิทธิภาพ 2 ตัวใน 3 ตัว คือ ค่า F-Statistic (79.234) และค่า RS (0.87445) ซึ่งเป็นตัววัดประสิทธิภาพของข้อมูลระหว่างกลุ่มหรือภายนอกกลุ่ม ในขณะเดียวกันตัววัดประสิทธิภาพ Silhouette ซึ่งเป็นตัววัดประสิทธิภาพภายในกลุ่มกลับให้ค่าที่ดีที่สุดจากการจัดกลุ่มโดยใช้อัลกอริทึม K-Means แบบให้น้ำหนักคุณลักษณะด้วยวิธีการ TF-IDF

(K-means(TF-IDF)) ดังนั้นจึงสามารถสรุปได้ว่าการจัดกลุ่มโดยใช้อัลกอริทึม K-Means ที่ให้น้ำหนักคุณลักษณะแบบ Logarithm Weight จะให้ผลดีต่อการจัดกลุ่มเมื่อวัดประสิทธิภาพของข้อมูลระหว่างกลุ่มสำหรับชุดข้อมูลชุดนี้

หลังจากที่ได้ผลการจัดกลุ่มข้อมูลที่ดีเป็นที่น่าพอใจแล้ว สิ่งที่ควรคำนึงต่อ คือ การหาความเชี่ยวชาญหรือคำสำคัญหลักประจำกลุ่มแต่ละกลุ่มตามผลที่อัลกอริทึมจัดกลุ่มให้ ในกรณีที่ผู้เชี่ยวชาญมีคำสำคัญมากกว่า 1 คำ หรือมีผู้เชี่ยวชาญหลายคนมีคำสำคัญที่คาบเกี่ยวกันจะสรุปผลที่ได้จากการจัดกลุ่มเป็นตัวชี้นำว่าผู้เชี่ยวชาญคนดังกล่าวควรอยู่ในกลุ่มคำสำคัญใด

1.4 การสกัดคำสำคัญและพิจารณาผู้เชี่ยวชาญของแต่ละกลุ่ม

หลังจากการจัดกลุ่มในขั้นตอนที่ 1.3 อัลกอริทึม K-Means แบบให้น้ำหนักคุณลักษณะด้วยวิธีการ Logarithm Weight ให้ผลในการจัดกลุ่มดีที่สุด ในการนำผลการจัดกลุ่มมาใช้ประโยชน์ควรพิจารณาจำนวนสมาชิกในแต่ละกลุ่มว่ากลุ่มดังกล่าวประกอบไปด้วยสมาชิกกี่แถว ซึ่งมีดังนี้

ตารางที่ 16 จำนวนสมาชิก (แถว) ที่ประกอบในแต่ละกลุ่ม (ข้อมูล 3,194 แถว 258 คุณลักษณะ)

กลุ่มที่	จำนวนสมาชิกในกลุ่ม (แถว)
1	50
2	249
3	63
4	11
5	43
6	2,671
7	107

เมื่อพิจารณาจำนวนสมาชิกในแต่ละกลุ่มแล้ว ในขั้นตอนต่อไป คือ การสกัดคำสำคัญที่เป็นไปได้ในแต่ละกลุ่มทั้ง 7 กลุ่ม จากผลการจัดกลุ่มโดยใช้อัลกอริทึม K-Means แบบให้น้ำหนักคุณลักษณะด้วยวิธีการ Logarithm Weight โดยพิจารณาจากความถี่คำสำคัญในแต่ละกลุ่มสูงสุด 5

อันดับแรก เนื่องจากคำสำคัญในอันดับต่อไปมักมีคำสำคัญที่เท่ากันหลายคำ คำสำคัญที่ทำการสกัดแล้วได้แสดงตามตารางที่ 17

จากตารางที่ 17 ในกรณีที่ความถี่คำสำคัญเท่ากัน จะพิจารณาจากรหัสคำสำคัญก่อนหลังตามที่ได้มีการกำหนดไว้ก่อนแล้ว ซึ่งเมื่อพิจารณาแล้วมักจะเป็นคำที่เรียงไปตามลำดับพยัญชนะทั้งในภาษาไทยและภาษาอังกฤษ หากพิจารณาร่วมกับตารางที่ 16 จะพบว่าบางกลุ่มมีความถี่ของคำสำคัญน้อยแต่จำนวนสมาชิกกลับมีมาก ดังเช่น กลุ่มที่ 6 ซึ่งจะกล่าวถึงในส่วนการวิจารณ์

จากผลการทดลองจริงที่ได้จากอัลกอริทึม K-Means แบบให้น้ำหนักคุณลักษณะด้วยวิธีการ Logarithm Weight ในตารางที่ 18 แสดงการคิดเปอร์เซ็นต์ของผู้เชี่ยวชาญในแต่ละกลุ่มว่ากลุ่มใดมีผู้เชี่ยวชาญในกลุ่มสายงานใดมาก โดยวิเคราะห์จากจำนวนแถวและคิดคำนวณออกมาเป็นเปอร์เซ็นต์จากกลุ่มสายงานทั้ง 9 กลุ่มสายงานข้างต้น ซึ่งสามารถนำผลการคิดเปอร์เซ็นต์ผู้เชี่ยวชาญนี้ไปใช้ในการวิเคราะห์เพื่อเชื่อมโยงกับคำสำคัญที่สกัดได้จากตารางที่ 17 สำหรับการวิเคราะห์เพิ่มเติมเกี่ยวกับความเชื่อมโยงของทั้งสองตารางนี้จะกล่าวถึงในส่วนการวิจารณ์ต่อไป

ตารางที่ 17 คำสำคัญที่ผ่านการสกัดจากความถี่คำสูงสุด 5 อันดับแรก (N = 3,194)

กลุ่ม 1 (50)		กลุ่ม 2 (249)		กลุ่ม 3 (63)		กลุ่ม 4 (11)		กลุ่ม 5 (43)		กลุ่ม 6 (2,671)		กลุ่ม 7 (107)	
คำ	ความถี่	คำ	ความถี่	คำ	ความถี่	คำ	ความถี่	คำ	ความถี่	คำ	ความถี่	คำ	ความถี่
เศรษฐศาสตร์	157	ป่า	282	Rice	82	Corn	22	สอน	71	Economic	124	ไฟฟ้า	212
Economic	143	Generate	125	ร้อน	47	หวาน	14	Student	63	สอน	108	Industrial	14
ทุน	61	ปลา	117	คาร์บอน	24	Baby	10	Admin	39	เศรษฐศาสตร์	95	ไร้สาย	12
เงิน	60	ไข่	72	ไดออกไซด์	24	Sweet	10	ช่วยสอน	34	Genetic	91	Fire	10
Finance	36	ท่องเที่ยว	39	สตาร์ช	24	Rice	8	เรียน	34	Industrial	89	Safe	10

หมายเหตุ N คือ จำนวนแถวของข้อมูล

ในกลุ่มที่ 2 อันดับคำสำคัญที่ 3 เป็นคำว่า “ทั่วไป” ซึ่งหากพิจารณาแล้วพบว่าเป็นคำที่ไม่สื่อถึงความเชี่ยวชาญ จึงตัดคำนี้ออกไปและพิจารณาคำในอันดับถัดไปแทน

ในกลุ่มที่ 4 คำว่า “Baby” เมื่อเปรียบเทียบกับชื่อผลงานวิจัยในชุดข้อมูลที่ใช้ในการทดลอง ในที่นี้หมายถึง ต้นอ่อนของพืช

ตารางที่ 18 ผู้เชี่ยวชาญจากกลุ่มสายงานต่าง ๆ ทั้ง 9 กลุ่มสายงานที่อยู่ในแต่ละกลุ่ม

	กลุ่ม 1		กลุ่ม 2		กลุ่ม 3		กลุ่ม 4		กลุ่ม 5		กลุ่ม 6		กลุ่ม 7	
กลุ่มสายงาน	**แถว**	**%**	**แถว**	**%**	**แถว**	**%**	**แถว**	**%**	**แถว**	**%**	**แถว**	**%**	**แถว**	**%**
วิทยาศาสตร์	5	10	17	6.83	8	12.7	0	0	4	9.3	387	14.49	11	10.28
วิศวกรรม	10	20	19	7.63	12	19.05	1	9.09	**15**	**34.88**	400	14.98	23	21.5
เกษตรฯ	**16**	**32**	50	20.08	**13**	**20.63**	**4**	**36.36**	9	20.93	**526**	**19.69**	**24**	**22.43**
บริหารฯ	3	6	17	6.83	4	6.35	1	9.09	0	0	195	7.3	11	10.28
สังคมศาสตร์	2	4	17	6.83	6	9.52	0	0	3	6.98	179	6.7	13	12.15
ศึกษาศาสตร์	5	10	34	13.65	9	14.29	3	27.27	6	13.95	371	13.89	13	12.15
สัตว์แพทย์ศาสตร์	6	12	14	5.62	3	4.76	2	18.18	5	11.63	137	5.13	3	2.8
สถาบันวิจัย มก.	1	2	29	11.65	6	9.52	0	0	1	2.33	218	8.16	3	2.8
สำนักงานภายในมหาวิทยาลัย	2	4	**52**	**20.88**	2	3.17	0	0	0	0	258	9.66	6	5.61
รวม	50	100	249	100	63	100	11	100	43	100	2,671	100	107	100

หมายเหตุ แถว คือ จำนวนผู้เชี่ยวชาญที่กลุ่มสายงานนั้น ๆ ประกอบอยู่ในกลุ่ม

จากตารางที่ 18 ผู้เชี่ยวชาญจากกลุ่มสายงานต่าง ๆ 3 อันดับแรกตามลำดับ ดังนี้

กลุ่มที่ 1 ได้แก่ ผู้เชี่ยวชาญจากกลุ่มสายงานเกษตร อุตสาหกรรมเกษตร (32%), กลุ่มสายงานวิศวกรรม (20%) และกลุ่มสายงานสัตว์แพทย์ศาสตร์ (12%)

กลุ่มที่ 2 ได้แก่ ผู้เชี่ยวชาญจากกลุ่มสำนักงานภายในมหาวิทยาลัย (20.88%), กลุ่มสายงานเกษตร อุตสาหกรรมเกษตร (20.08%) และกลุ่มสายงานศึกษาศาสตร์ (13.65%)

กลุ่มที่ 3 ได้แก่ ผู้เชี่ยวชาญจากกลุ่มสายงานเกษตร อุตสาหกรรมเกษตร (20.63%), กลุ่มสายงานวิศวกรรม (19.05%) และกลุ่มสายงานศึกษาศาสตร์ (14.29%)

กลุ่มที่ 4 เป็นกลุ่มที่เล็กที่สุด ได้แก่ ผู้เชี่ยวชาญจากกลุ่มสายงานเกษตร อุตสาหกรรมเกษตร (36.36%), กลุ่มสายงานศึกษาศาสตร์ (27.27%) และกลุ่มสายงานสัตว์แพทย์ศาสตร์ (18.18%)

กลุ่มที่ 5 ได้แก่ ผู้เชี่ยวชาญจากกลุ่มสายงานวิศวกรรม (34.88%), กลุ่มสายงานเกษตร อุตสาหกรรมเกษตร (20.93%) และกลุ่มสายงานศึกษาศาสตร์ (13.95%)

กลุ่มที่ 6 เป็นกลุ่มที่ใหญ่ที่สุด ได้แก่ ผู้เชี่ยวชาญจากกลุ่มสายงานเกษตร อุตสาหกรรมเกษตร (19.69%), กลุ่มสายงานวิศวกรรม (14.98%) และกลุ่มสายงานวิทยาศาสตร์ (14.49%)

กลุ่มที่ 7 ได้แก่ ผู้เชี่ยวชาญจากกลุ่มสายงานเกษตร อุตสาหกรรมเกษตร (22.43%), กลุ่มสายงานวิศวกรรม (21.5%) และอันดับที่สามร่วมกัน คือ กลุ่มสายงานสังคมศาสตร์และกลุ่มสายงานศึกษาศาสตร์ (12.15%)

เหตุจำเป็นที่ต้องประเมินผู้เชี่ยวชาญจากกลุ่มสายงานต่าง ๆ เพียง 3 อันดับแรก เนื่องจากข้อมูลหลังจากอันดับ 4 เป็นต้นไปจะมีโอกาสเป็นข้อมูลที่มีจำนวนแถวในกลุ่มสายงานต่าง ๆ เท่ากัน เมื่อย้อนกลับไปพิจารณาในตารางที่ 16 สามารถพิจารณาได้ว่าในกลุ่มที่ 6 เป็นกลุ่มที่ใหญ่ซึ่งประกอบไปด้วยสมาชิกมากถึงหลักพันแถว คือ 2,671 แถว ทำให้ยากต่อการระบุคำสำคัญได้ และ

เมื่อพิจารณาในตารางที่ 17 พบว่ามีคำสำคัญที่ปะปนกันและไม่มีความเกี่ยวข้องกันมากนัก จึงมีแนวคิดในการนำกลุ่มใหญ่นี้มาจัดกลุ่มใหม่เพื่อลดขนาดของกลุ่มลง สำหรับการจัดกลุ่มใหม่จะกล่าวถึงในข้อ 1.5

ในตารางที่ 19 แสดงผลการทดลองโดยพิจารณารวมจากคำสำคัญและผู้เชี่ยวชาญจากกลุ่มสายงานต่าง ๆ จากการจัดกลุ่ม

ตารางที่ 19 คำสำคัญและผู้เชี่ยวชาญจากกลุ่มสายงานต่าง ๆ ในการจัดกลุ่มผู้เชี่ยวชาญ

กลุ่ม	คำสำคัญ	กลุ่มสายงานผู้เชี่ยวชาญตามสายงาน
1	เศรษฐศาสตร์, Economic, ทุน, เงิน, Finance	เกษตรฯ, วิศวกรรม, สัตว์แพทย์ศาสตร์
2	ป่า, Generate, ปลา, ไข่, ท่องเที่ยว	สำนักงานภายใน, เกษตรฯ, ศึกษาศาสตร์
3	Rice, ร้อน, คาร์บอน, ไดออกไซด์, สตาร์ช	เกษตรฯ, วิศวกรรม, ศึกษาศาสตร์
4	Corn, หวาน, Baby, Sweet, Rice	เกษตรฯ, ศึกษาศาสตร์, สัตว์แพทย์ศาสตร์
5	สอน, Student, Admin, ช่วยสอน, เรียน	วิศวกรรม, เกษตรฯ, ศึกษาศาสตร์
6	Economic, สอน, เศรษฐศาสตร์, Genetic, Industrial	เกษตรฯ, วิศวกรรม, วิทยาศาสตร์
7	ไฟฟ้า, Industrial, ไร้สาย, Fire, Safe	เกษตรฯ, วิศวกรรม, สังคมศาสตร์, ศึกษาศาสตร์

ในตารางที่ 19 จะสังเกตได้ว่าถึงแม้ผู้เชี่ยวชาญจากกลุ่มสายงานเกษตร อุตสาหกรรมเกษตรจะปรากฏกระจายไปทั่วทุกกลุ่ม แต่เมื่อย้อนกลับไปพิจารณาคำสำคัญ เราอาจสรุปได้ว่าผู้เชี่ยวชาญจากกลุ่มสายงานเกษตร อุตสาหกรรมเกษตรมีความเกี่ยวเนื่องกับงานวิจัยในกลุ่มสายงานอื่นด้วย เช่น ในกลุ่มที่ 1 ผู้เชี่ยวชาญจากกลุ่มสายงานเกษตร อุตสาหกรรมเกษตรอาจทำงานวิจัยด้านการพิจารณาต้นทุนในการทำดัดแปลงพันธุ์พืช จึงทำให้ต้องเกี่ยวข้องกับงานวิจัยในกลุ่มด้านบริหารเศรษฐศาสตร์ เป็นต้น และอีกส่วนหนึ่งที่ทำให้กลุ่มสายงานผู้เชี่ยวชาญที่ได้ขัดแย้งกับคำสำคัญที่สกัดได้ คือ เมื่อมองย้อนไปยังงานวิจัยของสมาชิกภายในกลุ่มสามารถสังเกตได้ว่าผู้เชี่ยวชาญหนึ่ง

คนจากสายงานเศรษฐศาสตร์มีจำนวนงานวิจัยค่อนข้างมากกว่าผู้เชี่ยวชาญจากกลุ่มสายงานอื่น ๆ ปัญหาอีกจุดหนึ่งจากตารางที่ 19 คือ เราไม่ทราบว่าผู้เชี่ยวชาญจากสถาบันกลุ่มวิจัยและสำนักงานภายในมหาวิทยาลัยดำเนินงานวิจัยเฉพาะในด้านใด จึงสามารถมองได้ว่า 2 กลุ่มสายงานนี้สามารถวิจัยได้กว้างทุกด้าน

1.5 การจัดกลุ่มครั้งที่ 2

ในขั้นตอนนี้เป็นเพียงการทดลองเสริมเพื่อพยายามที่จะลดขนาดของกลุ่มที่ 6 ซึ่งเป็นกลุ่มใหญ่ เนื่องจากมีสมาชิกในกลุ่มจำนวน 2,671 แถว ทำให้ยากต่อการสกัดคำสำคัญดังที่ได้กล่าวไว้ในข้อ 1.4 โดยอัลกอริทึมที่ใช้ในการทดลองเสริมจะเป็นอัลกอริทึม K-Means แบบไม่ให้คุณลักษณะก่อน เนื่องจากตามทฤษฎีโดยทั่วไปการให้น้ำหนักคุณลักษณะมักทำให้ข้อมูลมีระยะทางที่ใกล้กว่าการไม่ให้น้ำหนักคุณลักษณะ กล่าวคือ ข้อมูลจะเกาะกลุ่มแน่นขึ้น ในการทดลองจัดกลุ่มครั้งที่ 2 จะทำการประมวลผลหาจำนวนกลุ่มที่เหมาะสมตั้งแต่ 2 กลุ่มไปจนถึง 10 กลุ่ม และใช้ตัววัดประสิทธิภาพเดิมทั้งสามตัววัดประสิทธิภาพ คือ ค่า F-Statistic, ค่า Silhouette และค่า RS สำหรับผลการทดลองการจัดกลุ่มครั้งที่ 2 แสดงในตารางที่ 20

ตารางที่ 20 ผลการทดลองที่ได้จากการจัดกลุ่มครั้งที่ 2

Number of cluster	F-Statistic	Silhouette	RS
2	NaN	0.41246	NaN
3	NaN	0.55306	NaN
4	NaN	0.507	NaN
5	NaN	0.44181	NaN
6	NaN	0.53519	NaN
7	NaN	0.42108	NaN
8	NaN	0.51609	NaN
9	NaN	0.55707	NaN
10	NaN	0.476	NaN

จากตารางที่ 20 พบว่าค่า F-Statistic และค่า RS ให้ค่า NaN (Not a Number) ทั้งหมด ซึ่งหมายถึงเกิดจำนวนค่าอนันต์ (Infinity) ทำให้ไม่สามารถหาค่าออกมาเป็นตัวเลขได้ มีเพียงค่า Silhouette เท่านั้นที่ยังคงให้ผลการวัดประสิทธิภาพออกมาเป็นตัวเลขได้ จากการวิเคราะห์เฉพาะค่า Silhouette พบว่าให้จำนวนกลุ่มที่ดีที่สุดจำนวน 9 กลุ่ม จึงทำการวิเคราะห์ต่อว่าในสมาชิกทั้ง 9 กลุ่มนี้ประกอบไปด้วยคำสำคัญใดบ้างและสมาชิกในแต่ละกลุ่มมีจำนวนเท่าใด สำหรับการวิเคราะห์ค่า F-Statistic และค่า RS ในการให้ค่า NaN จะกล่าวถึงต่อในส่วนการวิจารณ์เช่นกัน

จากผลการทดลองจัดกลุ่มครั้งที่ 2 พบว่าค่า F-Statistic และค่า RS ให้ค่า NaN ทั้งหมด ซึ่งหมายถึงเกิดระยะทางเป็นค่าอนันต์ ทำให้ไม่สามารถหาค่าออกมาเป็นตัวเลขได้ จึงหยุดแนวคิดการจัดกลุ่มซ้ำเพียงเท่านี้ สำหรับการพิจารณาสมาชิกในแต่ละกลุ่มจากผลการจัดกลุ่มครั้งที่ 2 ในกลุ่มที่ 7 สามารถสกัดคำสำคัญออกมาได้เพียง 4 คำเท่านั้น เนื่องจากมีจำนวนสมาชิกน้อยเพียง 2 แถวเท่านั้น และยังคงปรากฏกลุ่มที่มีขนาดใหญ่อีก คือ กลุ่มที่ 9 มีจำนวนสมาชิกถึง 2,097 แถวดังที่ได้แสดงตามตารางที่ 21

ตารางที่ 21 จำนวนสมาชิกที่ได้จากการจัดกลุ่มครั้งที่ 2 จำนวน 9 กลุ่ม

กลุ่มที่	จำนวนสมาชิกในกลุ่ม (แถว)
1	152
2	97
3	31
4	46
5	88
6	56
7	2
8	102
9	2,097

ในขั้นตอนต่อมาจะเป็นการสกัดความเชี่ยวชาญและพิจารณาผู้เชี่ยวชาญจากกลุ่มสายงานต่าง ๆ ในแต่ละกลุ่ม

ตารางที่ 22 คำสำคัญและผู้เชี่ยวชาญจากกลุ่มสายงานต่าง ๆ ในการจัดกลุ่มผู้เชี่ยวชาญในการจัดกลุ่มครั้งที่ 2

กลุ่ม	คำสำคัญ	กลุ่มสายงานผู้เชี่ยวชาญตามสายงาน
1	สอน, Industrial, Social, คลัง, บรรยากาศ	วิศวกรรม, เกษตรฯ, ศึกษาศาสตร์
2	ปลา, ร้อน, กล้วย, ท่องเที่ยว, เค็ม	เกษตรฯ, วิทยาศาสตร์, ศึกษาศาสตร์
3	ผัก, สมุนไพร, Consume, Eucalyptus, มือ	สถาบันวิจัย, เกษตรฯ, สำนักงานภายใน
4	สิ่งแวดล้อม, ขยะ, ตะกอน, Tourist, Metal	เกษตรฯ, วิทยาศาสตร์, ศึกษาศาสตร์
5	Admin, เงิน, Finance, ธุรกิจ, ธนาคาร	เกษตรฯ, วิศวกรรม, สำนักงานภายใน, วิทยาศาสตร์, สัตว์แพทย์ศาสตร์
6	Genetic, ยีน, Molecular, Mutation, China	เกษตรฯ, วิศวกรรม, ศึกษาศาสตร์
7	Cluster, Mine, สอน, Mill	เกษตรฯ, ศึกษาศาสตร์
8	Economic, เศรษฐศาสตร์, คหกรรม, ธุรกิจ, Admin	วิทยาศาสตร์, เกษตรฯ, ศึกษาศาสตร์
9	ธุรกิจ, Social, สมุนไพร, พฤกษ, โปรตีน	เกษตรฯ, วิทยาศาสตร์, วิศวกรรม, ศึกษาศาสตร์

จากตารางที่ 22 พบว่าในกลุ่มที่ 1 กลุ่มที่ 5 กลุ่มที่ 6 กลุ่มที่ 7 กลุ่มที่ 8 และกลุ่มที่ 9 ยังคงมีการปะปนกันของคำสำคัญและผู้เชี่ยวชาญในสายงานต่าง ๆ กันสูง

วิจารณ์

2. วิจารณ์

2.1 การคัดเลือกคุณลักษณะ

ข้อมูลที่ได้ในเบื้องต้นหลังผ่านกระบวนการการนับความถี่คำแล้ว (กระบวนการ Preprocess) มีจำนวนคุณลักษณะที่มากถึง 971 คุณลักษณะ ดังนั้นจำนวนข้อมูลที่ใช้ในการทดลองจัดกลุ่มมีจำนวนคุณลักษณะมาก เพื่อลดปริมาณการใช้ทรัพยากร เช่น หน่วยความจำ จึงจำเป็นต้องมีการนำอาเทคนิคการคัดเลือกคุณลักษณะมาใช้ แต่สิ่งสำคัญที่ต้องคำนึง คือ ต้องไม่ให้ประสิทธิภาพในการจัดกลุ่มแย่ลง สำหรับเทคนิคที่นำมาใช้ในการคัดเลือกคุณลักษณะ คือ CFS ร่วมกับวิธีการ Genetic Search

การนำเอาเทคนิคการคัดเลือกคุณลักษณะมาใช้เพื่อตัดคุณลักษณะที่มีผลต่อข้อมูลน้อยที่สุด กล่าวคือ บางคุณลักษณะอาจจะเคยปรากฏเพียง 2 แถว หรือ 3 แถว จากข้อมูลทั้งชุด จากตารางที่ 10 ซึ่งประมวลผลสูงสุดที่จำนวนประชากร 350 เนื่องจากลักษณะของค่า Merit ที่สูงขึ้นแล้วค่อย ๆ ลงต่ำและกลับมาสูงใหม่อีก แต่ไม่อาจดีกว่าจำนวนประชากร 100 แนวโน้มดังกล่าวทำให้ไม่อาจคาดการณ์ได้ว่าหากทำการประมวลผลต่อไปที่จำนวนประชากรที่เท่าใดถึงให้ผลการทดลองดีกว่าจำนวนประชากร 100 ซึ่งอาจต้องประมวลผลต่อไปอีกยาวนาน จึงหยุดที่จำนวนประชากร 350

2.2 การหาจำนวนกลุ่ม

หลังจากที่ข้อมูลลดขนาดของคุณลักษณะลงแล้ว ในขั้นตอนต่อไปจะเป็นการหาจำนวนกลุ่มที่เหมาะสม ในที่นี้จะใช้อัลกอริทึม SOM ในการหาจำนวนกลุ่ม ซึ่งผลการทดลองได้แสดงดังตารางที่ 11 จำนวนกลุ่มที่เหมาะสมอยู่ที่ 7 กลุ่ม โดยตัดสินจากค่า F-Statistic (32.08) และค่า RS (0.75513) สิ่งที่ต้องคำนึงต่อ คือ ข้อมูลที่ยังไม่ผ่านการคัดเลือกคุณลักษณะอาจดีกว่าข้อมูลที่ผ่านการคัดเลือกคุณลักษณะได้ ดังนั้นจึงทำการทดลองกับอัลกอริทึม SOM ที่จำนวนกลุ่ม 7 กลุ่มอีกครั้งจากผลการทดลองในตารางที่ 12 หน้า 54 ทำให้สามารถมั่นใจได้ว่าข้อมูลที่ผ่านการคัดเลือกคุณลักษณะดีกว่าข้อมูลที่ยังไม่ผ่านการคัดเลือกคุณลักษณะ โดยตัดสินจากค่า F-Statistic (32.08)

และค่า Silhouette (0.35024) ถึงแม้ว่าค่า F-Statistic จะดีกว่าไม่มากนักก็ตาม (ประมาณ 0.01) สิ่งที่น่าสังเกตอีกจุดหนึ่งของตารางที่ 12 นี้ คือ ค่า Silhouette จากที่เป็นค่าติดลบเมื่อข้อมูลยังไม่ผ่านการคัดเลือกคุณลักษณะมาเป็นค่าบวก สามารถบ่งบอกได้ว่าหลังจากที่ข้อมูลผ่านกระบวนการคัดเลือกคุณลักษณะแล้วทำให้ข้อมูลเกาะกลุ่มกันแน่นหนากว่าข้อมูลที่ยังไม่ผ่านการคัดเลือกคุณลักษณะอย่างเห็นได้ชัด ดังนั้นค่า Silhouette จึงเป็นตัวช่วยชี้ขาดในการสนับสนุนผลการทดลองร่วมกับค่า F-Statistic

2.3 การจัดกลุ่ม

จากผลการจัดกลุ่มข้อมูลผู้เชี่ยวชาญในตารางที่ 15 หน้า 57 พบว่า อัลกอริทึม K-Means แบบมีการให้น้ำหนักคุณลักษณะด้วยวิธีการ Logarithm Weight ให้ประสิทธิภาพในการจัดกลุ่มดีที่สุดจากตัววัดประสิทธิภาพของข้อมูลระหว่างกลุ่ม คือ ค่า F-Statistic (79.234) และค่า RS (0.87445) ในขณะที่ตัววัดประสิทธิภาพภายในกลุ่มอย่างค่า Silhouette กลับให้ค่าที่ดีที่สุดจากการจัดกลุ่มโดยใช้อัลกอริทึม K-Means แบบให้น้ำหนักคุณลักษณะด้วยวิธีการ TF-IDF คือ 0.68768 แสดงว่าข้อมูลที่ได้จากการจัดกลุ่มนี้ถูกแบ่งแยกกลุ่มกันอย่างชัดเจน ส่วนการจัดกลุ่มโดยใช้อัลกอริทึม K-Means ด้วยการให้น้ำหนักคุณลักษณะด้วยวิธีการ TF-IDF เมื่อเปรียบเทียบจากในตารางที่ 15 สามารถสรุปได้ว่าให้ผลการจัดกลุ่มดีขึ้นเมื่อวัดประสิทธิภาพของข้อมูลภายในกลุ่มเนื่องจากให้ค่า Silhouette ดีที่สุด จากผลการทดลองโดยให้น้ำหนักคุณลักษณะด้วยวิธีการ Logarithm Weight มีแนวโน้มเป็นไปในทางเดียวกันกับงานวิจัยของ Lan (Lan *et al.*, 2005) คือ การให้น้ำหนักคุณลักษณะโดยใช้วิธีการ Logarithm Weight ให้ผลดีกว่าการไม่ให้น้ำหนักคุณลักษณะและการให้น้ำหนักคุณลักษณะด้วยวิธีการ TF-IDF

ในการจัดกลุ่มข้อมูลซึ่งเป็นขั้นตอนลำดับถัดมาได้นำอัลกอริทึมการจัดกลุ่มแบบแบ่งส่วน คือ อัลกอริทึม K-Means และอัลกอริทึม Fuzzy C-Means มาใช้ เนื่องจากอัลกอริทึมแบบแบ่งส่วนเป็นอัลกอริทึมที่ทำได้ง่ายและรวดเร็ว จากงานวิจัยที่ผ่านมา (วงกต, 2551) ได้อ้างอิงผลการทดลองว่าอัลกอริทึม K-Means ทำงานได้รวดเร็ว ใช้รอบการทำงานน้อยจึงถือได้ว่ามีประสิทธิภาพดีระดับหนึ่ง และอัลกอริทึม Fuzzy C-Means ในงานวิจัยที่ผ่านมา (ศศิธร, 2550) ได้อ้างอิงว่าให้ประสิทธิภาพในการจัดกลุ่มดีที่สุด จึงได้นำเอาสองอัลกอริทึมนี้มาใช้ในการทดลองสำหรับกรณีศึกษานี้

สำหรับอัลกอริทึม Fuzzy C-Means ในตารางที่ 14 หน้า 56 ค่า Silhouette ของข้อมูลที่ผ่านการให้น้ำหนักคุณลักษณะ มีเพียงวิธีการ TF-IDF เท่านั้นที่ดีกว่าการจัดกลุ่มแบบไม่ให้คุณลักษณะ คือ 0.53488 ส่วนค่า F-Statistic และค่า RS ของการจัดกลุ่มแบบให้น้ำหนักคุณลักษณะทุกวิธีการดีกว่าการจัดกลุ่มแบบไม่ให้คุณลักษณะ จึงสามารถสรุปได้ว่าสำหรับอัลกอริทึม Fuzzy C-Means ที่ให้น้ำหนักคุณลักษณะจะให้ผลดีขึ้นกับการแบ่งข้อมูลระหว่างกลุ่มหรืออีกนัยหนึ่ง ข้อมูลระหว่างกลุ่มถูกแบ่งกันอย่างชัดเจนมากขึ้น และเมื่อพิจารณาแนวโน้มการวัดประสิทธิภาพจึงเป็นไปในแนวทางเดียวกับอัลกอริทึม K-Means คือ ถึงอย่างไรการให้น้ำหนักคุณลักษณะยังคงส่งผลให้ประสิทธิภาพในการจัดกลุ่มดีขึ้น

นอกจากนี้ในอัลกอริทึม K-Means ที่ผ่านการให้น้ำหนักคุณลักษณะทุกเทคนิคจากตารางที่ 15 ซึ่งให้ผลการจัดกลุ่มที่ดีกว่าอัลกอริทึม Fuzzy C-Means จากผลการทดลองดังกล่าวขัดแย้งกับการศึกษาของศศิธร (ศศิธร, 2550) ที่รายงานว่าอัลกอริทึม Fuzzy C-Means ดีกว่าอัลกอริทึม K-Means อาจสืบเนื่องมาจากข้อมูลมีการเกาะกลุ่มกันแน่นหนามาก (พิจารณาจากค่า Silhouette ของข้อมูลที่ผ่านการให้น้ำหนักคุณลักษณะเทียบกับค่า Silhouette ของข้อมูลที่ไม่ผ่านการคัดเลือกคุณลักษณะ) จึงเป็นไปได้สูงที่จะสุ่มจุดศูนย์กลางอยู่ในกลุ่มที่เกาะกลุ่มกันพอดี ส่วนอัลกอริทึม Fuzzy C-Means อาจถือว่าให้ประสิทธิภาพดีในระดับหนึ่งแต่อาจไม่เพียงพอที่จะเทียบกับอัลกอริทึม K-Means ได้ อาจสืบเนื่องมาจากการคำนวณค่า Fuzzy Pseudo Partition ที่ต้องมีการกำหนดค่า p ซึ่งมีอิทธิพลต่อผลการจัดกลุ่ม

สิ่งที่น่าสังเกตจากตารางที่ 15 หน้า 57 อีกส่วนหนึ่ง คือ การวัดประสิทธิภาพภายในกลุ่มโดยใช้ Silhouette สามารถคาดการณ์ว่าข้อมูลน่าจะเกาะกลุ่มกันภายในกลุ่มค่อนข้างดี เพราะค่า Silhouette ของทุกวิธีการไม่เป็นค่าติดลบ

ในตารางที่ 16 หน้า 58 แสดงจำนวนสมาชิกโดยรวมของแต่ละกลุ่ม ปัญหาที่สามารถพบได้จากตารางนี้ คือ มีกลุ่มหนึ่งกลุ่มที่มีขนาดที่ใหญ่มาก มีจำนวนสมาชิกถึง 2,671 แถว ซึ่งในกรณีศึกษานี้หากจำนวนสมาชิกภายในกลุ่มมากเกินไปจะเป็นผลทำให้การระบุคำสำคัญเป็นไปได้ยากและมีโอกาสสูงที่จะมีคำสำคัญที่ปะปนกัน จึงมีแนวคิดที่จะทดลองจัดกลุ่มซ้ำ ซึ่งจะวิจารณ์ต่อในส่วนที่ 2.5 หน้า 71

2.4 การสกัดคำสำคัญและพิจารณาผู้เชี่ยวชาญของแต่ละกลุ่ม

เมื่อจัดกลุ่มจนได้สมาชิกในแต่ละกลุ่มแล้ว ต่อมาจะทำการสกัดคำสำคัญเฉพาะในแต่ละกลุ่ม ในกรณีศึกษานี้จะใช้วิธีการที่สะดวกที่สุด คือ การพิจารณาจากความถี่คำสูงสุด 5 อันดับ เนื่องจากหากพิจารณาความถี่สูงสุดในอันดับที่มากกว่านี้จะทำให้มีโอกาสพบคำสำคัญหลายคำปรากฏอยู่ในความถี่คำที่เท่ากัน ทำให้คำสำคัญที่ได้มีโอกาสขัดแย้งกันเองภายในกลุ่ม

จากในตารางที่ 17 หน้า 60 พบว่าบางกลุ่มมีคำสำคัญที่เหมือนกัน แต่จะพิจารณาให้ผู้เชี่ยวชาญถูกชี้ขาดโดยผลการจัดกลุ่มเนื่องจากผู้เชี่ยวชาญคนนั้น ๆ อาจมีคำสำคัญอื่นที่สัมพันธ์กับคำสำคัญในกลุ่มที่ถูกจัดให้อยู่

หากพิจารณาตารางที่ 16 ร่วมกับตารางที่ 17 จะพบว่าบางกลุ่มมีความถี่ของคำสำคัญน้อยแต่จำนวนสมาชิกกลับมีมาก ดังเช่น กลุ่มที่ 6 ทั้งนี้ความถี่คำสำคัญดังกล่าวมีการกระจายไปยังสมาชิกในกลุ่มด้วยความถี่คำที่น้อยต่อข้อมูลหนึ่งแถว เช่น คำว่า “Economic” อาจมีปรากฏในสมาชิกทั้ง 124 แถว (ตามจำนวนความถี่ที่สกัดได้) ในขณะที่คำว่า “สอน” อาจมีทั้งปรากฏคาบเกี่ยวบางแถวใน 124 แถวเดียวกับคำแรกหรืออาจมีโอกาสปรากฏที่แถวอื่นที่ไม่มีคำแรกปรากฏหรือสมาชิกบางแถวอาจมีคำสำคัญนั้น ๆ เพียงคำเดียวในแถวนั้น ๆ เป็นต้น หรืออาจมีอีกกรณีหนึ่งที่ทำให้เกิดกลุ่มใหญ่ คือ ข้อมูลถูกตัดคุณลักษณะไปในขั้นตอนการคัดเลือกคุณลักษณะ จากเดิมมีเพียงหนึ่งคุณลักษณะอยู่แล้ว เมื่อถูกตัดคุณลักษณะออกไปอีกจึงมีโอกาสสูงที่จะถูกจัดให้อยู่ในกลุ่มใหญ่ และในขณะที่บางกลุ่มมีความถี่ของคำสำคัญมากแต่กลับมีจำนวนสมาชิกน้อยก็จะมีลักษณะความถี่คำสำคัญที่ตรงกันข้าม คือ ข้อมูลในหนึ่งแถวมีความถี่ของคำสำคัญนั้น ๆ มาก เช่น แถวหนึ่ง ๆ อาจมีคำว่า “ป่า” ปรากฏถึง 3 ครั้ง เป็นต้น

ในตารางที่ 18 หน้า 61 ซึ่งได้แสดงเปอร์เซ็นต์ของผู้เชี่ยวชาญในแต่ละกลุ่ม เมื่อพิจารณาร่วมกับคำสำคัญจะพบว่าคณะต้นสังกัดของผู้เชี่ยวชาญกับคำสำคัญที่แสดงความเชี่ยวชาญของกลุ่มไม่ตรงกัน เนื่องจากผู้เชี่ยวชาญบางคนมีงานวิจัยอยู่เป็นจำนวนมากจึงยังผลให้ในขั้นตอนการสกัดคำสำคัญมีโอกาสที่จะเกิดคำสำคัญที่เด่นไปทางนั้น ทำให้คลาดเคลื่อนกับสายงาน อีกสิ่งหนึ่งที่น่าสังเกต คือ เนื่องจากมีผู้เชี่ยวชาญจากกลุ่มสายงานเกษตร อุตสาหกรรมเกษตรจำนวนมากและมีความเชี่ยวชาญหลากหลายจึงทำให้มีโอกาสปรากฏกระจายไปตามกลุ่มต่าง ๆ ได้มากกว่า

ผู้เชี่ยวชาญจากกลุ่มสายงานอื่น แต่สามารถสังเกตได้อีกนัยหนึ่ง คือ ถึงแม้ผู้เชี่ยวชาญจากกลุ่มสายงานเกษตร อุตสาหกรรมเกษตรจะกระจายไปยังกลุ่มต่าง ๆ มากแต่อาจเป็นคำสำคัญที่สัมพันธ์กับกลุ่มสายงานอื่น ๆ ภายในกลุ่มซึ่งสามารถสังเกตได้จากตารางที่ 19 หน้า 63

2.5 การจัดกลุ่มครั้งที่ 2

จากผลการจัดกลุ่มครั้งที่ 2 โดยใช้อัลกอริทึม K-Means แบบไม่ให้น้ำหนักคุณลักษณะ ในตารางที่ 20 หน้า 64 พบว่าค่า F-Statistic และค่า RS ซึ่งเป็นตัววัดประสิทธิภาพของข้อมูลระหว่างกลุ่มทำให้ค่าเป็น NaN ทั้งหมด จึงทำให้สามารถระบุได้ว่าไม่อาจคำนวณระยะทางระหว่างกลุ่มได้ ยังคงมีเพียงค่า Silhouette เท่านั้นที่สามารถคำนวณระยะทางของข้อมูลภายในกลุ่มได้ คือ 0.55707 จากการทดลองใช้อัลกอริทึม K-Means แบบไม่ให้น้ำหนักคุณลักษณะก่อนซึ่งได้ให้ผลการทดลองที่ติดค่า NaN จึงสามารถคาดการณ์ได้ว่าหากทำการจัดกลุ่มโดยให้น้ำหนักคุณลักษณะด้วยเทคนิคต่าง ๆ จะทำให้ระยะทางยิ่งใกล้มาก เนื่องจากตัววัดระยะทางของข้อมูลระหว่างกลุ่มให้ระยะทางเป็นค่าอนันต์ตั้งแต่ข้อมูลยังไม่ผ่านกระบวนการให้น้ำหนักคุณลักษณะ ซึ่งอาจมีข้อผิดพลาดในการประมวลผลเกิดขึ้น ดังนั้นจึงทำการจัดกลุ่มครั้ง 2 โดยใช้อัลกอริทึม K-Means แบบไม่ให้น้ำหนักคุณลักษณะเท่านั้น ในการพิจารณาความเชี่ยวชาญและผู้เชี่ยวชาญในแต่ละกลุ่มจากตารางที่ 22 พบว่าในกลุ่มที่ 1 กลุ่มที่ 5 กลุ่มที่ 6 กลุ่มที่ 7 กลุ่มที่ 8 และกลุ่มที่ 9 ยังคงมีการปะปนกันของคำสำคัญและผู้เชี่ยวชาญในสายงานต่าง ๆ กันสูงมากกว่าการจัดกลุ่มในครั้งแรก ทำให้เพิ่มน้ำหนักความเชื่อมั่นจากผลการจัดกลุ่มครั้งที่ 2 ว่าให้ผลการจัดกลุ่มที่มีแนวโน้มไม่ดีร่วมกับการวัดประสิทธิภาพจากค่า Silhouette

สรุปและข้อเสนอแนะ

สรุป

งานวิจัยนี้เป็นการศึกษาการจัดกลุ่มผู้เชี่ยวชาญในมหาวิทยาลัยของรัฐบาลแห่งหนึ่งโดยอาศัยเทคนิคการทำเหมืองข้อมูล โดยเทคนิคการทำเหมืองข้อมูลที่นำมาใช้ คือ การจัดกลุ่ม ซึ่งการจัดกลุ่มที่นำมาใช้ ได้แก่ การจัดกลุ่มแบบอัลกอริทึม 2 ขั้นตอน คือ ขั้นตอนการหาจำนวนกลุ่มที่เหมาะสม และขั้นตอนการจัดกลุ่มสำหรับการวัดประสิทธิภาพของกลุ่มวัดจากค่า F-Statistic ค่า Silhouette และค่า R-Squared (RS)

ชุดข้อมูลเบื้องต้นที่ใช้ในการทดลองนี้ได้รวบรวมจาก 3 ฐานข้อมูล คือ ฐานข้อมูลบุคลากร ฐานข้อมูลผลงานวิจัยและฐานข้อมูลวิทยานิพนธ์ การจัดการกับข้อมูลเบื้องต้น การกำหนดคุณลักษณะจากคำสำคัญและแถวของข้อมูล โดยในที่นี้ได้กำหนดให้ผู้เชี่ยวชาญหนึ่งคนเท่ากับหนึ่งแถวและกำหนดให้คุณลักษณะหนึ่งคุณลักษณะเท่ากับหนึ่งคอลัมน์ จุดเด่นของข้อมูลชุดนี้ คือ ข้อมูลได้ผ่านการกำหนดคำสำคัญโดยผู้เชี่ยวชาญ จึงสามารถนำคำสำคัญมาใช้ได้ ดังนั้นข้อมูลเบื้องต้นมีจำนวนแถวเท่ากับ 3,194 แถว และจำนวนคุณลักษณะเท่ากับ 971 คุณลักษณะ

ในกรณีศึกษานี้ได้สังเกตเห็นว่าข้อมูลที่นำมาใช้มีคุณลักษณะขนาดใหญ่ คือ 3,194 แถว 971 คุณลักษณะ จึงได้เล็งเห็นว่าการคัดเลือกคุณลักษณะน่าจะเป็นหนทางหนึ่งที่ช่วยลดขนาดของคุณลักษณะลง โดยกำหนดคำตอบของปัญหาไว้ว่า *ข้อมูลหลังจากที่ผ่านกระบวนการคัดเลือกคุณลักษณะแล้วจะต้องไม่ให้ประสิทธิภาพในการจัดกลุ่มที่แย่กว่าข้อมูลที่ยังไม่ผ่านการคัดเลือกคุณลักษณะ* จึงได้นำเอาวิธีการ CFS และ GA มาช่วยในการคัดเลือกคุณลักษณะ จากผลการทดลองพบว่าจำนวนคุณลักษณะเท่ากับ 258 คุณลักษณะ ให้ค่าความแข็งแรงของประชากรดีที่สุด กล่าวคือคุณลักษณะของประชากรในกรณีศึกษามีความสัมพันธ์กันดีที่สุด ดังนั้นข้อมูลที่ใช้ในการทดลองจนถึงขั้นตอนนี้ คือ 3,194 แถว 258 คุณลักษณะ

หลังจากที่ข้อมูลถูกลดขนาดของคุณลักษณะลงแล้ว ในขั้นตอนต่อไปจะเป็นการหาจำนวนกลุ่มที่เหมาะสม ในที่นี้จะใช้อัลกอริทึม SOM ในการหาจำนวนกลุ่ม โดยจำนวนกลุ่มที่เหมาะสมของกรณีศึกษานี้เท่ากับ 7 กลุ่ม นอกจากนี้ยังได้ทำการทดลองเปรียบเทียบประสิทธิภาพในการจัด

กลุ่มระหว่างข้อมูลที่ผ่านการคัดเลือกคุณลักษณะแล้ว (โดย CFS และ GA) กับข้อมูลที่ยังไม่ผ่านการคัดเลือกคุณลักษณะ จากผลการทดลองพบว่าข้อมูลหลังจากที่ผ่านกระบวนการคัดเลือกคุณลักษณะให้ประสิทธิภาพในการจัดกลุ่มดีขึ้น ถึงแม้ว่าค่า F-Statistic จะดีกว่าข้อมูลที่ยังไม่ผ่านการคัดเลือกคุณลักษณะเล็กน้อยก็ตาม แต่หากพิจารณาค่า F-Statistic และค่า Silhouette ร่วมกันจะเห็นได้ว่าค่า Silhouette จึงเป็นตัวช่วยชี้ขาดในการสนับสนุนผลการทดลอง เนื่องจากค่า Silhouette ดีขึ้นอย่างเห็นได้ชัด (จากค่าติดลบเป็นค่าติดบวก) จึงถือได้ว่าคำตอบของปัญหานี้ คือ ข้อมูลที่ผ่านการคัดเลือกคุณลักษณะเป็นสิ่งที่ยอมรับได้ และนำมาใช้ในการทดลองในขั้นตอนต่อไปด้วยจำนวนแถว 3,194 แถว 258 คุณลักษณะ ด้วยจำนวนกลุ่มที่เหมาะสมจำนวน 7 กลุ่ม

เทคนิคการให้น้ำหนักคุณลักษณะเป็นเทคนิคหนึ่งที่สามารถนำมาใช้ประยุกต์ในการทำงานด้านการทำเหมืองข้อมูล โดยเฉพาะการจัดกลุ่ม ดังนั้นในการจัดกลุ่มจึงได้นำเอาเทคนิคการให้น้ำหนักคุณลักษณะมาประยุกต์ใช้ร่วมกัน คือ (1) อัลกอริทึม K-Means แบบไม่ให้น้ำหนักคุณลักษณะ, (2) อัลกอริทึม K-Means แบบให้น้ำหนักคุณลักษณะด้วยวิธีการ TF-IDF, (3) อัลกอริทึม K-Means แบบให้น้ำหนักคุณลักษณะด้วยวิธีการ Logarithm Weight, (4) อัลกอริทึม K-Means แบบให้น้ำหนักคุณลักษณะด้วยวิธีการ Augmented Weight, (5) อัลกอริทึม Fuzzy C-Means แบบไม่ให้น้ำหนักคุณลักษณะ, (6) อัลกอริทึม Fuzzy C-Means แบบให้น้ำหนักคุณลักษณะด้วยวิธีการ TF-IDF, (7) อัลกอริทึม Fuzzy C-Means แบบให้น้ำหนักคุณลักษณะด้วยวิธีการ Logarithm Weight และ (8) อัลกอริทึม Fuzzy C-Means แบบให้น้ำหนักคุณลักษณะด้วยวิธีการ Augmented Weight จากผลการทดลองพบว่าอัลกอริทึม K-Means แบบให้น้ำหนักคุณลักษณะด้วยวิธีการ Logarithm Weight ให้ประสิทธิภาพในการจัดกลุ่มดีที่สุด

เมื่อจัดกลุ่มจนได้สมาชิกในแต่ละกลุ่มแล้ว ขั้นตอนต่อไป คือ การสกัดคำสำคัญในแต่ละกลุ่มออกมา ในที่นี้จะพิจารณาจากความถี่คำสูงสุด 5 อันดับแรก เนื่องจากหากพิจารณาอันดับความถี่คำสำคัญมากกว่า 5 อันดับจะมีโอกาสสูงที่จะเจอคำที่มีความถี่เท่ากัน ซึ่งทำให้การพิจารณาคำสำคัญเป็นไปได้ยาก เนื่องจากคำสำคัญที่ปรากฏอาจเป็นคำที่ไม่เกี่ยวเนื่องกันหรือเป็นคำสำคัญที่ขัดแย้งกันเอง คำสำคัญภายในกลุ่มไม่เป็นไปในทางเดียวกัน ทั้งนี้ในการกำหนดว่าจะพิจารณาคำสำคัญกี่อันดับนั้นจึงขึ้นอยู่กับลักษณะปัญหาด้วย จากคำสำคัญที่สามารถสกัดได้จากการจัดกลุ่มครั้งที่ 1 สามารถสรุปได้ดังนี้

กลุ่มที่ 1 คำสำคัญ ได้แก่ เศรษฐศาสตร์, Economic, ทุน, เงิน, Finance
ประกอบด้วยผู้เชี่ยวชาญจำนวน 50 คน จากกลุ่มสายงานเกษตรฯ, วิศวกรรม, สัตว์แพทย์ศาสตร์

กลุ่มที่ 2 คำสำคัญ ได้แก่ ป่า, Generate, ปลา, ไข่, ท่องเที่ยว
ประกอบด้วยผู้เชี่ยวชาญจำนวน 249 คน จากกลุ่มสายงานสำนักงานภายใน, เกษตรฯ, ศึกษาศาสตร์

กลุ่มที่ 3 คำสำคัญ ได้แก่ Rice, ร้อน, คาร์บอน, ไดออกไซด์, สตาร์ช
ประกอบด้วยผู้เชี่ยวชาญจำนวน 63 คน จากกลุ่มสายงานเกษตรฯ, วิศวกรรม, ศึกษาศาสตร์

กลุ่มที่ 4 คำสำคัญ ได้แก่ Corn, หวาน, Baby, Sweet, Rice
ประกอบด้วยผู้เชี่ยวชาญจำนวน 11 คน จากกลุ่มสายงานเกษตรฯ, ศึกษาศาสตร์, สัตว์แพทย์ศาสตร์

กลุ่มที่ 5 คำสำคัญ ได้แก่ สอน, Student, Admin, ช่วยสอน, เรียน
ประกอบด้วยผู้เชี่ยวชาญจำนวน 43 คน ผู้เชี่ยวชาญจากกลุ่มสายงานวิศวกรรม, เกษตรฯ, ศึกษาศาสตร์

กลุ่มที่ 6 คำสำคัญ ได้แก่ Economic, สอน, เศรษฐศาสตร์, Genetic, Industrial
ประกอบด้วยผู้เชี่ยวชาญจำนวน 2,671 คน จากกลุ่มสายงานเกษตรฯ, วิศวกรรม, วิทยาศาสตร์

กลุ่มที่ 7 คำสำคัญ ได้แก่ ไฟฟ้า, Industrial, ไร้สาย, Fire, Safe
ประกอบด้วยผู้เชี่ยวชาญจำนวน 107 คน จากกลุ่มสายงานเกษตรฯ, วิศวกรรม, สังคมศาสตร์, ศึกษาศาสตร์

เมื่อพิจารณาจากจำนวนสมาชิกที่ประกอบอยู่ในกลุ่มจะพบว่า มีกลุ่มที่ 6 ที่มีขนาดกลุ่มที่ใหญ่ มีจำนวนสมาชิกถึง 2,671 แถว ซึ่งจัดได้ว่าเป็นกลุ่มที่มีขนาดใหญ่มาก เนื่องจากทำให้ไม่สามารถแยกแยะผู้เชี่ยวชาญได้ชัดเจน ในการแก้ไขปัญหาสมาชิกของกลุ่มที่มีขนาดใหญ่ วิธีการจัดการกับข้อมูลกลุ่มใหญ่ได้ง่ายที่สุด จะทดลองทำการจัดกลุ่มครั้งที่ 2 ในที่นี้จะทดลองใช้กับอัลกอริทึม K-Means แบบไม่ให้น้ำหนักคุณลักษณะก่อน เนื่องจากหากผลการจัดกลุ่มโดยไม่ให้น้ำหนักคุณลักษณะมีผลการประมวลผลออกมาไม่ดี จะสามารถหยุดการประมวลผลได้ทันทีโดยที่

ไม่ต้องทำการประมวลผลต่อกับอัลกอริทึมอื่น จากผลการทดลองค่า F-Statistic และ RS ไม่สามารถให้ค่าที่วัดออกมาได้ จึงหยุดการประมวลผลในการจัดกลุ่มครั้งที่ 2 ถึงแม้ว่าจะมีกลุ่มใหญ่เกิดขึ้นอีกก็ตาม และเมื่อพิจารณาคำสำคัญและผู้เชี่ยวชาญภายในกลุ่มแต่ละกลุ่มยังคงมีการปะปนกันของสายงานและความเชี่ยวชาญที่ไม่ควรจะเกี่ยวข้องกัน

ข้อเสนอแนะ

สามารถทดลองจัดกลุ่มผู้เชี่ยวชาญโดยอาศัยพจนานุกรมคำเหมือนภาษาไทย (Thesaurus) มาช่วยการจัดกลุ่มให้ละเอียดมากขึ้น แต่ในปัจจุบันนี้พจนานุกรมคำเหมือนภาษาไทยที่ใช้ในด้านต่าง ๆ ยังมีให้ใช้อ้างอิงน้อย ที่มีอยู่ตอนนี้ คือ พจนานุกรมคำเหมือนศัพท์ด้านเกษตร โดยมหาวิทยาลัยเกษตรศาสตร์ เท่านั้น นอกจากนี้ยังสามารถนำเอา Ontology มาช่วยในการกำหนดน้ำหนักของคำสำคัญเพื่อใช้ในการจัดกลุ่มต่อในขั้นต่อไปได้

เอกสารและสิ่งอ้างอิง

วงกต พจน์พงศ์สรรค์. 2551. **การจัดกลุ่มข้อมูลด้านความปลอดภัยของอาหารโดยวิธีการแบบสองขั้นตอน.** วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์.

สุคนธ์ทิพย์ วงศ์พันธ์. 2551. **การเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะที่เหมาะสมและอัลกอริทึมเพื่อจำแนกพฤติกรรมการกระทำความผิดของนักเรียนระดับอาชีวศึกษา.** วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์.

ศศิธร มงคลศรีพัฒนา. 2550. **การจัดกลุ่มศูนย์บริการและถ่ายทอดเทคโนโลยีทางการเกษตรประจำตำบล ในประเทศไทยโดยใช้อัลกอริทึม 2 ขั้นตอนคือ SOM และ Fuzzy C-Means.** วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์.

Aarabi, A., F. Wallois and R. Grebe. 2005. Automated Neonatal Seizure Detection: A Multistage Classification System through Feature Selection based on Relevance and Redundancy Analysis. **International Federation of Clinical Neurophysiology (IFCN)** 117 (2): 328-340.

Cios, K.J., W. Pedrycz, R.W. Swiniarski and A. Kurgan. 2007. **Data Mining: A Knowledge Discovery Approach**. Springer Science + Business Media, LLC.

Fanizzi, N., C. d'Amato and F. Esposoto. 2007. Randomized Metric Induction and Evolutionary Conceptual Clustering for Semantic Knowledge Bases, pp. 51-60. *In* **Conference on Information and Knowledge Management (CIKM).** Lisbon, Portugal.

Fent, T. 1999. **Using Genetics Based Machine learning to find Strategies for Product Placement in a dynamic Market.** Available Source: http://epub.wu.ac.at/694/1/document.pdf, August 31, 2009.

Ginsparg, P. 2009. **IR 5: Scoring, Term Weighting, The Vector Space Model II.** Available Source: http://www.infosci.cornell.eduCoursesinfo43002009faslides05.pdf, January 10, 2010.

Gong, Y. and X. Liu. 2001. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis, pp. 19-25. *In* **Annual ACM Conference on Research and Development in Information Retrieval.** New Orleans, USA.

Hall, M. 1999. **Correlation-based Feature Selection for Machine Learning.** Ph.D. Thesis, University of Waikato. Available Source: http://www.cs.waikato.ac.nz/~mhall/thesis.pdf, December 21, 2009.

_____ and L. A. Smith. 1997. Feature Subset Selection: A Correlation Based Filter Approach, pp. 855-858. *In* **Neural Information Processing and Intelligent Information Systems**. New Zealand.

Han. J and M. Kamber. 2001. **Data Mining: concept and techniques**. Academic Press.

Kovacs, F., C. Legany and A. Babos. 2006. Cluster Validity Measurement Techniques, pp. 388-393. *In* **World Scientific and Engineering Academy and Society (WSEAS)**.

Lan, M., C. L. Tan, H. B. Low and S. Y. Sung. 2005. A Comprehensive Comparative Study on Term Weighting Schemes for Text Categorization with Support Vector Machines, poster session pp. 1032-1033. *In* **International World Wide Web Conference Special interest tracks and posters of the 14th international conference on World Wide Web.** Japan.

Madahvi, M., M. H. Chehreghani, H. Abolhassani and R. Forsati. 2008. Novel meta-heuristic algorithms for clustering web documents. **Applied Mathematics and Computation**. 201: 441-451.

Manning, C. D., P. Raghavan and H. Schutze. 2008. **Introduction to Information Retrieval**. Available Source: http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf, June 12, 2008.

Matignon, R. 2007. **Data mining using SAS Enterprise miner.** Wiley – Interscience Press.

Nhien–An Le–Khac, L. M. Aouad and M–Tahar Kechadi. 2007. Knowledge Map: Toward a new approach supporting the knowledge management in Distributed Data Mining, pp. 67-72. *In* **International Council of the Aeronautical Sciences (ICAS)**. Alaska.

Polettini, N. 2004. **The Vector Space Model in Information Retrieval-Term Weighting Problem**. Available Source: http://sra.itc.it/people/polettini/PAPERS/Polettini_Information_Retrieval.pdf, December 2009.

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

Tibshirani, R. G. Walther and T. Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society, Series B**. 63: 411-423.

Wang, Y., I. Tetko., M. Hall., E. Frank., A. Facius., K. Mayer and H. Mewes. 2004. Gene Selection from Microarray Data for Cancer Classification—A Machine Learning approach. **Computational Biology and Chemistry.** 29 (1): 37-46.

Wongpun, S. and Srivihok, A. 2008. Comparison of Attribute Selection Techniques and Algorithms in Classifying Bad Behaviors of Vocational Education Students, pp. 526-531. *In* **Proceeding of Second IEEE International Conference on Digital Ecosystems and Technologies.** Phitsanuloke, Thailand.

Zhang, W., T. Yoshida and X. Tang. 2008. TFIDF, LSI and muti-word in Information Retrieval and Text Categorization, pp. 108-113. *In* **IEEE conference on Systems, Man and Cybernetics**. Singapore.

Zhao, H. and W. Lu. 2007. Using Document Weight Combining Method for Enterprise Expert Mining, pp. 3721-3723. *In* **Wireless Communications, Networking and Mobile Computing (WiCom)**. New York, USA.

Zheng, X., P. He and F. Yuan. 2003. Algorithm of documents clustering based in Minimum Spanning Tree, pp. 199-203. *In* **Machine Learning and Cybernetics**. China.

ภาคผนวก

ตารางผนวกที่ 1 คำสำคัญที่มีการกำหนดรหัสอ้างอิงคำสำคัญ (ข้อมูล 3,194 แถว 971 คุณลักษณะ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
1	Account	244	ศิลปะ	487	ขมิ้นชัน	730	สุกร
2	Admin	245	เศรษฐศาสตร์	488	ขยะ	731	สุขภาพ
3	Advertise	246	สถาปัตยกรรม	489	ข่าว	732	สุญญากาศ
4	Agriculture	247	สถิติ	490	ข้าว	733	สุนัข
5	Agro	248	สภาพแวดล้อม	491	ข้าวโพด	734	สุรา
6	Anatomic	249	สรีรวิทยา	492	ขิง	735	ไส้เดือน
7	Animal	250	สวน	493	เข้มข้น	736	หญ้า
8	Aqua	251	สหกรณ์	494	เขื่อน	737	หญิง
9	Architecture	252	สอน	495	โขง	738	หนอน
10	Area	253	สังคม	496	ไข่	739	หน่อไม้
11	Art	254	สัตว์	497	ไขมัน	740	หนังสือ
12	Bank	255	สากล	498	ไข้หวัด	741	หนู
13	Bio	256	สาธารณสุข	499	คน	742	หม่อน
14	Business	257	สาธารณะ	500	ครอบครัว	743	หมัก
15	Chem	258	สารสนเทศ	501	ครัว	744	หมู่บ้าน

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
16	Child	259	สิ่งทอ	502	ครู	745	ห้วย
17	China	260	สิ่งแวดล้อม	503	คลอง	746	หวาย
18	Civil	261	สื่อสาร	504	คลอไรด์	747	หอมมะลิ
19	Clinic	262	สุขศึกษา	505	คลื่น	748	หอมระเหย
20	Communicate	263	โสตทัศนศึกษา	506	ความดัน	749	หอย
21	Computer	264	หัตถ	507	คอนกรีต	750	หัวใจ
22	Crop	265	อณู	508	คอมโพสิท	751	หัวหิน
23	Culture	266	อนุบาล	509	คะน้า	752	หิน
24	Economic	267	อวกาศยาน	510	ค้า	753	เห็ด
25	Educate	268	ออกแบบ	511	คามาลดูเลนซิส	754	เห็บ
26	Electric	269	อักษร	512	คาร์บอน	755	เหมือง
27	Energy	270	อังกฤษ	513	คาร์โบไฮเดรต	756	เหล็ก
28	Engineer	271	อาคาร	514	คุ้มกัน	757	ไหม
29	English	272	อาชีวศึกษา	515	คู่มือ	758	ไหม้

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
30	Entomopathogenic	273	อายุรศาสตร์	516	เค็ม	759	องุ่น
31	Environ	274	อาหาร	517	เครียด	760	อนุภาค
32	Family	275	อินเตอร์เน็ต	518	เครือข่าย	761	อนุมูล
33	Farm	276	อินทรีย์	519	เครื่องดื่ม	762	อนุรักษ์
34	Finance	277	อิเล็กทรอนิกส์	520	เครื่องมือ	763	อลูมิเนียม
35	Fish	278	อุดมศึกษา	521	แคดเมียม	764	อสุจิ
36	Food	279	อุตสาหกรรม	522	แคลเซียม	765	ออกซิเจน
37	Foreign	280	อุทยาน	523	โครงการหลวง	766	ออนไลน์
38	Forest	281	Acid	524	โครงข่าย	767	ออมทรัพย์
39	France	282	Artificial	525	ไคโตซาน	768	อ้อย
40	Generate	283	Asia	526	จระเข้	769	ออสโมซิส
41	Genetic	284	Atmosphere	527	จริยธรรม	770	อะไมเลส
42	Geographic	285	Avian	528	จลน์	771	อันตราย
43	German	286	bacteria	529	จุลินทรีย์	772	อากาศ

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
44	Govern	287	Basin	530	เจล	773	อ่างเก็บน้ำ
45	Harvest	288	Benefit	531	เจลาติน	774	อาชีพ
46	Health	289	body	532	เจ้าพระยา	775	อาทิตย์
47	History	290	Breed	533	ช้าง	776	อ่าน
48	Home	291	broiler	534	ชายฝั่ง	777	อายุ
49	Horticulture	292	Build	535	ชายเลน	778	อ่าว
50	Hospital	293	Cadmium	536	ชีวิต	779	อินฟาเรด
51	Human	294	camaldulensis	537	ชื้น	780	อิมัลชัน
52	Industrial	295	carbon	538	ชุดวาริน	781	อีรี่
53	Internet	296	Catfish	539	เชื้อ	782	อุณหภูมิ
54	Irrigate	297	Cell	540	เชื้อเพลิง	783	อุทก
55	Japan	298	Chicken	541	เชื้อรา	784	อุปทาน
56	Language	299	Chitosan	542	ซอฟต์แวร์	785	เอกชน
57	Law	300	Clarian	543	ซัลเฟอร์	786	เอชไอวี

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
58	Life	301	Climate	544	ซิลิกอน	787	เอทานอล
59	Logistic	302	Clone	545	ซีเมนต์	788	เอทิลีน
60	Machineries	303	Cluster	546	ซีโอไลต์	789	เอนไซม์
61	Manage	304	coat	547	ซื้อ	790	เอ็นไซม์
62	Marin	305	coconut	548	เซลล์	791	แอซิด
63	Market	306	Coli	549	เซลลูโลส	792	แอนแทรคโนส
64	Math	307	Concrete	550	โซเดียม	793	แอมโมเนีย
65	Meat	308	Consume	551	โซล	794	แอลกอฮอล์
66	Mechanic	309	Cost	552	ไซยาไนด์	795	แอลฟา
67	Medicine	310	Cotton	553	ฐานข้อมูล	796	โอโซน
68	Metal	311	Curcuma	554	ดอก	797	ไอน้ำ
69	Micro	312	Custom	555	ดอง	798	ไอโอดีน
70	Molecular	313	degrade	556	ด่าง	799	ฮอร์โมน
71	Nature	314	Diesel	557	ดาวเทียม	800	ไฮโดร

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
72	Network	315	Diospyros	558	ดิจิตอล	801	Accident
73	Nutrit	316	dioxide	559	ดีเซล	802	Accumulate
74	Paper	317	Disease	560	ดีเอ็นเอ	803	Agrotour
75	Pathologic	318	Divers	561	ไดออกไซด์	804	Air
76	Philosophic	319	DNA	562	ไดอะทอไมต์	805	Alien
77	Physic	320	Domestic	563	ตะกอน	806	Ammonia
78	Plant	321	Dry	564	ตะกั่ว	807	Andrograph
79	Polite	322	Drug	565	ตะไคร้	808	Antagonist
80	Pollute	323	Dye	566	ถั่ว	809	Antimicrobial
81	Polymer	324	Egg	567	ถ่าน	810	Antioxidant
82	Populate	325	Enzyme	568	ทรัพย์สิน	811	Apple
83	Print	326	Ethanol	569	ทราย	812	Asparagus
84	Psychology	327	Ethylene	570	ทอง	813	Baby
85	Radio	328	Eucalyptus	571	ท้องถิ่น	814	Bacillus

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
86	Read	329	Fabric	572	ทานตะวัน	815	Banana
87	Recreate	330	Fabricius	573	ทุเรียน	816	Book
88	Region	331	Field	574	เทอร์โม	817	Budget
89	Resource	332	Film	575	โทรศัพท์	818	Calcium
90	Rural	333	Fingerprint	576	ธรรม	819	Cane
91	Sanskrit	334	Fire	577	ธาตุ	820	Carbohydrate
92	School	335	flour	578	นก	821	Cattle
93	Science	336	Flow	579	นักศึกษา	822	Cement
94	Sell	337	flower	580	นาโน	823	Chromatography
95	Social	338	Fresh	581	น้ำดอกไม้	824	Comic
96	Soil	339	Fruit	582	น้ำตาล	825	Commerce
97	Space	340	Fungi	583	น้ำมัน	826	Commiss
98	Sport	341	GAME	584	นิล	827	Construct
99	Statist	342	Gas	585	นิเวศน์	828	Corn

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
100	Structure	343	GIS	586	นิสิต	829	Crab
101	Sugar	344	gloeosporioide	587	เนื้อ	830	Crocodile
102	Teach	345	Gold	588	ไนโตรเจน	831	Cultivar
103	Technology	346	Graduate	589	บรรยากาศ	832	Curriculum
104	Telecommun	347	Grass	590	บัว	833	Democracy
105	Textile	348	GROUND	591	บ้าน	834	Deposit
106	Transport	349	HIV	592	บุคลากร	835	Ethic
107	Tropic	350	hydro	593	บุคลิกภาพ	836	Export
108	Urban	351	Influenza	594	เบต้า	837	Factor
109	Vocation	352	Investment	595	เบียร์	838	Fatigue
110	Water	353	ion	596	แบคทีเรีย	839	Feed
111	Wildlife	354	IRON	597	ใบ	840	Female
112	Wood	355	lactic	598	ปนเปื้อน	841	Fertile
113	กฎหมาย	356	Land	599	ประชาชน	842	Fever

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
114	ก่อสร้าง	357	larva	600	ประสาท	843	Fit
115	กายวิภาค	358	Leaf	601	ปลอดภัย	844	Fraction
116	การเมือง	359	Learn	602	ปลา	845	Fuel
117	กีฏวิทยา	360	light	603	ปลีก	846	Fund
118	กีฬา	361	Liquid	604	ปลูก	847	Gamma
119	เก็บเกี่ยว	362	Lotus	605	ป่วย	848	Grade
120	เกษตร	363	machine	606	ปาล์ม	849	Head
121	ขาย	364	Macrobrachium	607	ปุ๋ย	850	Heat
122	เขตร้อน	365	Maize	608	ปู	851	Herb
123	คณิต	366	Mango	609	ปูน	852	Image
124	คลัง	367	Mangrove	610	เปิด	853	Import
125	คลินิก	368	media	611	แป้ง	854	Income
126	คหกรรม	369	Methane	612	โปรตีน	855	Internet
127	คอมพิวเตอร์	370	Methionine	613	โปรไบโอติก	856	Japan

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
128	เคมี	371	milk	614	ผงชูรส	857	Juice
129	เครื่องกล	372	Mill	615	ผลไม้	858	Magazine
130	โค	373	miner	616	ผัก	859	Membrane
131	โครงสร้าง	374	Mine	617	ผิวหนัง	860	Metropolitan
132	โฆษณา	375	Miticide	618	ผึ้ง	861	Moisture
133	เงิน	376	monodon	619	ไผ่	862	Mont
134	จัดการ	377	morphology	620	ฝน	863	Multimedia
135	จิตวิทยา	378	Motor	621	ฝรั่ง	864	Municipal
136	จีน	379	Mushroom	622	ฝ้าย	865	Muscle
137	ชนบท	380	mussel	623	แฝก	866	Music
138	ชลประทาน	381	Mutation	624	พรรณ	867	Nanotube
139	ช่างยนต์	382	Neural	625	พริก	868	New
140	ชีววิทยา	383	niloticus	626	พลวัตร	869	Nile
141	ชุมชน	384	Noodle	627	พลศาสตร์	870	Nitrogen

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
142	เซรามิก	385	oil	628	พลาสติก	871	nurse
143	ญี่ปุ่น	386	Oreochromis	629	พาณิชย์	872	Nutrient
144	ดนตรี	387	Oryzae	630	พารา	873	Organic
145	ดิน	388	ostreatus	631	พีซีอาร์	874	Oxidation
146	ดุริยางค์	389	oxide	632	พื้นที่	875	Palm
147	เด็ก	390	Park	633	พื้นบ้าน	876	paniculata
148	ตลาด	391	PCR	634	เพจ	877	Parallel
149	ต่างประเทศ	392	Penaeus	635	เพศ	878	Parent
150	ทรัพยากร	393	People	636	แพ	879	Pasak
151	ท่องเที่ยว	394	pesticide	637	แพลงก์ตอน	880	Past
152	ทะเล	395	Phylogenetic	638	โพแทสเซียม	881	Person
153	ทั่วไป	396	Pig	639	ฟาง	882	Pigment
154	ทุน	397	Pineapple	640	ฟ่าง	883	Plastic
155	เทคโนโลยี	398	Pleurotus	641	ฟ้าทะลายโจร	884	Pond

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
156	โทรคมนาคม	399	Pouch	642	ฟ้าผ่า	885	Posit
157	ธนาคาร	400	Power	643	ฟิล์ม	886	Powder
158	ธรณี	401	Price	644	ไฟ	887	Prawn
159	ธรรมชาติ	402	Probiotic	645	ภาษี	888	Prison
160	ธุรกิจ	403	Project	646	ภูเขา	889	Radiation
161	นม	404	Protein	647	มลพิษ	890	Reactor
162	นวัตกรรม	405	Pulp	648	มหาวิทยาลัย	891	Regress
163	นันทนาการ	406	RAPD	649	มอเตอร์	892	Rehabilitation
164	นา	407	Registration	650	มะขาม	893	Release
165	นาฏศิลป์	408	Residue	651	มะเขือเทศ	894	Remote
166	นานาชาติ	409	resin	652	มะพร้าว	895	Reservoir
167	น้ำ	410	Rice	653	มะม่วง	896	Retent
168	นิติ	411	River	654	มังคุด	897	Rock
169	นิเทศ	412	root	655	มัน	898	Royal

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
170	นิวเคลียร์	413	rosenbergii	656	มัลติมีเดีย	899	Rubber
171	เนื้อ	414	Salt	657	มาบบอน	900	Safe
172	แนะแนว	415	Sanctuary	658	มิติ	901	Saline
173	บรรณารักษ์	416	Sauce	659	มือ	902	Save
174	บริการ	417	screen	660	มุก	903	Scale
174	บริหาร	418	sediment	661	มูล	904	Secretariat
176	บัญชี	419	Seed	662	มูลฝอย	905	Secure
177	บิน	420	Shrimp	663	เมมเบรน	906	Shelf
178	ปกครอง	421	Silk	664	แม่กลอง	907	Shell
179	ปฐพี	422	Song	665	แม่น้ำ	908	Sludge
180	ปฐมวัย	423	soybean	666	แมลงภู่	909	Smoke
181	ประชากร	424	species	667	แมว	910	Sodium
182	ประชาสัมพันธ์	425	Spodoptera	668	แม่เหล็ก	911	Solar
183	ประถมศึกษา	426	Stabil	669	ไมโคร	912	Speech

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
184	ประเทศ	427	starch	670	ยา	913	Spot
185	ประมง	428	Steel	671	ยาง	914	Squamosa
186	ประวัติศาสตร์	429	Stress	672	ยาสูบ	915	Stock
187	ปรัชญา	430	Suppl	673	ยิปซั่ม	916	Student
188	ป่า	431	swine	674	ยีน	917	Sunflower
189	ปิโตรเคมี	432	Temperature	675	ยีสต์	918	Sweet
190	ผ้า	433	texture	676	ยุติธรรม	919	Term
191	ฝรั่งเศส	434	Tiger	677	ยูคาลิปตัส	920	Think
192	พณิชยการ	435	Trade	678	ยูเรีย	921	Tilapia
193	พยาธิ	436	Trait	679	เยาวชน	922	Tourist
194	พยาบาล	437	Trat	680	รถ	923	Traffic
195	พฤกษ	438	Tree	681	ร้อน	924	Transfer
196	พลศึกษา	439	Trichoderma	682	ระบาด	925	Treatment
197	พลังงาน	440	Tsunami	683	ระบาย	926	Undergraduate

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
198	พอลิเมอร์	441	tuna	684	ราก	927	Universe
199	พันธุ	442	Uvaria	685	ราคา	928	Variat
200	พิมพ์	443	Vegetable	686	ราบ	929	White
201	พิษ	444	Vehicle	687	รายได้	930	กัมมันตรังสี
202	พืช	445	Vibrio	688	รำ	931	เกม
203	เพลง	446	Vigna	689	เรือ	932	เกาะ
204	เพาะ	447	Village	690	แรงงาน	933	เขียน
205	แพทย์	448	viridis	691	โรงเรือน	934	คดี
206	ฟาร์ม	449	Virtual	692	ไร้สาย	935	ช่วยสอน
207	ฟิสิกส์	450	Virus	693	ฤดู	936	ชั้น
208	ไฟฟ้า	451	Volatile	694	ลม	937	ดุก
209	ภาพถ่าย	452	Web	695	ลาว	938	แดง
210	ภาษา	453	Wetland	696	ลำต้น	939	ตราสารหนี้
211	ภูมิศาสตร์	454	Wireless	697	ลำไส้	940	เถ้า

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
212	เภสัช	455	Write	698	ลิสง	941	เทศบาล
213	โภชนา	456	Zeolite	699	ลุ่ม	942	เทียม
214	มนุษย์	457	กรด	700	เลือด	943	นักเรียน
215	มวลชน	458	กรม	701	แลคติก	944	นำเข้า
216	มัธยมศึกษา	459	กระดาษ	702	แล้ง	945	นิเวศ
217	เมือง	460	กระทง	703	โลก	946	บริโภค
218	แมลง	461	กระเพาะ	704	โลชั่น	947	ประชาธิปไตย
219	โมเลกุล	462	กล้วย	705	ไลเปส	948	พลัง
220	ไม้	463	กล้วยไม้	706	วัคซีน	949	ภูมิ
221	เยอรมัน	464	กล้อง	707	วังน้ำเขียว	950	โภชนาการ
222	โยธา	465	กล้ามเนื้อ	708	วิตามิน	951	มูลค่า
223	รังสี	466	กลายพันธุ์	709	เว็บ	952	เม็ด
224	รัฐ	467	กลิ่น	710	แวนนาไม	953	เย็น
225	แร่	468	ก๋วยเตี๋ยว	711	ไวน์	954	รส

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
226	โรค	469	กวาวเครือขาว	712	ศาล	955	ราษฎร
227	โรงงาน	470	ก๊าซ	713	สตรอเบอรี่	956	เรียน
228	โรงเรียน	471	กาแฟ	714	สตาร์ช	957	โรง
229	โรงแรม	472	ก้ามกราม	715	สบู่ดำ	958	โรงพยาบาล
230	ไร่	473	กุ้ง	716	ส้ม	959	ลีลา
231	เลขานุการ	474	กุนเชียง	717	สมรรถภาพ	960	วรรณกรรม
232	โลจิสติกส์	475	กุลาดำ	718	สมุนไพร	961	ส่งออก
233	โลหะ	476	เกลือ	719	สวาย	962	เสพติด
234	วรรณคดี	477	เกสร	720	สหรัฐอเมริกา	963	เสียง
235	วัฒนธรรม	478	แกมมา	721	สะอาด	964	หลักทรัพย์
236	วัสดุ	479	แกลบ	722	สับปะรด	965	หวาน
237	วารสาร	480	แก้ว	723	สาลี	966	ออกกำลังกาย
238	วิทยาศาสตร์	481	ไก่	724	สาหร่าย	967	ออกซิเดชัน
239	วิศวกรรม	482	ไกลโคไซด์	725	สำปะหลัง	968	อัคคีภัย

ตารางผนวกที่ 1 (ต่อ)

รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ	รหัส	คำสำคัญ
240	เวช	483	ขนม	726	สินค้า	969	อารมณ์
241	ไวรัส	484	ขนมจีน	727	สึนามิ	970	อินเทอร์เน็ต
242	ศัตรู	485	ขนส่ง	728	สืบพันธุ์	971	อุปสงค์
243	ศาสนา	486	ขมิ้น	729	สื่อ		

หมายเหตุ รหัส หมายถึง รหัสของคำสำคัญ

จากตารางผนวกที่ 1 จะสังเกตได้ว่าคำบางคำมีความหมายเหมือนกัน เช่น “Sport” และ “กีฬา” ทั้งนี้สืบเนื่องจากในงานวิจัยบางงานวิจัยมีชื่องานวิจัยเฉพาะภาษาอังกฤษ บางงานวิจัยมีชื่อเฉพาะภาษาไทยจึงต้องกำหนดคำที่มีความหมายเหมือนกันแต่เขียนคนละภาษากันเป็นคำสำคัญทั้งคู่

ตารางผนวกที่ 2 คุณลักษณะที่ถูกคัดเลือกโดยใช้อัลกอริทึม CFS และ GA (ข้อมูล 3,194 แถว 258 คุณลักษณะ)

รหัส	คำสำคัญ	TF	รหัส	คำสำคัญ	TF	รหัส	คำสำคัญ	TF	รหัส	คำสำคัญ	TF
2	Admin	157	293	Cadmium	11	545	ซีเมนต์	7	831	Cultivar	8
6	Anatomic	12	294	camaldulensis	10	546	ซีโอไลต์	23	832	Curriculum	21
10	Area	10	299	Chitosan	10	549	เซลลูโลส	7	833	Democracy	9
16	Child	32	303	Cluster	15	561	ไดออกไซด์	29	835	Ethic	6
17	China	38	306	Coli	8	563	ตะกอน	42	846	Fund	9
24	Economic	318	307	Concrete	12	564	ตะกั่ว	11	847	Gamma	8
30	Entomopathogenic	12	308	Consume	36	568	ทรัพย์สิน	17	848	Grade	5
32	Family	9	309	Cost	45	571	ท้องถิ่น	21	852	Image	10
34	Finance	82	315	Diospyros	6	574	เทอร์โม	4	854	Income	11
40	Generate	136	316	dioxide	10	576	ธรรม	4	856	Japan	4
41	Genetic	125	321	Dry	9	588	ไนโตรเจน	25	860	Metropolitan	12
45	Harvest	12	326	Ethanol	7	589	บรรยากาศ	16	875	Palm	11
49	Horticulture	17	328	Eucalyptus	23	590	บัว	8	876	paniculata	7
52	Industrial	130	334	Fire	42	602	ปลา	165	878	Parent	8

ตารางผนวกที่ 2 (ต่อ)

รหัส	คำสำคัญ	TF	รหัส	คำสำคัญ	TF	รหัส	คำสำคัญ	TF	รหัส	คำสำคัญ	TF
67	Medicine	18	342	Gas	20	605	ป่วย	13	879	Pasak	5
68	Metal	23	344	gloeosporioide	10	609	ปูน	9	880	Past	8
70	Molecular	22	345	Gold	7	612	โปรตีน	64	889	Radiation	10
72	Network	38	352	Investment	26	616	ผัก	75	895	Reservoir	9
75	Pathologic	21	362	Lotus	5	619	ไผ่	18	896	Retent	9
95	Social	89	364	Macrobrachium	10	628	พลาสติก	32	897	Rock	6
98	Sport	18	365	Maize	12	634	เพจ	12	900	Safe	30
104	Telecommun	4	366	Mango	17	640	ฟ่าง	8	902	Save	20
105	Textile	5	369	Methane	7	659	มือ	12	904	Secretariat	5
106	Transport	6	371	milk	18	664	แม่กลอง	9	905	Secure	10
107	Tropic	13	372	Mill	17	669	ไมโคร	16	906	Shelf	5
112	Wood	17	374	Mine	10	673	ยิปซั่ม	7	908	Sludge	9
114	ก่อสร้าง	11	376	monodon	44	674	ยีน	62	909	Smoke	5
124	คลัง	28	381	Mutation	10	676	ยุติธรรม	5	910	Sodium	10

ตารางผนวกที่ 2 (ต่อ)

รหัส	คำสำคัญ	TF	รหัส	คำสำคัญ	TF	รหัส	คำสำคัญ	TF	รหัส	คำสำคัญ	TF
126	คหกรรม	62	382	Neural	15	681	ร้อน	87	913	Spot	6
133	เงิน	111	384	Noodle	13	682	ระบาด	7	916	Student	77
139	ช่างยนต์	6	393	People	32	683	ระบาย	9	917	Sunflower	9
142	เซรามิก	10	401	Price	18	692	ไร้สาย	15	918	Sweet	12
151	ท่องเที่ยว	93	406	RAPD	7	693	ฤดู	5	920	Think	10
153	ทั่วไป	153	409	resin	4	694	ลม	6	921	Tilapia	7
154	ทุน	90	410	Rice	109	697	ลำไส้	9	922	Tourist	35
157	ธนาคาร	31	412	root	11	702	แล้ง	11	923	Traffic	10
160	ธุรกิจ	143	415	Sanctuary	7	704	โลชั่น	7	924	Transfer	24
164	นา	24	420	Shrimp	35	708	วิตามิน	16	925	Treatment	40
173	บรรณารักษ์	9	438	Tree	16	712	ศาล	9	926	Undergraduate	8
178	ปกครอง	6	442	Uvaria	5	713	สตรอเบอรี่	28	927	Universe	28
186	ประวัติศาสตร์	24	448	viridis	5	714	สตาร์ช	31	929	White	13
188	ป่า	306	449	Virtual	6	718	สมุนไพร	63	935	ช่วยสอน	34

ตารางผนวกที่ 2 (ต่อ)

รหัส	คำสำคัญ	TF	รหัส	คำสำคัญ	TF	รหัส	คำสำคัญ	TF	รหัส	คำสำคัญ	TF
192	พณิชยการ	9	456	Zeolite	21	728	สืบพันธุ์	20	936	ชั้น	5
195	พฤกษ	28	462	กล้วย	34	735	ไส้เดือน	10	939	ตราสารหนี้	4
204	เพาะ	4	465	กล้ามเนื้อ	10	739	หน่อไม้	24	940	เถ้า	5
208	ไฟฟ้า	237	467	กลิ่น	10	757	ไหม	35	942	เทียม	5
209	ภาพถ่าย	8	472	ก้ามกราม	31	758	ไหม้	10	944	นำเข้า	10
215	มวลชน	9	479	แกลบ	15	767	ออมทรัพย์	8	945	นิเวศ	15
219	โมเลกุล	31	486	ขมิ้น	9	770	อะไมเลส	6	946	บริโภค	39
229	โรงแรม	12	488	ขยะ	37	773	อ่างเก็บน้ำ	21	947	ประชาธิปไตย	7
231	เลขานุการ	17	493	เข้มข้น	5	777	อายุ	7	949	ภูมิ	13
232	โลจิสติกส์	5	495	โขง	13	784	อุปทาน	13	951	มูลค่า	12
245	เศรษฐศาสตร์	281	496	ไข่	87	786	เอชไอวี	8	954	รส	6
249	สรีรวิทยา	21	500	ครอบครัว	8	803	Agrotour	6	956	เรียน	38
252	สอน	181	505	คลื่น	12	811	Apple	9	957	โรง	8
256	สาธารณสุข	12	507	คอนกรีต	11	813	Baby	13	960	วรรณกรรม	7

ตารางผนวกที่ 2 (ต่อ)

รหัส	คำสำคัญ	TF	รหัส	คำสำคัญ	TF	รหัส	คำสำคัญ	TF	รหัส	คำสำคัญ	TF
260	สิ่งแวดล้อม	116	512	คาร์บอน	41	814	Bacillus	13	961	ส่งออก	18
263	โสตทัศนศึกษา	11	516	เค็ม	48	815	Banana	6	962	เสพติด	8
277	อิเล็กทรอนิกส์	39	519	เครื่องดื่ม	10	818	Calcium	5	964	หลักทรัพย์	12
278	อุดมศึกษา	17	520	เครื่องมือ	7	821	Cattle	11	965	หวาน	18
281	Acid	30	523	โครงการหลวง	6	822	Cement	4	966	ออกกำลังกาย	16
283	Asia	11	524	โครงข่าย	10	823	Chromatography	7	968	อัคคีภัย	11
285	Avian	7	525	ไคโตซาน	13	827	Construct	18	969	อารมณ์	15
287	Basin	18	532	เจ้าพระยา	16	828	Corn	31			
292	Build	11	537	ชื้น	19	830	Crocodile	6			

หมายเหตุ TF หมายถึง ความถี่ของคำสำคัญ (MAX TF = 318 และ MIN TF = 4)

ในการคัดเลือกคุณลักษณะจากตารางผนวกที่ 2 จะสังเกตว่าคุณลักษณะที่ไม่ถูกเลือกเพราะคุณลักษณะดังกล่าวอาจปรากฏเพียงไม่กี่แถวของข้อมูลทั้งชุด อาจจะหนึ่งหรือสองแถว จึงมีความเป็นไปได้ที่คุณลักษณะที่ปรากฏน้อยแถวจะไม่มีความสัมพันธ์กับคุณลักษณะอื่น ทำให้โอกาสสูงที่จะถูกตัดออกจากข้อมูลหลังผ่านกระบวนการคัดเลือกคุณลักษณะแล้ว

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	นางสาวจารุวรรณ กาญจนศุภวรรณ
วัน เดือน ปี ที่เกิด	23 สิงหาคม 2528
สถานที่เกิด	โรงพยาบาลศิริราช เขตบางกอกน้อย กรุงเทพมหานคร
ประวัติการศึกษา	วท.บ. (วิทยาการคอมพิวเตอร์) มหาวิทยาลัยหอการค้าไทย (พ.ศ. 2549)
ตำแหน่งหน้าที่การงานปัจจุบัน	-
สถานที่ทำงานปัจจุบัน	-
ผลงานดีเด่นและรางวัลทางวิชาการ	1. นำเสนอและตีพิมพ์บทความวิชาการเรื่อง “Enterprise Expert Mining of Public University in Thailand”, *In* MIWAI2009, Mahasarakham, Dec 10-11, 2009 2. นำเสนอและตีพิมพ์บทความวิชาการเรื่อง “การแบ่งกลุ่มผู้เชี่ยวชาญในมหาวิทยาลัยของรัฐโดยใช้เทคนิคการจัดกลุ่มแบบ 2 ขั้นตอนและการให้น้ำหนักคุณลักษณะของข้อมูล”, *In* JCSSE2010, Bangkok, May 12-14, 2010 3. นำเสนอและตีพิมพ์บทความวิชาการเรื่อง “Segmentation of University Experts by K-Means Algorithm and Feature Weighting Techniques”, *In* INCEB2010, Bangkok, Nov 18-19, 2010 4. นำเสนอและตีพิมพ์บทความวิชาการเรื่อง “Clustering of Experts in State University by Using Fuzzy C-Means Algorithm and Feature Weighting Techniques”, *In* ICIMT2010, Hong Kong, Dec 28-30,2010
ทุนการศึกษาที่ได้รับ	-