การลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ

Feature Reduction using FCA for Web Page Classification

วิรัตน์ ชูนุ้ย

Wirat Choonui

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์
มหาวิทยาลัยสงขลานครินทร์

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computer Science
Prince of Songkla University
2555

ชื่อวิทยานิพนธ์ การลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ
ผู้เขียน นายวิรัตน์ ชูนุ้ย
สาขาวิชา วิทยาการคอมพิวเตอร์

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

.....
(ผู้ช่วยศาสตราจารย์ ดร.วิภาดา เวทย์ประสิทธิ์)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

.....
(ผู้ช่วยศาสตราจารย์ ดร.ลัดดา ปรีชาวีรกุล)

คณะกรรมการสอบ

.....ประธานกรรมการ
(ดร.นพมาศ ปักเข็ม)

.....กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ลัดดา ปรีชาวีรกุล)

.....กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ศิริรัตน์ วณิชโยบล)

.....กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.วิภาดา เวทย์ประสิทธิ์)

บัณฑิตวิทยาลัย มหาวิทยาลัยสงขลานครินทร์ อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษา ตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

.....
(ศาสตราจารย์ ดร.อมรรัตน์ พงศ์ดารา)
คณบดีบัณฑิตวิทยาลัย

ชื่อวิทยานิพนธ์ การลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ
ผู้เขียน นายวิรัตน์ ชูนุ้ย
สาขาวิชา วิทยาการคอมพิวเตอร์
ปีการศึกษา 2554

บทคัดย่อ

จำนวนเว็บเพจมีอัตราการเพิ่มขึ้นอย่างมหาศาลทำให้การสืบค้นข้อมูลให้ได้ตรงกับความต้องการของผู้ใช้มีความยุ่งยากและเสียเวลามาก การจำแนกประเภทเว็บเพจเป็นวิธีการหนึ่งที่ช่วยแก้ปัญหาดังกล่าว อย่างไรก็ตามหากเว็บเพจมีจำนวนมากจะทำให้ขนาดลักษณะเฉพาะซึ่งเป็นข้อมูลนำเข้ามีขนาดใหญ่ตามไปด้วย วิทยานิพนธ์นี้ได้นำเสนอแบบจำลองการลดขนาดลักษณะเฉพาะโดยใช้ Formal Concept Analysis สำหรับการจำแนกประเภทเว็บเพจ (Feature Reduction using FCA for Web Page Classification: FR_FCA_WPC) พัฒนาโปรแกรมด้วย Visual C#.Net โดยมีขั้นตอนการทำงานแบ่งออกเป็น 4 ขั้นตอนหลักคือ 1) การเตรียมข้อมูลเว็บเพจ 2) การเลือกลักษณะเฉพาะโดยใช้ Information Gain (IG) 3) การเลือกลักษณะเฉพาะโดยใช้ FCA และ 4) การจำแนกประเภทและการประเมินผล ทำการทดลองกับชุดข้อมูลเว็บเพจมาตรฐานจาก CMU จำนวน 2 ชุดคือ ชุดข้อมูล 7Sectors และชุดข้อมูล BankResearch ใช้ลักษณะเฉพาะจากข้อความและหัวเรื่อง จำแนกประเภทด้วย Multi-Layer Perceptron Neural Networks (MLP) และ Support Vector Machine (SVM) ทำการเปรียบเทียบประสิทธิภาพระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ที่นำเสนอกับวิธี IG พบว่าวิธี FR_FCA_WPC สามารถลดขนาดลักษณะเฉพาะได้ดีกว่าและให้ค่า F-measure ที่สูงที่สุด

Thesis Title Feature Reduction using FCA for Web Page Classification

Author Mr. Wirat Choonui

Major Program Computer Science

Academic Year 2011

ABSTRACT

The number of web pages has an increasing rate eminently. Then finding information to meet the need of users is complex and time-consuming. Web Page Classification is one of methods to solve such problems. However, if the number of web pages is large then the number of appropriate feature input data will be large also. This thesis proposed a model of Feature Reduction using Formal Concept Analysis for Web Page Classification (FR_FCA_WPC). Visual C#.Net was used for programming development with four main steps that were 1) Web Page Preprocessing, 2) Feature Selection using Information Gain (IG), 3) Feature Selection using FCA, and 4) Classification and Evaluation. Two data sets of 7Sectors data set and BankResearch data set of web page benchmark from CMU were used for this study. The study used features from text and title. The classification was done by Multi-Layer Perceptron (MLP) Neural Networks and Support Vector Machine (SVM). In comparing the efficiencies between the feature reduction proposed with IG, the study found that proposed FR_FCA_WPC method could reduce more features and gave the highest value of F-measure.

สารบัญ

สารบัญ (ต่อ)

รายการตาราง

รายการตาราง (ต่อ)

ตาราง **หน้า**

รายการภาพประกอบ

รายการภาพประกอบ (ต่อ)

ภาพประกอบ **หน้า**

รายการภาพประกอบ (ต่อ)

ภาพประกอบ **หน้า**

รายการภาพประกอบ (ต่อ)

ภาพประกอบ **หน้า**

รายการภาพประกอบ (ต่อ)

ภาพประกอบ **หน้า**

รายการภาพประกอบ (ต่อ)

ภาพประกอบ **หน้า**

บทที่ 1

บทนำ

ปัจจุบันอินเตอร์เน็ตได้รับความนิยมอย่างแพร่หลายและมีการพัฒนาที่รวดเร็ว ส่งผลให้จำนวนเว็บเพจ (Web Page) มีอัตราเพิ่มขึ้นอย่างมหาศาลทำให้ประสิทธิภาพในการค้นหาข้อมูลเข้าสู่ภาวะวิกฤต โดยเฉพาะการสืบค้นข้อมูลด้วยเสิร์ชเอ็นจิน (Search Engine) จะทำได้ยากและได้ผลลัพธ์ที่ไม่ตรงกับความต้องการของผู้ใช้ เนื่องจากในการค้นหาแต่ละครั้งจะได้ผลลัพธ์ออกมาเป็นจำนวนมากซึ่งมีทั้งข้อมูลที่ตรงกับความต้องการและไม่ตรงกับความต้องการ โดยผู้ใช้จะต้องตัดสินใจเลือกข้อมูลจากผลลัพธ์จนกว่าจะได้ข้อมูลที่ตรงกับความต้องการจริงทำให้ยุ่งยากและเสียเวลามาก วิธีการหนึ่งที่ช่วยแก้ปัญหาดังกล่าวก็คือ การจำแนกประเภทเว็บเพจ (Web Page Classification) ซึ่งเป็นการจัดเว็บเพจให้เป็นกลุ่มตามความสนใจ เพื่อช่วยสนับสนุนให้ผู้ใช้มีความสะดวกรวดเร็วในการค้นหาข้อมูลมากยิ่งขึ้น

ในการจำแนกประเภทเว็บเพจด้วยวิธีการต่าง ๆ จะมีการแทนเอกสาร (Document Representation) โดยใช้คำ (Words) ที่อยู่ภายในเอกสารเป็นลักษณะเฉพาะ (Features) อย่างไรก็ตามในแต่ละเว็บเพจนั้นประกอบด้วยคำเป็นจำนวนมากจึงทำให้ขนาดของลักษณะเฉพาะซึ่งเป็นข้อมูลเข้า (Input Features) ของการจำแนกประเภทเว็บเพจมีขนาดใหญ่ตามไปด้วย ซึ่งส่งผลให้การจำแนกประเภทเว็บเพจมีความซับซ้อนมากยิ่งขึ้น ดังนั้นการนำเทคนิคการลดขนาดลักษณะเฉพาะ (Feature Reduction) มาใช้ก็เป็นแนวทางหนึ่งที่สามารถช่วยแก้ปัญหาดังกล่าวนี้ได้

Formal Concept Analysis (FCA) เป็นทฤษฎีการวิเคราะห์ข้อมูลโดยกำหนดเป็นโครงสร้างคอนเซ็ปต์ (Conceptual Structure) และเป็นเทคนิคในการจัดกลุ่มคอนเซ็ปต์ (Concept) ที่พัฒนาด้วยพื้นฐานทางคณิตศาสตร์ ซึ่งช่วยให้สามารถค้นหาความสัมพันธ์ของข้อมูลและวิเคราะห์ข้อมูลที่มีโครงสร้างที่ซับซ้อนได้ง่ายขึ้น

งานวิจัยนี้ได้นำวิธีการ FCA มาใช้หาความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารเว็บเพจเพื่อลดขนาดลักษณะเฉพาะสำหรับการจำแนกประเภทเว็บเพจ เพื่อให้ได้ผลการจำแนกประเภทที่ถูกต้องมากยิ่งขึ้น

1.1 การตรวจเอกสาร

เทคนิคที่ใช้ในการสร้างแบบจำลองในการลดขนาดลักษณะเฉพาะสำหรับการจำแนกประเภทเว็บเพจ ได้แก่ เทคนิคการจำแนกประเภทเว็บเพจ การลดมิติข้อมูล (Dimensionality Reduction) และ Formal Concept Analysis ดังรายละเอียดต่อไปนี้

1.1.1 การจำแนกประเภทเว็บเพจ (Web Page Classification)

การจำแนกประเภทเว็บเพจ (Web Page Classification) เป็นกระบวนการในการกำหนดประเภทให้กับเว็บเพจ โดยจะมีการเรียนรู้แบบมีผู้สอน (Supervised Learning) ซึ่งจะต้องมีการสอน (Training) เพื่อให้ได้โมเดลจำแนกประเภท (Classifier) สำหรับนำไปใช้จำแนกประเภทตัวอย่างที่ไม่ทราบในอนาคต (Qi and Davison, 2009) โดยทั่วไปงานวิจัยที่เกี่ยวข้องทางด้านการจำแนกประเภทเว็บเพจจะเป็นงานวิจัยที่เน้นการลดเวลาในการประมวลผลและเพิ่มประสิทธิภาพในการจำแนกประเภทด้วยเทคนิคต่าง ๆ เช่น เทคนิคการโหวต (Voting) เทคนิคการสรุปความ (Summarization) และการลดขนาดของลักษณะเฉพาะ (Feature Reduction) เป็นต้น

Chen และ Hsieh (2006) เสนอวิธีการจำแนกเว็บเพจด้วย SVM โดยมีการเลือกลักษณะเฉพาะด้วยวิธี Latent Semantic Analysis (LSA) และวิธี Web Page Feature Selection (WPFS) ที่มีชื่อว่า WVSVM (Weighted Voting Support Vector Machine) ซึ่งในส่วนแรกจะใช้ LSA หาความสัมพันธ์ของความหมาย (Semantic) ระหว่างเอกสารที่ถูกเลือก และในส่วนที่สองจะใช้ WPFS ทำการวิเคราะห์เพื่อเลือกลักษณะเฉพาะจากเนื้อหาของเว็บเพจโดยตรง จากนั้นนำลักษณะเฉพาะที่ได้จากทั้ง 2 ส่วนไปใช้เป็นข้อมูลนำเข้าของอัลกอริทึม SVM สำหรับใช้สอนและทดสอบตามลำดับ ในขั้นตอนสุดท้ายจะใช้แบบแผนการโหวต (Voting Schema) พิจารณาจัดประเภทให้กับเว็บเพจ โดยชุดข้อมูลทดลองที่ใช้เป็นข้อมูลข่าวกีฬาจากเว็บไซต์ Udndata จำนวน 1,724 หน้า ซึ่งผลการทดลองแสดงให้เห็นว่าวิธี WVSVM ให้ค่าความถูกต้องมากที่สุด

Shen และคณะ (2007) ได้นำเสนออัลกอริทึมการสรุปความแบบอัตโนมัติ (Web Page Summarization) สำหรับการเตรียมข้อมูลในการจำแนกประเภทเว็บเพจ ซึ่งสกัดข้อมูลจากหัวข้อหลัก (Main Topic) ของเว็บเพจ โดยการวิเคราะห์เค้าโครงของเว็บเพจ (Page-layout) เพื่อเพิ่มค่าความถูกต้องในการจำแนกประเภท และประเมินประสิทธิภาพในการจำแนกประเภทโดยเปรียบเทียบกับอัลกอริทึมอื่นๆ ทั้งแบบมีผู้สอนและไม่มีผู้สอน โดยใช้ชุดข้อมูลจากเว็บ LookSmart จำนวน 153,019 หน้าแบ่งเป็น 64 ประเภท ทำการทดสอบแบบไขว้ข้าม 10

กลุ่ม พบว่าการนำวิธีการสรุปความอัตโนมัติแบบต่าง ๆ มาผสมกันสามารถเพิ่มประสิทธิภาพในการจำแนกประเภทได้ดีขึ้น

Yin และคณะ (2008) นำเสนออัลกอริทึมเพื่อลดขนาดลักษณะเฉพาะของเว็บเพจและสกัดกฎการจำแนกประเภทโดยใช้การลดขนาดลักษณะด้วยทฤษฎีรัฟเซต (Rough Set) ในการทดลองใช้ชุดข้อมูลจากเว็บ www.sohu.com จำนวน 1,000 หน้า แบ่งเป็นข้อมูลสอนระบบ 760 หน้า และข้อมูลทดสอบ 240 หน้า ผลการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอสามารถลดขนาดลักษณะเฉพาะได้ดี และได้กฎการจำแนกประเภทที่เข้าใจง่าย ให้ค่าความถูกต้องที่ดีและจำแนกได้เร็วกว่าวิธีอื่น

Thamrongrat และคณะ (2009) ได้นำเสนออัลกอริทึมการโหวตในการจำแนกประเภทเว็บเพจโดยใช้วิธีแบบมัลติคลาสซัพพอร์ตเวกเตอร์แมชชีน (Voting Algorithm of Multi-Class SVM for Web Page Classification) ที่มีชื่อว่า VAMSVM_WPC ทำการทดลองกับชุดข้อมูลเว็บเพจมาตรฐานจาก CMU โดยใช้ลักษณะเฉพาะจากข้อความ (Text) และหัวเรื่อง (Title) ผลการทดลองแสดงให้เห็นอัลกอริทึม 1vsAll_Voting ให้ค่า F-measure ที่ดีที่สุด

1.1.2 การลดมิติข้อมูล (Dimensionality Reduction)

การลดมิติข้อมูล (Dimensionality Reduction) (Radovanovic, 2006) เป็นการลดขนาดลักษณะเฉพาะและกำจัดข้อมูลที่ไม่สำคัญออก ซึ่งหากเว็บเพจมีลักษณะเฉพาะเป็นจำนวนมากจะทำให้มีอุปสรรคต่อตัวจำแนกประเภท (Classifier) ต้องสิ้นเปลืองเนื้อที่และเวลาในการประมวลผล ดังนั้นการลดลักษณะเฉพาะจึงเป็นขั้นตอนหนึ่งที่สำคัญในการเตรียมข้อมูล (Data Preprocessing) ก่อนการสร้างตัวจำแนกประเภท

การกำจัดคำหยุด (Stopping) เป็นการกำจัดคำที่ไม่มีประโยชน์สำหรับการวิเคราะห์ซึ่งจะพิจารณาจากรายการคำหยุด (Stoplist) เช่น คำว่า “I” “the” และ “with” เป็นต้น

การหารากศัพท์ (Stemming) เป็นการหารูปเดิมของคำ โดยแปลงรูปแบบของคำที่มีความหมายคล้ายกันให้เป็นคำเดียวกัน เช่น คำว่า “computer” “computing” “computational” แปลงเป็น “comput” เป็นต้น ดังนั้นวิธีการนี้ไม่เพียงแต่จะช่วยลดจำนวนของลักษณะเฉพาะลง แต่ยังสามารถหาความสัมพันธ์ของคำระหว่างกันได้อีกด้วย ซึ่งอัลกอริทึมที่นิยมใช้กันก็คือ Porter Stemmer (Porter, 1980)

ในการจำแนกประเภทมีแนวทางในการลดขนาดลักษณะเฉพาะ 2 วิธีด้วยกัน คือ การเลือกลักษณะเฉพาะ (Feature Selection) และการสกัดลักษณะเฉพาะ (Feature Extraction)

1.1.2.1 การเลือกลักษณะเฉพาะ (Feature Selection) เป็นการเลือกเซตของลักษณะเฉพาะใหม่จากเซตของลักษณะเฉพาะเดิม โดยที่เซตของ

ลักษณะเฉพาะใหม่ที่ได้จะเป็นเซตย่อย (Subset) ของเซตลักษณะเฉพาะเดิม ซึ่งวิธีการเลือกลักษณะเฉพาะมี 2 วิธีได้แก่ Wrapper และ Filter

1) Wrapper เป็นวิธีการอย่างง่ายซึ่งจะสร้างเซตของลักษณะเฉพาะใหม่ขึ้นมาโดยการเพิ่มหรือลดคุณลักษณะจากเซตของลักษณะเฉพาะเดิม จากนั้นสร้างตัวจำแนกประเภทจากเซตของลักษณะเฉพาะใหม่ที่ได้แล้วประเมินประสิทธิภาพของตัวจำแนก ซึ่งเซตใหม่ที่ให้ผลลัพธ์ที่มีประสิทธิภาพมากที่สุดจะถูกเลือกใช้

2) Filter เป็นวิธีการที่พยายามเลือกลักษณะเฉพาะที่มีค่าความสำคัญมากที่สุดและเป็นค่าที่คำนวณง่าย วิธีการวัดค่าความสำคัญมีหลายวิธีการด้วยกัน ซึ่งวิธีการที่ง่ายที่สุดได้แก่ Term Frequency (TF) เป็นวิธีการวัดโดยนับจำนวนความถี่ของลักษณะเฉพาะ (หรือ Term) ที่ปรากฏอยู่ในเอกสาร วิธีการที่นำทฤษฎี Information มาใช้เพื่อพิจารณาเลือกลักษณะเฉพาะได้แก่ วิธี Information Gain และ Gain Ratio ซึ่งจะเลือกลักษณะเฉพาะที่มีค่า Information Gain สูงสุดหรือมีค่า Entropy น้อยที่สุด และอีกวิธีคือ Chi square ซึ่งเป็นการวัดโดยใช้สถิติประมาณความสัมพันธ์ร่วมระหว่างลักษณะเฉพาะกับคลาสของลักษณะเฉพาะ นอกจากนี้ยังมีวิธีการที่แตกต่างออกไปก็คือ Relief เป็นการสุ่มตัวอย่างจากชุดข้อมูลทดลองแล้วนำมาวางในกลุ่มที่ใกล้เคียงกันมากที่สุด (พิจารณาจากระยะห่าง) ซึ่งมีอยู่ 2 กลุ่ม โดยที่กลุ่มแรกเป็นตัวอย่างบวกและอีกกลุ่มเป็นตัวอย่างลบ และใช้ค่าของลักษณะเฉพาะเพื่อปรับปรุงความสัมพันธ์ระหว่างกัน ต่อมา Relief ได้ถูกพัฒนาเป็น ReliefF (RF) ซึ่งช่วยสนับสนุนชุดข้อมูลทดลองที่เป็นแบบมัลติคลาสและข้อมูลที่มีสิ่งรบกวน (Noisy)

งานวิจัยที่เกี่ยวข้องกับการเลือกลักษณะเฉพาะ ได้แก่ Zhang และคณะ (2007) ได้ใช้วิธีการเลือกลักษณะเฉพาะแบบ Information Gain ในการเลือกลักษณะเฉพาะสำหรับการจัดกลุ่มประเภทเว็บเพจแบบอัตโนมัติโดยใช้วิธีการ Principal Component Analysis (PCA) Indra และคณะ (2008) ได้รวมเทคนิคการเลือกลักษณะเฉพาะเพื่อหาลักษณะเฉพาะที่มีคุณภาพและมีจำนวนน้อยที่สุด ทดลองโดยใช้ชุดข้อมูลมาตรฐานจาก WebKB พรพล และคณะ (2552) ได้นำเสนอการจำแนกประเภทเว็บเพจโดยวิธีการลดขนาดลักษณะเฉพาะและเปรียบประสิทธิภาพระหว่างวิธีการเลือกลักษณะเฉพาะแบบ ReliefF Information Gain และ Chi Square ทำการจำแนกประเภทโดยเปรียบเทียบวิธีมัลติคลาสซัพพอร์ตเวกเตอร์แมชชีนแบบ 1vs1 และ 1vsAll

1.1.2.2 การสกัดลักษณะเฉพาะ (Feature Extraction) เป็นการสร้างหรือสังเคราะห์เซตของลักษณะเฉพาะใหม่จากเซตของลักษณะเฉพาะเดิม โดยที่เซตของคุณลักษณะเฉพาะใหม่ที่ได้มีขนาดเล็กกว่าเซตของลักษณะเฉพาะเดิม ซึ่งวิธีการสกัดลักษณะเฉพาะมี 2 วิธีคือ Term Clustering และ Latent Semantic Indexing

1) Term Clustering เป็นวิธีการที่ใช้เทคนิคการจัดกลุ่ม (Clustering Techniques) โดยสร้างกลุ่มของคำที่มีความหมายเกี่ยวข้องกันเข้าไว้ด้วยกันแล้วใช้กลุ่มของคำที่สร้างขึ้นมาแทนคำเหล่านั้น

2) Latent Semantic Indexing (LSI) เป็นวิธีการที่ใช้ตารางเมตริกซ์ระหว่างเอกสารกับลักษณะเฉพาะ (Documents Features Matrix) เข้ามาจัดการ โดยที่ LSI จะนำเทคนิค Singular Value Decomposition (SVD) มาใช้สำหรับแปลงเมตริกซ์ให้มีขนาดของมิติเล็กลง ในขณะเดียวกันก็ยังคงเก็บความสัมพันธ์ร่วมระหว่างลักษณะเฉพาะเอาไว้

1.1.3 Formal Concept Analysis

Formal Concept Analysis (FCA) (Obitko *et al*., 2004; Wille, 2005; Priss, 2006) เป็นทฤษฎีของการวิเคราะห์ข้อมูลโดยกำหนดเป็นโครงสร้างแนวคิด (Conceptual Structure) จากชุดข้อมูล (Data set) และแสดงโครงสร้างในรูปภาพของคอนเซ็ปต์แลททิส (Concept Lattice) ซึ่งช่วยให้การวิเคราะห์โครงสร้างที่ซับซ้อนและการค้นหาความสัมพันธ์กันของข้อมูลทำได้ง่ายขึ้น FCA เป็นเทคนิคการจัดกลุ่มคอนเซ็ปต์ (Conceptual Clustering) ที่พัฒนาด้วยพื้นฐานทางคณิตศาสตร์และประสบความสำเร็จอย่างกว้างขวางในการประยุกต์ใช้งานทางด้านการแพทย์ จิตวิทยา ห้องสมุด วิศวกรรมซอฟต์แวร์ ระบบนิเวศน์ และด้านอื่น ๆ ที่เกี่ยวข้องกับการวิเคราะห์ข้อมูล รวมไปถึงด้านการค้นคืนสารสนเทศ (Information Retrieval) และด้านการค้นหาความรู้ในฐานข้อมูล (Knowledge Discovery in Database: KDD) (Lakhal and Stumme, 2005)

การสร้างแนวคิดตามทฤษฎี FCA จะขึ้นอยู่กับความเข้าใจทางปรัชญา โดยแนวคิดหรือคอนเซ็ปต์ (Concept) ที่ตั้งขึ้นประกอบด้วย 2 ส่วน (Obitko *et al*., 2004) คือ “Extension” ประกอบด้วยออบเจ็กต์ (Objects) ทั้งหมดของคอนเซ็ปต์ที่มีคุณลักษณะ (Attributes) ร่วมกัน และ “Intension” ประกอบด้วยคุณลักษณะทั้งหมดของออบเจ็กต์ที่ใช้ร่วมกัน โดยคอนเซ็ปต์ต่าง ๆ ได้มาจากคอนเท็กซ์ที่กำหนดมาให้และสามารถนำเสนอให้อยู่ในรูปแบบของลำดับชั้น (Hierarchy) ได้

งานวิจัยที่เกี่ยวข้องกับ FCA ส่วนใหญ่จะเป็นงานวิจัยทางด้านออนโทโลจี (Ontology) ซึ่งเป็นการสร้างกลุ่มของคำที่มีโครงสร้างแบบลำดับชั้นและมีความสัมพันธ์ในเชิงความหมายสำหรับอธิบายขอบเขตเนื้อหาที่สนใจ โดยนำทฤษฎี FCA มาเป็นพื้นฐานในการสร้างอัลกอริทึมเพื่อออกแบบออนโทโลจี Haav (2004) ได้นำเสนอวิธีการแบบกึ่งอัตโนมัติ (Semi-Automatic) สำหรับออกแบบและสกัดออนโทโลจีจากคอนเซ็ปต์แลททิสที่สร้างด้วย FCA Hwang และคณะ (2005) ใช้ FCA เป็นพื้นฐานในการพัฒนาออนโทโลจี เพื่อแสดงให้เห็นว่า

FCA สามารถช่วยแสดงออนโทโลจีในแลททิสซึ่งเข้าใจได้ง่าย และเป็นตัวชี้นำในการสร้างออนโทโลจี Wang และ He (2006) นำเสนอวิธีการสำหรับสร้างออนโทโลจีบนพื้นฐานของ FCA โดยแสดงตัวอย่างโดเมนของสัตว์ และสามารถนำออนโทโลจีที่ได้ไปแทนใน SWRL (Semantic Web Rule Language) ได้ นอกจากนี้ Cure และ Jeansoulin (2008) ได้นำเสนอวิธีการรวมออนโทโลจีบนพื้นฐานของ FCA โดยรวมเอาตัวอย่างที่มีความสัมพันธ์กันจากสองออนโทโลจีแล้วสร้างเป็นคอนเซ็ปต์ของออนโทโลจีใหม่และสามารถลดความซ้ำซ้อนของคอนเซ็ปต์ลงได้

1.2 วัตถุประสงค์ของโครงการ

เพื่อวิเคราะห์ ออกแบบ และสร้างแบบจำลองการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ

1.3 ขอบเขตการดำเนินการวิจัย

1) วิเคราะห์ ออกแบบ และสร้างแบบจำลองการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ

2) พัฒนาโปรแกรมการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ เพื่อทดสอบกับชุดข้อมูลเว็บเพจมาตรฐานจาก CMU 2 ชุดข้อมูล ได้แก่

- ชุดข้อมูล 7Sectors (WebKB, 2012) ประกอบด้วย 7 กลุ่ม ได้แก่ 1) Energy 2) Financial 3) Healthcare 4) Materials 5) Technology 6) Transportation และ 7) Utilities
- ข้อมูล BankResearch (StatLib, 2012) แบ่งออกเป็น 11 คลาส ได้แก่ 1) Commercial Banks 2) Building Societies 3) Insurance Agencies 4) Java 5) C 6) Visual Basic 7) Astronomy 8) Biology 9) Soccer 10) Motor Racing และ 11) Sport

3) ประเมินผลของการจำแนกประเภทด้วยค่า F-measure

1.4 ขั้นตอนการดำเนินการวิจัยและระยะเวลาการดำเนินการวิจัย

1) ศึกษางานวิจัยและเอกสารที่เกี่ยวข้องสำหรับการเลือกลักษณะเฉพาะและการจำแนกประเภทเว็บเพจ

2) ศึกษาเทคโนโลยีและเครื่องมือสนับสนุนสำหรับงานวิจัย

3) วิเคราะห์ ออกแบบ และสร้างแบบจำลองการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ

4) เตรียมชุดข้อมูลเว็บเพจสำหรับนำมาทดสอบวัดประสิทธิภาพของโปรแกรมในการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ

5) พัฒนาโปรแกรมสำหรับการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจตามที่ได้ออกแบบไว้

6) ทำการทดลองด้วยโปรแกรมที่ได้พัฒนาพร้อมทั้งวิเคราะห์และประเมินผล

7) ทดสอบและติดตั้งโปรแกรมการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ

8) จัดทำเอกสารประกอบโปรแกรมการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ

ระยะเวลาการดำเนินการวิจัยสามารถแสดงได้ดังตารางที่ 1.1

ตารางที่ 1.1 ระยะเวลาการดำเนินการวิจัย

กิจกรรม/ขั้นตอนการดำเนินงาน	เดือน																											
	2552								2553								2554									2555			
	5	6	7	8	9	10	11	12	1	2	3	4	5	...	11	12	1	2	3	4	5	...	10	11	12	1	2	3	4
1. ศึกษางานวิจัยและเอกสารที่เกี่ยวข้อง																													
2. ศึกษาวิธีการเตรียมข้อมูลจากเว็บเพจ																													
3. ศึกษาเทคโนโลยีและเครื่องมือสนับสนุน																													
4. วิเคราะห์และออกแบบระบบ																													
5. พัฒนาและทดสอบระบบ																													
6. เขียนบทความวิจัย																													
7. จัดทำเอกสารวิทยานิพนธ์																													

1.5 สถานที่และเครื่องมือที่ใช้ทำวิจัย

1.5.1 สถานที่ทำวิจัย

ห้องปฏิบัติการคอมพิวเตอร์ CS207 ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่

1.5.2 เครื่องมือที่ใช้ทำวิจัย

1) ด้านฮาร์ดแวร์
   - เครื่องคอมพิวเตอร์จำนวน 1 เครื่อง
   - หน่วยความจำ 2 GB
   - ฮาร์ดดิสก์ 320 GB
   - เครื่องพิมพ์จำนวน 1 เครื่อง
2) ด้านซอฟต์แวร์
   - โปรแกรมประยุกต์ WEKA
   - โปรแกรมประยุกต์ C#.Net
   - โปรแกรม ConExp (Concept Explorer)
   - ระบบปฏิบัติการ Microsoft Windows XP

1.6 ประโยชน์ที่คาดว่าจะได้รับ

ได้แบบจำลองการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ

บทที่ 2

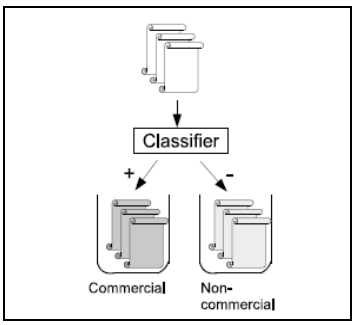
ทฤษฎีที่เกี่ยวข้อง

ทฤษฎีที่เกี่ยวข้องในงานวิจัยนี้ประกอบด้วย การจำแนกประเภทเว็บเพจ (Web Page Classification) การแทนเอกสารแบบ Bag-of-Words การให้ค่าน้ำหนักคำด้วย TF-IDF การกำจัดคำหยุด (Stopping) การหารากศัพท์ของคำ (Stemming) การลดขนาดลักษณะเฉพาะ (Feature Reduction) ทฤษฎี Formal Concept Analysis (FCA) โครงข่ายประสาทเทียม (Artificial Neural Networks) Support Vector Machine (SVM) และการประเมินผลของตัวจำแนกประเภท (Classifier Evaluation)

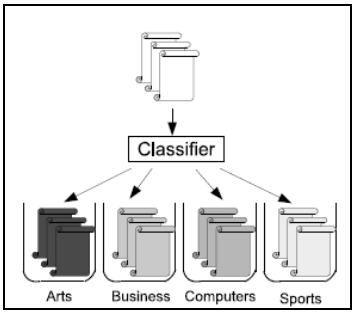
2.1 การจำแนกประเภทเว็บเพจ (Web Page Classification)

การจำแนกประเภท (Classification) เป็นวิธีการที่มีบทบาทสำคัญมากในงานการจัดการสารสนเทศและการค้นหาเอกสาร สำหรับการจำแนกประเภทจากเนื้อหาของเว็บเพจนั้นเป็นสิ่งสำคัญอย่างหนึ่งที่ช่วยสนับสนุนเรื่องต่าง ๆ บนเว็บไม่ว่าจะเป็นการพัฒนาเว็บไดเร็กทอรี่ (Web Directory) การวิเคราะห์เว็บลิงค์ (Web Link) การโฆษณาบนเว็บ และการวิเคราะห์โครงสร้างของเว็บ เป็นต้น ซึ่งการจำแนกประเภทเว็บเพจสามารถช่วยเพิ่มประสิทธิภาพในการค้นหาเอกสารเว็บได้ดียิ่งขึ้น

การจำแนกประเภทเว็บเพจหรือที่รู้จักกันคือ การจัดหมวดหมู่ของเว็บเพจ (Web Page Categorization) เป็นกระบวนการในการกำหนดประเภทให้กับเว็บเพจ โดยทั่วไปการจำแนกประเภทจะจัดเป็นปัญหาการเรียนรู้แบบมีผู้สอน (Supervised Learning) ซึ่งจะต้องมีการสอน (Training) เพื่อให้ได้โมเดลจำแนก (Classifier) สำหรับนำไปใช้จำแนกตัวอย่างที่ไม่ทราบในอนาคต โดยพื้นฐานของการจำแนกประเภทสามารถแบ่งได้เป็น 2 แบบคือ การจำแนกประเภทแบบไบนารี (Binary Classification) และการจำแนกประเภทแบบมัลติคลาส (Multi-Class Classification) (Qi and Davison, 2009) ซึ่งการจำแนกประเภทแบบไบนารีจะแบ่งตัวอย่างออกเป็น 2 คลาส (Class) คือคลาสที่เป็นตัวอย่างบวกและคลาสที่เป็นตัวอย่างลบ ดังภาพประกอบ 2.1 ส่วนการจำแนกประเภทแบบมัลติคลาสจะแบ่งตัวอย่างได้มากกว่า 2 คลาส ดังภาพประกอบ 2.2

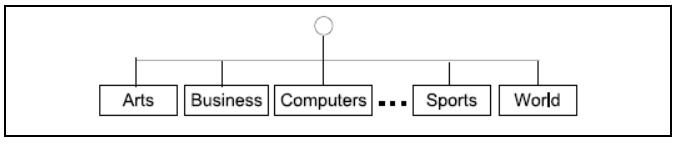


ภาพประกอบ 2.1 การจำแนกประเภทแบบไบนารี

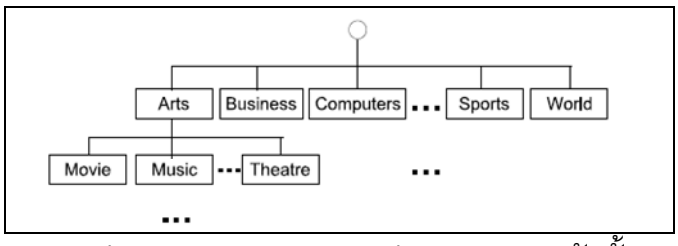


ภาพประกอบ 2.2 การจำแนกประเภทแบบมัลติคลาส

สำหรับการจำแนกประเภทเว็บเพจสามารถแบ่งออกได้เป็น 2 แบบคือ การจำแนกประเภทแบบแนวราบ (Flat Classification) และการจำแนกประเภทแบบลำดับชั้น (Hierarchical Classification) โดยที่การจำแนกประเภทแบบแนวราบจะพิจารณาแบ่งตามแนวขนาน (Parallel) กล่าวคือหนึ่งคลาสไม่สามารถแบ่งเป็นคลาสย่อยได้ แสดงดังภาพประกอบ 2.3 ส่วนการจำแนกแบบลำดับชั้นจะพิจารณาการแบ่งกลุ่มของคลาสในลักษณะลำดับชั้นที่คล้ายกับโครงสร้างต้นไม้ ซึ่งในแต่ละคลาสสามารถแบ่งเป็นคลาสย่อยได้อีก ดังภาพประกอบ 2.4



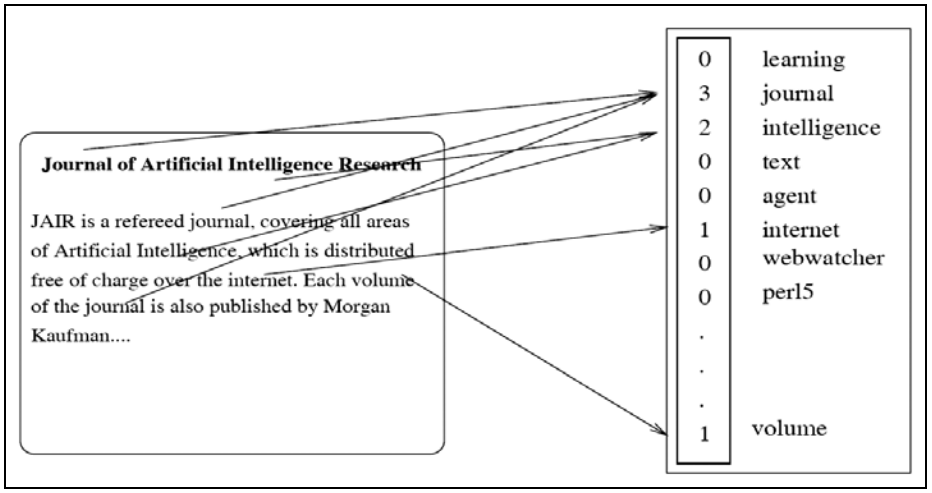
ภาพประกอบ 2.3 การจำแนกประเภทแบบแนวราบ



ภาพประกอบ 2.4 การจำแนกประเภทแบบลำดับชั้น

2.2 การแทนเอกสารแบบ Bag-of-Words

เนื่องจากเอกสาร HTML มีลักษณะเป็นกึ่งโครงสร้าง (Semi-Structured) จึงจำเป็นต้องแปลงเอกสารให้อยู่ในรูปที่สามารถนำไปใช้ในการเรียนรู้ด้วยเครื่องคอมพิวเตอร์ได้ ซึ่งวิธีการหนึ่งที่ทำได้ก็คือการแทนเอกสาร (Document Representation) ในงานวิจัยนี้ได้ใช้วิธีการแทนเอกสารแบบ Bag-of-Words (BOW) ซึ่งเป็นการนำคำ (Word) ที่อยู่ภายในเอกสาร (Document) มาใช้เป็นลักษณะเฉพาะ เมื่อกำหนดให้ T เป็นเซตของคำทั้งหมดที่ปรากฏอย่างน้อยหนึ่งครั้งในเอกสาร D ดังนั้นในการแทน BOW ของเอกสาร d_j ก็คือเวกเตอร์ของน้ำหนัก $\vec{w}_j = (w_{j1}, w_{j2}, ..., w_{j|T|})$ ซึ่งการแทน BOW มีได้หลายแบบขึ้นอยู่กับค่าของน้ำหนัก ตัวอย่างการแทนค่าแบบไบนารี (Binary) จะมีค่าน้ำหนัก $w_{jk} \in \{0,1\}$ นั่นคือ ถ้า $w_{jk} = 1$ หมายความว่า คำที่ k ปรากฏอยู่ในเอกสาร d_j และถ้า $w_{jk} = 0$ หมายความว่า คำที่ k ไม่ปรากฏอยู่ในเอกสาร d_j และตัวอย่างการแทนค่าแบบ Term Frequency (TF) จะมีค่าน้ำหนัก $w_{jk} = TF_{jk}$ นั่นคือความถี่ของคำที่ k ปรากฏอยู่ในเอกสารที่ j ภาพประกอบ 2.5 แสดงตัวอย่างการแทน BOW ด้วย TF (Radovanovic, 2006)



ภาพประกอบ 2.5 ตัวอย่างการแทนเอกสารแบบ Bag-of-Words ด้วย TF

2.3 การให้ค่าน้ำหนักคำด้วย TF-IDF

การให้ค่าน้ำหนักคำ (Word Weighting) เป็นการแทนคำในเอกสารด้วยค่าน้ำหนัก (Weight Value) ในงานวิจัยนี้ได้ใช้การให้ค่าน้ำหนักคำด้วย TF-IDF (Term Frequency–Inverse Document Frequency) (Joachims, 1997) ซึ่งเป็นวิธีที่คำนวณน้ำหนัก

จากความถี่ของคำ t_k ที่ปรากฏในเอกสาร d_j และพิจารณาความถี่ของคำ t_k ที่ปรากฏในเอกสารอื่นร่วมด้วย โดยให้ค่าน้ำหนักคำที่ปรากฏในเอกสารหลายๆ ฉบับมีค่าน้ำหนักต่ำ เนื่องจากเป็นคำที่ไม่แสดงถึงลักษณะเฉพาะของเอกสารนั้นๆ ซึ่งสามารถคำนวณได้จากสมการที่ (2.1)

$$IDF_k = \log\left(\frac{n}{DF_k}\right) \tag{2.1}$$

โดยที่ IDF_k คือ ส่วนกลับของความถี่เอกสารที่มีคำที่ k ปรากฏอยู่
DF_k คือ ความถี่เอกสารที่มีคำที่ k ปรากฏอยู่
n คือ จำนวนเอกสารทั้งหมด

การแทนค่าน้ำหนักคำด้วย TF-IDF เป็นการหาความสมดุลระหว่างความถี่ของคำกับความถี่ของเอกสารร่วมกัน ซึ่งสามารถคำนวณหาค่าน้ำหนักได้จากสมการที่ (2.2)

$$w_{jk} = TF_{jk} \times IDF_k \tag{2.2}$$

โดยที่ w_{jk} คือ ค่าน้ำหนักของคำที่ k ที่ปรากฏในเอกสารที่ j
TF_{jk} คือ ความถี่ของคำที่ k ที่ปรากฏในเอกสารที่ j

2.4 การกำจัดคำหยุด (Stopping)

การกำจัดคำหยุด (Stopping) เป็นการกำจัดคำที่ไม่มีประโยชน์สำหรับการวิเคราะห์ออก เนื่องจากเอกสารเว็บเพจจะประกอบด้วยคำเป็นจำนวนมากและมีบางคำที่มีความเกี่ยวข้องกับเอกสารน้อย ซึ่งคำเหล่านี้จัดอยู่ในกลุ่มของรายการคำหยุด (Stoplist) (Frakes and Yates, 1992) แสดงตัวอย่างดังภาพประกอบ 2.6 ดังนั้นในงานวิจัยนี้ได้นำเทคนิคการกำจัดคำหยุดมาใช้โดยจะพิจารณาตัดคำจากรายการคำหยุด เมื่อกำจัดคำเหล่านี้ออกไปจะทำให้จำนวนคำที่เป็นลักษณะเฉพาะมีขนาดลดลงซึ่งช่วยให้การจำแนกประเภทมีประสิทธิภาพมากยิ่งขึ้น

a	about	after	again	ago	all
almost	also	always	am	an	and
another	any	anybody	anyhow	anyone	anything
anyway	are	as	at	away	back
be	became	because	been	before	being
between	but	by	came	can	cannot
come	could	did	do	does	doing
done	down	each	else	even	ever
every	everyone	everything	for	from	front
get	getting	go	goes	going	gone
got	gotten	had	has	have	having
he	her	here	him	his	how
i	if	in	into	is	isn't
it	just	last	least	left	less
let	like	make	many	may	maybe
me	mine	more	most	much	my
myself	never	no	none	not	now
of	off	on	one	onto	or
our	ourselves	out	over	per	put
putting	same	saw	see	seen	shall
she	should	so	some	somebody	someone
something	stand	such	sure	take	than
that	the	their	them	then	there
these	they	this	those	through	till
to	too	two	unless	until	up
upon	us	very	was	we	went
were	what	whatever	what's	when	where
whether	which	while	who	whoever	whom
whose	why	will	with	within	without
won't	would	wouldn't	x	y	yet
you	you'd	you'll	your	you're	yours
yourself	yourselves	you've	z		

ภาพประกอบ 2.6 ตัวอย่างรายการคำหยุด (Stoplist)

2.5 การหารากศัพท์ของคำ (Stemming)

การหารากศัพท์ของคำ (Stemming) เป็นการหารูปเดิมของคำซึ่งเป็นคำที่ยังไม่ได้เติมคำต่อท้าย (Suffixes) โดยแปลงรูปแบบของคำที่มีความหมายคล้ายกันรวมให้เป็นคำเดียวกัน ยกตัวอย่างเช่นคำว่า “connected” “connecting” “connection” “connections” จะมีความหมายคล้ายกันและมีคำต่อท้ายด้วย –ed –ing –ion –ions ตามลำดับ ดังนั้นสามารถแปลงรูปแบบของคำได้โดยตัดคำต่อท้ายเหล่านี้ออกก็จะได้รูปเดิมของคำคือ “connect” เป็นต้น ซึ่งวิธีการนี้สามารถช่วยลดจำนวนของลักษณะเฉพาะลงได้ ดังนั้นงานวิจัยนี้ได้ใช้อัลกอริทึมที่นิยมใช้ในการหารากศัพท์ของคำภาษาอังกฤษก็คือ อัลกอริทึม Porter (Porter, 1980) ซึ่งมีกฎ (Rules) สำหรับการตัดคำต่อท้ายโดยกำหนดเป็นรูปแบบดัง (2.3) และมีขั้นตอนแสดงดังภาพประกอบ 2.7

$$(\text{Condition})\ S1 \rightarrow S2 \qquad (2.3)$$

โดยที่ S1 คือคำต่อท้าย
S2 คือคำที่ใช้แทนที่
Condition คือเงื่อนไขในการแทนที่

Step 1a:

รูปแบบกฎ	ตัวอย่างผลลัพธ์
SSES → SS	caresses → caress
IES → I	ponies → poni
	ties → ti
SS → SS	carress → carress
S →	cats → cat

Step 1b:

รูปแบบกฎ	ตัวอย่างผลลัพธ์
(m > 0) EED → EE	feed → feed
	agreed → agree
(*v*) ED →	plastered → plaster
	bled → bled
(*v*) ING →	motoring → motor
	sing → sing

ภาพประกอบ 2.7 ขั้นตอนการทำงานของอัลกอริทึม Porter

Step 1b.1:

รูปแบบกฎ			ตัวอย่างผลลัพธ์		
AT	→	ATE	feed	→	feed
			agreed	→	agree
BL	→	BLE	plastered	→	plaster
			bled	→	bled
IZ	→	IZE	motoring	→	motor
			sing	→	sing
(*d and not (*L or *S or *Z))	→	single letter	hopp(ing)	→	hop
			tann(ed)	→	tan
			fall(ing)	→	fall
			hiss(ing)	→	hiss
			fizz(ed)	→	fizz
(m = 1 and *o)	→	E	fail(ing)	→	fail
			fil(ing)	→	file

Step 1c:

รูปแบบกฎ			ตัวอย่างผลลัพธ์		
(*v*) Y	→	I	feed	→	feed
			agreed	→	agree

Step 2:

รูปแบบกฎ			ตัวอย่างผลลัพธ์		
(m > 0) ATIONAL	→	ATE	relational	→	relate
(m > 0) TIONAL	→	TION	conditional	→	condition
			rational	→	rational
(m > 0) ENCI	→	ENCE	valenci	→	valence
(m > 0) ANCI	→	ANCE	hesitanci	→	hesitance
(m > 0) IZER	→	IZE	digitizer	→	digitize
(m > 0) ABLI	→	ABLE	conformabli	→	conformable
(m > 0) ALLI	→	AL	radicalli	→	radical
(m > 0) ENTLI	→	ENT	differentli	→	different
(m > 0) ELI	→	E	vileli	→	vile
(m > 0) OUSLI	→	OUS	analogousli	→	analogous
(m > 0) IZATION	→	IZE	vietnamization	→	vietnamize
(m > 0) ATION	→	ATE	predication	→	predicate
(m > 0) ATOR	→	ATE	operator	→	operate
(m > 0) ALISM	→	AL	feudalism	→	feudal
(m > 0) IVENESS	→	IVE	decisiveness	→	decisive
(m > 0) FULNESS	→	FUL	hopefulness	→	hopeful
(m > 0) OUSNESS	→	OUS	callousness	→	callous
(m > 0) ALITI	→	AL	formaliti	→	formal
(m > 0) IVITI	→	IVE	sensitiviti	→	sensitive
(m > 0) BILITI	→	BLE	sensibiliti	→	sensible

ภาพประกอบ 2.7 ขั้นตอนการทำงานของอัลกอริทึม Porter (ต่อ)

Step 3:

รูปแบบกฎ			ตัวอย่างผลลัพธ์		
(m > 0) ICATE	→	IC	triplicate	→	triplic
(m > 0) ATIVE	→		formative	→	form
(m > 0) ALIZE	→	AL	formalize	→	formal
(m > 0) ICITI	→	IC	electriciti	→	electric
(m > 0) ICAL	→	IC	electrical	→	electric
(m > 0) FUL	→		hopeful	→	hope
(m > 0) NESS	→		goodness	→	good

Step 4:

รูปแบบกฎ		ตัวอย่างผลลัพธ์		
(m > 1) AL	→	revival	→	reviv
(m > 1) ANCE	→	allowance	→	allow
(m > 1) ENCE	→	inference	→	infer
(m > 1) ER	→	airliner	→	airlin
(m > 1) IC	→	gyroscopic	→	gyroscop
(m > 1) ABLE	→	adjustable	→	adjust
(m > 1) IBLE	→	defensible	→	defens
(m > 1) ANT	→	irritant	→	irrit
(m > 1) EMENT	→	replacement	→	replac
(m > 1) MENT	→	adjustment	→	adjust
(m > 1) ENT	→	dependent	→	depend
(m > 1) and (*S or *T) ION	→	adoption	→	adopt
(m > 1) OU	→	homologou	→	homolog
(m > 1) ISM	→	communism	→	commun
(m > 1) ATE	→	activate	→	activ
(m > 1) ITI	→	angulariti	→	angular
(m > 1) OUS	→	homologous	→	homolog
(m > 1) IVE	→	effective	→	effect
(m > 1) IZE	→	bowdlerize	→	bowdler

Step 5a:

รูปแบบกฎ		ตัวอย่างผลลัพธ์		
(m > 1) E	→	probate	→	probat
		rate	→	rate
(m = 1 and not *o) E	→	cease	→	ceas

Step 5b:

รูปแบบกฎ		ตัวอย่างผลลัพธ์		
(m > 1 and *d and *L)	→ single letter	controll	→	control
		roll	→	roll

ภาพประกอบ 2.7 ขั้นตอนการทำงานของอัลกอริทึม Porter (ต่อ)

2.6 การลดขนาดลักษณะเฉพาะ (Feature Reduction)

การจำแนกประเภทเว็บเพจด้วยวิธีการต่างๆ นิยมใช้การแทนเอกสารแบบ BOW โดยใช้คำที่อยู่ภายในเอกสารเป็นลักษณะเฉพาะ ซึ่งหากเว็บเพจมีลักษณะเฉพาะเป็นจำนวนมากจะทำให้มีอุปสรรคต่อตัวจำแนกประเภท (Classifier) ทำให้ต้องสิ้นเปลืองเนื้อที่และเวลาในการประมวลผล ซึ่งการปรับปรุงประสิทธิภาพอัลกอริทึมที่ใช้ในการเรียนรู้ก็ไม่สามารถจัดการปัญหาลักษณะเฉพาะที่มีขนาดใหญ่ได้ ยิ่งไปกว่านั้นอัลกอริทึมในการจำแนกประเภทหลายตัวยังเกิดปัญหา Overfitting กับข้อมูลการสอน (Training Data) ที่มีขนาดลักษณะเฉพาะมากอีกด้วย ดังนั้นการลดลักษณะเฉพาะจึงเป็นขั้นตอนหนึ่งที่สำคัญซึ่งเป็นการเตรียมข้อมูล (Preprocessing Data) ก่อนการสร้างตัวจำแนกประเภท

สำหรับการจำแนกประเภทเว็บนั้นมีแนวทางในการลดขนาดลักษณะเฉพาะ 2 วิธีด้วยกันคือ การเลือกลักษณะเฉพาะ (Feature Selection) เป็นการเลือกเซตของลักษณะเฉพาะใหม่จากเซตของลักษณะเฉพาะเดิม โดยที่เซตของลักษณะเฉพาะใหม่ที่ได้จะเป็นเซตย่อย (Subset) ของเซตลักษณะเฉพาะเดิม และอีกวิธีคือการสกัดลักษณะเฉพาะ (Feature Extraction) เป็นการสร้างหรือสังเคราะห์เซตของลักษณะเฉพาะใหม่จากเซตของลักษณะเฉพาะเดิม โดยที่เซตของคุณลักษณะเฉพาะใหม่ที่ได้จะแตกต่างจากเซตของลักษณะเฉพาะเดิมและเป็นตัวแทนเซตที่มีขนาดเล็กกว่า ในงานวิจัยนี้ได้นำการเลือกลักษณะด้วยวิธี Information Gain มาใช้ในขั้นตอนการลดลักษณะเฉพาะ ซึ่งเป็นวิธีที่ง่ายและรวดเร็ว

2.6.1 Information Gain

Information Gain (IG) เป็นวิธีการเลือกลักษณะเฉพาะที่นำทฤษฎี Information มาใช้เป็นตัววัดเพื่อพิจารณาเลือกลักษณะเฉพาะโดยใช้วิธีการจัดอันดับ (Ranking) ลักษณะเฉพาะตามค่า Information Gain ซึ่งสามารถคำนวณได้จากสมการที่ (2.4)

$$IG(Y,X) = H(Y) - H(Y \mid X) \tag{2.4}$$

กำหนดให้ Y คือ คลาส

X คือ ลักษณะเฉพาะ

$H(Y)$ คือ ค่าเอนโทรปีของ Y

$H(Y \mid X)$ คือ ค่าเอนโทรปีของ Y เมื่อมีเงื่อนไขตาม X

การหาค่า $H(Y)$ แสดงได้ดังสมการที่ (2.5) และการหาค่า $H(Y \mid X)$ แสดงได้ดังสมการที่ (2.6)

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 p(y) \tag{2.5}$$

$$H(Y \mid X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y \mid x) \log_2 p(y \mid x) \tag{2.6}$$

โดยที่ $p(y)$ คือ ความน่าจะเป็นของค่า y
$p(x)$ คือ ความน่าจะเป็นของค่า x
$p(y \mid x)$ คือ ความน่าจะเป็นของค่า y เมื่อมีเงื่อนไขตามค่า x

2.7 Formal Concept Analysis

Formal Concept Analysis (FCA) เป็นทฤษฎีของการวิเคราะห์ข้อมูลโดยกำหนดเป็นโครงสร้างแนวคิดและเป็นการจัดกลุ่มแนวคิดหรือคอนเซ็ปต์ที่พัฒนาด้วยพื้นฐานทางคณิตศาสตร์และแสดงโครงสร้างในรูปภาพของคอนเซ็ปต์แลททิสซึ่งช่วยให้การวิเคราะห์โครงสร้างที่ซับซ้อนและการค้นหาความสัมพันธ์กันของข้อมูลทำได้ง่ายขึ้น โดยใน FCA จะมีการแทนชุดข้อมูล (Data Set) ในรูปแบบของฟอร์มัลคอนเท็กซ์ (Formal Context) ซึ่งเป็นโครงสร้างพื้นฐานของ FCA ที่ใช้สำหรับอธิบายความสัมพันธ์ของข้อมูล

นิยามที่ 2.1 ฟอร์มัลคอนเท็กซ์ (Formal Context)

กำหนดความสัมพันธ์ $K = (G, M, I)$ โดย G เป็นเซตของออบเจ็กต์ (Object) และ M เป็นเซตของคุณลักษณะ (Attribute) ส่วน I เป็นความสัมพันธ์เชิงคู่ (Binary Relation) ระหว่าง G และ M และถ้าเขียนเป็น $(g, m) \in I$ จะอ่านได้ว่า “ออบเจ็กต์ g มีคุณลักษณะ m”

ฟอร์มัลคอนเท็กซ์สามารถเขียนแทนเป็นตารางเมตริกซ์ได้โดยที่ส่วนหัวของแถว (Row) จะแทนด้วยออบเจ็กต์ และส่วนหัวของสดมภ์ (Column) จะแทนด้วยคุณลักษณะ ส่วนกากบาทในแถว g และสดมภ์ m หมายความว่าออบเจ็กต์ g มีคุณลักษณะ m ดังตารางที่ 2.1 แสดงตัวอย่างของฟอร์มัลคอนเท็กซ์ประกอบด้วย 4 ออบเจ็กต์ และ 7 คุณลักษณะ (Hwang *et al*., 2005)

ตารางที่ 2.1 ตัวอย่างฟอร์มัลคอนเท็กซ์

	a1	a2	a3	a4	a5	a6	a7
o1		x				x	x
o2			x	x	x		
o3			x		x	x	x
o4	x		x				

นิยามที่ 2.2 ตัวดำเนินการ $'$ ($'$ Operation)
เมื่อเซต $A \subseteq G$ ของออบเจ็กต์ จะได้ว่า

$$A' = \{m \in M \mid \forall_g \in A : (g,m) \in I\} \tag{2.7}$$

เมื่อเซต $B \subseteq M$ ของคุณลักษณะ จะได้ว่า

$$B' = \{g \in G \mid \forall_m \in B : (g,m) \in I\} \tag{2.8}$$

นั่นคือ A' เป็นเซตของคุณลักษณะร่วมของออบเจ็กต์ทั้งหมดใน A และ B' เป็นเซตของออบเจ็กต์ทั้งหมดที่มีคุณลักษณะร่วมใน B

นิยามที่ 2.3 ฟอร์มัลคอนเซ็ปต์ (Formal Concept)
ฟอร์มัลคอนเซ็ปต์ C จากฟอร์มัลคอนเท็กซ์ (G,M,I) คือเซตคู่ (A,B) โดยที่ $A \subseteq G$, $B \subseteq M$, $A' = B$ และ $B' = A$ ซึ่งจะเรียกเซต A และ B ว่า “Extension” และ “Intension” ของคอนเซ็ปต์ C ตามลำดับ นั่นคือ Extension แทนกลุ่มของออบเจ็กต์ และ Intension แทนกลุ่มของคุณลักษณะร่วม ซึ่งจากตัวอย่างฟอร์มัลคอนเท็กซ์ในตารางที่ 2.1 จะได้ 9 ฟอร์มัลคอนเซ็ปต์ แสดงดังตารางที่ 2.2

ตารางที่ 2.2 ตัวอย่างฟอร์มัลคอนเซ็ปต์

No.	Formal concept	Extension	Intension
1	(Ø,{a1,a2,a3,a4,a5,a6,a7})	Ø	{a1,a2,a3,a4,a5,a6,a7}
2	({o1},{a2,a6,a7})	{o1}	{a2,a6,a7}
3	({o2},{a3,a4,a5})	{o2}	{a3,a4,a5}
4	({o3},{a3,a5,a6,a7})	{o3}	{a3,a5,a6,a7}
5	({o4},{a1,a3})	{o4}	{a1,a3}
6	({o1,o3},{a6,a7})	{o1,o3}	{a6,a7}
7	({o2,o3}, {a3,a5})	{o2,o3}	{a3,a5}
8	({o2,o3,o4},{a3})	{o2,o3,o4}	{a3}
9	({o1,o2,o3,o4},Ø)	{o1,o2,o3,o4}	Ø

นิยามที่ 2.4 คอนเซ็ปต์แลททิส (Concept Lattice)

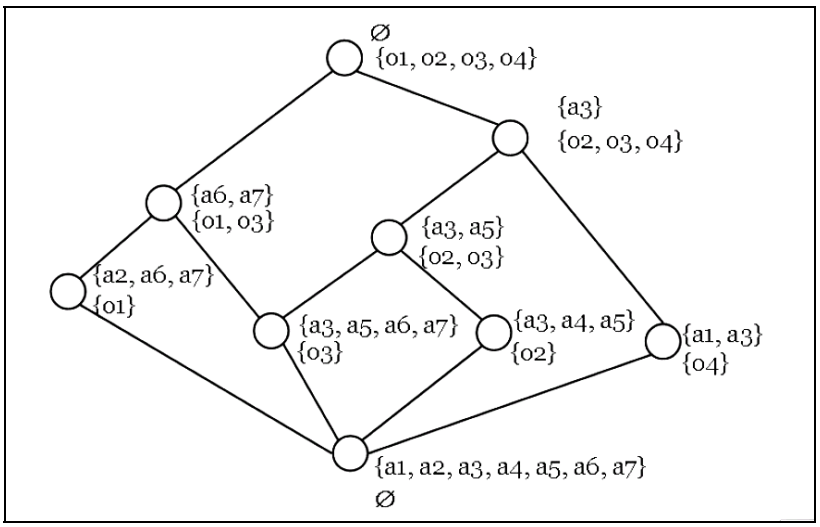
จากความสัมพันธ์ของฟอร์มัลคอนเท็กซ์ $K=(G,M,I)$ กำหนดให้คอนเซ็ปต์ $C_1=(A_1,B_1)$ และ $C_2=(A_2,B_2)$ จะได้ความสัมพันธ์ Subconcept-Superconcept ดังสมการ (2.9)

$$(A_1,B_1)\leq(A_2,B_2)\Leftrightarrow A_1\subseteq A_2(\Leftrightarrow B_1\supseteq B_2) \tag{2.9}$$

นั่นคือคอนเซ็ปต์ (A_1,B_1) จะมีขนาดเล็กกว่า (A_2,B_2) ก็ต่อเมื่อ A_1 เป็นเซตย่อยของ A_2 (และ B_2 เป็นเซตย่อยของ B_1) โดยเซตของคอนเซ็ปต์ทั้งหมดใน K จะถูกจัดลำดับชั้นด้วยความสัมพันธ์เชิงลำดับชั้นเรียกว่า คอนเซ็ปต์แลททิส

การสร้างคอนเซ็ปต์แลททิสจะแทนด้วยไลน์ไดอะแกรม (Line Diagram) ซึ่งประกอบด้วยโหนด (Node) และเส้นเชื่อม (Link) โดยที่แต่โหนดจะแทนด้วยฟอร์มัลคอนเซ็ปต์ ส่วน Intension และ Extension ของฟอร์มัลคอนเซ็ปต์จะระบุไว้ด้านบนและด้านล่างของโหนดตามลำดับ เส้นเชื่อมที่เชื่อมต่อ 2 โหนดเข้าด้วยกันจะแทนความสัมพันธ์ Subconcept-Superconcept ระหว่างกัน ซึ่งจากฟอร์มัลคอนเท็กซ์ในตารางที่ 2.1 สามารถสร้างคอนเซ็ปต์แลททิสได้ดังภาพประกอบ 2.8 (Hwang *et al*., 2005)

จากคอนเซ็ปต์แลททิส (ภาพประกอบ 2.8) สามารถหาออบเจ็กต์กับคุณลักษณะที่สัมพันธ์กันได้ ตัวอย่างเช่น ถ้าต้องการหาว่าออบเจ็กต์ใดที่มีคุณลักษณะเป็น a1 และ a3 ให้พิจารณาคอนเซ็ปต์แลททิสจากโหนดบนลงล่างจะได้ออบเจ็กต์ o4 และถ้าต้องการหาว่าคุณลักษณะใดถูกใช้ร่วมกับออบเจ็กต์ o1 และ o3 ให้พิจารณาจากโหนดล่างขึ้นบนจะได้คุณลักษณะ a6 และ a7 เป็นต้น



ภาพประกอบ 2.8 ตัวอย่างคอนเซ็ปต์แลททิส

2.8 โปรแกรม ConExp (Concept Explorer)

Concept Explorer หรือ ConExp (Yevtushenko, 2000) เป็นโปรแกรมที่พัฒนาฟังก์ชันการทำงานพื้นฐานขึ้นสำหรับการศึกษาวิจัยเกี่ยวกับ FCA ซึ่งโปรแกรม ConExp มีฟังก์ชันการทำงานดังนี้

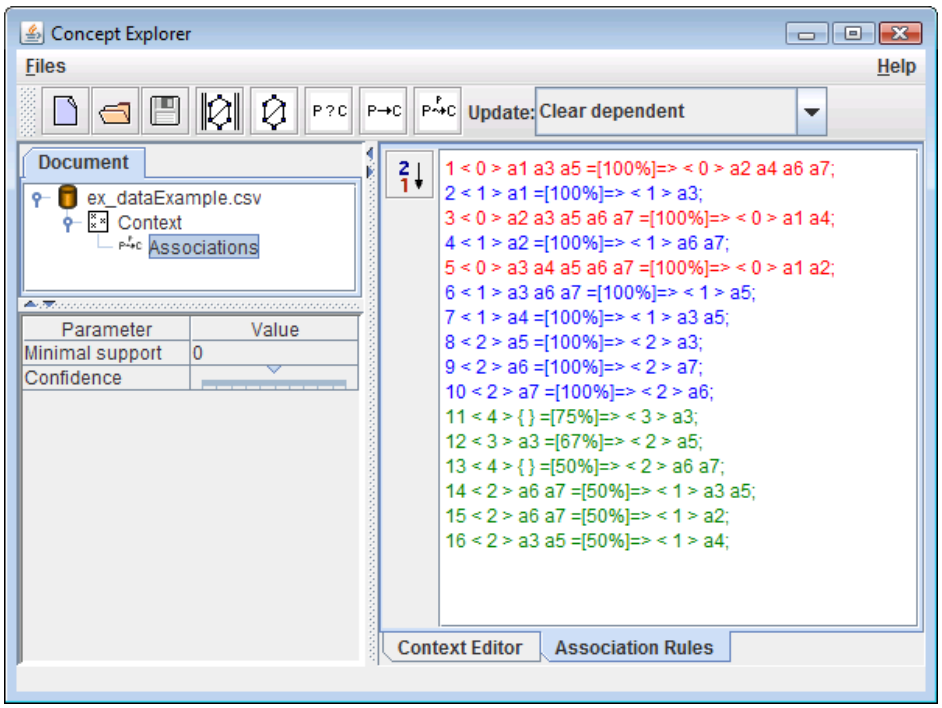
1) สร้างและแก้ไขคอนเท็กซ์
2) สร้างคอนเซ็ปต์แลททิสจากคอนเท็กซ์
3) หาลักษณะเฉพาะที่สัมพันธ์กันที่เกิดขึ้นจริงในคอนเท็กซ์
4) หากฎความสัมพันธ์พื้นฐานที่เป็นจริงในคอนเท็กซ์

งานวิจัยนี้ได้ใช้ฟังก์ชันการหากฎความสัมพันธ์ที่เกิดขึ้นจากฟอร์มัลคอนเท็กซ์ ซึ่งแสดงอยู่ในรูปแบบดังภาพประกอบ 2.9

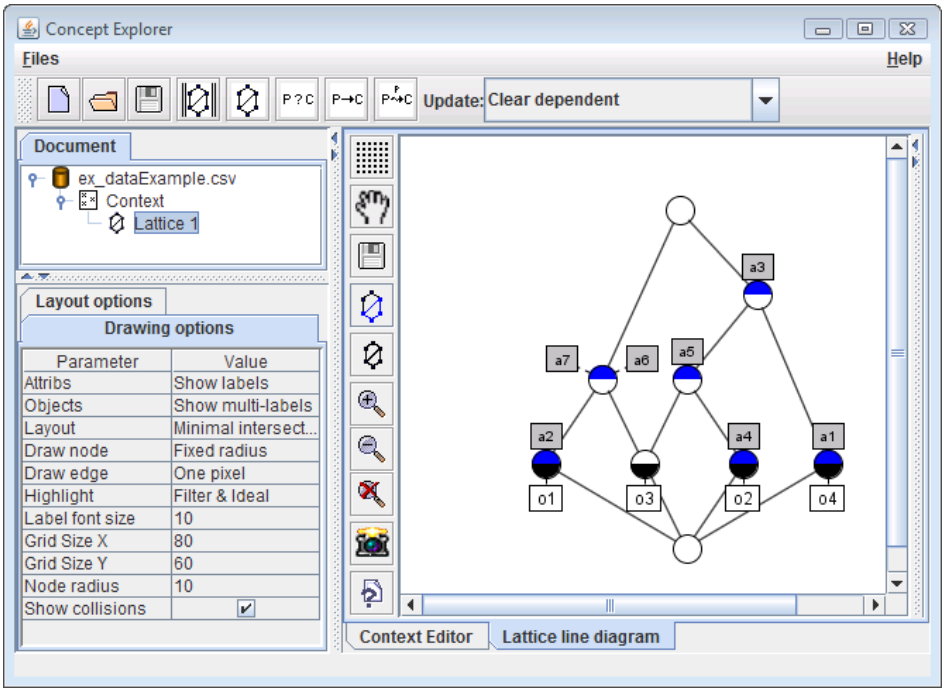
No <Number of objects, for which premise holds> Premise =[Rule confidence]=>
<Number of objects, for which premise and conclusion holds> Conclusion

ภาพประกอบ 2.9 รูปแบบของกฎความสัมพันธ์ที่ได้จากโปรแกรม ConExp

จากตัวอย่างฟอร์มัลคอนเท็กซ์ในตารางที่ 2.1 เมื่อใช้โปรแกรม ConExp สามารถหากฎความสัมพันธ์ที่ได้ดังภาพประกอบ 2.10 และแสดงเป็นคอนเซ็ปต์แลททิสได้ดังภาพประกอบ 2.11



ภาพประกอบ 2.10 ตัวอย่างหน้าต่างผลลัพธ์กฎความสัมพันธ์ด้วยโปรแกรม ConExp



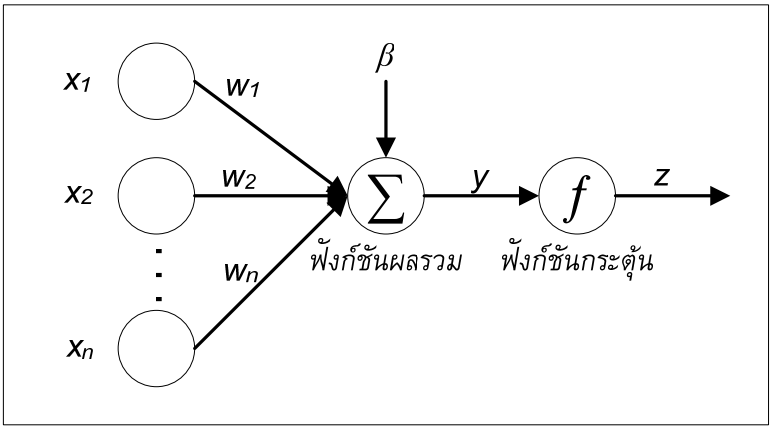
ภาพประกอบ 2.11 ตัวอย่างหน้าต่างผลลัพธ์คอนเซ็ปต์แลททิสด้วยโปรแกรม ConExp

จากหน้าต่างผลลัพธ์กฎความสัมพันธ์ (ภาพประกอบ 2.10) สามารถปรับค่าพารามิเตอร์ Minimal Support และ Confidence ที่ต้องการได้ ซึ่งค่า Minimal Support คือจำนวนออบเจ็กต์ขั้นต่ำที่สัมพันธ์กับคุณลักษณะ และค่า Confidence คือค่าความเชื่อมั่นของกฎ

2.9 โครงข่ายประสาทเทียม (Artificial Neural Networks)

โครงข่ายประสาทเทียม (Artificial Neural Networks) เป็นรูปแบบการประมวลผลโดยเลียนแบบการทำงานของสมองมนุษย์ที่ประกอบด้วยเซลล์ประสาท (Neuron) เป็นจำนวนมาก ข้อดีของโคร่งข่ายประสาทเทียมคือสามารถทำนายค่าที่มีความแม่นยำสูง ทนทานต่อความผิดพลาด (Fault Tolerant) และสามารถรองรับข้อมูลที่ไม่สมบูรณ์ได้ (ณสิทธิ์, 2551)

การประมวลผลที่เลียนแบบการทำงานของเซลล์ประสาทของมนุษย์ประกอบด้วยหน่วยประมวลผลย่อยหรือ Perceptron หลายหน่วยเชื่อมต่อกันเป็นโครงข่าย โดยที่ Perceptron เป็นหน่วยประมวลผลที่เล็กที่สุดของโครงข่ายประสาทเทียมและมีองค์ประกอบดังภาพประกอบ 2.12 ซึ่งประกอบด้วย ฟังก์ชันผลรวม (Summation Function) ทำหน้าที่หาผลรวมของผลคูณระหว่างค่าน้ำหนักข้อมูลเข้ากับค่าข้อมูลเข้าสามารถคำนวณได้ดังสมการที่ (2.10) และฟังก์ชันกระตุ้น (Activation Function) ทำหน้าที่แปลงผลลัพธ์จากฟังก์ชันผลรวมให้อยู่ในช่วงค่าที่ต้องการ ตัวอย่างฟังก์ชันกระตุ้นมีหลายแบบเช่น ฟังก์ชันสเต็ป (Step Function) ฟังก์ชันเส้นตรง (Linear Function) ฟังก์ชันล็อกซิกมอยด์ (Log-Sigmoid Function) และฟังก์ชันแทนซิกมอยด์ (Tan-Sigmoid Function) เป็นต้น

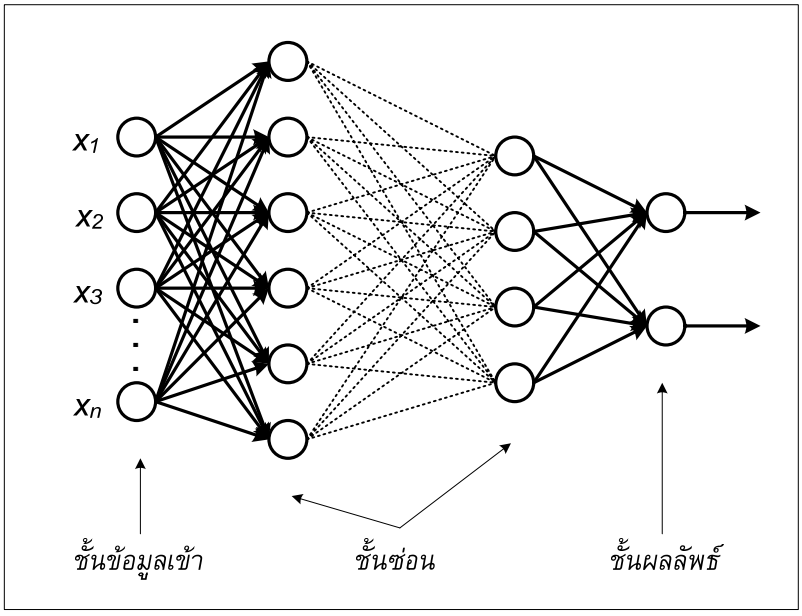


ภาพประกอบ 2.12 องค์ประกอบของ Perceptron

$$y = \sum_{i=1}^{n} x_i w_i + \beta \qquad (2.10)$$

โดยที่ y คือผลลัพธ์ของฟังก์ชันผลรวม
x_i คือค่าข้อมูลเข้าตัวที่ i
n คือจำนวนข้อมูลเข้าทั้งหมด
w_i คือค่าน้ำหนักของข้อมูลเข้าตัวที่ i
β คือค่าความโน้มเอียง

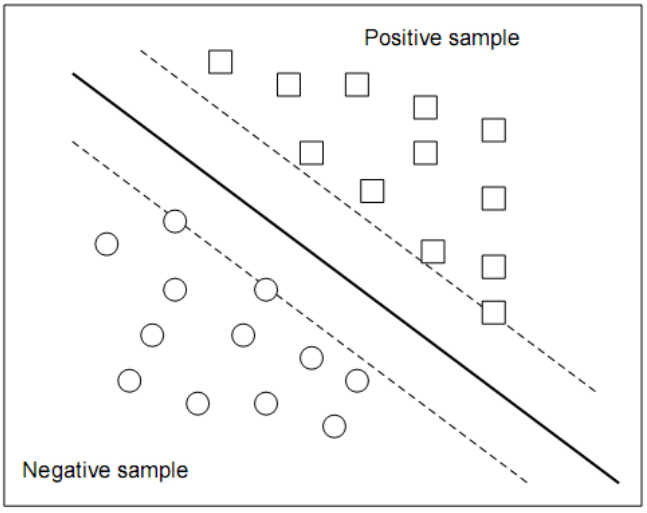
งานวิจัยนี้ได้ใช้โครงข่ายประสาทเทียมแบบ Multi-Layer Perceptron (MLP) (Kotsiantis, 2007) ซึ่งเป็นรูปแบบหนึ่งของโครงข่ายประสาทเทียมที่นำเอา Perceptron หลายหน่วยมาเชื่อมต่อกันเป็นโครงข่ายเพื่อเพิ่มประสิทธิภาพในการจำแนกประเภท โครงสร้างของโครงข่ายประสาทเทียมแบบ MLP ประกอบด้วย 3 ชั้น คือ ชั้นข้อมูลเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และชั้นผลลัพธ์ (Output Layer) โดยที่ชั้นข้อมูลเข้ามี 1 ชั้น ชั้นซ่อนมีกี่ชั้นก็ได้ และชั้นผลลัพธ์มี 1ชั้น และในแต่ละชั้นจะมี Perceptron กี่หน่วยก็ได้ แสดงตัวอย่างดังภาพประกอบ 2.13



ภาพประกอบ 2.13 ตัวอย่างโครงข่ายประสาทเทียมแบบ Multi-Layer Perceptron (MLP)

2.10 Support Vector Machine (SVM)

Support Vector Machine (SVM) (Chen *et al*, 2006; Kotsiantis, 2007) มีแนวคิดหลักคือการหาระนาบการตัดสินใจในการแบ่งข้อมูลออกเป็นสองส่วน ซึ่งใช้สำหรับข้อมูลที่มีมิติของข้อมูลสูง แสดงดังภาพประกอบ 2.14



ภาพประกอบ 2.14 ระนาบตัดสินใจของ SVM

กำหนดให้ (x_i, y_i), …, (x_n, y_n) เป็นตัวอย่างที่ใช้สำหรับการสอน n คือจำนวนข้อมูลตัวอย่าง m คือจำนวนมิติข้อมูลเข้า และ y คือผลลัพธ์มีค่า +1 หรือ -1 ดังสมการที่ (2.11)

$$(x_i, y_i),\ldots,(x_n, y_n) \text{ เมื่อ } x \in R^m,\ y \in \{+1,-1\} \tag{2.11}$$

สำหรับปัญหาเชิงเส้นที่มีมิติข้อมูลขนาดสูงได้ถูกแบ่งเป็น 2 กลุ่มด้วยระนาบตัดสินใจซึ่งคำนวณได้ดังสมการที่ (2.12) และจำแนกประเภทด้วยสมการที่ (2.13)

$$(w \cdot x) + b = 0 \tag{2.12}$$

$$(w \cdot x) + b > 0 \text{ ถ้า } y_i = +1 \text{ และ } (w \cdot x) + b < 0 \text{ ถ้า } y_i = -1 \tag{2.13}$$

โดยที่ w คือค่าน้ำหนัก

b คือค่า bias

2.11 การประเมินประสิทธิภาพตัวจำแนกประเภท (Classifier Evaluation)

การประเมินประสิทธิภาพของตัวจำแนกสามารถแบ่งออกเป็น 3 ด้านดังนี้คือ ด้านประสิทธิภาพในการสอนระบบ (Training Efficiency) ด้านประสิทธิภาพในการจำแนก (Classification Efficiency) และความถูกต้องในการจำแนก (Correctness of Classification) ซึ่งประสิทธิภาพในการสอนระบบและการจำแนกประเภทจะวัดจากความเร็วในการประมวลผลและการใช้เนื้อที่หน่วยความจำเป็นหลัก (Radovanovic, 2006) การวัดประสิทธิภาพนิยมใช้วิธีทางด้านการค้นคืนสารสนเทศ ซึ่งการวัดประสิทธิภาพของระบบจะประเมินจากข้อมูลที่ได้จากการจำแนกประเภทแสดงเป็นตาราง Confusion Matrix (Kohavi and Provost, 1998) ประกอบด้วยข้อมูลที่เป็นค่าจริง (Actual) และข้อมูลที่เป็นค่าทำนาย (Predicted) ดังตารางที่ 2.3 แสดง Confusion Matrix สำหรับการจำแนกประเภท 2 คลาส

ตารางที่ 2.3 Confusion Matrix

Value		**Predicted**	
		Negative	**Positive**
Actual	**Negative**	***a***	***b***
	Positive	***c***	***d***

กำหนดให้ *a* คือจำนวนตัวอย่างที่ทำนายได้ถูกต้องว่าตัวอย่างเป็น Negative
b คือจำนวนตัวอย่างที่ทำนายผิดว่าตัวอย่างเป็น Positive
c คือจำนวนตัวอย่างที่ทำนายผิดว่าตัวอย่างเป็น Negative
d คือจำนวนตัวอย่างที่ทำนายได้ถูกต้องว่าตัวอย่างเป็น Positive

จาก Confusion Matrix ในตารางที่ 2.3 สามารถคำนวณค่าต่าง ๆ ได้ดังนี้

1) *Accuracy (AC)* เป็นสัดส่วนของจำนวนทั้งหมดที่ตัวจำแนกประเภททำนายได้ถูกต้อง สามารถคำนวณได้ดังสมการที่ (2.14)

$$AC = \frac{a + d}{a + b + c + d} \tag{2.14}$$

2) *Recall (R)* หรือ *True Positive (TP)* เป็นสัดส่วนของจำนวนตัวอย่างที่ทำนายได้ถูกต้อง กรณีที่ค่าจริงเป็น Positive สามารถคำนวณได้ดังสมการที่ (2.15)

$$R = \frac{d}{c + d} \tag{2.15}$$

3) *Precision (P)* เป็นสัดส่วนของจำนวนตัวอย่างที่ทำนายได้ถูกต้อง กรณีที่ค่าทำนายเป็น Positive สามารถคำนวณได้ดังสมการที่ (2.16)

$$P = \frac{d}{b + d} \tag{2.16}$$

การวัดประสิทธิภาพด้วยค่า Accuracy ในสมการที่ (2.14) นั้นอาจไม่เหมาะสมเมื่อจำนวนตัวอย่างที่เป็น Negative มากกว่าจำนวนตัวอย่างที่เป็น Positive ยกตัวอย่างเช่น กำหนดให้มีจำนวนตัวอย่าง 1,000 ตัวอย่าง ซึ่งมี 995 ตัวอย่างเป็น Negative และอีก 5 ตัวอย่างเป็น Positive จะเห็นว่าระบบจะจำแนกได้ค่า Accuracy เป็น 99.5% ซึ่งมีค่าจริงเป็น Negative แต่ค่าจริงที่เป็น Positive น้อยมาก ดังนั้นในงานวิจัยนี้ได้ใช้ตัววัดประสิทธิภาพคือ F-Measure (VanRijsbergen, 1979; Sasaki, 2007) ซึ่งจะมีการประเมินค่า P และ R ร่วมด้วย แสดงได้ดังสมการที่ (2.17)

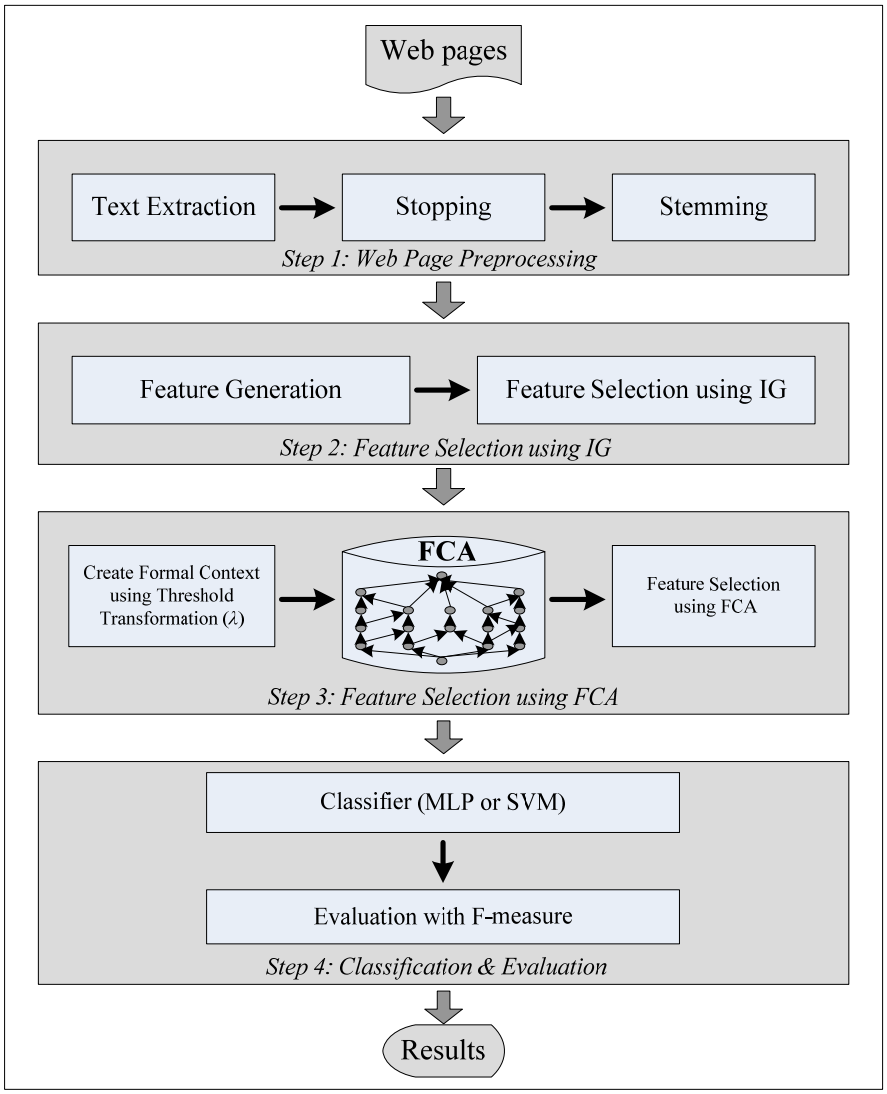
$$F_\beta = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R} \tag{2.17}$$

โดยที่ β คือพารามิเตอร์มีค่าอยู่ระหว่าง $1 - \infty$ เพื่อใช้ควบคุมความสมดุลระหว่าง P และ R เมื่อ $\beta = 1$ จะได้ว่า F_1 มีค่าเท่ากับค่าเฉลี่ยฮาร์โมนิกของ P และ R และถ้า $\beta > 1$ จะได้ว่า F มีค่า *Recall* สูง และถ้า $\beta < 1$ จะได้ว่า F มีค่า *Precision* สูง ยกตัวอย่างเช่น $F_0 = P$ เป็นต้น

บทที่ 3

แบบจำลองการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ

วิทยานิพนธ์นี้ได้ใช้เทคนิค FCA วิเคราะห์หาความสัมพันธ์ระหว่างลักษณะเฉพาะและเอกสารเว็บเพจเพื่อลดขนาดลักษณะเฉพาะสำหรับการจำแนกประเภทเว็บเพจเพื่อให้ได้ผลการจำแนกประเภทที่ถูกต้องมากยิ่งขึ้น โดยสร้างแบบจำลองการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ (Feature Reduction using FCA for Web Page Classification: FR_FCA_WPC) แสดงได้ดังภาพประกอบ 3.1

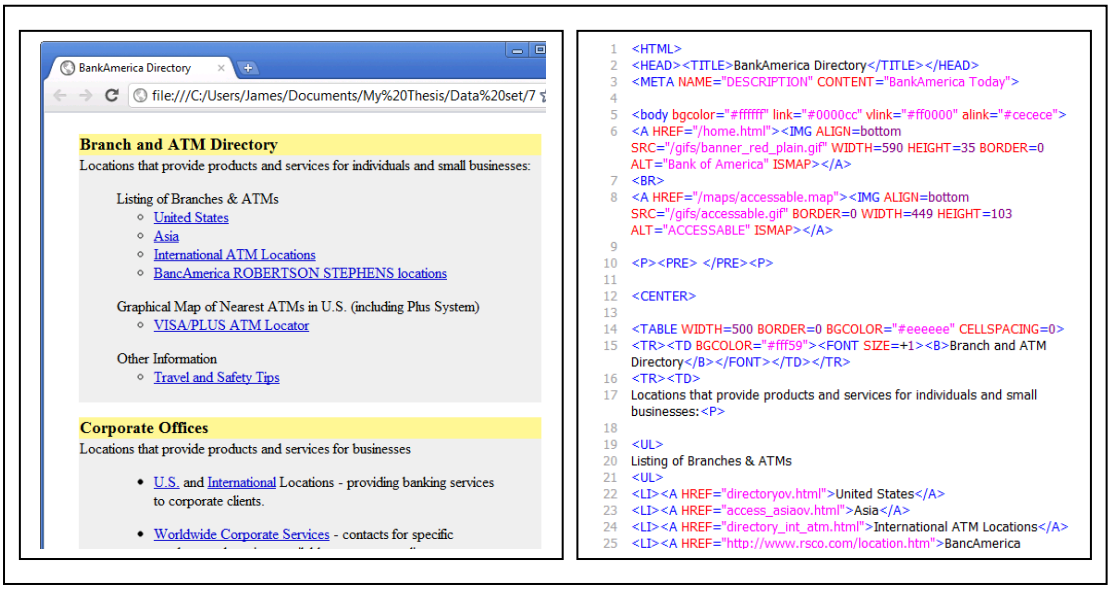


ภาพประกอบ 3.1 แบบจำลองของระบบ FR_FCA_WPC

ซึ่งแบ่งการทำงานออกเป็น 4 ขั้นตอนหลักได้แก่ ขั้นตอนที่ 1 คือ การเตรียมข้อมูลเว็บเพจ (Web Page Preprocessing) ขั้นตอนที่ 2 คือการเลือกลักษณะเฉพาะโดยใช้ IG (Feature Selection using IG) ขั้นตอนที่ 3 คือการเลือกลักษณะเฉพาะโดยใช้ FCA (Feature Selection using FCA) และขั้นตอนที่ 4 คือการจำแนกประเภทและการประเมินผล (Classification and Evaluation) โดยรายละเอียดการทำงานของแบบจำลอง FR_FCA_WPC ในแต่ละขั้นตอนมีดังนี้

3.1 ขั้นตอนการเตรียมข้อมูลเว็บเพจ (Web Page Preprocessing)

โดยทั่วไปเว็บเพจจะประกอบไปด้วยข้อความ (Text) และหัวเรื่อง (Title) แท็ก HTML (HTML tags) ไฮเปอร์ลิงค์ (Hyperlinks) และข้อความเชื่อมโยง (Anchor text) (ข้อความที่ถูกคลิกเพื่อเชื่อมโยงไปยังเว็บเพจอื่นซึ่งอยู่ระหว่างแท็ก <A> และ </A>) รวมถึงข้อความอื่น ๆ ที่มองผ่านเว็บบราวเซอร์ได้ (Qi and Davison, 2009) ดังภาพประกอบ 3.2 ดังนั้นจำเป็นต้องมีการเตรียมข้อมูลเว็บเพจเพื่อกำจัดสิ่งที่ไม่มีประโยชน์สำหรับการวิเคราะห์ออกไป ซึ่งขั้นตอนการเตรียมข้อมูลเว็บเพจแสดงได้ดังภาพประกอบ 3.3



ภาพประกอบ 3.2 ตัวอย่างเว็บเพจ

ขั้นตอนที่ 1 การเตรียมข้อมูลเว็บเพจ (Web Page Preprocessing)
1.1 สกัดข้อความ (Text) และหัวเรื่อง (Title) จากหน้าเว็บเพจ
1.2 กำจัดคำหยุด (Stopping)
1.3 หารากศัพท์ของคำ (Stemming) โดยใช้อัลกอริทึม Porter

ภาพประกอบ 3.3 ขั้นตอนการเตรียมข้อมูลเว็บเพจ (Web Page Preprocessing)

ขั้นตอนที่ 1.1 เป็นการสกัดข้อความ (Text Extraction) โดยกำจัดแท็ก HTML รูปภาพ สื่อมัลติมีเดีย ออกให้เหลือเฉพาะข้อความและหัวเรื่องเท่านั้น จากตัวอย่างเว็บเพจ (ภาพประกอบ 3.2) สามารถสกัดข้อความและหัวเรื่องแสดงได้ดังภาพประกอบ 3.4 และภาพประกอบ 3.5 ตามลำดับ

Branch and ATM Directory Locations that provide products and services for individuals and small businesses: Listing of Branches & ATMs United States Asia International ATM Locations BancAmerica ROBERTSON TEPHENS locations Graphical Map of Nearest ATMs in U.S. (including Plus System) VISA/PLUS ATM Locator Other Information Travel and Safety Tips Corporate Offices Locations that provide products and services for businesses U.S.International Locations - providing banking services to corporate clients. Worldwide Corporate Services - contacts for specific products and services available to corporate clients.

ภาพประกอบ 3.4 ตัวอย่างข้อความที่สกัดได้จากเว็บเพจ

BankAmerica Directory

ภาพประกอบ 3.5 ตัวอย่างหัวเรื่องที่สกัดได้จากเว็บเพจ

ขั้นตอนที่ 1.2 นำคำจากส่วนของข้อความและส่วนของหัวเรื่องที่ได้จากขั้นตอนการสกัดข้อความมากำจัดคำหยุด (Stopping) ซึ่งจะพิจารณาจากรายการคำหยุด (Stoplist) อย่างเช่น คำว่า “a” “an” “the” “I” และ “with” เป็นต้น ซึ่งคำเหล่านี้ไม่ได้เป็นข้อมูลที่มีประโยชน์สำหรับการวิเคราะห์จึงควรกำจัดออก

ขั้นตอนที่ 1.3 นำคำที่ได้จากขั้นตอนที่ 1.2 มาหารากศัพท์ของคำโดยใช้อัลกอริทึม Porter

3.2 ขั้นตอนการเลือกลักษณะเฉพาะโดยใช้ IG (Feature Selection using IG)

หลังจากผ่านการเตรียมข้อมูลเว็บเพจในขั้นตอนที่ 1 แล้วจะนำคำที่ได้จากส่วนของข้อความและหัวเรื่องมาสร้างเป็นลักษณะเฉพาะ (Feature Generation) และเลือกลักษณะเฉพาะโดยใช้ IG (Feature Selection using IG) เพื่อนำลักษณะเฉพาะที่ได้ไปใช้ในขั้นตอนต่อไป ซึ่งขั้นตอนที่ 2 แสดงได้ดังภาพประกอบ 3.6

ขั้นตอนที่ 2 การเลือกลักษณะเฉพาะโดยใช้ IG (Feature Selection using IG)
2.1 สร้างลักษณะเฉพาะจากส่วนของข้อความและหัวเรื่อง 2.1.1 สร้าง Document-Term Matrix จากส่วนของข้อความและหัวเรื่อง 2.1.2 ให้ค่าน้ำหนักคำด้วย TF-IDF 2.1.3 เลือกคำที่มีความถี่เอกสาร (Document Frequency) มากกว่าหรือเท่ากับ DF Threshold 2.1.4 รวมลักษณะเฉพาะที่ได้จากส่วนของข้อความและหัวเรื่องเข้าด้วยกัน
2.2 เลือกลักษณะเฉพาะโดยใช้ IG 2.2.1 ใช้ตัวกรองลักษณะเฉพาะด้วยวิธี Information Gain (IG) 2.2.2 กำหนดจำนวนลักษณะเฉพาะที่ต้องการเท่ากับ *r* เช่น *r* = 50 เป็นต้น

ภาพประกอบ 3.6 ขั้นตอนการเลือกลักษณะเฉพาะโดยใช้ IG (Feature Selection using IG)

ขั้นตอนที่ 2.1 นำคำจากส่วนของข้อความและส่วนของหัวเรื่องที่ได้จากขั้นตอนการเตรียมข้อมูลมาสร้างลักษณะเฉพาะในรูปของ Document-Term Matrix ดังตารางที่ 3.1 และให้ค่าน้ำหนักของคำด้วยวิธีการ TF-IDF ด้วยสมการที่ (2.2) ในบทที่ 2 จากนั้นเลือกคำที่มีค่าความถี่เอกสาร (Document Frequency: DF) มากกว่าหรือเท่ากับค่าที่กำหนดซึ่งมี 2 ค่าคือ DF Threshold ของข้อความและ DF Threshold ของหัวเรื่อง ดังนั้นคำที่มีค่าน้อยกว่าค่าที่กำหนดจะถูกกำจัดออก สุดท้ายรวมลักษณะเฉพาะที่ได้จากส่วนของข้อความและส่วนของหัวเรื่องเข้าด้วยกัน

ตารางที่ 3.1 Document-Term Matrix

Terms / Documents	t_1	t_2	t_3	…	t_m
d_1	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$	…	$w_{1,m}$
d_2	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$	…	$w_{2,m}$
d_3	$w_{3,1}$	$w_{3,2}$	$w_{3,3}$	…	$w_{3,m}$
…	…	…	…	…	…
d_n	$w_{n,1}$	$w_{n,2}$	$w_{n,3}$	…	$w_{n,m}$

ขั้นตอนที่ 2.2 นำลักษณะเฉพาะที่ได้มาเลือกลักษณะเฉพาะโดยใช้วิธี Information Gain (IG) โดยกำหนดให้จำนวนลักษณะเฉพาะที่ต้องการเท่ากับ r เช่น $r = 50$ เป็นต้น

ตัวอย่างชุดข้อมูลเว็บเพจที่เลือกลักษณะเฉพาะโดยใช้วิธี IG โดยกำหนดให้จำนวนลักษณะเฉพาะที่ต้องการเท่ากับ $r = 6$ แสดงดังตารางที่ 3.2 ประกอบด้วยเอกสารเว็บเพจจำนวน 10 เว็บเพจ มีลักษณะเฉพาะจำนวน 6 ลักษณะเฉพาะ และคลาสจำนวน 2 คลาส ในตารางจะแสดงค่าน้ำหนัก (จากค่า TF-IDF) ระหว่างลักษณะเฉพาะ (f) กับเอกสารเว็บเพจ (d) ยกตัวอย่างเช่น ลักษณะเฉพาะ f1 มีค่าน้ำหนักเท่ากับ 1.3 ในเอกสาร d1 เป็นต้น

ตารางที่ 3.2 ตัวอย่างผลลัพธ์ชุดข้อมูลเว็บเพจที่เลือกลักษณะเฉพาะโดยใช้วิธี IG

Features / Web pages	f1	f2	f3	f4	f5	f6	class
d1	1.3	0	0	0	0	0	c1
d2	2.9	0	0	1.3	0	0	c1
d3	2.9	0	0	1.3	0	0.7	c1
d4	1.5	0	0	1.3	0	0.7	c1
d5	1.5	0.2	0	1.3	0	0.7	c1
d6	0	0	3.0	0	0	0	c2
d7	0	0	2.6	0	1.8	0	c2
d8	0	0	1.6	0	1.8	0	c2
d9	0.2	1.6	1.6	0	1.5	0	c2
d10	0.2	1.6	1.6	0	1.5	0	c2

ข้อสังเกตจากตัวอย่างผลลัพธ์ชุดข้อมูลเว็บเพจที่เลือกลักษณะเฉพาะโดยใช้วิธี IG ในตารางที่ 3.2 สามารถคำนวณค่าน้ำหนักเฉลี่ย (Mean) ระหว่างลักษณะเฉพาะ f1 ถึง f6

กับเอกสารเว็บเพจ d1 ถึง d10 โดยการนำผลรวมของค่าน้ำหนักทั้งหมด (ทุกเซลล์) หารด้วยผลคูณระหว่างจำนวนลักษณะเฉพาะกับจำนวนเอกสารเว็บเพจ จากตัวอย่างดังกล่าวค่าน้ำหนักเฉลี่ยมีค่าเท่ากับ 38.2 / (6*10) = 0.6

3.3 ขั้นตอนการเลือกลักษณะเฉพาะโดยใช้ FCA (Feature Selection using FCA)

ในขั้นตอนนี้จะใช้ FCA มาพิจารณาความสัมพันธ์ระหว่างลักษณะเฉพาะ (คำ) กับเอกสารเว็บเพจเพื่อเลือกลักษณะเฉพาะ โดยนำลักษณะเฉพาะที่เลือกด้วยวิธี IG ให้มีขนาดลักษณะเฉพาะเท่ากับ r จากนั้นเลือกลักษณะเฉพาะโดยใช้ FCA ที่สามารถลดขนาดลักษณะเฉพาะลงเท่ากับ s โดยที่ $s \leq r$ โดยมีขั้นตอนแสดงได้ดังภาพประกอบ 3.7

ขั้นตอนที่ 3 การเลือกลักษณะเฉพาะโดยใช้ FCA (Feature Selection using FCA)
3.1 นำลักษณะเฉพาะที่เลือกได้จากขั้นตอนที่ 2 แปลงให้อยู่ในรูปฟอร์มัลคอนเท็กซ์ด้วยฟังก์ชัน Threshold Transformation ดังนิยามที่ 3.1 โดยกำหนดค่า λ
3.2 ใช้โปรแกรม ConExp วิเคราะห์หากฎความสัมพันธ์จากฟอร์มัลคอนเท็กซ์ที่ได้
3.3 เลือกลักษณะเฉพาะโดยพิจารณาจากกฎความสัมพันธ์ที่ได้ตามค่า Minimal Support และค่า Confidence ที่กำหนด

ภาพประกอบ 3.7 ขั้นตอนการเลือกลักษณะเฉพาะโดยใช้ FCA
(Feature Selection using FCA)

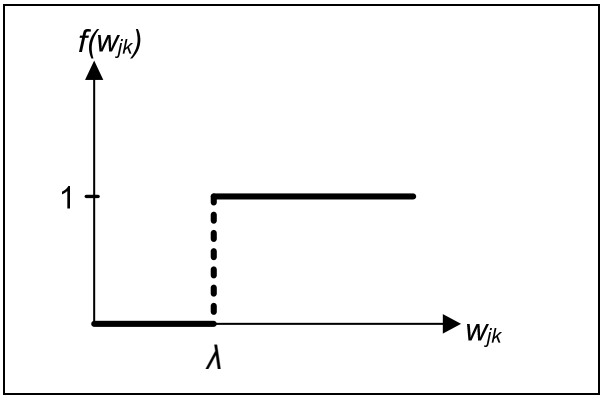
ขั้นตอนที่ 3.1 เป็นการสร้างฟอร์มัลคอนเท็กซ์โดยนำชุดข้อมูลเว็บเพจที่เลือกลักษณะเฉพาะด้วยวิธี IG มาผ่านฟังก์ชัน Threshold Transformation ดังนิยามที่ 3.1 เพื่อแปลงข้อมูลเว็บเพจให้อยู่ในรูปฟอร์มัลคอนเท็กซ์

นิยามที่ 3.1 ฟังก์ชัน Threshold Transformation

กำหนดให้ λ คือเกณฑ์น้ำหนักที่ใช้กำหนดความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารว่าสัมพันธ์กันหรือไม่ ซึ่งสามารถเขียนเป็นฟังก์ชันกำหนดความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารได้ดังสมการที่ (3.1) และแสดงกราฟดังภาพประกอบ 3.8

$$f(w_{jk}) = \begin{cases} 1; & \text{if } w_{jk} > \lambda \\ 0; & \text{if } w_{jk} \leq \lambda \end{cases} \quad (3.1)$$

โดยที่ w_{jk} คือค่าน้ำหนักของลักษณะเฉพาะที่ *k* ที่สัมพันธ์กับเอกสารที่ *j*



ภาพประกอบ 3.8 ฟังก์ชัน Threshold Transformation

จากสมการที่ (3.1) ผลลัพธ์ของ $f(w_{jk})$ จะให้ค่าเป็น 1 หรือ 0 มีดังนี้คือถ้า $f(w_{jk}) = 1$ แสดงว่าลักษณะเฉพาะที่ *k* สัมพันธ์กับเอกสารที่ *j* และถ้า $f(w_{jk}) = 0$ แสดงว่าลักษณะเฉพาะที่ *k* ไม่สัมพันธ์กับเอกสารที่ *j*

ในการแปลงข้อมูลเว็บเพจให้อยู่ในรูปฟอร์มัลคอนเท็กซ์นั้นจะต้องกำหนดเกณฑ์น้ำหนักของความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสาร โดยแบ่งเป็น 2 กรณี ดังนี้

1) กรณี $\lambda = 0$

เมื่อกำหนด $\lambda = 0$ หมายถึง ต้องการเลือกทุกความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารที่มีค่าน้ำหนักมากกว่า 0 มาแปลงเป็นฟอร์มัลคอนเท็กซ์ นั่นคือ ถ้าค่าน้ำหนัก w_{jk} มีค่ามากกว่า 0 ก็จะได้ว่าลักษณะเฉพาะที่ *k* นั้นสัมพันธ์กับเอกสารที่ *j*

จากตัวอย่างชุดข้อมูลเว็บเพจที่เลือกลักษณะเฉพาะโดยใช้ IG ในตารางที่ 3.2 พิจารณาลักษณะเฉพาะ f1 มีค่าน้ำหนักเท่ากับ 1.3 ในเอกสาร d1 ซึ่งเมื่อผ่านฟังก์ชัน Threshold Transformation และกำหนด $\lambda = 0$ จะได้ $f(w_{1,1}) = 1$ นั่นคือ $w_{1,1} > 0$ จะได้ว่าลักษณะเฉพาะ f1 สัมพันธ์กับเอกสาร d1 ดังนั้นจากตารางที่ 3.2 เมื่อผ่านฟังก์ชัน Threshold Transformation และกำหนด $\lambda = 0$ จะได้ดังตารางที่ 3.3 และสามารถแปลงเป็นฟอร์มัลคอนเท็กซ์ได้ดังตารางที่ 3.4

ตารางที่ 3.3 ตัวอย่างชุดข้อมูลเว็บเพจเมื่อผ่านฟังก์ชัน Threshold Transformation ($\lambda = 0$)

Features / Web pages	f1	f2	f3	f4	f5	f6	c1	c2
d1	1	0	0	0	0	0	1	0
d2	1	0	0	1	0	0	1	0
d3	1	0	0	1	0	1	1	0
d4	1	0	0	1	0	1	1	0
d5	1	1	0	1	0	1	1	0
d6	0	0	1	0	0	0	0	1
d7	0	0	1	0	1	0	0	1
d8	0	0	1	0	1	0	0	1
d9	1	1	1	0	1	0	0	1
d10	1	1	1	0	1	0	0	1

ตารางที่ 3.4 ตัวอย่างฟอร์มัลคอนเท็กซ์ที่สร้างจากชุดข้อมูลเว็บเพจ ($\lambda = 0$)

Features / Web pages	f1	f2	f3	f4	f5	f6	c1	c2
d1	x						x	
d2	x			x			x	
d3	x			x		x	x	
d4	x			x		x	x	
d5	x	x		x		x	x	
d6			x					x
d7			x		x			x
d8			x		x			x
d9	x	x	x		x			x
d10	x	x	x		x			x

พิจารณาตัวอย่างฟอร์มัลคอนเท็กซ์ในตารางที่ 3.4 แสดงตัวอย่างฟอร์มัลคอนเท็กซ์ที่สร้างจากตารางที่ 3.2 เมื่อกำหนด $\lambda = 0$ ซึ่งประกอบด้วยเอกสารเว็บเพจ 10 เว็บเพจ มีลักษณะเฉพาะ 6 ลักษณะเฉพาะ และลักษณะเฉพาะคลาส 2 คลาส ซึ่งเป็นตารางเมตริกซ์ส่วนหัวของสดมภ์แทนด้วยลักษณะเฉพาะ (f1 f2 f3 … f6) และลักษณะเฉพาะคลาส (c1 และ c2) และส่วนหัวของแถวแทนด้วยเอกสารเว็บเพจ (d1 d2 d3 …

d10) ส่วนกากบาท (x) ในแถวและสดมภ์จะแทนความสัมพันธ์ระหว่างเอกสารเว็บเพจกับลักษณะเฉพาะและคลาส

2) กรณีปรับค่า λ

การกำหนดค่าน้ำหนักความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารในการสร้างฟอร์มัลคอนเท็กซ์นั้นสามารถปรับค่า λ เพิ่มขึ้นเพื่อเลือกความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารที่มีค่าน้ำหนักสูงขึ้น ซึ่งจะได้ว่าลักษณะเฉพาะที่ *k* สัมพันธ์กับเอกสารที่ *j* นั้นจะต้องมีค่าน้ำหนัก w_{jk} มากกว่าค่า λ ที่กำหนด ดังนั้นถ้าค่าน้ำหนักความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารใดน้อยกว่าหรือเท่ากับค่า λ ที่กำหนดก็จะถูกตัดออกไป เช่น เมื่อกำหนด $\lambda = 0.5$ หมายถึง เลือกความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารที่มีค่าน้ำหนักมากกว่า 0.5 นั่นคือ ถ้าค่าน้ำหนัก w_{jk} มีค่ามากกว่า 0.5 ก็จะได้ว่าลักษณะเฉพาะที่ *k* สัมพันธ์กับเอกสารที่ *j*

จากตัวอย่างชุดข้อมูลเว็บเพจในตารางที่ 3.2 เมื่อผ่านฟังก์ชัน Threshold Transformation และกำหนด $\lambda = 0.5$ ดังตารางที่ 3.5 ซึ่งจะเห็นว่าลักษณะเฉพาะที่ *k* สัมพันธ์กับเอกสารที่ *j* นั้นจะต้องมีค่าน้ำหนัก w_{jk} มากกว่า 0.5 เช่น ลักษณะเฉพาะ f1 มีค่าน้ำหนักเท่ากับ 1.3 ในเอกสาร d1 ซึ่งจะได้ $f(w_{1,1}) = 1$ นั่นคือ $w_{1,1} > 0.5$ แสดงว่าลักษณะเฉพาะ f1 สัมพันธ์กับเอกสาร d1 แต่ลักษณะเฉพาะ f1 ไม่สัมพันธ์กับเอกสาร d9 และ d10 นั่นคือ $w_{9,1} \leq 0.5$ และ $w_{10,1} \leq 0.5$ ตามลำดับ และจากตารางที่ 3.5 สามารถแปลงเป็นฟอร์มัลคอนเท็กซ์ได้ดังตารางที่ 3.6

ตารางที่ 3.5 ตัวอย่างชุดข้อมูลเว็บเพจเมื่อผ่านฟังก์ชัน Threshold Transformation ($\lambda = 0.5$)

Features / Web pages	f1	f2	f3	f4	f5	f6	c1	c2
d1	1	0	0	0	0	0	1	0
d2	1	0	0	1	0	0	1	0
d3	1	0	0	1	0	1	1	0
d4	1	0	0	1	0	1	1	0
d5	1	0	0	1	0	1	1	0
d6	0	0	1	0	0	0	0	1
d7	0	0	1	0	1	0	0	1
d8	0	0	1	0	1	0	0	1
d9	0	1	1	0	1	0	0	1
d10	0	1	1	0	1	0	0	1

ตารางที่ 3.6 ตัวอย่างฟอร์มัลคอนเท็กซ์ที่สร้างจากชุดข้อมูลเว็บเพจ ($\lambda = 0.5$)

Features / Web pages	f1	f2	f3	f4	f5	f6	c1	c2
d1	x						x	
d2	x			x			x	
d3	x			x		x	x	
d4	x			x		x	x	
d5	x			x		x	x	
d6			x					x
d7			x		x			x
d8			x		x			x
d9		x	x		x			x
d10		x	x		x			x

จากตารางที่ 3.2 เมื่อผ่านฟังก์ชัน Threshold Transformation และกำหนด $\lambda = 1.0$ จะได้ดังตารางที่ 3.7 ซึ่งจะพบว่าลักษณะเฉพาะ f6 ไม่สัมพันธ์กับเอกสารเว็บเพจใดเลย นั่นคือ $w_{1,6} \leq 1.0$, $w_{2,6} \leq 1.0$, ..., $w_{10,6} \leq 1.0$ ดังนั้นทำให้จำนวนลักษณะเฉพาะที่ได้ลดลงคือ f1 f2 f3 f4 และ f5 (ตัด f6 ออกได้) และจากตารางที่ 3.7 สามารถแปลงเป็นฟอร์มัลคอนเท็กซ์ได้ดังตารางที่ 3.8

ตารางที่ 3.7 ตัวอย่างชุดข้อมูลเว็บเพจเมื่อผ่านฟังก์ชัน Threshold Transformation ($\lambda = 1.0$)

Features / Web pages	f1	f2	f3	f4	f5	f6	c1	c2
d1	1	0	0	0	0	0	1	0
d2	1	0	0	1	0	0	1	0
d3	1	0	0	1	0	0	1	0
d4	1	0	0	1	0	0	1	0
d5	1	0	0	1	0	0	1	0
d6	0	0	1	0	0	0	0	1
d7	0	0	1	0	1	0	0	1
d8	0	0	1	0	1	0	0	1
d9	0	1	1	0	1	0	0	1
d10	0	1	1	0	1	0	0	1

ตารางที่ 3.8 ตัวอย่างฟอร์มัลคอนเท็กซ์ที่สร้างจากชุดข้อมูลเว็บเพจ ($\lambda = 1.0$)

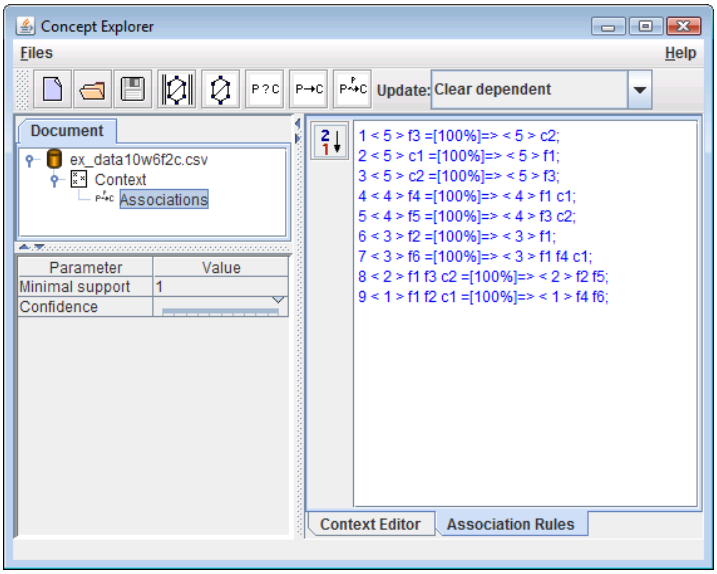
Features / Web pages	f1	f2	f3	f4	f5	f6	c1	c2
d1	x						x	
d2	x			x			x	
d3	x			x			x	
d4	x			x			x	
d5	x			x			x	
d6			x					x
d7			x		x			x
d8			x		x			x
d9		x	x		x			x
d10		x	x		x			x

ขั้นตอนที่ 3.2 ใช้โปรแกรม ConExp วิเคราะห์หาความสัมพันธ์จากฟอร์มัลคอนเท็กซ์ที่ได้จากขั้นตอนที่ 3.1 ซึ่งจากตัวอย่างฟอร์มัลคอนเท็กซ์ ($\lambda = 0$) ในตารางที่ 3.4 สามารถวิเคราะห์หาความสัมพันธ์ได้ผลลัพธ์ออกมาดังภาพประกอบ 3.9 และแสดงเป็นคอนเซ็ปต์แลททิสได้ดังภาพประกอบ 3.10

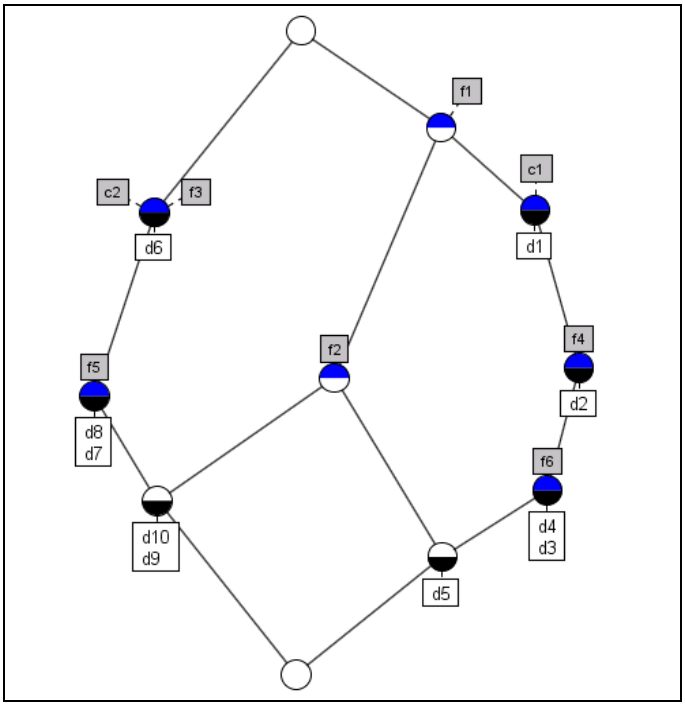
จากตัวอย่างผลลัพธ์ที่ได้จากการวิเคราะห์หากฎความสัมพันธ์ด้วยโปรแกรม ConExp (ภาพประกอบ 3.9) ซึ่งกำหนดค่า Minimal Support เท่ากับ 1 (เลือกลักษณะเฉพาะที่สัมพันธ์กับเอกสารเว็บเพจอย่างน้อย 1 เว็บเพจ) และค่า Confidence เท่ากับ 100% จะได้กฎความสัมพันธ์ออกมา 9 กฎ โดยเรียงกฎตามค่า Support จากมากไปน้อย

เมื่อพิจารณากฎ 1 <5> f3 =[100%]=> <5> c2; แสดงให้เห็นว่าลักษณะเฉพาะ f3 สัมพันธ์กับเอกสารเว็บเพจจำนวน 5 เว็บเพจ และเอกสารเว็บเพจทั้ง 5 เว็บเพจอยู่ในคลาส c2 จากคอนเซ็ปต์แลททิสจะเห็นว่าลักษณะเฉพาะ f3 สัมพันธ์กับเอกสารเว็บเพจ d6 d7 d8 d9 และ d10 ซึ่งสามารถสรุปได้ว่าลักษณะเฉพาะ f3 เป็นลักษณะเฉพาะเด่นของคลาส c2 เนื่องจากลักษณะเฉพาะ f3 มีความสัมพันธ์กับเอกสารเว็บเพจในคลาส c2 สูงที่สุด

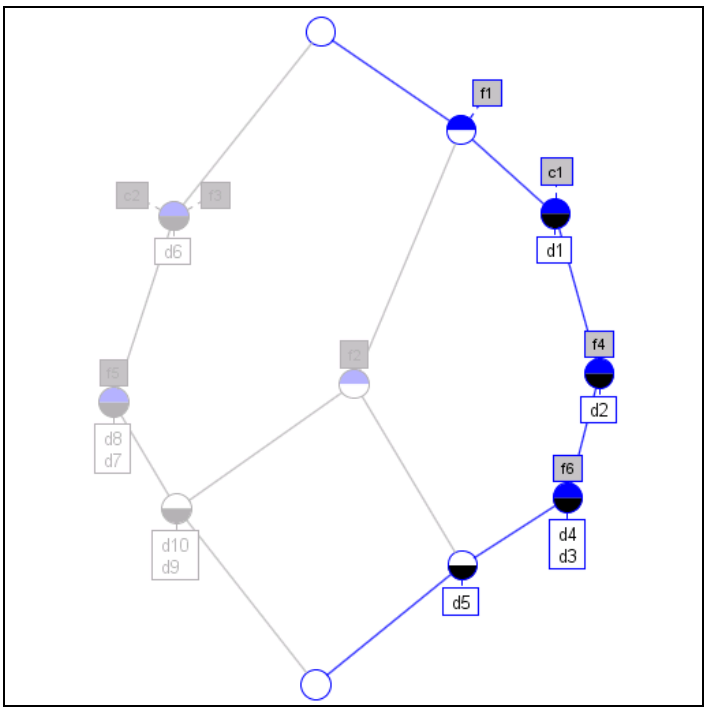
จากคอนเซ็ปต์แลททิส (ภาพประกอบ 3.10) เมื่อพิจารณาความสัมพันธ์ระหว่างคลาสจะเห็นว่าสามารถแบ่งคลาสออกเป็น c1 และ c2 ได้ชัดเจน ซึ่งคลาส c1 ประกอบด้วยเอกสารเว็บเพจ d1 d2 d3 d4 และ d5 และมีลักษณะเฉพาะที่สัมพันธ์กับคลาส c1 คือ f1 f4 และ f6 แสดงได้ดังภาพประกอบ 3.11 ส่วนคลาส c2 ประกอบด้วยเอกสารเว็บเพจ d6 d7 d8 d9 และ d10 และมีลักษณะเฉพาะที่สัมพันธ์กับคลาส c2 คือ f3 และ f5 แสดงได้ดังภาพประกอบ 3.12



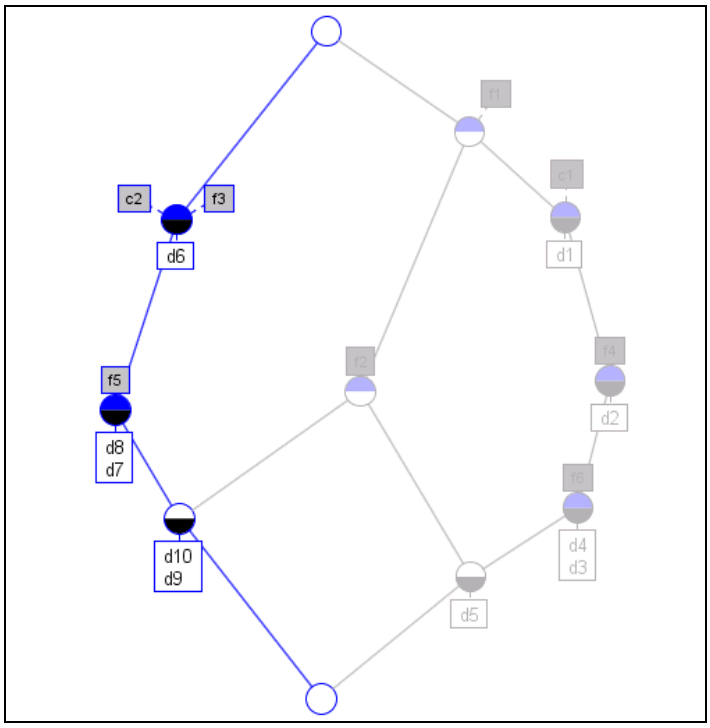
ภาพประกอบ 3.9 ผลลัพธ์กฎความสัมพันธ์ของฟอร์มัลคอนเท็กซ์ ($\lambda = 0$)



ภาพประกอบ 3.10 ผลลัพธ์คอนเซ็ปต์แลททิสของฟอร์มัลคอนเท็กซ์ ($\lambda = 0$)

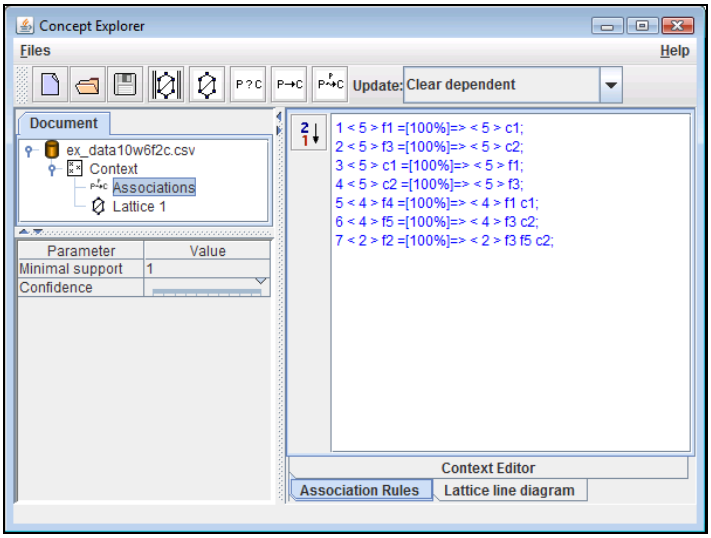


ภาพประกอบ 3.11 คอนเซ็ปต์แลททิสแสดงลักษณะเฉพาะและเอกสารเว็บเพจที่สัมพันธ์กับคลาส c1 ของฟอร์มัลคอนเท็กซ์ ($\lambda = 0$)

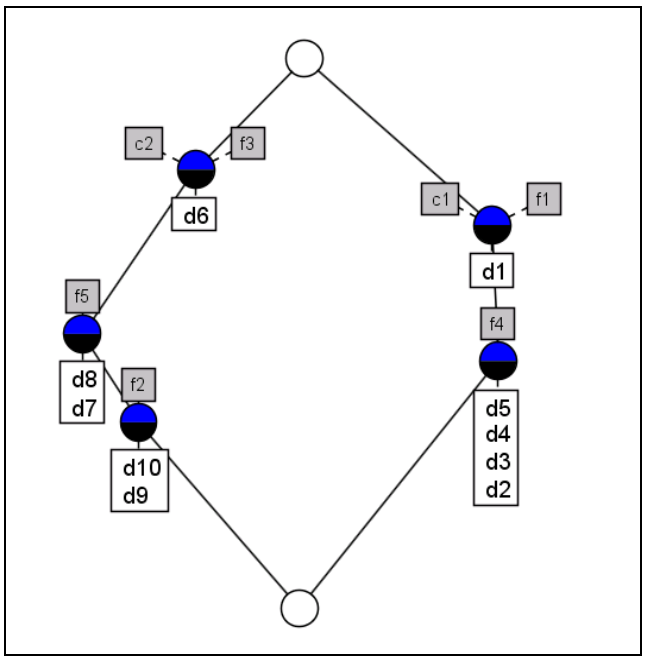


ภาพประกอบ 3.12 คอนเซ็ปต์แลททิสแสดงลักษณะเฉพาะและเอกสารเว็บเพจที่สัมพันธ์กับคลาส c2 ของฟอร์มัลคอนเท็กซ์ ($\lambda = 0$)

จากตัวอย่างฟอร์มัลคอนเท็กซ์ในตารางที่ 3.8 เมื่อปรับค่า $\lambda = 1.0$ สามารถวิเคราะห์หากฎความสัมพันธ์ได้ผลลัพธ์ออกมาดังภาพประกอบ 3.13 และแสดงเป็นคอนเซ็ปต์แลททิสได้ดังภาพประกอบ 3.14



ภาพประกอบ 3.13 ผลลัพธ์กฎความสัมพันธ์ของฟอร์มัลคอนเท็กซ์ ($\lambda = 1.0$)



ภาพประกอบ 3.14 ผลลัพธ์คอนเซ็ปต์แลททิสของฟอร์มัลคอนเท็กซ์ ($\lambda = 1.0$)

จากตัวอย่างผลลัพธ์กฎความสัมพันธ์ของฟอร์มัลคอนเท็กซ์ ($\lambda = 1.0$) ที่ได้จากการวิเคราะห์หาความสัมพันธ์ด้วยโปรแกรม ConExp (ภาพประกอบ 3.13) ซึ่งกำหนดค่า Minimal Support เท่ากับ 1 และค่า Confidence เท่ากับ 100% จะได้กฎความสัมพันธ์ออกมา 7 กฎ จะเห็นว่าจำนวนกฎที่ได้ลดลง (จาก $\lambda = 0$) เนื่องจากน้ำหนักความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารบางตัวถูกตัดออกไป และทำให้ลักษณะเฉพาะ f6 ถูกตัดออกไปด้วย เพราะไม่สัมพันธ์กับเอกสารใดเลย

ขั้นตอนที่ 3.3 จากกฎความสัมพันธ์ที่ได้จากขั้นตอนที่ 3.2 จะเลือกลักษณะเฉพาะโดยพิจารณาจากกฎที่มีลักษณะเฉพาะ (f) สัมพันธ์กับคลาส (c) เท่านั้น ดังนั้น จากภาพประกอบ 3.9 กฎความสัมพันธ์ของฟอร์มัลคอนเท็กซ์ ($\lambda = 0$) เมื่อพิจารณาเลือกลักษณะเฉพาะที่สัมพันธ์กับคลาสจะได้ดังตารางที่ 3.9 และจากภาพประกอบ 3.13 กฎความสัมพันธ์ของฟอร์มัลคอนเท็กซ์ ($\lambda = 1.0$) เมื่อพิจารณาเลือกลักษณะเฉพาะที่สัมพันธ์กับคลาสจะได้ดังตารางที่ 3.10

ตารางที่ 3.9 ตัวอย่างลักษณะเฉพาะที่สัมพันธ์กับคลาสจากกฎความสัมพันธ์ของฟอร์มัลคอนเท็กซ์ ($\lambda = 0$)

No.	ลักษณะเฉพาะ	คลาส	จำนวนเอกสารเว็บเพจ
1	f3	c2	5
2	f1	c1	5
3	f4 f1	c1	4
4	f5 f3	c2	4
5	f6 f1 f4	c1	3
6	f1 f3 f2 f5	c2	2
7	f1 f2 f4 f6	c1	1

ตารางที่ 3.10 ตัวอย่างลักษณะเฉพาะที่สัมพันธ์กับคลาสจากกฎความสัมพันธ์ของฟอร์มัลคอนเท็กซ์ ($\lambda = 1.0$)

No.	ลักษณะเฉพาะ	คลาส	จำนวนเอกสารเว็บเพจ
1	f1	c1	5
2	f3	c2	5
3	f4 f1	c1	4
4	f5 f3	c2	4
5	f2 f3 f5	c2	2

การเลือกลักษณะเฉพาะจะใช้การจัดอันดับโดยเรียงตามลักษณะเฉพาะที่สัมพันธ์กับจำนวนเอกสารเว็บเพจมากที่สุดไปหาน้อยที่สุด ซึ่งลักษณะเฉพาะที่ดีที่สุดจะถูกเลือกในอันดับแรก ๆ จากตารางที่ 3.9 ลักษณะเฉพาะ f3 และ f1 ในลำดับที่ 1 และ 2 สัมพันธ์กับเอกสารเว็บเพจมากที่สุดเท่ากับ 5 เว็บเพจ ดังนั้นลักษณะเฉพาะ f3 และ f1 จะถูกเลือกก่อน จากนั้นพิจารณาในลำดับถัดไปที่มีจำนวนเอกสารเว็บเพจน้อยลงมาตามลำดับ ลักษณะเฉพาะ f4 และ f5 ซึ่งสัมพันธ์กับเอกสารเว็บเพจจำนวน 4 เว็บเพจและลักษณะเฉพาะ f6 และ f2 ที่สัมพันธ์กับเอกสารเว็บเพจจำนวน 3 เว็บเพจและ 2 เว็บเพจตามลำดับ ดังนั้นอันดับของลักษณะเฉพาะที่ได้จากการเลือกด้วยวิธีการ FCA คือ f3 f1 f4 f5 f6 และ f2 ตามลำดับแสดงได้ดังตารางที่ 3.11

จากตารางที่ 3.10 ลักษณะเฉพาะ f1 และ f3 ในลำดับที่ 1 และ 2 สัมพันธ์กับเอกสารเว็บเพจมากที่สุดเท่ากับ 5 เว็บเพจ ดังนั้นลักษณะเฉพาะ f1 และ f3 จะถูกเลือกก่อน จากนั้นพิจารณาในลำดับถัดไปที่มีจำนวนเอกสารเว็บเพจน้อยลงมาตามลำดับ ลักษณะเฉพาะ f4 และ f5 ซึ่งสัมพันธ์กับเอกสารเว็บเพจจำนวน 4 เว็บเพจ และลักษณะเฉพาะ f2 ที่สัมพันธ์กับเอกสารเว็บเพจจำนวน 2 เว็บเพจ ดังนั้นอันดับของลักษณะเฉพาะที่ได้จากการเลือกด้วยวิธีการ FCA คือ f1 f3 f4 f5 และ f2 ตามลำดับแสดงได้ดังตารางที่ 3.12

ตารางที่ 3.11 ตัวอย่างลักษณะเฉพาะที่ได้จากการเลือกด้วยวิธีการ FCA ($\lambda = 0$)

No.	ลักษณะเฉพาะ	จำนวนเอกสารเว็บเพจ
1	f3	5
2	f1	5
3	f4	4
4	f5	4
5	f6	3
6	f2	2

ตารางที่ 3.12 ตัวอย่างลักษณะเฉพาะที่ได้จากการเลือกด้วยวิธีการ FCA ($\lambda = 1.0$)

No.	ลักษณะเฉพาะ	จำนวนเอกสารเว็บเพจ
1	f1	5
2	f3	5
3	f4	4
4	f5	4
5	f2	2

3.4 ขั้นตอนการจำแนกประเภทและการประเมินผล (Classification and Evaluation)

งานวิจัยนี้ได้แบ่งขั้นตอนการจำแนกประเภทและการประเมินผลของตัวจำแนกประเภท แสดงดังภาพประกอบ 3.15

ขั้นตอนที่ 4 การจำแนกประเภทและการประเมินผล (Classification and Evaluation)
4.1 การจำแนกประเภท 4.1.1 เลือกตัวจำแนกประเภทด้วย MLP หรือ SVM 4.1.2 เลือกการทดสอบแบบ 10-folds Cross Validation
4.2 การประเมินผลของตัวจำแนกประเภท 4.2.1 ประเมินผลของตัวจำแนกประเภทด้วยค่า F-measure

ภาพประกอบ 3.15 ขั้นตอนการจำแนกประเภทและการประเมินผล (Classification and Evaluation)

ขั้นตอนที่ 4.1 ใช้ตัวจำแนกประเภท (Classifier) คือ Multi-Layer Perceptron Neural Networks (MLP) และ Support Vector Machine (SVM) ทดสอบแบบ 10-folds Cross Validation โดยใช้โปรแกรม WEKA (Hall *et al*., 2009)

ขั้นตอนที่ 4.2 ประเมินผลของตัวจำแนกประเภทด้วยค่า F-measure จากสมการ (2.17) โดยในการทดลองของงานวิจัยนี้ได้ใช้พารามิเตอร์ $\beta = 1$ ซึ่งจะได้ F_1 มีค่าเฉลี่ยระหว่าง P และ R ที่สมดุลกัน ดังสมการที่ (3.2)

$$F_1 = \frac{2PR}{P + R} \tag{3.2}$$

บทที่ 4

โปรแกรมการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ

โปรแกรมการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจได้พัฒนาขึ้นตามแนวคิดและขั้นตอนการทำงานของแบบจำลองการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ (Feature Reduction using Formal Concept Analysis for Web Page Classification: FR_FCA_WPC) ซึ่งการทำงานของโปรแกรมอธิบายด้วยผังการทำงานของโปรแกรมและส่วนประกอบของโปรแกรมพร้อมผลการทำงานของโปรแกรม ดังรายละเอียดต่อไปนี้

4.1 ผังการทำงานของโปรแกรม

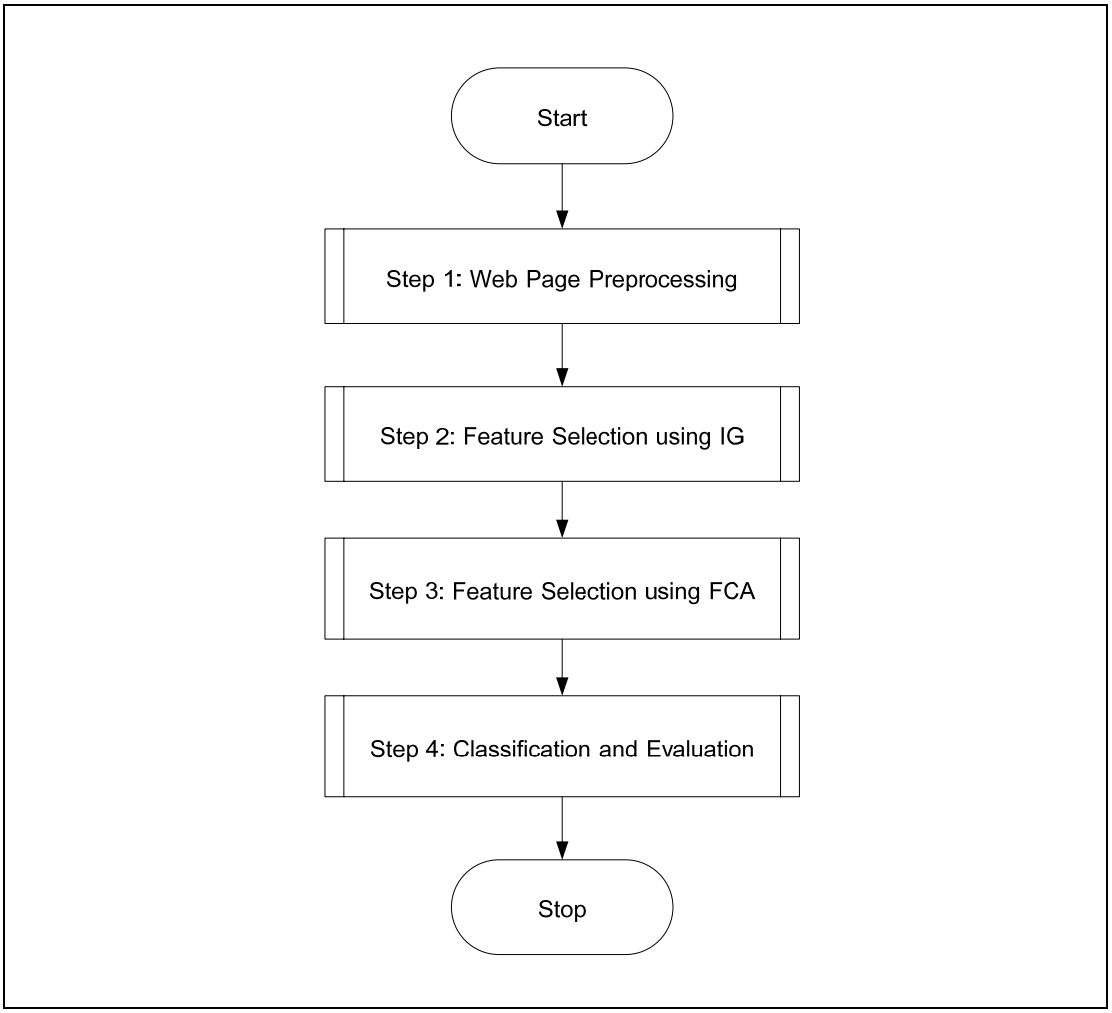
4.1.1 ผังงานโปรแกรมหลักของโปรแกรม FR_FCA_WPC แสดงดังภาพประกอบ 4.1 โดยขั้นตอนการทำงานของโปรแกรมประกอบด้วย 4 ขั้นตอนหลักคือ ขั้นตอนที่ 1 การเตรียมข้อมูลเว็บเพจ (Step 1: Web Page Preprocessing) ขั้นตอนที่ 2 การเลือกลักษณะเฉพาะโดยใช้ IG (Step 2: Feature Selection using IG) ขั้นตอนที่ 3 การเลือกลักษณะเฉพาะโดยใช้ FCA (Step 3: Feature Selection using FCA) และขั้นตอนที่ 4 การจำแนกประเภทและการประเมินผล (Step 4: Classification and Evaluation)

4.1.2 ผังงานโปรแกรมย่อยของ Step 1: Web Page Preprocessing แสดงดังภาพประกอบ 4.2 เป็นขั้นตอนการเตรียมข้อมูลเว็บเพจโดยการสกัดข้อความและหัวเรื่องจากหน้าเว็บเพจ จากนั้นหารากศัพท์ของคำโดยใช้อัลกอริทึม Porter และกำจัดคำหยุดโดยพิจารณาจาก Stoplist

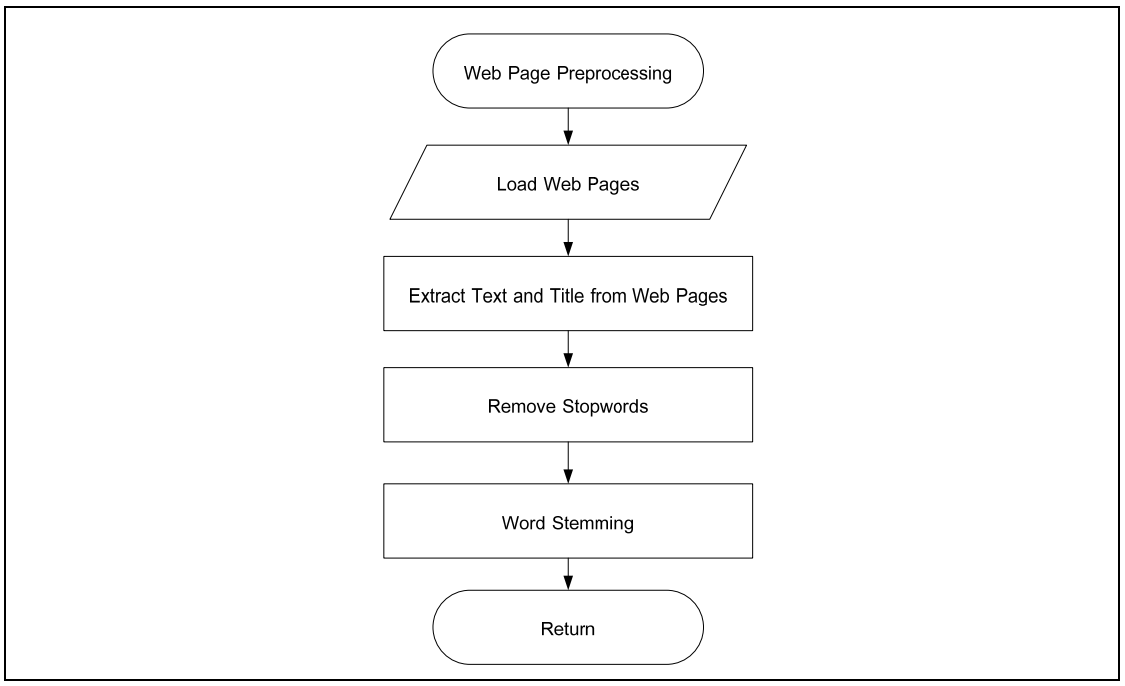
4.1.3 ผังงานโปรแกรมย่อยของ Step 2: Feature Selection using IG แสดงดังภาพประกอบ 4.3 เป็นขั้นตอนการสร้างลักษณะเฉพาะจากส่วนของข้อความและหัวเรื่อง โดยสร้าง Document-Term Matrix จากส่วนของข้อความและหัวเรื่อง และให้ค่าน้ำหนักคำด้วย TF-IDF จากนั้นเลือกคำที่มีความถี่เอกสารมากกว่าหรือเท่ากับค่า Threshold ที่กำหนด และรวมลักษณะเฉพาะที่ได้จากส่วนของข้อความและหัวเรื่องเข้าด้วยกัน สุดท้ายเลือกลักษณะเฉพาะด้วยวิธี IG โดยระบุจำนวนลักษณะเฉพาะที่ต้องการ (*r*)

4.1.4 ผังงานโปรแกรมย่อยของ Step 3: Feature Selection using FCA แสดงดังภาพประกอบ 4.4 เป็นขั้นตอนการเลือกลักษณะเฉพาะโดยใช้ FCA โดยระบุค่าเกณฑ์น้ำหนัก (λ) เพื่อแปลงข้อมูลลักษณะเฉพาะที่เลือกด้วยวิธี IG ให้อยู่ในรูปฟอร์มัลคอนเท็กซ์ จากนั้นใช้โปรแกรม ConExp วิเคราะห์หาความสัมพันธ์จากฟอร์มัลคอนเท็กซ์ที่ได้และเลือกลักษณะเฉพาะโดยพิจารณาจากกฎความสัมพันธ์ที่ได้

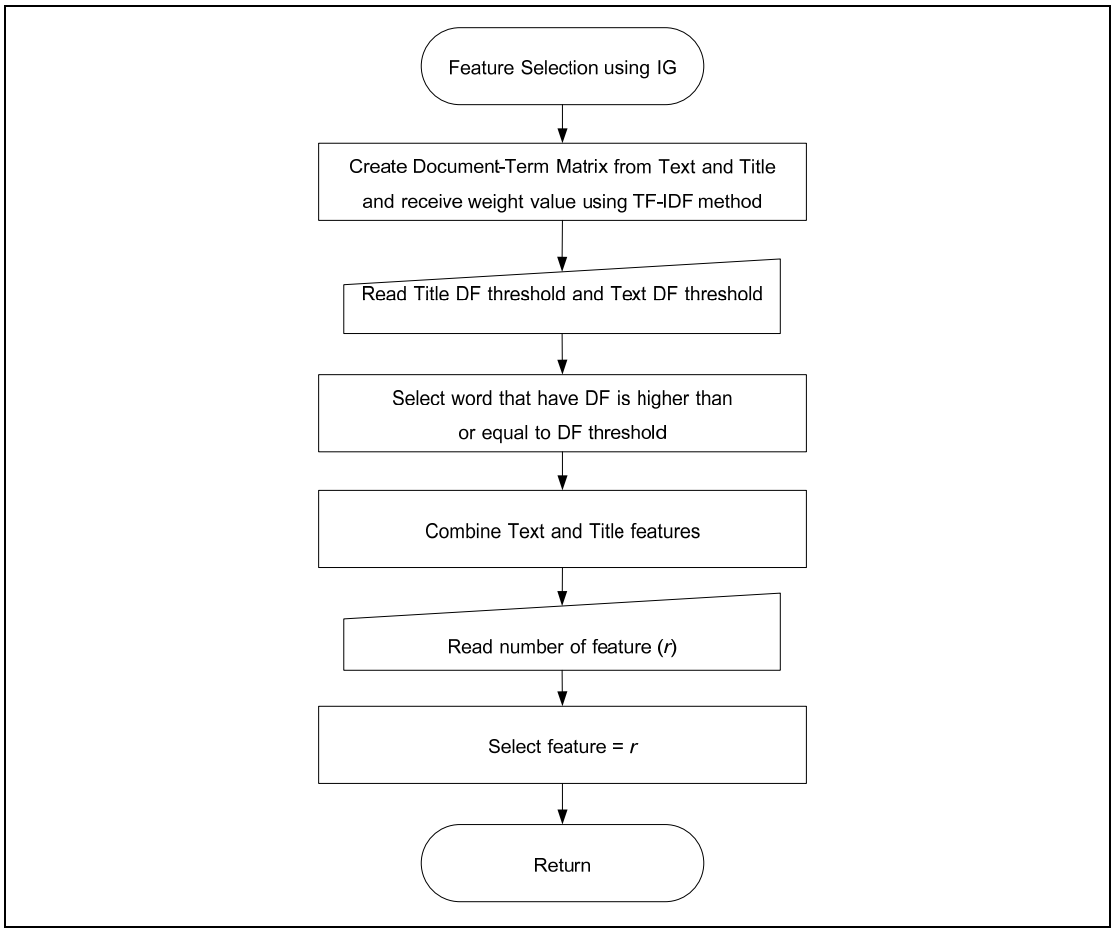
4.1.5 ผังงานโปรแกรมย่อยของ Step 4: Classification and Evaluation แสดงดังภาพประกอบ 4.5 เป็นขั้นตอนการจำแนกประเภทและการประเมินผลโดยกำหนดตัวจำแนกประเภทและระบุจำนวน Fold ของการทดสอบแบบ Cross Validation และประเมินผลการจำแนกด้วยค่า F-measure



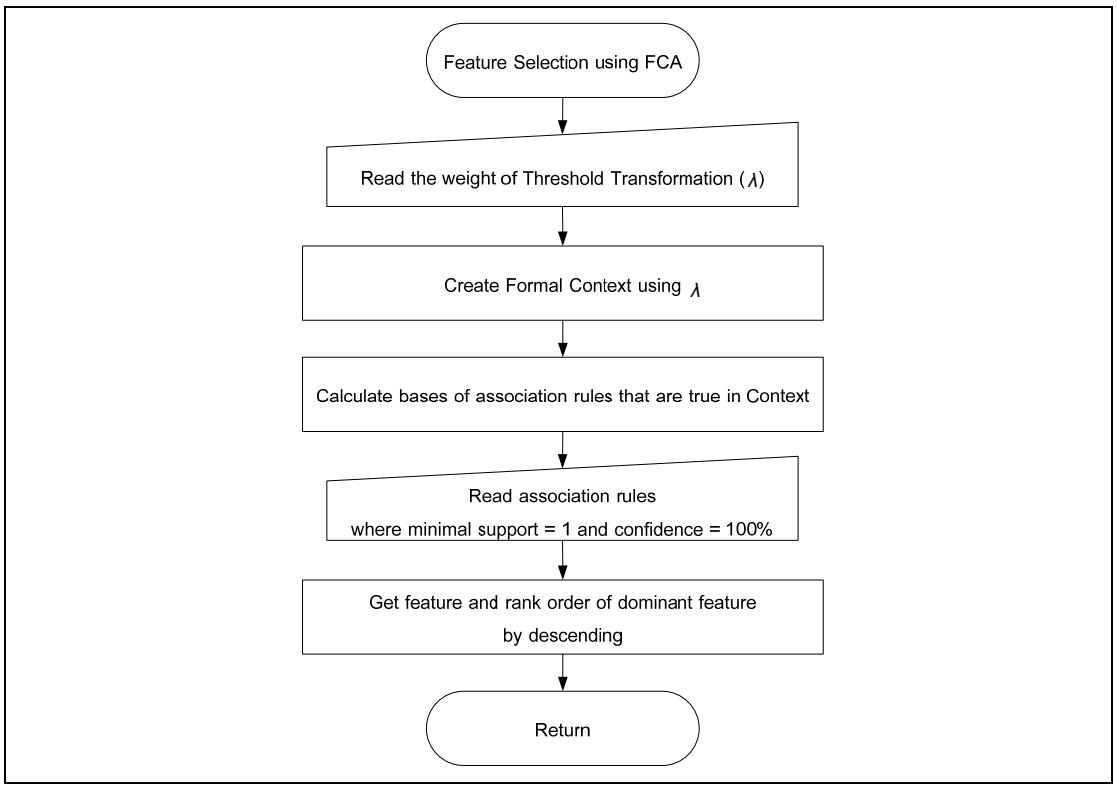
ภาพประกอบ 4.1 ผังการทำงานของโปรแกรม FR_FCA_WPC



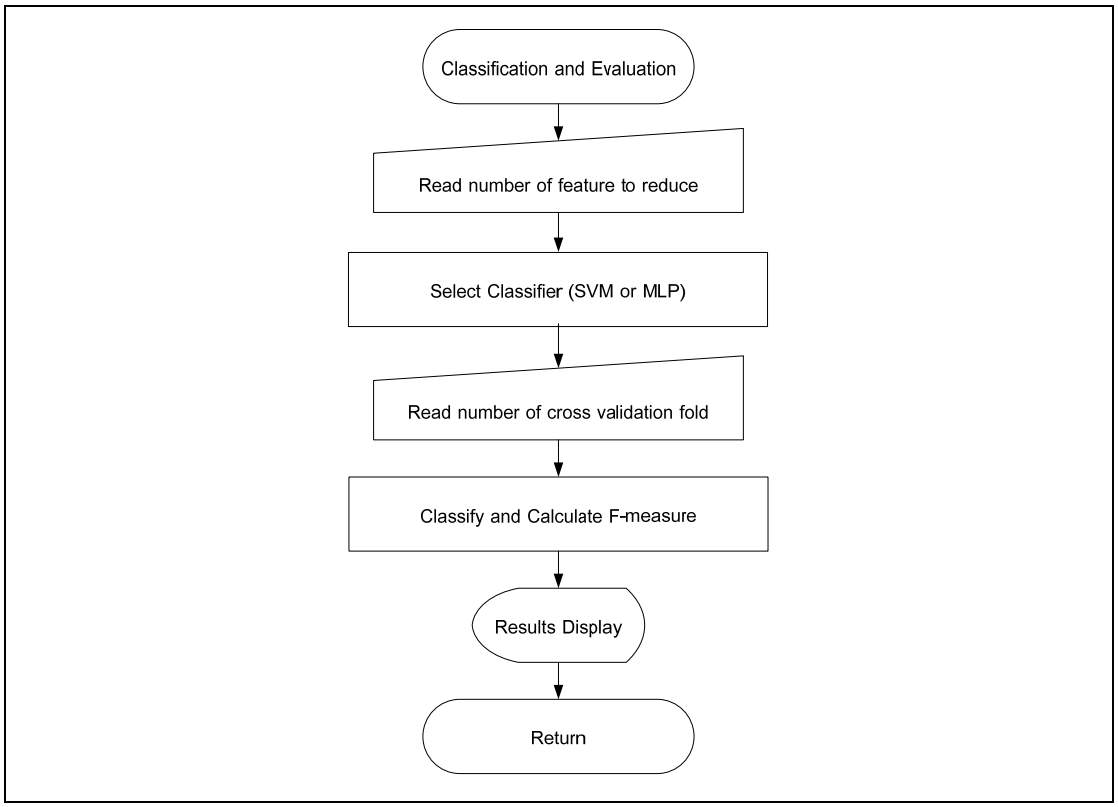
ภาพประกอบ 4.2 ผังการทำงานของโปรแกรม Step 1: Web Page Preprocessing



ภาพประกอบ 4.3 ผังการทำงานของโปรแกรม Step 2: Feature Selection using IG



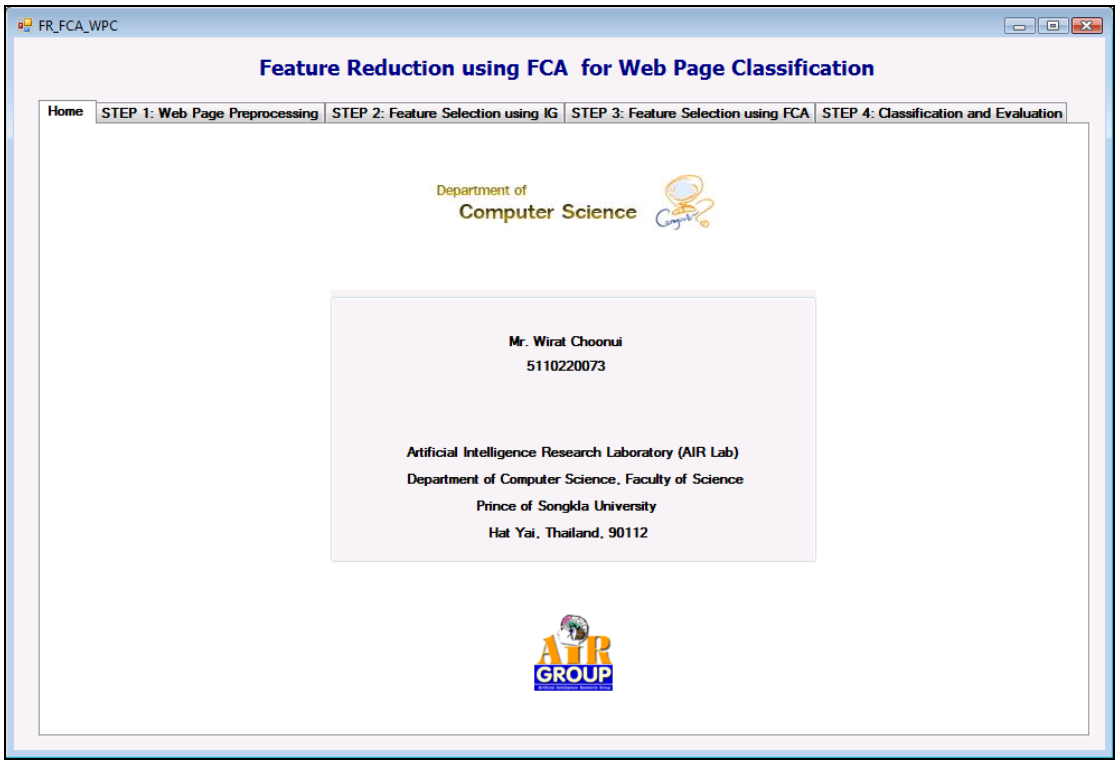
ภาพประกอบ 4.4 ผังการทำงานของโปรแกรม Step 3: Feature Selection using FCA



ภาพประกอบ 4.5 ผังการทำงานของโปรแกรม Step 4: Classification and Evaluation

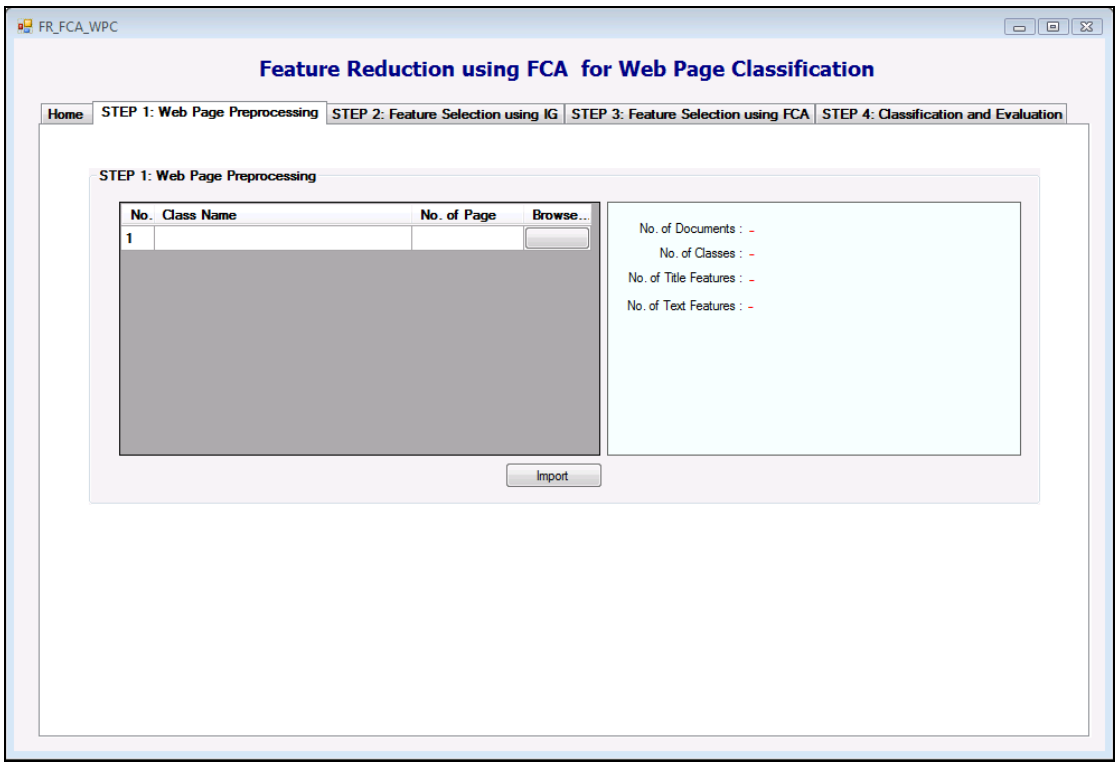
4.2 ส่วนประกอบของโปรแกรม

เมื่อเปิดโปรแกรมจะปรากฏหน้าจอหลักซึ่งประกอบด้วยขั้นตอนการทำงานแบ่งออกเป็น 4 ส่วน ดังภาพประกอบ 4.6 โดยในแต่ละส่วนมีขั้นตอนการทำงานดังต่อไปนี้

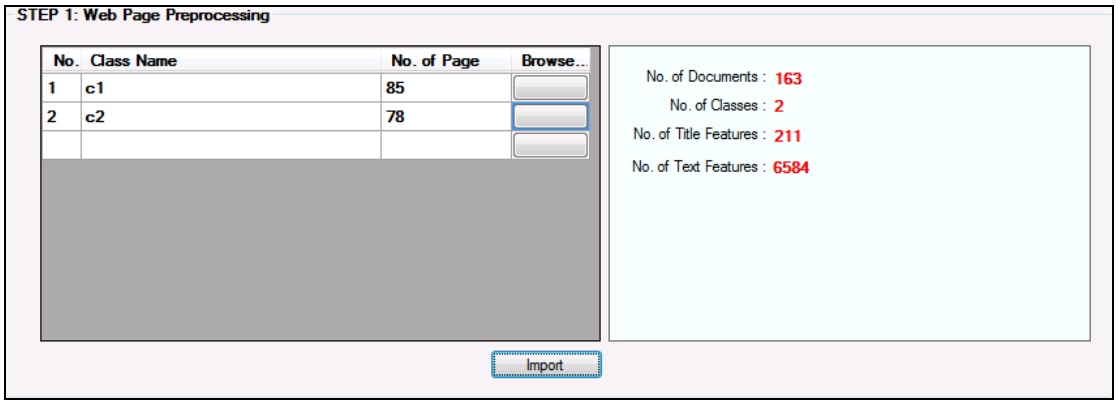


ภาพประกอบ 4.6 หน้าจอหลักของโปรแกรม

4.2.1 ส่วนการทำงานของ Step 1: Web Page Preprocessing แสดงดังภาพประกอบ 4.7 เป็นส่วนของการนำข้อมูลเข้าซึ่งผู้ใช้จะต้องเลือกชุดข้อมูลเว็บเพจโดยกดปุ่ม [ ] ในคอลัมน์ Browse เพื่อเลือกไฟล์เว็บเพ็จของแต่ละคลาสพร้อมระบุชื่อคลาสในคอลัมน์ Class Name เมื่อป้อนข้อมูลครบแล้วให้กดปุ่ม [Import] โปรแกรมจะทำการสกัดข้อความและหัวเรื่องจากหน้าเว็บเพจ หารากศัพท์ของคำและกำจัดคำหยุดโดยอัตโนมัติ และแสดงรายละเอียดผลลัพธ์เป็นจำนวนเว็บเพจ จำนวนคลาส จำนวนคำของหัวเรื่องและจำนวนคำของข้อความเนื้อหา ตัวอย่างผลลัพธ์แสดงดังภาพประกอบ 4.8



ภาพประกอบ 4.7 หน้าจอการทำงานของ Step 1: Web Page Preprocessing

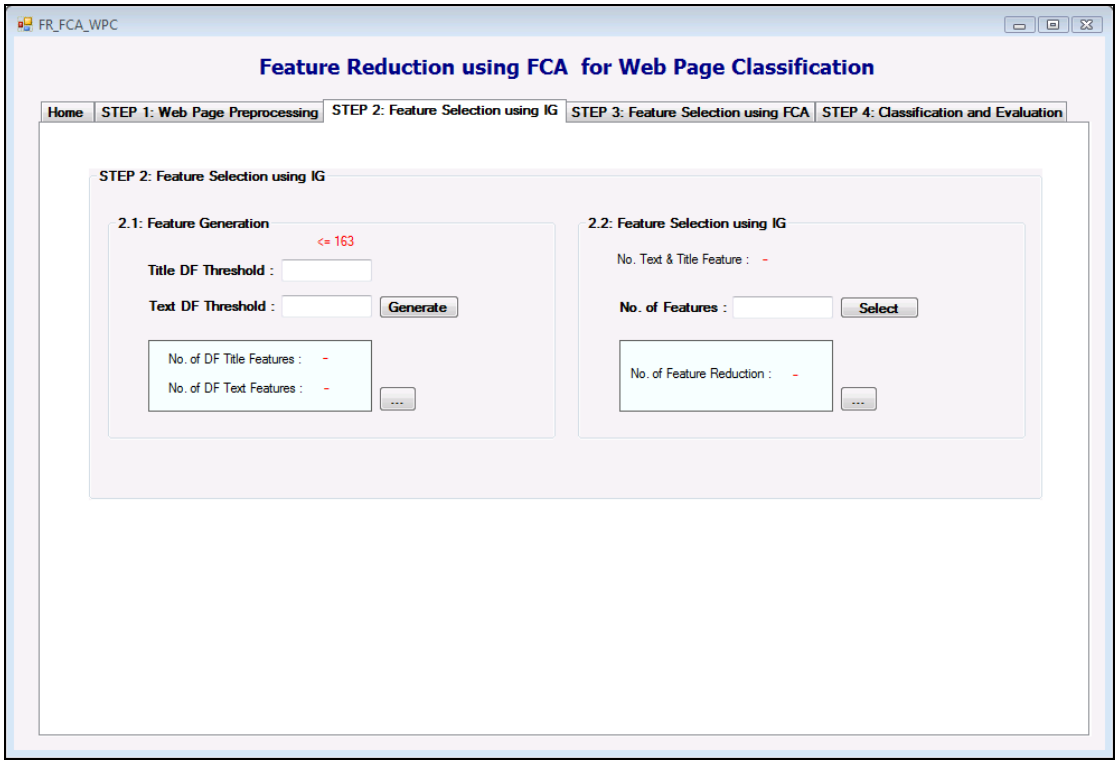


ภาพประกอบ 4.8 ตัวอย่างผลลัพธ์การทำงานของ Step 1: Web Page Preprocessing

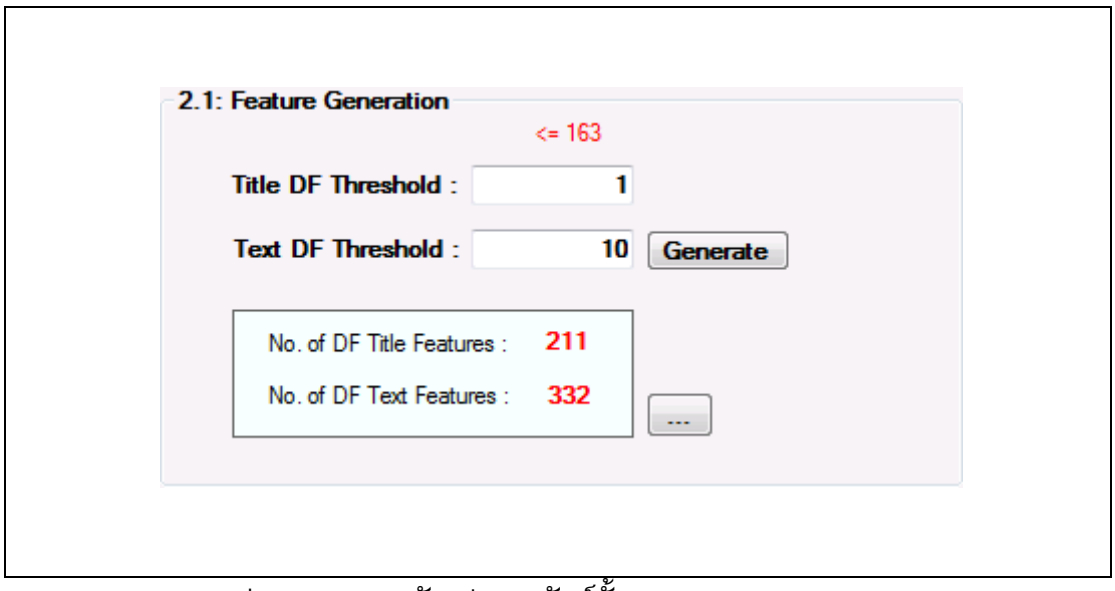
4.2.2 ส่วนการทำงานของ Step 2: Feature Selection using IG แสดงดังภาพประกอบ 4.9 เป็นขั้นตอนการเลือกลักษณะเฉพาะโดยใช้ IG ซึ่งแบ่งออกเป็น 2 ขั้นตอนย่อยดังนี้

1) Feature Generation เป็นขั้นตอนการนำคำที่ได้จากหัวเรื่องและข้อความเนื้อหามาสร้างเป็นลักษณะเฉพาะโดยจะเลือกคำที่มีค่า Document Frequency (DF) มากกว่าหรือเท่ากับค่า DF Threshold ซึ่งผู้ใช้จะต้องกำหนดค่า Title DF Threshold และ Text DF Threshold โดยป้อนค่าเป็นตัวเลขระหว่าง 0 ถึง จำนวนของเว็บเพจ

ทั้งหมด เมื่อกำหนดเสร็จแล้วกดปุ่ม Generate โปรแกรมจะทำการสร้างลักษณะเฉพาะโดยอัตโนมัติและแสดงจำนวนลักษณะเฉพาะของหัวเรื่องและจำนวนลักษณะเฉพาะของข้อความที่ได้ ตัวอย่างผลลัพธ์แสดงดังภาพประกอบ 4.10 และผู้ใช้สามารถกดปุ่ม ... เพื่อดูไฟล์ผลลัพธ์ได้

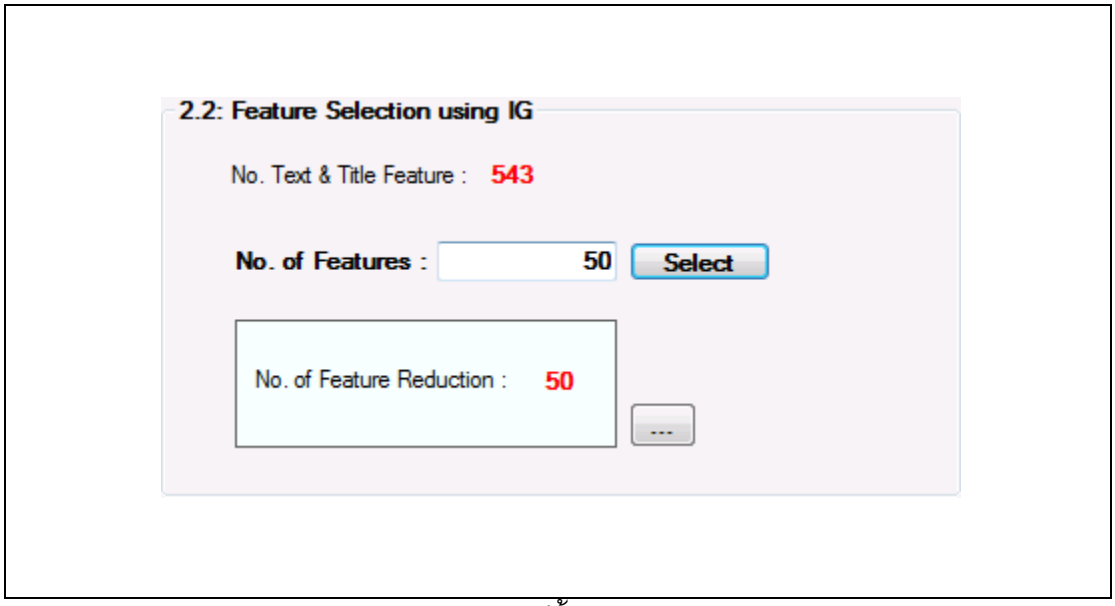


ภาพประกอบ 4.9 หน้าจอการทำงานของ Step 2: Feature Selection using IG



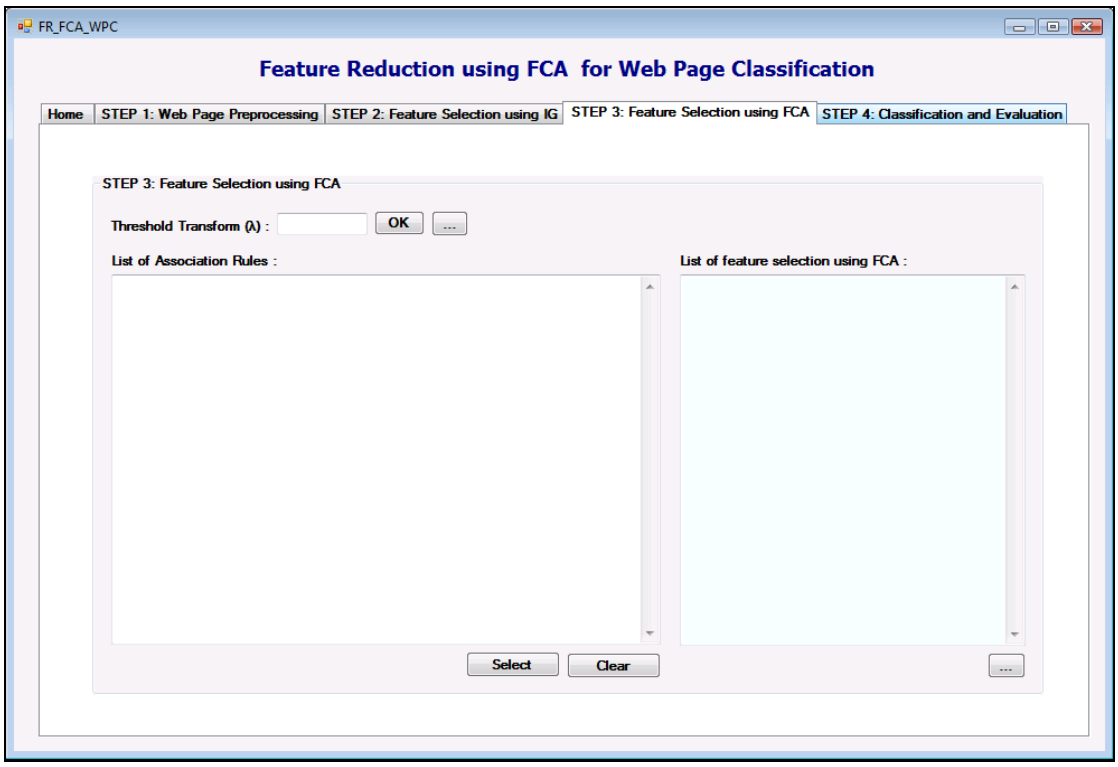
ภาพประกอบ 4.10 ตัวอย่างผลลัพธ์ขั้นตอน Feature Generation

2) Feature Selection using IG เป็นการนำลักษณะเฉพาะที่ได้ทั้งหมดมาเลือกด้วย IG โดยผู้ใช้จะต้องกำหนดจำนวนลักษณะเฉพาะที่ต้องการ (r) เช่น r = 30 เป็นต้น จากนั้นกดปุ่ม Select โปรแกรมจะทำการเลือกและแสดงจำนวนลักษณะเฉพาะโดยอัตโนมัติ ตัวอย่างผลลัพธ์แสดงดังภาพประกอบ 4.11 และผู้ใช้สามารถกดปุ่ม ... เพื่อดูไฟล์ผลลัพธ์ได้

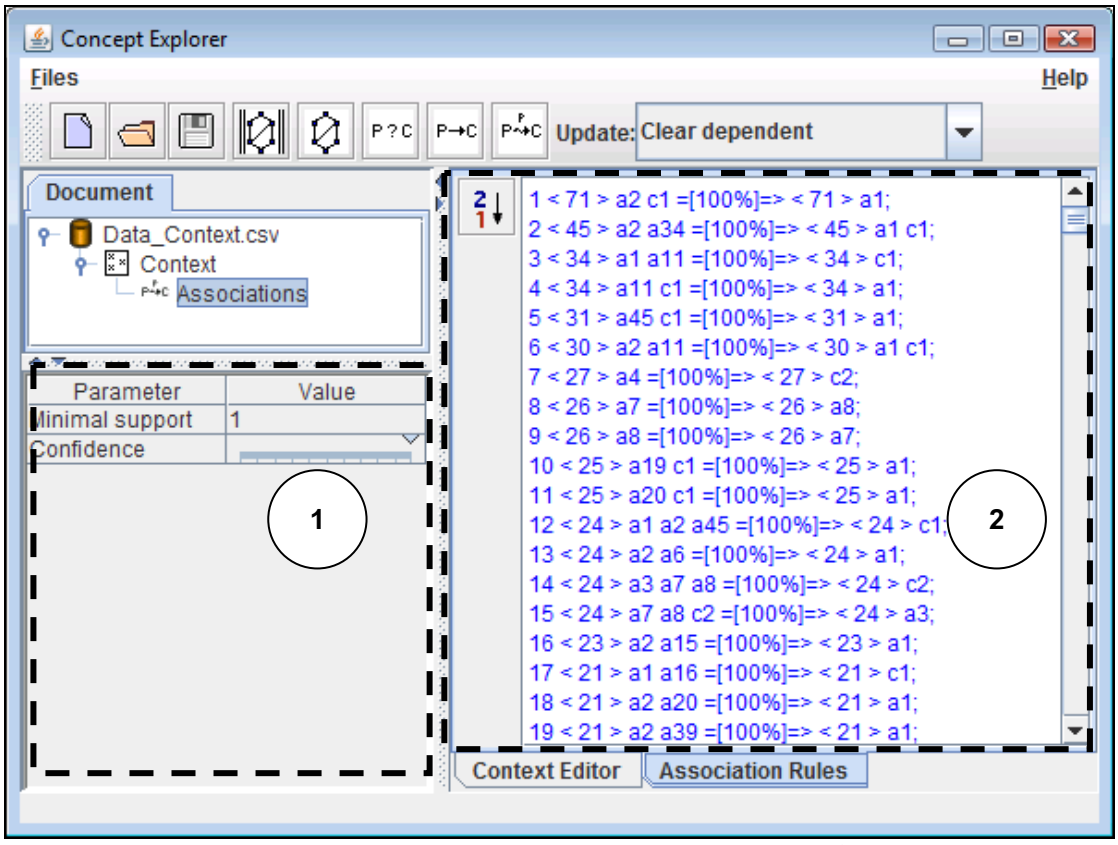


ภาพประกอบ 4.11 ตัวอย่างผลลัพธ์ขั้นตอน Feature Selection using IG

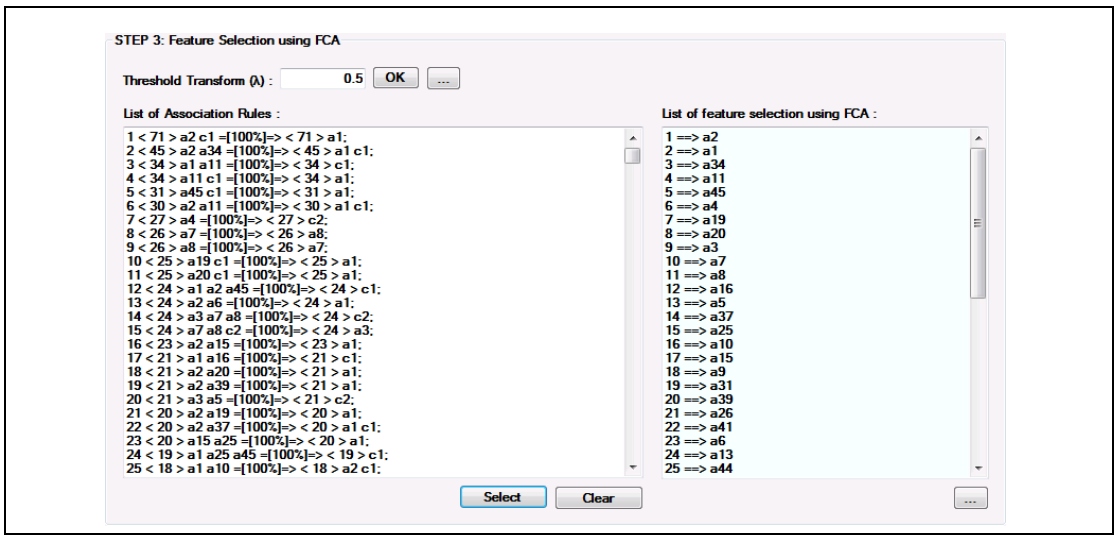
4.2.3 ส่วนการทำงานของ Step 3: Feature Selection using FCA แสดงดังภาพประกอบ 4.12 เป็นขั้นตอนการเลือกลักษณะเฉพาะโดยใช้ FCA โดยนำข้อมูลลักษณะเฉพาะที่เลือกด้วยวิธี IG แปลงให้อยู่ในรูปฟอร์มัลคอนเท็กซ์ โดยผู้ใช้จะต้องป้อนค่า Threshold Transform (λ) จากนั้นกดปุ่ม OK โปรแกรมจะทำการแปลงข้อมูลให้เป็นฟอร์มัลคอนเท็กซ์โดยอัตโนมัติ จากนั้นใช้โปรแกรม ConExp คำนวณ Association Rules จากฟอร์มัลคอนเท็กซ์ที่ได้ ดังภาพประกอบ 4.13 ซึ่งในส่วนที่ 1 ผู้ใช้จะต้องกำหนดค่า Minimal Support และค่า Confidence และในส่วนที่ 2 เรียงลำดับกฎที่ได้ตามจำนวนเอกสารจากมากไปน้อยโดยกดปุ่ม จากนั้นคัดลอกกฎทั้งหมดป้อนในกล่อง List of Association Rules แล้วกดปุ่ม Select โปรแกรมจะทำการเลือกลักษณะเฉพาะและแสดงลักษณะเฉพาะที่เลือกมาได้โดยอัตโนมัติ ตัวอย่างผลลัพธ์แสดงดังภาพประกอบ 4.14



ภาพประกอบ 4.12 หน้าจอการทำงานของ Step 3: Feature Selection using FCA

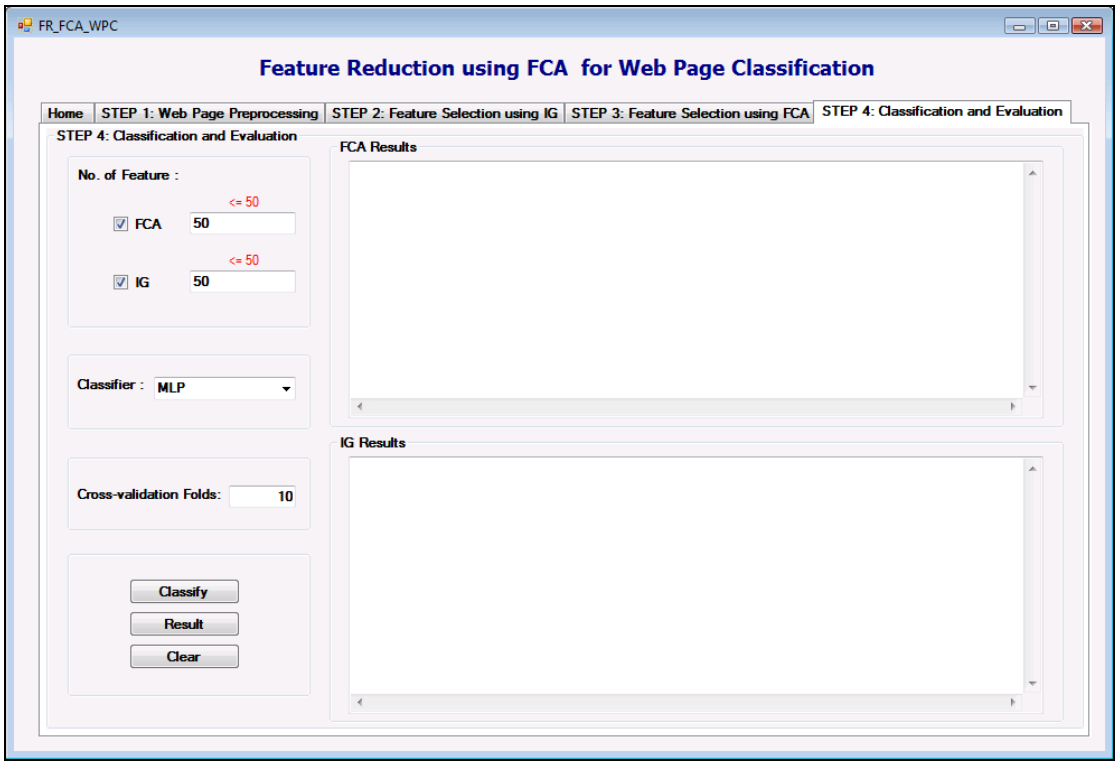


ภาพประกอบ 4.13 หน้าจอการคำนวณหา Association Rules ด้วยโปรแกรม ConExp



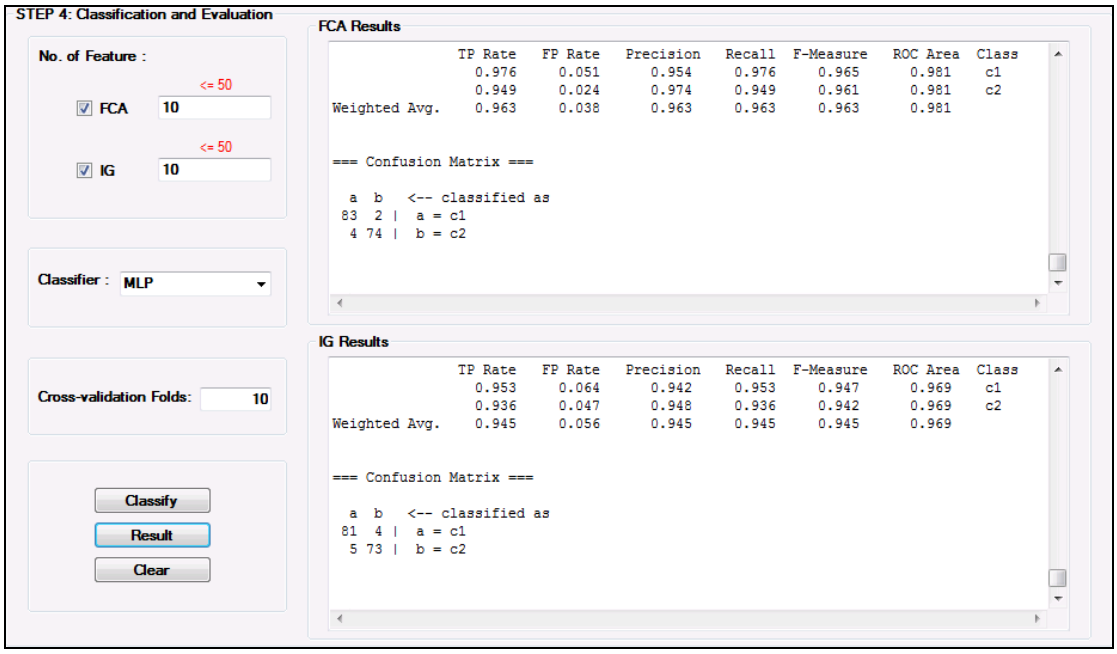
ภาพประกอบ 4.14 ตัวอย่างผลลัพธ์การทำงานของ Step 3: Feature Selection using FCA

4.2.4 ส่วนการทำงานของ Step 4: Classification and Evaluation แสดงดังภาพประกอบ 4.15 เป็นขั้นตอนการจำแนกประเภทและการประเมินผลโดยผู้ใช้จะต้องกำหนดจำนวนลักษณะเฉพาะนำเข้าที่ต้องการ กำหนดจำนวน Fold ของการทดสอบแบบ Cross Validation และเลือกตัวจำแนกประเภท (Classifier) จากนั้นกดปุ่ม Classify โปรแกรมจะทำการจำแนกประเภทโดยอัตโนมัติ



ภาพประกอบ 4.15 หน้าจอการทำงานของ Step 4: Classification and Evaluation

ผู้ใช้จะต้องกดปุ่ม Result เพื่อแสดงผลลัพธ์การจำแนกประเภท ตัวอย่างผลลัพธ์แสดงดังภาพประกอบ 4.16



ภาพประกอบ 4.16 ตัวอย่างผลลัพธ์การทำงานของ Step 4: Classification and Evaluation

บทที่ 5

ผลการทดลองและวิจารณ์

บทนี้ได้นำเสนอผลลัพธ์ที่ได้จากการทดลองตามแบบจำลองการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ (Feature Reduction using FCA for Web Page Classification: FR_FCA_WPC) การทดลองใช้ชุดข้อมูลเว็บเพจมาตรฐานจาก CMU และประเมินผลการจำแนกประเภทด้วยค่า F-measure

5.1 ชุดข้อมูลเว็บเพจ

5.1.1 ชุดข้อมูล 7Sectors

ชุดข้อมูล 7Sectors เป็นชุดข้อมูลเว็บเพจมาตรฐานจาก CMU (WebKB, 2012) ซึ่งมีจำนวน 4,581 เว็บเพจ ประกอบด้วย 7 กลุ่ม ได้แก่ 1) Materials จำนวน 949 เว็บเพจ 2) Energy จำนวน 355 เว็บเพจ 3) Financial จำนวน 964 เว็บเพจ 4) Healthcare จำนวน 399 เว็บเพจ 5) Technology จำนวน 1,099 เว็บเพจ 6) Transportation จำนวน 515 เว็บเพจ และ 7) Utilities จำนวน 300 เว็บเพจ แสดงรายละเอียดดังตารางที่ 5.1

ตารางที่ 5.1 ชุดข้อมูล 7Sectors

ข้อมูล	จำนวนเว็บเพจ
1. basic.materials.sector	
1.1 chemical.manufacturing.industry	100
1.2 chemicals.plastics.and.rubber.industry	83
1.3 containers.and.packaging.industry	100
1.4 fabricated.plastic.and.rubber.industry	90
1.5 forestry.and.wood.products.industry	77
1.6 gold.and.silver.industry	100
1.7 iron.and.steel.industry	100
1.8 metal.and.mining.industry	100
1. 9 misc.fabricated.products.industry	100
1.10 non.metallic.mining.industry	0
1.11 paper.and.paper.products.industry	99

ตารางที่ 5.1 ชุดข้อมูล 7Sectors (ต่อ)

ข้อมูล	จำนวนเว็บเพจ
2. energy.sector	
2.1 coal.industry	50
2.2 oil.and.gas.integrated.industry	101
2.3 oil.and.gas.operations.industry	102
2.4 oil.well.services.and.equipment.industry	102
3. financial.sector	
3.1 misc.financial.services.industry	90
3.2 investment.services.industry	100
3.3 insurance.sector	
3.3.1 accident.and.health.insurance.industry	85
3.3.2 life.insurance.industry	100
3.3.3 misc.insurance.industry	100
3.3.4 property.and.casualty.insurance.industry	100
3.4 consumer.financial.services.industry	100
3.5 banking.sector	
3.5.1 money.center.banks.industry	89
3.5.2 regional.banks.industry	100
3.5.3 s.and.ls.savings.banks.industry	100
4. healthcare.sector	
4.1 biotechnology.and.drugs.industry	100
4.2 healthcare.facilities.industry	99
4.3 major.drugs.industry	100
4.4 medical.equipment.and.supplies.industry	100
5. technology.sector	
5.1 communications.equipment.industry	100
5.2 computer.sector	
5.2.1 computer.hardware.industry	100
5.2.2 computer.networks.industry	100
5.2.3 computer.peripherals.industry	100
5.2.4 computer.services.industry	100
5.2.5 computer.storage.devices.industry	100
5.2.6 software.and.programming.industry	99
5.3 electronic.instruments.and.controls.industry	100
5.4 office.equipment.industry	100
5.5 scientific.and.technical.instruments.industry	100
5.6 semiconductors.industry	100

ตารางที่ 5.1 ชุดข้อมูล 7Sectors (ต่อ)

ข้อมูล	จำนวนเว็บเพจ
6. transportation.sector	
6.1 air.courier.industry	99
6.2 airline.industry	105
6.3 misc.transportation.industry	80
6.4 railroad.industry	95
6.5 trucking.industry	97
6.6 water.transportation.industry	39
7. utilities.sector	
7.1 electric.utilities.industry	100
7.2 natural.gas.industry	100
7.3 water.utilities.industry	100

5.1.2 ชุดข้อมูล BankResearch

ชุดข้อมูล BankResearch เป็นชุดข้อมูลเว็บเพจ StatLib Datasets Archive จาก CMU (StatLib, 2012) ซึ่งมีจำนวนเว็บเพจทั้งหมด 11,000 เว็บเพจ แบ่งออกเป็น 11 คลาส ได้แก่ 1) Commercial Banks 2) Building Societies 3) Insurance Agencies 4) Java 5) C 6) Visual Basic 7) Astronomy 8) Biology 9) Soccer 10) Motor Racing และ 11) Sport โดยแต่ละคลาสมีจำนวนเว็บเพจ 1,000 เว็บเพจ

5.1.3 การแบ่งชุดข้อมูลเว็บเพจสำหรับการทดลอง

1) ชุดข้อมูล A เป็นชุดข้อมูล 7Sectors จาก CMU โดยเลือกมาจำนวน 2 คลาส ได้แก่ 1) containers and packaging industry จำนวน 85 เว็บเพจ และ 2) oil and gas integrated industry จำนวน 78 เว็บเพจ รวมเป็นจำนวน 163 เว็บเพจ (เลือกเฉพาะส่วนที่เป็นไฟล์ html และ htm)

2) ชุดข้อมูล B เป็นชุดข้อมูล 7Sectors จาก CMU โดยเลือกมาจำนวน 2 คลาส ได้แก่ 1) regional banks industry จำนวน 72 เว็บเพจ และ 2) healthcare facilities industry จำนวน 79 เว็บเพจ รวมเป็นจำนวน 151 เว็บเพจ (เลือกเฉพาะส่วนที่เป็นไฟล์ html และ htm)

3) ชุดข้อมูล C เป็นชุดข้อมูล 7Sectors จาก CMU โดยสุ่มจาก 2 กลุ่ม แบ่งเป็น 2 คลาส ได้แก่ 1) healthcare sector จำนวน 100 เว็บเพจ และ 2) transportation sector จำนวน 100 เว็บเพจ รวมเป็นจำนวน 200 เว็บเพจ

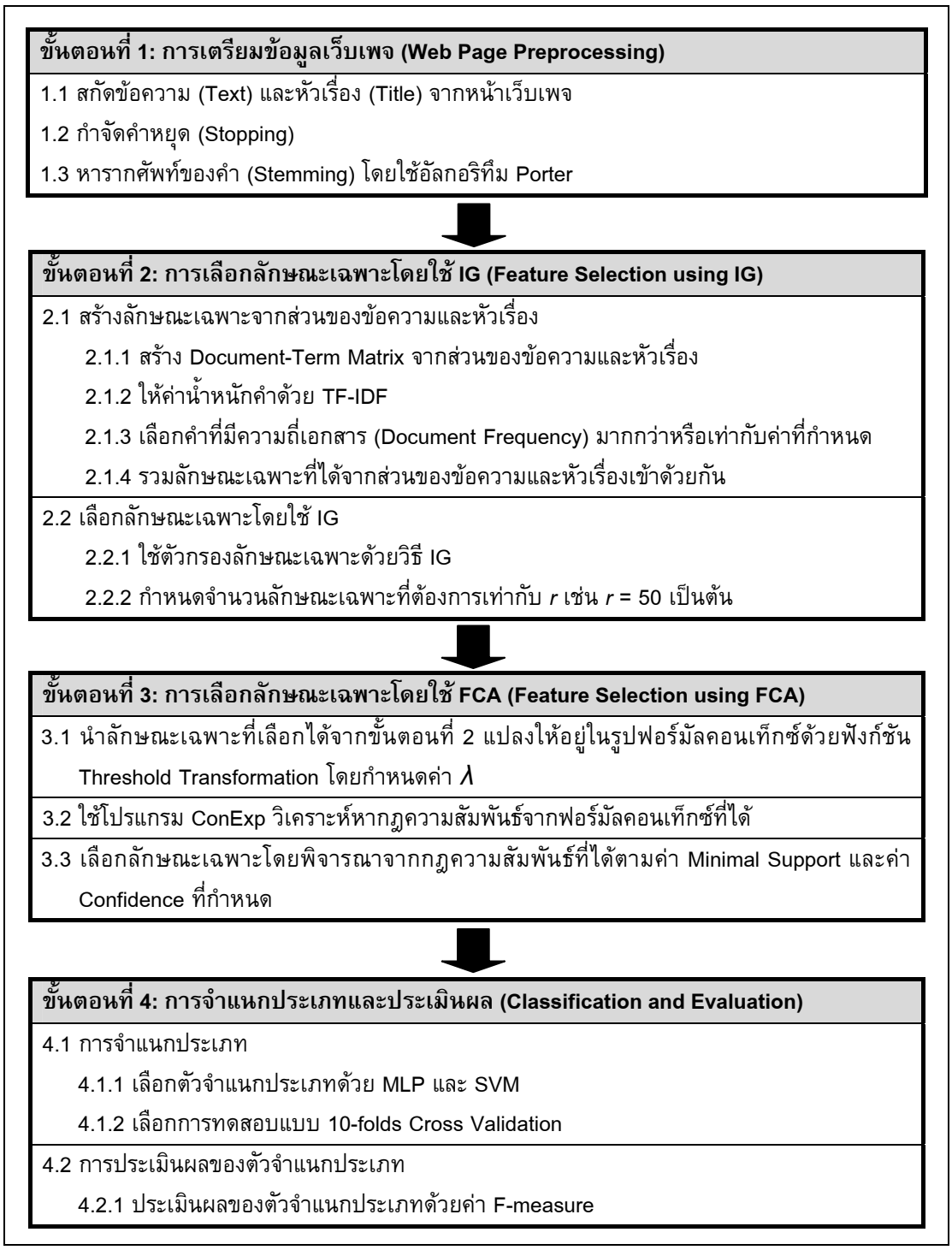
4) ชุดข้อมูล D เป็นชุดข้อมูล 7Sectors จาก CMU โดยสุ่มจาก 3 กลุ่ม แบ่งเป็น 3 คลาส ได้แก่ 1) financial sector จำนวน 30 เว็บเพจ 2) healthcare sector จำนวน 30 เว็บเพจ และ 3) transportation sector จำนวน 30 เว็บเพจ รวมเป็นจำนวน 90 เว็บเพจ

5) ชุดข้อมูล E เป็นชุดข้อมูล BankResearch ของ StatLib Datasets Archive จาก CMU โดยสุ่มเอกสารเว็บเพจมาเป็นตัวแทนแบ่งเป็น 2 คลาส ได้แก่ 1) CommercialBanks จำนวน 50 เว็บเพจ และ 2) Sport จำนวน 50 เว็บเพจ รวมเป็นจำนวน 100 เว็บเพจ

6) ชุดข้อมูล F เป็นชุดข้อมูล BankResearch ของ StatLib Datasets Archive จาก CMU โดยสุ่มเอกสารเว็บเพจมาเป็นตัวแทนแบ่งเป็น 2 คลาส ได้แก่ 1) Building จำนวน 50 เว็บเพจ และ 2) Insurance จำนวน 50 เว็บเพจ รวมเป็นจำนวน 100 เว็บเพจ

5.2 การทดลองการลดขนาดลักษณะเฉพาะโดยใช้ FCA

การทดลองได้ทำตามแบบจำลองการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ (Feature Reduction using FCA for Web Page Classification: FR_FCA_WPC) ดังภาพประกอบ 3.1 (บทที่ 3) ซึ่งแบ่งการทำงานออกเป็น 4 ขั้นตอนหลักประกอบด้วย ขั้นตอนที่ 1 คือการเตรียมข้อมูลเว็บเพจ ขั้นตอนที่ 2 คือการเลือกลักษณะเฉพาะโดยใช้ IG ขั้นตอนที่ 3 คือการเลือกลักษณะเฉพาะโดยใช้ FCA และขั้นตอนที่ 4 คือการจำแนกประเภทและการประเมินผล การทดลองได้ใช้ชุดข้อมูล 7Sectors คือชุดข้อมูล A B C และ D และชุดข้อมูล BankResearch คือชุดข้อมูล E และ F โดยทำการทดลองตามขั้นตอนดังภาพประกอบ 5.1



ภาพประกอบ 5.1 ขั้นตอนการทดลองการลดขนาดลักษณะเฉพาะโดยใช้ FCA

5.2.1 การทดลองชุดข้อมูล A

ขั้นตอนที่ 1: การเตรียมข้อมูลเว็บเพจ

1.1 นำเอกสารเว็บเพจทั้ง 163 เว็บเพจมาสกัดข้อความจากหน้าเว็บเพจ โดยกำจัดแท็ก HTML รูปภาพ สื่อมัลติมีเดีย ออกให้เหลือเฉพาะข้อความและหัวเรื่องเท่านั้น

1.2 นำคำจากส่วนของข้อความและส่วนของหัวเรื่องที่ได้มากำจัดคำหยุด (Stopping) ซึ่งจะพิจารณาจากรายการคำหยุด (Stoplist)

1.3 นำคำที่ได้จากขั้นตอนที่ 1.2 มาหารากศัพท์ของคำ

จากชุดข้อมูล A ประกอบด้วยเอกสารเว็บเพจ 2 คลาส จำนวน 163 เว็บเพจ หลังจากผ่านขั้นตอนการเตรียมข้อมูลเว็บเพจแล้วจะได้คำจากข้อความจำนวน 6,584 คำ และจากหัวเรื่องจำนวน 211 คำ

ขั้นตอนที่ 2: การเลือกลักษณะเฉพาะโดยใช้ IG

2.1 นำคำจากส่วนของข้อความและส่วนของหัวเรื่องที่ได้จากขั้นตอนการเตรียมข้อมูลมาสร้างลักษณะเฉพาะในรูปของ Document-Term Matrix โดยให้ค่าน้ำหนักของคำด้วยวิธีการ TF-IDF จากนั้นเลือกคำที่มีค่าความถี่เอกสาร (Document Frequency: DF) มากกว่าหรือเท่ากับค่าที่กำหนด (Threshold) ดังนั้นคำที่มีค่าน้อยกว่าค่าที่กำหนดจะถูกกำจัดออก ซึ่งจากชุดข้อมูล A เมื่อเลือกคำด้วยค่า DF Threshold ของหัวเรื่องเท่ากับ 1 และของข้อความเท่ากับ 10 จะได้ลักษณะเฉพาะจากหัวเรื่องจำนวน 211 ลักษณะเฉพาะ และได้ลักษณะเฉพาะจากข้อความจำนวน 332 ลักษณะเฉพาะ รวมลักษณะเฉพาะที่ได้จากหัวเรื่องและข้อความเข้าด้วยกันเป็นจำนวน 543 ลักษณะเฉพาะ

2.2 นำลักษณะเฉพาะที่ได้มากรองเพื่อเลือกลักษณะเฉพาะด้วยวิธี Information Gain (IG) ในชุดข้อมูลนี้ได้เลือกจำนวนลักษณะเฉพาะที่ต้องการ (r) เท่ากับ 50 ดังนั้นจากลักษณะเฉพาะจำนวน 543 ลักษณะเฉพาะ จะถูกกรองเหลือเพียง 50 ลักษณะเฉพาะ เพื่อใช้ในขั้นตอนต่อไป

ขั้นตอนที่ 3: การเลือกลักษณะเฉพาะโดยใช้ FCA

3.1 นำข้อมูลลักษณะเฉพาะที่เลือกจากขั้นตอนที่ 2 มาสร้างเป็นฟอร์มัลคอนเท็กซ์ โดยในชุดข้อมูลนี้ได้ใช้พารามิเตอร์ $\lambda = 0$ $\lambda = 0.5$ $\lambda = 1.0$ $\lambda = 1.5$ และ $\lambda = 2.0$ สำหรับแปลงข้อมูลลักษณะเฉพาะให้อยู่ในรูปฟอร์มัลคอนเท็กซ์

3.2 ใช้โปรแกรม ConExp วิเคราะห์หากฎความสัมพันธ์จากฟอร์มัลคอนเท็กซ์ที่ได้ ซึ่งกำหนดค่า Minimal Support เท่ากับ 1 และค่า Confidence เท่ากับ 100% และเรียงกฎตามค่า Support จากมากไปน้อย

3.3 เลือกลักษณะเฉพาะโดยสกัดจากกฎความสัมพันธ์ที่ได้

จากชุดข้อมูล A จะได้ผลลัพธ์จากขั้นตอนการเลือกลักษณะเฉพาะโดยใช้ FCA แสดงดังตารางที่ 5.2 เมื่อปรับค่า λ เพิ่มขึ้นจะทำให้ความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารเว็บเพจบางตัวถูกตัดออกไป จึงทำให้จำนวนคอนเซ็ปต์ จำนวนกฎความสัมพันธ์ และจำนวนลักษณะเฉพาะ (*s*) ที่ได้มีแนวโน้มลดลง

ตารางที่ 5.2 ผลการเลือกลักษณะเฉพาะโดยใช้ FCA ชุดข้อมูล A

พารามิเตอร์	ชุดข้อมูล A		
	จำนวนคอนเซ็ปต์	จำนวนกฎความสัมพันธ์	จำนวนลักษณะเฉพาะ
$\lambda = 0$	1,788	1,090	50
$\lambda = 0.5$	1,928	1,298	50
$\lambda = 1.0$	1,612	1,030	47
$\lambda = 1.5$	957	893	47
$\lambda = 2.0$	901	825	47

ขั้นตอนที่ 4: การจำแนกประเภทและประเมินผล

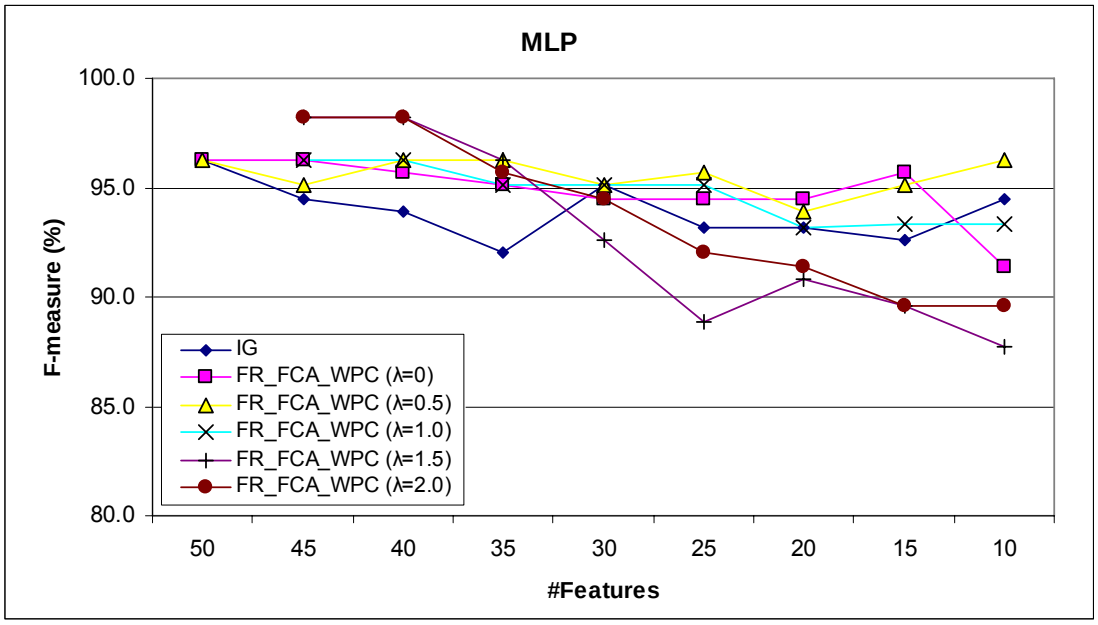
4.1 นำลักษณะเฉพาะที่ได้จากขั้นตอนที่ 3 มาลดขนาดลักษณะเฉพาะลงโดยตัดลักษณะเฉพาะจากอันดับหลังสุดออกทีละ 5 ลักษณะเฉพาะ ซึ่งจะได้ขนาดลักษณะเฉพาะนำเข้าเป็น 50 45 40 35 30 25 20 15 และ 10 ตามลำดับ (ข้อสังเกต หากจำนวนลักษณะเฉพาะที่เลือกด้วย FCA มีจำนวนเท่ากับ 47 จะตัดลักษณะโดยเริ่มต้นเป็น 45) แล้วนำไปจำแนกประเภทด้วย MLP และ SVM ทดสอบแบบ 10-folds Cross Validation

4.2 ประเมินผลด้วยค่า F-measure จากการเลือกลักษณะเฉพาะโดยใช้ FCA เปรียบเทียบกับการเลือกลักษณะเฉพาะโดยใช้ IG

จากชุดข้อมูล A จะได้ผลการทดลองดังตารางที่ 5.3 และ 5.4 และแสดงกราฟเปรียบเทียบดังภาพประกอบ 5.2 และ 5.3 ตามลำดับ (ข้อสังเกต จากตารางผลการทดลองเมื่อเปรียบค่า F-measure ที่ขนาดลักษณะเฉพาะนำเข้าเท่ากันระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC และวิธี IG **ตัวอักษรหนา** หมายถึงวิธีนั้นให้ค่า F-measure สูงกว่าหรือเท่ากันกับอีกวิธี)

ตารางที่ 5.3 ผลการจำแนกประเภทด้วย MLP ของชุดข้อมูล A

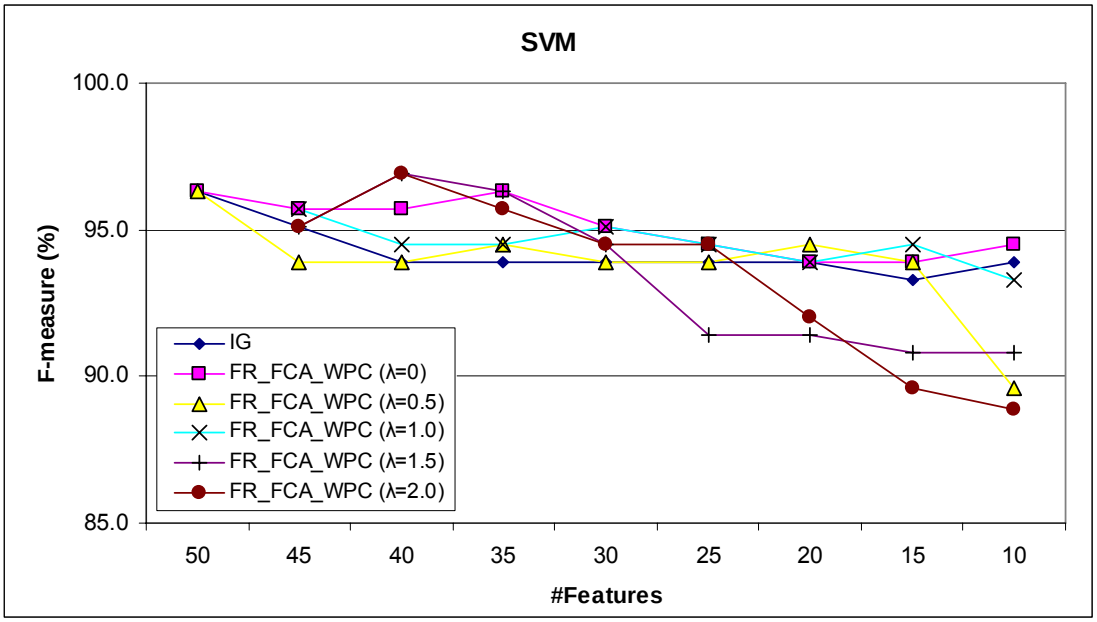
จำนวนลักษณะเฉพาะ (#Features)	ค่า F-measure (%) จำแนกประเภทด้วย MLP ชุดข้อมูล A					
	IG	FR_FCA_WPC				
		$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$
50	***96.3***	**96.3**	**96.3**	-	-	-
45	94.5	**96.3**	**95.1**	**96.3**	***98.2***	***98.2***
40	93.9	**95.7**	**96.3**	**96.3**	***98.2***	***98.2***
35	92.0	**95.1**	**96.3**	**95.1**	**96.3**	**95.7**
30	**95.1**	94.5	**95.1**	**95.1**	92.6	94.5
25	93.2	**94.5**	**95.7**	**95.1**	88.9	92.0
20	93.2	**94.5**	**93.9**	**93.2**	90.8	91.4
15	92.6	**95.7**	**95.1**	**93.3**	89.6	89.6
10	94.5	91.4	**96.3**	93.3	87.7	89.6



ภาพประกอบ 5.2 เปรียบเทียบผลการจำแนกประเภทด้วย MLP ของชุดข้อมูล A

ตารางที่ 5.4 ผลการจำแนกประเภทด้วย SVM ของชุดข้อมูล A

จำนวนลักษณะเฉพาะ (#Features)	ค่า F-measure (%) จำแนกประเภทด้วย SVM ชุดข้อมูล A					
	IG	FR_FCA_WPC				
		$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$
50	***96.3***	**96.3**	**96.3**	-	-	-
45	95.1	**95.7**	93.9	**95.7**	**95.1**	**95.1**
40	93.9	**95.7**	**93.9**	**94.5**	***96.9***	***96.9***
35	93.9	**96.3**	**94.5**	**94.5**	**96.3**	**95.7**
30	93.9	**95.1**	**93.9**	**95.1**	**94.5**	**94.5**
25	93.9	**94.5**	**93.9**	**94.5**	91.4	**94.5**
20	93.9	**93.9**	**94.5**	**93.9**	91.4	92.0
15	93.3	**93.9**	**93.9**	**94.5**	90.8	89.6
10	93.9	**94.5**	89.6	93.3	90.8	88.9



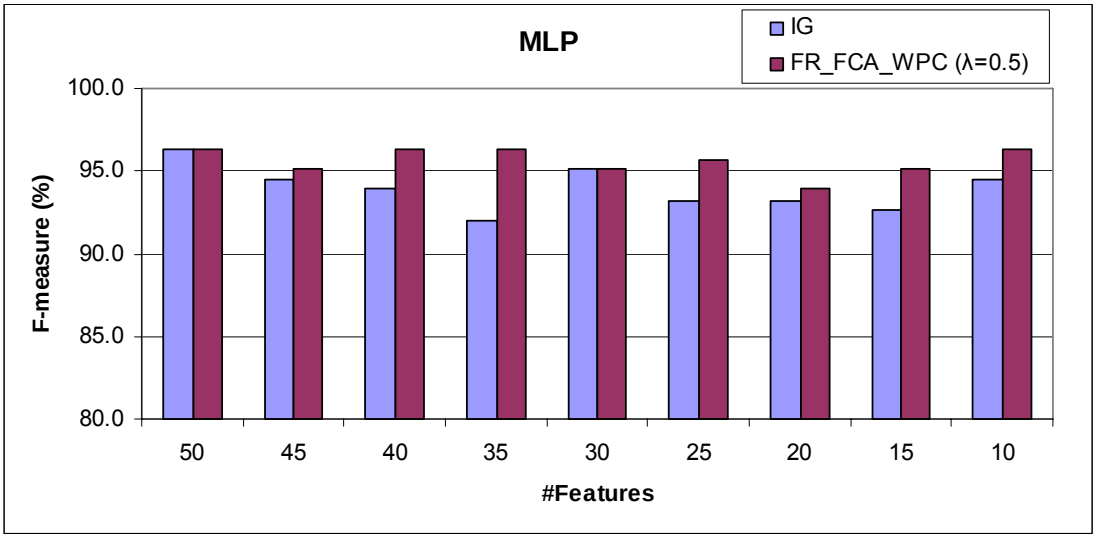
ภาพประกอบ 5.3 เปรียบเทียบผลการจำแนกประเภทด้วย SVM ของชุดข้อมูล A

จากผลการทดลองของชุดข้อมูล A สามารถอธิบายได้ 2 ประเด็นดังนี้

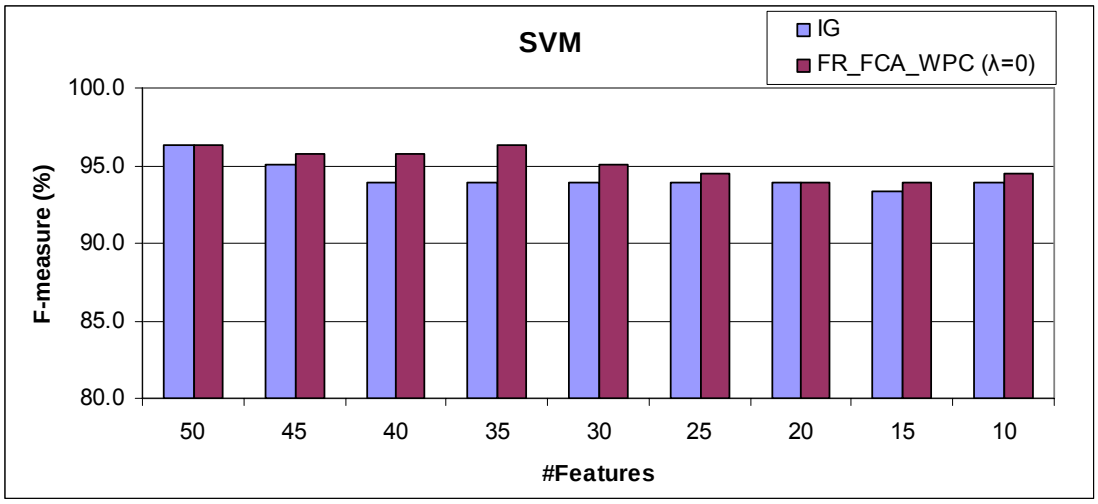
1) ประเด็นเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC กับวิธี IG

ผลการจำแนกประเภทด้วย MLP จากตารางที่ 5.3 และภาพประกอบ 5.2 จะเห็นว่าที่ขนาดลักษณะเฉพาะเท่ากับ 45 และ 40 วิธี FR_FCA_WPC ($\lambda = 1.5$ และ $\lambda = 2.0$) ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 98.2% ในขณะที่วิธี IG ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 96.3% ที่ขนาดลักษณะเฉพาะเท่ากับ 50 และจากภาพประกอบ 5.4 เมื่อเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0.5$) กับวิธี IG จะเห็นว่าวิธี FR_FCA_WPC ($\lambda = 0.5$) ส่วนใหญ่ให้ค่า F-measure ที่สูงกว่าวิธี IG

ผลการจำแนกประเภทด้วย SVM จากตารางที่ 5.4 และภาพประกอบ 5.3 จะเห็นว่าที่ขนาดลักษณะเฉพาะเท่ากับ 40 วิธี FR_FCA_WPC ($\lambda = 1.5$ และ $\lambda = 2.0$) ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 96.9% ในขณะที่วิธี IG ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 96.3% ที่ขนาดลักษณะเฉพาะเท่ากับ 50 และจากภาพประกอบ 5.5 เมื่อเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จะเห็นว่าวิธี FR_FCA_WPC ($\lambda = 0$) ส่วนใหญ่ให้ค่า F-measure ที่สูงกว่าวิธี IG



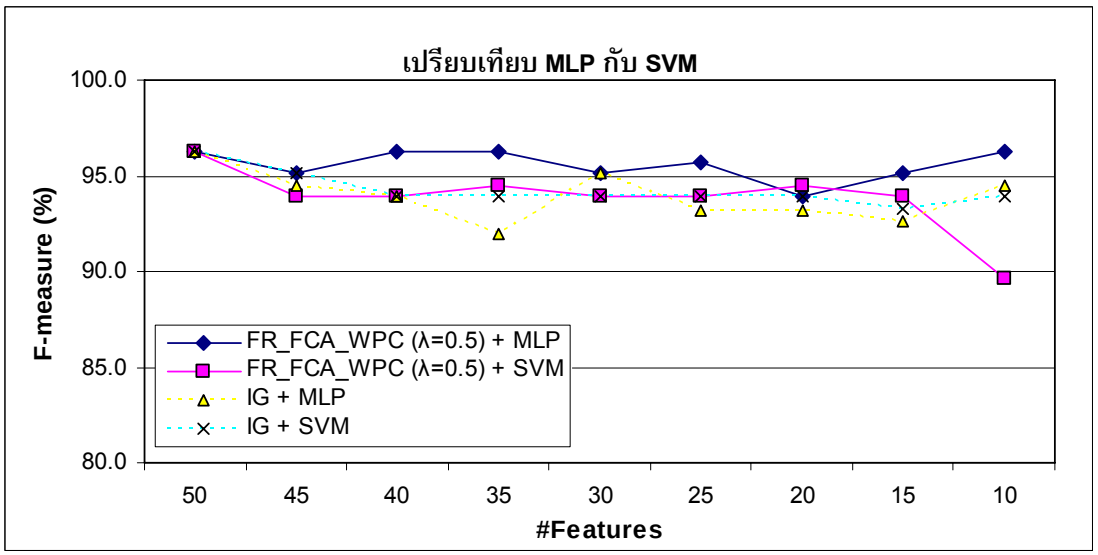
ภาพประกอบ 5.4 เปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0.5$) กับวิธี IG จำแนกประเภทด้วย MLP ของชุดข้อมูล A



ภาพประกอบ 5.5 เปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จำแนกประเภทด้วย SVM ของชุดข้อมูล A

2) ประเด็นเปรียบเทียบตัวจำแนกประเภท

จากภาพประกอบ 5.6 เมื่อเปรียบเทียบค่า F-measure จำแนกประเภทด้วย MLP และ SVM ลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0.5$) และวิธี IG พบว่าเมื่อลดขนาดลักษณะเฉพาะลง MLP ให้ค่า F-measure ที่สูงกว่า SVM เช่นที่ขนาดลักษณะเฉพาะเท่ากับ 10 วิธี FR_FCA_WPC ($\lambda = 0.5$) ที่จำแนกประเภทด้วย MLP ให้ค่า F-measure สูงกว่าที่จำแนกด้วย SVM



ภาพประกอบ 5.6 เปรียบเทียบค่า F-measure จำแนกประเภทด้วย MLP และ SVM ลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0.5$) และวิธี IG ของชุดข้อมูล A

5.2.2 การทดลองชุดข้อมูล B

ขั้นตอนที่ 1: การเตรียมข้อมูลเว็บเพจ

จากชุดข้อมูล B ประกอบด้วยเอกสารเว็บเพจ 2 คลาส จำนวน 151 เว็บเพจ หลังจากผ่านขั้นตอนการเตรียมข้อมูลเว็บเพจแล้วจะได้คำจากข้อความจำนวน 9,608 คำ และจากหัวเรื่องจำนวน 202 คำ

ขั้นตอนที่ 2: การเลือกลักษณะเฉพาะโดยใช้ IG

นำคำที่ได้จากข้อความและหัวเรื่องมาเลือกค่าความถี่เอกสารด้วยค่า DF Threshold ของหัวเรื่องเท่ากับ 1 และของข้อความเท่ากับ 10 จะได้ลักษณะเฉพาะจากหัวเรื่องจำนวน 202 ลักษณะเฉพาะ และได้ลักษณะเฉพาะจากข้อความจำนวน 404 ลักษณะเฉพาะ รวมลักษณะเฉพาะที่ได้จากหัวเรื่องและข้อความเข้าด้วยกันเป็นจำนวน 606 ลักษณะเฉพาะ

จากนั้นนำลักษณะเฉพาะที่ได้มากรองเพื่อเลือกลักษณะเฉพาะด้วยวิธี IG โดยในชุดข้อมูลนี้ได้เลือกจำนวนลักษณะเฉพาะเท่ากับ 50 เพื่อใช้ในขั้นตอนต่อไป

ขั้นตอนที่ 3: การเลือกลักษณะเฉพาะโดยใช้ FCA

นำข้อมูลลักษณะเฉพาะที่เลือกจากขั้นตอนที่ 2 มาสร้างเป็นฟอร์มัลคอนเท็กซ์ โดยในชุดข้อมูลนี้ได้ใช้พารามิเตอร์ $\lambda = 0$ $\lambda = 0.5$ $\lambda = 1.0$ $\lambda = 1.5$ และ $\lambda = 2.0$ แปลงข้อมูลลักษณะเฉพาะให้อยู่ในรูปฟอร์มัลคอนเท็กซ์ จากนั้นใช้โปรแกรม ConExp วิเคราะห์หากฎความสัมพันธ์จากฟอร์มัลคอนเท็กซ์ที่ได้ ซึ่งกำหนดค่า Minimal Support เท่ากับ 1 และค่า Confidence เท่ากับ 100% และเรียงกฎตามค่า Support จากมากไปน้อย และเลือกลักษณะเฉพาะโดยสกัดจากกฎความสัมพันธ์ที่ได้

จากชุดข้อมูล B จะได้ผลลัพธ์จากขั้นตอนการเลือกลักษณะเฉพาะโดยใช้ FCA แสดงดังตารางที่ 5.5 เมื่อปรับค่า λ เพิ่มขึ้นจะทำให้ความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารเว็บเพจบางตัวถูกตัดออกไป จึงทำให้จำนวนคอนเซ็ปต์ จำนวนกฎความสัมพันธ์ และจำนวนลักษณะเฉพาะที่ได้มีแนวโน้มลดลง

ตารางที่ 5.5 ผลการเลือกลักษณะเฉพาะโดยใช้ FCA ชุดข้อมูล B

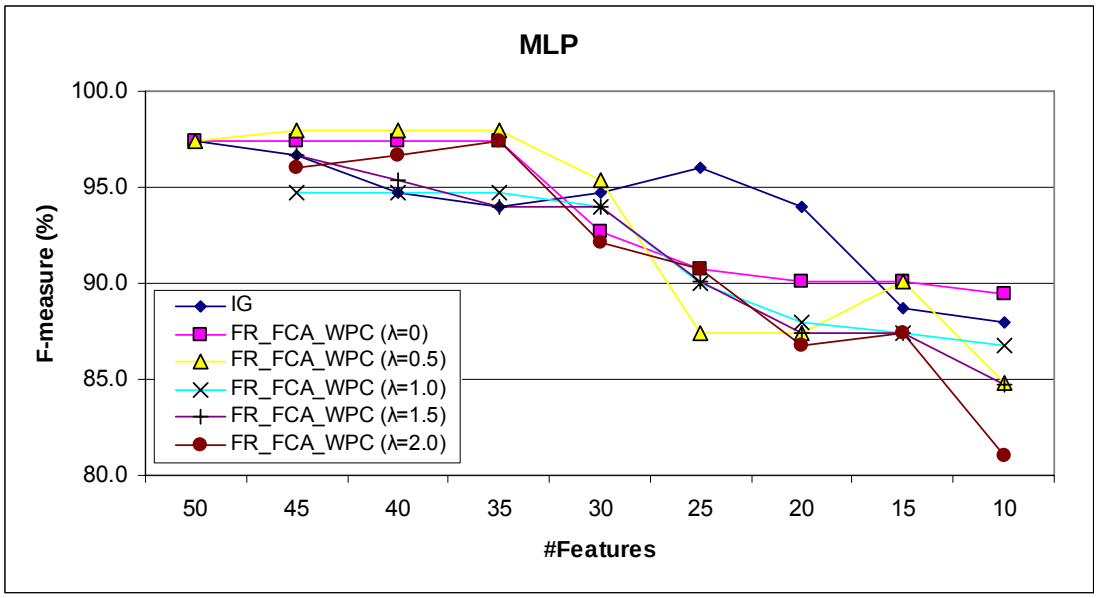
พารามิเตอร์	ชุดข้อมูล B		
	จำนวนคอนเซ็ปต์	จำนวนกฎความสัมพันธ์	จำนวนลักษณะเฉพาะ
$\lambda = 0$	7,414	2,434	50
$\lambda = 0.5$	9,230	3,091	50
$\lambda = 1.0$	6,950	2,389	47
$\lambda = 1.5$	4,347	1,655	46
$\lambda = 2.0$	2,545	1,234	45

ขั้นตอนที่ 4: การจำแนกประเภทและประเมินผล

นำลักษณะเฉพาะที่ได้จากขั้นตอนที่ 3 มาลดขนาดลักษณะเฉพาะลงโดยตัดลักษณะเฉพาะจากอันดับหลังสุดออกทีละ 5 ลักษณะเฉพาะ ซึ่งจะได้ขนาดลักษณะเฉพาะนำเข้าเป็น 50 45 40 35 30 25 20 15 และ 10 ตามลำดับ แล้วนำไปจำแนกประเภทด้วย MLP และ SVM ทดสอบแบบ 10-folds Cross Validation และประเมินผลด้วยค่าเฉลี่ย F-measure จากชุดข้อมูล B ได้ผลการทดลองดังตารางที่ 5.6 และ 5.7 และกราฟเปรียบเทียบดังภาพประกอบ 5.7 และ 5.8 ตามลำดับ

ตารางที่ 5.6 ผลการจำแนกประเภทด้วย MLP ของชุดข้อมูล B

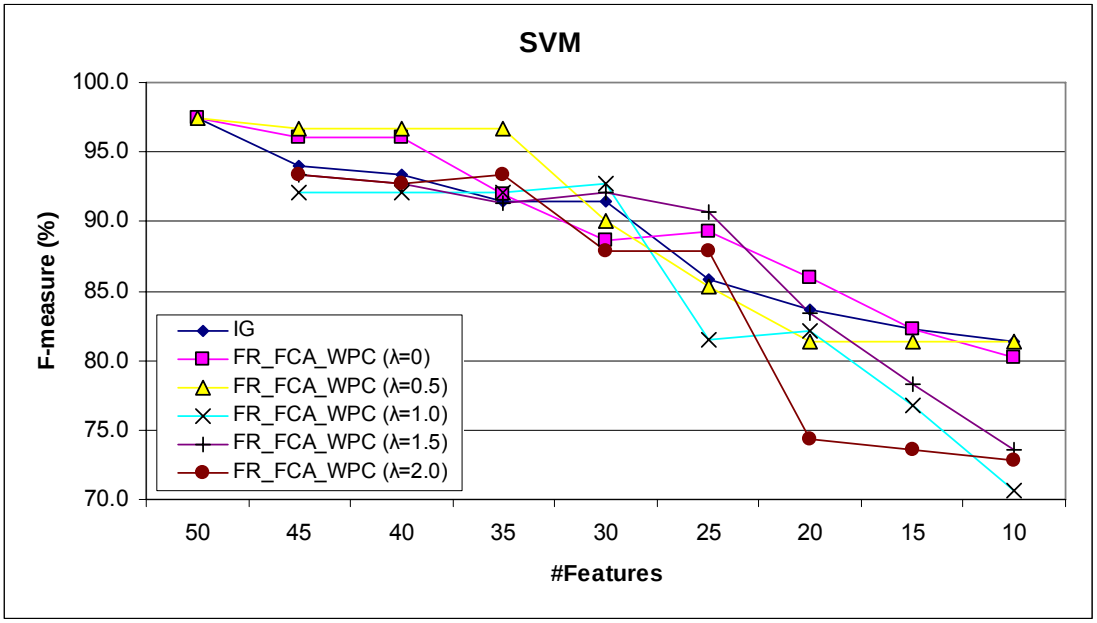
จำนวนลักษณะเฉพาะ (#Features)	ค่า F-measure (%) จำแนกประเภทด้วย MLP ชุดข้อมูล B					
	IG	FR_FCA_WPC				
		$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$
50	***97.4***	**97.4**	**97.4**	-	-	-
45	96.7	**97.4**	***98.0***	94.7	**96.7**	96.0
40	94.7	**97.4**	***98.0***	**94.7**	**95.4**	**96.7**
35	94.0	**97.4**	***98.0***	**94.7**	**94.0**	**97.4**
30	94.7	92.7	**95.4**	94.0	94.0	92.1
25	**96.0**	90.7	87.4	90.0	90.1	90.7
20	**94.0**	90.1	87.4	88.0	87.4	86.8
15	88.7	**90.1**	**90.1**	87.4	87.4	87.4
10	88.0	**89.4**	84.8	86.8	84.7	81.0



ภาพประกอบ 5.7 เปรียบเทียบผลการจำแนกประเภทด้วย MLP ของชุดข้อมูล B

ตารางที่ 5.7 ผลการจำแนกประเภทด้วย SVM ของชุดข้อมูล B

จำนวนลักษณะเฉพาะ (#Features)	ค่า F-measure (%) จำแนกประเภทด้วย SVM ชุดข้อมูล B					
	IG	FR_FCA_WPC				
		$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$
50	***97.4***	***97.4***	***97.4***	-	-	-
45	94.0	**96.0**	**96.7**	92.1	93.4	93.4
40	93.4	**96.0**	**96.7**	92.1	92.7	92.7
35	91.4	**92.0**	**96.7**	**92.1**	91.3	**93.4**
30	91.4	88.7	90.0	**92.7**	**92.1**	87.9
25	85.8	**89.3**	85.3	81.5	**90.7**	**87.9**
20	83.6	**85.9**	81.4	82.1	83.4	74.4
15	**82.2**	**82.2**	81.4	76.8	78.3	73.6
10	**81.4**	80.2	**81.4**	70.6	73.6	72.8



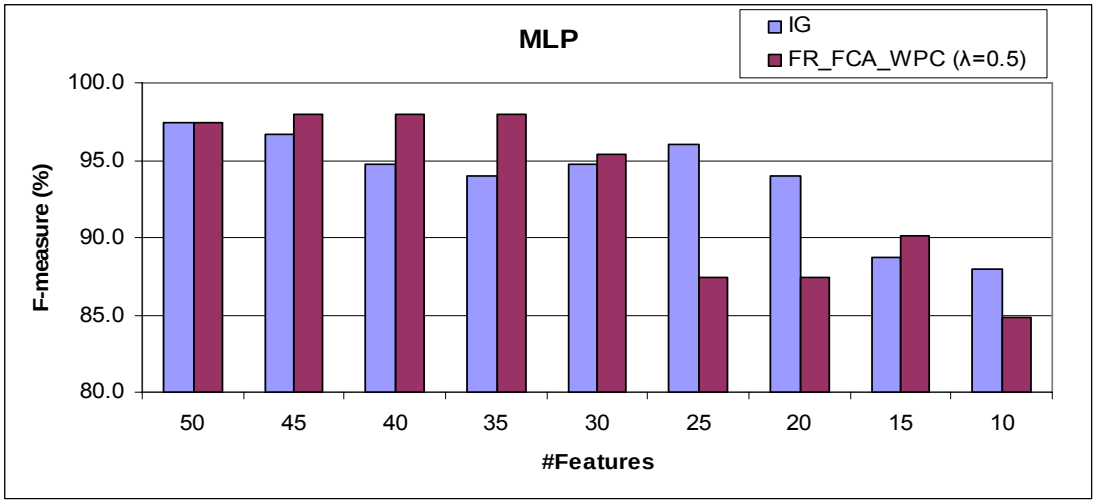
ภาพประกอบ 5.8 เปรียบเทียบผลการจำแนกประเภท SVM ของชุดข้อมูล B

จากผลการทดลองของชุดข้อมูล B สามารถอธิบายได้ 2 ประเด็นดังนี้

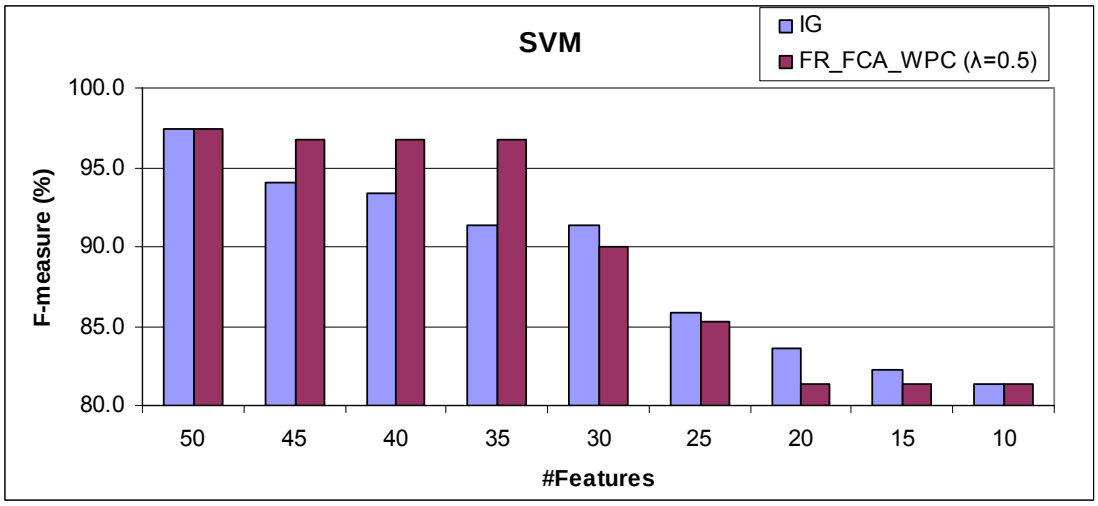
1) ประเด็นเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC กับวิธี IG

ผลการจำแนกประเภทด้วย MLP จากตารางที่ 5.6 และภาพประกอบ 5.7 จะเห็นว่าที่ขนาดลักษณะเฉพาะเท่ากับ 45 40 และ 35 วิธี FR_FCA_WPC ($\lambda = 0.5$) ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 98% ในขณะที่วิธี IG ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 97.4% ที่ขนาดลักษณะเฉพาะเท่ากับ 50 และจากภาพประกอบ 5.9 เมื่อเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0.5$) กับวิธี IG จะเห็นว่าที่ขนาดลักษณะเฉพาะเท่ากับ 45 40 และ 35 วิธี FR_FCA_WPC ($\lambda = 0.5$) ให้ค่า F-measure ที่สูงกว่าวิธี IG

ผลการจำแนกประเภทด้วย SVM จากตารางที่ 5.7 และภาพประกอบ 5.8 จะเห็นว่าที่ขนาดลักษณะเฉพาะเท่ากับ 50 วิธี FR_FCA_WPC ($\lambda = 0$ และ $\lambda = 0.5$) และวิธี IG ให้ค่า F-measure ที่สูงที่สุดเท่ากันคือ 97.4% และจากภาพประกอบ 5.10 เมื่อเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0.5$) กับวิธี IG จะเห็นว่าที่ขนาดลักษณะเฉพาะเท่ากับ 45 40 และ 35 วิธี FR_FCA_WPC ($\lambda = 0.5$) ให้ค่า F-measure ที่สูงกว่าวิธี IG



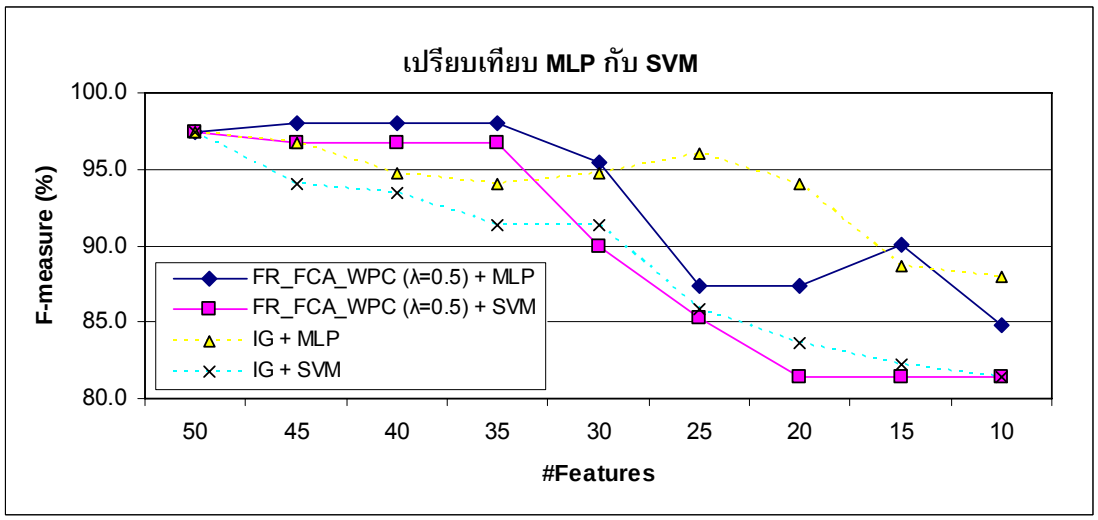
ภาพประกอบ 5.9 เปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0.5$) กับวิธี IG จำแนกประเภทด้วย MLP ของชุดข้อมูล B



ภาพประกอบ 5.10 เปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0.5$) กับวิธี IG จำแนกประเภทด้วย SVM ของชุดข้อมูล B

2) ประเด็นเปรียบเทียบตัวจำแนกประเภท

จากภาพประกอบ 5.11 เมื่อเปรียบเทียบค่า F-measure จำแนกประเภทด้วย MLP และ SVM ลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0.5$) และวิธี IG พบว่าเมื่อลดขนาดลักษณะเฉพาะลง MLP ยังคงให้ค่า F-measure ที่สูงกว่า SVM เช่นที่ขนาดลักษณะเฉพาะเท่ากับ 35 วิธี FR_FCA_WPC ($\lambda = 0.5$) ที่จำแนกประเภทด้วย MLP ให้ค่า F-measure สูงถึง 98% ซึ่งสูงกว่าที่จำแนกด้วย SVM



ภาพประกอบ 5.11 เปรียบเทียบค่า F-measure จำแนกประเภทด้วย MLP และ SVM ลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0.5$) และวิธี IG ของชุดข้อมูล B

5.2.3 การทดลองชุดข้อมูล C

ขั้นตอนที่ 1: การเตรียมข้อมูลเว็บเพจ

จากชุดข้อมูล C ประกอบด้วยเอกสารเว็บเพจ 2 คลาส จำนวน 200 เว็บเพจ หลังจากผ่านขั้นตอนการเตรียมข้อมูลเว็บเพจแล้วจะได้คำจากข้อความจำนวน 13,715 คำ และจากหัวเรื่องจำนวน 355 คำ

ขั้นตอนที่ 2: การเลือกลักษณะเฉพาะโดยใช้ IG

นำคำที่ได้จากข้อความและหัวเรื่องมาเลือกค่าความถี่เอกสารด้วยค่า DF Threshold ของหัวเรื่องเท่ากับ 1 และของข้อความเท่ากับ 15 จะได้ลักษณะเฉพาะจากหัวเรื่องจำนวน 355 ลักษณะเฉพาะ และได้ลักษณะเฉพาะจากข้อความจำนวน 413 ลักษณะเฉพาะ รวมลักษณะเฉพาะที่ได้จากหัวเรื่องและข้อความเข้าด้วยกันเป็นจำนวน 768 ลักษณะเฉพาะ

จากนั้นนำลักษณะเฉพาะที่ได้มากรองเพื่อเลือกลักษณะเฉพาะด้วยวิธี IG โดยในชุดข้อมูลนี้ได้เลือกจำนวนลักษณะเฉพาะเท่ากับ 50 เพื่อใช้ในขั้นตอนต่อไป

ขั้นตอนที่ 3: การเลือกลักษณะเฉพาะโดยใช้ FCA

นำข้อมูลลักษณะเฉพาะที่เลือกจากขั้นตอนที่ 2 มาสร้างเป็นฟอร์มัลคอนเท็กซ์ โดยในชุดข้อมูลนี้ได้ใช้พารามิเตอร์ $\lambda = 0$ $\lambda = 0.5$ $\lambda = 1.0$ $\lambda = 1.5$ และ $\lambda = 2.0$ แปลงข้อมูลลักษณะเฉพาะให้อยู่ในรูปฟอร์มัลคอนเท็กซ์ จากนั้นใช้โปรแกรม ConExp วิเคราะห์หากฎความสัมพันธ์จากฟอร์มัลคอนเท็กซ์ที่ได้ ซึ่งกำหนดค่า Minimal Support เท่ากับ 1 และค่า Confidence เท่ากับ 100% และเรียงกฎตามค่า Support จากมากไปน้อย และเลือกลักษณะเฉพาะโดยสกัดจากกฎความสัมพันธ์ที่ได้

จากชุดข้อมูล C จะได้ผลลัพธ์จากขั้นตอนการเลือกลักษณะเฉพาะโดยใช้ FCA แสดงดังตารางที่ 5.8 เมื่อปรับค่า λ เพิ่มขึ้นจะทำให้ความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารเว็บเพจบางตัวถูกตัดออกไป จึงทำให้จำนวนคอนเซ็ปต์ จำนวนกฎความสัมพันธ์ และจำนวนลักษณะเฉพาะที่ได้มีแนวโน้มลดลง

ตารางที่ 5.8 ผลการเลือกลักษณะเฉพาะโดยใช้ FCA ชุดข้อมูล C

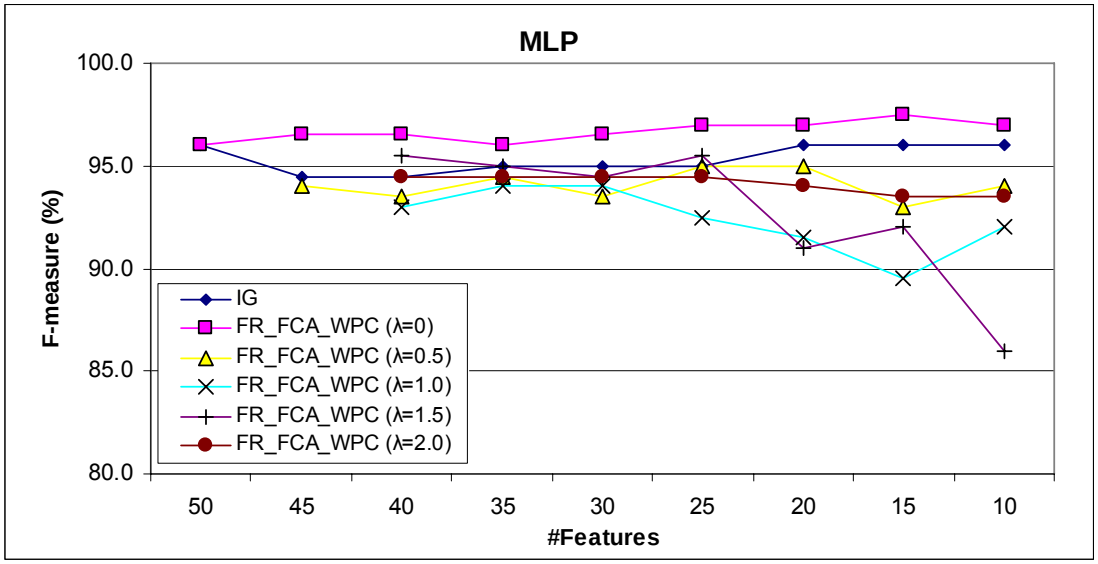
พารามิเตอร์	ชุดข้อมูล C		
	จำนวนคอนเซ็ปต์	จำนวนกฎความสัมพันธ์	จำนวนลักษณะเฉพาะ
$\lambda = 0$	5,129	2,205	50
$\lambda = 0.5$	4,660	2,220	49
$\lambda = 1.0$	3,346	2,030	44
$\lambda = 1.5$	3,177	1,909	44
$\lambda = 2.0$	1,213	857	42

ขั้นตอนที่ 4: การจำแนกประเภทและประเมินผล

นำลักษณะเฉพาะที่ได้จากขั้นตอนที่ 3 มาลดขนาดลักษณะเฉพาะลงโดยตัดลักษณะเฉพาะจากอันดับหลังสุดออกทีละ 5 ลักษณะเฉพาะ ซึ่งจะได้ขนาดลักษณะเฉพาะนำเข้าเป็น 50 45 40 35 30 25 20 15 และ 10 ตามลำดับ แล้วนำไปจำแนกประเภทด้วย MLP และ SVM ทดสอบแบบ 10-folds Cross Validation และประเมินผลด้วยค่าเฉลี่ย F-measure จากชุดข้อมูล C ได้ผลการทดลองดังตารางที่ 5.9 และ 5.10 และกราฟเปรียบเทียบดังภาพประกอบ 5.12 และ 5.13 ตามลำดับ

ตารางที่ 5.9 ผลการจำแนกประเภทด้วย MLP ของชุดข้อมูล C

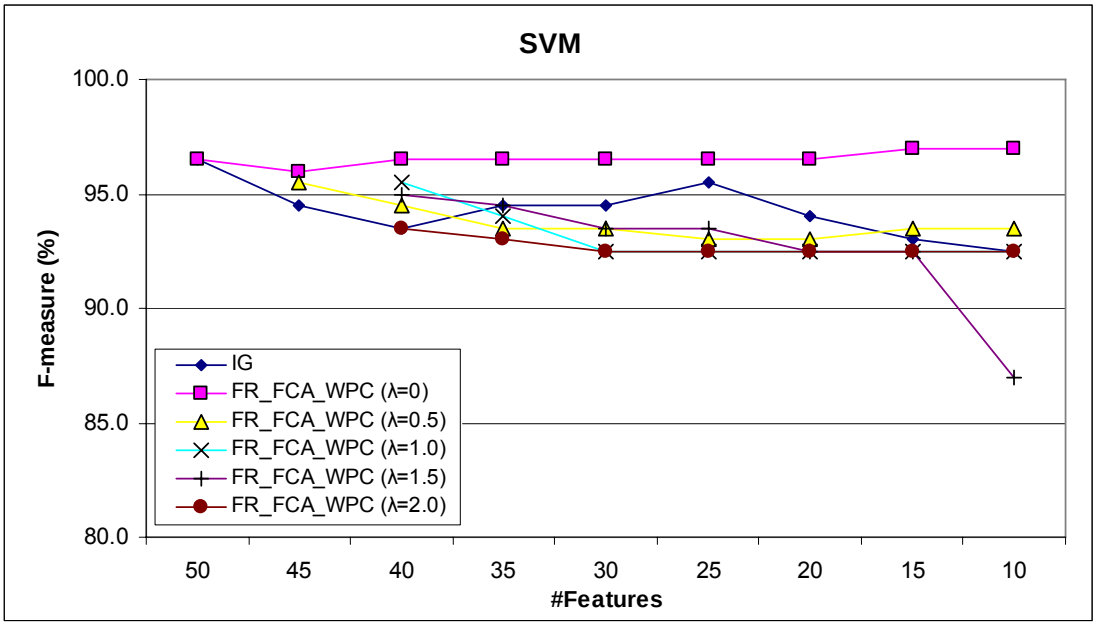
จำนวนลักษณะเฉพาะ (#Features)	ค่า F-measure (%) จำแนกประเภทด้วย MLP ชุดข้อมูล C					
	IG	FR_FCA_WPC				
		$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$
50	***96.0***	**96.0**	-	-	-	-
45	94.5	**96.5**	94.0	-	-	-
40	94.5	**96.5**	93.5	93.0	**95.5**	**94.5**
35	95.0	**96.0**	94.5	94.0	**95.0**	94.5
30	95.0	**96.5**	93.5	94.0	94.5	94.5
25	95.0	**97.0**	**95.0**	92.5	**95.5**	94.5
20	*96.0*	**97.0**	95.0	91.5	91.0	94.0
15	*96.0*	***97.5***	93.0	89.5	92.0	93.5
10	*96.0*	**97.0**	94.0	92.0	86.0	93.5



ภาพประกอบ 5.12 เปรียบเทียบผลการจำแนกประเภทด้วย MLP ของชุดข้อมูล C

ตารางที่ 5.10 ผลการจำแนกประเภทด้วย SVM ของชุดข้อมูล C

จำนวนลักษณะเฉพาะ (#Features)	ค่า F-measure (%) จำแนกประเภทด้วย SVM ชุดข้อมูล C					
	IG	FR_FCA_WPC				
		$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$
50	***96.5***	**96.5**	-	-	-	-
45	94.5	**96.0**	**95.5**	-	-	-
40	93.5	**96.5**	**94.5**	**95.5**	**95.0**	**93.5**
35	94.5	**96.5**	93.5	94.0	**94.5**	93.0
30	94.5	**96.5**	93.5	92.5	93.5	92.5
25	95.5	**96.5**	93.0	92.5	93.5	92.5
20	94.0	**96.5**	93.0	92.5	92.5	92.5
15	93.0	***97.0***	**93.5**	92.5	92.5	92.5
10	92.5	***97.0***	**93.5**	**92.5**	87.0	**92.5**



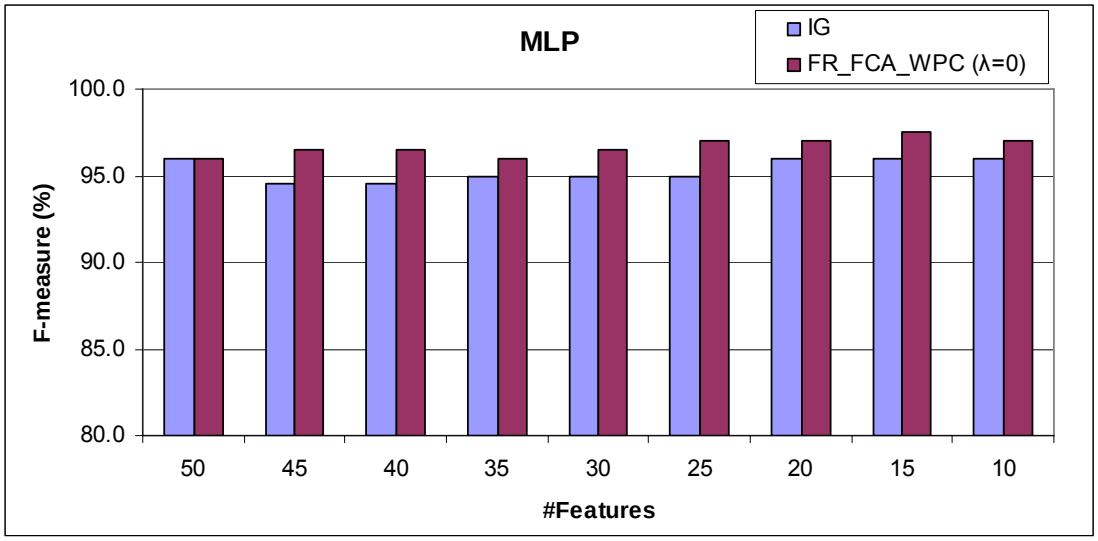
ภาพประกอบ 5.13 เปรียบเทียบผลการจำแนกประเภทด้วย SVM ของชุดข้อมูล C

จากผลการทดลองของชุดข้อมูล C สามารถอธิบายได้ 2 ประเด็นดังนี้

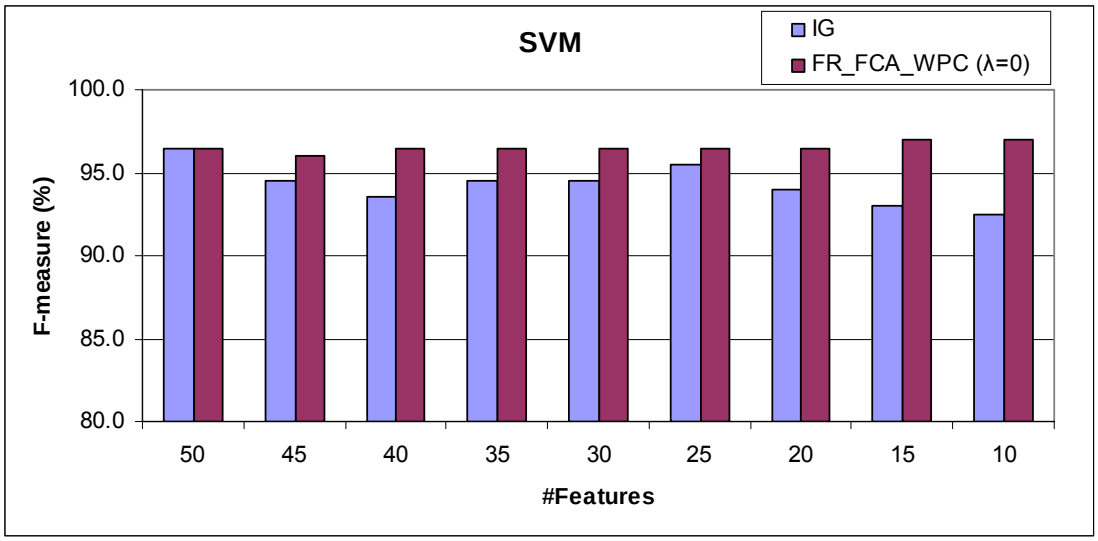
1) ประเด็นเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC กับวิธี IG

ผลการจำแนกประเภทด้วย MLP จากตารางที่ 5.9 และภาพประกอบ 5.12 จะเห็นว่าที่ขนาดลักษณะเฉพาะเท่ากับ 15 วิธี FR_FCA_WPC ($\lambda = 0$) ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 97.5% ในขณะที่วิธี IG ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 96% ที่ขนาดลักษณะเฉพาะเท่ากับ 50 20 15 และ 10 และจากภาพประกอบ 5.14 เมื่อเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จะเห็นว่าวิธี FR_FCA_WPC ($\lambda = 0$) ให้ค่า F-measure ที่สูงกว่าวิธี IG โดยเฉพาะที่ขนาดลักษณะเฉพาะเท่ากับ 10 วิธี FR_FCA_WPC ($\lambda = 0$) ยังให้ค่า F-measure สูงถึง 97% แต่วิธี IG ให้ค่า F-measure เท่ากับ 96%

ผลการจำแนกประเภทด้วย SVM จากตารางที่ 5.10 และภาพประกอบ 5.13 จะเห็นว่าที่ขนาดลักษณะเฉพาะเท่ากับ 15 และ 10 วิธี FR_FCA_WPC ($\lambda = 0$) ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 97% ในขณะที่วิธี IG ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 96.5% ที่ขนาดลักษณะเฉพาะเท่ากับ 50 และจากภาพประกอบ 5.15 เมื่อเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จะเห็นว่าวิธี FR_FCA_WPC ($\lambda = 0$) ให้ค่า F-measure ที่สูงกว่าวิธี IG โดยเฉพาะที่ขนาดลักษณะเฉพาะเท่ากับ 10 วิธี FR_FCA_WPC ($\lambda = 0$) ยังให้ค่า F-measure สูงถึง 97% แต่วิธี IG ให้ค่า F-measure เพียงแค่ 92.5%



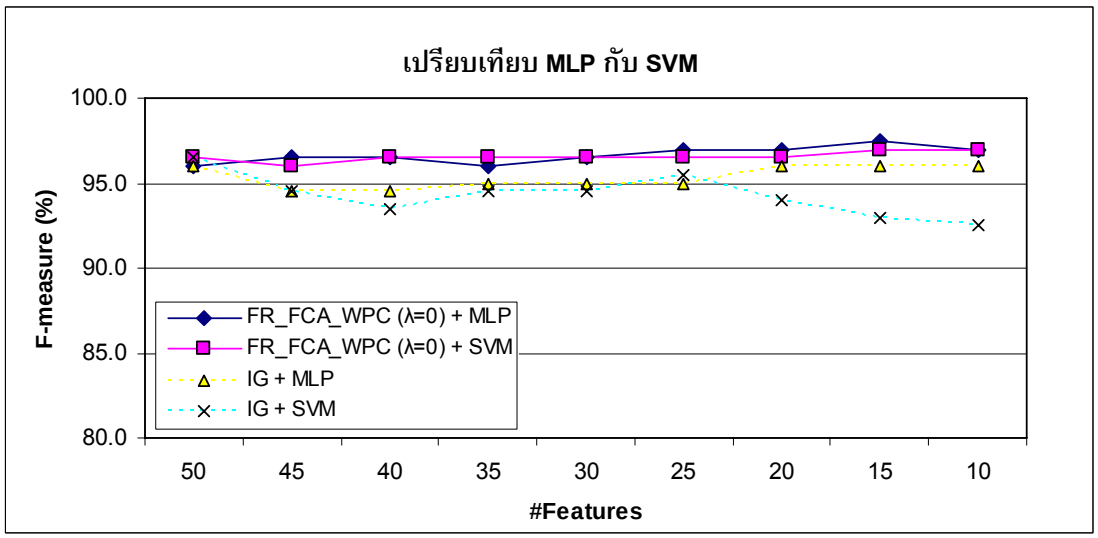
ภาพประกอบ 5.14 เปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จำแนกประเภทด้วย MLP ของชุดข้อมูล C



ภาพประกอบ 5.15 เปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จำแนกประเภทด้วย SVM ของชุดข้อมูล C

2) ประเด็นเปรียบเทียบตัวจำแนกประเภท

จากภาพประกอบ 5.16 เมื่อเปรียบเทียบค่า F-measure จำแนกประเภทด้วย MLP และ SVM ลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) และวิธี IG พบว่า วิธี FR_FCA_WPC ($\lambda = 0$) ที่จำแนกประเภทด้วย MLP และ SVM ให้ค่า F-measure ที่ใกล้เคียงกัน



ภาพประกอบ 5.16 เปรียบเทียบค่า F-measure จำแนกประเภทด้วย MLP และ SVM ลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) และวิธี IG ของชุดข้อมูล C

5.2.4 การทดลองชุดข้อมูล D

ขั้นตอนที่ 1: การเตรียมข้อมูลเว็บเพจ

จากชุดข้อมูล D ประกอบด้วยเอกสารเว็บเพจ 3 คลาส จำนวน 90 เว็บเพจ หลังจากผ่านขั้นตอนการเตรียมข้อมูลเว็บเพจแล้วจะได้คำจากข้อความจำนวน 5,986 คำ และจากหัวเรื่องจำนวน 168 คำ

ขั้นตอนที่ 2: การเลือกลักษณะเฉพาะโดยใช้ IG

นำคำที่ได้จากข้อความและหัวเรื่องมาเลือกค่าความถี่เอกสารด้วยค่า DF Threshold ของหัวเรื่องเท่ากับ 1 และของข้อความเท่ากับ 10 จะได้ลักษณะเฉพาะจากหัวเรื่องจำนวน 168 ลักษณะเฉพาะ และได้ลักษณะเฉพาะจากข้อความจำนวน 178 ลักษณะเฉพาะ รวมลักษณะเฉพาะที่ได้จากหัวเรื่องและข้อความเข้าด้วยกันเป็นจำนวน 346 ลักษณะเฉพาะ

จากนั้นนำลักษณะเฉพาะที่ได้มากรองเพื่อเลือกลักษณะเฉพาะด้วยวิธี IG โดยในชุดข้อมูลนี้ได้เลือกจำนวนลักษณะเฉพาะเท่ากับ 50 เพื่อใช้ในขั้นตอนต่อไป

ขั้นตอนที่ 3: การเลือกลักษณะเฉพาะโดยใช้ FCA

นำข้อมูลลักษณะเฉพาะที่เลือกจากขั้นตอนที่ 2 มาสร้างเป็นฟอร์มัลคอนเท็กซ์ โดยในชุดข้อมูลนี้ได้ใช้พารามิเตอร์ $\lambda = 0$ $\lambda = 0.5$ $\lambda = 1.0$ $\lambda = 1.5$ และ $\lambda = 2.0$ แปลงข้อมูลลักษณะเฉพาะให้อยู่ในรูปฟอร์มัลคอนเท็กซ์ จากนั้นใช้โปรแกรม ConExp วิเคราะห์หากฎความสัมพันธ์จากฟอร์มัลคอนเท็กซ์ที่ได้ ซึ่งกำหนดค่า Minimal Support เท่ากับ 1 และค่า Confidence เท่ากับ 100% และเรียงกฎตามค่า Support จากมากไปน้อย และเลือกลักษณะเฉพาะโดยสกัดจากกฎความสัมพันธ์ที่ได้

จากชุดข้อมูล D จะได้ผลลัพธ์จากขั้นตอนการเลือกลักษณะเฉพาะโดยใช้ FCA แสดงดังตารางที่ 5.11 เมื่อปรับค่า λ เพิ่มขึ้นจะทำให้ความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารเว็บเพจบางตัวถูกตัดออกไป จึงทำให้จำนวนคอนเซ็ปต์ จำนวนกฎความสัมพันธ์ และจำนวนลักษณะเฉพาะที่ได้มีแนวโน้มลดลง

ตารางที่ 5.11 ผลการเลือกลักษณะเฉพาะโดยใช้ FCA ชุดข้อมูล D

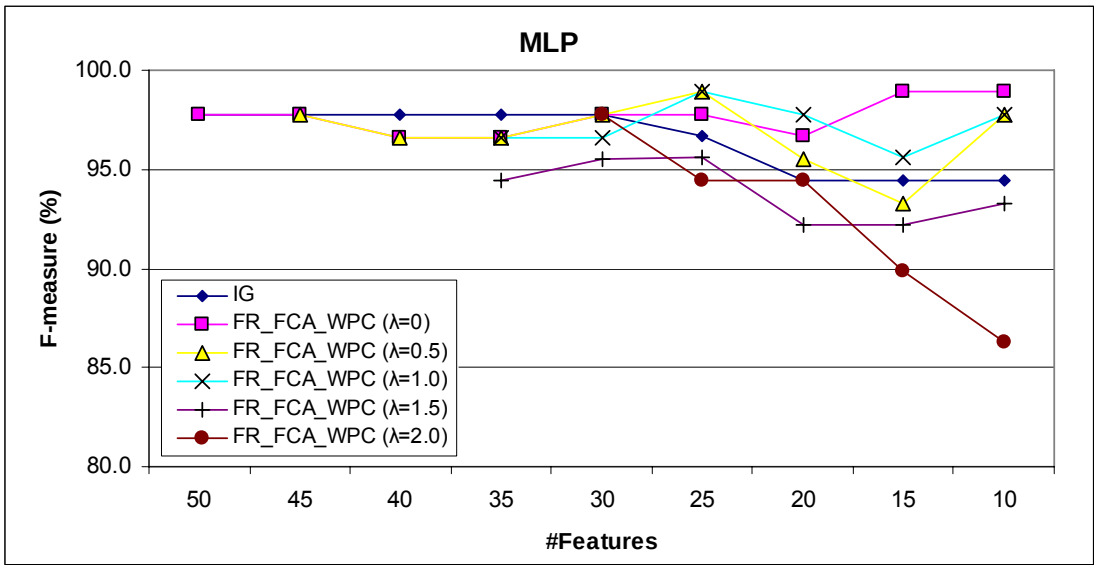
พารามิเตอร์	ชุดข้อมูล D		
	จำนวนคอนเซ็ปต์	จำนวนกฎความสัมพันธ์	จำนวนลักษณะเฉพาะ
$\lambda = 0$	923	648	50
$\lambda = 0.5$	874	656	49
$\lambda = 1.0$	739	604	39
$\lambda = 1.5$	450	429	36
$\lambda = 2.0$	299	256	34

ขั้นตอนที่ 4: การจำแนกประเภทและประเมินผล

นำลักษณะเฉพาะที่ได้จากขั้นตอนที่ 3 มาลดขนาดลักษณะเฉพาะลงโดยตัดลักษณะเฉพาะจากอันดับหลังสุดออกทีละ 5 ลักษณะเฉพาะ ซึ่งจะได้ขนาดลักษณะเฉพาะนำเข้าเป็น 50 45 40 35 30 25 20 15 และ 10 ตามลำดับ แล้วนำไปจำแนกประเภทด้วย MLP และ SVM ทดสอบแบบ 10-folds Cross Validation และประเมินผลด้วยค่าเฉลี่ย F-measure จากชุดข้อมูล D ได้ผลการทดลองดังตารางที่ 5.12 และ 5.13 และกราฟเปรียบเทียบดังภาพประกอบ 5.17 และ 5.18 ตามลำดับ

ตารางที่ 5.12 ผลการจำแนกประเภทด้วย MLP ของชุดข้อมูล D

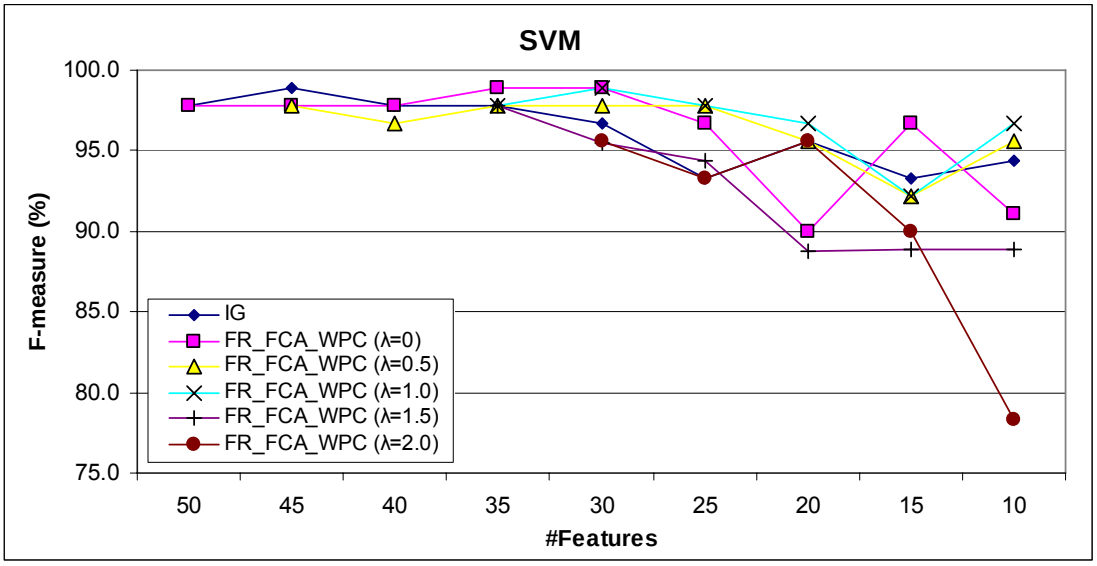
จำนวนลักษณะเฉพาะ (#Features)	ค่า F-measure (%) จำแนกประเภทด้วย MLP ชุดข้อมูล D					
	IG	FR_FCA_WPC				
		$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$
50	***97.8***	**97.8**	-	-	-	-
45	***97.8***	**97.8**	**97.8**	-	-	-
40	***97.8***	96.6	96.6	-	-	-
35	***97.8***	96.6	96.6	96.6	94.4	-
30	***97.8***	**97.8**	**97.8**	96.6	95.5	**97.8**
25	96.7	**97.8**	***98.9***	***98.9***	95.6	94.4
20	94.4	**96.7**	**95.5**	**97.8**	92.2	**94.4**
15	94.4	***98.9***	93.3	**95.6**	92.2	89.9
10	94.4	***98.9***	**97.8**	**97.8**	93.3	86.3



ภาพประกอบ 5.17 เปรียบเทียบผลการจำแนกประเภทด้วย MLP ของชุดข้อมูล D

ตารางที่ 5.13 ผลการจำแนกประเภทด้วย SVM ของชุดข้อมูล D

จำนวนลักษณะเฉพาะ (#Features)	ค่า F-measure (%) จำแนกประเภทด้วย SVM ชุดข้อมูล D					
	IG	FR_FCA_WPC				
		$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$
50	**97.8**	**97.8**	-	-	-	-
45	***98.9***	97.8	97.8	-	-	-
40	**97.8**	**97.8**	96.7	-	-	-
35	97.8	***98.9***	**97.8**	**97.8**	**97.8**	-
30	96.7	***98.9***	**97.8**	***98.9***	95.5	95.6
25	93.3	**96.7**	**97.8**	**97.8**	**94.4**	**93.3**
20	95.6	90.0	**95.6**	**96.7**	88.8	**95.6**
15	93.3	**96.7**	92.2	92.2	88.9	90.0
10	94.4	91.1	**95.6**	**96.7**	88.9	78.3



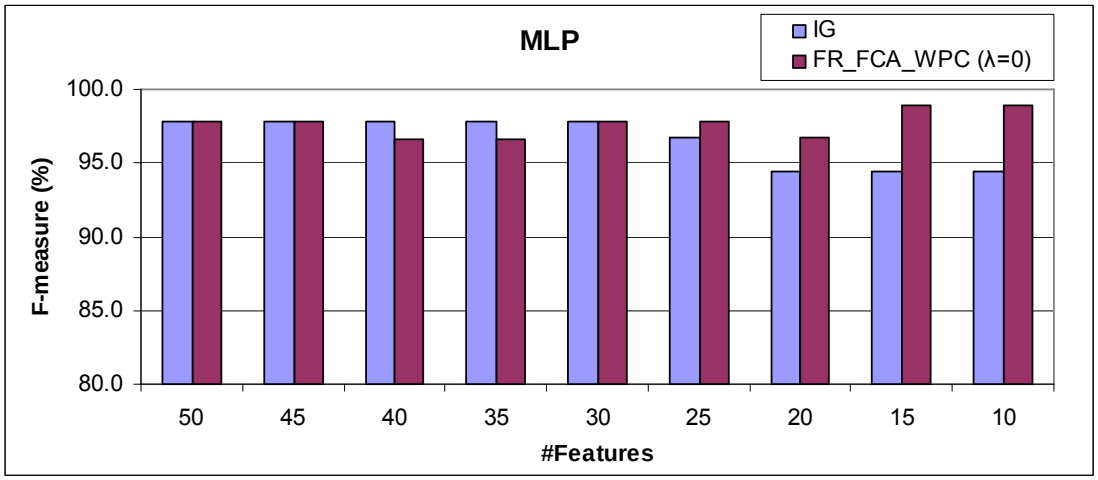
ภาพประกอบ 5.18 เปรียบเทียบผลการจำแนกประเภทด้วย SVM ของชุดข้อมูล D

จากผลการทดลองของชุดข้อมูล D สามารถอธิบายได้ 2 ประเด็นดังนี้

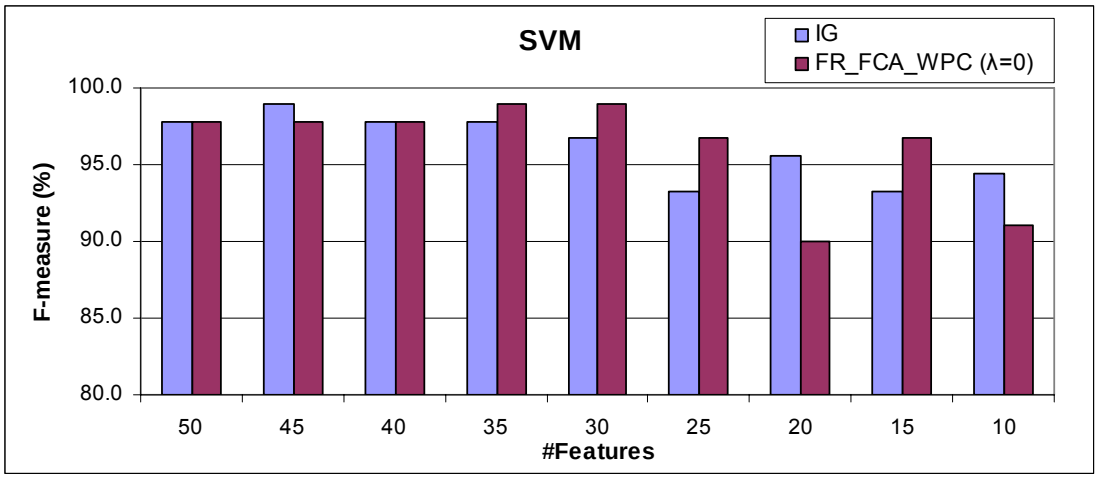
1) ประเด็นเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC กับวิธี IG

ผลการจำแนกประเภทด้วย MLP จากตารางที่ 5.12 และภาพประกอบ 5.17 จะเห็นว่าที่ขนาดลักษณะเฉพาะเท่ากับ 25 15 และ 10 วิธี FR_FCA_WPC ($\lambda = 0$ $\lambda = 0.5$ หรือ $\lambda = 1.0$) ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 98.9% ในขณะที่วิธี IG ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 97.8% ที่ขนาดลักษณะเฉพาะเท่ากับ 50 45 40 35 และ 30 และจากภาพประกอบ 5.19 เมื่อเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จะเห็นว่าวิธี FR_FCA_WPC ($\lambda = 0$) ให้ค่า F-measure สูงกว่าวิธี IG ที่ขนาดลักษณะเฉพาะเท่ากับ 25 20 15 และ 10 โดยเฉพาะที่ขนาดลักษณะเฉพาะเท่ากับ 10 วิธี FR_FCA_WPC ($\lambda = 0$) ยังให้ค่า F-measure สูงถึง 98.9% แต่วิธี IG ให้ค่า F-measure เพียงแค่ 94.4%

ผลการจำแนกประเภทด้วย SVM จากตารางที่ 5.13 และภาพประกอบ 5.18 จะเห็นว่าที่ขนาดลักษณะเฉพาะเท่ากับ 35 และ 30 วิธี FR_FCA_WPC ($\lambda = 0$ หรือ $\lambda = 1.0$) ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 98.9% ซึ่งเท่ากันกับวิธี IG ที่ขนาดลักษณะเฉพาะเท่ากับ 45 และจากภาพประกอบ 5.20 เมื่อเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จะเห็นว่าวิธี FR_FCA_WPC ($\lambda = 0$) ให้ค่า F-measure สูงกว่าวิธี IG ที่ขนาดลักษณะเฉพาะเท่ากับ 35 30 25 และ 15 โดยเฉพาะที่ขนาดลักษณะเฉพาะเท่ากับ 30 วิธี FR_FCA_WPC ($\lambda = 0$) ยังให้ค่า F-measure สูงถึง 98.9% แต่วิธี IG ให้ค่า F-measure เพียงแค่ 96.7%



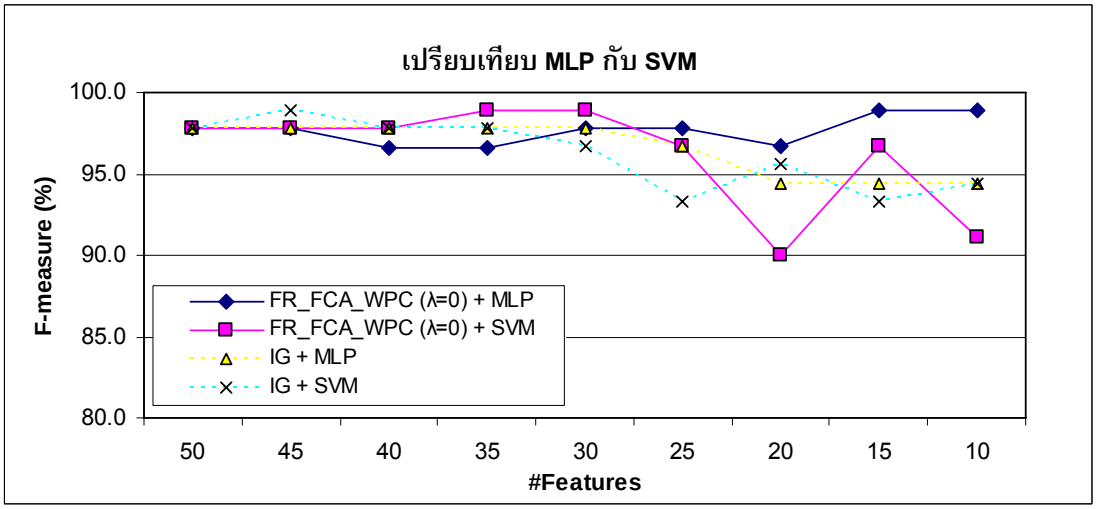
ภาพประกอบ 5.19 เปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จำแนกประเภทด้วย MLP ของชุดข้อมูล D



ภาพประกอบ 5.20 เปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จำแนกประเภทด้วย SVM ของชุดข้อมูล D

2) ประเด็นเปรียบเทียบตัวจำแนกประเภท

จากภาพประกอบ 5.21 เมื่อเปรียบเทียบค่า F-measure จำแนกประเภทด้วย MLP และ SVM ลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) และวิธี IG พบว่าเมื่อลดขนาดลักษณะเฉพาะลง MLP ยังคงให้ค่า F-measure ที่สูงกว่า SVM เช่นที่ขนาดลักษณะเฉพาะเท่ากับ 10 วิธี FR_FCA_WPC ($\lambda = 0$) ที่จำแนกประเภทด้วย MLP ให้ค่า F-measure สูงถึง 98.9% ซึ่งสูงกว่าที่จำแนกด้วย SVM



ภาพประกอบ 5.21 เปรียบเทียบค่า F-measure จำแนกประเภทด้วย MLP และ SVM ลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) และวิธี IG ของชุดข้อมูล D

5.2.5 การทดลองชุดข้อมูล E

ขั้นตอนที่ 1: การเตรียมข้อมูลเว็บเพจ

จากชุดข้อมูล E ประกอบด้วยเอกสารเว็บเพจ 2 คลาส จำนวน 100 เว็บเพจ หลังจากผ่านขั้นตอนการเตรียมข้อมูลเว็บเพจแล้วจะได้คำจากข้อความจำนวน 33,973 คำ และจากหัวเรื่องจำนวน 224 คำ

ขั้นตอนที่ 2: การเลือกลักษณะเฉพาะโดยใช้ IG

นำคำที่ได้จากข้อความและหัวเรื่องมาเลือกค่าความถี่เอกสารด้วยค่า DF Threshold ของหัวเรื่องเท่ากับ 1 และของข้อความเท่ากับ 20 จะได้ลักษณะเฉพาะจากหัวเรื่องจำนวน 224 ลักษณะเฉพาะ และได้ลักษณะเฉพาะจากข้อความจำนวน 371 ลักษณะเฉพาะ รวมลักษณะเฉพาะที่ได้จากหัวเรื่องและข้อความเข้าด้วยกันเป็นจำนวน 595 ลักษณะเฉพาะ

จากนั้นนำลักษณะเฉพาะที่ได้มากรองเพื่อเลือกลักษณะเฉพาะด้วยวิธี IG โดยในชุดข้อมูลนี้ได้เลือกจำนวนลักษณะเฉพาะเท่ากับ 50 เพื่อใช้ในขั้นตอนต่อไป

ขั้นตอนที่ 3: การเลือกลักษณะเฉพาะโดยใช้ FCA

นำข้อมูลลักษณะเฉพาะที่เลือกจากขั้นตอนที่ 2 มาสร้างเป็นฟอร์มัลคอนเท็กซ์ โดยในชุดข้อมูลนี้ได้ใช้พารามิเตอร์ $\lambda = 0$ $\lambda = 0.5$ $\lambda = 1.0$ $\lambda = 1.5$ และ $\lambda = 2.0$ แปลงข้อมูลลักษณะเฉพาะให้อยู่ในรูปฟอร์มัลคอนเท็กซ์ จากนั้นใช้โปรแกรม ConExp วิเคราะห์หากฎความสัมพันธ์จากฟอร์มัลคอนเท็กซ์ที่ได้ ซึ่งกำหนดค่า Minimal Support เท่ากับ 1 และค่า Confidence เท่ากับ 100% และเรียงกฎตามค่า Support จากมากไปน้อย และเลือกลักษณะเฉพาะโดยสกัดจากกฎความสัมพันธ์ที่ได้

จากชุดข้อมูล E จะได้ผลลัพธ์จากขั้นตอนการเลือกลักษณะเฉพาะโดยใช้ FCA แสดงดังตารางที่ 5.14 เมื่อปรับค่า λ จาก $\lambda = 0$ เพิ่มขึ้นเป็น $\lambda = 0.5$ จำนวนคอนเซ็ปต์ จำนวนกฎความสัมพันธ์ และจำนวนลักษณะเฉพาะที่ได้ยังไม่เปลี่ยนแปลง เนื่องจากไม่มีความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารใดถูกตัดออกไปเลย และเมื่อปรับค่าเป็น $\lambda = 1.0$ $\lambda = 1.5$ และ $\lambda = 2.0$ จะทำให้จำนวนที่ได้มีแนวโน้มลดลง

ตารางที่ 5.14 ผลการเลือกลักษณะเฉพาะโดยใช้ FCA ชุดข้อมูล E

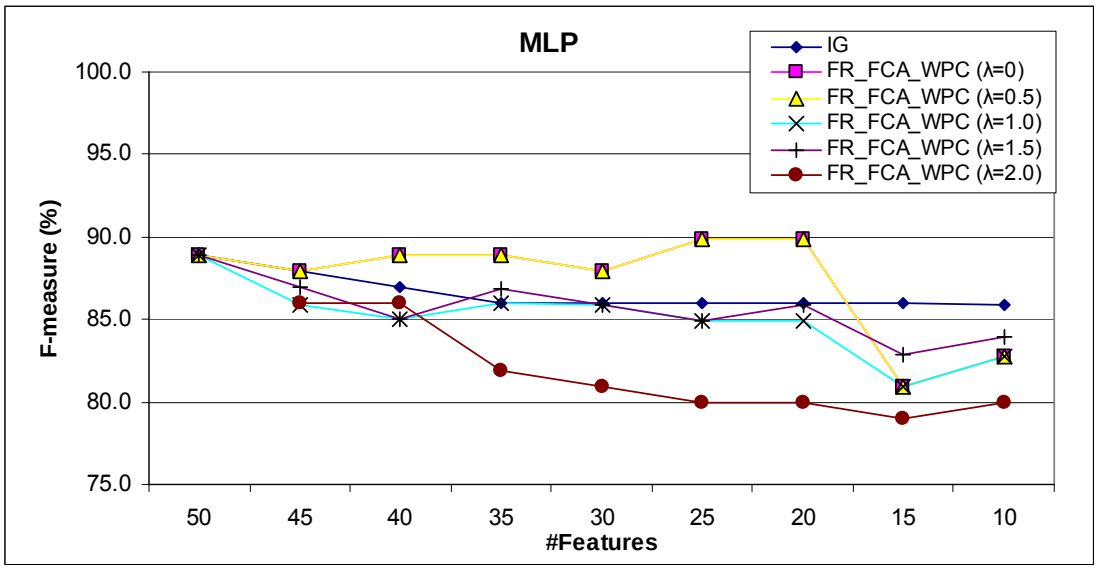
พารามิเตอร์	ชุดข้อมูล E		
	จำนวนคอนเซ็ปต์	จำนวนกฎความสัมพันธ์	จำนวนลักษณะเฉพาะ
$\lambda = 0$	14,237	2,966	50
$\lambda = 0.5$	14,237	2,966	50
$\lambda = 1.0$	12,292	2,841	50
$\lambda = 1.5$	8,815	2,337	50
$\lambda = 2.0$	4,878	1,681	49

ขั้นตอนที่ 4: การจำแนกประเภทและประเมินผล

นำลักษณะเฉพาะที่ได้จากขั้นตอนที่ 3 มาลดขนาดลักษณะเฉพาะลงโดยตัดลักษณะเฉพาะจากอันดับหลังสุดออกทีละ 5 ลักษณะเฉพาะ ซึ่งจะได้ขนาดลักษณะเฉพาะนำเข้าเป็น 50 45 40 35 30 25 20 15 และ 10 ตามลำดับ แล้วนำไปจำแนกประเภทด้วย MLP และ SVM ทดสอบแบบ 10-folds Cross Validation และประเมินผลด้วยค่าเฉลี่ย F-measure จากชุดข้อมูล E ได้ผลการทดลองดังตารางที่ 5.15 และ 5.16 และกราฟเปรียบเทียบดังภาพประกอบ 5.22 และ 5.23 ตามลำดับ

ตารางที่ 5.15 ผลการจำแนกประเภทด้วย MLP ของชุดข้อมูล E

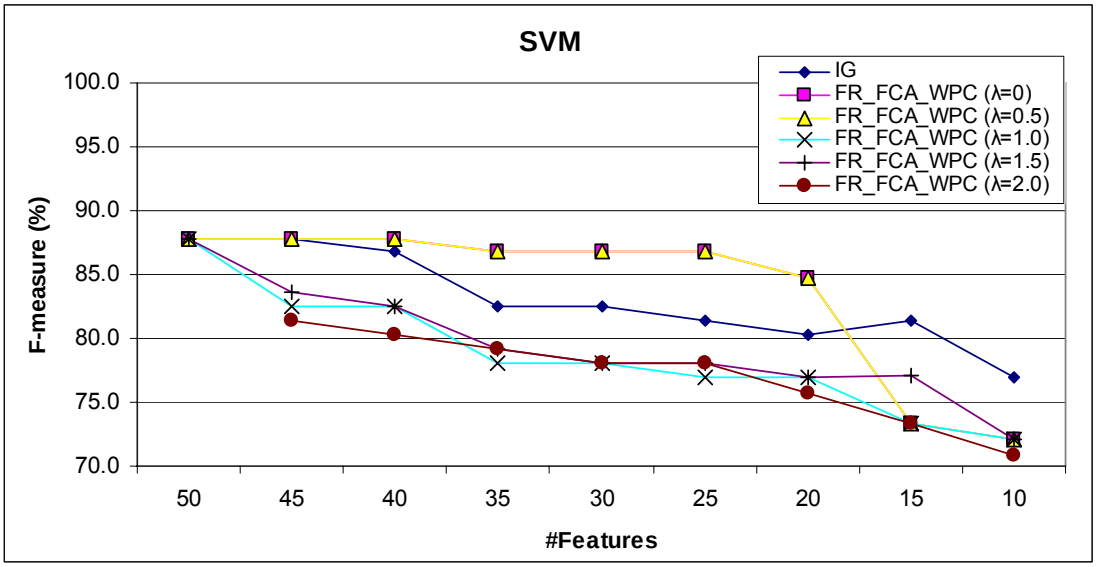
จำนวนลักษณะเฉพาะ (#Features)	ค่า F-measure (%) จำแนกประเภทด้วย MLP ชุดข้อมูล E					
	IG	FR_FCA_WPC				
		$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$
50	***88.9***	**88.9**	**88.9**	**88.9**	**88.9**	-
45	**87.9**	**87.9**	**87.9**	85.9	87.0	86.0
40	87.0	**88.9**	**88.9**	85.0	85.0	86.0
35	86.0	**88.9**	**88.9**	**86.0**	**86.9**	81.9
30	86.0	**87.9**	**87.9**	85.9	85.9	80.9
25	86.0	***89.9***	***89.9***	84.9	84.9	80.0
20	86.0	***89.9***	***89.9***	84.9	85.9	80.0
15	**86.0**	80.9	80.9	80.9	82.9	79.0
10	**85.9**	82.8	82.8	82.8	83.9	80.0



ภาพประกอบ 5.22 เปรียบเทียบผลการจำแนกประเภทด้วย MLP ของชุดข้อมูล E

ตารางที่ 5.16 ผลการจำแนกประเภทด้วย SVM ของชุดข้อมูล E

จำนวนลักษณะเฉพาะ (#Features)	ค่า F-measure (%) จำแนกประเภทด้วย SVM ชุดข้อมูล E					
	IG	FR_FCA_WPC				
		$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$
50	***87.8***	***87.8***	***87.8***	***87.8***	***87.8***	-
45	***87.8***	***87.8***	***87.8***	82.5	83.6	81.4
40	86.8	***87.8***	***87.8***	82.5	82.5	80.3
35	82.5	**86.8**	**86.8**	78.0	79.2	79.2
30	82.5	**86.8**	**86.8**	78.0	78.0	78.0
25	81.4	**86.8**	**86.8**	76.9	78.0	78.0
20	80.3	**84.7**	**84.7**	76.9	76.9	75.7
15	**81.4**	73.3	73.3	73.3	77.1	73.3
10	**76.9**	72.1	72.1	72.1	72.1	70.9



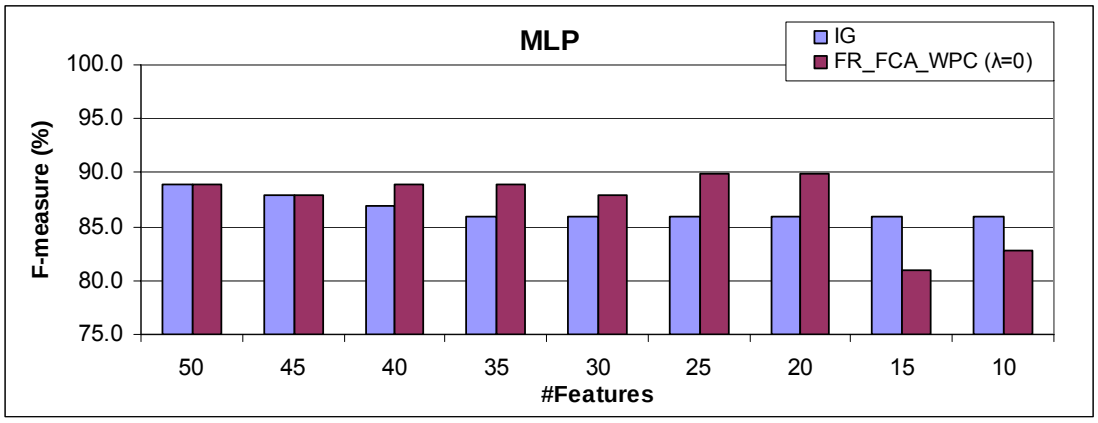
ภาพประกอบ 5.23 เปรียบเทียบผลการจำแนกประเภทด้วย SVM ของชุดข้อมูล E

จากผลการทดลองของชุดข้อมูล E สามารถอธิบายได้ 2 ประเด็นดังนี้

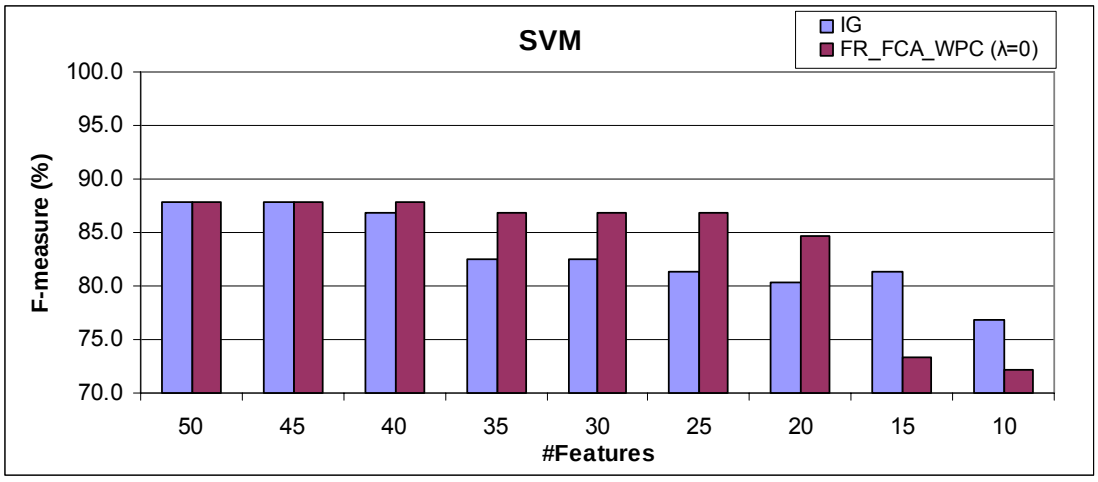
1) ประเด็นเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC กับวิธี IG

ผลการจำแนกประเภทด้วย MLP จากตารางที่ 5.15 และภาพประกอบ 5.22 จะเห็นว่าวิธี FR_FCA_WPC ที่ $\lambda = 0$ และ $\lambda = 0.5$ ให้ค่า F-measure เท่ากัน และที่ขนาดลักษณะเฉพาะเท่ากับ 25 และ 20 วิธี FR_FCA_WPC ($\lambda = 0$ และ $\lambda = 0.5$) ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 89.9% ในขณะที่วิธี IG ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 88.9% ที่ขนาดลักษณะเฉพาะเท่ากับ 50 และจากภาพประกอบ 5.24 เมื่อเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จะเห็นว่าวิธี FR_FCA_WPC ($\lambda = 0$) ให้ค่า F-measure สูงกว่าวิธี IG ที่ขนาดลักษณะเฉพาะเท่ากับ 40 35 30 25 และ 20 โดยเฉพาะที่ขนาดลักษณะเฉพาะเท่ากับ 20 วิธี FR_FCA_WPC ($\lambda = 0$) ยังให้ค่า F-measure สูงถึง 89.9% แต่วิธี IG ให้ค่า F-measure เพียงแค่ 86%

ผลการจำแนกประเภทด้วย SVM จากตารางที่ 5.16 และภาพประกอบ 5.23 จะเห็นว่าวิธี FR_FCA_WPC และวิธี IG ให้ค่า F-measure ที่สูงที่สุดเท่ากัน คือ 87.8% แต่วิธี FR_FCA_WPC ($\lambda = 0$ และ $\lambda = 0.5$) สามารถลดขนาดลักษณะเฉพาะอยู่ที่ 40 ซึ่งใช้จำนวนลักษณะเฉพาะน้อยกว่าวิธี IG ที่สามารถลดขนาดลักษณะเฉพาะอยู่ที่ 45 และจากภาพประกอบ 5.25 เมื่อเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จะเห็นว่าวิธี FR_FCA_WPC ($\lambda = 0$) ให้ค่า F-measure สูงกว่าวิธี IG ที่ขนาดลักษณะเฉพาะเท่ากับ 40 35 30 25 และ 20 เช่นที่ขนาดลักษณะเฉพาะเท่ากับ 25 วิธี FR_FCA_WPC ($\lambda = 0$) ยังให้ค่า F-measure สูงถึง 86.8% แต่วิธี IG ให้ค่า F-measure เพียงแค่ 81.4%



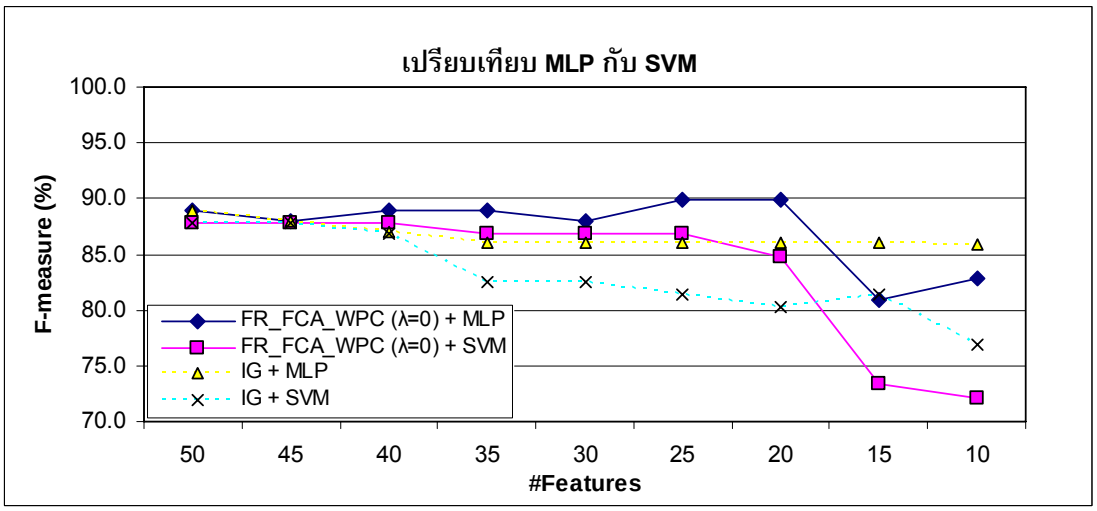
ภาพประกอบ 5.24 เปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จำแนกประเภทด้วย MLP ของชุดข้อมูล E



ภาพประกอบ 5.25 เปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จำแนกประเภทด้วย SVM ของชุดข้อมูล E

2) ประเด็นเปรียบเทียบตัวจำแนกประเภท

จากภาพประกอบ 5.26 เมื่อเปรียบเทียบค่า F-measure จำแนกประเภทด้วย MLP และ SVM ลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) และวิธี IG พบว่าเมื่อลดขนาดลักษณะเฉพาะลง MLP ยังคงให้ค่า F-measure ที่สูงกว่า SVM โดยเฉพาะที่ขนาดลักษณะเฉพาะเท่ากับ 20 วิธี FR_FCA_WPC ($\lambda = 0$) ที่จำแนกประเภทด้วย MLP ให้ค่า F-measure สูงถึง 89.9% ซึ่งสูงกว่าที่จำแนกด้วย SVM



ภาพประกอบ 5.26 เปรียบเทียบค่า F-measure จำแนกประเภทด้วย MLP และ SVM ลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) และวิธี IG ของชุดข้อมูล E

5.2.6 การทดลองชุดข้อมูล F

ขั้นตอนที่ 1: การเตรียมข้อมูลเว็บเพจ

จากชุดข้อมูล F ประกอบด้วยเอกสารเว็บเพจ 2 คลาส จำนวน 100 เว็บเพจ หลังจากผ่านขั้นตอนการเตรียมข้อมูลเว็บเพจแล้วจะได้คำจากข้อความจำนวน 16,335 คำ และจากหัวเรื่องจำนวน 180 คำ

ขั้นตอนที่ 2: การเลือกลักษณะเฉพาะโดยใช้ IG

นำคำที่ได้จากข้อความและหัวเรื่องมาเลือกค่าความถี่เอกสารด้วยค่า DF Threshold ของหัวเรื่องเท่ากับ 1 และของข้อความเท่ากับ 20 จะได้ลักษณะเฉพาะจากหัวเรื่องจำนวน 180 ลักษณะเฉพาะ และได้ลักษณะเฉพาะจากข้อความจำนวน 413 ลักษณะเฉพาะ รวมลักษณะเฉพาะที่ได้จากหัวเรื่องและข้อความเข้าด้วยกันเป็นจำนวน 593 ลักษณะเฉพาะ

จากนั้นนำลักษณะเฉพาะที่ได้มากรองเพื่อเลือกลักษณะเฉพาะด้วยวิธี IG โดยในชุดข้อมูลนี้ได้เลือกจำนวนลักษณะเฉพาะเท่ากับ 50 เพื่อใช้ในขั้นตอนต่อไป

ขั้นตอนที่ 3: การเลือกลักษณะเฉพาะโดยใช้ FCA

นำข้อมูลลักษณะเฉพาะที่เลือกจากขั้นตอนที่ 2 มาสร้างเป็นฟอร์มัลคอนเท็กซ์ โดยในชุดข้อมูลนี้ได้ใช้พารามิเตอร์ $\lambda = 0$ $\lambda = 0.5$ $\lambda = 1.0$ $\lambda = 1.5$ และ $\lambda = 2.0$ แปลงข้อมูลลักษณะเฉพาะให้อยู่ในรูปฟอร์มัลคอนเท็กซ์ จากนั้นใช้โปรแกรม ConExp วิเคราะห์หากฎความสัมพันธ์จากฟอร์มัลคอนเท็กซ์ที่ได้ ซึ่งกำหนดค่า Minimal Support เท่ากับ 1 และค่า Confidence เท่ากับ 100% และเรียงกฎตามค่า Support จากมากไปน้อย และเลือกลักษณะเฉพาะโดยสกัดจากกฎความสัมพันธ์ที่ได้

จากชุดข้อมูล F จะได้ผลลัพธ์จากขั้นตอนการเลือกลักษณะเฉพาะโดยใช้ FCA แสดงดังตารางที่ 5.17 เมื่อปรับค่า λ จาก $\lambda = 0$ เพิ่มขึ้นเป็น $\lambda = 0.5$ จำนวนคอนเซ็ปต์ จำนวนกฎความสัมพันธ์ และจำนวนลักษณะเฉพาะที่ได้ยังไม่เปลี่ยนแปลง เนื่องจากไม่มีความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารใดถูกตัดออกไปเลย และเมื่อปรับค่าเป็น $\lambda = 1.0$ $\lambda = 1.5$ และ $\lambda = 2.0$ จะทำให้จำนวนที่ได้มีแนวโน้มลดลง

ตารางที่ 5.17 ผลการเลือกลักษณะเฉพาะโดยใช้ FCA ชุดข้อมูล F

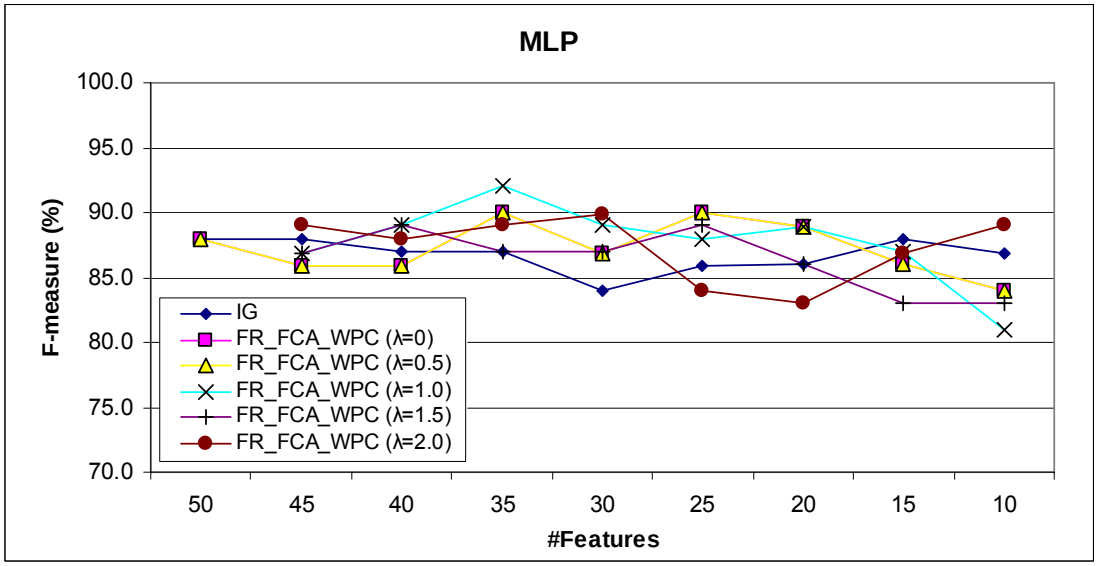
พารามิเตอร์	ชุดข้อมูล F		
	จำนวนคอนเซ็ปต์	จำนวนกฎความสัมพันธ์	จำนวนลักษณะเฉพาะ
$\lambda = 0$	44,289	5,932	50
$\lambda = 0.5$	44,289	5,932	50
$\lambda = 1.0$	31,846	5,576	48
$\lambda = 1.5$	19,181	4,274	48
$\lambda = 2.0$	8,894	2,593	47

ขั้นตอนที่ 4: การจำแนกประเภทและประเมินผล

นำลักษณะเฉพาะที่ได้จากขั้นตอนที่ 3 มาลดขนาดลักษณะเฉพาะลงโดยตัดลักษณะเฉพาะจากอันดับหลังสุดออกทีละ 5 ลักษณะเฉพาะ ซึ่งจะได้ขนาดลักษณะเฉพาะนำเข้าเป็น 50 45 40 35 30 25 20 15 และ 10 ตามลำดับ แล้วนำไปจำแนกประเภทด้วย MLP และ SVM ทดสอบแบบ 10-folds Cross Validation และประเมินผลด้วยค่าเฉลี่ย F-measure จากชุดข้อมูล F ได้ผลการทดลองดังตารางที่ 5.18 และ 5.19 และกราฟเปรียบเทียบดังภาพประกอบ 5.27 และ 5.28 ตามลำดับ

ตารางที่ 5.18 ผลการจำแนกประเภทด้วย MLP ของชุดข้อมูล F

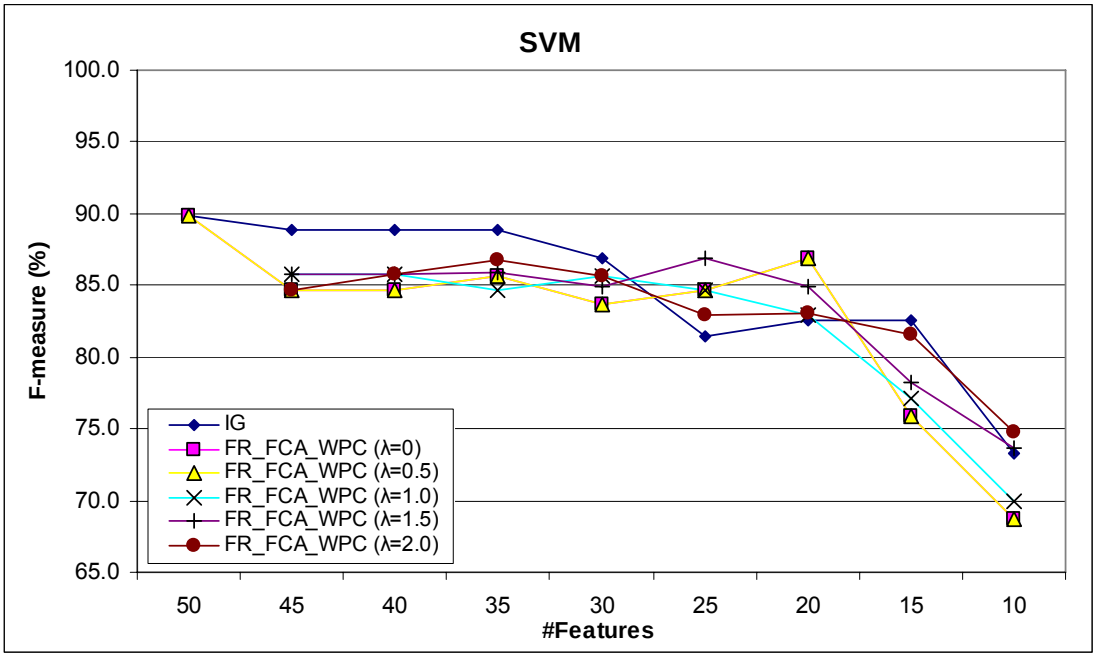
จำนวนลักษณะเฉพาะ (#Features)	ค่า F-measure (%) จำแนกประเภทด้วย MLP ชุดข้อมูล F					
	IG	FR_FCA_WPC				
		$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$
50	***87.9***	**87.9**	**87.9**			
45	*87.9*	85.9	85.9	86.9	86.9	**89.0**
40	87.0	85.9	85.9	**89.0**	**89.0**	**88.0**
35	87.0	**90.0**	**90.0**	***92.0***	**87.0**	**89.0**
30	84.0	**86.9**	**86.9**	**89.0**	**87.0**	**89.9**
25	85.9	**90.0**	**90.0**	87.9	**89.0**	84.0
20	86.0	**88.9**	**88.9**	**88.9**	**86.0**	83.0
15	***87.9***	86.0	86.0	87.0	83.0	86.9
10	86.8	84.0	84.0	81.0	83.0	**89.0**



ภาพประกอบ 5.27 เปรียบเทียบผลการจำแนกประเภทด้วย MLP ของชุดข้อมูล F

ตารางที่ 5.19 ผลการจำแนกประเภทด้วย SVM ของชุดข้อมูล F

จำนวนลักษณะเฉพาะ (#Features)	ค่า F-measure (%) จำแนกประเภทด้วย SVM ชุดข้อมูล F					
	IG	FR_FCA_WPC				
		$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$
50	***89.9***	***89.9***	***89.9***			
45	**88.9**	84.7	84.7	85.8	85.8	84.7
40	**88.9**	84.7	84.7	85.8	85.8	85.8
35	**88.9**	85.7	85.7	84.7	85.9	86.8
30	**86.9**	83.7	83.7	85.7	84.9	85.7
25	81.5	**84.7**	**84.7**	**84.7**	**86.9**	**82.9**
20	82.6	**86.9**	**86.9**	**82.9**	**84.9**	**83.0**
15	**82.6**	75.9	75.9	77.1	78.2	81.6
10	73.3	68.7	68.7	70.0	73.6	**74.8**



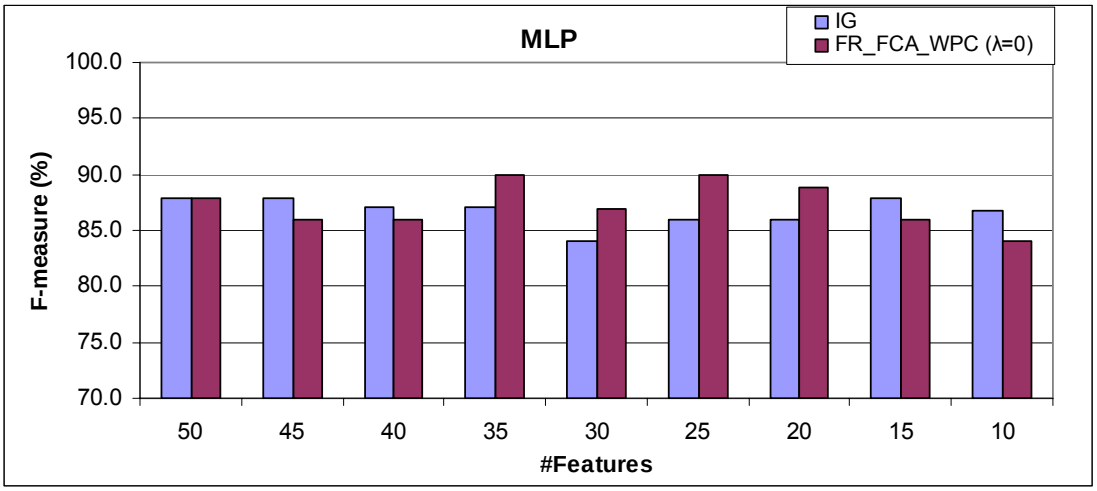
ภาพประกอบ 5.28 เปรียบเทียบผลการจำแนกประเภทด้วย SVM ของชุดข้อมูล F

จากผลการทดลองของชุดข้อมูล F สามารถอธิบายได้ 2 ประเด็นดังนี้

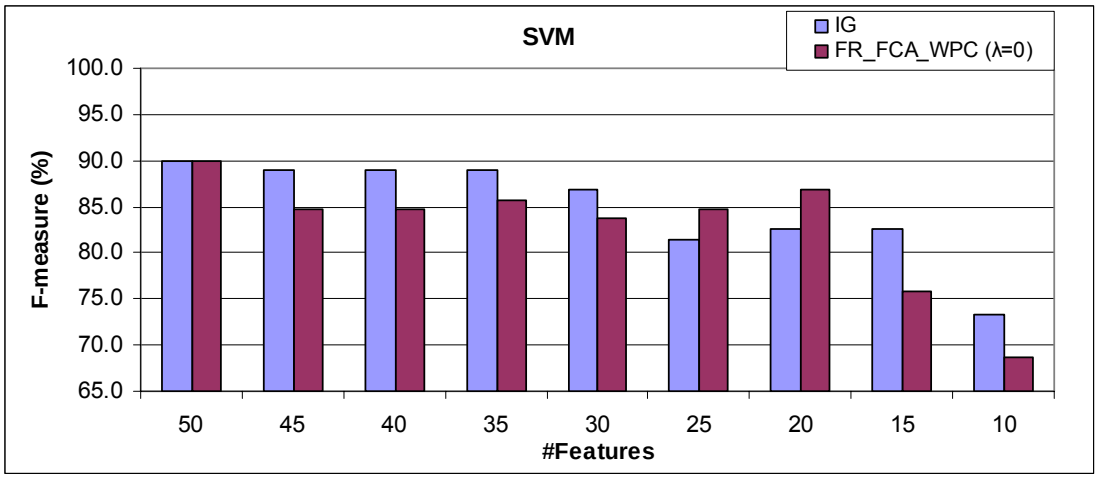
1) ประเด็นเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC กับวิธี IG

ผลการจำแนกประเภทด้วย MLP จากตารางที่ 5.18 และภาพประกอบ 5.27 จะเห็นว่าวิธี FR_FCA_WPC ที่ $\lambda = 0$ และ $\lambda = 0.5$ ให้ค่า F-measure เท่ากัน และที่ขนาดลักษณะเฉพาะเท่ากับ 35 วิธี FR_FCA_WPC ($\lambda = 1.0$) ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 92% ในขณะที่วิธี IG ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 87.9% ที่ขนาดลักษณะเฉพาะเท่ากับ 50 45 และ 15 และจากภาพประกอบ 5.29 เมื่อเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จะเห็นว่าวิธี FR_FCA_WPC ($\lambda = 0$) ให้ค่า F-measure สูงกว่าวิธี IG ที่ขนาดลักษณะเฉพาะเท่ากับ 35 30 25 และ 20 โดยเฉพาะที่ขนาดลักษณะเฉพาะเท่ากับ 25 วิธี FR_FCA_WPC ($\lambda = 0$) ยังให้ค่า F-measure สูงถึง 90% แต่วิธี IG ให้ค่า F-measure เพียงแค่ 85.9%

ผลการจำแนกประเภทด้วย SVM จากตารางที่ 5.19 และภาพประกอบ 5.28 จะเห็นว่าที่ขนาดลักษณะเฉพาะเท่ากับ 50 วิธี FR_FCA_WPC ($\lambda = 0$ และ $\lambda = 0.5$) และวิธี IG ให้ค่า F-measure ที่สูงที่สุดเท่ากันคือ 89.9% และจากภาพประกอบ 5.30 เมื่อเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จะเห็นว่าวิธี FR_FCA_WPC ($\lambda = 0$) ให้ค่า F-measure สูงกว่าวิธี IG ที่ขนาดลักษณะเฉพาะเท่ากับ 25 และ 20 เช่นที่ขนาดลักษณะเฉพาะเท่ากับ 20 วิธี FR_FCA_WPC ($\lambda = 0$) ให้ค่า F-measure เท่ากับ 86.9% แต่วิธี IG ให้ค่า F-measure เท่ากับ 82.6%



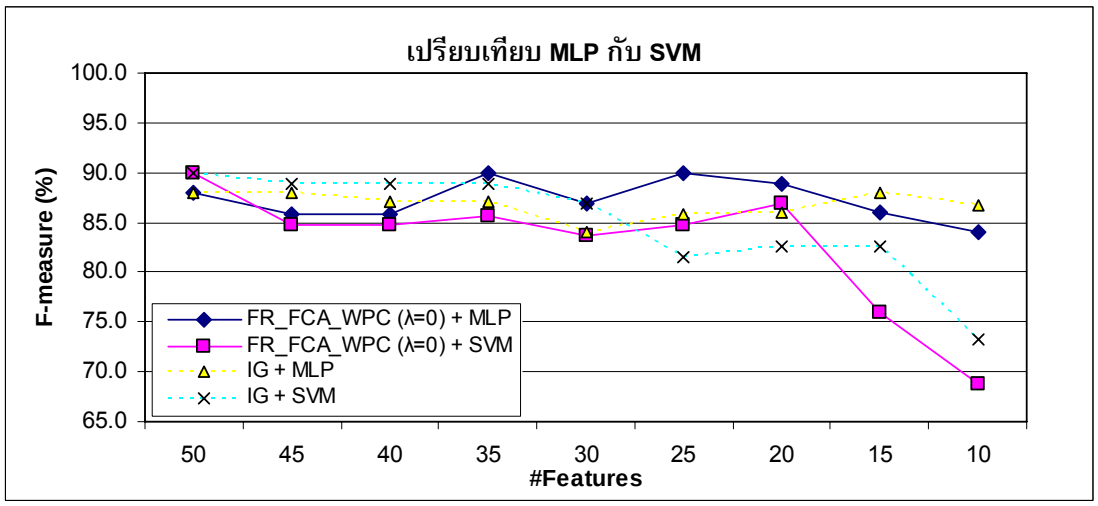
ภาพประกอบ 5.29 เปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จำแนกประเภทด้วย MLP ของชุดข้อมูล F



ภาพประกอบ 5.30 เปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) กับวิธี IG จำแนกประเภทด้วย SVM ของชุดข้อมูล F

2) ประเด็นเปรียบเทียบตัวจำแนกประเภท

จากภาพประกอบ 5.31 เมื่อเปรียบเทียบค่า F-measure จำแนกประเภทด้วย MLP และ SVM ลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) และวิธี IG พบว่าเมื่อลดขนาดลักษณะเฉพาะลง MLP ยังคงให้ค่า F-measure ที่สูงกว่า SVM เช่นที่ขนาดลักษณะเฉพาะเท่ากับ 25 วิธี FR_FCA_WPC ($\lambda = 0$) ที่จำแนกประเภทด้วย MLP ให้ค่า F-measure สูงถึง 90% ซึ่งสูงกว่าที่จำแนกด้วย SVM



ภาพประกอบ 5.31 เปรียบเทียบค่า F-measure จำแนกประเภทด้วย MLP และ SVM ลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ($\lambda = 0$) และวิธี IG ของชุดข้อมูล F

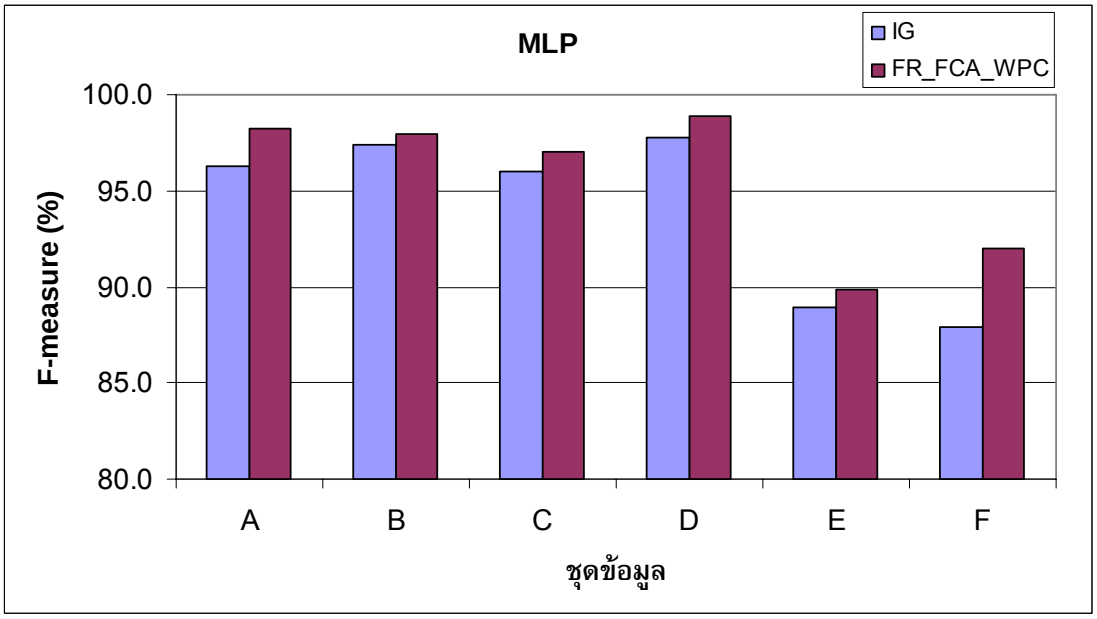
5.2.7 เปรียบเทียบผลการทดลองและวิจารณ์

5.2.7.1 เปรียบเทียบค่า F-measure ที่สูงที่สุด

จากผลการทดลองของชุดข้อมูลทั้ง 6 ชุด เมื่อเปรียบเทียบค่า F-measure ที่สูงที่สุดระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC และวิธี IG จำแนกประเภทด้วย MLP แสดงดังตารางที่ 5.20 และกราฟเปรียบเทียบดังภาพประกอบ 5.32

ตารางที่ 5.20 เปรียบเทียบค่า F-measure ที่สูงที่สุดระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC และวิธี IG จำแนกประเภทด้วย MLP

ชุดข้อมูล	ผลการจำแนกประเภทด้วย MLP				
	IG_MLP		FR_FCA_WPC_MLP		
	ค่า F-measure ที่สูงที่สุด (%)	จำนวนลักษณะเฉพาะ (#Features)	ค่า F-measure ที่สูงที่สุด (%)	จำนวนลักษณะเฉพาะ (#Features)	ค่า λ
A	96.3	50	**98.2**	**40**	1.5
B	97.4	50	**98.0**	**35**	0.5
C	96.0	**10**	**97.0**	15	0
D	97.8	30	**98.9**	**10**	0
E	88.9	50	**89.9**	**20**	0
F	87.9	**15**	**92.0**	35	1.0



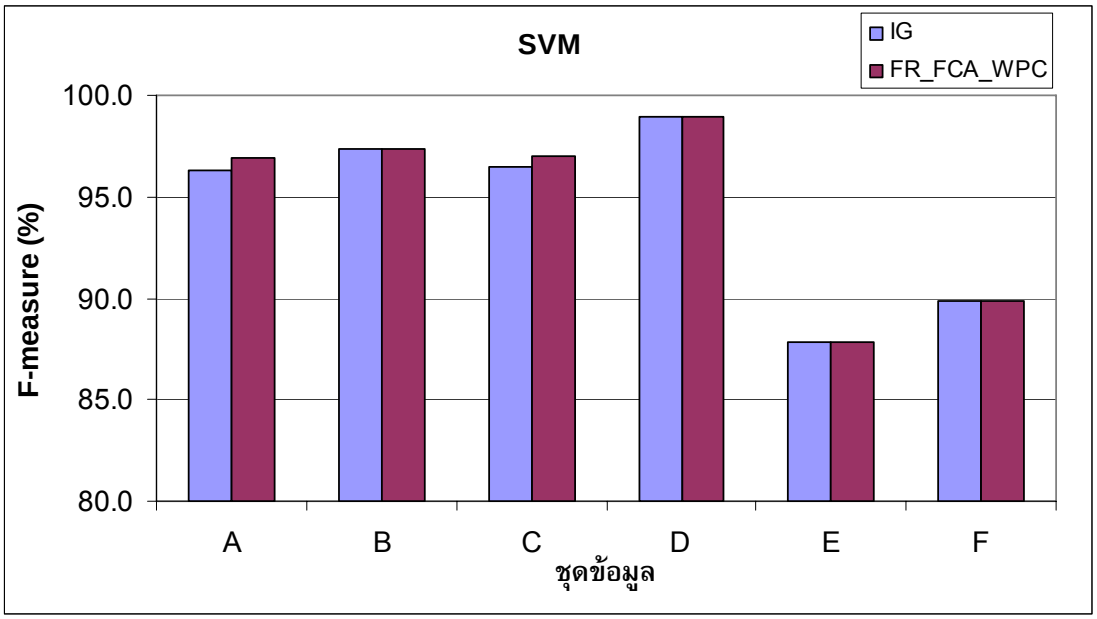
ภาพประกอบ 5.32 กราฟเปรียบเทียบค่า F-measure ที่สูงที่สุดระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC และวิธี IG จำแนกประเภทด้วย MLP

ผลการจำแนกด้วย MLP จากตารางที่ 5.20 และภาพประกอบ 5.32 เมื่อเปรียบเทียบค่า F-measure ที่สูงที่สุดระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC และวิธี IG ของชุดข้อมูลทั้ง 6 พบว่า วิธี FR_FCA_WPC ให้ค่า F-measure สูงกว่าวิธี IG ในทุกชุดข้อมูล โดยเฉพาะในชุดข้อมูล A B D และ E วิธี FR_FCA_WPC สามารถลดขนาดลักษณะเฉพาะได้ดีกว่าและยังให้ค่า F-measure ที่สูงกว่าวิธี IG เช่นในชุดข้อมูล A วิธี FR_FCA_WPC ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 98.2% ซึ่งสูงกว่าวิธี IG ที่ให้ค่า F-measure ที่สูงที่สุดเพียง 96.3% และวิธี FR_FCA_WPC สามารถลดจำนวนลักษณะเฉพาะลงเหลือ 40 ในขณะที่วิธี IG ลดจำนวนลักษณะเฉพาะลงเหลือ 50 ซึ่งใช้จำนวนลักษณะเฉพาะนำเข้ามากกว่า เป็นต้น และเมื่อปรับค่า λ เพิ่มขึ้น พบว่าวิธี FR_FCA_WPC ยังสามารถให้ค่า F-measure ที่สูงที่สุดได้ดังในชุดข้อมูล A B และ F วิธี FR_FCA_WPC ที่ $\lambda = 1.5$ $\lambda = 0.5$ และ $\lambda = 1.0$ ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 98.2% 98% และ 92% ตามลำดับ

ในทำนองเดียวกันจากผลการทดลองของชุดข้อมูลทั้ง 6 ชุด เมื่อเปรียบเทียบค่า F-measure ที่สูงที่สุดระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC และวิธี IG จำแนกประเภทด้วย SVM แสดงดังตารางที่ 5.21 และกราฟเปรียบเทียบดังภาพประกอบ 5.33

ตารางที่ 5.21 เปรียบเทียบค่า F-measure ที่สูงที่สุดระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC และวิธี IG จำแนกประเภทด้วย SVM

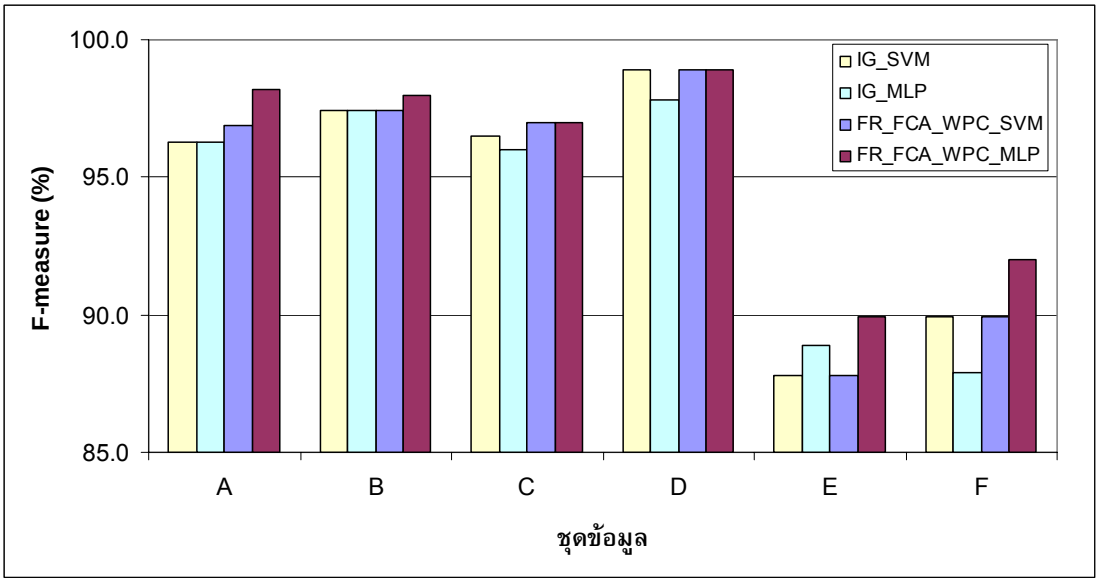
ชุดข้อมูล	ผลการจำแนกประเภทด้วย SVM				
	IG_SVM		FR_FCA_WPC_SVM		
	ค่า F-measure ที่สูงที่สุด (%)	จำนวนลักษณะเฉพาะ (#Features)	ค่า F-measure ที่สูงที่สุด (%)	จำนวนลักษณะเฉพาะ (#Features)	ค่า λ
A	96.3	50	**96.9**	**40**	1.5
B	**97.4**	**50**	**97.4**	**50**	0
C	96.5	50	**97.0**	**10**	0
D	**98.9**	45	**98.9**	**30**	1.0
E	**87.8**	45	**87.8**	**40**	0
F	**89.9**	**50**	**89.9**	**50**	0



ภาพประกอบ 5.33 กราฟเปรียบเทียบค่า F-measure ที่สูงที่สุดระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC และวิธี IG จำแนกประเภทด้วย SVM

ผลการจำแนกประเภทด้วย SVM จากตารางที่ 5.21 และภาพประกอบ 5.33 เมื่อเปรียบเทียบค่า F-measure ที่สูงที่สุดระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC และวิธี IG ของชุดข้อมูลทั้ง 6 พบว่าวิธี FR_FCA_WPC และวิธี IG ให้ค่า F-measure ที่สูงที่สุดใกล้เคียงกัน แต่ในชุดข้อมูล A และ C วิธี FR_FCA_WPC สามารถลดขนาดลักษณะเฉพาะได้ดีกว่าและยังให้ค่า F-measure ที่สูงกว่าวิธี IG เช่นในชุดข้อมูล A วิธี FR_FCA_WPC ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 96.9% ซึ่งสูงกว่าวิธี IG ที่ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 96.3% และวิธี FR_FCA_WPC สามารถลดจำนวนลักษณะเฉพาะลงเหลือ 40 ในขณะที่วิธี IG ลดจำนวนลักษณะเฉพาะลงเหลือ 50 ซึ่งใช้จำนวนลักษณะเฉพาะนำเข้ามากกว่า เป็นต้น ส่วนในชุดข้อมูล B D E และ F วิธี FR_FCA_WPC และวิธี IG ให้ค่า F-measure ที่สูงที่สุดเท่ากัน แต่ในชุดข้อมูล D และ E วิธี FR_FCA_WPC สามารถลดขนาดลักษณะเฉพาะได้ดีกว่าวิธี IG เช่นในชุดข้อมูล D วิธี FR_FCA_WPC สามารถลดจำนวนลักษณะเฉพาะลงเหลือ 30 ในขณะที่วิธี IG ลดจำนวนลักษณะเฉพาะลงเหลือ 45 ซึ่งใช้จำนวนลักษณะเฉพาะนำเข้ามากกว่า เป็นต้น และเมื่อปรับค่า λ เพิ่มขึ้น วิธี FR_FCA_WPC ยังสามารถให้ค่า F-measure ที่สูงที่สุดได้ดังในชุดข้อมูล A และ D วิธี FR_FCA_WPC ที่ $\lambda = 1.5$ และ $\lambda = 1.0$ ให้ค่า F-measure ที่สูงที่สุดเท่ากับ 96.9% และ 98.9% ตามลำดับ

จากภาพประกอบ 5.34 เมื่อเปรียบเทียบตัวจำแนกประเภทระหว่าง MLP และ SVM จะเห็นว่าในชุดข้อมูล A B E และ F การลดขนาดลักษณะด้วยวิธี FR_FCA_WPC ที่จำแนกประเภทด้วย MLP (FR_FCA_WPC_MLP) ให้ค่า F-measure ที่สูงที่สุดเมื่อเปรียบเทียบกับการลดขนาดลักษณะด้วยวิธี FR_FCA_WPC ที่จำแนกประเภทด้วย SVM (FR_FCA_WPC_SVM) การลดขนาดลักษณะด้วยวิธี IG ที่จำแนกประเภทด้วย MLP (IG_MLP) และการลดขนาดลักษณะด้วยวิธี IG ที่จำแนกประเภทด้วย SVM (IG_SVM) ส่วนในชุดข้อมูล C จะเห็นว่า FR_FCA_WPC_MLP และ FR_FCA_WPC_SVM ให้ค่า F-measure ที่สูงที่สุดเท่ากัน และในชุดข้อมูล D จะเห็นว่า FR_FCA_WPC_MLP FR_FCA_WPC_SVM และ IG_SVM ให้ค่า F-measure ที่สูงที่สุดเท่ากัน ดังนั้นสามารถสรุปได้ว่าการลดขนาดลักษณะด้วยวิธี FR_FCA_WPC ที่จำแนกประเภทด้วย MLP ให้ค่า F-measure ที่มีแนวโน้มสูงที่สุด



ภาพประกอบ 5.34 กราฟเปรียบเทียบค่า F-measure ที่สูงที่สุด ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC และวิธี IG จำแนกประเภทด้วย MLP และ SVM

5.2.7.2 การพิจารณาค่า λ ที่เหมาะสมในการทดลอง

จากการคำนวณหาค่าน้ำหนักเฉลี่ยระหว่างลักษณะเฉพาะกับเอกสารเว็บเพจในแต่ละชุดข้อมูล (A ถึง F) ที่เลือกลักษณะเฉพาะโดยใช้ IG สามารถสรุปได้ดังตารางที่ 5.22 กำหนดให้ *m* คือจำนวนลักษณะเฉพาะที่เลือกด้วย IG *n* คือจำนวนเอกสารเว็บเพจ และค่าน้ำหนักเฉลี่ย (Mean) คำนวณได้จากผลรวมของค่าน้ำหนัก (TF-IDF) ทั้งหมดทุกเซลล์หารด้วยผลคูณระหว่างจำนวนลักษณะเฉพาะกับจำนวนเอกสารเว็บเพจ (*m*n*) ตัวอย่างเช่น ชุดข้อมูล A มีจำนวนลักษณะเฉพาะเท่ากับ 50 ลักษณะเฉพาะ จำนวนเอกสารเว็บเพจเท่ากับ 163 เว็บเพจ จำนวนเซลล์ทั้งหมดเท่ากับ (50*163) = 8,150 เซลล์ และผลรวม

ของค่าน้ำหนักทั้งหมดทุกเซลล์เท่ากับ 6,142.8 ดังนั้นสามารถคำนวณค่าน้ำหนักเฉลี่ยได้เท่ากับ 6,142.8 / 8,150 = 0.8 เป็นต้น

ตารางที่ 5.22 ค่าน้ำหนักเฉลี่ยระหว่างลักษณะเฉพาะกับเอกสารเว็บเพจในแต่ละชุดข้อมูล (A ถึง F) ที่เลือกลักษณะเฉพาะเฉพาะโดยใช้ IG

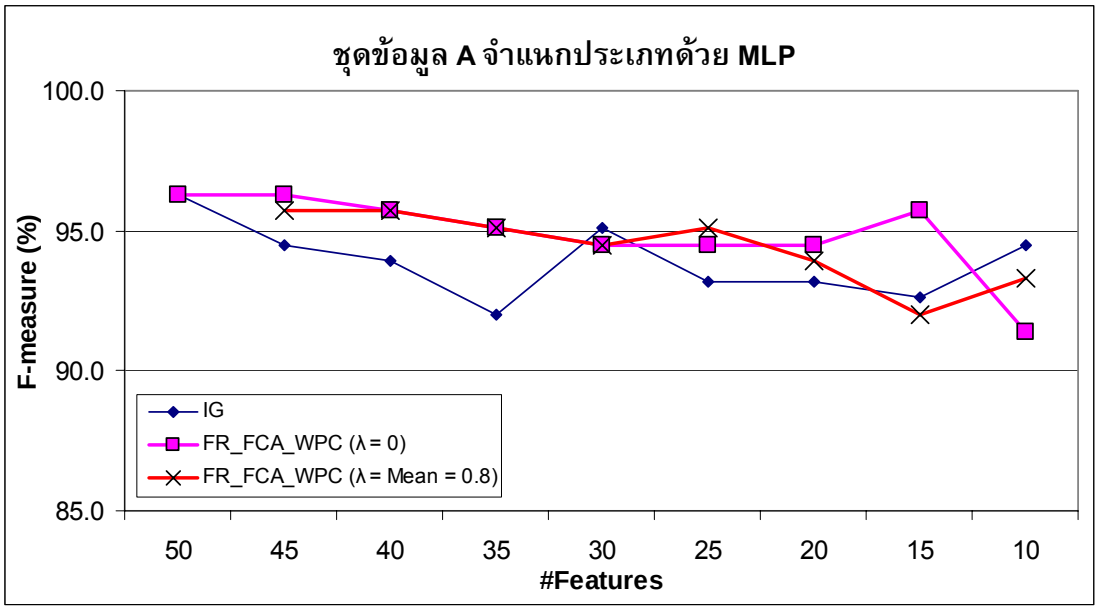
ผลการทดลอง	ชุดข้อมูล					
	A	**B**	**C**	**D**	**E**	**F**
จำนวนลักษณะเฉพาะ (*m*)	50	50	50	50	50	50
จำนวนเอกสารเว็บเพจ (*n*)	163	151	200	90	100	100
จำนวนเซลล์ทั้งหมด (*mn)**	8,150	7,550	10,000	4,500	5,000	5,000
จำนวนเซลล์ที่มีค่าน้ำหนักเท่ากับ 0	6,715	5,690	7,943	3,473	3,593	3,382
จำนวนเซลล์ที่มีค่าน้ำหนักมากกว่า 0	1,435	1,910	2,057	1,027	1,407	1,618
ผลรวมของค่าน้ำหนักทั้งหมดทุกเซลล์	6,142.8	8,780.0	7,987.2	2,877.2	27,346.5	19,676.0
ค่าน้ำหนักเฉลี่ย (Mean)	**0.8**	**1.2**	**0.8**	**0.6**	**5.5**	**3.9**

จากการสังเกตข้อมูลเบื้องต้นพบว่าจำนวนเซลล์ที่มีค่าน้ำหนักเท่ากับ 0 มีค่อนข้างมากหรือเกินร้อยละ 50 ของจำนวนเซลล์ทั้งหมด ตัวอย่างเช่น จากตารางที่ 5.22 ชุดข้อมูล A มีจำนวนเซลล์ทั้งหมดเท่ากับ 8,150 เซลล์ จำนวนเซลล์ที่มีค่าน้ำหนักเท่ากับ 0 จำนวน 6,715 เซลล์ และจำนวนเซลล์ที่มีค่าน้ำหนักมากกว่า 0 จำนวน 1,435 เซลล์ เป็นต้น เมื่อจำนวนเซลล์ที่มีค่าน้ำหนักเท่ากับ 0 มีจำนวนมากทำให้ Mean มีค่าต่ำหรือเข้าใกล้ 0

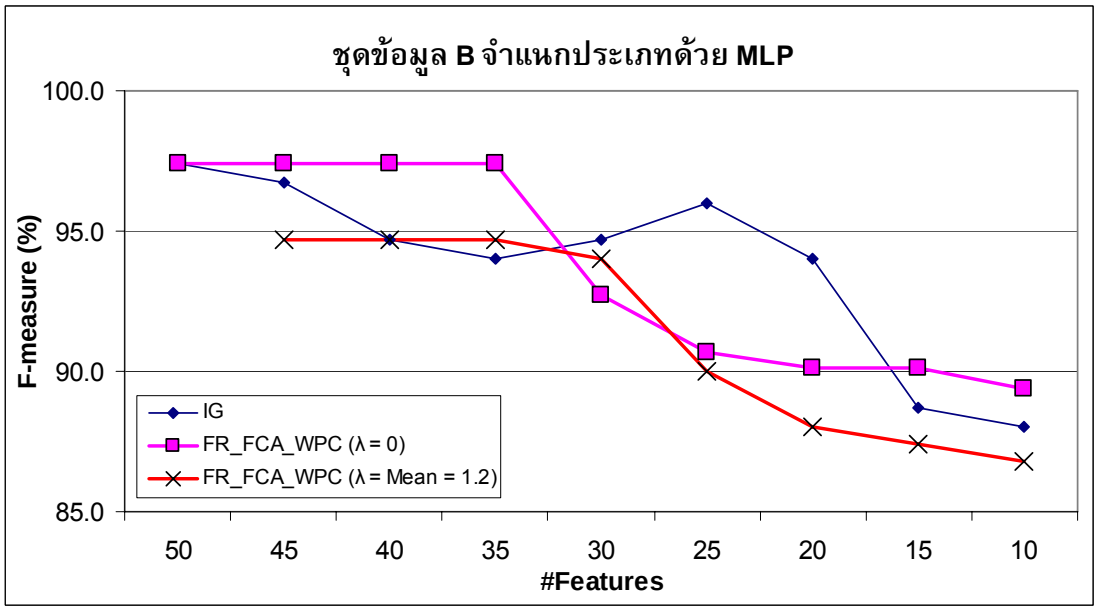
ในการทดลองกรณีที่ 1 กำหนดให้ λ = 0 หมายถึงเลือกความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารเว็บเพจที่มีค่าน้ำหนักมากกว่า 0 กรณีที่ 2 กำหนดให้ λ = Mean หมายถึงเลือกความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารเว็บเพจที่มีค่าน้ำหนักมากกว่าค่า Mean ซึ่งผลการทดลองเมื่อกำหนดค่า λ = 0 และ λ = Mean เปรียบเทียบกับวิธี IG จำแนกประเภทด้วย MLP ของชุดข้อมูล A ถึง F แสดงดังตารางที่ 5.23 และกราฟเปรียบเทียบดังภาพประกอบ 5.35 - 5.40 พบว่าวิธี FR_FCA_WPC ที่ λ = 0 ให้ค่า F-measure ที่มีแนวโน้มสูงกว่าวิธี IG ในทุกชุดข้อมูล และที่ λ = Mean วิธี FR_FCA_WPC ให้ค่า F-measure ที่มีแนวโน้มสูงกว่าวิธี IG ในชุดข้อมูล A และ F ตัวอย่างเช่น จากภาพประกอบ 5.35 ชุดข้อมูล A วิธี FR_FCA_WPC ที่ λ = 0 และ λ = Mean (0.8) ส่วนใหญ่ให้ค่า F-measure ที่สูงกว่าวิธี IG เป็นต้น

ตารางที่ 5.23 เปรียบเทียบค่า F-measure จำแนกประเภทด้วย MLP ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC (λ = 0 และ λ = Mean) กับวิธี IG ของชุดข้อมูล A ถึง F

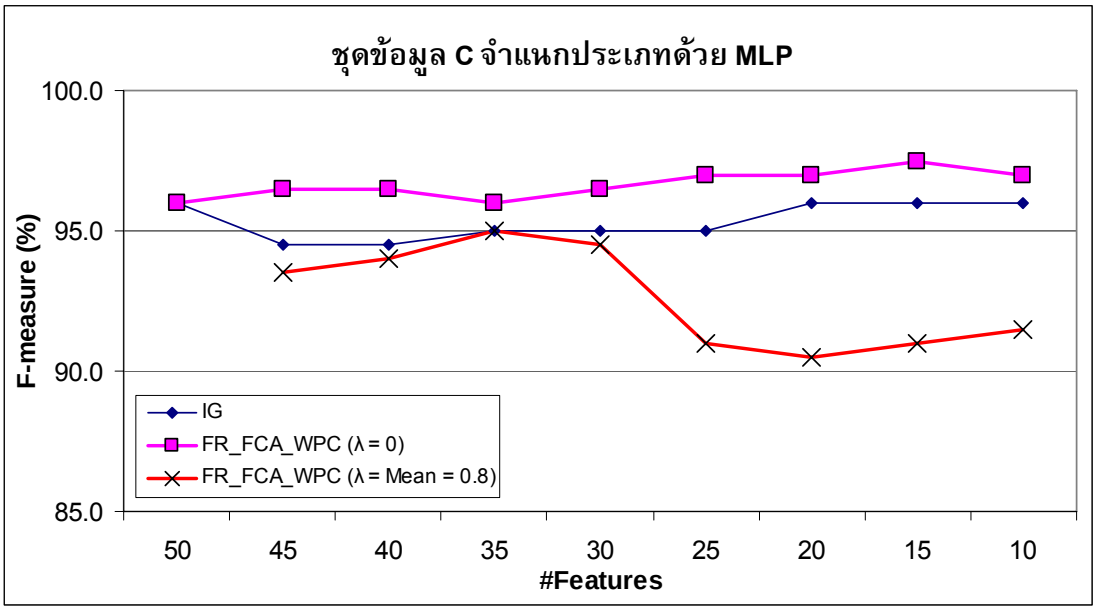
จำนวนลักษณะเฉพาะ (#Features)	ค่า F-measure (%) จำแนกประเภทด้วย MLP																	
	ชุดข้อมูล A			ชุดข้อมูล B			ชุดข้อมูล C			ชุดข้อมูล D			ชุดข้อมูล E			ชุดข้อมูล F		
	IG	FR_FCA_WPC		IG	FR_FCA_WPC		IG	FR_FCA_WPC		IG	FR_FCA_WPC		IG	FR_FCA_WPC		IG	FR_FCA_WPC	
		λ = 0	λ = 0.8 (Mean)		λ = 0	λ = 1.2 (Mean)		λ = 0	λ = 0.8 (Mean)		λ = 0	λ = 0.6 (Mean)		λ = 0	λ = 5.5 (Mean)		λ = 0	λ = 3.9 (Mean)
50	***96.3***	***96.3***	-	***97.4***	***97.4***	-	***96.0***	**96.0**	-	***97.8***	**97.8**	-	***88.9***	**88.9**	-	***87.9***	**87.9**	-
45	94.5	***96.3***	**95.7**	96.7	***97.4***	**94.7**	94.5	**96.5**	93.5	***97.8***	**97.8**	**97.8**	87.9	**87.9**	**88.0**	87.9	85.9	**88.9**
40	93.9	**95.7**	**95.7**	94.7	***97.4***	**94.7**	94.5	**96.5**	94.0	***97.8***	96.6	96.6	87.0	**88.9**	83.9	87.0	85.9	**88.9**
35	92.0	**95.1**	**95.1**	94.0	***97.4***	**94.7**	95.0	**96.0**	**95.0**	***97.8***	96.6	96.6	86.0	**88.9**	85.9	87.0	***90.0***	**88.0**
30	**95.1**	94.5	94.5	**94.7**	92.7	94.0	95.0	**96.5**	94.5	***97.8***	**97.8**	**97.8**	86.0	**87.9**	85.9	84.0	**86.9**	**90.0**
25	93.2	**94.5**	**95.1**	**96.0**	90.7	90.0	95.0	**97.0**	91.0	96.7	**97.8**	***98.9***	86.0	***89.9***	85.9	85.9	***90.0***	85.0
20	93.2	**94.5**	**93.9**	**94.0**	90.1	88.0	*96.0*	**97.0**	90.5	94.4	**96.7**	**95.5**	86.0	***89.9***	84.9	86.0	**88.9**	85.0
15	92.6	**95.7**	92.0	88.7	**90.1**	87.4	*96.0*	***97.5***	91.0	94.4	***98.9***	93.3	**86.0**	80.9	81.9	87.9	86.0	**88.0**
10	**94.5**	91.4	93.3	88.0	**89.4**	86.8	*96.0*	**97.0**	91.5	94.4	***98.9***	**97.8**	**85.9**	82.8	82.9	86.8	84.0	**89.9**



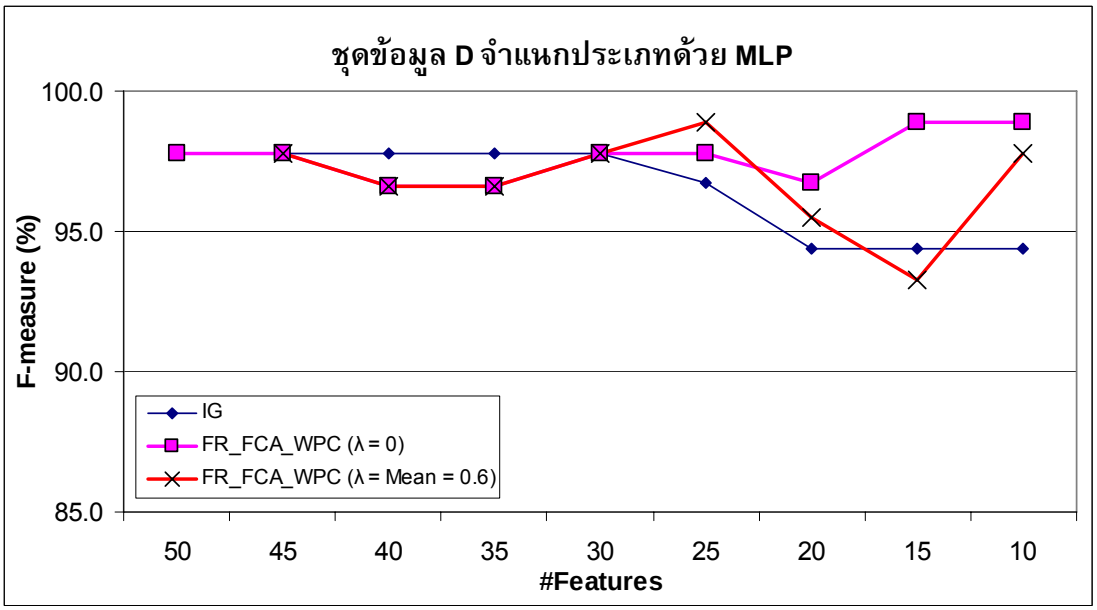
ภาพประกอบ 5.35 กราฟเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC (λ = 0 และ λ = Mean) กับวิธี IG จำแนกประเภทด้วย MLP ของชุดข้อมูล A



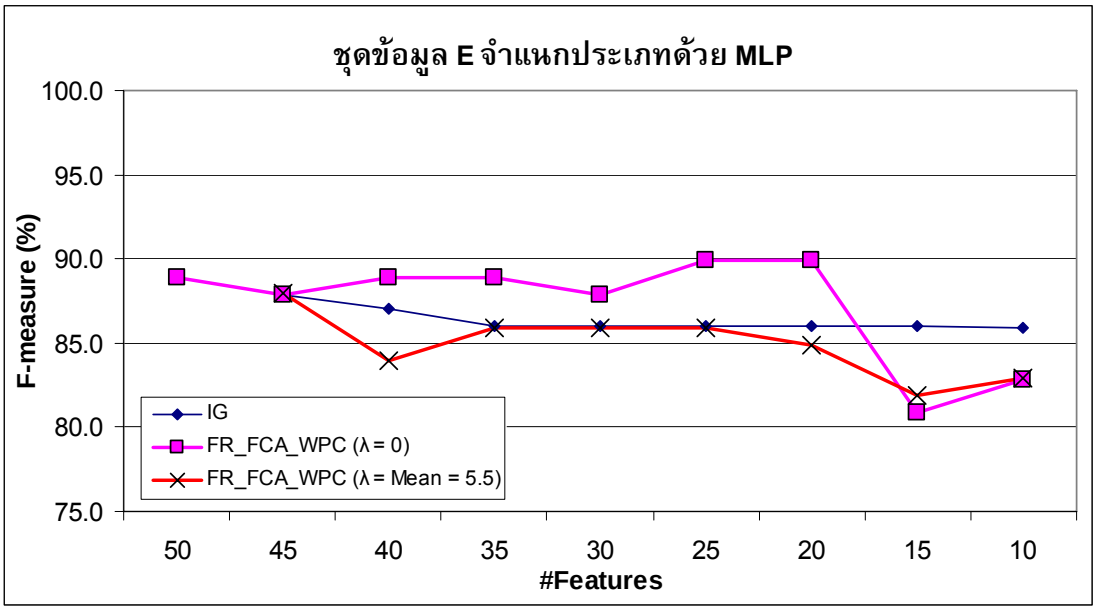
ภาพประกอบ 5.36 กราฟเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC (λ = 0 และ λ = Mean) กับวิธี IG จำแนกประเภทด้วย MLP ของชุดข้อมูล B



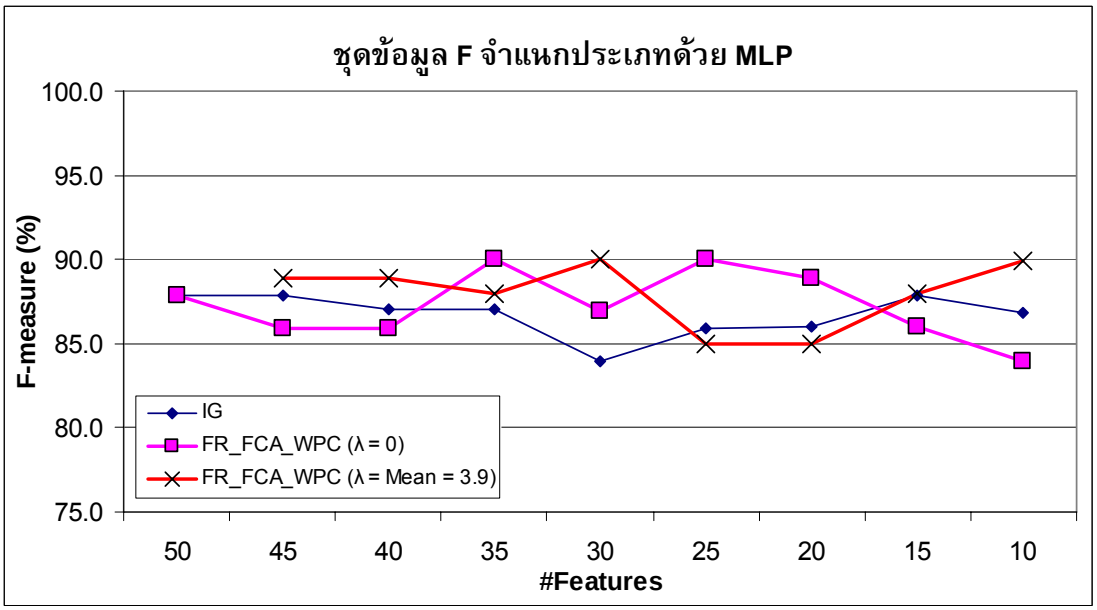
ภาพประกอบ 5.37 กราฟเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC (λ = 0 และ λ = Mean) กับวิธี IG จำแนกประเภทด้วย MLP ของชุดข้อมูล C



ภาพประกอบ 5.38 กราฟเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC (λ = 0 และ λ = Mean) กับวิธี IG จำแนกประเภทด้วย MLP ของชุดข้อมูล D



ภาพประกอบ 5.39 กราฟเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC (λ = 0 และ λ = Mean) กับวิธี IG จำแนกประเภทด้วย MLP ของชุดข้อมูล E



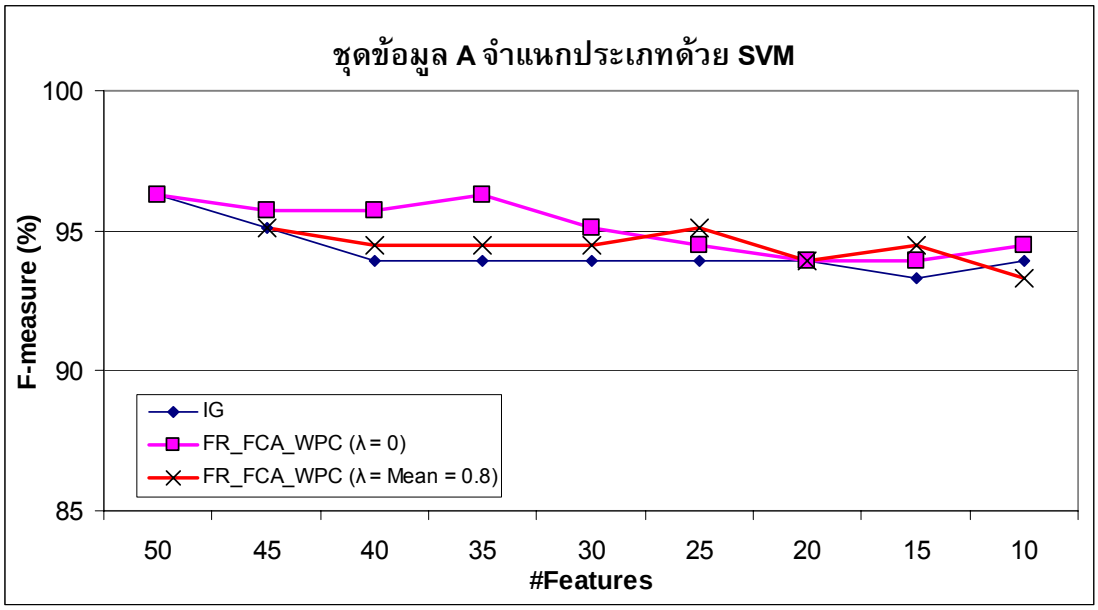
ภาพประกอบ 5.40 กราฟเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC (λ = 0 และ λ = Mean) กับวิธี IG จำแนกประเภทด้วย MLP ของชุดข้อมูล F

ผลการทดลองเมื่อกำหนดค่า λ = 0 และ λ = Mean เปรียบเทียบกับวิธี IG จำแนกประเภทด้วย SVM ของชุดข้อมูล A ถึง F แสดงดังตารางที่ 5.24 และกราฟเปรียบเทียบดังภาพประกอบ 5.41 - 5.46 พบว่าวิธี FR_FCA_WPC ที่ λ = 0 ให้ค่า F-measure ที่มีแนวโน้มสูงกว่าวิธี IG ในชุดข้อมูล A B C D และ E และที่ λ = Mean วิธี FR_FCA_WPC ให้ค่า F-measure ที่มีแนวโน้มสูงกว่าวิธี IG ในชุดข้อมูล A และ D ตัวอย่างเช่น จากภาพประกอบ 5.41 ชุดข้อมูล A วิธี FR_FCA_WPC ที่ λ = 0 และ λ = Mean (0.8) ส่วนใหญ่ให้ค่า F-measure ที่สูงกว่าวิธี IG เป็นต้น

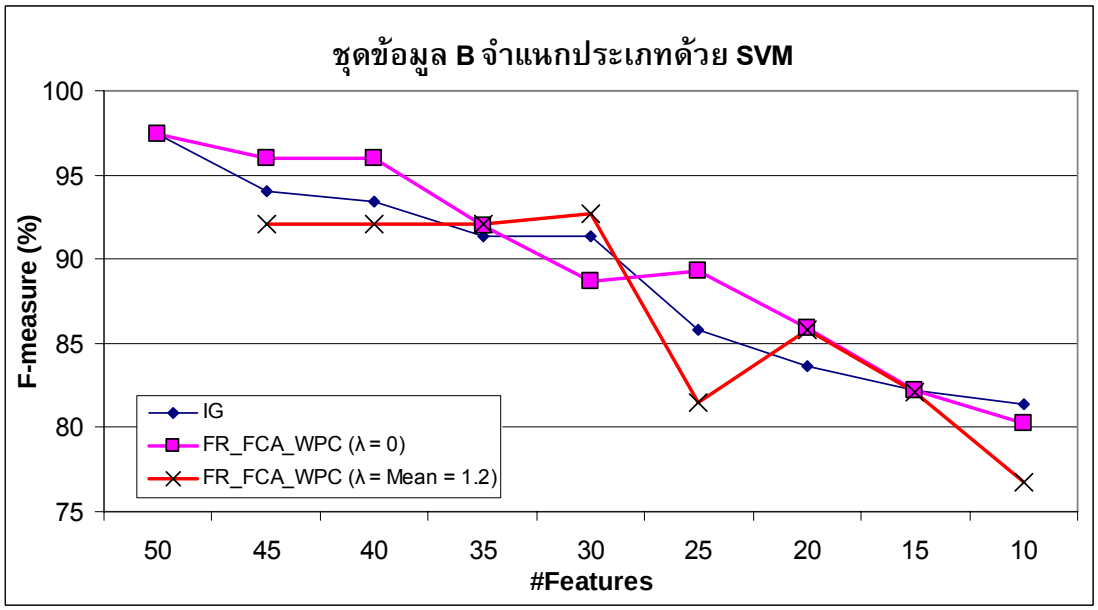
ข้อสังเกต จากผลการทดลองจะเห็นได้ว่าการกำหนดค่า λ = 0 และ λ = Mean วิธี FR_FCA_WPC สามารถลดขนาดลักษณะเฉพาะได้ดีกว่าและยังให้ค่า F-measure ที่สูงกว่าวิธี IG ทั้งการจำแนกประเภทด้วย MLP และ SVM ดังนั้นข้อเสนอแนะในการทดลองคือให้กำหนดค่า λ = 0 เป็นค่าเริ่มต้นและสามารถปรับค่า λ เพิ่มขึ้นได้จนถึงค่า Mean เนื่องจากค่า Mean เป็นค่าน้ำหนักเฉลี่ยระหว่างลักษณะเฉพาะกับเอกสารเว็บเพจซึ่งจะขึ้นอยู่กับชุดข้อมูลต่าง ๆ นั่นเอง

ตารางที่ 5.24 เปรียบเทียบค่า F-measure จำแนกประเภทด้วย SVM ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC (λ = 0 และ λ = Mean) กับวิธี IG ของชุดข้อมูล A ถึง F

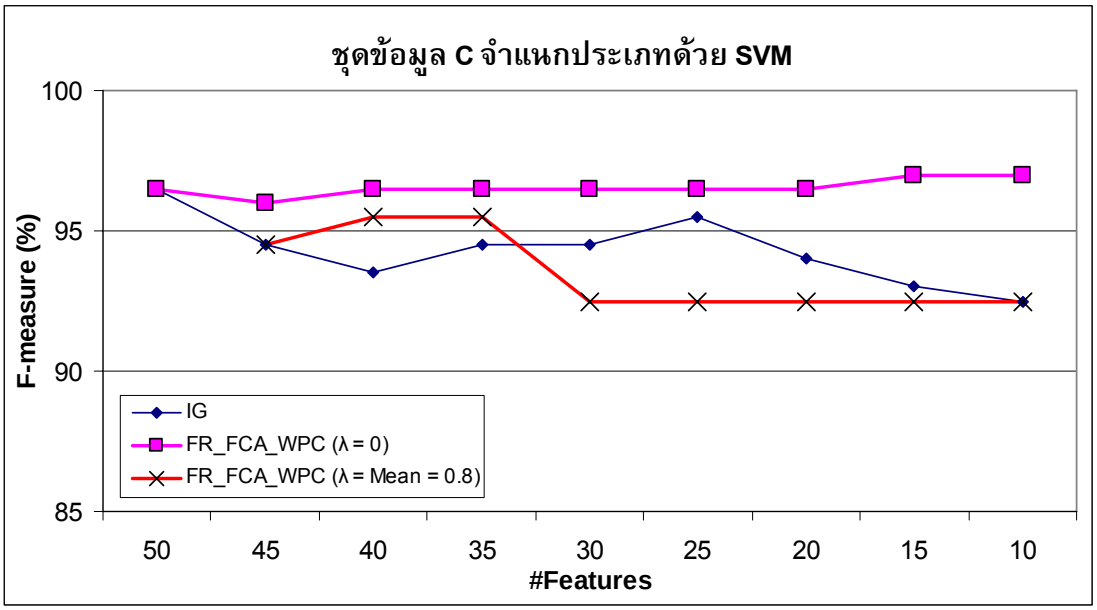
จำนวนลักษณะเฉพาะ (#Features)	ค่า F-measure (%) จำแนกประเภทด้วย SVM																	
	ชุดข้อมูล A			ชุดข้อมูล B			ชุดข้อมูล C			ชุดข้อมูล D			ชุดข้อมูล E			ชุดข้อมูล F		
	IG	FR_FCA_WPC		IG	FR_FCA_WPC		IG	FR_FCA_WPC		IG	FR_FCA_WPC		IG	FR_FCA_WPC		IG	FR_FCA_WPC	
		λ = 0	λ = 0.8 (Mean)		λ = 0	λ = 1.2 (Mean)		λ = 0	λ = 0.8 (Mean)		λ = 0	λ = 0.6 (Mean)		λ = 0	λ = 5.5 (Mean)		λ = 0	λ = 3.9 (Mean)
50	***96.3***	***96.3***	-	***97.4***	***97.4***	-	***96.5***	**96.5**	-	**97.8**	**97.8**	-	***87.8***	***87.8***	-	***89.9***	***89.9***	-
45	95.1	**95.7**	**95.1**	94.0	**96.0**	92.1	94.5	**96.0**	**94.5**	***98.9***	97.8	97.8	***87.8***	***87.8***	83.6	**88.9**	84.7	**88.9**
40	93.9	**95.7**	**94.5**	93.4	**96.0**	92.1	93.5	**96.5**	**95.5**	**97.8**	**97.8**	96.7	86.8	***87.8***	80.3	**88.9**	84.7	85.8
35	93.9	***96.3***	**94.5**	91.4	**92.0**	**92.1**	94.5	**96.5**	**95.5**	97.8	***98.9***	**97.8**	82.5	**86.8**	79.2	**88.9**	85.7	86.8
30	93.9	**95.1**	**94.5**	91.4	88.7	**92.7**	94.5	**96.5**	92.5	96.7	***98.9***	**97.8**	82.5	**86.8**	78.0	**86.9**	83.7	86.8
25	93.9	**94.5**	**95.1**	85.8	**89.3**	81.5	95.5	**96.5**	92.5	93.3	**96.7**	**97.8**	81.4	**86.8**	78.0	81.5	**84.7**	81.0
20	**93.9**	**93.9**	**93.9**	83.6	**85.9**	**85.8**	94.0	**96.5**	92.5	**95.6**	90.0	**95.6**	80.3	**84.7**	75.7	82.6	**86.9**	79.9
15	93.3	**93.9**	**94.5**	**82.2**	**82.2**	82.1	93.0	***97.0***	92.5	93.3	**96.7**	92.2	**81.4**	73.3	73.3	**82.6**	75.9	79.3
10	93.9	**94.5**	93.3	81.4	80.2	76.8	92.5	***97.0***	**92.5**	94.4	91.1	**95.6**	**76.9**	72.1	64.3	73.3	68.7	**80.0**



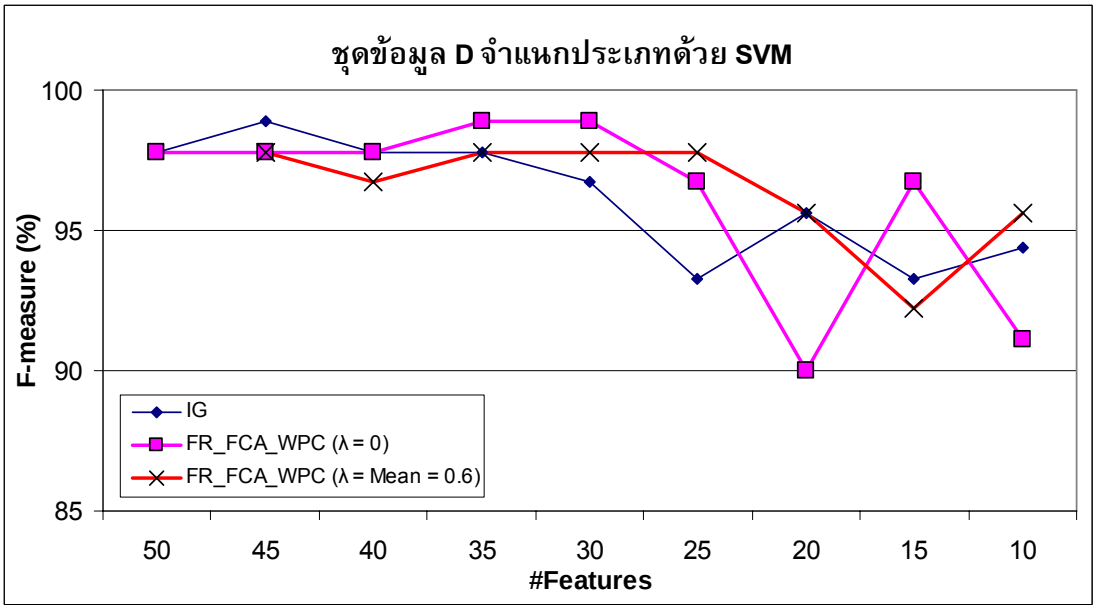
ภาพประกอบ 5.41 กราฟเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC (λ = 0 และ λ = Mean) กับวิธี IG จำแนกประเภทด้วย SVM ของชุดข้อมูล A



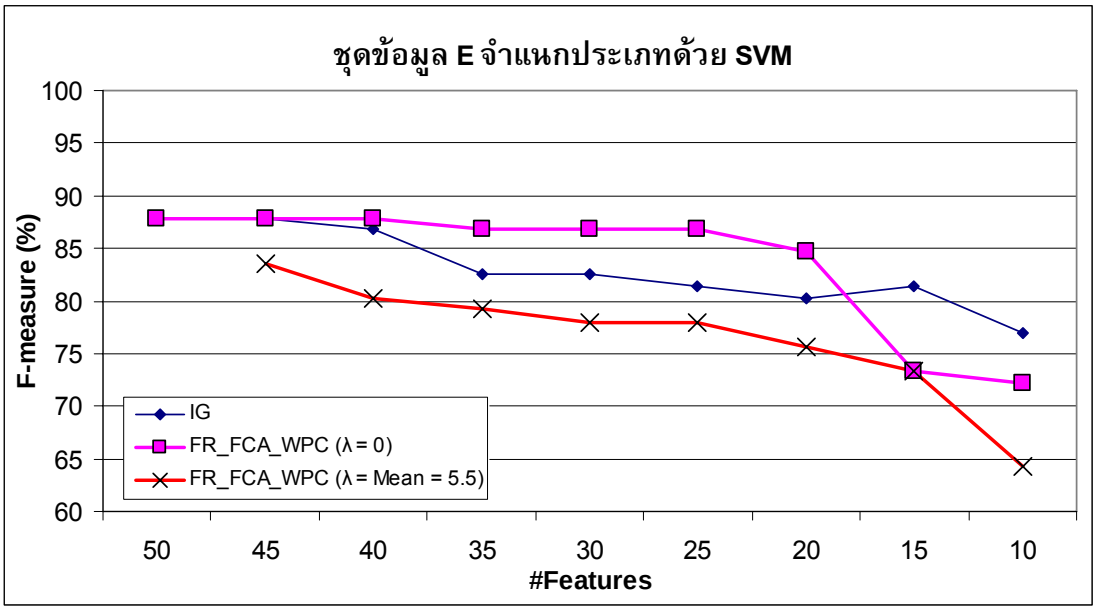
ภาพประกอบ 5.42 กราฟเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC (λ = 0 และ λ = Mean) กับวิธี IG จำแนกประเภทด้วย SVM ของชุดข้อมูล B



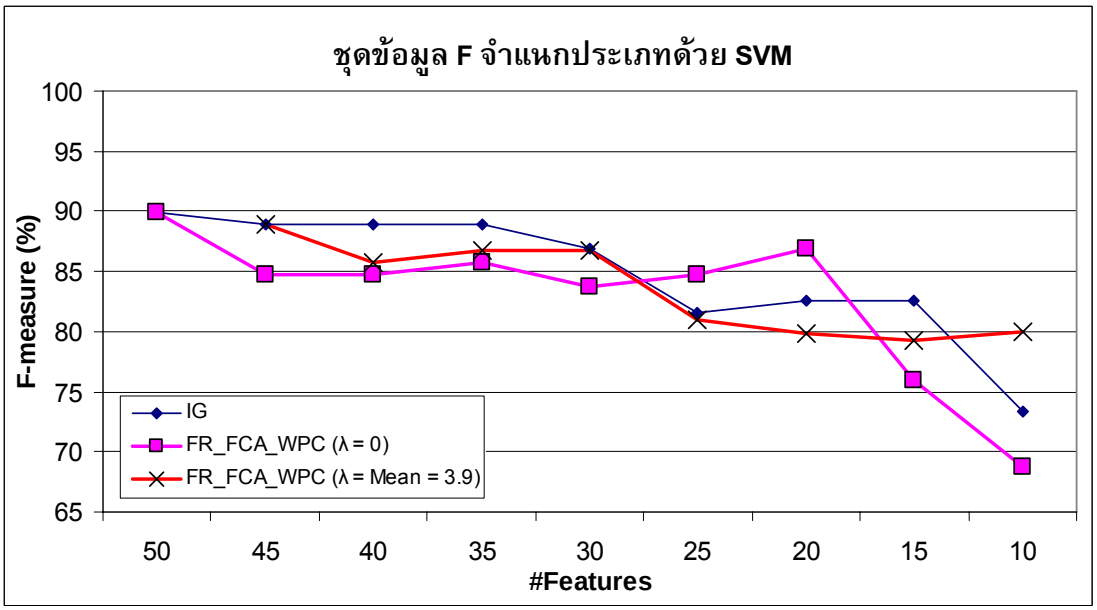
ภาพประกอบ 5.43 กราฟเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC (λ = 0 และ λ = Mean) กับวิธี IG จำแนกประเภทด้วย SVM ของชุดข้อมูล C



ภาพประกอบ 5.44 กราฟเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC (λ = 0 และ λ = Mean) กับวิธี IG จำแนกประเภทด้วย SVM ของชุดข้อมูล D



ภาพประกอบ 5.45 กราฟเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC (λ = 0 และ λ = Mean) กับวิธี IG จำแนกประเภทด้วย SVM ของชุดข้อมูล E



ภาพประกอบ 5.46 กราฟเปรียบเทียบค่า F-measure ระหว่างการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC (λ = 0 และ λ = Mean) กับวิธี IG จำแนกประเภทด้วย SVM ของชุดข้อมูล F

บทที่ 6

บทสรุปและข้อเสนอแนะ

6.1 สรุปผลการวิจัย

งานวิจัยนี้ได้นำเสนอแนวคิดใหม่ในการลดขนาดลักษณะเฉพาะโดยใช้วิธีการ Formal Concept Analysis (FCA) สำหรับการจำแนกประเภทเว็บเพจ โดยสร้างแบบจำลองการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ (Feature Reduction using FCA for Web Page Classification: FR_FCA_WPC) งานวิจัยนี้ได้ใช้วิธีการเลือกลักษณะเฉพาะโดยใช้ Information Gain (IG) เปรียบเทียบกับวิธี FR_FCA_WPC ที่นำเสนอ โดยทดสอบกับชุดข้อมูลเว็บเพจมาตรฐานจาก CMU จำนวน 2 ชุด คือ ชุดข้อมูล 7Sectors และชุดข้อมูล BankResearch จำแนกประเภทเว็บเพจด้วย Multi-Layer Perceptron Neural Networks (MLP) และ Support Vector Machine (SVM) ทดสอบแบบ 10-folds Cross Validation และประเมินผลด้วยค่า F-measure

ผู้ทำการวิจัยได้พัฒนาโปรแกรมตามแบบจำลองที่นำเสนอเพื่อลดขนาดลักษณะเฉพาะสำหรับการจำแนกประเภทเว็บเพจจากชุดข้อมูลเว็บเพจดังกล่าว ซึ่งผู้ใช้สามารถใช้งานได้ง่ายด้วยส่วนติดต่อผู้ใช้ในรูปแบบกราฟิก (Graphic User Interface) โปรแกรมการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจนี้ได้พัฒนาด้วย Visual C#.Net ทำงานร่วมกับโปรแกรม Concept Explorer (ConExp) ในการวิเคราะห์หาความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารเว็บเพจ สำหรับเลือกลักษณะเฉพาะด้วย FCA และโปรแกรม WEKA แบบ Command Line Interface ดังภาคผนวก ก สำหรับเลือกลักษณะเฉพาะด้วย IG และจำแนกประเภท

ผลการทดลองของงานวิจัยนี้ในการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับจำแนกประเภทเว็บเพจ เรื่อง “การจำแนกประเภทเว็บเพจโดยวิธีการลดขนาดลักษณะเฉพาะด้วย FCA” ได้รับการตีพิมพ์ใน The 7th International Joint Conference on Computer Science and Software Engineering (JCSSE 2010) วันที่ 12-14 พฤษภาคม 2553 ดังภาคผนวก ข และผลการทดลองเกี่ยวกับการปรับค่าน้ำหนักของลักษณะเฉพาะที่สัมพันธ์กับเอกสารเว็บเพจเพื่อเลือกลักษณะเฉพาะโดยใช้ FCA เรื่อง “Feature Reduction Using Formal Concept Analysis for Web Page Classification” ได้รับการตีพิมพ์ใน The 5th PSU-UNS International Conference on Engineering and Technology (ICET 2011) วันที่ 2-3 พฤษภาคม 2554 ดังภาคผนวก ค

ผลการทดลองตามแบบจำลองการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจสามารถสรุปได้ 4 ประเด็นดังต่อไปนี้

1) ประเด็นความสามารถในการลดขนาดลักษณะเฉพาะ จากผลการทดลองพบว่าวิธี FR_FCA_WPC สามารถลดขนาดลักษณะเฉพาะลงได้และยังให้ค่า F-measure ที่สูง

2) ประเด็นเปรียบเทียบประสิทธิภาพการลดขนาดลักษณะเฉพาะระหว่างวิธี FR_FCA_WPC กับวิธี IG พบว่าการลดขนาดลักษณะด้วยวิธี FR_FCA_WPC สามารถลดขนาดลักษณะเฉพาะได้ดีกว่าและให้ค่า F-measure สูงกว่าวิธี IG สำหรับการจำแนกประเภทด้วย MLP และให้ค่า F-measure สูงกว่าหรือเท่ากับวิธี IG สำหรับการจำแนกประเภทด้วย SVM

3) ประเด็นเปรียบเทียบตัวจำแนกประเภทระหว่าง MLP และ SVM จากผลการทดลองพบว่าการลดขนาดลักษณะเฉพาะด้วยวิธี FR_FCA_WPC ที่จำแนกประเภทด้วย MLP ให้ค่า F-measure สูงกว่าที่จำแนกประเภทด้วย SVM

4) ประเด็นการปรับค่า λ จากผลการทดลองพบว่าค่า λ มีผลที่ให้ค่า F-measure สูงขึ้นได้ โดยจะต้องเลือกค่า λ ที่เหมาะสมกับชุดข้อมูล กล่าวคือ ในการกำหนดค่า λ ในขั้นตอนวิธีของ FCA ถ้า $\lambda = 0$ หมายถึงจะพิจารณาความสัมพันธ์ที่มีค่าน้ำหนัก TF-IDF ระหว่างลักษณะเฉพาะกับเอกสารเว็บเพจทั้งหมด แต่เมื่อปรับค่า λ เพิ่มมากขึ้นจะพิจารณาความสัมพันธ์ที่มีค่าน้ำหนักมากขึ้นตามไปด้วย ซึ่งค่า λ ที่เหมาะสมจะขึ้นอยู่กับชุดข้อมูลโดยเสนอแนะให้ค่า λ อยู่ระหว่าง 0 ถึง Mean

ข้อสังเกต วิธีที่นำเสนออาจต้องใช้เวลาในการคำนวณหาความสัมพันธ์ระหว่างลักษณะเฉพาะกับเอกสารเว็บเพจในขั้นตอนการเลือกลักษณะเฉพาะโดยใช้ FCA ซึ่งเป็นการทำงานในลักษณะของการเตรียมข้อมูล (Preprocessing) ก่อนการจำแนกประเภท

6.2 ปัญหาและอุปสรรค

เนื่องจากข้อมูลเว็บเพจที่นำมาทดสอบอยู่ในรูปแบบที่ไม่สามารถนำมาใช้วิเคราะห์ได้ทันทีจึงต้องใช้เวลานานและมีกระบวนการที่ซับซ้อนในการเตรียมข้อมูลให้อยู่ในรูปแบบที่พร้อมใช้งาน ซึ่งต่างจากบางข้อมูลอื่นในการทำเหมืองข้อมูลที่สามารถนำมาใช้งานได้ทันที

6.3 ข้อเสนอแนะ

1) เนื่องจากเว็บเพจในปัจจุบันอาจมีการใช้คำใหม่ ๆ หรือสัญลักษณ์ที่ไม่มีความหมายในการวิเคราะห์ ดังนั้นควรมีการเพิ่มข้อมูลคำใหม่ใน Stoplist ให้ทันสมัยเพื่อให้การดำเนินการในขั้นตอนการเตรียมข้อมูลเว็บเพจมีประสิทธิภาพสูงสุด

2) แบบจำลองการลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจของงานวิจัยนี้สามารถนำไปปรับปรุงใช้กับเว็บเพจที่เป็นภาษาอื่น ๆ ได้ เช่น ภาษาไทย ภาษาจีน ภาษาญี่ปุ่น ซึ่งอาจได้ผลลัพธ์ที่แตกต่างกันขึ้นอยู่กับลักษณะเฉพาะที่สกัดได้จากเว็บเพจของภาษานั้น ๆ

บรรณานุกรม

ณสิทธิ์ เหล่าเส็น. 2551. แบบจำลองการกรองข้อมูลอากาศที่มีสิ่งรบกวนโดยใช้โครงข่ายประสาทเทียม. วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต, มหาวิทยาลัยสงขลานครินทร์ สงขลา.

พรพล ธรรมรงค์รัตน์ และคณะ. 2552. การจำแนกประเภทเว็บเพจโดยวิธีการลดขนาดลักษณะเฉพาะและมัลติคลาสซัพพอร์ตเวกเตอร์แมชชีน. Proceedings of the Conference on Knowledge and Smart Technologies 2009. 24-25 กรกฎาคม 2552. หน้า 103-108.

Chen, R. C. and Hsieh, C. H. 2006. Web page classification based on a support vector machine using a weighted vote schema, Expert Systems with Applications, 31: 427–435.

Cure, O. and Jeansoulin, R. 2008. An FCA-based Solution for Ontology Mediation, Proceeding of the 2nd International workshop on Ontologies and information systems for the semantic web (ONISW'08), pp. 39-46.

Frankes, W. B., and Yates, R. B. 1992. Information retrieval data structure & algorithm New Jersey: Prentice Hall.

Haav, H. M. 2004. A Semi-automatic Method to Ontology Design by Using FCA, CLA 2004, pp. 13-24.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten I. H. 2009. The WEKA Data Mining Software: An Update, SIGKDD Explorations, 11: 10-18.

Hwang, S. H., Kim, H. G. and Yang, H. S. 2005. A FCA-Based Ontology Construction for the Design of Class Hierarchy, ICCSA 2005, LNCS 3482, pp. 827-835.

Indra, M., Rajaram, R., and Selvakuberan, K. 2008. Generating best features for web page classification. Webology, 5(1), Article 52.

Joachims, T. 1997. A Probabilistic Analysis of The Rocchio Algorithm with TFIDF for Text Categorization. Proceedings of ICML-97 International Conference on Machine Learning.

Kohavi R., and Provost F. 1998. Glossary of terms, Machine Learning, Vol. 30, No. 2/3, pp. 271-274.

Kotsiantis, S.B. 2007. Supervised Machine Learning: A Review of Classification Techniques, Informatica. 31: 249-268.

Lakhal, L., and Stumme, G. 2005. Efficient Mining of Association Rules Based on Formal Concept Analysis. Formal Concept Analysis, LNAI 3626, pp. 180–195.

Obitko, M., Snasel, V. and Smid, J. 2004. Ontology Design with Formal Concept Analysis, CLA 2004, pp. 111-119.

Porter. M. F. 1980. An algorithm for suffix stripping. Program, 14(3): 130–137.

Priss, U. 2006. Formal Concept Analysis in Information Science. Annual Review of Information Science and Technology, pp. 521-543.

Qi, X. and Davison, D. 2009. Web Page Classification: Features and Algorithms, ACM Comput. Surv. 41, 2, pp. 1-31.

Radovanovic, M. 2006. Machine Learning in Web Mining. Master's thesis, Department of Mathematics and Informatics, Faculty of Science, University of Novi Sad, Serbia.

Sasaki, Y. 2007. The truth of the F-measure. Version: 26th October, pp.1-5.

Shen, D., Yang, Q. amd Chen, Z. 2007. Noise reduction through summarization for Web-page classification, Information Processing and Management, 43: 1735–1747.

StatLib. 2012. StatLib Datasets Archive. http://lib.stat.cmu.edu/datasets/. (accessed 01/02/2012).

Thamrongrat, P., Preechaveerakul, L. and Wettayaprasit W. 2009. A Novel Voting Algorithm of Multi-Class SVM for Web Page Classification. In Proceedings The 2nd IEEE International Conference on Computer Science and Information Technology (ICCSIT 2009), Beijing, China, August 8-11, pp. 327-330.

Van Rijsbergen, C.J. 1979. Information Retrieval. London: Butterworths.

Wang, J. and He, K. 2006. Towards Representing FCA-based Ontologies in Semantic Web Rule Language, Proceeding of The Sixth IEEE International Conference on Computer and Information Technology (CIT'06), 20-22, pp. 1-5.

WebKB. 2012. CMU World Wide Knowledge Base project. http://www.cs.cmu.edu/~WebKB/. (accessed 01/02/2012).

Wille, R. 2005. Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies. Formal Concept Analysis, LNAI3626, Springer Verlag, pp. 1-33.

Yevtushenko, S. A. 2000. System of data analysis "Concept Explorer". Proceedings of the 7th national conference on Artificial Intelligence KII-2000, Russian, pp. 127-134.

Yin, S., Wang, F., Xie, Z. and Qiu, Y. 2008. Study on Web-Page Classification Algorithm Based on Rough Set Theory. International Symposiums on Information Processing, 118, pp. 202-206.

Zhang, R., Shepherd, M., Duffy, J. and Watters C. 2007. Automatic Web Page Categorization using Principal Component Analysis. Proceedings of the 40th Hawaii International Conference on System Sciences, pp. 1-10.

ภาคผนวก

ภาคผนวก ก

การใช้งาน Command Line Interface ใน WEKA-3-6

โปรแกรม WEKA (Waikato Environment for Knowledge Analysis) เป็นโปรแกรมที่รวบรวมเอาอัลกอริทึมเกี่ยวกับการเรียนรู้ของเครื่องมาใช้ในการวิเคราะห์ข้อมูลในการทำเหมืองข้อมูล สามารถดาวน์โหลดได้จาก http://www.cs.waikato.ac.nz/ml/weka/ ส่วนประกอบของโปรแกรม WEKA มีทั้งแบบ User Interface และ Command Line Interface ในงานวิจัยนี้ได้ใช้แบบ Command Line Interface ซึ่งสามารถเรียกใช้จากโปรแกรมที่พัฒนาได้

ก.1 โครงสร้างของ WEKA

โครงสร้างของ WEKA ประกอบด้วยแพคเกจในการทำงานหลายแพคเกจ ซึ่งในงานวิจัยนี้ได้ใช้แพคเกจ AttributeSelection และ Classifiers ดังภาพประกอบ ก.1 และ ก.2 ตามลำดับ

Package weka.attributeSelection

Interface Summary

AttributeEvaluator	Interface for classes that evaluate attributes individually.
AttributeTransformer	Abstract attribute transformer.
ErrorBasedMeritEvaluator	Interface for evaluators that calculate the "merit" of attributes/subsets as the error of a learning scheme
RankedOutputSearch	Interface for search methods capable of producing a ranked list of attributes.
StartSetHandler	Interface for search methods capable of doing something sensible given a starting set of attributes.
SubsetEvaluator	Interface for attribute subset evaluators.

Class Summary

ASEvaluation	Abstract attribute selection evaluation class
ASSearch	Abstract attribute selection search class.
AttributeSelection	Attribute selection class.
AttributeSetEvaluator	Abstract attribute set evaluator.
BestFirst	BestFirst: Searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility.
CfsSubsetEval	CfsSubsetEval : Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred. For more information see: M.
CheckAttributeSelection	Class for examining the capabilities and finding problems with attribute selection schemes.
ChiSquaredAttributeEval	ChiSquaredAttributeEval : Evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class. Valid options are:

ภาพประกอบ ก.1 แพคเกจ AttributeSelection

ClassifierSubsetEval	Evaluates attribute subsets on training data or a seperate hold out testing set.
ConsistencySubsetEval	ConsistencySubsetEval : Evaluates the worth of a subset of attributes by the level of consistency in the class values when the training instances are projected onto the subset of attributes.
CostSensitiveASEvaluation	Abstract base class for cost-sensitive subset and attribute evaluators.
CostSensitiveAttributeEval	A meta subset evaluator that makes its base subset evaluator cost-sensitive.
CostSensitiveSubsetEval	A meta subset evaluator that makes its base subset evaluator cost-sensitive.
ExhaustiveSearch	ExhaustiveSearch : Performs an exhaustive search through the space of attribute subsets starting from the empty set of attrubutes.
FilteredAttributeEval	Class for running an arbitrary attribute evaluator on data that has been passed through an arbitrary filter (note: filters that alter the order or number of attributes are not allowed).
FilteredSubsetEval	Class for running an arbitrary subset evaluator on data that has been passed through an arbitrary filter (note: filters that alter the order or number of attributes are not allowed).
GainRatioAttributeEval	GainRatioAttributeEval : Evaluates the worth of an attribute by measuring the gain ratio with respect to the class. GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute). Valid options are:
GeneticSearch	GeneticSearch: Performs a search using the simple genetic algorithm described in Goldberg (1989). For more information see: David E.
GreedyStepwise	GreedyStepwise : Performs a greedy forward or backward search through the space of attribute subsets.
HoldOutSubsetEvaluator	Abstract attribute subset evaluator capable of evaluating subsets with respect to a data set that is distinct from that used to initialize/ train the subset evaluator.
InfoGainAttributeEval	InfoGainAttributeEval : Evaluates the worth of an attribute by measuring the information gain with respect to the class. InfoGain(Class,Attribute) = H(Class) - H(Class | Attribute). Valid options are:
LatentSemanticAnalysis	Performs latent semantic analysis and transformation of the data.
LFSMethods	
LinearForwardSelection	LinearForwardSelection: Extension of BestFirst.
OneRAttributeEval	OneRAttributeEval : Evaluates the worth of an attribute by using the OneR classifier. Valid options are:
PrincipalComponents	Performs a principal components analysis and transformation of the data.
RaceSearch	Races the cross validation error of competing attribute subsets.
RandomSearch	RandomSearch : Performs a Random search in the space of attribute subsets.
Ranker	Ranker : Ranks attributes by their individual evaluations.
RankSearch	RankSearch : Uses an attribute/subset evaluator to rank all attributes.
ReliefFAttributeEval	ReliefFAttributeEval : Evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class.
ScatterSearchV1	Class for performing the Sequential Scatter Search.
SubsetSizeForwardSelection	SubsetSizeForwardSelection: Extension of LinearForwardSelection.
SVMAttributeEval	SVMAttributeEval : Evaluates the worth of an attribute by using an SVM classifier.
SymmetricalUncertAttributeEval	SymmetricalUncertAttributeEval : Evaluates the worth of an attribute by measuring the symmetrical uncertainty with respect to the class.
UnsupervisedAttributeEvaluator	Abstract unsupervised attribute evaluator.
UnsupervisedSubsetEvaluator	Abstract unsupervised attribute subset evaluator.
WrapperSubsetEval	WrapperSubsetEval: Evaluates attribute sets by using a learning scheme.

ภาพประกอบ ก.1 แพคเกจ AttributeSelection (ต่อ)

Package weka.classifiers

Interface Summary

IntervalEstimator	Interface for classifiers that can output confidence intervals
IterativeClassifier	Interface for classifiers that can induce models of growing complexity one step at a time.
Sourcable	Interface for classifiers that can be converted to Java source.
UpdateableClassifier	Interface to incremental classification models that can learn using one instance at a time.

Class Summary

BVDecompose	Class for performing a Bias-Variance decomposition on any classifier using the method specified in: Ron Kohavi, David H.
BVDecomposeSegCVSub	This class performs Bias-Variance decomposion on any classifier using the sub-sampled cross-validation procedure as specified in (1). The Kohavi and Wolpert definition of bias and variance is specified in (2). The Webb definition of bias and variance is specified in (3). Geoffrey I.
CheckClassifier	Class for examining the capabilities and finding problems with classifiers.
CheckSource	A simple class for checking the source generated from Classifiers implementing the weka.classifiers.Sourcable interface.
Classifier	Abstract classifier.
CostMatrix	Class for storing and manipulating a misclassification cost matrix.
Evaluation	Class for evaluating machine learning models.
IteratedSingleClassifierEnhancer	Abstract utility class for handling settings common to meta classifiers that build an ensemble from a single base learner.
MultipleClassifiersCombiner	Abstract utility class for handling settings common to meta classifiers that build an ensemble from multiple classifiers.
RandomizableClassifier	Abstract utility class for handling settings common to randomizable classifiers.
RandomizableIteratedSingleClassifierEnhancer	Abstract utility class for handling settings common to randomizable meta classifiers that build an ensemble from a single base learner.
RandomizableMultipleClassifiersCombiner	Abstract utility class for handling settings common to randomizable meta classifiers that build an ensemble from multiple classifiers based on a given random number seed.
RandomizableSingleClassifierEnhancer	Abstract utility class for handling settings common to randomizable meta classifiers that build an ensemble from a single base learner.
SingleClassifierEnhancer	Abstract utility class for handling settings common to meta classifiers that use a single base learner.

ภาพประกอบ ก.2 แพคเกจ Classifiers

ก.2 การเขียนคำสั่งใน WEKA

ในงานวิจัยนี้ได้เลือกใช้คลาส InfoGainAttributeEval จากแพคเกจ AttributeSelection สำหรับการเลือกลักษณะเฉพาะด้วยวิธี IG และเลือกใช้คลาสฟังก์ชัน (Function) ได้แก่ ฟังก์ชัน MultilayerPerceptron สำหรับการจำแนกประเภทด้วย MLP และฟังก์ชัน SMO สำหรับจำแนกประเภทด้วย SVM จากแพคเกจ Classifiers ซึ่งการเขียนคำสั่งใช้งานแบบ Command Line ใน WEKA ประกอบด้วยพารามิเตอร์ต่าง ๆ ที่ใช้ในการทำงานดังนี้

1) **AttributeSelection** มีพารามิเตอร์ที่จำเป็นในการใช้งานดังนี้

-S คือเรียกใช้ class การจัดลำดับของลักษณะเฉพาะแบบ Ranker จาก weak.attributeSelection.Ranker

-N คือกำหนดจำนวนลักษณะเฉพาะที่ต้องการ

-E คือกำหนดเรียกใช้คลาส InfoGainAttributeEval จาก weka.attributeSelection.InfoGainAttributeEval

-i คือไฟล์ Input

-o คือไฟล์ Output

ตัวอย่างการใช้คำสั่งสำหรับการเลือกลักษณะเฉพาะแสดงดังภาพประกอบ ก.3

```
java weka.filters.supervised.attribute.AttributeSelection -S "weka.attributeSelection.Ranker -N 50"
-E "weka.attributeSelection.InfoGainAttributeEval" -i feature_input.arff -o IG_feature50.arff
```

ภาพประกอบ ก.3 ตัวอย่างการเขียนคำสั่งเรียกใช้ AttributeSelection

จากภาพประกอบ ก.3 ตัวอย่างการเขียนคำสั่งเรียกใช้ AttributeSelection ต้องการเลือกลักษณะเฉพาะด้วยวิธี IG ซึ่งจำนวนลักษณะเฉพาะที่ต้องเท่ากับ 50 โดยมีไฟล์ Input คือ feature_input.arff และไฟล์ Output คือ IG_feature50.arff

2) Classifiers มีพารามิเตอร์ที่จำเป็นในการใช้งานดังนี้

-x คือจำนวน Fold ในการแบ่งข้อมูลแบบ Cross Validation

-t คือไฟล์ Input

-i คือแสดงรายละเอียดการประเมินประสิทธิภาพจากค่า True-Positive False-Positive Precision Recall และ F-measure ซึ่งคำนวณได้จากตาราง Confusion Matrix

ตัวอย่างการใช้คำสั่งสำหรับการจำแนกประเภทแสดงดังภาพประกอบ ก.4

```
java weka.classifiers.functions.MultilayerPerceptron -x 10 -t FCA_feature.arff –i > FCA_MLP.txt
```

ภาพประกอบ ก.4 ตัวอย่างการเขียนคำสั่งเรียกใช้ Classifiers

จากภาพประกอบ ก.4 ตัวอย่างการเขียนคำสั่งเรียกใช้ Classifiers ต้องการจำแนกประเภทด้วย MLP โดยทดสอบแบบ 10-Fold Cross Validation โดยมีไฟล์ Input คือ FCA_feature.arff และไฟล์ Output คือ FCA_MLP.txt

ภาคผนวก ข

ผลงานวิจัยที่ได้รับการตีพิมพ์ในงานประชุมวิชาการ JCSSE 2010

เรื่อง การจำแนกประเภทเว็บเพจโดยวิธีการลดขนาดลักษณะเฉพาะด้วย FCA

Conference The 7th International Joint Conference on Computer Science and Software Engineering (JCSSE 2010)

สถานที่ มหาวิทยาลัยรามคำแหง ประเทศไทย

วันที่ 12-14 พฤษภาคม 2553

การจำแนกประเภทเว็บเพจโดยวิธีการลดขนาดลักษณะเฉพาะด้วย FCA
Web Page Classification Using FCA Feature Reduction

วิรัตน์ ชูนุ้ย[1] วิภาดา เวทย์ประสิทธิ์[1] และ ลัดดา ปรีชาวีรกุล[2]
ห้องปฏิบัติการวิจัยปัญญาประดิษฐ์[1] ห้องปฏิบัติการวิจัยเทคโนโลยีระบบสารสนเทศและโปรแกรมประยุกต์[2] คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ จังหวัดสงขลา
E-mail: wirat218@hotmail.com[1], wwetayaprasit@yahoo.com[1], ladda.p@psu.ac.th[2]

บทคัดย่อ

จำนวนเว็บเพจมีอัตราเพิ่มขึ้นอย่างมหาศาลทำให้การสืบค้นข้อมูลให้ได้ตรงกับความต้องการของผู้ใช้มีความยุ่งยากและเสียเวลามาก การจำแนกประเภทเว็บเพจเป็นวิธีการหนึ่งที่ช่วยแก้ปัญหาดังกล่าว อย่างไรก็ตามหากจำนวนเว็บเพจมีจำนวนมากก็จะทำให้ขนาดลักษณะเฉพาะซึ่งเป็นข้อมูลเข้ามีขนาดใหญ่ตามไปด้วย บทความนี้จึงได้นำเสนอการจำแนกประเภทเว็บเพจโดยวิธีการลดขนาดลักษณะเฉพาะด้วย FCA ทดลองกับชุดข้อมูลเว็บเพจมาตรฐานจาก CMU ใช้ลักษณะเฉพาะจากข้อความและหัวเรื่อง จำแนกประเภทด้วย MLP Neural Networks เปรียบเทียบประสิทธิภาพระหว่างการเลือกลักษณะเฉพาะด้วยวิธีการที่นำเสนอกับการเลือกลักษณะเฉพาะแบบ Information Gain ผลการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอสามารถลดขนาดลักษณะเฉพาะได้มากกว่าโดยที่ค่า F-measure ยังคงสูงหรือสูงกว่าเดิม

คำสำคัญ: การจำแนกประเภทเว็บเพจ, การลดขนาดลักษณะเฉพาะ, Formal Concept Analysis

Abstract

A number of web pages were increasing immensely which brought to the difficulty and time-consuming of finding information that exactly matched the need of users. The classification of web pages would be one method to solve such problem. However, if the high number of the web pages was, the larger of the input data would be. Then this paper proposed Web Page Classification Using FCA Feature Reduction (WPC_FCA_FR) to study with the benchmark web page data set from CMU. The study used unique features from text and title to be classified by MLP Neural Networks comparing on the efficiency between feature selection methods of the proposed method and Information Gain. The result of the study indicated that this proposed method could reduce more features while F-measure was high or higher value.

Key Words: Web Page Classification, Feature Reduction, Formal Concept Analysis

1. บทนำ

ปัจจุบันอินเตอร์เน็ตได้รับความนิยมอย่างแพร่หลายและมีการพัฒนาที่รวดเร็ว ส่งผลให้จำนวนเว็บเพจ (Web Page) มีอัตราเพิ่มขึ้นอย่างมหาศาลทำให้การสืบค้นข้อมูลให้ได้ข้อมูลที่ตรงกับความต้องการจริงมีความยุ่งยากและเสียเวลามาก วิธีการหนึ่งที่ช่วยแก้ปัญหาดังกล่าวก็คือ การจำแนกประเภทเว็บเพจ (Web Page Classification) [1] ซึ่งเป็นการจัดเว็บเพจให้เป็นกลุ่มตามความสนใจ เพื่อช่วยสนับสนุนให้ผู้ใช้สืบค้นข้อมูลได้ถูกต้องและสะดวกรวดเร็วมากยิ่งขึ้น อย่างไรก็ตามในแต่ละเว็บเพจประกอบด้วยคำ (Words) เป็นจำนวนมากจึงทำให้ขนาดของลักษณะเฉพาะ (Features) ของข้อมูลเข้า (Input Features) มีขนาดใหญ่ตามไปด้วย ซึ่งส่งผลให้การจำแนกประเภทเว็บเพจมีความซับซ้อนมากยิ่งขึ้น ดังนั้นแนวทาง

หนึ่งที่ช่วยแก้ปัญหานี้คือ การลดขนาดลักษณะเฉพาะ (Feature Reduction) เพื่อช่วยเพิ่มประสิทธิภาพของการจำแนกประเภทเว็บเพจ

งานวิจัยที่เกี่ยวข้องกับการเลือกลักษณะเฉพาะ ได้แก่ Zhang และคณะ [2] ได้ใช้วิธีการเลือกลักษณะเฉพาะแบบ Information Gain เพื่อเลือกลักษณะเฉพาะสำหรับการจัดกลุ่มประเภทเว็บเพจแบบอัตโนมัติโดยใช้วิธีการ Principal Component Analysis (PCA) Indra และคณะ [3] ได้รวมเทคนิคการเลือกลักษณะเฉพาะเพื่อหาลักษณะเฉพาะที่มีคุณภาพดีและมีจำนวนน้อยที่สุด ทดลองโดยใช้ชุดข้อมูลมาตรฐานจาก WebKB พรพลและคณะ [4] ได้นำเสนอการจำแนกประเภทเว็บเพจโดยวิธีการลดขนาดลักษณะเฉพาะและเปรียบประสิทธิภาพระหว่างวิธีการเลือกลักษณะเฉพาะแบบ ReliefF Information Gain และ Chi Square ทำการจำแนกประเภทโดยเปรียบเทียบวิธีมัลติคลาสซัพพอร์ตเวกเตอร์แมชชีนแบบ 1vs1 และ 1vsAll นอกจากนี้ Thamrongrat และคณะ [5] ยังได้นำเสนออัลกอริทึมการโหวตในการจำแนกประเภทเว็บเพจโดยใช้วิธีแบบมัลติคลาสซัพพอร์ตเวกเตอร์แมชชีน (Voting Algorithm of Multi-Class SVM for Web Page Classification) ที่มีชื่อว่า VAMSVM_WPC ทำการทดลองกับชุดข้อมูลเว็บเพจมาตรฐานจาก CMU โดยใช้ลักษณะเฉพาะจากข้อความ (Text) และหัวเรื่อง (Title) และประเมินประสิทธิภาพด้วยค่า F-measure Park และ Zhang [6] ได้ใช้ตัวจำแนก Multi-Layer Perceptron (MLP) ในการจำแนกประเภทเว็บเพจแบบอัตโนมัติ ทดลองโดยใช้ชุดข้อมูล NIPS 2000 และชุดข้อมูล WebKB

บทความนี้จึงได้นำเสนอการจำแนกประเภทเว็บเพจโดยวิธีการลดขนาดลักษณะเฉพาะด้วยวิธีการ Formal Concept Analysis (FCA) ทดลองกับชุดข้อมูลเว็บเพจมาตรฐานจาก CMU จำแนกประเภทด้วย Multi-Layer Perceptron Neural Networks และประเมินประสิทธิภาพด้วยค่า F-measure

บทความนี้ประกอบด้วยส่วนต่าง ๆ ดังนี้ ส่วนที่ 2 Formal Concept Analysis ส่วนที่ 3 ขั้นตอนการจำแนกประเภทเว็บเพจ ส่วนที่ 4 วิธีการทดลองและผลการทดลอง และส่วนที่ 5 บทสรุป

2. Formal Concept Analysis

Formal Concept Analysis (FCA) [7] เป็นการกำหนดกรอบแนวคิด (Conceptual Framework) สำหรับการสร้างโครงสร้างข้อมูล วิเคราะห์ข้อมูล และแสดงข้อมูลในรูปแบบของแผนภาพความสัมพันธ์ที่เรียกว่า คอนเซ็ปต์แลททิส (Concept Lattice) ซึ่งจะทำให้สามารถเข้าใจข้อมูลนั้นได้ง่ายยิ่งขึ้น โดยใน FCA จะมีการแทนชุดข้อมูล (Data Set) ด้วยฟอร์มัลคอนเท็กซ์ (Formal Context) ซึ่งเป็นโครงสร้างพื้นฐานของ FCA

นิยาม 1 ฟอร์มัลคอนเท็กซ์ (Formal Context)

กำหนดความสัมพันธ์ $K := (G, M, I)$ โดย G เป็นเซตของออบเจ็กต์ (Object) และ M เป็นเซตของคุณลักษณะ (Attribute) ส่วน I เป็นความสัมพันธ์เชิงคู่ (Binary Relation) ระหว่าง G และ M และถ้าเขียนเป็น $(g, m) \in I$ จะอ่านได้ว่า "ออบเจ็กต์ g มีคุณลักษณะ m"

ฟอร์มัลคอนเท็กซ์สามารถเขียนแทนเป็นตารางเมตริกซ์ได้โดยที่ส่วนหัวของแถว (Row) จะแทนด้วยออบเจ็กต์ และส่วนหัวของคอลัมน์ (Column) จะแทนด้วยคุณลักษณะ ส่วนกากบาทในแถว g และคอลัมน์ m หมายความว่าออบเจ็กต์ g มีคุณลักษณะ m ดังตารางที่ 1 แสดงตัวอย่างของฟอร์มัลคอนเท็กซ์ประกอบด้วย 4 ออบเจ็กต์ และ 7 คุณลักษณะ [8]

ตารางที่ 1 ตัวอย่างของฟอร์มัลคอนเท็กซ์

Attributes / Objects	a1	a2	a3	a4	a5	a6	a7
o1		x				x	x
o2			x	x	x		
o3			x		x	x	x
o4	x		x				

นิยาม 2 ตัวดำเนินการ $'$ ($'$ Operation)

เมื่อเซต $A \subseteq G$ ของออบเจ็กต์ จะได้ว่า

$$A' = \{m \in M \mid \forall_g \in A : (g,m) \in I\} \quad (1)$$

เมื่อเซต $B \subseteq M$ ของคุณลักษณะ จะได้ว่า

$$B' = \{g \in G \mid \forall_m \in B : (g,m) \in I\} \quad (2)$$

นั่นคือ A' เป็นเซตของคุณลักษณะร่วมของออบเจ็กต์ทั้งหมดใน A และ B' เป็นเซตของออบเจ็กต์ทั้งหมดที่มีคุณลักษณะร่วมใน B

นิยาม 3 ฟอร์มัลคอนเซ็ปต์ (Formal Concept)

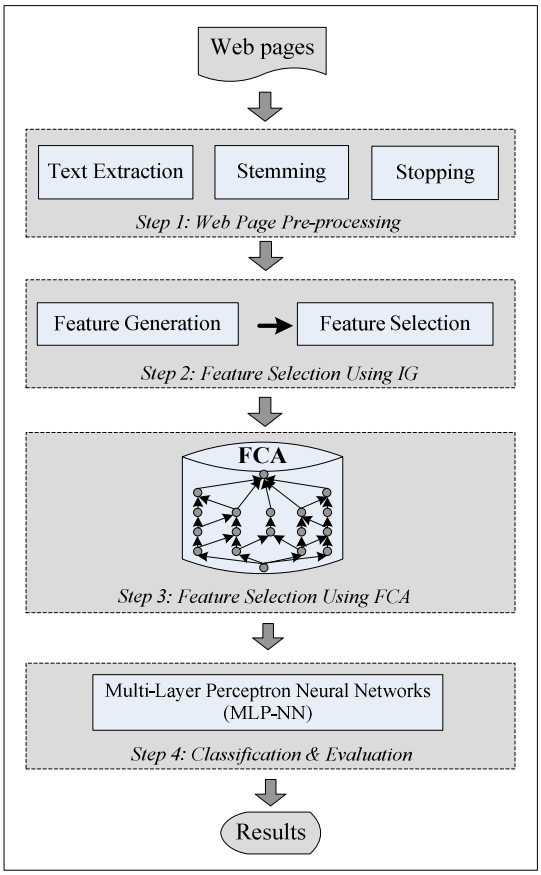
ฟอร์มัลคอนเซ็ปต์ C จากฟอร์มัลคอนเท็กซ์ (G,M,I) คือคู่อันดับ (A,B) โดยที่ $A \subseteq G$, $B \subseteq M$, $A' = B$ และ $B' = A$ ซึ่งจะเรียกเซต A และ B ว่า “Extent” และ “Intent” ของคอนเซ็ปต์ C ตามลำดับ นั่นคือ Extent แทนกลุ่มของออบเจ็กต์ และ Intent แทนกลุ่มของคุณลักษณะร่วม ซึ่งจากตัวอย่างฟอร์มัลคอนเท็กซ์ในตารางที่ 1 จะได้ 9 ฟอร์มัลคอนเซ็ปต์ ดังนี้ ({o1,o2,o3,o4},Ø), ({o2,o3,o4},{a3}), ({o1,o3},{a6,a7}), ({o2,o3},{a3,a5}), ({o1},{a2,a6,a7}),({o3},{a3,a5,a6,a7}),{o2},{a3,a4,a5}), ({o4},{a1,a3}), (Ø,{a1,a2,a3,a4,a5,a6,a7})

3. ขั้นตอนการจำแนกประเภทเว็บเพจ

บทความนี้ได้นำเสนอขั้นตอนการจำแนกประเภทเว็บเพจโดยใช้วิธีการลดขนาดลักษณะเฉพาะด้วย FCA (Web Page Classification Using FCA Feature Reduction: WPC_FCA_FR) ซึ่งมีโครงสร้างการทำงานของระบบดังรูปที่ 1 และมีขั้นตอนการทำงานทั้งหมด 4 ขั้นตอน ได้แก่ การเตรียมข้อมูลเว็บเพจ (Web Page Pre-processing) การเลือกลักษณะเฉพาะด้วย IG (Feature Selection Using IG) การเลือกลักษณะเฉพาะด้วย FCA (Feature Selection Using FCA) และการจำแนกประเภทและการประเมินผล (Classification and Evaluation) โดยการทำงานในแต่ละขั้นตอนมีดังนี้

3.1 ขั้นตอนที่ 1 การเตรียมข้อมูลเว็บเพจ

เป็นขั้นตอนการสกัดข้อความ (Text) และหัวเรื่อง (Title) จากหน้าเว็บเพจ จากนั้นหารากศัพท์ (Stemming) ของคำโดยใช้อัลกอริทึม Potter Stemmer และกำจัดคำหยุด (Stopwords) [9]



รูปที่ 1 โครงสร้างการทำงานของระบบ WPC_FCA_FR

3.2 ขั้นตอนที่ 2 การเลือกลักษณะเฉพาะด้วย IG

หลังจากผ่านขั้นตอนที่ 1 ข้อมูลที่ได้จะถูกนำมาสร้างเป็นลักษณะเฉพาะ (Feature Generation) โดยสร้างอยู่ในรูปของตารางความถี่ของคำ (Term Frequency) เพื่อเลือกคำที่มีค่าความถี่เอกสาร (Document Frequency) มากกว่าค่าที่กำหนด (Threshold) [9] จากนั้นนำคำที่ได้มาให้ค่าน้ำหนัก (Weighting) ด้วยวิธีการ TF-IDF แล้วนำค่าที่ได้สร้างให้อยู่ในรูปของเมตริกซ์เอกสาร (Document Matrix) [9] จากลักษณะเฉพาะที่ได้นำมาเลือกลักษณะเฉพาะด้วยวิธี

Information Gain (IG) เพื่อลดขนาดลักษณะเฉพาะให้เหมาะสมต่อการจำแนกประเภท

3.3 ขั้นตอนที่ 3 การเลือกลักษณะเฉพาะด้วย FCA

เริ่มต้นจากการเลือกลักษณะเฉพาะด้วยวิธี IG ขนาด *m* จำนวน จากนั้นใช้วิธีการ FCA ลดขนาดลักษณะเฉพาะลง *n* จำนวน โดยที่ $n < m$ และยังคงให้ค่าความถูกต้องเพิ่มขึ้นหรือเท่าเดิม แนวทางในการเลือกลักษณะเฉพาะด้วยวิธีการ FCA คือจะพิจารณาจากความสัมพันธ์ระหว่างเอกสารเว็บเพจกับลักษณะเฉพาะและคลาส (Class) โดยมีขั้นตอนดังนี้

1) นำชุดข้อมูลเว็บเพจที่เลือกลักษณะเฉพาะด้วยวิธี IG แปลงเป็นฟอร์มัลคอนเท็กซ์ตามนิยาม 1 ตารางที่ 2 แสดงตัวอย่างฟอร์มัลคอนเท็กซ์ของชุดข้อมูลเว็บเพจประกอบด้วยเอกสารเว็บเพจ 10 เว็บเพจ มีลักษณะเฉพาะ 6 ลักษณะเฉพาะ และลักษณะเฉพาะคลาส 2 คลาส ซึ่งเป็นตารางเมตริกซ์ส่วนหัวของคอลัมน์แทนด้วยลักษณะเฉพาะ (f1 f2 f3 … f6) และลักษณะเฉพาะคลาส (c1 และ c2) และส่วนหัวของแถวแทนด้วยเอกสารเว็บเพจ (d1 d2 d3 … d10) ส่วนกากบาทในแถวและคอลัมน์จะแทนความสัมพันธ์ระหว่างเอกสารเว็บเพจกับลักษณะเฉพาะและคลาส

2) สร้างฟอร์มัลคอนเซ็ปต์จากฟอร์มัลคอนเท็กซ์ในข้อ 1) ตามนิยาม 2 และ 3 จะได้ Extent ของคอนเซ็ปต์ คือ เซตของเอกสารเว็บเพจ (d) และ Intent ของคอนเซ็ปต์ คือ เซตของลักษณะเฉพาะ (f และ c) จากตัวอย่างในตารางที่ 2 จะได้ฟอร์มัลคอนเซ็ปต์ที่มีลักษณะเฉพาะสัมพันธ์กับลักษณะเฉพาะคลาสทั้งหมด 7 ฟอร์มัลคอนเซ็ปต์ และเรียงตามจำนวนสมาชิกใน Extent จากมากไปน้อยได้ดังตารางที่ 3 จะเห็นว่าฟอร์มัลคอนเซ็ปต์ที่ 1 มี Extent ของคอนเซ็ปต์ คือ {d1,d2,d3,d4,d5} ซึ่งมีจำนวนสมาชิกมากที่สุดเท่ากับ 5 และมี Intent ของคอนเซ็ปต์ คือ {f1,c1} ซึ่งหมายความว่า ลักษณะเฉพาะ f1 สัมพันธ์กับเอกสารเว็บเพจทั้งหมด 5 เว็บเพจในคลาส c1 ดังนั้น f1 จึงเป็นลักษณะเฉพาะเด่นของคลาส c1

ตารางที่ 2 ตัวอย่างฟอร์มัลคอนเท็กซ์ของชุดข้อมูลเว็บเพจ

Features / Web pages	f1	f2	f3	f4	f5	f6	c1	c2
d1	x						x	
d2	x			x			x	
d3	x			x		x	x	
d4	x			x		x	x	
d5	x	x		x		x	x	
d6			x					x
d7			x		x			x
d8			x		x			x
d9	x	x	x		x			x
d10	x	x	x		x			x

ตารางที่ 3 ตัวอย่างฟอร์มัลคอนเซ็ปต์ของชุดข้อมูลเว็บเพจ

ลำดับ	ฟอร์มัลคอนเซ็ปต์	จำนวนสมาชิกใน Extent
1	({d1,d2,d3,d4,d5}, {f1,c1})	5
2	({d6,d7,d8,d9,d10}, {f3,c2})	5
3	({d2,d3,d4,d5}, {f1,f4,c1})	4
4	({d7,d8,d9,d10}, {f3,f5,c2})	4
5	({d3,d4,d5}, {f1,f4,f6,c1})	3
6	({d9,d10}, {f1,f2,f3,f5,c2})	2
7	({d5}, {f1,f2,f4,f6,c1})	1

3) การเลือกลักษณะเฉพาะที่ดีที่สุดจะพิจารณาจากฟอร์มัลคอนเซ็ปต์ที่มีจำนวนสมาชิก (เอกสารเว็บเพจ) ใน Extent ที่มากที่สุดก่อน จากตารางที่ 3 ฟอร์มัลคอนเซ็ปต์ที่ 1 และ 2 มีจำนวนสมาชิกใน Extent ที่มากที่สุดคือ 5 ดังนั้นลักษณะเฉพาะ f1 และ f3 จะถูกเลือกก่อน จากนั้นพิจารณาฟอร์มัลคอนเซ็ปต์ที่มีจำนวนสมาชิกใน Extent น้อยลงมาตามลำดับ จะได้ลักษณะเฉพาะ f4 และ f5 ซึ่งมีจำนวนสมาชิกใน Extent เท่ากับ 4 และได้ลักษณะเฉพาะ f6 และ f2 ที่มีจำนวนสมาชิกใน Extent เท่ากับ 3 และ 2 ตามลำดับ ดังนั้นลักษณะเฉพาะที่ดีที่สุดที่ได้จากการเลือกด้วยวิธีการ FCA คือ f1 f3 f4 f5 f6 และ f2 ตามลำดับ

3.4 ขั้นตอนที่ 4 การจำแนกประเภทและการประเมินผล

การทดลองนี้ใช้ตัวจำแนกประเภท (Classifier) คือ Multi-Layer Perceptron Neural Networks (MLP-NN) [10] และประเมินประสิทธิภาพด้วยค่าเฉลี่ย F-measure

4. วิธีการทดลองและผลการทดลอง

4.1 ชุดข้อมูลทดสอบ

บทความนี้ใช้ชุดข้อมูลเว็บเพจจาก CMU [11] จำนวน 4,581 เว็บเพจ แบ่งเป็น 7 คลาส ได้แก่ Energy Financial Healthcare Materials Technology Transportation และ Utilities และสร้างชุดข้อมูลเพื่อใช้สำหรับทดลองออกเป็น 2 ชุด โดยชุดข้อมูลที่ 1 สุ่มข้อมูลเว็บเพจมาทั้งหมด 90 เว็บเพจ จาก 3 คลาส ได้แก่ 1) Financial จำนวน 30 เว็บเพจ 2) Healthcare จำนวน 30 เว็บเพจ และ 3) Transportation จำนวน 30 เว็บเพจ และชุดข้อมูลที่ 2 สุ่มข้อมูลเว็บเพจมาจากทั้ง 7 คลาส คลาสละ 30 เว็บเพจ รวมเป็น 210 เว็บเพจ ทดสอบแบบ 10-fold Cross Validation โดยใช้โปรแกรม Weka 3.6.1 [12]

4.2 ผลการทดลอง

จากข้อมูลเว็บเพจทั้ง 2 ชุด เมื่อผ่านขั้นตอนการเตรียมข้อมูลเว็บเพจ ชุดข้อมูลที่ 1 จะได้ลักษณะเฉพาะจากข้อความจำนวน 5,996 ลักษณะเฉพาะ และจากหัวเรื่องจำนวน 171 ลักษณะเฉพาะ ส่วนชุดข้อมูลที่ 2 จะได้ลักษณะเฉพาะจากข้อความจำนวน 10,861 ลักษณะเฉพาะ และจากหัวเรื่องจำนวน 354 ลักษณะเฉพาะ หลังจากเลือกลักษณะเฉพาะด้วยค่าความถี่เอกสารโดยใช้ Threshold เท่ากับ 10 จะได้ลักษณะเฉพาะข้อความและหัวเรื่องรวมกันของชุดข้อมูลที่ 1 และ 2 เท่ากับ 182 และ 853 ลักษณะเฉพาะ ตามลำดับ จากนั้นกำหนดลักษณะเฉพาะด้วยวิธี IG จำนวน 100 ลักษณะเฉพาะ เพื่อนำไปใช้ในขั้นการตอนการเลือกลักษณะเฉพาะด้วยวิธี WPC_FCA_FR ซึ่งตัวอย่างลักษณะเฉพาะที่เด่นๆ จำนวน 30 ลักษณะเฉพาะที่ได้จากชุดข้อมูลที่ 1 แสดงดังรูปที่ 2

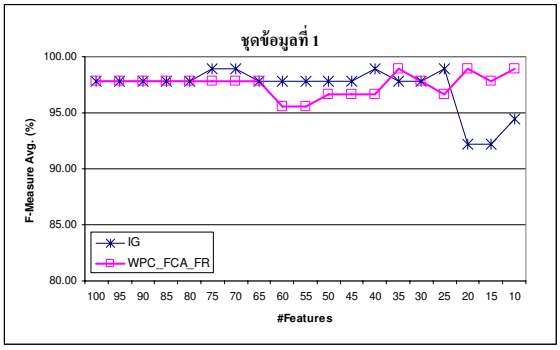
loan byte freight health site care capit return investor financi medic product cash airborn quarter depot credit prefer manag internet stock system servic person includ price custom market support compress

รูปที่ 2 ตัวอย่างลักษณะเฉพาะที่ได้จากชุดข้อมูลที่ 1

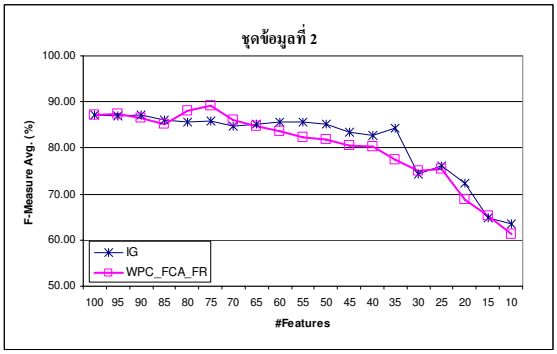
ตารางที่ 4 ค่าเฉลี่ย F-measure จากการจำแนกด้วย MLP-NN

จำนวนลักษณะเฉพาะ (Features)	ค่าเฉลี่ย F-measure (%)			
	ชุดข้อมูลที่ 1		ชุดข้อมูลที่ 2	
	IG	WPC_FCA_FR	IG	WPC_FCA_FR
100	97.80	97.80	87.10	87.10
95	97.80	97.80	87.00	87.30
90	97.80	97.80	87.10	86.60
85	97.80	97.80	86.10	85.10
80	97.80	97.80	85.60	88.00
75	98.90*	97.80	85.80	**89.20***
70	98.90*	97.80	84.80	86.10
65	97.80	97.80	85.10	84.70
60	97.80	95.57	85.70	83.70
55	97.80	95.57	85.70	82.40
50	97.80	96.68	85.20	81.80
45	97.80	96.68	83.30	80.50
40	98.90*	96.68	82.80	80.30
35	97.80	**98.90***	84.20	77.40
30	97.80	97.80	74.40	74.90
25	98.90*	96.68	76.20	75.40
20	92.22	**98.90***	72.40	68.70
15	92.22	97.80	64.80	65.20
10	94.43	**98.90***	63.40	61.30

* ค่าเฉลี่ย F-measure ที่สูงที่สุด



รูปที่ 3 เปรียบเทียบค่าเฉลี่ย F-measure ในชุดข้อมูลที่ 1



รูปที่ 4 เปรียบเทียบค่าเฉลี่ย F-measure ในชุดข้อมูลที่ 2

จากลักษณะเฉพาะ 100 ลักษณะเฉพาะ ทำการเลือกลักษณะเฉพาะด้วยวิธี WPC_FCA_FR โดยลดจำนวนลงทีละ 5 ลักษณะเฉพาะ จนเหลือ 10 ลักษณะเฉพาะ จากนั้น

จำแนกประเภทด้วย MLP-NN และเปรียบเทียบค่าเฉลี่ย F-measure จากการเลือกลักษณะเฉพาะด้วยวิธี WPC_FCA_FR กับการเลือกลักษณะเฉพาะด้วยวิธี IG ซึ่งผลการทดลองแสดงดังตารางที่ 4

4.2.1 ประสิทธิภาพด้านการจำแนกประเภท

จากตารางที่ 4 เมื่อพิจารณาจากค่าเฉลี่ย F-measure ที่ได้จากการจำแนกด้วย MLP-NN จะเห็นว่าผลการทดลองในชุดข้อมูลที่ 1 การเลือกลักษณะเฉพาะด้วยวิธี WPC_FCA_FR และวิธี IG ให้ค่าเฉลี่ย F-measure ที่สูงที่สุดเท่ากันคือ 98.90 % และผลการทดลองในชุดข้อมูลที่ 2 การเลือกลักษณะเฉพาะด้วยวิธี WPC_FCA_FR ให้ค่าเฉลี่ย F-measure ที่สูงที่สุดคือ 89.20% ซึ่งสูงกว่าวิธี IG ที่ให้ค่าเฉลี่ย F-measure สูงที่สุดเพียง 87.10% ซึ่งกราฟเปรียบเทียบค่าเฉลี่ย F-measure แสดงดังรูปที่ 3 และ 4

4.2.2 ประสิทธิภาพด้านการลดขนาดลักษณะเฉพาะ

จากตารางที่ 4 จะเห็นว่าผลการทดลองในชุดข้อมูลที่ 1 ที่ค่าเฉลี่ย F-measure สูงที่สุดเท่ากับ 98.90% การเลือกลักษณะเฉพาะด้วยวิธี WPC_FCA_FR สามารถลดขนาดลักษณะเฉพาะเหลือน้อยที่สุดเท่ากับ 10 ซึ่งมีจำนวนน้อยกว่าวิธี IG ที่ลดได้เหลือน้อยที่สุดเท่ากับ 25 และผลการทดลองในชุดข้อมูลที่ 2 การเลือกลักษณะเฉพาะด้วยวิธี IG ที่ค่าเฉลี่ย F-measure สูงที่สุดเท่ากับ 87.10% สามารถลดขนาดลักษณะเฉพาะเหลือน้อยที่สุดเท่ากับ 90 ส่วนวิธี WPC_FCA_FR ที่ค่าเฉลี่ย F-measure สูงที่สุดเท่ากับ 89.20% สามารถลดขนาดลักษณะเฉพาะได้เท่ากับ 75 ซึ่งสามารถลดได้มากกว่าและยังให้ค่าเฉลี่ย F-measure สูงกว่าอีกด้วย

5. บทสรุป

บทความนี้ได้นำเสนอการจำแนกประเภทโดยวิธีการลดขนาดลักษณะเฉพาะด้วย FCA ทำการทดลองจำแนกประเภทด้วย MLP-NN เปรียบเทียบค่าเฉลี่ย F-measure ที่ได้จากการเลือกลักษณะเฉพาะด้วยวิธี IG กับการเลือกลักษณะเฉพาะด้วยวิธีที่นำเสนอ ซึ่งผลการทดลองแสดงให้เห็นว่าการเลือกลักษณะเฉพาะด้วยวิธีที่นำเสนอสามารถลดลักษณะเฉพาะได้มากกว่าและให้ค่าเฉลี่ย F-measure ที่สูงที่สุดในทั้ง 2 ชุดข้อมูล ดังนั้นเราสามารถใช้วิธีการ FCA มาช่วยลดลักษณะเฉพาะได้ ซึ่งช่วยลดเวลาและทรัพยากรในการประมวลผลโดยที่ค่าความถูกต้องยังคงสูงหรือสูงกว่าเดิม

7. กิตติกรรมประกาศ

งานวิจัยนี้ได้รับการสนับสนุนจากคณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่

8. เอกสารอ้างอิง

[1] X. Qi, and D. Davison, “Web Page Classification: Features and Algorithms,” ACM Computing Surveys. 41, 2, 2006, pp. 1-31.

[2] R. Zhang, M. Shepherd, J. Duffy, and C. Watters, “Automatic Web Page Categorization using Principal Component Analysis,” Proceedings of the 40th Hawaii International Conference on System Sciences, 2007, pp. 1-10.

[3] M. Indra Devi, R. Rajaram, and K. Selvakuberan, “Generating best features for web page classification,” Webology, 5(1), Article 52, 2008. Available at: http://www.webology.ir/2008/v5n1/a52.html (February, 2010).

[4] พรพล ธรรมรงค์รัตน์ ลัดดา ปรีชาวีรกุล และวิภาดา เวทย์-ประสิทธิ์, “การจำแนกประเภทเว็บเพจโดยวิธีการลดขนาดลักษณะเฉพาะและมัลติคลาสซัพพอร์ตเวกเตอร์แมชชีน,” Proceedings of the Conference on Knowledge and Smart Technologies 2009, 24-25 กรกฎาคม, 2009, หน้า 103-108.

[5] P. Thamrongrat, L. Preechaveerakul, and W. Wettayaprasit, “A Novel Voting Algorithm of Multi-Class SVM for Web Page Classification,” In Proceedings The 2nd IEEE International Conference on Computer Science and Information Technology (ICCSIT 2009), Beijing, China, August 8-11, 2009, pp. 1-5.

[6] S. B. Park, and B. T. Zhang, “Automatic webpage classification enhanced by unlabeled data,” In Proceedings of the 4th international conference on intelligent data engineering and automated learning, Hong-Kong, China, 2003, pp. 821-825.

[7] R. Wille,R, “Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies,” Springer Verlag, LNAI3626, 2005, pp. 1-33.

[8] S.-H Hwang, H.-G. Kim, and H.-S. Yang, “A FCA-Based Ontology Construction for the Design of Class Hierarchy,” ICCSA 2005, LNCS 3482, 2005, pp. 827-835.

[9] พรพล ธรรมรงค์รัตน์ ลัดดา ปรีชาวีรกุล และวิภาดา เวทย์-ประสิทธิ์, “การจำแนกประเภทเว็บเพจโดยใช้ค่าความถี่เอกสารและซัพพอร์ตเวกเตอร์แมชชีน,” The 12th National Computer Science and Engineering Conference (NCSEC 2008), 2008, หน้า 498-504.

[10] S. B. Kotsiantis, I. B. Zaharakis and P. E. Pintelas, “Machine learning: a review of classification and combining techniques,” Artificial Intelligence Review, vol. 26, 2006, pp. 159-190.

[11] WebKB. [Online]. Available at: http:// www.cs.cmu.edu/~WebKB/ (February, 2010).

[12] Weka. [Online]. Available at: http://www.cs.waikato.ac.nz/ml/weka/ (February, 2010).

ภาคผนวก ค

ผลงานวิจัยที่ได้รับการตีพิมพ์ในงานประชุมวิชาการ ICET 2011

เรื่อง Feature Reduction Using Formal Concept Analysis for Web Page Classification

Conference The 5th PSU-UNS International Conference on Engineering and Technology (ICET 2011)

สถานที่ ภูเก็ต ประเทศไทย

วันที่ 2-3 พฤษภาคม 2554

Feature Reduction Using Formal Concept Analysis for Web Page Classification

Wirat Choonui[1], Wiphada Wettayaprasit[1], Ladda Preechaveerakul[2]
[1]Artificial Intelligence Research Laboratory, [2]*i*STAR Research Laboratory
Prince of Songkla University, Faculty of Science, Department of Computer Science, Thailand
email: wirat218@hotmail.com, wwettayaprasit@yahoo.com, ladda.p@psu.ac.th

Abstract: *Web page classification is one way to encourage users to conveniently find accurate information. However, if the number of web pages was large then the number of input data would be very large also. As a result, web page classification is complex with time-consuming. This paper proposed feature reduction using formal concept analysis for web page classification. The study used the benchmark web page data sets from CMU. The feature selection performance is comparing between the proposed method and the Information Gain. The result demonstrated that the proposed method could reduce more features while F-measure was still very high.*

Key Words: ***Web Page/ Formal Concept Analysis/ Feature Reduction/ Classification***

1. INTRODUCTION

Internet is widely popular and rapid development. The number of web pages increased everyday while searching information to meet the need of users is time-consuming and more complicated. One way to solve this problem is using web page classification [1]. Web page is organized into groups according to the interesting of users and can help users to conveniently find accurate information. Since each web page contains many words, then the feature size of input data is also very large. As a result, web page classification will be very complex. One way to solve this problem is using feature reduction to optimize web page classification.

2. RELATED WORK

Research related to feature selection such as Zhang *et al*. [2] used the Information Gain (IG) and principal component analysis to select feature for automatically grouping web page types. Xin *et al*. [3] used the ReliefF feature selection method for selecting relevant words to improve the classification performance. Indra *et al*. [4] combined the feature selection techniques to select good quality feature with the least amount by using benchmark web page data set from WebKB. Thamrongrat *et al*. [5] used ReliefF, Information Gain, Chi Square, and Gain Ratio for classifying web pages and proposed voting algorithm of multi-class SVM for web page classification called VAMSVM_WPC. The study examined the benchmark web page data sets from CMU by using the features of the text and title and evaluated the performance with *F*-measure.

This paper proposed web page classification by reducing the number of features using Formal Concept Analysis (FCA) tested with the benchmark web page data sets from CMU. The evaluation by *F*-measure was used for this study.

3. FORMAL CONCEPT ANALYSIS

Formal Concept Analysis (FCA) [6,7] is a theory of data analysis which identifies conceptual structures among data sets. These structures are graphically represented as conceptual lattices in order to make them more understandable by allowing the analysis of complex structures and the discovery of dependencies within the data. Data sets are represented as formal context, which is the basic structure of FCA. This paper used FCA for finding the relationship between document (web page) and term (word) to select the appropriate features.

Definition 1. *Formal Context*

A formal context constitutes a triple (G, M, I), where G is the set of objects, M is the set of attributes, and I is binary relation defined between G and M. If an object g has an attribute m then $g \in G$ is related with I to m which is indicated by the relationship $(g,m) \in I$. This means that g includes m.

A formal context can be written as a cross-table. The head of the row is represented by objects and the head of the column is represented by features. An 'x' in the row and the column indicate that the object has the corresponding attributes. For example, Table 1 represents the formal context that consists of 4 objects and 7 attributes. Object 4 is corresponding to attribute 1 and attribute 3 [8].

Definition 2. *′ Operation*

Given A is the set of objects ($A \subseteq G$) and B is the set of attributes ($B \subseteq M$), consider the dual sets A' and B' as follows.

$$A' = \{m \in M \mid \forall_g \in A : (g, m) \in I\} \quad (1)$$

$$B' = \{g \in G \mid \forall_m \in B : (g, m) \in I\} \quad (2)$$

A' is the set defined by the attributes applying to all objects belongs to A and B' is the object with all attributes belongs to B.

Definition 3. *Formal Concept*

A formal concept of the formal context (G, M, I) is defined as a pair (A, B) where $A \subseteq G$, $B \subseteq M$, $A' = B$, and $B' = A$. A is the set of objects called the *extent*, and B is the set of attributes called the *intent*. Table 2 shows the formal concepts which can be generated for the formal context from Table 1. For example in row no. 8, the intent is the set of attributes {a1, a3}, and the extent is the set of object {o4}. The formal concept is ({o4},{a1,a3}).

4. THE PROPOSED APPROACH

This paper proposed Feature Reduction Using Formal Concept Analysis for Web Page Classification (FR_FCA_WPC) model as shown in Fig. 1. The FR_FCA_WPC has 4 steps: 1) Web Page Pre-processing, 2) Feature Selection Using IG, 3) Feature Selection Using FCA, and 4) Classification and Evaluation.

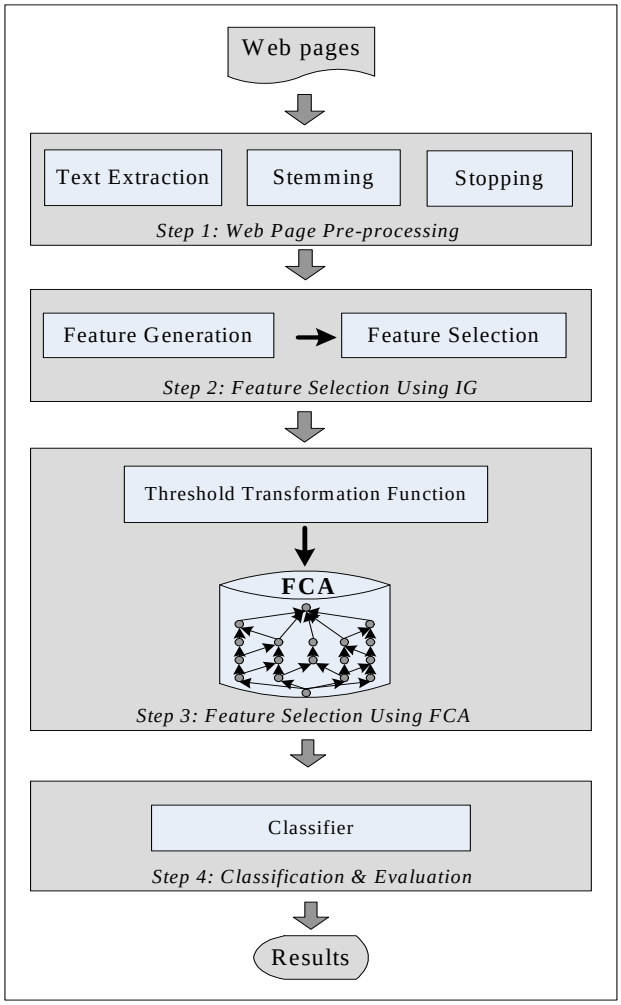


Fig. 1. *FR_FCA_WPC model*

Table 1. *An example of the formal context*

	a1	a2	a3	a4	a5	a6	a7
o1		x				x	x
o2			x	x	x		
o3			x		x	x	x
o4	x		x				

Table 2. *The formal concepts*

No.	Formal concept	Intent	Extent
1	({o1,o2,o3,o4},Ø)	Ø	{o1,o2,o3,o4}
2	({o2,o3,o4},{a3})	{a3}	{o2,o3,o4}
3	({o1,o3},{a6,a7})	{a6,a7}	{o1,o3}
4	({o2,o3}, {a3,a5})	{a3,a5}	{o2,o3}
5	({o1},{a2,a6,a7})	{a2,a6,a7}	{o1}
6	({o3},{a3,a5,a6,a7})	{a3,a5,a6,a7}	{o3}
7	({o2},{a3,a4,a5})	{a3,a4,a5}	{o2}
8	({o4},{a1,a3})	{a1,a3}	{o4}
9	(Ø,{a1,a2,a3,a4,a5,a6,a7})	{a1,a2,a3,a4,a5,a6,a7}	Ø

4.1 Web Page Pre-processing

This step is the process of extracting words from text and title of the web pages. Then words are stemmed using Porter stemming algorithm and stop words are eliminated.

4.2 Feature Selection Using IG

The words from pre-processing step will generate in term of document matrix as shown in Table 3, and will receive weight value using *tf-idf* method as (3).

$$w_{j,k} = tf_{j,k} \times idf_k \quad (3)$$

where $w_{j,k}$ is the weight of document j with word k (k = 1, 2, 3, ..., m, where m is the total number of words from all documents, and j = 1, 2, 3, ..., n where n is the total number of documents), $tf_{j,k}$ is the term frequency of word k occurs in document j, and idf_k is the inverse document frequency of word k. The inverse document frequency of a word is low if it occurs in many documents and is the highest if the word occurs only once [8].

The features are selected from words with document frequency rather than predefined threshold. Finally, the IG feature selection method is used to select the features for reducing the number of features.

Table 3. *The document-term matrix*

Terms / Documents	t_1	t_2	t_3	…	t_m
d_1	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$	...	$w_{1,m}$
d_2	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$	...	$w_{2,m}$
d_3	$w_{3,1}$	$w_{3,2}$	$w_{3,3}$	...	$w_{3,m}$
…	...	...	...	...	...
d_n	$w_{n,1}$	$w_{n,2}$	$w_{n,3}$	…	$w_{n,m}$

4.3 Feature Selection Using FCA

In this step, FCA is used for examining the relationship between document (web page) and term (word) to select the appropriate features. Let r be the number of features selected by IG, and s be the number of features selected by FCA. The algorithm will try to reduce the number of features with $s \leq r$ while maintain the same or higher accuracy. The proposed steps are as follows.

4.3.1 Formal Context Generation

The features selected by IG will convert to formal context using threshold transformation function in Definition 4.

Definition 4. *Threshold Transformation Function*

Let λ be the weight assigned as the threshold to determine the relationship between words and documents. Equation (4) means that if $w_{j,k}$ is greater than λ, then $f(w_{j,k})$ is 1, otherwise $f(w_{j,k})$ is 0. The threshold transformation function is shown in Fig 2.

$$f(w_{j,k}) = \begin{cases} 1; & if\ w_{j,k} > \lambda \\ 0; & if\ w_{j,k} \leq \lambda \end{cases} \quad (4)$$

The threshold transformation function $f(w_{j,k})$ uses a threshold of λ causing the output to be either 1 or 0, i.e., if $f(w_{j,k}) = 1$ ($w_{j,k} > \lambda$) indicates that $(d_j,t_k) \in I$ (document j is related to word k), and if $f(w_{j,k}) = 0$ ($w_{j,k} \leq \lambda$) indicates that $(d_j,t_k) \notin I$ (document j is not related to word k).

Table 4 shows an example of the results for feature selection using IG. The weight of document d1 (j = 1) and feature f1 (k = 1) is 1.3 ($w_{1,1}$ = 1.3). Table 5 shows an example of the formal context which is generated from Table 4 where λ = 0. It contains the document of 10 web pages (d1, d2, d3, ..., d10), 6 features (f1, f2, f3, ..., f6), and 2 classes (c1 and c2). An 'x' in row and column indicates that the document is related to the feature.

For example in Table 4 and Table 5, $w_{1,1}$ = 1.3, $w_{3,6}$ = 0.7, $w_{2,6}$ = 0, if λ = 0, then $f(w_{1,1})$ = 1 because $w_{1,1} > 0$, $f(w_{3,6})$ = 1 because $w_{3,6} > 0$, and $f(w_{2,6})$ = 0 because $w_{2,6} \leq 0$. In Table 6, if λ = 1.0, then $f(w_{1,1})$ = 1 because $w_{1,1} > 1.0$, $f(w_{3,6})$ = 0 because $w_{3,6} \leq 1.0$, and $f(w_{2,6})$ = 0 because $w_{2,6} \leq 1.0$. Note that if λ = 1.0, we can eliminate feature f6 because $w_{1,6} \leq 1.0$, $w_{2,6} \leq 1.0$, $w_{3,6} \leq 1.0$, …, $w_{10,6} \leq 1.0$.

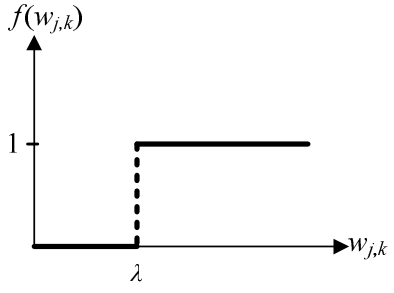


Fig. 2. *Threshold Transformation Function*

Table 4. *An example of document-term matrix*

Features / Web pages	f1	f2	f3	f4	f5	f6	class
d1	1.3	0	0	0	0	0	c1
d2	2.9	0	0	1.3	0	0	c1
d3	2.9	0	0	1.3	0	0.7	c1
d4	1.5	0	0	1.3	0	0.7	c1
d5	1.5	0.2	0	1.3	0	0.7	c1
d6	0	0	3.0	0	0	0	c2
d7	0	0	2.6	0	1.8	0	c2
d8	0	0	1.6	0	1.8	0	c2
d9	0.2	1.6	1.6	0	1.5	0	c2
d10	0.2	1.6	1.6	0	1.5	0	c2

Table 5. *An example of the formal context (λ = 0)*

	f1	f2	f3	f4	f5	f6	c1	c2
d1	x						x	
d2	x			x			x	
d3	x			x		x	x	
d4	x			x		x	x	
d5	x	x		x		x	x	
d6			x					x
d7			x		x			x
d8			x		x			x
d9	x	x	x		x			x
d10	x	x	x		x			x

Table 6. *An example of the formal context (λ = 1.0)*

	f1	f2	f3	f4	f5	c1	c2
d1	x					x	
d2	x			x		x	
d3	x			x		x	
d4	x			x		x	
d5	x			x		x	
d6			x				x
d7			x		x		x
d8			x		x		x
d9		x	x		x		x
d10		x	x		x		x

4.3.2 Formal Concept Generation

We generate the formal concept from formal context in Table 5 by using Definition 2 and 3. The extent of the concept is a set of document (d) and the intent of the concept is a set of feature (f) and class (c). In Table 5, the formal context can be generated to the formal concept, which have 7 concepts that the intent includes the feature of class which sorted by the number of members in extent by descending, as shown in Table 7.

Table 7. *An example of the formal concept (λ = 0)*

No.	Formal concept	The number of members in extent
1	({d1,d2,d3,d4,d5}, {f1,c1})	5
2	({d6,d7,d8,d9,d10}, {f3,c2})	5
3	({d2,d3,d4,d5}, {f1,f4,c1})	4
4	({d7,d8,d9,d10}, {f3,f5,c2})	4
5	({d3,d4,d5}, {f1,f4,f6,c1})	3
6	({d9,d10}, {f1,f2,f3,f5,c2})	2
7	({d5}, {f1,f2,f4,f6,c1})	1

For example, formal concept no. 1 in Table 7, the extent of concept is {d1, d2, d3, d4, d5} and the intent of concept is {f1, c1}. The members in extent equals to 5, which means that feature f1 is related to all 5 documents belonging to class c1, and feature f1 is the dominant feature of class c1.

4.3.3 Feature Selection

The features can be selected from the formal concept with the highest number of members in the extent. For example in Table 7, the formal concepts no.1 and no. 2 have the same highest number of member in the extent equals to 5, therefore feature f1 and f3 are selected first. Then feature f4 and f5 are selected, next feature f6 and f2 are selected. Finally, the sequence features that are selected by FCA are f1, f3, f4, f5, f6, and f2, respectively.

4.4 Classification and Evaluation

The experiment used Multi-Layer Perceptron Neural Networks (MLP-NN) [10] to classify the features and tested with 10-folds Cross Validation using WEKA 3.6.1 [11]. The classification performance will be evaluated with the *F*-measure.

5. EXPERIMENTAL RESULTS

5.1 Data sets

This paper used the benchmark web page data set from CMU [12]. There are 4,581 web pages divided into 7 classes: Energy, Financial, Technology, Healthcare, Materials, Utilities, and Transportation. The data set D1 is 90 web pages, which was randomly selected from financial 30 web pages, Healthcare 30 web pages, and Transportation 30 web pages. The data set D2 composes of 200 web pages divided into 2 classes, which was randomly selected from Financial 100 web pages and 100 web pages from the remaining 6 classes.

5.2 Experimental results

The data set D1 has 5,996 features from texts and 171 features from titles. The data set D2 has 9,968 features from texts and 304 features from titles. After the features are selected from words with document frequency, then the data set D1 has 353 features and the data set D2 has 802 features. The next step is using IG to reduce the features to 30. After that, reduce the features with λ=0, λ=0.5, and λ=1.0. The experimental results using the proposed FR_FCA_WPC method comparing with IG method for the data set D1 and D2 are shown in Table 8 and 9, respectively.

5.2.1 The classification performance

From data set D1 in Table 8, the *F*-measure of FR_FCA_WPC method (λ=0 and λ=0.5) and IG method is the same at 98.9%. The *F*-measure of FR_FCA_WPC method at λ=1.0 is higher at 100.0% when compares with IG method at 98.9% where the number of feature is 25. From data set D2 in Table 9, the *F*-measure of FR_FCA_WPC method at λ=0, λ=0.5, and λ=1.0 which are 92.0%, 90.5%, and 91.5%, is higher when compares with IG method (89.0%).

From data set D1, Fig. 3 shows that the FR_FCA_WPC method comparing with the difference λ parameter, IG method and the number of features, the FR_FCA_WPC method at λ=1.0 gives the highest *F*-measure for the number of features at 25 and 20, at λ=0 and λ=0.5 give the highest *F*-measure at the number of features at 15 and 10. But IG method has the lowest *F*-measure. From data set D2, Fig. 4 shows that the FR_FCA_WPC method at λ=0 give the highest *F*-measure at the number of features at 30, 20 and 10 while most of IG method has the lowest *F*-measure.

5.2.2 The feature reduction performance

The FR_FCA_WPC method can reduce the number of features less than the IG method. For example in Table 8 from data set D1, For IG method the highest *F*-measure is 98.9% (the number of feature is 25), the FR_FCA_WPC method can reduce the number of features to 10, at λ=0 and λ=0.5, while the *F*-measure is still the same at 98.9%. In Table 9 from data set D2, the highest *F*-measure of IG method is 89% with the number of feature at 30, at λ=0 , λ=0.5, and λ=1.0, the FR_FCA_WPC can reduce the number of feature to 25 while the *F*-measure are higher than IG method at 89.5%, 90.5%, and 89.5%, respectively.

Table 8. *The F-measure for the data set D1*

#Features	The *F*-measure for the data set D1 (%)			
	IG	FR_FCA_WPC		
		λ=0	λ=0.5	λ=1.0
30	97.8	97.8	97.8	97.8
25	**98.9**	97.8	97.8	**100.0**
20	92.2	95.5	95.5	97.8
15	92.2	97.8	97.8	94.4
10	94.4	**98.9**	**98.9**	97.8

Table 9. *The F-measure for the data set D*2

#Features	The *F*-measure for the data set D2 (%)			
	IG	FR_FCA_WPC		
		λ=0	λ=0.5	λ=1.0
30	**89.0**	**92.0**	**90.5**	**91.5**
25	86.0	89.5	**90.5**	89.5
20	86.4	88.5	87.5	88.0
15	84.9	85.5	87.5	82.5
10	83.0	86.5	84.0	83.5

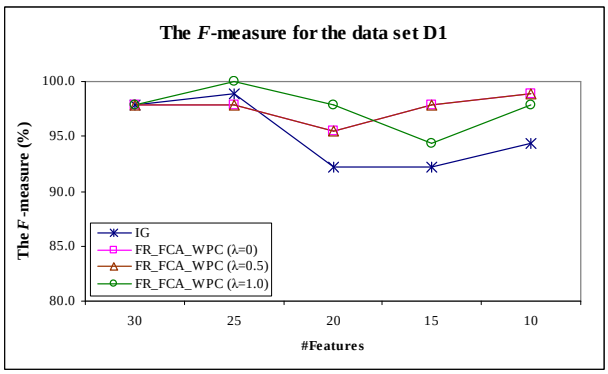


Fig. 3. *Comparing the F-measure for data set D1*

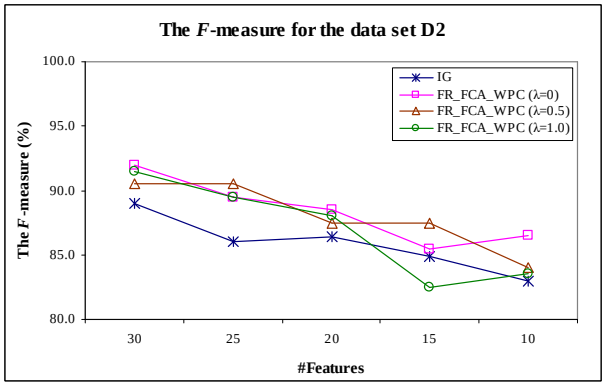


Fig. 4. *Comparing the F-measure for data set D2*

5.2.3 The performance of λ

The results of experiment for both data set D1 and D2 show that the highest *F*-measure of FR_FCA_WPC method depends on the different weight values of λ. For example, the highest *F*-measure of data set D1 at λ=0, λ=0.5, and λ=1.0 are 98.9%, 98.9% and 100.0%, respectively. The highest *F*-measure of data set D2 at λ=0, λ=0.5, and λ=1.0 are 92.0%, 90.5% and 91.0%, respectively. Note that the suitable λ needs to be considered. If λ is high then the number of features will be small, but it may give the lower *F*-measure.

6. CONCLUSIONS

This paper presents a method to reduce the number of features using FCA for classifying web pages. The experiment used MLP-NN classifier. The experimental results show the *F*-measure of FR_FCA_WPC for both data sets are higher when compares with IG method. Therefore, FR_FCA_WPC method can reduce the number of features less than IG method while the *F*-measure is still very high.

7. REFERENCES

[1] X. Qi, and D. Davison, "Web Page Classification: Features and Algorithms", ACM Computing Surveys. 41, 2, 2006, pp. 1-31.

[2] R. Zhang, M. Shepherd, J. Duffy, and C. Watters, "Automatic Web Page Categorization using Principal Component Analysis", *in Proc. of the 40th Hawaii International Conference on System Sciences*, 2007, pp. 1-10.

[3] J. Xin, L. Rongyan, S. Xian, B. Rongfang, "Automatic Web Pages Categorization with ReliefF and Hidden Naive Bayes", *in Proc. of the 2007 ACM Symposium on Applied Computing*, Seoul Korea, 2007, pp. 617-621.

[4] M. Indra Devi, R. Rajaram, and K. Selvakuberan, "Generating best features for web page classification," Webology, 5(1), Article 52, 2008.

[5] P. Thamrongrat, L. Preechaveerakul, and W. Wettayaprasit, "A Novel Voting Algorithm of Multi-Class SVM for Web Page Classification", *in Proc. of the 2nd IEEE International Conference on Computer Science and Information Technology (ICCSIT 2009)*, Beijing, China, August 8-11, 2009, pp. 327-330.

[6] R. Wille, "Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies", Springer Verlag, LNAI3626, 2005, pp. 1-33.

[7] U. Priss. "Formal Concept Analysis in Information Science", Annual Review of Information Science and Technology, Vol. 40, No. 1, 2006, pp. 521-543.

[8] S. H Hwang, H. G. Kim, and H. S. Yang, "A FCA-Based Ontology Construction for the Design of Class Hierarchy", *ICCSA 2005*, LNCS 3482, 2005, pp. 827-835.

[9] L. P. Jing, H. K. Huang, and H. B. Shi, "Improved feature selection approach TFIDF in text mining", *Proceeding of the First International Conference on Machine Learning and Cybernetics*, Beijing, 4-5 November, 2002, pp. 994-996.

[10] S. B. Kotsiantis, I. B. Zaharakis and P. E. Pintelas, "Machine learning: a review of classification and combining techniques", Artificial Intelligence Review, Vol. 26, 2006, pp. 159-190.

[11] WEKA. [Online]. Available at: http://www.cs.waikato.ac.nz/ml/weka/ (February, 2011).

[12] WebKB. [Online]. Available at: http://www.cs.cmu.edu/~WebKB/ (February, 2011).

ประวัติผู้เขียน

ชื่อ สกุล นายวิรัตน์ ชูนุ้ย

รหัสประจำตัวนักศึกษา 5110220073

วุฒิการศึกษา

วุฒิ	ชื่อสถาบัน	ปีที่สำเร็จการศึกษา
วท.บ. (คณิตศาสตร์ประยุกต์)	มหาวิทยาลัยสงขลานครินทร์	2545

การตีพิมพ์เผยแพร่ผลงาน

วิรัตน์ ชูนุ้ย, วิภาดา เวทย์ประสิทธิ์ และ ลัดดา ปรีชาวีรกุล. 2553. การจำแนกประเภทเว็บเพจโดยวิธีการลดขนาดลักษณะเฉพาะด้วย FCA. The 7th International Joint Conference on Computer Science and Software Engineering (JCSSE 2010). กรุงเทพมหานคร ประเทศไทย. หน้า 235-240.

Choonui, W., Wettayaprasit, W., and Preechaveerakul, L. 2011. Feature Reduction Using Formal Concept Analysis for Web Page Classification. Proceeding of The 5th PSU-UNS International Conference on Engineering and Technology (ICET-2011). Phuket, Thailand, pp. 320-324.