การคัดกรองสแปมเมลโดยประยุกต์ใช้การจำแนกประเภทแบบฟัซซีและนาอีฟเบเซียน

FUZZY AND NAIVE BAYESIAN CLASSIFICATION FOR SPAM MAIL FILTERING

โดย

นิวภรณ์ หัวนาวิลาส

NIWAPORN HUANAWILAS

อาจารย์ที่ปรึกษา

รศ.ดร.อาริต ธรรมโน

รายงานนี้เป็นส่วนหนึ่งของวิชาศึกษาอิสระ 2
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคฤดูร้อน ปีการศึกษา 2553

FUZZY AND NAIVE BAYESIAN CLASSIFICATION FOR SPAM MAIL FILTERING

A REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF THE COURSE

INDEPENDENT STUDY 2

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

KING MONGKUT`S INSTITUTE OF TECHNOLOGY LADKRABANG

Summer/ 2010

ใบรับรองการศึกษาอิสระ 2 (Independent Study 2)

เรื่อง

การคัดกรองสแปมเมลโดยประยุกต์ใช้การจำแนกประเภท แบบฟัซซีและนาอีฟเบเซียน

Fuzzy and Naive Bayesian Classification for Spam Mail Filtering

นางสาวนิวภรณ์ หัวนาวิลาส

รหัสประจำตัว 52660513

ขอรับรองว่ารายงานฉบับนี้ ข้าพเจ้าไม่ได้คัดลอกมาจากที่ใด
รายงานฉบับนี้ได้รับการตรวจสอบและอนุมัติให้เป็นส่วนหนึ่งของ
การศึกษาวิชาการศึกษาอิสระ 2 หลักสูตรวิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศ)
ภาคฤดูร้อน ปีการศึกษา 2553

.....อาจารย์ที่ปรึกษา
(รศ.ดร.อาริต ธรรมโน)

.....กรรมการสอบ
(รศ.ดร.วรพจน์ กรีสุระเดช)

.....กรรมการสอบ
(ผศ.ดร.พรฤดี เนติโสภากุล)

หัวข้อ	การคัดกรองสแปมเมลโดยประยุกต์ใช้การจำแนกประเภทแบบฟัซซีและนาอีฟเบเซียน
นักศึกษา	นางสาวนิวภรณ์ หัวนาวิลาส
รหัสนักศึกษา	52660513
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	เทคโนโลยีระบบสารสนเทศ
ปีการศึกษา	2553
อาจารย์ที่ปรึกษา	รศ.ดร.อาริต ธรรมโน

บทคัดย่อ

อีเมลเป็นสื่อที่สำคัญในการติดต่อสื่อสารในระบบธุรกิจปัจจุบัน เนื่องจากเป็นวิธีการติดต่อสื่อสารที่สะดวก รวดเร็วและมีต้นทุนต่ำ บ่อยครั้งที่พบว่าข้อมูลได้รับมานั้นเข้าข่ายเป็นสแปมเมล ในการนี้ได้มีงานศึกษาไม่น้อยที่ได้นำเสนอเทคนิคการกรองสแปมเมลแบบต่างๆอยู่เสมอ งานศึกษานี้มีวัตถุประสงค์เพื่อพัฒนาโปรแกรมคัดกรองสแปมเมล 2 เทคนิควิธีคือ เทคนิคการกรองสแปมแบบเบเซียนซึ่งใช้กระบวนการจำแนกประเภทแบบนาอีฟเบเซียน และเทคนิคฟัซซี-นาอีฟเบเซียน เพื่อเปรียบเทียบประสิทธิภาพและความสามารถในการคัดกรองสแปมเมลจากแบบจำลองที่สร้างขึ้นจากทั้งสองวิธี

ในการพัฒนาการจำแนกประเภทแบบฟัซซี-นาอีฟเบเซียน ผู้ศึกษาได้ประยุกต์ใช้ฟัซซีลอจิกร่วมกับนาอีฟเบเซียน ซึ่งการรวมทฤษฎีทั้งสองเข้าด้วยกันนั้นเป็นการประสานข้อดีของทั้งสองวิธีเข้าด้วยกัน ในการทดสอบความสามารถการแยกประเภทข้อมูลจากแบบจำลองที่สร้างขึ้นได้ทำการศึกษาเปรียบเทียบกับตัวกรองสแปมแบบเบเซียนจากการทดสอบข้อมูลชุดอีเมลตัวอย่าง ผลการทดสอบประสิทธิภาพแสดงให้เห็นว่า ตัวกรองสแปมแบบฟัซซี-นาอีฟเบเซียนใช้เวลาในการเรียนรู้ข้อมูลนานกว่าตัวกรองตัวกรองสแปมแบบเบเซียน แต่อย่างไรก็ตามเทคนิคการกรองสแปมแบบฟัซซี-นาอีฟเบเซียน มีความถูกต้องแม่นยำในการแยกข้อมูลสูงกว่าเทคนิคตัวกรองสแปมเบเซียนแบบเดิม

Title	Fuzzy and Naive Bayesian Classification for Spam Mail Filtering
Student	Miss Niwaporn Huanawilas
Student ID.	52660513
Degree	Master of Science
Program	Information Technology
Major	Information System Technology
Academic Year	2010
Advisor	Associate Professor Arit Thummano Ph.D.

ABSTRACT

E-mail is an important communication in today's business due to its rapidity, low cost and convenience. In some situation, emails received often became spam which directly affects computer system. There are a number of research papers reporting spam filtering algorithm. With this regards, the objectives of this research is to develop Spam Filtering Systems from Bayesian Spam filtering (using Naive Bayesian Classifier) and Fuzzy-Naive Bayesian Classification, and to compare the performance of the two techniques.

In the development process of Fuzzy-Naive Bayesian Classifier, a classification model employed the Fuzzy Logic (a theory about the uncertainty of data) and Naive Bayesian Classifier (a statistical method for classification). The synergy of two models results in a hybrid system which takes benefits of two techniques. To evaluate the performance, the developed system was compared with Bayesian Spam Filtering technique by using spam corpus dataset. From the experimental simulations, the result showed that Fuzzy and Naive Bayesian technique outperformed Bayesian Spam Filtering technique.

กิตติกรรมประกาศ

ขอกราบขอบพระคุณอาจารย์ รศ.ดร. อาริต ธรรมโน อาจารย์ที่ปรึกษาโครงการ ที่ได้กรุณาสละเวลาให้ความรู้ คำปรึกษา แนะนำมาโดยตลอด ทำให้โครงการพัฒนาระบบงานนี้สำเร็จลุล่วงไปได้ด้วยดี

นอกจากนี้ขอกราบขอบพระคุณ คณาจารย์ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังทุกท่าน ที่ได้ถ่ายทอดวิชาความรู้ในด้านต่างๆ ทำให้ผู้พัฒนาสามารถนำความรู้ที่ได้มาประยุกต์ใช้ในการทำสารนิพนธ์นี้ได้อย่างมีประสิทธิภาพ

ขอกราบขอบพระคุณหัวหน้างาน ที่ได้ส่งเสริมและสนับสนุนเรื่องการศึกษาด้วยดีตลอดมา ตลอดจนขอกราบขอบพระคุณบุพการี บุคคลในครอบครัว และเพื่อนๆคณะเทคโนโลยีสารสนเทศทุกท่านซึ่งได้ให้กำลังใจในการทำงานแก่ผู้จัดทำเสมอมา ข้าพเจ้าหวังเป็นอย่างยิ่งว่าโครงการฉบับนี้จะเป็นประโยชน์ต่อผู้อ่าน หรือผู้ทำการศึกษาไม่มากก็น้อย

นิวภรณ์ หัวนาวิลาส

สารบัญ

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

สารบัญรูป(ต่อ)

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

อีเมลเป็นสื่อที่สำคัญในการติดต่อสื่อสารในระบบธุรกิจปัจจุบัน แต่อีเมลก็ได้นำมาซึ่งปัญหาต่างๆตามมาเช่นกัน เช่น การโจมตีทางอีเมล (Bomb Mail) การหลอกลวงหรือปลอมแปลงผู้ส่ง (Email Spoofing) การกระจายไวรัสคอมพิวเตอร์ผ่านทางอีเมล แม้กระทั่งการโฆษณาชวนเชื่อหรือสแปมเมล (Spam Mail) การส่งสแปมเมลเริ่มแพร่หลายเนื่องจากค่าใช้จ่ายในการส่งข้อความผ่านทางระบบอิเล็กทรอนิกส์นั้นมีค่าใช้จ่ายไม่มากนักเมื่อเทียบการการส่งข้อความชักชวนทางอื่น เช่นทางจดหมาย หรือการโฆษณาทางสื่อต่างๆ ทำให้ผู้ส่งข้อความเชิญชวนประหยัดค่าใช้จ่ายในการส่งข้อความ

สแปมเมลได้ก่อความรำคาญใจแก่ผู้รับข้อความเป็นอย่างมาก ต้องเสียเวลาในการอ่านอีเมลเพื่อคัดแยกอีเมลดี (Ham) และอีเมลขยะ (Spam) ทั้งยังอาจกระทบประสิทธิภาพการสื่อสารเครือข่ายคอมพิวเตอร์หรือประสิทธิภาพระบบอีเมลที่ต้องรับส่ง หรือพื้นที่ในเก็บสแปมเมลมากขึ้น การคัดกรองสแปมเมลนั้นทำได้หลากหลายวิธีด้วยกัน และหนึ่งในเทคนิคที่ได้รับความนิยมคือ เทคนิคการคัดกรองสแปมแบบเซียน (Bayesian Spam Filtering) เป็นวิธีที่ได้รับความนิยม เนื่องจากเป็นวิธีการกรองสแปมเมลตามวิธีทางสถิติ ประยุกต์ใช้กระบวนการจำแนกประเภทแบบนาอีฟเบเซียนในการระบุ สแปมเมลโดยการตรวจจับข้อความในอีเมล ซึ่งวิธีการคัดกรองสแปมแบบเบเซียนนี้มีความถูกต้องแม่นยำในการคัดกรองสแปมเมลสูงถึง 99.9 เปอร์เซ็นต์

จากข้อมูลข้างต้น สามารถกล่าวได้ว่าเทคนิคการจำแนกประเภทแบบนาอีฟเบเซียน เป็นเทคนิคที่มีความสามารถในการใช้ทำนายลักษณะหรือข้อมูลที่ไม่ทราบล่วงหน้า จึงมีแนวคิดที่จะศึกษาและเปรียบเทียบประสิทธิภาพการคัดกรองสแปมเมล ระหว่างเทคนิคการจำแนกประเภทแบบนาอีฟเบเซียน (Naive Bayesian) และเทคนิคการจำแนกประเภทแบบฟัซซี-นาอีฟเบเซียน (Fuzzy-Naive Bayesian Classifier) ซึ่งเป็นเทคนิคที่เกิดขึ้นใหม่ในงานศึกษานี้ โดยนำทฤษฎีฟัซซีเซ็ตมาประกอบกับเทคนิคนาอีฟเบเซียน ทั้งนี้เพื่อนำผลที่ได้มาวิเคราะห์หาวิธีที่เหมาะสม และนำไปใช้ในการพิจารณาคัดกรองชุดอีเมลตัวอย่างให้มีความถูกต้องและแม่นยำมากยิ่งขึ้นต่อไป

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

1) เพื่อสร้างโปรแกรมการคัดกรองสแปมเมลโดยใช้เทคนิคการการจำแนกประเภทแบบนาอีฟเบเซียน และเทคนิคการจำแนกประเภทแบบฟัซซี-นาอีฟเบเซียน เพื่อใช้ในการจำแนกประเภทอีเมลออกเป็น 2 ประเภทคือ สแปมและแฮม
2) เพื่อเปรียบเทียบประสิทธิภาพและความถูกต้องแม่นยำของโปรแกรมการคัดกรองสแปมเมลที่สร้างขึ้นจากทั้ง 2 เทคนิควิธีที่พัฒนาขึ้น

1.3 ทฤษฎีที่ใช้ในการศึกษา

1) ทฤษฎีแรกเริ่มที่เป็นพื้นฐานหลักที่ก่อให้เกิดการศึกษานี้คือ ทฤษฎีความน่าจะเป็น (Bayes' Theorem) และเทคนิคการคัดกรองสแปมแบบเบเซียน ซึ่งเป็นการประยุกต์ใช้กระบวนการจำแนกประเภทแบบนาอีฟเบเซียน (Naive Bayesian Classifier) ในการจำแนกประเภทข้อความ (Text Classification) การจำแนกประเภทแบบนาอีฟเบเซียนนี้เป็นกระบวนการจำแนกประเภททางสถิติในรูปแบบความน่าจะเป็น ภายใต้สมมุติฐานที่เหตุการณ์แต่ละเหตุการณ์ไม่ขึ้นตรงต่อกันอย่างมีเงื่อนไข และเป็นกระบวนการที่ต้องได้รับการฝึกสอนให้รู้จักประเภทที่จะจำแนก โดยใช้หลักการการวิเคราะห์คำศัพท์ที่ปรากฏในอีเมลตัวอย่างจำนวนหนึ่ง
2) ทฤษฎีประกอบที่นำมาประยุกต์ใช้ เพื่อช่วยให้การจำแนกประเภทข้อความชัดเจนยิ่งขึ้นคือ ทฤษฎีระบบฟัซซี (Fuzzy System) เครื่องมือช่วยในการตัดสินใจภายใต้ความไม่แน่นอนของข้อมูล เพื่อช่วยปรับค่าความน่าจะเป็นสแปมของคำศัพท์แต่ละคำให้เหมาะสม โดยประยุกต์ให้มีฟัซซีเซต 2 เซต คือเซตสแปมและเซตแฮม

1.4 ขอบเขตการศึกษา

1) ศึกษาการคัดกรองสแปมแบบเบเซียน กระบวนการจำแนกประเภทแบบนาอีฟเบเซียน และทฤษฎีระบบฟัซซี
2) พัฒนาโปรแกรมคัดกรองสแปมเมลโดยใช้ตัวกรองสแปมแบบเบเซียน ซึ่งเป็นการประยุกต์ใช้กระบวนการจำแนกประเภทแบบนาอีฟเบเซียน

3) ทดสอบการจำแนกประเภทอีเมลของโปรแกรมคัดกรองสแปมเมลแบบเบเซียน ด้วยชุดตัวอย่างอีเมลจำนวนหนึ่ง เพื่อวัดประสิทธิภาพการทำงานและความถูกต้องในการจำแนกประเภทอีเมล
4) ปรับปรุงโปรแกรมคัดกรองสแปมเมล โดยนำระบบฟัซซีเข้ามาประยุกต์ใช้ร่วมกับตัวกรองสแปมแบบเบเซียน เพื่อเพิ่มประสิทธิภาพในการคัดกรองสแปมเมล ซึ่งเรียกเทคนิคที่พัฒนาขึ้นใหม่นี้ว่าเทคนิคการจำแนกประเภทแบบฟัซซี-นาอีฟเบเซียน
5) ทดสอบการจำแนกประเภทอีเมลของโปรแกรมคัดกรองสแปมเมลแบบฟัซซี-นาอีฟเบเซียน เพื่อวัดประสิทธิภาพการทำงานและความถูกต้องในการจำแนกประเภทอีเมล
6) เปรียบเทียบประสิทธิภาพการทำงานและความถูกต้องในการจำแนกประเภทอีเมลระหว่างโปรแกรมคัดกรองสแปมเมลที่พัฒนาขึ้นทั้ง 2 เทคนิควิธี

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ต่อผู้พัฒนาระบบ

1) ได้ศึกษาและเข้าใจระบบอีเมลและโพรโทคอลต่างๆที่เกี่ยวข้อง
2) ได้ศึกษาและเข้าใจพฤติกรรมของสแปมเมล และเทคนิคการกำจัดสแปมเมล
3) ได้ศึกษาและเข้าใจกระบวนการทำงานของตัวกรองสแปมแบบเบเซียน เทคนิคการจำแนกประเภทแบบนาอีฟเบเซียน และทฤษฎีระบบฟัซซี
4) ได้ศึกษาและเข้าใจการวิเคราะห์ ออกแบบและพัฒนาโปรแกรมการคัดกรองสแปมเมล

1.6 ขั้นตอนการศึกษา

โครงงานฉบับนี้ได้แบ่งเนื้อหาออกเป็น 5 บทด้วยกันคือ

บทที่ 1 กล่าวถึงความเป็นมาของการศึกษา ความมุ่งหมายและวัตถุประสงค์ ทฤษฎีที่ใช้ในการศึกษา รวมถึงขอบเขตและขั้นตอนการศึกษา

บทที่ 2 กล่าวถึงทฤษฎีพื้นฐานที่เกี่ยวข้องในการพัฒนาโปรแกรมการคัดกรองสแปมเมล

บทที่ 3 กล่าวถึงการขั้นตอนวิธีการดำเนินการต่างๆ

บทที่ 4 กล่าวถึงผลการดำเนินงานในการพัฒนาและทดสอบโปรแกรมคัดกรองสแปมเมล

บทที่ 5 กล่าวถึงบทสรุปของผลการทดลองและข้อเสนอแนะ

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีพื้นฐานต่างๆที่เกี่ยวข้องในการพัฒนาโปรแกรมการคัดกรองสแปมเมล ซึ่งประกอบด้วย ระบบไปรษณีย์อิเล็กทรอนิกส์ กล่าวถึงโครงสร้างและโพรโทคอลที่เกี่ยวข้อง สแปมเมลและเทคนิคการกำจัดสแปม กระบวนการจำแนกแบบนาอีฟเบเซียน และการคัดกรองสแปมแบบเบเซียน รวมทั้งทฤษฎีฟัซซีเซต ฟัซซีลอจิกและระบบฟัซซี ซึ่งเนื้อหาทั้งหมดนี้จำเป็นสำหรับการพัฒนาและประเมินประสิทธิภาพของโปรแกรมการคัดกรองสแปมเมล โดยแบ่งออกเป็นหัวข้อต่างๆ ดังต่อไปนี้

1) ไปรษณีย์อิเล็กทรอนิกส์
2) สแปมและเทคนิคการกำจัดสแปมเมล
3) ทฤษฎีของเบย์
4) การจำแนกประเภทแบบนาอีฟเบเซียน
5) การคัดกรองสแปมแบบเบเซียน
6) ฟัซซีเซต
7) ฟัซซีลอจิก
8) ระบบฟัซซี

2.1 ไปรษณีย์อิเล็กทรอนิกส์

ไปรษณีย์อิเล็กทรอนิกส์หรืออีเมล (Email) เป็นโปรแกรมประยุกต์เพื่อการสื่อสารแลกเปลี่ยนข้อความผ่านอินเทอร์เน็ตหรือภายในเครือข่ายคอมพิวเตอร์ ในการสื่อสารนั้นข้อความอีเมลจะต้องประกอบด้วยเนื้อหา ที่อยู่ของผู้ส่ง และที่อยู่ของผู้รับ ซึ่งต้องมีหนึ่งผู้รับเป็นอย่างน้อย แต่เดิมอีเมลเป็นการแลกเปลี่ยนสื่อสารเฉพาะข้อความ ภายหลังได้พัฒนาให้สามารถแนบแฟ้มสื่อประสมอื่นๆได้โดยใช้โพรโทคอล Multipurpose Internet Mail Extensions (MIME)

2.1.1 รูปแบบข้อความอีเมล

ภายใต้มาตรฐานการสื่อสารแลกเปลี่ยนข้อความผ่านอินเทอเน็ต RFC 822 มีข้อกำหนดให้ข้อความอีเมลมีส่วนประกอบ 2 ส่วนหลักคือ ส่วนหัวอีเมล (Header) และส่วนข้อความ (Body)

2.1.1.1 ส่วนหัวอีเมล

ประกอบด้วยข้อมูลเกี่ยวกับเส้นทางการรับส่ง อีเมลผู้รับ อีเมลผู้ส่ง วันเวลา หัวเรื่อง และรูปแบบการเข้ารหัส (Encode) ซึ่งข้อมูลบางส่วนในส่วนเฮดเดอร์จะปรากฏให้ผู้รับเห็นเมื่อเปิดอ่านด้วยโปรแกรมเปิดอีเมล ข้อมูลดังกล่าวได้แก่อีเมลผู้รับ อีเมลผู้ส่ง วันเวลาและหัวเรื่องจดหมาย ฯลฯ ดังตัวอย่างแสดงในรูปที่ 2.1 และในแต่ละข้อความส่วนหัวอีเมลประกอบด้วยเขตข้อความที่สำคัญได้แก่

ตารางที่ 2.1 ตารางเขตข้อความส่วนหัวอีเมล

เขตข้อความ	ข้อมูล
From:	ที่อยู่อีเมลผู้ส่ง และอาจจะประกอบด้วย ชื่อและนามสกุลเมลแอดเดรส เช่น test@example.com
To:	ที่อยู่อีเมลผู้รับ และอาจจะประกอบด้วย ชื่อและนามสกุล สามารถมีได้มากกว่า 1 คน แยกกันด้วย เครื่องหมาย ","
Subject:	สรุปเนื้อความของอีเมลเพื่อให้ผู้รับสามารถเข้าใจเนื้อหาของข้อความคร่าวๆ
Date:	วันและเวลาจากเครื่องผู้ส่ง
Message-ID:	หมายเลขข้อความ ถูกกำหนดอัตโนมัติเพื่อป้องกันการส่งซ้ำและเพื่ออ้างอิงการตอบกลับข้อความนั้น

```
From: "Kelly J. Weadock" <kelly@litware.com>
To: <anton@proseware.com>
Cc: <tim@cpandl.com>
Subject: Review of staff assignments
Date: Wed, 12 Dec 2007 13:38:31 -0800
Message-ID: <MAILbbnewS5IqCRL00000013@mail.litware.com>
```

รูปที่ 2.1 ตัวอย่างส่วนหัวอีเมล

2.1.1.2 ส่วนเนื้อหาอีเมล

ประกอบด้วยข้อความหรือข้อมูลเนื้อหาจดหมาย (Messages) ที่ผู้ส่งต้องการส่ง เนื้อหาอีเมลถูกกำหนดให้ใช้มาตรฐานการแทนรหัสตัวอักขระแบบ 7-bit

ASCII คือใช้เลขฐานสอง 7 หลักในการแทนตัวอักขระหนึ่งตัว เช่น ตัวอักษร A แทนด้วยรหัส 1000001

หากมีแนบเอกสารหรือสื่อประสมมาอื่นๆกับอีเมล เอกสารนั้นต้องผ่านการเข้ารหัสด้วยโพรโทคอล MIME เมื่อผ่านกระบวนการเข้ารหัสแล้ว ข้อมูลทั้งหมดจะอยู่ในรูปแบบข้อความ (Text) และถูกส่งมาในส่วนเนื้อหาอีเมลนี้เช่นกัน เมื่อรับข้อความอีเมลด้วยโปรแกรมเปิดอ่านอีเมล โปรแกรมนี้จะทำหน้าที่ถอดรหัสให้ส่วนข้อมูลเอกสารแนบที่ถูกเข้ารหัสดังกล่าว ให้กลับไปเป็นรูปแบบที่ผู้ส่งได้แนบมากับอีเมลตามเดิม

เนื่องจากการส่งข้อมูลแบบ ASCII นั้นไม่สามารถส่งข้อมูลประเภทสื่อประสม (Multimedia) เช่น รูปภาพ แฟ้มเสียง จึงมีการกำหนดรูปแบบข้อมูลมาตรฐานเพิ่มเติม นั่นคือ MIME เป็นมาตรฐานที่กำหนดรูปแบบเพิ่มเติม เพื่อรองรับอีเมลที่บรรจุข้อความและเอกสารแนบ (Attachment) หากมีการแนบเอกสารมากับอีเมลนั้น ต้องทำการแปลงให้เป็นข้อความก่อน แล้วประกาศประเภทของเอกสารและรูปแบบการเข้ารหัสไว้ในส่วนหัวอีเมลด้วย ดังแสดงในรูปที่ 2.2 ข้อมูล MIME ที่ต้องระบุในส่วนหัวอีเมลมีดังนี้

- Content-Type: ระบุชนิดเนื้อหา
- Content-Transfer-Encoding: ระบุชนิดการเข้ารหัส

```
From: "Kelly J. Weadock" <kelly@litware.com>
To: <anton@proseware.com>
Cc: <tim@cpandl.com>
Subject: Review of staff assignments
Date: Wed, 12 Dec 2007 13:38:31 -0800
MIME-Version: 1.0
Content-Type: image/jpeg
Content-Transfer-Encoding: base64
```

รูปที่ 2.2 ตัวอย่างการระบุข้อมูล MIME ในส่วนหัวอีเมล

2.1.2 ระบบอีเมล (Email System)

ระบบอีเมลประกอบด้วยองค์ประกอบหลัก 3 ส่วนคือ

2.1.2.1 เครื่องให้บริการอีเมล (Mail Server)

โปรแกรมบริการอีเมลที่ติดตั้งบนฝั่งเครื่องให้บริการทำหน้าที่ให้บริการรับส่งอีเมลระหว่าง จัดเก็บ บริหารจัดการตู้จดหมาย (Mailbox) ของผู้ใช้แต่ละคน และจัดลำดับในการส่งข้อความอีเมล ตัวอย่างโปรแกรมบริการอีเมล เช่น Microsoft Exchange และ Lotus Notes

2.1.2.2 โปรแกรมสำหรับรับ เขียนและอ่านอีเมล (Mail User Agent: MUA)

โปรแกรมที่ติดตั้งบนฝั่งผู้ใช้ เนื่องจากข้อความอีเมลทั้งหมดของผู้ใช้ถูกเก็บไว้ในตู้จดหมายบนเครื่องให้บริการอีเมล ดังนั้นจึงต้องมีโปรแกรมทำหน้าที่เชื่อมต่อผู้ใช้กับระบบอีเมล เพื่อรับ เขียนหรืออ่านอีเมลเหล่านั้น MUA ติดต่อกับ MTA ผ่านโพรโทคอล Post Office Protocol (POP) และ Internet Message Access Protocol (IMAP) ตัวอย่างโปรแกรมเชื่อมต่อระบบอีเมล เช่น Microsoft Outlook และ Mozilla Thunderbird

2.1.2.3 โพรโทคอลสำหรับถ่ายโอนข้อความอีเมล (Simple Mail Transfer Protocol: SMTP)

โพรโทคอลในชั้นแอพพลิเคชัน ใช้ในการโอนถ่ายอีเมลระหว่างเครื่องให้บริการอีเมล และจากผู้ส่งไปยังตู้จดหมายของผู้รับบนเครื่องให้บริการอีเมล

2.2 สแปม

สแปมคือการส่งข้อความที่ผู้รับไม่พึงประสงค์จะรับ ผู้รับไม่ได้ร้องขอ ผ่านทางระบบอิเล็กทรอนิกส์ โดยส่วนมากจะทำให้เกิดความไม่พึงพอใจต่อผู้รับข้อความ ข้อความสแปมนั้นถูกส่งไปได้หลายรูปแบบหลายช่องทางการสื่อสาร รูปแบบหนึ่งที่เป็นที่รู้จักกันคือ สแปมอีเมล (Email Spam) เป็นการส่งข้อความสแปมผ่านทางระบบอีเมล เรียกข้อความอีเมลที่ผู้รับไม่พึงประสงค์รับเหล่านั้นว่าเป็น สแปมเมล (Spam Mail) หรือ อีเมลขยะ (Junk Mail) จุดประสงค์ของการส่งสแปมเมลก็เพื่อโฆษณา เชิญชวนให้ซื้อสินค้าและบริการ หรือแนะนำเว็บทางการค้าต่างๆ โดยผู้รับไม่ทราบว่าผู้ส่งนั้นเป็นใคร

2.2.1 ลักษณะทั่วไปของสแปมเมล

- มักเป็นอีเมลโฆษณาเว็บไซต์ หรือการให้บริการอย่างใดอย่างหนึ่ง ซึ่งการบริการหลายอย่างนั้น บ่อยครั้งไม่สามารถโฆษณาผ่านช่องทางปกติได้โดยสะดวก

อย่างเช่นโฆษณาเกี่ยวกับการเสริมความงามทั้งหญิงและชาย โฆษณาขายซอฟต์แวร์ หนังสือ เป็นต้น

- ชื่อผู้ส่งมักไม่สามารถค้นหาที่มาได้ ด้วยเหตุที่การส่งอีเมลขยะเสี่ยงต่อข้อกฎหมายและเสี่ยงต่อเว็บไซต์ของผู้ส่งที่จะต้องถูกปิดลงจากการละเมิดการใช้งาน
- มีจุดเชื่อมต่อที่แจ้งให้ผู้รับสามารถถอนชื่อออก เมื่อผู้รับเข้าเว็บไซต์นั้นเพื่อทำการถอดถอนบริการ มักถูกใช้เป็นช่องทางการจรวจสอบว่าอีเมลผู้รับนั้นใช้งานอยู่จริง
- หัวข้ออีเมลมักจะเป็นข้อความที่น่าสนใจ หรือหลอกลวงให้ผู้รับเปิดอ่านให้ได้

2.2.2 เทคนิคการกำจัดสแปมเมล

ในปัจจุบันได้มีการพัฒนาเทคนิคการกำจัดสแปมเมลขึ้นมาหลากหลายวิธี เพื่อให้สามารถตรวจจับสแปมเมลได้อย่างมีประสิทธิภาพมากขึ้น ซึ่งเทคนิคที่ได้รับความนิยมสามารถสรุปและรวบรวมได้ดังนี้

2.2.2.1 แบล็คลิสต์ (Blacklist)

มีการรวมตัวกันขององค์กรที่ช่วยรวบรวมรายชื่อเครื่องให้บริการอีเมล ที่มีพฤติกรรมการส่งสแปมเมล หรือเปิดช่องทางให้ผู้ส่งสแปมเข้ามาใช้เพื่อส่งสแปมเมลได้ แล้วตั้งเป็นแบล็คลิสต์สากล เพื่อแจกจ่ายข้อมูลให้เครื่องให้บริการอีเมลที่ร้องขอรายชื่อทั่วโลก แต่วิธีนี้ยังไม่ค่อยได้ผลนักเพราะรายชื่อที่มีการแบล็คลิสต์ไว้มีเพียงน้อยนิดเมื่อเทียบกับเครื่องให้บริการอีเมล ที่มีการส่งสแปมที่เกิดขึ้นใหม่ทุกวัน อีกทั้งองค์กรที่ทำแบล็คลิสต์บางรายใช้วิธีไม่เหมาะสมจนทำให้เครื่องให้บริการอีเมลที่ดี ติดอยู่ในแบล็คลิสต์ด้วยเช่นกัน

2.2.2.2 สแปมแพทเทิร์น (Spam Pattern)

วิธีสแปมแพทเทิร์น วิธีเดียวกับการตรวจจับไวรัสคอมพิวเตอร์ เป็นการนำสแปมเมลที่เคยพบ มาคำนวณหารูปแบบเฉพาะตัว หรือที่เรียกว่าซิกเนเจอร์ (Signature) แล้วรวบรวมเป็นฐานข้อมูลเพื่อใช้ในการตรวจสอบกับอีเมลฉบับอื่นๆ ที่ผ่านเข้ามาในระบบ หากมีการตรวจพบว่าอีเมลฉบับใดตรงกับซิกเนเจอร์ที่รวบรวมไว้ ก็แสดงว่าอีเมลฉบับนั้นเป็นสแปมเมล

สแปมเมลนั้นสามารถสร้างได้ง่ายและรวดเร็วกว่าไวรัสมาก แม้แต่บุคคลทั่วไปที่ไม่มีความรู้ทางด้านเทคนิคมากนัก ก็สามารถส่งสแปมเมลจำนวนมากได้

รูปแบบของสแปมเมลจึงมีความหลากหลาย และยากต่อการทำซิกเนเจอร์ให้ทันต่อสแปมที่เกิดขึ้น และทำให้ฐานข้อมูลของซิกเนเจอร์มีขนาดใหญ่ ถ้าหากอีเมลขาเข้าทุกฉบับที่ได้รับจะต้องผ่านการเปรียบเทียบกับซิกเนเจอร์จำนวนมากครั้ง ย่อมส่งผลให้ประสิทธิภาพของระบบอีเมลลดลง

2.2.2.3 การคำนวณแบบเบเซียน

วิธีนี้เป็นวิธีที่ได้รับการยอมรับทั่วโลกว่าดีที่สุดในปัจจุบัน ใช้หลักการการวิเคราะห์คำศัพท์ที่ปรากฏในอีเมลตัวอย่างจำนวนหนึ่ง ซึ่งอีเมลตัวอย่างจะแบ่งเป็น 2 กลุ่มคือ กลุ่มอีเมลปกติและกลุ่มอีเมลที่เป็นสแปม จะใช้เป็นข้อมูลสำหรับฝึกสอนระบบให้เรียนรู้และจำแนกอีเมลนั้นๆ ว่าอยู่ในกลุ่มของสแปมหรือแฮม หากคำศัพท์ใดมีปรากฏอยู่ในสแปมเมลจำนวนมาก อีเมลที่มีคำๆ นั้นอยู่ก็มีความน่าจะเป็นที่จะเป็นสแปมสูงนั่นเอง

ระบบเบเซียนมีความแม่นยำสูงมากถึง 99.9 เปอร์เซ็นต์ แต่ข้อเสียของวิธีนี้ก็มีเช่นกันคือ การจะได้มาซึ่งความน่าจะเป็นที่เหมาะสม จะต้องรวบรวมอีเมลที่เป็น สแปมและแฮมจำนวนมากเพื่อฝึกสอนระบบ และให้การตัดสินใจของระบบเหมาะสมกับสภาพแวดล้อมของผู้ใช้งาน เช่น ในประเทศไทย การใช้ระบบที่ถูกฝึกสอนมาจากภาษาต่างประเทศ จะไม่สามารถแยกแยะอีเมลที่เป็นภาษาไทยได้

2.2.2.4 การกรองแบบฐานกฎ (Rule-based Filtering)

วิธีนี้เป็นการกำหนดคะแนนความเป็นสแปมให้กับ คำ วลี และคุณสมบัติต่างๆ ของอีเมล แล้วตั้งเป็นกฎขึ้นมาตรวจสอบกับอีเมลทุกฉบับ เมื่อมีอีเมลผ่านเข้ามาในระบบก็จะถูกตรวจสอบว่าตรงกับกฎข้อใดบ้าง ระบบจะรวบรวมคะแนนที่ใช้เป็นเกณฑ์ในการตัดสินหากค่ามากเกินเกณฑ์กำหนด ก็จะถูกตัดสินว่าเป็นสแปม

วิธีนี้เป็นวิธีการที่นิยมใช้มากที่สุดในปัจจุบัน เนื่องจากเข้าใจง่าย ติดตั้งง่ายและมีความแม่นยำสูงพอสมควร (ประมาณ 90 เปอร์เซ็นต์) แต่วิธีนี้มีข้อเสียคืออีเมลที่ไม่ใช่สแปมมักถูกตัดสินเป็นสแปมบ่อยๆ เพราะคะแนนของกฎที่ตั้งไว้ไม่ได้เหมาะสมกับอีเมลทุกประเภท ข้อเสียอีกประการหนึ่งคือ กฎที่กำหนดไว้มีค่าคะแนนคงที่ตายตัว และผู้ส่งสแปมก็สามารถหาซอฟต์แวร์เหล่านั้นมาใช้งานได้

เช่นกัน เมื่อ สแปมเมอร์ต้องการส่งสแปม ก็ทำการทดสอบสแปมเมลของตนเสียก่อน และปรับแก้เนื้อหาให้สามารถผ่านระบบการกรองแบบกฎได้

2.2.2.5 การกรองแบบชาเลนจ์เรสปอนส์ (Challenge-Response Filtering)

วิธีนี้เป็นการเพิ่มขั้นตอนในการยืนยันตัวเองให้กับการส่งอีเมล โดยเมื่อมีการส่งอีเมลไปยังผู้รับที่มีการป้องกันด้วยระบบการกรองแบบชาเลนจ์เรสปอนส์ ระบบจะตรวจสอบว่าผู้ส่งเป็นผู้ที่อยู่ในไวท์ลิสต์ (White list) หรือไม่ ถ้าใช่ก็จะส่งอีเมลไปถึงผู้รับตามปกติ แต่ถ้าไม่ใช่ ระบบจะทำการระงับการส่งอีเมลฉบับนั้นไว้ชั่วคราว แล้วส่งอีเมลตอบกลับไปยังผู้ส่ง ให้เข้ามายืนยันตัวเอง โดยการเข้ามากรอกแบบฟอร์มที่ระบบเตรียมไว้ ระบบจึงส่งอีเมลฉบับนั้นไปยังผู้รับ

วิธีนี้ก่อให้เกิดความลำบากต่อผู้ส่งสแปมที่ส่งอีเมลไปให้ผู้รับนับพันหมื่นฉบับ เนื่องจากไม่สามารถเสียเวลากับการกรอกแบบฟอร์มเพื่อส่งอีเมลตอบกลับไปยังผู้รับรายใดรายหนึ่งได้ และเมื่อไม่มีการกรอกแบบฟอร์มภายในระยะเวลาที่กำหนด ระบบก็จะลบอีเมลฉบับนั้นทิ้งไป วิธีนี้จึงสามารถป้องกันสแปมเมลได้ในอัตราสูงถึงเกือบ 100 เปอร์เซ็นต์ แต่เป็นการเพิ่มขั้นตอนในการส่งอีเมลที่ไม่เป็นสากล และผู้ใช้งานอินเทอร์เน็ตทั่วไปจำนวนมากไม่รู้จักระบบนี้ จึงทำให้ข่าวสารสำคัญๆ ที่ส่งถึงผู้รับมีความเสี่ยงที่จะสูญหายหรือได้รับอีเมลฉบับนั้นช้าเกินไปก็ได้ ในบางครั้งวิธีชาเลนจ์เรสปอนส์ จึงถูกนำไปใช้ร่วมกับวิธีอื่นๆ เช่น วิธีการกรองแบบเบเซียน โดยระบบจะทำการชาเลนจ์เรสปอนส์ เฉพาะอีเมลที่ถูกตัดสินว่าเป็นสแปมเมลจากวิธี เบเซียนเท่านั้น

2.2.2.6 เอฟเอฟบี (FFB: Filters that Fight Back)

สแปมเมลจำนวนมากมีเป้าหมายเพื่อการโฆษณา จะมีการสอดแทรกลิงค์ของเว็บไซต์ที่ขายสินค้าและบริการนั้น หากผู้ได้รับสแปมเมลทุกคนเข้าไปเยี่ยมชมเว็บไซต์เหล่านั้นพร้อมกันทุกคน จะทำให้เครื่องบริการของเว็บไซต์เหล่านั้นต้องรับภาระหนักจนอาจต้องหยุดให้บริการ

แนวความคิดนี้มีผู้นำไปคิดและพัฒนาต่อจนกลายเป็นระบบเอฟเอฟบี หากระบบกรองสแปมเมลมีความสามารถในการโจมตีเว็บไซต์เหล่านั้นอัตโนมัติ โดยผู้รับอีเมลไม่ได้รับผลกระทบใดๆ ก็จะเป็นการทำให้เครื่องบริการของสแปมเมลรับภาระหนักเกินไปจนต้องหยุดให้บริการ ดังนั้นผู้ที่ส่งสแปมเมล จะต้องเพิ่ม

ประสิทธิภาพของเครื่องที่ใช้งานและระบบเครือข่ายเพื่อรองรับการโจมตีของระบบเอฟเอฟบี ซึ่งนั่นหมายถึงค่าใช้จ่ายที่เพิ่มขึ้น ข้อเสียของวิธีนี้คือการโต้ตอบกลับเช่นนี้เป็นการเป็นการโจมตีเครือข่ายอย่างหนึ่ง ซึ่งผู้ใช้บางคนเห็นว่าเป็นสิ่งที่ไม่ควรทำ แม้ว่าเป้าหมายจะเป็นผู้ส่งสแปมก็ตาม

2.2.2.7 แอดเดรสการ์ด (AddressGuard™)

วิธีนี้เป็นวิธีการป้องกันสแปมเมลที่คิดค้นโดย Yahoo.com และใช้ได้กับผู้สมัครสมาชิกอีเมลของ Yahoo เท่านั้น โดยมีหลักการคือ สร้างชื่ออีเมลแอดเดรสปลอมขึ้นมา โดยชื่ออีเมลแอดเดรสปลอมนี้จะไม่บ่งบอกถึงชื่อ หรืออีเมลแอดเดรสจริงของผู้ใช้ ชื่ออีเมลแอดเดรสปลอมนี้สามารถสร้างได้ถึง 500 ชื่อ ถ้าผู้ใช้จำเป็นต้องให้ชื่ออีเมลแอดเดรสบนอินเทอร์เน็ต ผู้ใช้สามารถให้ชื่ออีเมลแอดเดรสปลอมเหล่านี้ได้ เมื่อมีอีเมลส่งถึงผู้ใช้ที่แอดเดรสปลอม ระบบจำทำการส่งต่ออีเมลนั้นไปยังอีเมลแอดเดรสจริงของผู้ใช้ แต่ถ้าเริ่มมีการส่งสแปมเมลมายังชื่ออีเมลแอดเดรสปลอมนี้เมื่อใด ระบบจะยกเลิกอีเมลปลอมนั้นอัตโนมัติ โดยอีเมลแอดเดรสจริงของผู้ใช้งานจะยังคงใช้งานได้ตามปกติ

วิธีนี้ถือเป็นการป้องกันสแปมเมลโดยตัดช่องทางการส่งสแปมเมลซ้ำๆ จำนวนมากได้ แต่ไม่ได้ช่วยคัดกรองสแปมเมลในตู้จดหมายของผู้ใช้ จัดเป็นการป้องกันที่ต้นเหตุมากกว่าปลายเหตุ ซึ่งอาจต้องใช้วิธีอื่นๆ ดังที่กล่าวมาแล้วร่วมกันเพื่อให้การป้องกันสแปมเมลมีประสิทธิภาพมากขึ้น

2.3 ทฤษฎีของเบย์

เป็นการหาค่าความน่าจะเป็นของเหตุการณ์แบบมีเงื่อนไข กำหนดให้ h_i เป็นสมมุติฐาน $P(h_i)$ เป็นความน่าจะเป็นก่อนที่สมมุติฐาน h_i เกิดขึ้นเมื่อไม่อยู่ภายใต้เหตุการณ์ใดๆ หรือความน่าจะเป็นก่อน (Prior Probability) $P(h_i|D)$ เป็นความน่าจะเป็นของสมมุติฐาน h_i จะถูกต้องภายใต้เหตุการณ์ D หรือความน่าจะเป็นภายหลัง (Posterior Probability) $P(D|h_i)$ เป็นความน่าจะเป็นที่ได้จากเหตุการณ์ D โดยกำหนดสมมุติฐาน h_i ถูกต้องและ i เป็นจำนวนกรณีที่เป็นไปได้ของสมมุติฐาน ดังสมการที่ 2.1

$$P(h_i/D) = \frac{P(D/h_i)P(h_i)}{\sum_{j=1}^{m} P(D/h_j)P(h_j)} \quad (2.1)$$

จากกฎของทฤษฎีความน่าจะเป็นทั้งหมด (Theorem of total probability) ถ้ากำหนดให้ *เหตุการณ์* $h_1, h_2, ..., h_n$ ไม่เกิดร่วมกัน และ $\sum_{i=1}^{n} P(h_i) = 1$ แล้วค่าน่าจะเป็น $P(D)$ จะได้ดังสมการ

$$P(D) = \sum_{i=1}^{n} P(D/h_i)P(h_i) \quad (2.2)$$

แทนในสมการข้างต้นจึงได้ว่า

$$P(h_i/D) = \frac{P(D/h_i)P(h_i)}{P(D)} \quad (2.3)$$

ซึ่งสมมุติฐานที่ดีที่สุดคือ สมมุติฐานภายหลังมากที่สุด (Maximum A Posterior hypothesis: MAP)

$$h_{map} = \underset{h \in H}{argmax}\, P(h/D) \quad (2.4)$$

$$h_{map} = \underset{h \in H}{argmax} \frac{P(h/D)P(h)}{P(D)} \quad (2.5)$$

$$h_{map} = \underset{h \in H}{argmax}\, P(h/D)(h) \quad (2.6)$$

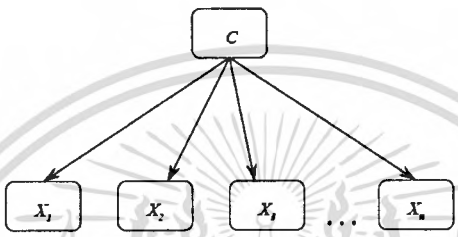
เนื่องจาก $P(D)$ เป็นค่าคงที่และไม่เกิดร่วมกับเหตุการณ์ใดๆ จึงสามารถละในการคำนวณได้

2.4 การจำแนกประเภทแบบนาอีฟเบเซียน

ในการศึกษาครั้งนี้เลือกศึกษาการคัดกรองสแปมแบบเบเซียน ซึ่งเป็นการประยุกต์ใช้การจำแนกประเภทแบบนาอีฟเบเซียน

2.4.1 หลักการการจำแนกประเภทแบบนาอีฟเบเซียน

การจำแนกประเภทแบบนาอีฟเบเซียนเป็นการประยุกต์ทฤษฎีของเบย์ (Bayes' theorem) โดยการคำนวณหาความน่าจะเป็นของแต่ละกลุ่มข้อมูล ซึ่งในที่นี้เรียกว่าคลาส (Class) เมื่อกำหนดคุณสมบัติหรือแอททริบิว (Attribute) และค่าของแอททริบิวแต่ละตัวให้ การทำนายจะคำนวณหาค่าความน่าจะเป็นของทุกๆคลาสเพื่อเปรียบเทียบกัน แล้วเลือกค่าความน่าจะเป็นที่สูงที่สุดของคลาสใดๆ มาเป็นผลของการทำนายเพียงค่าเดียว โดยที่ถือว่าแอททริบิวแต่ละตัวมีความเป็นอิสระต่อกัน (Conditional independence) สามารถแสดงได้ดังภาพ เมื่อ C เป็นคลาสและ $x_1, x_2, \ldots x_n$ คือ แอททริบิวแต่ละตัว



รูปที่ 2.3 คลาสและแอททริบิว

จากสมมุติฐานที่แต่ละเหตุการณ์ไม่มีความเกี่ยวข้องกัน (Naive assumption) สมมุติต้องการจำแนก $d = (x_1, x_2, \ldots, x_n)$ ลงกลุ่ม c_j โดยที่ c_j เป็นกลุ่มหนึ่งในเซตของกลุ่ม C กลุ่มของ d คือกลุ่ม C_{MAP} ซึ่งคำนวณได้จากทฤษฎีของเบย์ ดังนี้

$$C_{map} = \underset{cj \in C}{argmax} \frac{P(x_1, x_2, \ldots, x_n / c_j) P(c_j)}{P(x_1, x_2, \ldots, x_n)} \tag{2.7}$$

$$C_{map} = \underset{cj \in C}{argmax}\, P(x_1, x_2, \ldots, x_n / c_j) P(c_j) \tag{2.8}$$

เมื่อสมมติให้ค่าของแอทริบิวต่างๆเป็นอิสระจากกันจะได้ว่า

$$P(x_1, x_2, \ldots, x_n / c_j) = \prod_{i=1}^{n} P(x_i / c_j) \tag{2.9}$$

ดังนั้นการคำนวณค่าความน่าจะเป็นของการจำแนกแบบนาอีฟเบเซียนจึงได้ดังสมการ

$$C_{NB} = \underset{cj \in C}{argmax}\, P(c_j) \prod_{i=1}^{n} P(x_i / c_j) \tag{2.10}$$

2.4.2 ความน่าจะเป็นมีค่าเป็นศูนย์

ในกรณีที่เหตุการณ์นั้นไม่เคยเกิดขึ้นมาก่อน เมื่อทำการคำนวณหาความน่าจะเป็นจะมีค่าเป็นศูนย์ ซึ่งส่งผลให้การคำนวณเกิดการผิดพลาดได้ ดังนั้นจึงต้องทำการปรับค่า วิธีการปรับค่าต่างๆที่นิยมใช้ดังนี้

- Direct Estimate

$$P(A_i/C) = \frac{N_{ic}}{N_c} \tag{2.11}$$

- Laplace Estimate

$$P(A_i/C) = \frac{N_{ic} + 1}{N_c + c} \tag{2.12}$$

- m-estimate

$$P(A_i/C) = \frac{N_{ic} + mp}{N_c + m} \tag{2.13}$$

โดยที่ c คือ จำนวนคลาส

p คือ ความน่าจะเป็นก่อน

m คือ ค่าคงที่ค่าหนึ่ง เช่น จำนวนแอททริบิว

2.4.3 การสร้างแบบจำลองจากชุดการเรียนรู้และการทำนาย

2.4.3.1 เตรียมข้อมูลสำหรับเป็นชุดการเรียนรู้ (Training Dataset) ในการฝึกสอนระบบ

ตารางที่ 2.2 ข้อมูลชุดการเรียนรู้ จำนวน 14 เรคคอร์ด (records)

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Overcast	Cool	Normal	True	No
Sunny	Cool	Normal	True	Yes
Sunny	Mild	High	False	No

ตารางที่ 2.2 (ต่อ)

Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

จากตารางประกอบด้วย แอททริบิว *Play* เป็นคลาส ซึ่งมี 2 ค่า คือ

- *Play = Yes* จำนวน 9 เรคคอร์ด
- *Play = No* จำนวน 5 เรคคอร์ด

แอททริบิวสำหรับเป็นข้อมูลในการสร้างแบบจำลองจำนวน 4 แอททริบิว คือ

- แอททริบิว *Outlook* มีค่าที่เป็นไปได้คือ *Sunny, Overcast, Rainy*
- แอททริบิว *Temperature* มีค่าที่เป็นไปได้คือ *Hot, Mild, Cool*
- แอททริบิว *Humidity* มีค่าที่เป็นไปได้คือ *High, Normal*
- แอททริบิว *Windy* มีค่าที่เป็นไปได้คือ *True, False*

2.4.3.2 สร้างแบบจำลอง

สร้างแบบจำลองจากชุดการเรียนรู้ที่กำหนด โดยคำนวณค่าความน่าจะเป็นสำหรับแต่ละแอททริบิวในแต่ละคลาส โดยเริ่มจากพิจารณาแอททริบิว *Outlook* ทำการคำนวณความน่าจะเป็นโดยใช้สมการได้ผลดังนี้

$$P(Outlook = Sunny \mid Class = Yes) = \frac{2}{9}$$

$$P(Outlook = Sunny \mid Class = No) = \frac{3}{5}$$

ทำจนครบทุกค่าของแอททริบิว Outlook ถึง Windy เมื่อทำการคำนวณเสร็จแล้วจะได้แบบจำลองที่ประกอบด้วยค่าต่างๆ ดังนี้

ตารางที่ 2.3 แบบจำลองที่ได้จากชุดการเรียนรู้ที่กำหนด

Outlook	**Play=Yes**	**Play=No**
Sunny	2/9	3/5
Overcast	4/9	0/5
Rainy	3/9	2/5
Temperature	**Play=Yes**	**Play=No**
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5
Humidity	**Play=Yes**	**Play=No**
High	3/9	4/5
Normal	6/9	1/5
Windy	**Play=Yes**	**Play=No**
False	6/9	2/5
True	3/9	3/5
Play (Class)	**Yes**	**No**
	9/14	5/14

2.4.3.3 ทำนายข้อมูล

นำแบบจำลองที่สร้างได้จากชุดการเรียนรู้ไปทดสอบการทำนายกับข้อมูลชุดอื่น เพื่อหาค่าความแม่นยำ ความถูกต้องของแบบจำลองว่าสามารถยอมรับได้หรือไม่ การทำนายข้อมูลเป็นการนำแบบจำลองที่สร้างขึ้นมาใช้ในการทำนายข้อมูลที่ไม่ทราบคลาสมาก่อน โดยมีการกำหนดเงื่อนไขของเหตุการณ์ E ที่ต้องการทำนาย ดังตาราง โดยทำการคำนวณหาความน่าจะเป็นของเหตุการณ์ที่จะทำให้เกิดคลาสแต่ละตัว

ตารางที่ 2.4 ชุดข้อมูลสำหรับทดสอบการทำนายประเภท

Outlook	**Temperature**	**Humidity**	**Windy**	**Play**
Sunny	Cool	High	True	??

ทำการคำนวณ 2 กรณี คือ กรณีที่คลาส *(Play = Yes)* และ คลาส *(Play = No)*

- กรณีที่คลาส *(Play) = Yes*

$$P(Outlook = Sunny \mid Class = Yes) = \frac{2}{9}$$

$$P(Temperature = Cool \mid Class = Yes) = \frac{3}{9}$$

$$P(Humidity = High \mid Class = Yes) = \frac{3}{9}$$

$$P(Windy = True \mid Class = Yes) = \frac{3}{9}$$

$$P(Class = Yes) = \frac{9}{14}$$

จะได้

$$P(Class_{Yes} \mid E) = P(Sunny \mid Yes)P(Cool \mid Yes)P(High \mid Yes)P(True \mid Yes)P(Yes)$$

$$= \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}$$

$$= 0.0053$$

- กรณีที่คลาส (Play) = No

$$P(Outlook = Sunny \mid Class = No) = \frac{3}{5}$$

$$P(Temperature = Cool \mid Class = No) = \frac{1}{5}$$

$$P(Humidity = High \mid Class = No) = \frac{4}{5}$$

$$P(Windy = True \mid Class = No) = \frac{3}{5}$$

$$P(Class = No) = \frac{5}{14}$$

จะได้

$$P(Class_{No} \mid E) = P(Sunny \mid No)P(Cool \mid No)P(High \mid No)P(True \mid No)P(No)$$

$$= \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14}$$

$$= 0.0206$$

ดังนั้น

$$P(Class_{Yes} / E) = \frac{0.0053}{0.0053 + 0.0206}$$
$$= 20.5\%$$

$$P(Class_{No} / E) = \frac{0.0206}{0.0053 + 0.0206}$$
$$= 79.5\%$$

สามารถทำนายได้ว่า ในกรณีที่เกิดเหตุการณ์ ตามตารางที่ 2.4 ข้างต้นจะมี *Class = No*

2.5 การคัดกรองสแปมแบบเบเซียน

เป็นการประยุกต์ใช้การจำแนกประเภทแบบนาอีฟเบเซียน ในการจำแนกประเภทข้อความ (Text Classification) เพื่อจำแนกประเภทอีเมลว่าเป็นสแปมหรือแฮม ในการเรียนรู้นั้นนาอีฟเบเซียนจะกรองข้อความในอีเมลจำนวนหนึ่งจากชุดข้อมูลการเรียนรู้ เก็บคำศัพท์และสร้างแบบจำลองจากความน่าจะเป็น แล้วใช้ผลการเรียนรู้นั้นในการระบุประเภทอีเมล

2.5.1 การประยุกต์ใช้ทฤษฎีของเบย์ในการคัดกรองสแปมเมล

จากทฤษฎีของเบย์

$$P(A / B) = \frac{P(B / A)P(A)}{P(B)} \tag{2.14}$$

สามารถนำไปใช้ในการคำนวณในการเรียนรู้ เพื่อคัดแยกสแปมเมลโดยเขียนแทนการคำนวณได้ดังนี้

$$P(Spam / Word) = \frac{P(Word / Spam)P(Spam)}{P(Word)} \tag{2.15}$$

แปลว่าความน่าจะเป็นแบบสุ่มที่คำที่เป็น Spam จะอยู่ในเซตของเอกสาร เท่ากับ ความน่าจะเป็นของคำที่ปรากฏอยู่ในเซตของคำที่เป็น Spam คูณกับ ความน่าจะเป็น Spam หารด้วย คววมน่าจะเป็นของคำนั้น และเมื่อความน่าจะเป็นของคำ เท่ากับ ความน่าจะเป็นของคำที่ปรากฏอยู่ในเซตของ Spam คูณกับ ความน่าจะเป็น Spam บวกด้วย ความน่าจะเป็นของคำที่ปรากฏอยู่ในเซตของคำที่ไม่ใช้ Spam คูณกับ ความน่าจะเป็นของคำที่ไม่ใช่ Spam และเมื่อ *P(Word)* เขียนรูปสมการได้ดังนี้

$$P(Word) = P(Word \mid Spam)P(Spam) + P(Word \mid Ham)P(Ham) \quad (2.16)$$

สามารถเขียนสมการในการคำนวณค่าความน่าจะเป็นของสแปมใหม่ได้ดังนี้

$$P(Spam \mid Word) = \frac{P(Word \mid Spam)P(Spam)}{P(Word \mid Spam)P(Spam) + P(Word \mid Ham)P(Ham)} \quad (2.17)$$

สามารถเขียนเป็นสมการในรูปได้ว่า

$$P(Spam_i \mid Words) = \frac{P(Word \mid Spam_i)P(Spam_i)}{\sum_j P(Words \mid Spam_j)P(Spam_j)} \quad (2.18)$$

การนำไปใช้ ยกตัวอย่างกรณีที่รู้คำ 2 คำคือ word1 และ word2 ซึ่งค่าความน่าจะเป็นสแปมของแต่ละคำคือ a, b ตามลำดับ จะได้สมการดังนี้

$$P(Spam \mid word1, word2) = \frac{ab}{ab + (1-a)(1-b)} \quad (2.19)$$

กรณีที่รู้คำ 3 คำคือ word1, word2 และ word3 ซึ่งค่าความน่าจะเป็นสแปมของแต่ละคำคือ a, b, c ตามลำดับ จะได้สมการดังนี้

$$P(Spam \mid word1, word2, word3) = \frac{abc}{abc + (1-a)(1-b)(1-c)} \quad (2.20)$$

ค่าความน่าจะเป็นที่ได้จะมีค่าตั้งแต่ 0 - 1

- เมื่อ 0 แสดงว่ามีความน่าจะเป็น Spam 0%
- เมื่อ 1 แสดงว่ามีความน่าจะเป็น Spam 100%

จากกฎของเบย์ข้างต้น บอกถึงแนวทางในการนำกฎนี้ไปใช้ได้ว่า เมื่อรับอีเมลเข้ามาระบบต้องแยกข้อความในอีเมลนั้นออกเป็นคำๆ ก่อน เพื่อที่จะนำคำเหล่านั้นไปหาความน่าจะเป็นในเงื่อนไขที่ว่า ถ้ามีคำเหล่านั้นปรากฏในเอกสาร มีความน่าจะเป็นมากน้อยเพียงใดที่อีเมลฉบับนั้นจะเป็นสแปม แต่เนื่องจากคำเหล่านี้ ก่อนจะนำไปใช้กับกฎดังที่กล่าวมา จำเป็นต้องมีค่าตัวเลขที่แสดงถึงน้ำหนักของคำแต่ละคำนั้นว่า แต่ละคำมีน้ำหนัก

โอนเอียงไปทาง Spam หรือ Ham โดยค่าดังกล่าวเรียกว่าค่า Weight ในระบบจะเรียกค่า Weight ว่า Spamicity ซึ่งหาได้จากสมการดังนี้

$$Spamicity = \frac{SpamProbability_k}{SpamProbability_k + HamProbability_k} \tag{2.18}$$

โดย $SpamProbability = \frac{tf_k}{N_s}$ เมื่อ

Spam Probability คือ ความน่าจะเป็นสแปม คำในลำดับที่ k (Term ที่ k)

tf_k คือ ความถี่ของคำที่ k ปรากฏในเอกสารชุดการเรียนรู้ทั้งหมด

N_s คือ จำนวนเอกสารที่อยู่ในเซตที่เป็นสแปม (Spam) ในชุดการเรียนรู้ทั้งหมด

โดย $HamProbability = \frac{tf_k}{N_h}$ เมื่อ

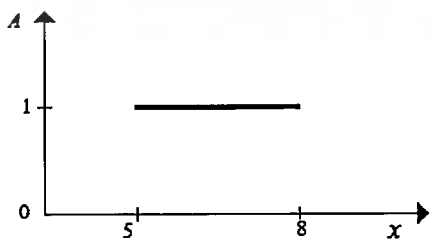
Ham Probability คือ ความน่าจะเป็นสแปม คำในลำดับที่ k (Term ที่ k)

tf_k คือ ความถี่ของคำที่ k ปรากฏในเอกสารชุดการเรียนรู้ทั้งหมด

N_h คือ จำนวนเอกสารที่อยู่ในเซตที่เป็นแฮม (Ham) ในชุดการเรียนรู้ทั้งหมด

2.6 ฟัซซีเซ็ต

ในทางคณิตศาสตร์ เซตแบบฉบับ (Classical Set) หรือเซตปกติ (Crisp Set) เป็นเซตที่มีค่าความเป็นสมาชิกเป็น 0 หรือ 1 {0, 1} เท่านั้น เซตในทฤษฎีเซตแบบฉบับจะมีขอบเขตที่ตัดขาดจากกันแบบทันทีทันใด เซตแบบฉบับมีการกำหนดค่าความเป็นสมาชิกตามแนวคิดเลขฐานสอง โดยที่ตัวแปรหนึ่งๆ จะมีค่าความเป็นสมาชิกเพียงสองค่า คือ 0 ไม่เป็นสมาชิก และ 1 เป็นสมาชิก ตัวอย่างเช่น เซต A เป็นเซตของค่าจำนวนจริงตั้งแต่ 5 ถึง 8 เขียนแทนเป็นช่วงได้ $A = \{5, 8\}$



รูปที่ 2.4 เซตปกติของ $A = \{5, 8\}$

จากเซต $A = \{5, 8\}$ ข้างต้น

- จำนวนจริง 6 อยู่ในเซต A จะมีค่าความเป็นสมาชิกของเซต A เป็น 1

- จำนวนจริง 4 ไม่ได้อยู่ในเซต A จะมีค่าความเป็นสมาชิกของเซต A เป็น 0

ฟัซซีเซตเป็นส่วนขยายของเซตปกติ โดยจะมีลักษณะตามฟังก์ชันความเป็นสมาชิก (Membership Function) ซึ่งจะใช้ฟังก์ชันความเป็นสมาชิกในการแปลงค่าสมาชิกของเอกภพสัมพัทธ์ให้อยู่ในช่วงตั้งแต่ 0 ถึง 1 [0, 1] โดยค่า 0 หมายถึง ไม่ได้มีส่วนร่วมอยู่ในสมาชิกของเซตเลย และ 1 หมายถึง มีส่วนที่เป็นสมาชิกของเซตอย่างสมบูรณ์

ค่าความเป็นสมาชิก 0 กับ 1 นั้นไม่ละเอียดและไม่ยืดหยุ่นพอสำหรับการนำไปใช้งานบางกรณี เช่น ถ้าเซต S เป็นเซตของผู้เยาว์ อายุ 0 - 20 ปี นั้นหมายความว่าถ้าอายุ 20 ปี 1 วันก็ไม่ถือว่าเป็นผู้เยาว์แล้ว ซึ่งดูแล้วเป็นการตีความที่ไม่ดีพอสำหรับกรณีนี้ ดังนั้นจึงมีการนิยามฟัซซีเซตผู้เยาว์ *(young)* ขึ้นเพื่อตอบคำถามที่ว่า x มีค่าความเป็นผู้เยาว์เท่าใด โดยต้องคิดค่าความเป็นสมาชิกที่อยู่ในฟัซซีเซต young ของแต่ละคน ฟังก์ชันความเป็นสมาชิกที่ง่ายที่สุดคือพิจารณาอายุคนดังนี้

$$young(x) = \begin{cases} 1 & if \quad age(x) \le 20 \\ (30 - age(x))/10 & if \quad 20 < age(x) \le 30 \\ 0 & if \quad age(x) > 30 \end{cases} \qquad (2.19)$$

จากสมการที่ 2.19 จะได้กราฟดังรูป

รูปที่ 2.5 กราฟฟัซซีเซต young

2.6.1 ฟังก์ชันความเป็นสมาชิก (Membership Function)

ถ้าให้ U เป็นเอกภพสัมพัทธ์ (Universal Set) ที่เป็นเซตปกติ โดยที่ $\mu_A(x)$ เป็นค่าความเป็นสมาชิกของ x ในเซตปกติ A ในเอกภพสัมพัทธ์ U จะได้ฟังก์ชันความเป็นสมาชิกดังนี้

$$\mu_{young}(x) = \begin{cases} 1 & if \quad and \quad only \quad if \quad x \in A \\ 0 & if \quad and \quad only \quad if \quad x \notin A \end{cases} \qquad (2.20)$$

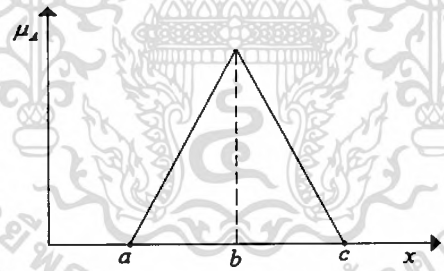
ฟัซซีเซต A ในเอกภพสัมพัทธ์ U และให้ $\mu_A(x)$ เป็นค่าความเป็นสมาชิกของ x ใน ฟัซซีเซต สามารถนิยามได้ดังนี้

$$A = \{(x, \mu_A(x)) \mid x \in U\} \tag{2.21}$$

ฟังก์ชันความเป็นสมาชิกใช้เพื่อแสดงขอบเขตของค่าความเป็นสมาชิกในแต่ละฟัซซีเซต ฟังก์ชันความเป็นสมาชิกที่ใช่ในฟัซซีลอจิกที่นิยม มีดังนี้

2.6.1.1 ฟังก์ชันความเป็นสมาชิกแบบสามเหลี่ยม (Triangular Membership Function)

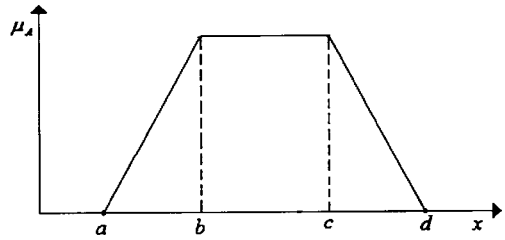
$$\mu_A(x) = \begin{cases} 0 & ;x < a \\ \dfrac{x-a}{b-a} & ;a \le x \le b \\ \dfrac{c-x}{c-b} & ;b \le x \le c \\ 0 & ;x > c \end{cases} \tag{2.22}$$



รูปที่ 2.6 ฟังก์ชันความเป็นสมาชิกแบบสามเหลี่ยม

2.6.1.2 ฟังก์ชันความเป็นสมาชิกแบบสี่เหลี่ยมคางหมู (Trapezoidal Membership Function)

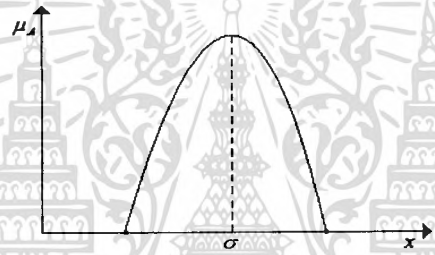
$$\mu_A(x) = \begin{cases} 0 & ;x < a \\ \dfrac{x-a}{b-a} & ;a \le x < b \\ 1 & ;b \le x < c \\ \dfrac{d-x}{d-c} & ;c \le x < d \\ 0 & ;x \ge d \end{cases} \tag{2.23}$$



รูปที่ 2.7 ฟังก์ชันความเป็นสมาชิกแบบสี่เหลี่ยมคางหมู

2.6.1.3 ฟังก์ชันความเป็นสมาชิกแบบเกาส์ (Gaussian Membership Function)

$$\mu_A(x) = \left\{ \exp\left(-\frac{(x-m)^2}{\sigma^2} \right) \right. \tag{2.24}$$



รูปที่ 2.8 ฟังก์ชันความเป็นสมาชิกแบบเกาส์

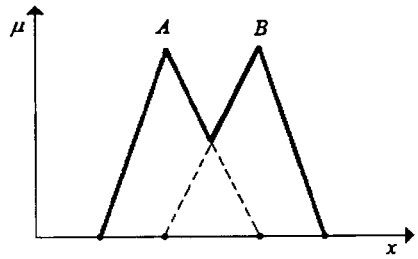
2.6.2 การดำเนินการในฟัซซีเซต (Operation of Fuzzy Set)

ให้ A และ B เป็นฟัซซีเซตใน U ที่มีความเป็นสมาชิกกับ μ_A และ μ_B ตามลำดับ การดำเนินการพื้นฐานของฟัซซีเซตมีดังนี้

2.6.2.1 การดำเนินการ Union หรือ OR

ฟังก์ชันความเป็นสมาชิกของฟัซซีเซต A และ B ที่ Union กัน ผลลัพธ์ที่ได้เป็นค่าที่มากที่สุดของทั้งสองฟังก์ชันความเป็นสมาชิกดังนี้

$$\mu_{A\cup B}(x) = \max[\mu_A(x), \mu_B(x)] \tag{2.25}$$

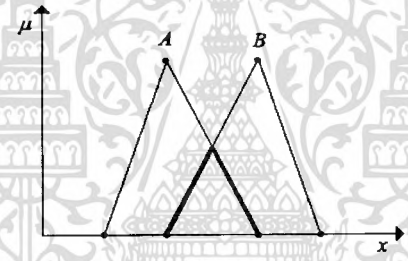


รูปที่ 2.9 การดำเนินการ Union หรือ OR

2.6.2.2 การดำเนินการ Intersection หรือ AND

ฟังก์ชันความเป็นสมาชิกของฟัซซีเซต A และ B ที่ Intersection กัน ผลลัพธ์ที่ได้เป็นค่าที่น้อยที่สุดของทั้งสองฟังก์ชันความเป็นสมาชิกดังนี้

$$\mu_{A\cap B}(x) = \min[\mu_A(x), \mu_B(x)] \tag{2.26}$$

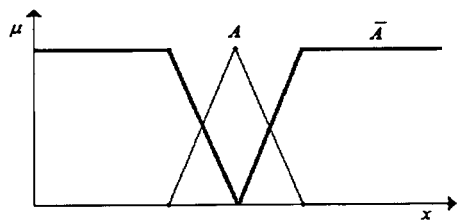


รูปที่ 2.10 การดำเนินการ Intersection หรือ AND

2.6.2.3 การดำเนินการ Complement หรือ NOT

ฟังก์ชันความเป็นสมาชิกของฟัซซีเซต A ที่มีฟังก์ชันความเป็นสมาชิก μ_A ผลลัพธ์ที่ได้เป็นนิเสธของฟังก์ชันความเป็นสมาชิกที่ระบุไว้

$$\mu_{A'}(x) = 1 - \mu_A(x) \tag{2.27}$$



รูปที่ 2.11 การดำเนินการ Complement หรือ NOT

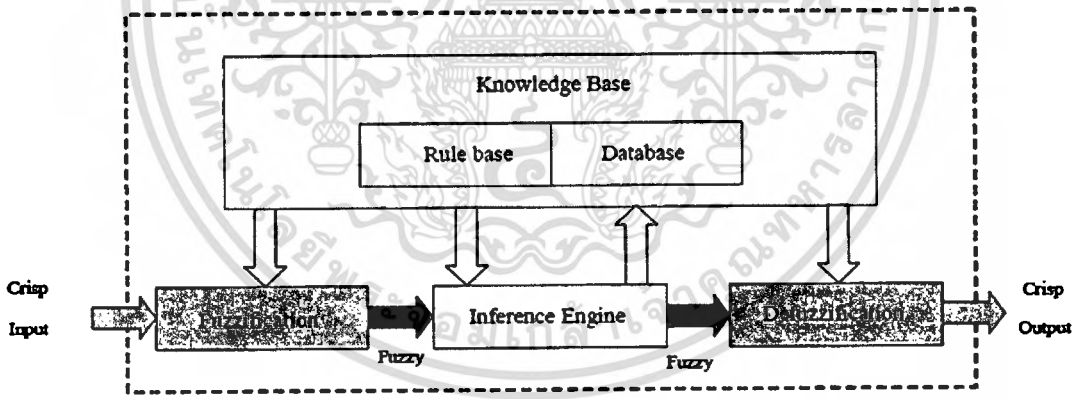
2.7 ฟัซซีลอจิก

ฟัซซีลอจิกหรือตรรกศาสตร์คลุมเครือ พัฒนาจากทฤษฎีฟัซซีเซตโดยเป็นการใช้เหตุผลแบบประมาณด้วยค่าความเป็นจริง ซึ่งแตกต่างจากการใช้เหตุผลแบบเด็ดขาดในลักษณะ ถูกหรือผิด ใช่หรือไม่ใช่ ของตรรกศาสตร์แบบเดิม (Classical Logic) ฟัซซีลอจิกนั้นถือเป็นการประยุกต์ใช้งานฟัซซีเซตเพื่อจำลองการตัดสินใจของผู้เชี่ยวชาญต่อปัญหาที่ซับซ้อน

ฟัซซีลอจิกเป็นเครื่องมือช่วยในการตัดสินใจภายใต้ความไม่แน่นอนของข้อมูล โดยยอมให้มีความยืดหยุ่นได้ ใช้หลักเหตุผลที่คล้ายการเลียนแบบวิธีความคิดที่ซับซ้อนของมนุษย์ ฟัซซีลอจิกมีลักษณะพิเศษกว่าตรรกะแบบจริงเท็จ (Boolean Logic) เป็นแนวคิดที่มีการต่อขยายในส่วนของความจริง (Partial True) โดยค่าความจริงจะอยู่ในช่วงระหว่างจริง (Completely True) หรือ 1 กับเท็จ (Completely False) หรือ 0 เป็นลักษณะ อาจจะ บางที ค่อนข้าง ซึ่งเป็นลักษณะที่ฟัซซีลอจิกยอมให้สมาชิกบางส่วนอยู่ในเซตได้ ส่วนตรรกศาสตร์เดิมจะมีค่าเป็นจริงกับเท็จเท่านั้น

2.8 ระบบฟัซซี

ระบบฟัซซีเป็นระบบที่ใช้หลักการของฟัซซีเซตและฟัซซีลอจิก พัฒนาขึ้นเป็นระบบควบคุมที่ใช้ในแก้ไขปัญหาต่างๆ



รูปที่ 2.12 โครงสร้างพื้นฐานของระบบฟัซซี

2.8.1 โครงสร้างพื้นฐานของระบบฟัซซี

โครงสร้างพื้นฐานของการประมวลผลแบบฟัซซี ประกอบด้วยส่วนที่สำคัญ 4 ส่วนดังนี้

2.8.1.1 ส่วนที่แปลงอินพุตทั่วไปให้เป็นการอินพุตแบบตัวแปรฟัซซี (Fuzzification) หรือในรูปแบบเซตฟัซซีหรือเรียกว่าเป็นตัวแปรเชิงภาษา (Linguistic Variable)

2.8.1.2 ฐานความรู้ (Knowledge base) เป็นส่วนที่จัดเก็บรวบรวมข้อมูลในการควบคุมประกอบ 2 ส่วนคือ ฐานกฎ (Rule base) และฐานข้อมูล

- ฐานกฎ (Rule base) ส่วนของการกำหนดวิธีการควบคุม ซึ่งได้จากผู้เชี่ยวชาญในรูปแบบของชุดข้อมูลแบบกฎเชิง โดยเงื่อนไขพื้นฐานนี้มักออกแบบให้อยู่ในรูปแบบดังนี้

  กฎข้อที่ i: x_1 เป็น $A_1^{(i)}$ และ...และ x_n เป็น $A_n^{(i)}$ แล้ว y เป็น $B^{(i)}$

- ฐานข้อมูลเป็นการจัดเตรียมข้อมูลส่วนที่จำเป็น ได้แก่ ชนิดของฟังก์ชันความเป็นสมาชิก และพารามิเตอร์ต่าง ๆ

2.8.1.3 เครื่องอนุมานหรือการตีความ (Inference Engine) เป็นส่วนที่ทำหน้าที่ตรวจสอบข้อเท็จจริงและกฎ เพื่อใช้ในการตีความหาเหตุผล เหมือนกลไกสำหรับควบคุมการใช้ความรู้ในการแก้ไขปัญหา รวมทั้งการกำหนดวิธีการของการตีความเพื่อหาคำตอบ

2.8.1.4 ส่วนที่แปลงการเอาต์พุตให้อยู่ในช่วงที่เหมาะสม (Defuzzification) เป็นการทำการแปลงข้อมูลที่อยู่ในรูปแบบฟัซซีให้เป็นค่าที่สรุปผลหรือค่าการควบคุมระบบ

2.8.2 การประมวลผลของระบบฟัซซี

ขั้นตอนการประมวลผลของระบบฟัซซี โดยทั่วไปมีรูปแบบการทำงานเป็น 5 ส่วน ได้แก่

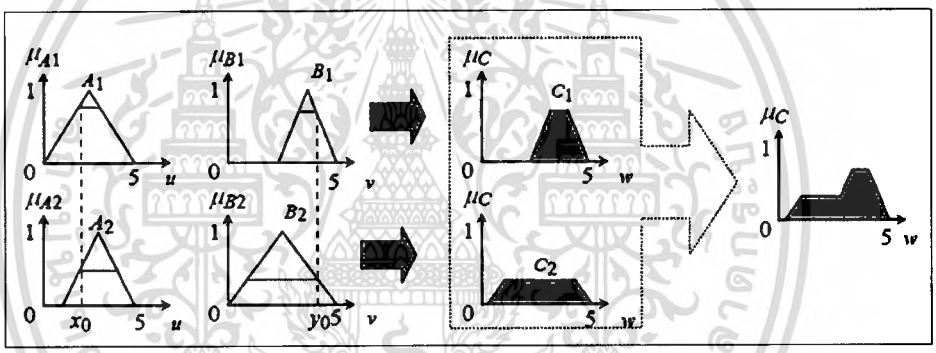
2.8.2.1 การแปลงค่าอินพุตทั่วไปเป็นค่าฟัซซี (Fuzzification) เป็นการคำนวณค่าฟัซซีผ่านฟังก์ชันความเป็นสมาชิกจากส่วนข้อสมมุติฐานของกฎฟัซซีเพื่อหาค่าดีกรีระหว่าง 0 ถึง 1

2.8.2.2 การรวมค่าฟัซซีจากส่วนข้อสมมุติฐาน (Combining) เป็นการรวมค่าฟัซซีจากฟังก์ชันความเป็นสมาชิกในส่วนข้อสมมุติฐานของกฎข้อเดียวกันเข้าด้วยกันโดยใช้ตัวดำเนินการ Fuzzy AND (min) หรือ Fuzzy OR (max) ทำเป็นค่าดีกรีความแข็งแรงเป็นค่าระหว่าง 0 ถึง 1 ส่งออกไปจากส่วนข้อสมมุติฐาน

2.8.2.3 การตีความ (Implication) เป็นส่วนที่ทำหน้าที่ตรวจสอบข้อเท็จจริงและกฎ เพื่อใช้ในการตีความหาเหตุผล เหมือนกลไกสำหรับควบคุมการใช้ความรู้ในการแก้ไขปัญหา รวมทั้งการกำหนดวิธีการของการตีความเพื่อหาคำตอบ วิธีการที่นิยมใช้ในการตีความ ได้แก่ Max-Min method และ Max-Dot method

- Max-Min method ใช้กฎการหาค่าต่ำสุด (Minimum Operation Rule) เป็นฟังก์ชันความเป็นสมาชิกสำหรับหาค่า โดยค่าความเป็นสมาชิกของผลลัพธ์ c จากกฎ 2 ข้อ ตามสมการที่ 2.28

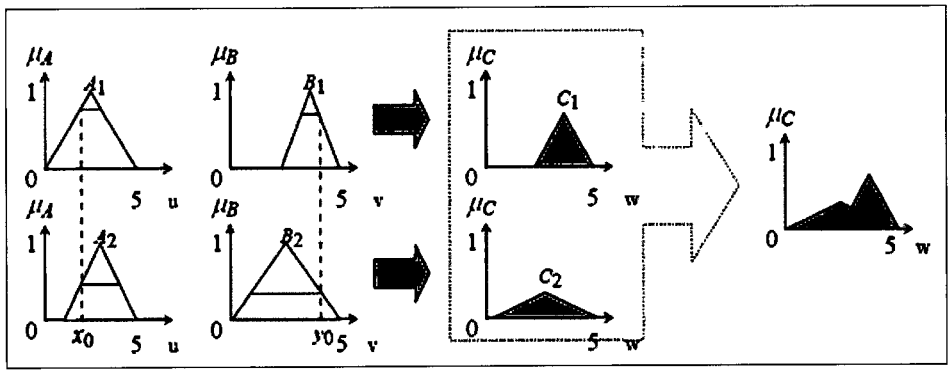
$$\mu_C(w) = \alpha_1 \wedge \mu_{C1}(w)) \vee \alpha_2 \wedge \mu_{C2}(w)) \tag{2.28}$$



รูปที่ 2.13 การประมวลผล Max-Min Inference กับค่า Crisp Input x_0 และ y_0

- Max-Dot method จะใช้กฎของผลคูณ (Product Operation Rule) เป็นฟังก์ชันความเป็นสมาชิกสำหรับหาค่า โดยค่าความเป็นสมาชิกของผลลัพธ์ C จากกฎทั้งหมด 2 ข้อ ตามสมการที่ 2.29

$$\mu_C(w) = \alpha_1 . \mu_{C1}(w)) \vee \alpha_2 . \mu_{C2}(w)) \tag{2.29}$$



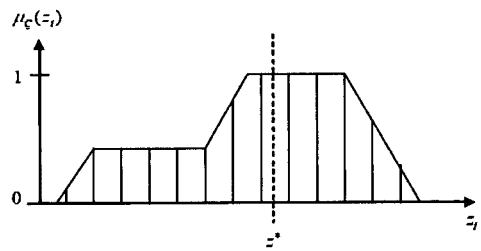
รูปที่ 2.14 การประมวลผล Max-Dot Inference กับค่า Crisp Input x_0 และ y_0

2.8.2.2 การรวมค่า ฟัซซีเอาต์พุตจากกฎทุกข้อ (Aggregation) เป็นการรวมค่าจากข้อตามหรือข้อสรุปของกฎทุกข้อเพื่อเป็นฟัซซีเซตของระบบทั้งหมดด้วยวิธี Fuzzy OR

2.8.2.3 การทำค่าฟัซซีเป็นค่าปกติ (Defuzzification) เป็นการทำค่าฟัซซีเอาต์พุตที่รวมจากกฎทุกข้อเป็นค่าปกติที่ใช้ในงานจริงในบางงาน เช่น ในระบบควบคุม วิธีการทำค่าฟัซซีให้เป็นค่าทั่วไปมีหลายวิธีด้วยกันดังนี้

- วิธีเฉลี่ยถ่วงน้ำหนัก (Weighted average method) หรือวิธีค่าพื้นที่กลาง (Centroid of area: COA) วิธีการหาค่าเฉลี่ยถ่วงน้ำหนักของพื้นที่ใต้กราฟฟัซซีซึ่งเป็นผลที่ได้จากการตีความ ค่าที่ได้จะประมาณเทียบเคียงค่าจุดศูนย์ถ่วงโดยรวม (Central of gravity: COG) จะหาได้จากการประมาณค่าจากสมการที่ 2.30

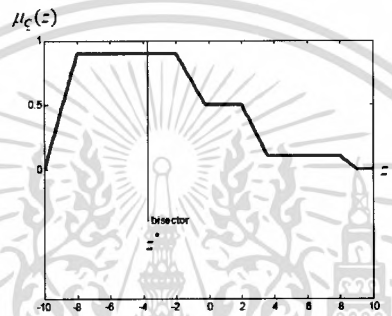
$$z^* = \frac{\sum_{i=1}^{n} \mu_C(Z_i).Z_i}{\sum_{i=1}^{n} \mu_C(Z_i)} \qquad (2.30)$$



รูปที่ 2.15 การแปลงค่าฟัซซีเป็นค่าจริงทั่วไปด้วยวิธีถ่วงน้ำหนัก

- วิธีแบ่งครึ่งของพื้นที่ (Bisector of area: BOA) ค่าเอาต์พุตที่ได้จะจากระบบฟัซซีเป็นค่าครึ่งหนึ่งของพื้นที่ใต้กราฟฟัซซี หาได้จากสมการที่ 2.31

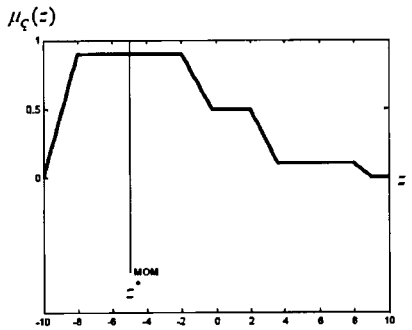
$$z^* = \left\{ Z_j \sum_{i=1}^{j} \mu_C(Z_i) \geq \frac{\sum_{i=1}^{n} \mu_C(Z_i}{2} \right. \tag{2.31}$$



รูปที่ 2.16 การแปลงค่าฟัซซีเป็นค่าจริงทั่วไปด้วยวิธีแบ่งครึ่งของพื้นที่

- วิธีค่าเฉลี่ยของค่าสูงสุด (Mean value of maximum: MOM) วิธีค่าเฉลี่ยของค่าสูงสุด เป็นการหาค่าเอาต์พุตที่จะเป็นค่าจริงทั่วไปจากระบบฟัซซี ที่คำนวณจากค่าเฉลี่ยของค่าในโดเมนจริงที่มีค่าดีกรีความเป็นสมาชิกสูงสุด ดังสมการที่ 2.32

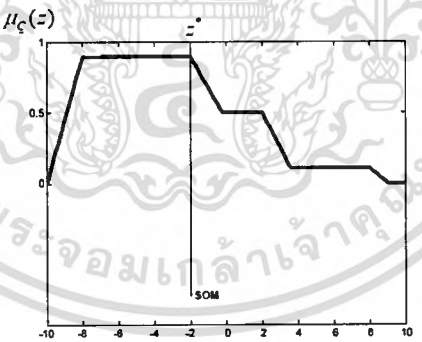
$$z^* = \left\{ \frac{\sum_{j=1}^{m} Z_j}{M} \middle| Z_j \in \max(\mu_C(Z_i)) \right. \tag{2.32}$$



รูปที่ 2.17 การแปลงค่าฟัซซีเป็นค่าจริงทั่วไปด้วยวิธีค่าเฉลี่ยของค่าสูงสุด

- วิธีค่าน้อยสุดของค่าสูงสุด (Smallest absolute value of maximum: SOM) การแปลงค่าฟัซซีเป็นค่าจริงทั่วไปด้วยวิธีค่าน้อยสุดของค่าสูงสุด เป็นการหาค่าเอาต์พุตที่จะเป็นค่าทั่วไปจากระบบฟัซซี ที่คำนวณจากค่าน้อยที่สุดของค่าขนาดในโดเมนจริงที่มีค่าดีกรีความเป็นสมาชิกสูงสุด ดังสมการที่ 2.33

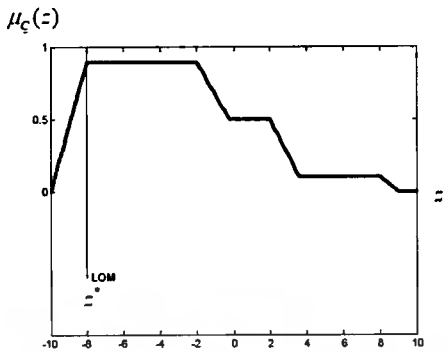
$$z^* = \left\{ Z_j \middle| J = \arg \min_{abs(Z_i)} \left[\max(\mu_C(Z_i)) \right] \right. \tag{2.33}$$



รูปที่ 2.18 การแปลงค่าฟัซซีเป็นค่าจริงทั่วไปด้วยวิธีวิธีค่าน้อยสุดของค่าสูงสุด

- วิธีค่ามากสุดของค่าสูงสุด (Largest absolute value of maximum: LOM) การแปลงค่าฟัซซีเป็นค่าจริงทั่วไปด้วยวิธีค่ามากสุดของค่าสูงสุด เป็นการหาค่าเอาต์พุตที่จะเป็นค่าทั่วไปจากระบบฟัซซี ที่คำนวณจากค่ามากที่สุดของค่าขนาดในโดเมนจริงที่มีค่าดีกรีความเป็นสมาชิกสูงสุด ดังสมการที่ 2.34

$$z^* = \left\{ Z_j \middle| J = \arg\max_{abs(Z_i)} \left[\max(\mu_C(Z_i)) \right] \right. \tag{2.34}$$



รูปที่ 2.19 การแปลงค่าฟัซซีเป็นค่าจริงทั่วไปด้วยวิธีค่ามากสุดของค่าสูงสุด

วิธีการแปลงค่าฟัซซีเป็นค่าจริงทั่วไปที่เลือกใช้คือวิธีเฉลี่ยถ่วงน้ำหนัก (Weighted average method) หรือวิธีค่าพื้นที่กลาง (Centroid of area: COA) โดยหาค่าเฉลี่ยถ่วงน้ำหนักของพื้นที่ใต้กราฟฟัซซีซึ่งเป็นผลที่ได้จากการตีความ

2.9 ข้อกำหนดอื่นๆที่ใช้ร่วมในระบบ

2.9.1 ขอบเขตของเนื้อความอีเมลใช้พิจารณาและการแบ่งคำ

ขอบเขตการพิจารณาเนื้อความจดหมาย เลือกพิจาณาทั้งส่วนหัว (Header) และเนื้อความของอีเมลที่เป็นตัวอักษรเท่านั้น ไม่ได้พิจารณาแฟ้มที่แนบมาด้วยทุกชนิด การพิจารณาแยกคำศัพท์จากอีเมล แยกจากเครื่องหมายต่อไปนี้

(	เครื่องหมายวงเล็บเปิด
)	เครื่องหมายวงเล็บปิด
:	เครื่องหมายทวิภาค
@	เครื่องหมายแอด
<	เครื่องหมายน้อยกว่า
>	เครื่องหมายมากกว่า
\n	ตัวอักษรแบบ white space - ขึ้นบรรทัดใหม่
\r	ตัวอักษรแบบ white space - เครื่องหมาย Return
\t	ตัวอักษรแบบ white space - เครื่องหมาย Tab
“	เครื่องหมายอัญประกาศคู่
‘	เครื่องหมายอัญประกาศเดี่ยว

,	เครื่องหมายจุลภาค
!	เครื่องหมายอัศเจรีย์
?	เครื่องหมายปรัศนีย์
#	เครื่องหมายชาร์ป
$	เครื่องหมายดอลลาร์
*	เครื่องหมายดอกจัน
+	เครื่องหมายบวก
/	เครื่องหมายหาร
{	วงเล็บปีกกาเปิด
}	วงเล็บปีกกาปิด
[	วงเล็บก้ามปูเปิด
]	วงเล็บก้ามปูปิด
=	เครื่องหมายเท่ากับ
;	เครื่องหมายอัฒภาค

กล่าวคือเมื่อโปรแกรมเปิดอ่านอีเมลและพบเครื่องหมาบดังรายการที่กำหนด จะทำการตัดคำโดยใช้เครื่องหมายดังกล่าวเป็นตัวแยก

บทที่ 3
วิธีการดำเนินงาน

วิธีการดำเนินงานในการพัฒนาโปรแกรมคัดกรองสแปมเมล โดยเทคนิคการแบ่งกลุ่มแบบเบเซียนประเภทนาอีฟเบเซียน และเทคนิคการแบ่งกลุ่มแบบฟัซซี-นาอีฟเบเซียน ผู้พัฒนาได้แบ่งขั้นตอนการดำเนินงานออกเป็น 6 ขั้นตอนคือ

1) ขั้นตอนการศึกษาเทคนิคการแบ่งกลุ่มแบบนาอีฟเบเซียนและการจัดกลุ่มแบบฟัซซี
2) ขั้นตอนการออกแบบอัลกอริธึมสำหรับเทคนิคการแบ่งกลุ่มแบบนาอีฟเบเซียน และเทคนิคการแบ่งกลุ่มแบบฟัซซี-นาอีฟเบเซียน
3) ขั้นตอนการวางแผนและออกแบบหน้าจอสำหรับโปรแกรมคัดกรองสแปมเมล
4) ขั้นตอนการพัฒนาโปรแกรมคัดกรองสแปมเมล
5) ขั้นตอนการทดสอบโปรแกรมคัดกรองสแปมเมล
6) ขั้นตอนการเปรียบเทียบประสิทธิภาพการทำงานและความถูกต้องในการจำแนกประเภทอีเมล ระหว่างโปรแกรมคัดกรองสแปมเมลที่พัฒนาขึ้นทั้ง 2 เทคนิควิธี เพื่อทำการสรุปผล

3.1 ขั้นตอนการศึกษาเทคนิคการแบ่งกลุ่มแบบเบเซียนประเภทนาอีฟเบเซียน และการจัดกลุ่มแบบฟัซซี

จากการทำการศึกษาเทคนิคการแบ่งกลุ่มแบบนาอีฟเบเซียน พบว่าเป็นเทคนิคที่ใช้ในการแบ่งกลุ่มชุดข้อมูลประเภทหนึ่งที่มีการนำไปประยุกต์ใช้หลายด้าน มักใช้ในการแบ่งกลุ่มข้อมูลในกรณีที่ทราบประเภทหรือคลาสของข้อมูลล่วงหน้า โดยจะทำการคำนวณหาความน่าจะเป็นของแต่ละคลาส เมื่อกำหนดแอททริบิวและค่าที่เป็นไปได้ของแอททริบิวแต่ละตัวมาให้ การทำนายจะคำนวณหาค่าความน่าจะเป็นของทุกๆคลาสมาเปรียบเทียบกัน แล้วเลือกค่าความน่าจะเป็นที่ดีที่สุดของคลาสใดๆ มาเป็นผลของการทำนายเพียงค่าเดียว โดยที่ถือว่าข้อมูลแต่ละตัวมีความเป็นอิสระต่อกัน

ส่วนการจัดกลุ่มแบบฟัซซีเป็นการจัดกลุ่มข้อมูลที่จะพิจารณาจากค่าความเป็นสมาชิก ดังนั้นข้อมูลชุดหนึ่งๆ สามารถเป็นสมาชิกได้มากกว่า 1 คลาส และข้อมูลที่ใช้ในการจัดกลุ่มแบบฟัซซีจะเป็นข้อมูลเชิงปริมาณ

3.2 ขั้นตอนการออกแบบอัลกอริธึม

3.2.1 อัลกอริธึมเทคนิคการแบ่งกลุ่มแบบนาอีฟเบเซียน

เทคนิคการแบ่งกลุ่มแบบแบบนาอีฟเบเซียน ประกอบด้วยขั้นตอนการสร้างแบบจำลองและขั้นตอนการทำนาย ดังแสดงในบทที่ 2 หัวข้อ 2.4 รายละเอียดขั้นตอนการทำงานมีดังนี้

3.2.1.1 การสร้างแบบจำลอง

1) ขั้นตอนการนำเข้าชุดข้อมูลเพื่อการเรียนรู้ เป็นการนำเข้าข้อมูลสู่ระบบทำการแยกคำ เก็บรายการคำศัพท์และจำนวนครั้งที่เจอแต่ละคำในกลุ่มอีเมลที่เป็นสแปมและกลุ่มที่เป็นแฮม
2) ขั้นตอนการเรียนรู้ เป็นการคำนวณโอกาสที่เจอคำศัพท์แต่ละคำในเซตสแปมและเซตแฮม ซึ่งสามารถเขียนเป็นอัลกอริธึมได้ดังนี้

```
For Each Distinct Value in Class
    For Each Attribute
        For Each Distinct Value in Attribute
            Count the Records (Xi)
        Next
        Calculate the Total Records For Each Class (Cj)
        Store the Value (Xi/Cj)) For Each Distinct Value in Attribute
    Next Attribute
Next Class
```

3.2.1.2 การทำนายข้อมูล เป็นการนำข้อมูลที่ต้องการทำนายหาคลาสโดยใช้แบบจำลองที่สร้างขึ้นข้างต้น โดยคำนวณหาค่าความน่าจะเป็นที่จะเกิดเหตุการณ์ตามเงื่อนไขที่นำเข้ามาทำนายของทั้ง 2 คลาส คือคลาสแปมและคลาสแฮม แล้วเลือกคลาสที่มีค่าความน่าจะเป็นสูงสุดให้เป็นคลาสของข้อมูลที่ทำนาย ซึ่งสามารถเขียนเป็นอัลกอริธึมดังนี้

```
For Each Distinct Value in Class
    For Each Attribute
        Calculate Probability Density from Value (Xi/Cj)
        Total Prob = Multiple Probability Attribute
    Next Attribute
```

Prob of C_j = Total Prob * (C_j/Total Class)

Next Class

When Predict use Max (Prob of C_j) to give final result

3.2.2 อัลกอริธึมเทคนิคการแบ่งกลุ่มแบบฟัซซี-นาอีฟเบเซียน

ในการนำเทคนิคการแบ่งกลุ่มแบบฟัซซีเข้ามาประยุกต์ใช้ร่วมในโปรแกรมคัดกรองสแปมเมล เพื่อปรับปรุงประสิทธิภาพการทำงานของเทคนิคการแบ่งกลุ่มแบบนาอีฟเบเซียนแบบเดิมนั้น ได้นำฟัซซีเข้ามาร่วมใช้ในขั้นตอนการเรียนรู้เพื่อสร้างแบบจำลอง โดยเรียนรู้จากชุดตัวอย่างอีเมลชุดใหม่ ร่วมกับค่าความน่าจะเป็นสแปมของคำศัพท์จากแบบจำลองที่สร้างขึ้นจากเทคนิคการแบ่งกลุ่มแบบนาอีฟเบเซียนเดิม โดยมีแนวคิดคือระบบทราบค่าความน่าจะเป็นสแปมของแต่ละคำศัพท์ เช่น คำว่า Sexy มีค่าความน่าจะเป็นสแปมเท่ากับ 0.89 นั่นคือมีค่าความน่าจะเป็นแฮมเท่ากับ 0.11 เมื่อพิจารณาประกอบกับทฤษฎีฟัซซีเซต สามารถประยุกต์ได้ว่า คำว่า Sexy มีค่าความเป็นสมาชิกเซตสแปมเท่ากับ 0.89 และมีค่าความเป็นสมาชิกเซตแฮมเท่ากับ 0.11 จึงได้นำเทคนิคการแบ่งกลุ่มแบบฟัซซีเข้ามาช่วยปรับค่าความน่าจะเป็นสแปมของคำศัพท์แต่ละคำให้เหมาะสมมากขึ้น

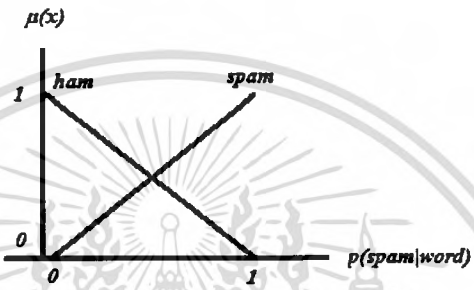
ในกระบวนการทำงาน ระบบจะทำการคำนวณค่าความเป็นสแปมของอีเมลแต่ละฉบับ จากสมการที่ 2.18 แล้วนำมาเปรียบเทียบกับคลาสของอีเมลฉบับนั้นๆ ในแฟ้มดัชนี เพื่อพิจารณาว่าคำใดที่จะต้องทำการปรับค่าความน่าจะเป็นในระบบฟัซซี ซึ่งจะพิจารณาปรับค่าเฉพาะกรณีที่คลาส และค่าความน่าจะเป็นสแปมของแต่ละอีเมลที่คำนวณได้มีความขัดแย้งกัน ดังนี้

- กรณีที่ 1 อีเมลมีคลาสเป็นแฮม แต่ค่าความน่าจะเป็นสแปมที่คำนวณได้มีค่ามากกว่า 0.5 ($p(spam|email) > 0.5$)
- กรณีที่ 2 อีเมลมีคลาสเป็นสแปม แต่ค่าความน่าจะเป็นสแปมที่คำนวณได้มีค่าต่ำกว่า หรือเท่ากับ 0.5 ($p(spam|email) \leq 0.5$)

ในกรณีที่อีเมลมีคลาสเป็นแฮมและมีค่าความน่าจะเป็นสแปมต่ำ หรืออีเมลมีคลาสเป็นสแปมและมีค่าความน่าจะเป็นสแปมสูง จะไม่ทำการปรับค่าใหม่อีกครั้ง เนื่องจากอีเมลฉบับนั้นถูกระบุประเภทได้ถูกต้องแล้ว และแสดงว่าค่าความน่าจะเป็นสแปมของคำศัพท์ภายในอีเมลนั้นมีความเหมาะสมแล้ว

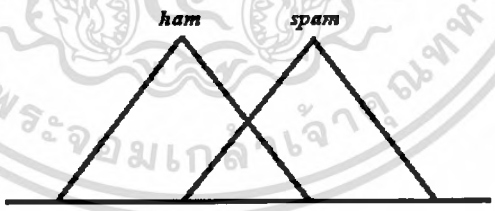
3.2.2.1 การออกแบบระบบฟัซซี

1) ออกแบบฟัซซีเซตและฟังก์ชันการนำเข้า (Input Function) โดยใช้ฟังก์ชันสี่เหลี่ยมคางหมูในการคำนวณค่าความเป็นสมาชิก และข้อมูลที่นำเข้าสู่ระบบฟัซซีคือค่าความเป็นสแปมของคำศัพท์แต่ละคำ หรือ *P(spam|word)* ซึ่งเป็นข้อมูลที่ได้จากแบบจำลองการแบ่งกลุ่มแบบเบเซียนประเภทนาอีฟเบเซียน และกำหนดช่วงข้อมูลนำเข้าให้อยู่ระหว่าง 0 – 1 ดังแสดงในรูปที่ 3.1



รูปที่ 3.1 ฟังก์ชันการนำเข้าของระบบฟัซซี

2) ฟังก์ชันผลลัพธ์ (Output Function) ใช้ฟังก์ชันสามเหลี่ยมในการรวมผลลัพธ์ และตัดสินว่าคำศัพท์นั้นเป็นสมาชิกของเซตสแปมหรือแฮม จากค่าความเป็นสมาชิกที่มากที่สุด



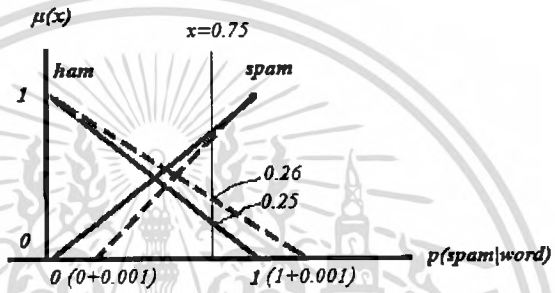
รูปที่ 3.2 ฟังก์ชันผลลัพธ์ของระบบฟัซซี

3.2.2.2 การสร้างแบบจำลอง

1) ขั้นตอนการนำเข้าข้อมูล เป็นการนำเข้าค่าความน่าจะเป็นสแปมของทุกคำศัพท์ในอีเมลตัวอย่างทีละฉบับ เพื่อคำนวณค่าความเป็นสมาชิกของเซตสแปมและแฮมของคำศัพท์แต่ละคำโดยใช้ฟังก์ชันสามเหลี่ยมตามเงื่อนไขที่ได้ระบุไว้

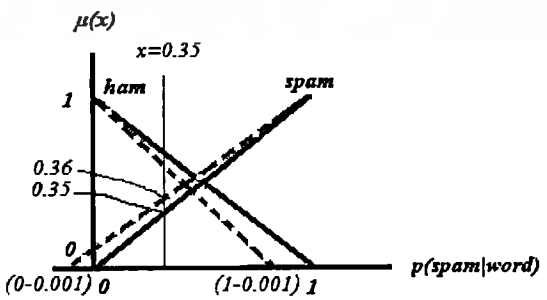
2) ขั้นตอนการเรียนรู้ คำนวณค่าความเป็นสมาชิกของคำศัพท์โดยใช้ฟังก์ชันสามเหลี่ยม และปรับค่าความเป็นสมาชิกกลุ่มให้เหมาะสม โดยหลักการปรับค่ามีหลักการปรับตามกรณีที่แตกต่างกันดังนี้

ก) กรณีที่ 1 อีเมลมีคลาสเป็นแฮม แต่ค่าความน่าจะเป็นสแปมที่คำนวณได้มีค่ามากกว่า 0.5 ปรับช่วงของฟังก์ชันความเป็นสมาชิกเพิ่มขึ้น 0.001 (+0.001) เพื่อเมื่อคำนวณค่าความเป็นสมาชิกใหม่ จะทำให้คำศัพท์เป็นสมาชิกของเซตแฮมสูงขึ้น ดังแสดงในรูปที่ 3.3



รูปที่ 3.3 การปรับช่วงฟังก์ชันความเป็นสมาชิกเพิ่มขึ้น +0.001

ข) กรณีที่ 2 อีเมลมีคลาสเป็นสแปม แต่ค่าความน่าจะเป็นสแปมที่คำนวณได้มีค่าต่ำกว่าหรือเท่ากับ 0.5 ปรับช่วงของฟังก์ชันความเป็นสมาชิกลดลง 0.001 (-0.001) เพื่อเมื่อคำนวณค่าความเป็นสมาชิกใหม่ จะทำให้คำศัพท์เป็นสมาชิกของเซตสแปมสูงขึ้น ดังแสดงในรูปที่ 3.4



รูปที่ 3.4 การปรับช่วงฟังก์ชันความเป็นสมาชิกลดลง -0.001

ในขั้นตอนการสร้างแบบจำลองของเทคนิคการแบ่งกลุ่มแบบฟัซซี-นาอีฟเบเซียน สามารถเขียนเป็นอัลกอริธึม ดังนี้

For Each Instance

Total Prob = Multiple Probability Attribute //From Naive Bayesian

For Each Class

Fuzzy Set = Ham, Spam

Membership Function = Triangular

Middle of Ham Set = b, 0

End of Ham Set = c, 1

Start of Spam Set = d, 0

Middle of Spam Set = e, 1

If Class = Ham and Total Prob > 0.5

End of Ham Set = c + 0.001

Start of Spam Set = d + 0.001

For Each Attribute

For Each Distinct Set

Calculate Membership Value (O_i)

When Set Membership use Max (O_i)

Next Set

Next Attribute

Else If Class = Spam and Total Prob $\leq$ 0.5

End of Ham Set = c - 0.001

Start of Spam Set = d - 0.001

For Each Attribute

For Each Distinct Set

Calculate Membership Value (O_i)

When Set Membership use Max (O_i)

Next Set

Next Attribute

Next Class

Next Instance

3.2.2.3 การทำนายข้อมูล จะทำนายโดยใช้หลักการเดียวกันกับการจำแนกประเภทแบบนาอีฟเบเซียน คือคำนวณหาค่าความน่าจะเป็นที่จะเกิดเหตุการณ์ตามข้อมูลที่นำเข้ามาทำนายของทั้ง 2 คลาส คือคลาสแปมและคลาสแฮม แล้วเลือกคลาสที่มีค่าความน่าจะเป็นสูงสุดให้เป็นคลาสของข้อมูลที่ทำนาย ซึ่งสามารถเขียนเป็นอัลกอริธึม ดังนี้

For Each Distinct Value in Class

For Each Attribute

Calculate Probability Density from Value (X_i/C_j)

Total Prob = Multiple Probability Attribute

Next Attribute

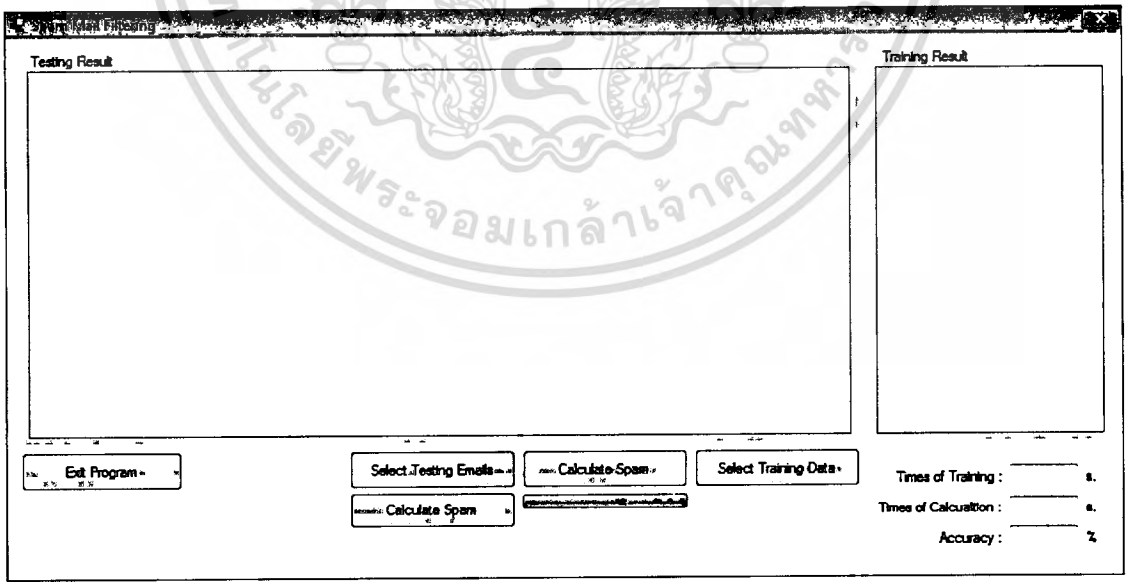
Prob of C_j = Total Prob * (C_j/Total Class)

Next Class

When Predict use Max (Prob of C_j) to give final result

3.3 ขั้นตอนการวางแผนและออกแบบหน้าจอสำหรับโปรแกรมคัดกรองสแปมเมล

จากขั้นตอนของการศึกษาและออกแบบอัลกอริธึม ผู้พัฒนาได้วางแผนและออกแบบหน้าจอสำหรับโปรแกรมคัดกรองสแปมเมลทั้ง 2 เทคนิค ให้มีส่วนประกอบต่างๆ ดังนี้



รูปที่ 3.5 หน้าจอสำหรับโปรแกรมคัดกรองสแปมเมล

3.4 ขั้นตอนการพัฒนาโปรแกรมคัดกรองสแปมเมล

เมื่อทำการวางแผนและออกแบบระบบแล้ว ขั้นต่อมาคือขั้นตอนของการพัฒนาโปรแกรมคัดกรองสแปมเมล โดยใช้โปรแกรมไมโครซอฟท์วิชวลซีชาร์ปดอทเน็ต (Microsoft Visual C#.net) เป็นเครื่องมือสำหรับพัฒนาโปรแกรม ซึ่งสามารถเขียนเป็นลำดับได้ดังนี้

3.4.1 ขั้นตอนของการสร้างหน้าจอ

ในขั้นตอนนี้จะยึดตามการออกแบบที่ได้วางแผนไว้ เป็นการสร้างหน้าจอที่เกี่ยวข้องกับระบบทั้งหมด รวมถึงการสร้างปุ่มควบคุมและเมนูต่างๆ สำหรับแต่ละหน้าจอ

3.4.2 ขั้นตอนการเขียนโปรแกรม

ในขั้นตอนนี้จะทำการเขียนโปรแกรมเพื่อควบคุมการทำงานของระบบ โดยยึดหลักตามอัลกอริธึมที่ได้ออกแบบไว้ ซึ่งจะแยกเป็น 2 โปรแกรม คือ โปรแกรมสำหรับระบบการคัดกรองสแปมเมลโดยเทคนิคการแบ่งกลุ่มแบบเบเซียนประเภทนาอีฟเบเซียน และโปรแกรมสำหรับระบบการคัดกรองสแปมเมลโดยเทคนิคการแบ่งกลุ่มแบบฟัซซี-นาอีฟเบเซียน เครื่องมือที่ใช้ในการพัฒนาระบบดังนี้

3.4.2.1 ฮาร์ดแวร์ที่ใช้ในการพัฒนาระบบมีคุณสมบัติดังนี้

- CPU : Intel Core2 Duo T7300 2.4 GHz
- RAM : 4 GB
- System Type : 32-bit Operating System
- Hard Disk : 250 GB

3.4.2.2 ซอฟท์แวร์ที่ใช้ในการพัฒนาระบบมีดังนี้

- Windows 7 Professional
- Visual C#.net
- Microsoft Visual Studio 2008

3.5 ขั้นตอนของการทดสอบโปรแกรมคัดกรองสแปมเมล

เมื่อพัฒนาโปรแกรมคัดกรองสแปมเมลเสร็จเรียบร้อยแล้ว จึงทำการทดสอบความถูกต้องของระบบโดยใช้ข้อมูลที่จัดเตรียมไว้ และแบ่งการทดสอบเป็นหัวข้อต่างๆ ดังนี้

3.5.1 การทดสอบการทำงานของโปรแกรม

แบ่งเป็นการตรวจสอบในเรื่องต่างๆ ดังนี้

1) ตรวจสอบลำดับการทำงานของโปรแกรม
2) ตรวจสอบการเชื่อมโยงระหว่างหน้าจอและปุ่มควบคุมต่างๆ

3) ตรวจสอบความถูกต้องของข้อความที่แสดงในแต่ละหน้าจอ

4) ตรวจสอบผลลัพธ์ที่ได้จากการคำนวณ

3.5.2 ข้อมูลที่ใช้ในการทดสอบโปรแกรม

ข้อมูลที่ใช้ในการฝึกสอนและทดสอบระบบ เป็นชุดอีเมลตัวอย่างจากหน่วยงานที่ให้บริการข้อมูลมาตรฐานและมาตรฐานการวัดสำหรับทำการวิจัยต่าง ๆ (National Institute of Standards and Technology: NIST) อีเมลแต่ละฉบับมีการระบุประเภทที่ถูกต้องไว้แล้วว่าเป็นสแปมหรือแฮมด้วยแฟ้มดัชนีที่ชื่อ Index.txt ซึ่งแฟ้มดัชนีนี้จะมาพร้อมกับแต่ละชุดอีเมลตัวอย่างที่นำมาศึกษา รายละเอียดชุดอีเมลตัวอย่างมีที่นำมาศึกษามีดังนี้

1) ชุดอีเมลที่ใช้ในการประชุมหัวข้อ "การเรียกค้นข้อความ" ปี 2549 (Text Retrieval Conference of National Institute of Standards and Technology (NIST): TREC 2006 Spam Corpora) ซึ่งประกอบด้วยข้อความอีเมลจำนวน 37,822 ฉบับ โดยอีเมลที่เป็นแฮมจำนวน 12,910 ฉบับ และอีเมลที่เป็นสแปมจำนวน 24,912 ฉบับ

2) ชุดอีเมลที่ใช้ในการประชุมหัวข้อ "การเรียกค้นข้อความ" ปี 2550 (Text Retrieval Conference of National Institute of Standards and Technology (NIST): TREC 2007 Spam Corpora) ซึ่งประกอบด้วยข้อความอีเมลจำนวน 75,419 ฉบับ โดยอีเมลที่เป็นแฮมจำนวน 25,220 ฉบับ และอีเมลที่เป็นสแปมจำนวน 50,199 ฉบับ

3.5.3 การออกแบบการทดลอง

การทดสอบจะสุ่มเลือกจากอีเมลตัวอย่างข้างต้นทั้ง 2 ชุดตัวอย่าง สำหรับตรวจสอบความน่าเชื่อถือของแบบจำลองที่สร้างขึ้น โดยการคำนวณหาค่าความถูกต้องของแบบจำลอง ซึ่งแบ่งการทดสอบได้เป็น 2 ประเภท คือ

1) การทดสอบโดยใช้ข้อมูลชุดเรียนรู้ (Evaluate on Training Set) โดยนำข้อมูลชุดการเรียนรู้มาทำการทำนาย จากนั้นจึงทำการเปรียบเทียบระหว่างคลาสเดิมและคลาสที่เกิดจากการทำนายว่ามีความถูกต้องเพียงใด

2) การทดสอบโดยใช้ข้อมูลชุดอื่น (Evaluate on Testing Set) โดยนำข้อมูลชุดอื่นที่ระบบไม่เคยทำการเรียนรู้มาก่อนมาทำการทำนาย จากนั้นจึงทำการเปรียบเทียบระหว่างคลาสเดิมและคลาสที่เกิดจากการทำนายว่ามีความถูกต้องเพียงใด ชุดข้อมูลที่แบ่งได้มีดังนี้

ตารางที่ 3.1 ชุดข้อมูลสำหรับฝึกสอนและทดสอบโปรแกรม

ชุดข้อมูล (Corpus)	จำนวนอีเมล	สแปม	ความน่าจะเป็นสแปม	แฮม	ความน่าจะเป็นแฮม
ชุดการเรียนรู้ (TREC 2007)	10,000	7078	0.71	2,922	0.29
ชุดทดสอบ1 (จากชุดเรียนรู้)	1,000	668	0.67	332	0.33
ชุดทดสอบ2 (TREC 2007)	2,000	1442	0.72	558	0.28
ชุดทดสอบ3 (TREC 2006)	3,000	2212	0.74	788	0.26

3.6 ขั้นตอนของการเปรียบเทียบประสิทธิภาพโปรแกรมคัดกรองสแปมเมล

เมื่อทำการทดสอบการทำงานและความถูกต้องของโปรแกรมคัดกรองสแปมเมลแล้ว จึงทำการบันทึกเวลาที่ใช้ในการประมวลผลทั้งในส่วนการฝึกสอนและทดสอบการจำแนกประเภทอีเมล หลังจากนั้นจะทำการหาค่าเปอร์เซ็นต์ความถูกต้องแม่นยำในการจำแนกประเภท (Accuracy) ซึ่งสามารถคำนวณได้จากสูตร ดังแสดงในสมการที่ 3.1

$$Accuracy = \frac{T\_Positive + T\_Negative}{(T\_Positive + F\_Positive) + (T\_Negative + F\_Negative)} \quad (3.1)$$

เมื่อ Accuracy คือ ค่าความถูกต้องในการจำแนกประเภทของแบบจำลอง

True Positive คือ จำนวนสแปมเมลที่ทำนายถูกว่าเป็นสแปม

True Negative คือ จำนวนแฮมเมลที่ทำนายถูกว่าเป็นแฮม

False Positive คือ จำนวนสแปมเมลที่ทำนายผิด

False Negative คือ จำนวนแฮมที่ทำนายผิด

บทที่ 4
ผลการดำเนินงาน

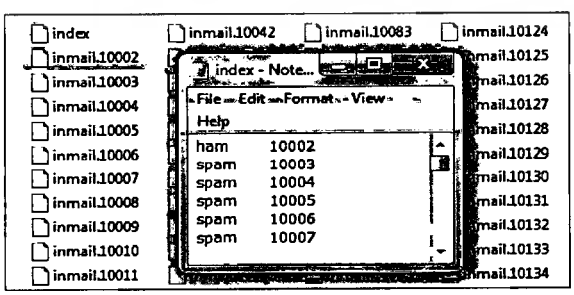
ในหัวข้อนี้จะกล่าวถึงเครื่องมือที่ใช้ในการพัฒนาระบบ ผลการพัฒนาโปรแกรมคัดกรอง สแปมเมล รวมถึงการออกแบบการทดลองเพื่อทดสอบโปรแกรมคัดกรอง สแปมเมลและผลการทดลอง ซึ่งจะกล่าวถึงดังนี้

4.1 ผลการพัฒนาโปรแกรมการคัดกรองสแปมเมล

โปรแกรมประยุกต์ที่พัฒนาขึ้นด้วยเทคนิคการแบ่งกลุ่มแบบนาอีฟเบเซียน และฟัซซี-นาอีฟเบเซียนนั้น มีกระบวนการทำงานโดยรวมของโปรแกรมเหมือนกัน ซึ่งแบ่งการทำงานออกเป็น 2 ส่วนคือ ส่วนการนำเข้าข้อมูลและฝึกสอนระบบ และส่วนการทดสอบระบบ จึงขอแสดงหน้าจอการทำงานเพียงเทคนิคเดียว ซึ่งสามารถแสดงหน้าจอส่วนต่างๆ ได้ดังนี้

4.1.1 ส่วนการเตรียมข้อมูลสำหรับฝึกสอนและทดสอบการทำนายประเภทอีเมลของโปรแกรม

1) จัดแบ่งและเก็บชุดข้อมูลแยกไว้เป็นโฟลเดอร์ตามที่ได้ออกแบบไว้ในตารางที่ 3.1 ชุดข้อมูลสำหรับฝึกสอนระบบให้เก็บไว้ในโฟลเดอร์ Training ภายในโปรแกรม
2) จัดเตรียมแฟ้มดัชนีสำหรับแต่ละชุดข้อมูล ในแฟ้มดัชนีจะระบุหมายเลขอีเมล และคลาสที่ถูกต้องของแต่ละอีเมลไว้ หมายเลขอีเมลในแฟ้มดัชนีจะต้องอ้างอิงถึงแฟ้มอีเมลที่มีอยู่จริง ดังรูปที่ 4.1 แฟ้ม inmail.10002 คืออีเมลตัวอย่างฉบับที่ 10002 และมีคลาสเป็นแฮม (ham) เก็บแฟ้มดัชนีของแต่ละชุดข้อมูลและชุดอีเมลนั้น ๆ ไว้ในโฟลเดอร์เดียวกันเพื่อความสะดวกในการเรียกใช้งาน

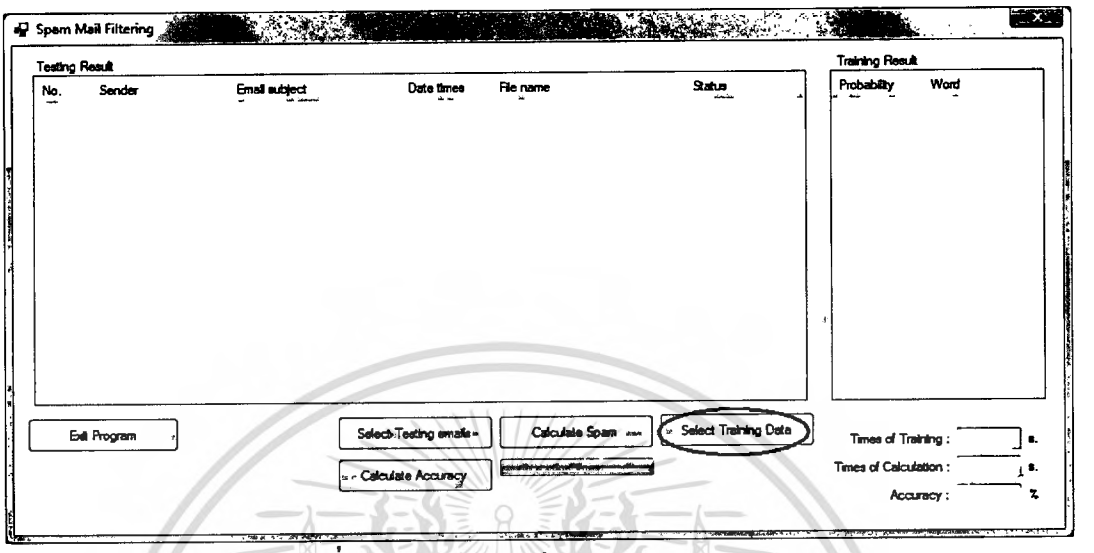


รูปที่ 4.1 การเตรียมข้อมูลสำหรับฝึกสอนระบบ

4.1.2 ส่วนนำเข้าข้อมูลและฝึกสอนระบบ

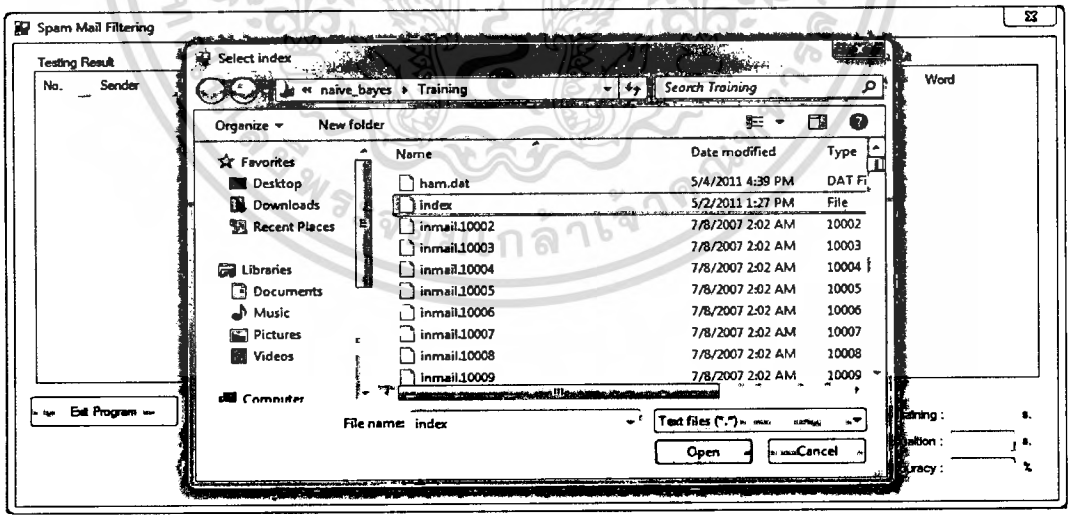
ขั้นตอนการใช้งานมีดังนี้

1) บนหน้าจอหลัก คลิกปุ่ม Select Training Data เพื่อเลือกข้อมูลสำหรับฝึกสอนระบบ



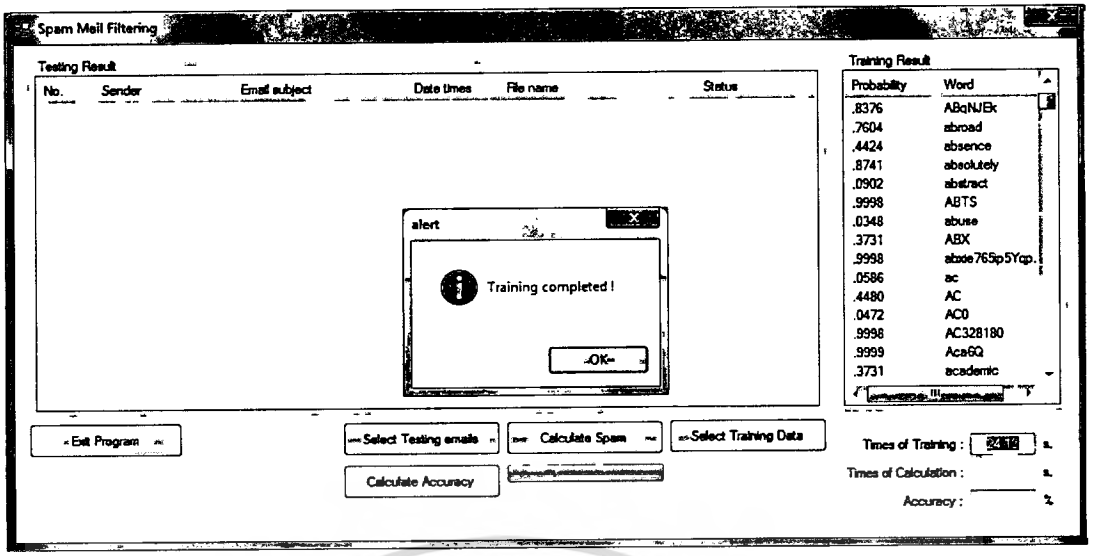
รูปที่ 4.2 การเข้าสู่การฝึกสอนระบบ

2) เลือกชุดข้อมูลที่จะนำเข้าฝึกสอนระบบโดยเลือกแฟ้ม Index จากโฟลเดอร์ Training ภายในโปรแกรม หลังจากนั้นบนหน้าจอหลักให้คลิกปุ่ม Caculate Spam



รูปที่ 4.3 การเลือกชุดข้อมูลนำเข้าฝึกสอนระบบโดยเลือกแฟ้ม Index

3) หลังจากเรียนรู้ครบทุกอีเมลแล้ว โปรแกรมจะแสดงคำศัพท์และความน่าจะเป็นสแปมของแต่ละคำที่คำนวณได้ พร้อมแสดงเวลาที่ใช้ในการเรียนรู้

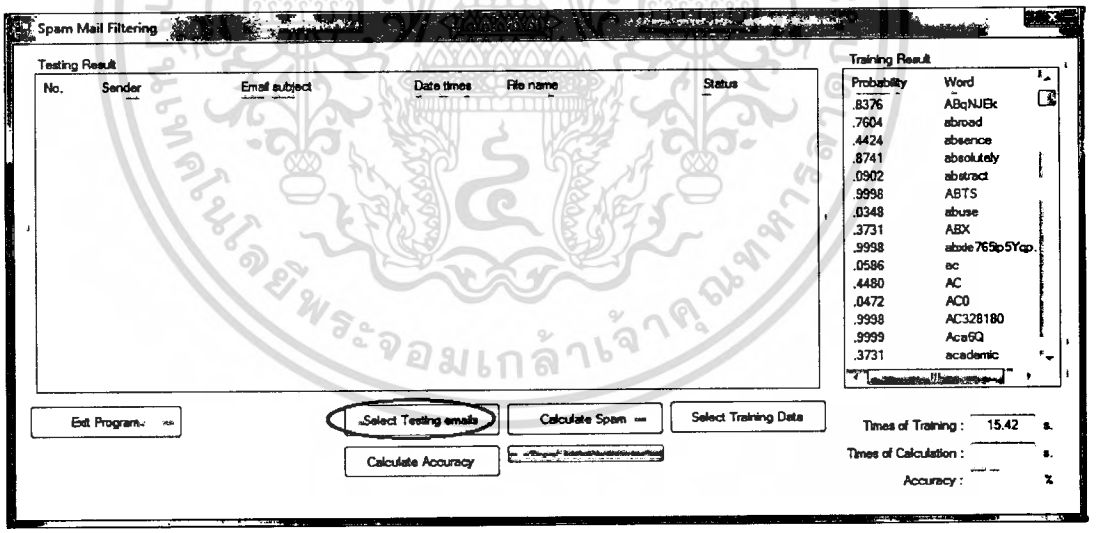


รูปที่ 4.4 แสดงผลและเวลาที่ใช้ในการเรียนรู้

4.1.3 ส่วนการทดสอบระบบ

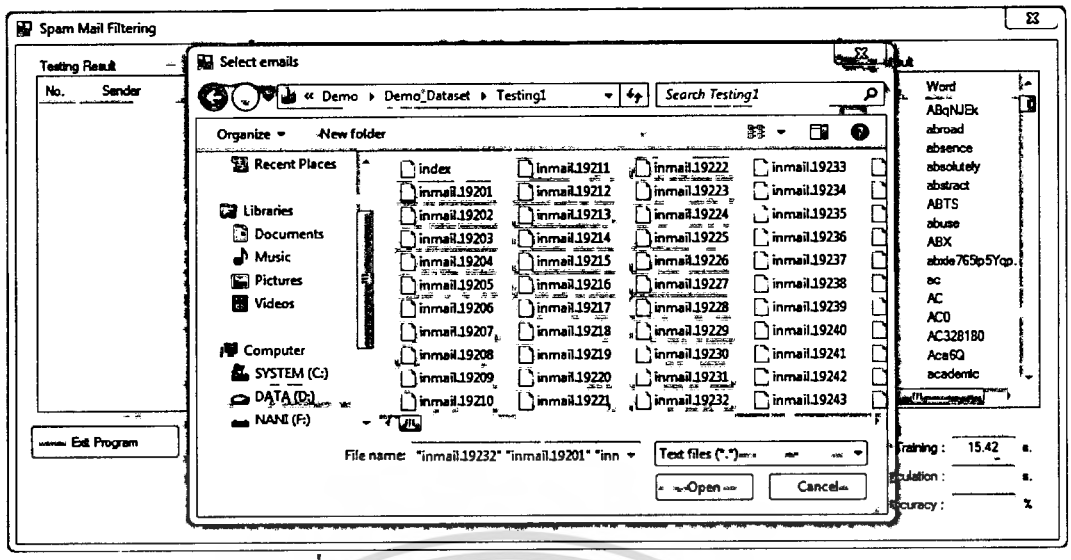
ขั้นตอนการใช้งานมีดังนี้

1) บนหน้าจอหลัก คลิกปุ่ม Select Testing Emails เพื่อเลือกข้อมูลสำหรับทดสอบระบบ



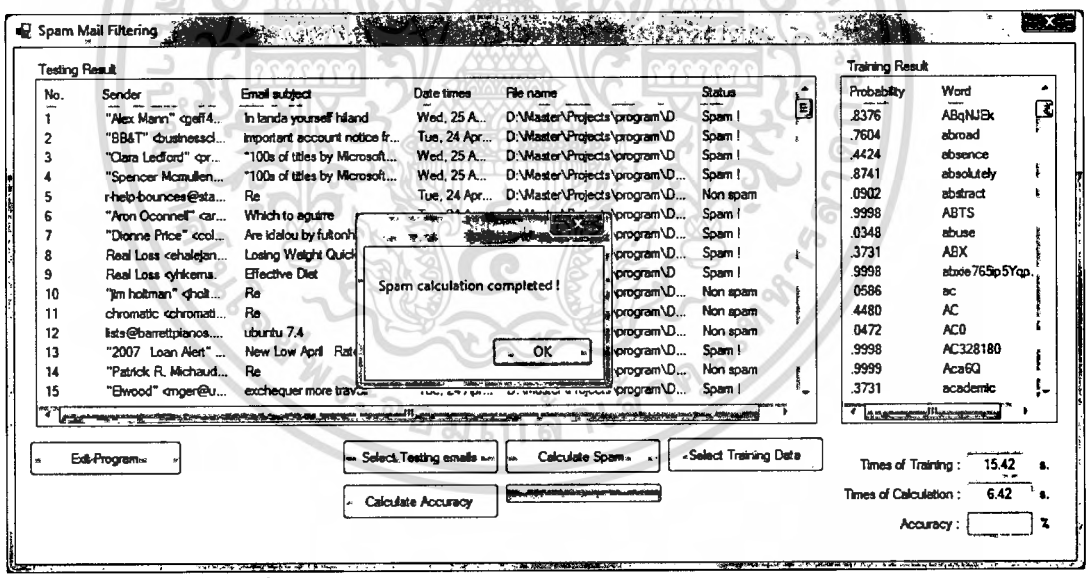
รูปที่ 4.5 เริ่มการทดสอบการทำนาย

2) เลือกข้อมูลสำหรับทดสอบระบบ หลังจากนั้นคลิกปุ่ม Calculate Spam บนหน้าจอหลัก



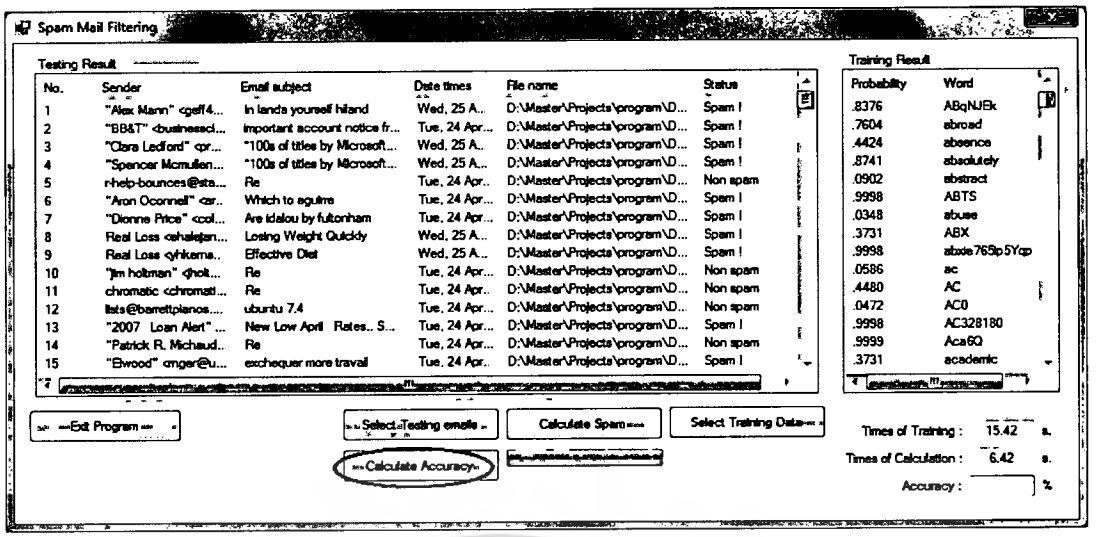
รูปที่ 4.6 การเลือกข้อมูลสำหรับทดสอบ

3) หลังจากทดสอบครบทุกอีเมลแล้ว โปรแกรมจะแสดงรายละเอียดอีเมลและผลการจำแนกประเภท พร้อมแสดงเวลาที่ใช้ในการทดสอบและเปอร์เซ็นต์ความถูกต้อง



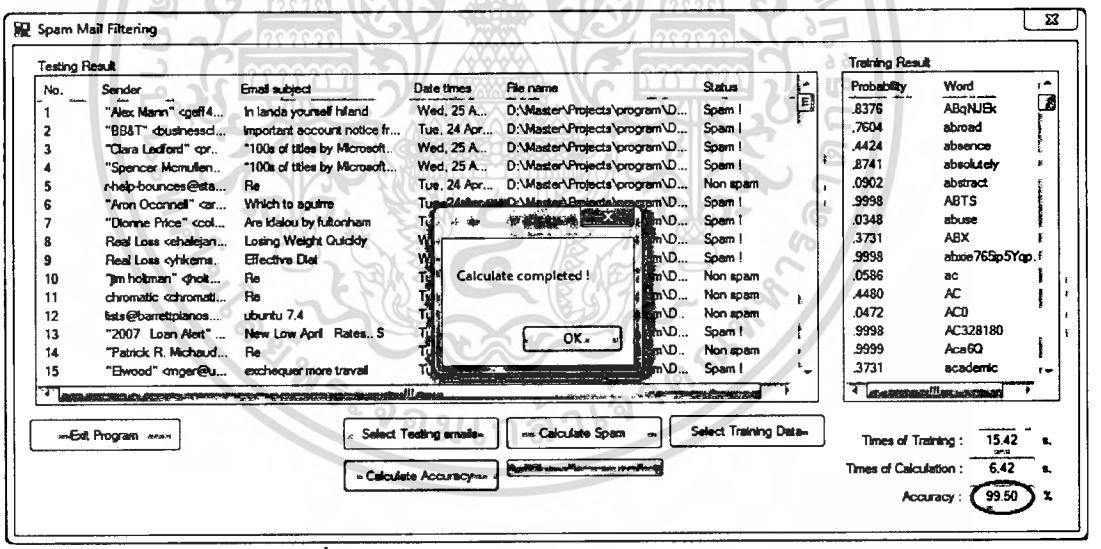
รูปที่ 4.7 ผลการทดสอบการจำแนกประเภทอีเมล

4) หลังจากคำนวณค่าความเป็นสแปมของอีเมล และทำนายประเภทของอีเมลแต่ละฉบับแล้วทำการคำนวณค่าความถูกต้องของการทำนายประเภทอีเมล โดยบนหน้าจอหลักคลิกปุ่ม Calculate Accuracy เพื่อเริ่มการคำนวณ



รูปที่ 4.8 ทำการคำนวณค่าความถูกต้องในการจำแนกประเภทอีเมล

5) เลือกแฟ้ม Index ของชุดทดสอบ เมื่อคำนวณเสร็จโปรแกรมจะแสดงผลคำนวณค่าความถูกต้องขึ้นมา



รูปที่ 4.9 แสดงผลคำนวณค่าความถูกต้อง

4.2 ผลการทดสอบโปรแกรมและเปรียบเทียบประสิทธิภาพของการจำแนกประเภทอีเมล

4.2.1 ผลการทดลอง

ผลการเปรียบเทียบประสิทธิภาพของโปรแกรมคัดกรองสแปมเมล ที่สร้างขึ้นระหว่างเทคนิคการแบ่งกลุ่มแบบนาอีฟเบเซียน และเทคนิคการแบ่งกลุ่มแบบฟัซซี-นาอีฟเบเซียน

ตารางที่ 4.1 การเปรียบเทียบเวลาที่ใช้ในการเรียนรู้

ชุดข้อมูล	จำนวนอีเมล	Naive Bayesian (s)	Fuzzy-Naive Bayesian (s)
เวลาที่ใช้ในการเรียนรู้	10,000	207.20	283.15

ตารางที่ 4.2 การเปรียบเทียบเวลาที่ใช้ในการทดสอบการจำแนกประเภท

ชุดข้อมูล	จำนวนอีเมล	Naive Bayesian (s)	Fuzzy-Naive Bayesian (s)
ชุดทดสอบ1 (จากชุดเรียนรู้)	1,000	72.62	71.82
ชุดทดสอบ 2 (TREC 2006)	2,000	136.32	137.72
ชุดทดสอบ 3 (TREC 2007)	3,000	222.52	223.15
ค่าเฉลี่ย		**143.82**	**144.23**

ตารางที่ 4.3 การเปรียบเทียบค่าความถูกต้องแม่นยำในการจำแนกประเภท

ชุดข้อมูล	จำนวนอีเมล	Naive Bayesian	Fuzzy-Naive Bayesian
ชุดทดสอบ1 (จากชุดเรียนรู้)	1,000	99.80%	99.90%
ชุดทดสอบ 2 (TREC 2006)	2,000	96.15%	98.90%
ชุดทดสอบ 3 (TREC 2007)	3,000	95.80%	98.10%
ค่าเฉลี่ย		**97.25%**	**98.97%**

บทที่ 5
สรุปผลและข้อเสนอแนะ

ผลจากการศึกษาการคัดกรองสแปมเมลโดยใช้เทคนิคการแบ่งกลุ่มแบบนาอีฟเบเซียน และการประยุกต์ใช้เทคนิคการจำแนกประเภทแบบฟัซซี-นาอีฟเบเซียน สามารถสรุปผลการทดลอง ปัญหาและข้อเสนอแนะได้ดังต่อไปนี้

5.1 สรุปผลการทดลอง

จากการพัฒนาโปรแกรมคัดกรองสแปมเมลโดยใช้เทคนิคการแบ่งกลุ่มแบบนาอีฟเบเซียน และเทคนิคการจำแนกประเภทแบบฟัซซี-นาอีฟเบเซียน พบว่าแบบจำลองที่สร้างขึ้นจากเทคนิคการแบ่งกลุ่มแบบนาอีฟเบเซียน มีประสิทธิภาพน้อยกว่าแบบจำลองที่สร้างขึ้นจากระบบที่ใช้เทคนิคฟัซซี-นาอีฟเบเซียนในทุกเงื่อนไขที่ทำการเปรียบเทียบ ซึ่งสามารถแสดงได้ดังนี้

เมื่อให้ระบบเรียนรู้ด้วยอีเมลชุดเดียวกันจำนวน 10,000 ฉบับ พบว่าเทคนิคการแบ่งกลุ่มแบบนาอีฟเบเซียน ใช้เวลาในการเรียนรู้ 207.20 วินาที และเทคนิคการแบ่งกลุ่มแบบฟัซซี-นาอีฟเบเซียนใช้เวลาในการเรียนรู้ 283.15 วินาที

เมื่อทดสอบประสิทธิภาพการจำแนกประเภทอีเมลด้วยชุดทดสอบต่างๆที่จัดแบ่งขึ้น พบว่าแบบจำลองของโปรแกรมคัดกรองสแปมเมลที่สร้างขึ้นด้วยเทคนิคการแบ่งกลุ่มแบบนาอีฟเบเซียน มีค่าความถูกต้องเฉลี่ย 97.25% และใช้เวลาในการจำแนกประเภทเฉลี่ย 143.82 วินาที ส่วนแบบจำลองที่สร้างขึ้นด้วยเทคนิคการแบ่งกลุ่มแบบฟัซซี-นาอีฟเบเซียน มีค่าความถูกต้องเฉลี่ย 144.23 วินาที และใช้เวลาในการจำแนกประเภทเฉลี่ย 98.97%

จากผลการเปรียบเทียบประสิทธิภาพการจำแนกประเภทอีเมล ระหว่างแบบจำลองที่สร้างขึ้นจากทั้ง 2 เทคนิควธี จะเห็นว่าเทคนิคการแบ่งกลุ่มแบบฟัซซี-นาอีฟเบเซียน สามารถสร้างแบบจำลองที่มีเปอร์เซ็นต์ความถูกต้องในการจำแนกประเภทที่ดีกว่า

ด้านระยะเวลาที่ใช้ในการสร้างแบบจำลอง เทคนิคการแบ่งกลุ่มแบบฟัซซี-นาอีฟเบเซียน จะใช้ระยะเวลามากกว่าเทคนิคการแบ่งกลุ่มแบบนาอีฟเบเซียน เนื่องจากมีการเรียนรู้โดยนำค่าความน่าจะเป็นสแปมของแต่ละคำศัพท์ ไปปรับค่าความเป็นสมาชิกของเซตสแปมและแฮมที่เหมาะสมสำหรับแต่ละคำศัพท์อีกครั้ง ในขณะที่เทคนิคการแบ่งกลุ่มแบบนาอีฟเบเซียนจะเก็บสถิติที่คำศัพท์ปรากฏ และคำนวณค่าความน่าจะเป็นสแปมเท่านั้น

5.2 ปัญหาและข้อจำกัด

5.2.1 โปรแกรมนี้ไม่สามารถจำแนกประเภทอีเมลที่เป็นภาษาไทยได้

เนื่องจากโปรแกรมได้รับการฝึกสอนเพื่อสร้างแบบจำลองจากชุดอีเมลตัวอย่างที่เป็นภาษาอังกฤษ กระบวนการตัดคำจึงได้ถูกออกแบบให้พิจารณาตัดคำตามเงื่อนไขที่กำหนดใช้กับภาษาอังกฤษเท่านั้น ซึ่งหากนำมาใช้กับอีเมลภาษาไทย เมื่อพิจารณาตัดคำภาษาไทย จะทำให้ได้คำศัพท์ที่ยาวและมีความถี่ที่ซ้ำกันน้อยมาก หรือไม่ซ้ำกันเลย การพิจารณาคัดแยกสแปมเมลภาษาไทยจึงไม่น่าเชื่อถือ รวมทั้งหากต้องการฝึกสอนอีเมลที่เป็นภาษาไทย จะต้องมีการเก็บรวบรวมตัวอย่างอีเมลภาษาไทยทั้งที่เป็นสแปมและแฮมจำนวนมากด้วยเช่นกัน

5.2.2 ปัญหาของตัวกรองสแปมแบบเบเซียน

ตัวกรองสแปมแบบเบเซียนยังคงมีปัญหาอื่นๆ เช่น ผลลัพธ์ลวง (False positive) คือจำแนกอีเมลผิดพลาดจากความเป็นจริง และการวางยา (Bayesian poisoning) คือการหลีกเลี่ยงการใช้คำที่มักจะถูกพบว่าเป็นสแปมในข้อความที่จะทำการสแปม หรือแม้กระทั่งเพิ่มคำที่มีค่าความน่าจะเป็นแฮมสูงเข้ามาในข้อความ เพื่อทำให้โอกาสที่อีเมลนั้นจะถูกตัดสินว่าเป็นสแปมลดลง

5.2.3 จำนวนอีเมลตัวอย่างในการฝึกสอนมีผลต่อความถูกต้องแม่นยำของระบบ

การจะได้มาซึ่งความน่าจะเป็นที่เหมาะสม จะต้องรวบรวมอีเมลที่เป็นสแปมและแฮมจำนวนมากเพื่อฝึกสอนระบบ และเพื่อให้การตัดสินใจของระบบเหมาะสมกับสภาพแวดล้อมของผู้ใช้งาน ทั้งนี้เนื่องจากหากมีจำนวนคำที่ใช้ในการฝึกสอนระบบเพิ่มมากขึ้น สามารถทำให้ความถูกต้องแม่นยำของระบบเพิ่มขึ้น

5.3 ข้อเสนอแนะ

5.3.1 พัฒนาระบบให้สามารถคัดกรองอีเมลภาษาไทยได้

เนื่องจากรูปแบบการเขียนภาษาไทยนั้น เป็นการเขียนคำแต่ละคำติดกันและเว้นวรรคเมื่อหมดประโยค เงื่อนไขการพิจารณาตัดคำจึงควรใช้เทคนิคกลไกการเรียนรู้การตัดคำสำหรับภาษาไทยโดยเฉพาะ เช่น โปรแกรม CU Thai Segmentation

5.3.2 หาแนวทางแก้ปัญหาของตัวกรองสแปมแบบเบเซียนที่มีอยู่

เนื่องจากตัวกรองสแปมแบบเบเซียนยังคงมีปัญหาอื่นๆ ดังกล่าวข้างต้น ข้อความอีเมลที่ถูกจำแนกผิดพลาดนั้น ล้วนแต่เป็นข้อความที่ผู้ส่งสารต้องการส่งถึงผู้รับทั้งสิ้น ซึ่งในบางกรณีอาจเป็นข้อความที่มีความสำคัญมาก หากข้อความนั้นไม่ถูกส่งถึงผู้รับสารอาจก่อให้เกิดความเสียหายได้ อีกทั้งปัญหาการวางยามีความสัมพันธ์โดยตรงกับผลลัพธ์ลวง เนื่องจากผู้ทำการสแปมก็หวังว่าการวางยาจะส่งผลให้ตัวกรองสแปมนั้นมีอัตราผลลัพธ์ลวงที่สูง ดังนั้นค่าผลลัพธ์ลวงถึงจัดได้ว่าเป็นความผิดพลาดอย่างหนึ่งที่ควรหลีกเลี่ยงหรือพยายามควบคุมให้เกิดน้อยที่สุด

บรรณานุกรม

กอบเกียรติ สระอุบล. 2552. "การกรองสแปมจากบอทเน็ต". สารนิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

ณัฏฐา ห่อประทุม. 2547. "ระบบแบ่งกลุ่มแบบฟัซซีเบเซียนและการประยุกต์ใช้ในการอนุมัติสินเชื่อ เบื้องต้น". สารนิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

นันทชัย สมัญญาภรณ์. 2550. "ระบบคัดกรองเมลขยะสำหรับเว็บเบสอีเมล". โครงงานพัฒนาระบบวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

พยุง มีสัจ, (ผู้รวบรวม). 2553. โครงข่ายประสาทเทียมและระบบฟัซซี. กรุงเทพฯ: สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

Graham Paul. 2002. A Plan for Spam. [Online] Available: http://www.paulgraham.com/spam.html.

Graham Paul. 2003. Better Bayesian Filtering. [Online] Available: http://www.paulgraham.com/better.html.

Ke Chen. 2010. Machine Learning Lectures: Naive Bayes. [Online] Available: http://www.cs.manchester.ac.uk/ugt/COMP24111/lectures.php.

National Institute of Standards and Technology (NIST). 2010. Text Retrieval Conference (TREC) Public Spam Corpora. [Online] Available: http://trec.nist.gov/data/spam.html.

ภาคผนวก ก

ก.1 ผลการทดลองโปรแกรมคัดกรองสแปมเมลที่สร้างขึ้นด้วยเทคนิคการแบ่งกลุ่มแบบฟัซซี-นาอีฟเบเซียน

ผลการทดลองประสิทธิภาพการทำงานและการจำแนกประเภทอีเมลของแบบจำลองที่สร้างขึ้นด้วยเทคนิคการแบ่งกลุ่มแบบฟัซซี-นาอีฟเบเซียนเมื่อปรับช่วงของฟัซซีเซตเพิ่มขึ้นและลดลงเป็น 0.002, 0.003 และ 0.005 ตามลำดับดังนี้

ก.1.1 ผลการทดลองเมื่อปรับช่วงของฟัซซีเซตเพิ่ม-ลด 0.002

ตารางที่ ก.1 เวลาที่ใช้ในการเรียนรู้ เมื่อปรับช่วงเพิ่ม-ลด 0.002

ชุดข้อมูล	จำนวนอีเมล	Fuzzy-Naive Bayesian (s)
เวลาที่ใช้ในการเรียนรู้	10,000	268.04

ตารางที่ ก.2 เวลาที่ใช้ในการทดสอบการจำแนก เมื่อปรับช่วงเพิ่ม-ลด 0.002

ชุดข้อมูล	จำนวนอีเมล	Fuzzy-Naive Bayesian (s)
ชุดทดสอบ1 (จากชุดเรียนรู้)	1,000	69.31
ชุดทดสอบ 2 (TREC 2006)	2,000	145.79
ชุดทดสอบ 3 (TREC 2007)	3,000	188.63
ค่าเฉลี่ย		**134.58**

ตารางที่ ก.3 ค่าความถูกต้องแม่นยำในการจำแนกประเภท เมื่อปรับช่วงเพิ่ม-ลด 0.002

ชุดข้อมูล	จำนวนอีเมล	Fuzzy-Naive Bayesian
ชุดทดสอบ1 (จากชุดเรียนรู้)	1,000	99.80%
ชุดทดสอบ 2 (TREC 2006)	2,000	95.40%
ชุดทดสอบ 3 (TREC 2007)	3,000	95.23%
ค่าเฉลี่ย		**96.81%**

ก.1.2 ผลการทดลองเมื่อปรับช่วงของฟัซซีเซตเพิ่ม-ลด 0.003

ตารางที่ ก.4 เวลาที่ใช้ในการเรียนรู้ เมื่อปรับช่วงเพิ่ม-ลด 0.003

ชุดข้อมูล	จำนวนอีเมล	Fuzzy-Naive Bayesian (s)
เวลาที่ใช้ในการเรียนรู้	10,000	271.35

ตารางที่ ก.5 เวลาที่ใช้ในการทดสอบการจำแนก เมื่อปรับช่วงเพิ่ม-ลด 0.003

ชุดข้อมูล	จำนวนอีเมล	Fuzzy-Naive Bayesian (s)
ชุดทดสอบ1 (จากชุดเรียนรู้)	1,000	58.53
ชุดทดสอบ 2 (TREC 2006)	2,000	146.89
ชุดทดสอบ 3 (TREC 2007)	3,000	194.70
ค่าเฉลี่ย		**133.37**

ตารางที่ ก.6 ค่าความถูกต้องแม่นยำในการจำแนกประเภท เมื่อปรับช่วงเพิ่ม-ลด 0.003

ชุดข้อมูล	จำนวนอีเมล	Fuzzy-Naive Bayesian
ชุดทดสอบ1 (จากชุดเรียนรู้)	1,000	98.60%
ชุดทดสอบ 2 (TREC 2006)	2,000	92.90%
ชุดทดสอบ 3 (TREC 2007)	3,000	93.27%
ค่าเฉลี่ย		**94.92%**

ก.1.3 ผลการทดลองเมื่อปรับช่วงของฟัซซีเซตเพิ่ม-ลด 0.005

ตารางที่ ก.7 เวลาที่ใช้ในการเรียนรู้ เมื่อปรับช่วงเพิ่ม-ลด 0.005

ชุดข้อมูล	จำนวนอีเมล	Fuzzy-Naive Bayesian (s)
เวลาที่ใช้ในการเรียนรู้	10,000	285.15

ตารางที่ ก.8 เวลาที่ใช้ในการทดสอบการจำแนก เมื่อปรับช่วงเพิ่ม-ลด 0.005

ชุดข้อมูล	จำนวนอีเมล	Fuzzy-Naive Bayesian (s)
ชุดทดสอบ1 (จากชุดเรียนรู้)	1,000	73.86

ตารางที่ ก.8 (ต่อ)

ชุดทดสอบ 2 (TREC 2006)	2,000	129.74
ชุดทดสอบ 3 (TREC 2007)	3,000	191.70
ค่าเฉลี่ย		**131.77**

ตารางที่ ก.9 ค่าความถูกต้องแม่นยำในการจำแนกประเภท เมื่อปรับช่วงเพิ่ม-ลด 0.005

ชุดข้อมูล	จำนวนอีเมล	Fuzzy-Naive Bayesian
ชุดทดสอบ1 (จากชุดเรียนรู้)	1,000	90.20%
ชุดทดสอบ 2 (TREC 2006)	2,000	83.05%
ชุดทดสอบ 3 (TREC 2007)	3,000	89.93%
ค่าเฉลี่ย		**87.72%**

ประวัติผู้เขียน

ชื่อ-นามสกุล	นางสาวนิวภรณ์ หัวนาวิลาส
สถานที่เกิด	จังหวัดพะเยา
ประวัติการศึกษา	สำเร็จการศึกษาระดับปริญญาตรี วิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร
ประวัติการทำงาน	- Worldwide Antimalarial Resistance Network: WWARN (พ.ศ. 2553 - ปัจจุบัน) ตำแหน่ง: IT Specialist - Western Digital (พ.ศ. 2549 - พ.ศ. 2553) ตำแหน่ง: IT Administrator - HCL Technologies, BPO (พ.ศ. 2547 - พ.ศ. 2549) ตำแหน่ง: Technical Support Professional - Tatung Thailand (พ.ศ. 2545 - พ.ศ. 2547) ตำแหน่ง: Software Engineer