ตัวกรองอีเมล์ขยะด้วยทฤษฎีเจเนติกอัลกอริทึม

SPAM MAIL FILTERING USING GENETIC ALGORITHM

อุษารัตน์ แสนป่ากดี
USARAT SANPAKDEE

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตรคอมพิวเตอร์
บัณฑิตวิทยาลัย
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2549

ISBN 974-15-2794-2

ตัวกรองอีเมล์ขยะด้วยทฤษฎีเจเนติกอัลกอริทึม

SPAM MAIL FILTERING USING GENETIC ALGORITHM

อุษารัตน์ แสนปากดี

USARAT SANPAKDEE

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2549

ISBN 974-15-2794-2

SPAM MAIL FILTERING USING GENETIC ALGORITHM

USARAT SANPAKDEE

A THESIS SUBMITTED PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF ENGINEERING IN COMPUTER ENGINEERING

SCHOOL OF GRADUATE STUDIES

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2006

ISBN 974-15-2794-2

หัวข้อวิทยานิพนธ์	ตัวกรองอีเมล์ขยะด้วยทฤษฎีเจเนติกอัลกอริทึม
นักศึกษา	นางสาวอุษารัตน์ แสนปากดี
รหัสประจำตัว	47060802
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
พ.ศ.	2549
อาจารย์ผู้ควบคุมวิทยานิพนธ์	ดร.สมศักดิ์ วลัยรัชต์

บทคัดย่อ

ในปัจจุบัน มีการคิดค้นวิธีการกำจัดอีเมล์ขยะขึ้นมาหลากหลายวิธี หนึ่งในวิธีที่นิยมคือการใช้เครื่องเรียนรู้ (Machine Learning) วิทยานิพนธ์นี้ได้นำเสนอ วิธีการกรองอีเมล์ขยะด้วยทฤษฎีเจเนติกอัลกอริทึม (Genetic Algorithm) ซึ่งกระบวนการหลักๆ ของเจเนติกอัลกอริทึมที่นำมาใช้ประกอบด้วย การคัดเลือก (Selection) การครอสโอเวอร์ (Crossover) และการมิวเตชัน (Mutation) กระบวนการเหล่านี้ถูกนำมาใช้การสร้างแม่แบบของอีเมล์ ซึ่งสืบทอดมาจากอีเมล์ที่มีอยู่เดิม แม่แบบของอีเมล์ที่ถูกสร้างขึ้นมาใหม่จะถูกคัดเลือกเฉพาะแม่แบบที่มีความเหมาะสมมากกว่าแม่แบบอื่นๆ เอาไว้ ตามหลักการของเจเนติกอัลกอริทึม (Fittest of Survival) ทั้งนี้ แม่แบบของอีเมล์ที่ถูกสร้างขึ้นมาใหม่ทำให้ตัวกรองได้เรียนรู้รูปแบบของอีเมล์ใหม่ๆ ส่งผลให้เกิดทางเลือกที่จะนำมาใช้ในการตัดสินใจว่าอีเมล์เป็นอีเมล์ประเภทใดเพิ่มมากขึ้น ซึ่งเมื่อนำตัวกรองอีเมล์ขยะที่นำเสนอมาทดสอบประสิทธิภาพโดยเปรียบเทียบกับตัวกรองอีเมล์ขยะโดยประยุกต์ใช้ทฤษฎีเบย์เซียนพบว่ามีประสิทธิที่ดีกว่าโดยมี ค่า Accuracy 87.05% ค่า Recall 88.50% และค่า Precision 86.35%

Thesis Title	Spam Mail Filtering Using Genetic Algorithm
Student	Miss Usarat Sanpakdee
Student ID.	47060802
Degree	Master of Engineering
Program	Computer Engineering
Year	2006
Thesis Advisor	Dr. Somsak Walairacth

ABSTRACT

Nowadays, there are several methods to eliminate spam mail. One of popular method is eliminating spam mail by using machine learning. In this paper, we present a spam mail filtering using genetic algorithm. We applied Genetic operator, crossover and mutation to create varieties of mails templates which inherit from old mails. Therefore, it saves time for filter to prepared training set and not need large training set to learn like others method. New mails templates which have more fitness will be selected according to "fittest of survival" of genetic principle. New mail templates are the result of new learning and new choices which improve the spam mail filter efficiency and decided whether the incoming e-mail is spam. In this thesis, we compare the efficiency of filter with Bayesian spam filter. We found that the propose method has more efficient than Bayesian spam filter. With the accuracy 87.05%, recall is 88.50%and precision 86.35%.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ ด้วยคำแนะนำและคำปรึกษาจาก ดร. สมศักดิ์ วลัยรัชต์ และ ดร. อรัญญา วลัยรัชต์ ซึ่งเป็นอาจารย์ผู้ควบคุมวิทยานิพนธ์ผู้ดำเนินการวิจัยขอขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณสมาชิกในครอบครัวทุกๆ ท่าน ที่ให้กำลังใจมาตลอด

ขอขอบพระคุณคณาจารย์ภาควิชาวิศวกรรมคอมพิวเตอร์ทุกท่าน ที่ได้ประสิทธิ์ประสาทวิชาความรู้ ตลอดระยะเวลาที่ศึกษาอยู่ในสถานศึกษาแห่งนี้

ขอขอบคุณเพื่อน ๆ พี่น้องนักศึกษาทุกคนที่ได้ให้ความช่วยเหลือและความรู้ รวมทั้งเป็นที่ปรึกษาเมื่อเกิดปัญหาในงานวิจัย

สุดท้ายขอขอบคุณ บัณฑิตวิทยาลัยที่ได้ให้ทุนสนับสนุนการทำวิทยานิพนธ์ครั้งนี้

คุณงามความดีและประโยชน์ที่ได้จากวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบแด่บิดามารดาอันเป็นที่รักและเคารพยิ่ง ตลอดจนผู้มีพระคุณทุกท่าน

อุษารัตน์ แสนปากดี

สารบัญ

หน้า

บทคัดย่อภาษาไทย I

บทคัดย่อภาษาอังกฤษ II

กิตติกรรมประกาศ III

สารบัญ IV

สารบัญตาราง VIII

สารบัญรูป IX

บทที่ 1 บทนำ 1

1.1 ความเป็นมาและความสำคัญของปัญหา 1

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา 2

1.3 สมมติฐานของการศึกษา 2

1.4 ทฤษฎีหรือแนวความคิดที่ใช้ในการวิจัย 2

1.5 ขอบเขตการวิจัย 2

1.6 ขั้นตอนของการศึกษา 3

1.7 ข้อจำกัดของการศึกษา 3

1.8 คำจำกัดความที่ใช้ในการศึกษา 3

1.9 เครื่องมือและอุปกรณ์ที่ใช้ในงานวิจัย 4

1.10 โครงสร้างของวิทยานิพนธ์ 4

บทที่ 2 งานวิจัยที่เกี่ยวข้อง 5

2.1 เทคนิคต่างๆในการกรองอีเมล์ขยะ 5

2.1.1 เทคนิคการกรองอีเมล์ขยะโดยใช้การกระตุ้นและตอบสนอง 5

2.1.2 เทคนิคการกรองอีเมล์ขยะโดยใช้การวิเคราะห์ส่วนหัวของอีเมล์ 5

2.1.3 เทคนิคการกรองอีเมล์ขยะโดยใช้ลายเซ็นดิจิตอล 6

2.1.4 เทคนิคการกรองอีเมล์ขยะโดยใช้บัญชีที่อยู่ 6

2.1.5 เทคนิคการกรองอีเมล์ขยะโดยใช้รายการคำสำคัญ ตอบสนอง 6

2.1.6 เทคนิคการกรองอีเมล์ขยะโดยการพิจารณาโอเพนรีเลย์ 7

2.1.7 เทคนิคการกรองอีเมล์ขยะโดยการพิจารณาโอเพนพร็อกซี่ 7

สารบัญ (ต่อ)

หน้า

2.2 การกรองอีเมล์ขยะโดยใช้เครื่องมือเรียนรู้และวิธีการทางสถิติ 7
2.2.1 การกรองอีเมล์ขยะโดยใช้ตัวกรองเบย์เซียน 7
2.2.2 การกรองอีเมล์ขยะโดยใช้การแบ่งแยกของมาร์คอฟเวียน 8

บทที่ 3 ความรู้พื้นฐานที่เกี่ยวข้องกับงานวิจัย 9
3.1 ทฤษฎีที่เกี่ยวข้องกับอีเมล์ 9
3.1.1 ความเป็นมาของอีเมล์ 9
3.1.2 ลักษณะการทำงานของระบบรับ-ส่งอีเมล์ 9
3.1.3 สถาปัตยกรรมของระบบเมล์ 10
3.1.4 โปรโตคอล 12
3.1.4.1 SMTP 12
3.1.4.2 POP 13
3.1.4.3 IMAP 14
3.1.5 คำนิยามของอีเมล์ขยะ 14
3.2 องค์ประกอบของอีเมล์ 15
3.2.1 ส่วนหัวเรื่อง 15
3.2.2 ส่วนเนื้อเรื่อง 16
3.3 ผลกระทบของอีเมล์ขยะ 16
3.4 ความรู้พื้นฐานเกี่ยวกับทฤษฎีเบย์เซียน 17
3.4.1 ทฤษฎีเบย์เซียน 18
3.4.2 การนำทฤษฎีเบย์เซียนมาใช้ในการกรองอีเมล์ขยะ 19
3.5 ความรู้พื้นฐานเกี่ยวกับแขนงการตัดสินใจ 20
3.6 ความรู้พื้นฐานเกี่ยวกับเจเนติกอัลกอริทึม 21
3.6.1 พันธุศาสตร์ทางชีววิทยากับเจเนติกอัลกอริทึม 22
3.6.1.1 พันธุศาสตร์ทางชีววิทยา 23
3.6.1.2 พันธุศาสตร์ทางเจเนติกอัลกอริทึม 23
3.6.2 ขั้นตอนการทำงานของเจเนติกอัลกอริทึม 24
3.6.2.1 การกำหนดรูปแบบโครโมโซม 24
3.6.2.2 ประชากร 25

สารบัญ (ต่อ)

หน้า

3.6.2.3 กำหนดฟังก์ชันความเหมาะสม 25
3.6.2.4 การวิเคราะห์ค่าความเหมาะสม 25
3.6.2.5 การคัดเลือก 25
3.6.2.6 การครอสโอเวอร์ 28
3.6.2.7 การมิวเตชัน 29
3.6.2.8 การสร้างประชากรรุ่นใหม่ 30
3.6.2.9 การกำหนดค่าตัวแปรต่างๆ 31

บทที่ 4 ตัวกรองอีเมล์ขยะด้วยทฤษฎีเจเนติกอัลกอริทึม 33
4.1 การเตรียมข้อมูล 34
4.2 การสร้างฐานข้อมูลของคำ 34
4.3 การประยุกต์ใช้เจเนติกอัลกอริทึม 35
4.3.1 การกำหนดรูปแบบของโครโมโซม 35
4.3.2 การประเมินค่าความเหมาะสม 38
4.3.3 การคัดเลือกประชากร 39
4.3.4 การครอสโอเวอร์และการมิวเตชัน 40
4.3.5 การทดสอบประสิทธิภาพของระบบ 41

บทที่ 5 ผลการทดลองและการวิเคราะห์ 44
5.1 ขั้นตอนการทดลองตัวกรองอีเมล์ขยะโดยใช้เจเนติกอัลกอริทึม 44
5.2 การปรับพารามิเตอร์ต่างๆ ในงานวิจัย 44
5.2.1 การเปรียบเทียบวิธีการคัดเลือกโครโมโซมระหว่างการคัดเลือกแบบวงล้อถ่วงน้ำหนักและการคัดเลือกแบบจัดลำดับ 45
5.2.2 การปรับเปอร์เซ็นต์การเลือกคู่โครโมโซมพ่อแม่ 48
5.2.3 การปรับค่ามิวเตชันพารามิเตอร์ 50
5.2.4 การปรับครอสโอเวอร์พารามิเตอร์ 53
5.2.5 การปรับค่าขีนเธร์โชว์ 54
5.3 ผลการทดลองการกรองอีเมล์ขยะโดยใช้เจเนติกอัลกอริทึม 59
5.4 เปรียบเทียบตัวกรองอีเมล์ขยะที่นำเสนอกับตัวกรองอีเมล์ขยะซึ่งประยุกต์

สารบัญ (ต่อ)

หน้า

ใช้ทฤษฎีเบย์เซียน 61

บทที่ 6 สรุปผลการทดลองและข้อเสนอแนะ 64

6.1 สรุปและวิเคราะห์ผลการดำเนินงานวิจัย 64

6.2 ปัญหาที่พบในวิทยานิพนธ์.....65

6.3 ข้อเสนอแนะและแนวทางในการพัฒนาต่อ 66

เอกสารอ้างอิง 67

ภาคผนวก ก ผลงานวิจัยที่ได้รับการตีพิมพ์.....69

ประวัติผู้เขียน 75

สารบัญตาราง

ตารางที่ หน้า

3.1 คำสำคัญที่ใช้ในกระบวนการเจเนติกอัลกอริทึม 24
3.2 ตัวอย่างการใช้แบบจำลองวงล้อถ่วงน้ำหนัก 27
4.1 แสดงตัวอย่างของคำเฉพาะในแต่ละกลุ่ม..... 35
4.2 แสดงการคิดค่าเฉลี่ยความน่าจะเป็นของคำเฉพาะในอีเมล์ตัวอย่าง..... 36
4.3 แสดงการแปลงค่าความน่าจะเป็นก่อนปรับและหลังปรับและค่าความน่าจะเป็นหลังปรับ เมื่อแปลงเป็นเลขฐานสอง..... 37
4.4 แสดงตัวอย่างโครโมโซมแม่แบบ และค่าความเหมาะสมของโครโมโซม..... 39
5.1 แสดงการเปรียบเทียบค่า Accuracy, Recall และ Precision ระหว่างการคัดเลือกแบบวงล้อถ่วงน้ำหนัก (Roulette Wheel) และการคัดเลือกแบบจัดลำดับ (Ranking)..... 46
5.2 แสดงค่า Accuracy, Recall และ Precision ที่การคัดเลือกคู่โครโมโซมพ่อแม่ที่เปอร์เซ็นต์ต่างๆ..... 49
5.3 แสดงค่า Accuracy, Recall และ Precision ของการมิวเตชันเมื่อผ่านไป 20, 40, 60, 80 Generation และเมื่อไม่ทำการมิวเตชันเลย (No)..... 50
5.4 แสดงค่า Accuracy, Recall และ Precision ของการมิวเตชันเมื่อปรับจำนวนโครโมโซมที่เปอร์เซ็นต์ต่างๆ..... 51
5.5 แสดงค่า Accuracy, Recall และ Precision ของการครอสโอเวอร์ทั้งสองแบบ..... 53
5.6 แสดงค่า Accuracy, Recall และ Precision ของยีนเธร์โชว์ที่แตกต่างกัน..... 55
5.7 แสดงค่า Accuracy, Recall และ Precision ของการปรับจำนวนคู่โครโมโซมพ่อแม่และยีนเธร์โชว์ที่ค่าต่างๆ..... 57
5.8 แสดงค่าเฉลี่ย Accuracy, Recall และ Precision ของการทดลองในแต่ละเซตและค่าเฉลี่ย Accuracy, Recall และ Precision ของทุกๆ เซต..... 59
5.9 แสดงค่า Accuracy, Recall และ Precision ของเซตข้อมูลทั้ง 5 เซต..... 61

สารบัญรูป

รูปที่ หน้า

2.1 แสดงลำดับขั้นในกระบวนการกระตุ้น- ตอบสนองและชนิดของการตอบสนองที่ต้องการ 5
3.1 แสดงลักษณะการทำงานของระบบอีเมล์ 10
3.2 แสดงสถาปัตยกรรมในทีซีพี/ไอพี 11
3.3 แสดงโปรโตคอลที่ใช้ในการรับส่งเมล์ 12
3.4 แสดงตัวอย่างของอีเมล์ขยะ 15
3.5 แสดงตัวอย่างของแขนงการตัดสินใจ 21
3.6 แสดงการเข้ารหัสแบบไบนารี 24
3.7 แสดงการเข้ารหัสแบบสลับลำดับ 24
3.8 แสดงการเข้ารหัสแบบค่า 24
3.9 แสดงการเข้ารหัสแบบทรี 25
3.10 แสดงการเลือกโครโมโซมตามค่าความเหมาะสม 26
3.11 แสดงกระบวนการครอสโอเวอร์แบบไบนารี 29
3.12 แสดงกระบวนการครอสโอเวอร์แบบตัวอักษร 29
3.13 การครอสโอเวอร์ของโครโมโซมแบบทรี 29
3.14 แสดงกระบวนการไบนารีมิวเตชัน 30
3.15 แสดงกระบวนการสร้างประชากรในรุ่นถัดไป 31
4.1 แสดงบล็อกไดอะแกรมของระบบ 33
4.2 แสดงรูปแบบของโครโมโซม 36
4.3 แสดงการเก็บผลการจำแนกของแต่ละแม่แบบ (Template) สำหรับใช้คำนวณหาค่าความเหมาะสม 38
4.4 แสดงแม่แบบของอีเมล์ทั้งหมดที่ได้จากกระบวนการเจเนติกอัลกอริทึมในรุ่นนี้ 41
4.5 แสดงการจำแนกอีเมล์โดยพิจารณาเปรียบเทียบจากค่าความน่าจะเป็นเฉลี่ยในยีน 42
5.1 แสดงการเปรียบเทียบค่า Accuracy, Recall และ Precision ระหว่างการคัดเลือกแบบวงล้อถ่วงน้ำหนัก (Roulette Wheel) และการคัดเลือกแบบจัดลำดับ (Ranking) 46
5.2 แสดงค่า Accuracy ของการการคัดเลือกแบบวงล้อถ่วงน้ำหนัก และการคัดเลือกแบบจัดลำดับ 47
5.3 แสดงค่า Recall ของการการคัดเลือกแบบวงล้อถ่วงน้ำหนัก และการคัดเลือกแบบจัดลำดับ 47

สารบัญรูป(ต่อ)

รูปที่ หน้า

5.4 แสดงค่า Precision ของการการคัดเลือกแบบวงล้อถ่วงน้ำหนัก และการคัดเลือกแบบจัดลำดับ 48
5.5 แสดงค่า Accuracy, Recall และ Precision ที่การคัดเลือกคู่โครโมโซมพ่อแม่ที่เปอร์เซ็นต์ต่างๆ 48
5.6 แสดงค่า Accuracy, Recall และ Precision ของการมิวเตชันเมื่อผ่านไป 20, 40, 60, 80 Generation และเมื่อไม่ทำการมิวเตชันเลย (No) 52
5.7 แสดงค่า Accuracy, Recall และ Precision ของการมิวเตชันเมื่อปรับจำนวนโครโมโซมที่เปอร์เซ็นต์ต่างๆ 52
5.8 แสดงค่า Accuracy, Recall และ Precision ของการครอสโอเวอร์ทั้งสองแบบ 54
5.9 แสดงค่า Accuracy, Recall และ Precision ของยีนเธร์โชว์ที่แตกต่างกัน 56
5.10 แสดงค่า Accuracy ของเซตข้อมูลทั้ง 5 เซตเมื่อใช้พารามิเตอร์ที่ได้จากการทดลอง 60
5.11 แสดงค่า Precision ของเซตข้อมูลทั้ง 5 เซตเมื่อใช้พารามิเตอร์ที่ได้จากการทดลอง 60
5.12 แสดงค่า Precision ของเซตข้อมูลทั้ง 5 เซตเมื่อใช้พารามิเตอร์ที่ได้จากการทดลอง 61
5.13 แสดงการเปรียบเทียบค่า Accuracy ระหว่างเจเนติกอัลกอริทึมและเบย์เซียน 62
5.14 แสดงการเปรียบเทียบค่า Recall ระหว่างเจเนติกอัลกอริทึมและเบย์เซียน 62
5.15 แสดงการเปรียบเทียบค่า Precision ระหว่างเจเนติกอัลกอริทึมและเบย์เซียน 62

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เนื่องจากในปัจจุบัน ปริมาณของอีเมล์ขยะได้เพิ่มจำนวนขึ้นอย่างรวดเร็ว ซึ่งเป็นเหตุให้เกิดปัญหากับทั้งผู้ใช้งานอีเมล์และผู้ให้บริการอินเตอร์เน็ต (Internet Service Provider) เป็นอย่างมาก ทั้งในเรื่องการก่อให้เกิดความน่ารำคาญ ก่อให้เกิดการสูญเสียช่องทางการติดต่อสื่อสาร สร้างภาระหนักให้กับแม่ข่าย(Server) [1] อีกทั้งยังอาจนำมาซึ่งไวรัสและหนอนอินเตอร์เน็ตอีกด้วย[2] ด้วยเหตุนี้หน่วยงานต่างๆ จึงได้ตระหนักถึงปัญหาของอีเมล์ขยะและได้มีการคิดค้นวิธีการต่างๆ เพื่อใช้ป้องกันหรือกำจัดอีเมล์ขยะขึ้นมามากมาย หนึ่งในวิธีการกำจัดอีเมล์ขยะที่เป็นที่นิยมเป็นอย่างยิ่งได้แก่ การใช้เครื่องมือเรียนรู้ (Machine Learning) ซึ่งอาศัยวิธีการต่างๆ เช่น Naïve Bays, Support Vector Machine, Decision Tree หรือ Rule Learning เป็นต้น จากการศึกษาข้อมูลของวิธีการต่างๆ ข้างต้น เราพบว่าวิธีที่นิยมใช้ในการสร้างตัวกรองอีเมล์ขยะคือ Naïve Bayes ซึ่งเป็นวิธีการทางสถิติที่อาศัยหลักการของความน่าจะเป็นโดยมีสมมติฐานที่ว่า ความน่าจะเป็นของคำแต่ละคำเป็นอิสระจากกัน วิธีการนี้เป็นที่นิยมเนื่องจากมีข้อดีคือสามารถนำไปประยุกต์ใช้งานได้ง่ายและมีประสิทธิภาพสูง แต่ก็มีข้อเสียคือ มีลักษณะของการเรียนรู้ที่ตายตัว จึงจำเป็นต้องปรับปรุงการเรียนรู้อย่างสม่ำเสมอ ทำให้เสียเวลาและใช้ชุดข้อมูลฝึกหัดจำนวนมากรวมทั้งประสิทธิภาพการกรองน้อยลงหากไม่มีการปรับปรุงการเรียนรู้อย่างสม่ำเสมอ จากปัญหาดังกล่าวเราจึงได้ศึกษาวิธีการในการกรองอีเมล์ขยะด้วยทฤษฎีเจเนติกอัลกอริทึม ซึ่งสามารถใช้กระบวนการทางเจเนติก เช่น การคัดเลือก (Selection) การครอสโอเวอร์(Crossover) และการมิวเตชัน(Mutation) เพื่อสร้างรูปแบบของอีเมล์ที่หลากหลายขึ้นมา เมื่อได้รูปแบบของอีเมล์เพิ่มขึ้น ก็จะส่งผลให้การเรียนรู้มีมากขึ้นทำให้ได้ผลลัพธ์การกรองอีเมล์ขยะดีขึ้นด้วย อีกทั้งไม่จำเป็นต้องใช้ชุดข้อมูลจำนวนมากเพื่อนำมาให้ระบบเรียนรู้ ซึ่งทำให้ประหยัดเวลาการทำงาน ผู้ใช้สามารถกำจัดอีเมล์ขยะได้รวดเร็วยิ่งขึ้นและเสี่ยงต่อไวรัสและหนอนอินเตอร์เน็ตน้อยลง ในส่วนของผู้ให้บริการอินเตอร์เนตก็จะรักษาทรัพยากรไว้ได้มากขึ้น ส่งผลให้การติดต่อสื่อสารเป็นไปอย่างรวดเร็วซึ่งจะนำมาซึ่งประโยชน์ของผู้ใช้บริการอินเตอร์เน็ตเป็นอย่างยิ่ง

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

งานวิจัยนี้ ได้ศึกษาทำการศึกษาโดยมีจุดมุ่งหมายและวัตถุประสงค์สำคัญดังนี้

1.2.1. ศึกษาตัวกรองอีเมล์ขยะโดยเทคนิคที่นิยมใช้กันทั่วไปได้แก่ การประยุกต์ใช้ทฤษฎีเบย์เซียน

1.2.2. ศึกษาและพัฒนาตัวกรองอีเมล์ขยะ (Spam Mail Filter) โดยประยุกต์ใช้ทฤษฎีเจเนติกอัลกอริทึม

1.2.3. ทดสอบประสิทธิภาพของการกรองอีเมล์ขยะโดยการประยุกต์ใช้เจเนติกอัลกอริทึม รวมทั้งวิเคราะห์ผลการการกรองอีเมล์ขยะที่ได้จากการใช้พารามิเตอร์ต่างๆ ในกระบวนการทางเจเนติก

1.2.4. ศึกษาผลลัพธ์ของการกรองอีเมล์ขยะโดยประยุกต์ใช้เจเนติกอัลกอริทึมเปรียบเทียบกับการประยุกต์ใช้ทฤษฎีเบย์เซียน

1.3 สมมติฐานของการศึกษา

การพัฒนาตัวกรองอีเมล์ขยะในงานนี้เป็นการนำเจเนติกอัลกอริทึมที่มีตัวดำเนินการทางเจเนติกมาสร้างรูปแบบของอีเมล์ให้มีความหลากหลายเพิ่มขึ้นมา โดยแต่ละรูปแบบของอีเมล์ที่ถูกสร้างขึ้นมาใหม่นั้นจะถูกนำไปเป็นกฎที่ใช้สำหรับกรองอีเมล์ต่อไป ซึ่งจะทำให้การกรองมีประสิทธิภาพมากขึ้น โดยที่ไม่ต้องแก้ไขปรับเปลี่ยนพารามิเตอร์ต่างๆของระบบทั้งหมด หรือต้องคอยเสียเวลาปรับปรุงชุดข้อมูลฝึกหัดอย่างสม่ำเสมอ ดังเช่นงานวิจัยที่ผ่านมา

1.4 ทฤษฎีหรือแนวความคิดที่ใช้ในการวิจัย

สำหรับการกรองอีเมล์ขยะโดยใช้ทฤษฎีเจเนติกอัลกอริทึมนี้ จะต้องอาศัยหลักการและทฤษฎีดังต่อไปนี้

1.4.1 ทฤษฎีเจเนติกอัลกอริทึม (Genetic Algorithm)

1.4.2 ระบบแขนงการตัดสินใจ (Decision Tree)

1.4.3 ทฤษฎีเบย์เซียน (Bayesian Theorem)

1.5 ขอบเขตการวิจัย

งานวิจัยนี้มีจุดประสงค์เพื่อที่จะสร้างตัวกรองอีเมล์ขยะซึ่งใช้อีเมล์ขยะและอีเมล์ดีในการให้ระบบเรียนรู้ ซึ่งมีขอบเขตการวิจัย ดังนี้

1.5.1 เป็นตัวกรองอีเมล์ขยะ ซึ่งใช้ทั้งอีเมล์ขยะและอีเมล์เป็นชุดเรียนรู้

1.5.2 ฐานข้อมูลคำศัพท์ (Data Dictionary) ที่นำมาใช้ เป็นฐานข้อมูลที่สร้างมาจากคลังอีเมล์ขยะที่นิยมใช้โดยทั่วไปในการเรียนรู้เพื่อสร้างตัวกรองอีเมล์ขยะ[3],[4] โดยฐานข้อมูลนี้ครอบคลุมคำศัพท์ที่เกี่ยวข้องกับอีเมล์ขยะในปัจจุบัน

1.6 ขั้นตอนของการศึกษา

สำหรับขั้นตอนของการทำการศึกษาวิจัย สามารถแบ่งออกเป็นลำดับได้ดังนี้

1.6.1 ศึกษาค้นคว้าผลงานวิจัยและเอกสารทางวิชาการในหัวข้อที่เกี่ยวข้อง ที่มีผู้ทำวิจัยมาแล้ว

1.6.2 กำหนดหัวข้อ เป้าหมาย วัตถุประสงค์ และขอบเขตของการวิจัย

1.6.3 ศึกษาทฤษฎีและหลักการที่เกี่ยวข้องกับการวิจัย

1.6.4 วิเคราะห์และออกแบบตัวกรองอีเมล์ขยะโดยใช้เจเนติกอัลกอริทึม

1.6.5 พัฒนาโปรแกรม ทำการทดลองพร้อมทั้งบันทึกผลที่ได้จากการทดลองในแต่ละขั้นตอน

1.6.6 ปรับค่าพารามิเตอร์ให้เหมาะสมกับการทดลอง

1.6.7 วิเคราะห์ เปรียบเทียบผลการทดลองที่ได้ และสรุปผลการทดลอง

1.6.8 จัดทำเอกสารประกอบวิทยานิพนธ์

1.7 ข้อจำกัดของการศึกษา

1.7.1 อีเมล์ที่นำมาใช้ทดลองจะพิจารณาเฉพาะส่วนของเนื้อหา (Body) เท่านั้น

1.7.2 อีเมล์ที่นำมาพิจารณาต้องมีคำต่างๆ ปรากฏอยู่ในส่วนของเนื้อหา อีเมล์ที่มีรูปภาพหรือไฟล์แนบ (Attach File) จะไม่ถูกนำมาพิจารณา

1.7.3 คำศัพท์ที่ใช้บ่งบอกความเป็นอีเมล์ขยะทั้งหมดจะถูกจัดเก็บไว้ในฐานข้อมูลคำ โดยจะถูกจัดเป็น 8 กลุ่ม สำหรับคำอื่นๆ ที่นอกเหนือจากนี้จะไม่นำมาใช้ในการพิจารณา

1.7.4 การดำเนินการต่างๆ ของกระบวนการทางเจเนติกอัลกอริทึมทำงานภายในกลุ่มหรือระหว่างกลุ่มที่สอดคล้องกัน

1.8 คำจำกัดความที่ใช้ในการศึกษา

1.8.1 อีเมล์ขยะจะถูกแปลงให้อยู่ในรูปแบบของโครโมโซมซึ่งจะประกอบไปด้วยยีนทั้งหมด 8 ยีน โดยยีนแต่ละยีนจะแสดงค่าความน่าจะเป็นที่จะเป็นขยะตามที่พบในกลุ่มของยีนนั้นๆ โดยเฉลี่ย

1.8.2 โครโมโซมแม่แบบ (Template) หมายถึงชุดของโครโมโซมทั้งหมดที่ได้จากกระบวนการเจเนติกอัลกอริทึม ซึ่งจะถูกนำไปใช้ในการพิจารณาว่าอีเมล์ทดสอบว่าเป็นอีเมล์ประเภทใด

1.9 เครื่องมือและอุปกรณ์ที่ใช้ในงานวิจัย

เครื่องมือและอุปกรณ์ที่ใช้ในการวิจัยในครั้งนี้ ได้แก่

1.9.1 เครื่องคอมพิวเตอร์ที่ใช้หน่วยประมวลผลกลาง (CPU) Intel Celeron 2.0 GHz หน่วยความจำ (RAM) 640 MB) จำนวน 1 เครื่อง

1.9.2 ระบบปฏิบัติการ Windows XP Professional

1.9.3 โปรแกรม Microsoft Visual Basic.Net เวอร์ชัน 6.0

1.9.4 โปรแกรม Microsoft Excel

1.9.5 โปรแกรมที่ใช้ในการตัดคำและนับความถี่ของคำ GNU Awk เวอร์ชัน3.1.3 [6]

1.9.6 โปรแกรมที่ใช้ในการหารากศัพท์ของคำ Word Stemming [7]

1.10 โครงสร้างของวิทยานิพนธ์

วิทยานิพนธ์ฉบับนี้แบ่งออกเป็น 6 บท แต่ละบทประกอบด้วยเนื้อหาดังต่อไปนี้

บทที่ 1 กล่าวถึง ความเป็นมาและความสำคัญของปัญหา ความมุ่งหมาย และวัตถุประสงค์ของการศึกษา สมมติฐานของการศึกษา รวมทั้งทฤษฎีหรือแนวคิดที่ใช้ในการศึกษา ขอบเขตของการศึกษา ขั้นตอนของการศึกษา ข้อตกลงเบื้องต้น ข้อจำกัดของการศึกษา และคำจำกัดความที่ใช้ในการศึกษา

บทที่ 2 กล่าวถึงผลงานวิจัยที่เกี่ยวข้องกับกรองอีเมล์ขยะ ได้แก่ เทคนิคต่างๆ ในการกรองอีเมล์ขยะ และการกรองอีเมล์ขยะโดยใช้เครื่องมือเรียนรู้และวิธีการทางสถิติ

บทที่ 3 กล่าวถึงองค์ความรู้ที่เป็นพื้นฐานในงานวิจัยนี้ ซึ่งได้แก่ทฤษฎีต่างๆที่เกี่ยวข้องกับอีเมล์ องค์ประกอบของอีเมล์ ความรู้พื้นฐานเกี่ยวกับเจเนติกอัลกอริทึม ระบบแขนงการตัดสินใจและความรู้พื้นฐานเกี่ยวกับทฤษฎีเบย์เซียน

บทที่ 4 กล่าวถึง การพัฒนาตัวกรองอีเมล์ขยะโดยใช้เจเนติกอัลกอริทึม ซึ่งประกอบไปด้วย การเตรียมข้อมูล (Data Preparation) การสร้างฐานข้อมูลคำ (Creating Data Dictionary) และ การประยุกต์ใช้เจเนติกอัลกอริทึม (Adopting Genetic Algorithm)

บทที่ 5 กล่าวถึง การทดลองและผลการทดลองการกรองอีเมล์ขยะโดยใช้เจเนติกอัลกอริทึม โดยใช้พารามิเตอร์ที่แตกต่างกัน การปรับหาพารามิเตอร์ที่เหมาะสม และ การกรองอีเมล์ขยะโดยใช้ทฤษฎีเบย์เซียนเพื่อการเปรียบเทียบผลการทดลองและประสิทธิภาพกับวิธีการที่นำเสนอ จากนั้นวิเคราะห์ผลที่ได้จากการศึกษาวิจัยการกรองอีเมล์ขยะโดยใช้เจเนติกอัลกอริทึม

บทที่ 6 กล่าวถึง บทสรุปและบทวิจารณ์ รวมถึงข้อเสนอแนะ และแนวทางในการพัฒนาตัวกรองอีเมล์ขยะต่อไปในอนาคต

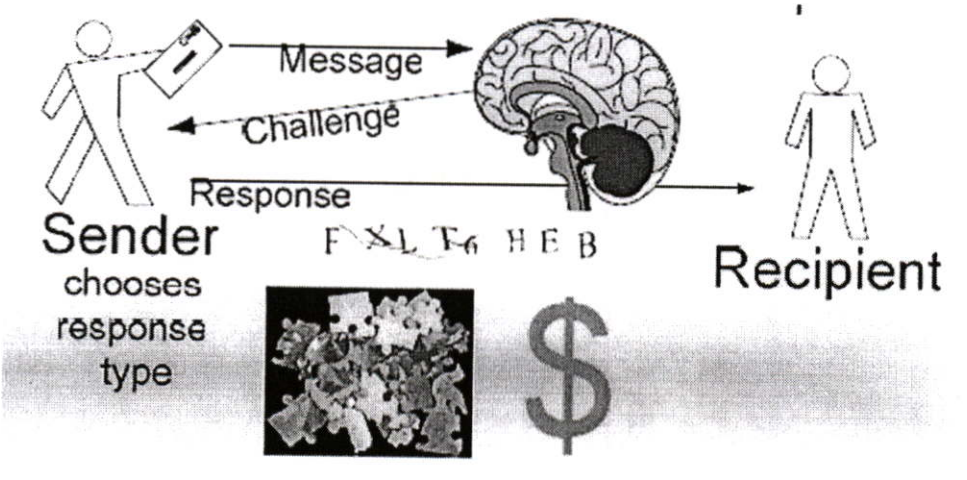
บทที่ 2
งานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการกรองอีเมล์ขยะ ซึ่งจะเริ่มด้วยเทคนิคต่างๆที่ใช้ในการกรองอีเมล์ขยะ และในส่วนของวิธีการในการกรองอีเมล์ขยะซึ่งเป็นที่นิยมใช้กันอย่างแพร่หลายซึ่งได้แก่การกรองอีเมล์ขยะโดยใช้เครื่องมือเรียนรู้และวิธีการทางสถิติ เช่นตัวกรองเบย์เซียนและตัวกรองที่ใช้การแบ่งแยกมาร์คอฟเวียน

2.1 เทคนิคต่างๆในการกรองอีเมล์ขยะ(Spam Filtering Techniques)

2.1.1 เทคนิคการกรองอีเมล์ขยะโดยใช้การกระตุ้นและตอบสนอง (Challenge and Respond)

การกระตุ้น (Challenge) คือการถามคำถามซึ่งผู้ส่งอีเมล์ต้องตอบอย่างถูกต้องก่อนที่อีเมล์จะถูกส่งไป โดยกรณีทั่วๆไปคือผู้ส่งจะต้องตอบคำถามหรือมีการตอบสนอง (Respond) ที่ถูกต้อง ระบบจึงจะทำการบันทึกผู้ใช้งานนี้ลงในไวท์ลิสต์ (White List) ถ้าตอบสนองไม่ถูกต้องก็จะทำการเก็บผู้ใช้งานนี้ไว้ในแบล็กลิสต์ (Black List) โดยที่การกระตุ้นจะต้องง่ายสำหรับคน แต่ยากสำหรับเครื่องจักรในการตอบสนองดังแสดงในรูปที่ 2.1 [1], [2]



รูปที่ 2.1 แสดงลำดับขั้นในกระบวนการกระตุ้น-ตอบสนองและชนิดของการตอบสนองที่ต้องการ

2.1.2 เทคนิคการกรองอีเมล์ขยะโดยใช้การวิเคราะห์ส่วนหัวของอีเมล์ (Header Analysis)

ส่วนหัวเรื่องของอีเมล์มักประกอบไปด้วย ชื่อเรื่อง วันที่ เวลา ที่อยู่อีเมล์(E-mail Address) ของทั้งผู้ส่งและผู้รับ โดยในส่วนของหัวเรื่องนี้สิ่งเดียวที่ไม่สามารถปลอมแปลงได้คือส่วนที่บอกว่าผู้รับได้รับอีเมล์นี้มาจากใคร จากการวิเคราะห์หัวเรื่องพบว่ามีข้อสังเกตหลายอย่างที่ใช้บ่งชี้ได้เป็นอย่าง

ดีว่าอีเมล์เป็นอีเมล์ขยะหรือไม่ ยกตัวอย่างเช่น ไอพีซึ่งไม่ตรงกับโดเมน (Domain Name) เป็นตัวบ่งชี้ว่าอีเมล์เกิดจากการปลอมแปลง ดังนั้นส่วนหัวเรื่องของอีเมล์มักถูกนำมาใช้ในการสร้างกฎสำหรับการเรียนรู้ข่าวขยะของตัวกรอง [1], [2]

2.1.3 เทคนิคการกรองอีเมล์ขยะโดยใช้ลายเซ็นดิจิตอล (Digital Signature)

ลายเซ็นดิจิตอลในอีเมล์นั้นเปรียบเสมือนกับลายพิมพ์นิ้วมือของมนุษย์ ดังนั้นอีเมล์แต่ละฉบับจะมีลายเซ็นดิจิตอลที่แตกต่างกัน ดังนั้นถ้าต้องการทราบว่าอีเมล์ใดเป็นอีเมล์ขยะ เราสามารถนำลายเซ็นดิจิตอลของอีเมล์นั้นมาเปรียบเทียบกับฐานข้อมูลลายเซ็นดิจิตอลที่ได้ถูกรวบรวมไว้ ถ้าพบว่าลายเซ็นดิจิตอลนั้นตรงกันก็สรุปได้ว่าอีเมล์นั้นๆ เป็นอีเมล์ขยะ แต่วิธีการนี้มีข้อเสียคือเมื่อเปลี่ยนแปลงอีเมล์เพียงเล็กน้อยลายเซ็นดิจิตอลของอีเมล์ก็จะเปลี่ยนไป ดังนั้น สแปมเมอร์ (Spammer) มักจะเพิ่มสุ่มข้อความสุ่มลงไปในอีเมล์เพียงเล็กน้อยเพื่อให้ลายเซ็นดิจิตอลของอีเมล์เปลี่ยนไป [1]

2.1.4 เทคนิคการกรองอีเมล์ขยะโดยใช้บัญชีที่อยู่ (Address Lists)

บัญชีที่อยู่มี 2 ประเภทด้วยกันคือ บัญชีขาว (White List) ซึ่งคือบัญชีที่อยู่อีเมล์ของผู้ใช้ที่ต้องการติดต่อด้วย และบัญชีดำ (Black List) คือบัญชีที่อยู่อีเมล์ของสแปมเมอร์หรือบุคคลที่ไม่ต้องการติดต่อด้วย บัญชีที่อยู่ไม่ใช่เพียงแต่ประกอบไปด้วยที่อยู่อีเมล์เท่านั้นแต่ยังสามารถรวมชื่อเมนและส่วนขยายของชื่อโดเมนเช่น ".org" หรือ "edu" ได้ ถ้าหากพิจารณาทางฝั่งแม่ข่าย บัญชีดำประกอบไปด้วยที่อยู่อีเมล์ของสแปมเมอร์ที่รู้จักกันดี บัญชีเหล่านั้นจะถูกเรียกว่าบัญชีหลุมดำเรียลไทม์ (Real-Time Blackhole List) ส่วนระบบเครือข่ายซึ่งอนุญาตให้สแปมเมอร์ใช้ระบบในการส่งอีเมล์ขยะหรือไม่มีการป้องกันสแปมเมอร์ที่เข้ามาใช้ระบบในทางที่ผิดก็จะถูกจัดไว้ในกลุ่มบัญชีหลุมดำเรียลไทม์ด้วยเช่นกัน เมื่อมีการใช้บัญชีที่อยู่ จะต้องแน่ใจว่าไม่ไม่มีการเพิ่มที่อยู่ที่ไม่แน่ใจว่าถูกต้องลงไป เพราะสแปมเมอร์มักปลอมแปลงหัวเรื่องเพื่อให้ดูคล้ายกับอีเมล์ที่ส่งมาจากที่อื่นๆ การใช้ที่อยู่ไอพีจึงปลอดภัยกว่าเนื่องจากยากต่อการปลอมแปลง[1]

2.1.5 เทคนิคการกรองอีเมล์ขยะโดยใช้รายการคำสำคัญ (Keyword Lists)

บัญชีคีย์เวิร์ดมี 2 ประเภทด้วยกันคือ บัญชีคีย์เวิร์ดขาว (Keyword Whitelists) ซึ่งคือบัญชีของคีย์เวิร์ดที่บ่งชี้ว่าไม่น่าจะเป็นอีเมล์ขยะ และบัญชีคีย์เวิร์ดดำ (Keyword Blacklists) ซึ่งคือบัญชีของคีย์เวิร์ดที่บ่งชี้ว่าน่าจะเป็นอีเมล์ขยะ คีย์เวิร์ดนั้นพบได้จากชื่อเรื่อง หัวเรื่อง หรือเนื้อความของอีเมล์ โดยคีย์เวิร์ดอาจประกอบไปด้วย รูปแบบที่เกิดจากการผสมของคำที่หลากหลายหรือรหัสต่างๆ บัญชีคีย์เวิร์ดนั้นต้องการการสร้างอย่างถูกต้องและทันต่อข่าวสารเสมอ ซึ่งเป็นเรื่องที่ต้องใช้เวลาและความพยายามเป็นอย่างสูง [1]

2.1.6 เทคนิคการกรองอีเมล์ขยะโดยการพิจารณาโอเพนรีเลย์ (Open Relays)

โอเพนรีเลย์คือ โอเพนรีเลย์เซิร์ฟเวอร์ หรืออีเมล์เซิร์ฟเวอร์ที่อนุญาตให้กลุ่มที่สาม (Third Party) ส่งผ่านข้อความ เช่น การรับหรือการส่งอีเมล์ซึ่งไม่ใช่ส่งถึงหรือมาจากผู้ใช้งานภายใน โอเพนรีเลย์สร้างขึ้นเพื่อให้มีความสามารถในการใช้งานกับผู้ใช้เคลื่อนที่ (Mobile Users) เพื่อติดต่อกับเครือข่ายโดยต้องผ่านผู้ให้บริการอินเตอร์เนตก่อน (Internet Service Provider) ซึ่งจะส่งต่อข้อความไปยังที่อยู่ของผู้ให้บริการอินเตอร์เนต จากนั้นจึงส่งต่อข้อความไปยังจุดหมายปลายทาง เทคโนโลยีนี้ไม่ดีนักเนื่องจากสแปมเมอร์ผู้ใช้ช่องทางนี้ในการส่งอีเมล์ได้เพิ่มจำนวนขึ้นอย่างรวดเร็ว[8] การบล็อกโดยวิธีโอเพนรีเลย์นั้นมีทั้งข้อดีและข้อเสีย ข้อดีคือลดจำนวนของอีเมล์ขยะลงได้มาก ส่วนข้อเสียคืออีเมล์ดีที่ส่งมาจากผู้ให้บริการอินเตอร์เนตก็อาจถูกบล็อกไปด้วยหรือเมล์เซิร์ฟเวอร์ที่ใช้ร่วมกันก็อาจไม่มีความปลอดภัยไปด้วย แต่ในปัจจุบันระบบความรับผิดชอบของผู้ดูแลระบบนั้นรวดเร็วมากต่อการแก้ไขการปรับแต่งที่ผิดพลาด(Misconfiguration) ซึ่งอาจจะทิ้งเซิร์ฟเวอร์ที่ถูกเปิดเผยออกมาเป็นโอเพนรีเลย์ดังนั้นปริมาณของอีเมล์ดีที่ถูกบล็อกควรจะลดน้อยลงด้วย[9]

2.1.7 เทคนิคการกรองอีเมล์ขยะโดยการพิจารณาโอเพนพร็อกซี่ (Open Proxy)

โอเพนพร็อกซี่คือพร็อกซี่เซิร์ฟเวอร์ซึ่งสามารถเข้าถึงได้โดยผู้ใช้งานอินเตอร์เนตโดยทั่วไป แล้วพร็อกซี่เซิร์ฟเวอร์จะอนุญาตให้ผู้ใช้รับและส่งเว็บเซอร์วิส (Web Service) เช่น DNS อีเมล์ เว็บเพจ ภายในกลุ่มเครือข่ายเท่านั้น แต่สำหรับโอเพนพร็อกซี่ ผู้ใช้ภายนอกได้รับมอบอำนาจให้เข้ามาในระบบเพื่อวัตถุประสงค์ที่จะทำกิจกรรมใดก็ตามที่ส่งผลเสียแก่ระบบได้ [10] ซึ่งการบล็อกโดยวิธีโอเพนพร็อกซี่นั้นมีข้อดีและข้อเสียคล้ายคลึงกับการกรองโดยใช้วิธีโอเพนรีเลย์ [9]

2.2 การกรองอีเมล์ขยะโดยใช้เครื่องมือเรียนรู้และวิธีการทางสถิติ (Machine Learning and Statistical Method)

2.2.1 การกรองอีเมล์ขยะโดยใช้ตัวกรองเบย์เซียน (Bayesian Filter)

เป็นวิธีการทางสถิติที่นักวิจัยใช้กันมากในการสร้างตัวกรองอีเมล์ขยะ ในปี 1998 Sahami และคณะ [11] ได้เขียนบทความซึ่งเป็นบทความแรกที่ใช้เครื่องมือเรียนรู้ในการต่อสู้กับอีเมล์ขยะโดยใช้เบย์อย่างง่าย (Naïve Bayes) โดยพิจารณา คำ วลี และโดเมนเป็นหลักโดยผลที่ได้นั้น ถ้าไม่มีการพิจารณาโดเมนจะต่ำกว่าพิจารณาโดเมนร่วมด้วย ต่อมาในปี 2002 Paul Graham [12] ซึ่งได้รับการกล่าวถึงอย่างมากในชุมชนโอเพนซอร์ส ได้ทำการปรับเปลี่ยนเวอร์ชันของเบย์อย่างง่ายให้แกร่งขึ้นโดยให้โทเคนในเมล์ดีได้รับคะแนนเพิ่มขึ้นเป็นสองเท่า ในเวลาทดสอบจะใช้โทเคน 15 โทเคนที่มีค่าความน่าจะเป็นสูงที่สุด มีการออกแบบและปรับหลายๆ อย่างเพื่อหลีกเลี่ยงการทายอีเมล์ดีให้เป็นอีเมล์ขยะ(False Positive)

ส่วนที่นำมาพิจารณาได้แก่ คำในส่วนของหัวเรื่องและเนื้อเรื่อง ซึ่งจากการทดลองนี้พบว่าได้ผลที่ดีกว่า[11] เล็กน้อย

2.2.2 การกรองอีเมล์ขยะโดยใช้การแบ่งแยกมาร์คอฟเวียน (Markovian Discrimination)

เนื่องจากปัญหาหลักของการกรองโดยใช้เบย์อย่างง่าย (Naïve Bays) จะสนใจนัยสำคัญของคำแต่ละคำ โดยมองว่าแต่ละคำนั้นเป็นอิสระจากกันโดยไม่พิจารณาคำอื่นๆรอบๆตัวมันเลย ดังนั้นมันอาจนำมาซึ่งการถูกโจมตีได้โดยง่าย เพียงแค่การพยายามสุ่มเพิ่มคำดีๆ (Good Word) เข้าไปเพื่อลดคะแนนในส่วนของคำที่น่าจะเป็นขยะ (Spam Word) การแบ่งแยกของมาร์คอฟเวียนจะพิจารณากลุ่มของคำที่พบในอีเมล์ขยะ ซึ่งระดับความรุนแรงของคำที่น่าจะเป็นขยะขึ้นอยู่กับระยะทางของคำในกลุ่มนั้นเทียบกับคำที่พบว่าเป็นขยะจริงๆ W. S. Yerazuris [13] ได้ทดลองโดยเปรียบเทียบวิธีการกรองโดยใช้เบย์อย่างง่ายและกรองโดยใช้การแบ่งแยกของมาร์คอฟเวียน พบว่าผลที่ดีขึ้นโดยมีความผิดพลาดน้อยกว่าการกรองโดยใช้เบย์อย่างง่าย แต่ในแง่ของนัยสำคัญแล้ว วิธีเบย์อย่างง่ายมีนัยสำคัญน้อยมาก

บทที่ 3
ความรู้พื้นฐานที่เกี่ยวข้องกับงานวิจัย

ในบทนี้จะกล่าวถึงองค์ความรู้ที่เป็นพื้นฐานในงานวิจัยนี้ ซึ่งได้แก่ ทฤษฎีที่เกี่ยวข้องกับอีเมล์ องค์ประกอบของอีเมล์ ความรู้พื้นฐานเกี่ยวกับทฤษฎีเบย์เซียน (Bayesian Theorem) ความรู้พื้นฐานเกี่ยวกับแขนงการตัดสินใจ (Decision Tree) และความรู้พื้นฐานเกี่ยวกับกระบวนการเจเนติกอัลกอริทึม (Genetic Algorithm)

3.1 ทฤษฎีที่เกี่ยวข้องกับอีเมล์

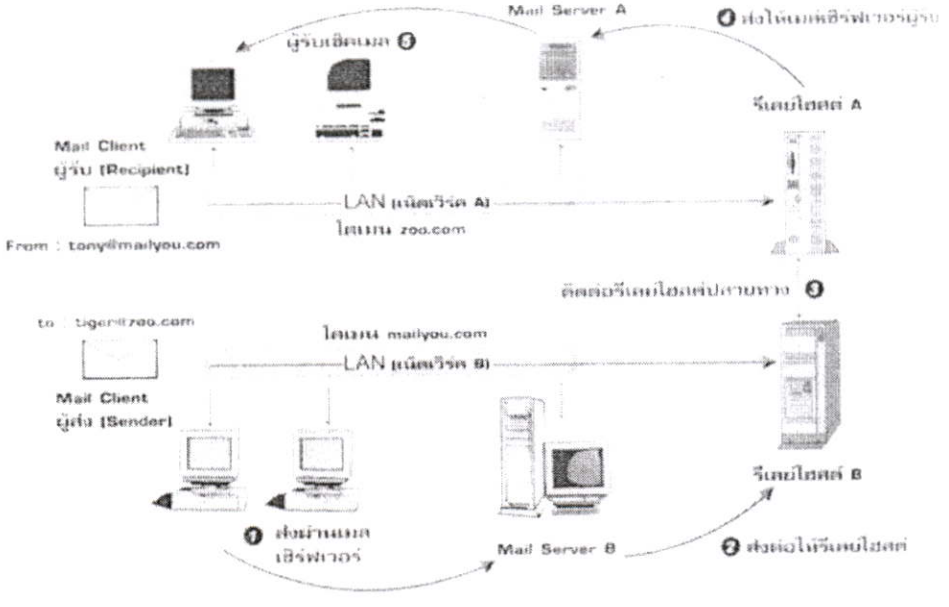
3.1.1 ความเป็นมาของอีเมล์

บทความ[14] ได้กล่าวว่า อีเมล์หรือจดหมายอิเล็กทรอนิกส์ ได้มีการใช้งานมานานแล้ว ตั้งแต่ยุคของเครื่องเมนเฟรมหรือมินิคอมพิวเตอร์ ซึ่งไอบีเอ็มได้พัฒนาระบบที่เรียกว่า PROFS (Professional Office System) ออกมาใช้งาน นอกจากนี้ยังมีบนระบบ UNIX อีกด้วย ต่อมาหลายค่ายก็ได้พัฒนาระบบอีเมล์ของตนขึ้นมา โดยส่วนใหญ่จะเป็นองค์ประกอบในแอพพลิเคชั่นที่ทำงานบนระบบเครือข่าย เช่น ไมโครซอฟต์เมล์(Microsoft Mail) ของไมโครซอฟต์, ซีซีเมล์ (CC Mail) ของโลตัส ซึ่งต่างก็ได้ใช้เทคโนโลยีของตนเองและเป็นระบบปิด ดังนั้นการส่งเมล์ไปยังผู้ใช้ที่มีระบบเมล์คนละค่ายกันจึงเป็นเรื่องที่ยุ่งยาก ในยุคต่อมาได้มีระบบเครือข่าย แลน (LAN) และแวน (WAN) ซึ่งต่างมีมาตรฐานและเป็นระบบเปิดมากขึ้น จึงมีการปรับเปลี่ยนการทำงานของระบบเมล์มาเป็นแบบไคลแอนท์-เซิร์ฟเวอร์ (Client-Server) ที่เป็นพื้นฐานที่ใช้กันในระบบยูนิกซ์ (Unix) และมีการพัฒนาอีเมล์เซิร์ฟเวอร์ (Email Server) ทั้งในรูปแบบที่เป็นการใช้งานผ่านระบบ LAN หรือ ใช้งานผ่าน modem ที่เชื่อมต่ออยู่กับระบบ WAN ทำให้ผู้ใช้ไม่สามารถมองเห็นและเข้าถึงไฟล์ในฮาร์ดดิสก์บนเครื่องเซิร์ฟเวอร์ได้ดังนั้นความปลอดภัยของระบบจึงมีมากขึ้น จนในปัจจุบัน ได้พัฒนาขึ้นมาเป็นระบบเวิร์คโฟล์ว (Workflow) ที่ใช้อีเมล์เป็นพื้นฐาน

3.1.2 ลักษณะการทำงานของระบบรับ-ส่งอีเมล์

การรับ-ส่งอีเมล์จะมีลักษณะดังรูปที่ 3.1 โดยเริ่มจากผู้ส่ง(Sender) ทำการสร้าง หรือ เขียนอีเมล์ขึ้นมาตามวัตถุประสงค์ที่ต้องการ จากนั้นทำการกดปุ่มส่งโดยผ่านโปรโตคอล(Protocol) ส่งไปยังเครื่องเมล์เซิร์ฟเวอร์ต้นทาง ดังขั้นตอนที่ 1จากนั้นเมล์เซิร์ฟเวอร์จะส่งไปยังเครื่องที่เป็นรีเลย์โฮสต์ (Relay Host) ต้นทาง เนื่องจากรีเลย์โฮสต์เป็นเครื่องที่สามารถติดต่อกับระบบเครือข่ายภายนอกได้ ดังขั้นตอนที่ 2 (โดยทั่วไปเมล์เซิร์ฟเวอร์อาจทำหน้าที่เป็นรีเลย์โฮสต์ในเครื่องเดียวกันก็ได้ ถ้าเป็นเช่นนี้ก็จะไม่มีขั้นตอนที่ 2 เกิดขึ้น) จากรีเลย์โฮสต์ต้นทางเมื่อได้รับเมล์มาแล้ว จะติดต่อ

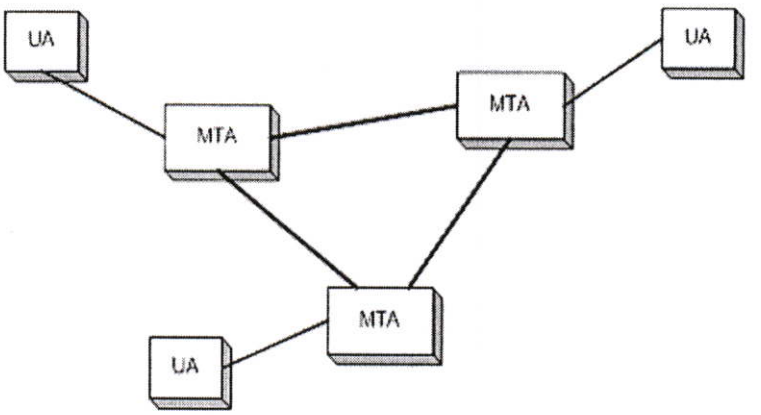
กับรีเลย์โฮสต์ปลายทางเพื่อส่งเมลฉบับนี้ไปดังขั้นตอนที่ 3 และในกรณีเดียวกัน ถ้าเครื่องรีเลย์โฮสต์ปลายทางกับเครื่องเมล์เซิร์ฟเวอร์ปลายทางเป็นเครื่องดียวกันจะไม่มีขั้นตอนที่ 4 เมื่อเมล์ไปถึงเมล์เซิร์ฟเวอร์ปลายทางเรียบร้อยแล้วนั่นคือจบกระบวนการส่งเมล์ เมื่อผู้รับเช็คเมล์ ไม่ว่าจะใช้วิธีไหนก็ตามก็จะต้องติดต่อกับเครื่องเมล์เซิร์ฟเวอร์ของตนเอง เพื่อนำเมล์ฉบับนั้นมาอ่าน ในที่นี้เครื่องเมล์เซิร์ฟเวอร์หรือเครื่องรีเลย์โฮสต์ก็จะทำหน้าที่เหมือนกับที่ทำการไปรษณีย์นั่นเอง



รูปที่ 3.1 แสดงลักษณะการทำงานของระบบอีเมล์

3.1.3 สถาปัตยกรรมของระบบเมล์

ทีซีพี/ไอพี (TCP/IP) มีโปรโตคอลที่สนับสนุนการรับ- ส่งอีเมล์หลายโปรโตคอล แต่โปรโตคอลที่นิยมใช้ในอินเตอร์เน็ตคือ SMTP (Simple Mail Transport Protocol) หน้าที่ของ SMTP คือการกำหนดกรรมวิธีและแบบแผนการนำส่งข้อความระหว่างผู้รับและผู้ส่ง โดย SMTP อาศัยทีซีพีเพื่อลำเลียงจดหมายผ่านพอร์ต 25 ระบบเมล์ที่ใช้ในทีซีพี/ไอพี มีองค์ประกอบสองส่วนคือตัวแทนผู้ใช้ (User Agent :UA) หรืออาจจะเรียกว่าตัวแทนผู้ใช้เมล์ (Mail User Agent : MUA) และตัวแทนขนส่งเมล์ (MTA :Mail Transfer Agent) ทั้งตัวแทนผู้ใช้และตัวแทนขนส่งเมล์เป็นชื่อที่นำมาจากระบบ X.400 ซึ่งเป็นมาตรฐานที่นานาชาติกำหนดไว้เพื่อการนำส่งอีเมล์



รูปที่ 3.2 แสดงสถาปัตยกรรมในทีซีพี/ไอพี

จากรูปที่ 3.2 ตัวแทนผู้ใช้ทำหน้าที่ในการติดต่อกับผู้ใช้เพื่อรับและส่งอีเมล์ ซึ่งรูปแบบของการติดต่อมี 3 แบบ ดังนี้

1. การติดต่อโดยตรงหรือประมวลผลบนเครื่องที่เก็บเมล์บ๊อกซ์ (Mailbox) อยู่เลย ซึ่งโปรแกรมที่ใช้ในการรับส่งเมล์ที่นิยมกันบน Linux/Unix ก็เช่น /bin/mail, mailx, pine, elm เป็นต้น โดยการใช้งานจริงอาจจะด้วยการเทลเนต (Telnet) จากเครื่องคอมพิวเตอร์ส่วนบุคคลเข้าไปยังเครื่องที่เป็นเมล์เซิร์ฟเวอร์แล้วใช้งานโปรแกรมดังกล่าวบนเมล์เซิร์ฟเวอร์

2. การทำงานแบบไคล์แอนท์-เซิร์ฟเวอร์โดยเครื่องที่เป็นเมล์ไคล์แอนท์จะติดต่อกับเครื่องเมล์เซิร์ฟเวอร์โดยผ่านโปรโตคอลสำหรับการจัดการโดยเฉพาะ เช่น POP3 (Post Office Protocol version 3) หรือ IMAP 4 (Internet Mail Access Protocol version 4) เพื่อให้ดึงเมล์จากเมล์บ๊อกซ์บนเซิร์ฟเวอร์ไปอ่านได้โดยสะดวกซึ่งโปรแกรมที่นิยมใช้งานเป็นเมล์ไคล์แอนท์ ได้แก่ Microsoft Outlook, Outlook Express, Endora, Netcape Mail เป็นต้น

3. การทำงานแบบเว็บเมล์ (Web Mail) เป็นการติดต่อระหว่างเว็บเซิร์ฟเวอร์ (Web Server) ที่มีโปรแกรมเว็บเมล์ติดตั้งอยู่กับเมล์เซิร์ฟเวอร์ผ่านโปรโตคอล ที่นิยมใช้กันส่วนใหญ่จะเป็น IMAP ซึ่งเว็บเซิร์ฟเวอร์กับเมล์เซิร์ฟเวอร์อาจจะเป็นเซิร์ฟเวอร์ตัวเดียวกันหรือคนละตัวกันก็ได้ โดยโปรแกรมที่เป็นเว็บเมล์ก็เช่นโปรแกรมที่ติดตั้งอยู่บนเว็บเซิร์ฟเวอร์ของ yahoo.com , hotmail.com เป็นต้น หรือถ้าเป็นโปรแกรมแบบฟรีก็เช่น Horde mail (www.horde.org), OpenWebmail (www.openwebmail.org) , SquirrelMail (www.squirrelmail.org) เป็นต้น โปรแกรมที่เป็นตัวแทนผู้ใช้ในแบบนี้ก็จะหมายถึงเบราเซอร์ (Browser) ที่รันอยู่บนเครื่องคอมพิวเตอร์ส่วนบุคค ที่ใช้ติดติดต่อไปยังเมล์เซิร์ฟเวอร์ผ่านเว็บเซิร์ฟเวอร์เพื่อดำเนินการในส่วนของการรับและส่งเมล์ ตัวแทนผู้ใช้แบบนี้จะต่างกับแบบที่ 2 คือไม่ต้องมีการใช้โปรโตคอล POP และ IMAP เพราะจะมีตัวกลางที่เป็นเว็บเซิร์ฟเวอร์เป็นตัวใช้งานโปรโตคอลดังกล่าวแทน อาจจะ

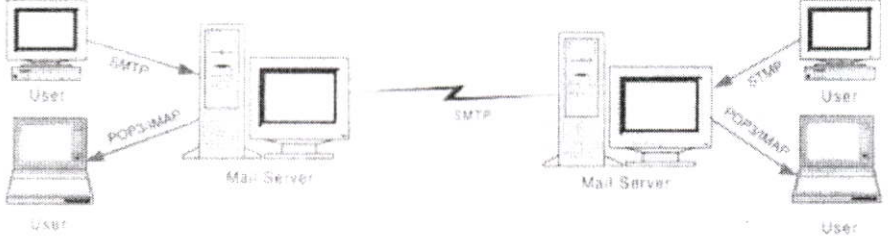
กล่าวได้ว่าตัวแทนผู้ใช้เป็นโปรแกรมอำนวยความสะดวกให้ผู้ใช้เขียน แก้ไข และส่งเมล์ รวมทั้งการเปิดอ่านเมล์ที่ได้รับ และจัดเก็บเมล์เพื่อนำมาใช้ภายหลัง

ตัวแทนขนส่งเมล์ คือส่วนที่ทำหน้าที่ในการรับและส่งอีเมล์ โดยจะรับจากตัวแทนผู้ใช้แล้วตรวจสอบว่าผู้รับปลายทางอยู่ในเครื่องเดียวกันหรือไม่ หากอยู่ในเครื่องเดียวกันก็จะส่งเมล์นั้นไว้ในเมล์บ๊อกซ์หรือโฟลเดอร์ที่เก็บเมล์ของผู้รับนั้น แต่หากอยู่คนละเครื่องจะส่งให้กับอีกกระบวนการหนึ่งเพื่อส่งต่อไปยังเครื่องนั้น ๆ ได้ต่อไป (กระบวนการที่ทำหน้าที่รับส่งเมล์ข้ามเครื่องนั้นอาจเป็น smtpd ที่ทำหน้าที่คอยแปลงเมล์ให้อยู่ในรูปของโปรโตคอล SMTP เพื่อให้สามารถส่งผ่านเครือข่ายทีซีพี/ไอพีได้) ในขณะเดียวกันก็ทำหน้าที่รับเมล์ที่ส่งเข้ามายังผู้รับในเครื่องนั้น แล้วทำการจัดส่งให้ผู้รับแต่ละคนอย่างถูกต้องด้วย ในส่วนนี้โปรแกรมที่นิยมกันก็เช่น Sendmail, Microsoft Mail, Microsoft Exchange

การจัดแบ่งออกเป็นตัวแทนผู้ใช้และตัวแทนขนส่งเมล์มีข้อดีคือ แยกงานของทั้งสองส่วนให้เป็นอิสระจากกัน หน้าที่ของตัวแทนผู้ใช้เน้นการทำงานกับผู้ใช้เพื่อให้ผู้ใช้อ่านเขียนเมล์ได้อย่างสะดวกโดยไม่ต้องยุ่งเกี่ยวกับการทำงานระดับล่างของโปรโตคอล ตัวแทนขนส่งเมล์ทำงานตาม SMTP เช่นการตรวจสอบความถูกต้องของที่อยู่ผู้รับผู้ส่ง รวมทั้งการหาเส้นทางและนำส่งเมล์ไปยังปลายทาง

3.1.4 โปรโตคอล

การที่เครื่องคอมพิวเตอร์ 2 เครื่องจะรับส่งเมล์กันได้ หรือผู้ใช้จะโหลดเมล์ไปอ่านที่เครื่องของตนเองนั้น จำเป็นต้องมีโปรโตคอลที่ใช้คุยกันระหว่างเครื่องทั้งสองคือ SMTP POP3 หรือ IMAP ดังรูปที่ 3.3



รูปที่ 3.3 แสดงโปรโตคอลที่ใช้ในการรับส่งเมล์

3.1.4.1 SMTP

SMTP หรือ Simple Mail Transfer Protocol เป็นโปรโตคอลที่ติดต่อกันระหว่างเครื่องที่เป็นโฮสต์กับโฮสต์ โดยโฮสต์ในที่นี้ทำหน้าที่เป็นเมล์เซิร์ฟเวอร์หรือผู้ให้บริการอีเมล์ ซึ่งจะมีกระบวนการที่ทำหน้าที่เป็นตัวแทนขนส่งเมล์ ทำงานอยู่บนทั้ง 2 ด้าน และรับส่งข้อมูลระหว่างกัน

โดยใช้ SMTP เมื่อได้รับเมล์มาแล้วก็จะเก็บเมล์เหล่านั้นไว้ในไดเร็คทอรี่ที่เป็นกล่องหรือตู้ไปรษณีย์ในเครื่องนั้น และรอจนกว่าผู้ใช้มาเปิดอ่าน ซึ่งมีได้ 3 วิธีด้วยกันคือ

- ผู้ใช้มีบัญชี (Account) บนเครื่องเมล์เซิร์ฟเวอร์ก็สามารถเปิดอ่านได้โดยใช้คำสั่งต่าง ๆ ของ Linux/Unix เช่น mail, pine และเมล์ที่ถูกอ่านจะถูกย้ายไปเก็บไว้ในเมล์บ๊อกซ์ของผู้ใช้แทนเมล์บ๊อกซ์ของระบบได้
- ผู้ใช้อยู่บนเครื่องไคล์แอนท์จะต้องโหลดเมล์ไปไว้ในเครื่องของตัวเองก่อนแล้วจึงเปิดอ่านได้
- ผู้ใช้รับส่งเมล์ผ่านตัวกลางที่เป็นเว็บเซิร์ฟเวอร์ซึ่งเมล์จะยังคงถูกเก็บไว้ที่เครื่องเมล์เซิร์ฟเวอร์

การทำงานของ SMTP จะทำหน้าที่ในการกำหนดว่าตัวแทนขนส่งเมล์แต่ละตัวจะติดต่อกันได้อย่างไรผ่านทางทีซีพี/ไอพี เมล์ที่ส่งไปนั้นอาจจะส่งตรงไปยังตัวแทนขนส่งเมล์ปลายทางเลย หรือว่าผ่านตัวแทนขนส่งเมล์หลายเครื่อง (หมายถึงผ่านรีเลย์โฮสต์หลายเครื่อง) โดยผ่านกระบวนการเก็บและส่งต่อ (Store and Forward) ก็ได้เช่นกัน

โปรโตคอล SMTP จะไม่สนใจว่าข้อความในเมล์เป็นอะไร แต่จำกัดว่า SMTP สามารถส่งได้แต่ข้อมูลที่เป็นข้อความแอสกี (ASCII) เท่านั้น ไม่สามารถส่งไฟล์ที่เป็นเพลง, หนัง, รูปภาพหรืออื่น ๆ ได้ ซึ่งถ้าเราต้องการส่งไฟล์เหล่านั้นผ่านทาง SMTP จะต้องแปลงไฟล์เหล่านั้นให้อยู่ในรูปของข้อความเสียก่อนและเมื่อส่งไปถึงปลายทางแล้วค่อยทำการแปลงกลับอีกที

นอกจากการใช้ SMTP เพื่อรับส่งเมล์ระหว่างเมล์เซิร์ฟเวอร์ด้วยกันแล้ว ยังใช้ในขณะที่เป็นไคล์แอนท์ส่งเมล์ไปยังเครื่องที่เป็นเมล์เซิร์ฟเวอร์ด้วย

3.1.4.2 POP

POP หรือ Post Office Protocol คือโปรโตคอลที่ออกแบบมาให้ใช้สำหรับการรับเมล์จากเครื่องที่เป็นเมล์เซิร์ฟเวอร์มายังเครื่องของผู้ใช้ โดยทางฝั่งเซิร์ฟเวอร์จะมีกระบวนการที่เป็น POP เซิร์ฟเวอร์ขณะที่ทางฝั่งผู้ใช้มี POP ไคล์แอนท์ซึ่งในบางโปรแกรมที่ผู้ใช้อ่านและเขียนเมล์นั้นจะมี POP ไคล์แอนท์ฝังอยู่ในตัวอยู่แล้วไม่ได้แยกออกมาเป็นโปรแกรมหนึ่ง เมื่อผู้ใช้เชื่อมต่อไปที่ POP เซิร์ฟเวอร์อีเมล์ที่อยู่บนเมล์เซิร์ฟเวอร์จะถูกส่งมาเก็บไว้ในเครื่องของผู้ใช้เลย ดังนั้นเมื่อผู้ใช้จัดการกับเมล์ เช่น ลบเมล์หรือส่งต่อเมล์ก็จะทำกับเมล์ที่อยู่บนเครื่องของผู้ใช้เอง ส่วนเมล์บนเมล์เซิร์ฟเวอร์จะถูกลบทิ้งไปเมื่อมีการส่งให้ผู้ใช้เรียบร้อยแล้ว เว้นเสียแต่ว่าได้กำหนดเพิ่มเติมไว้ที่โปรแกรมเมล์ไคล์แอนท์ว่าอย่าให้ลบเมล์ออกจากเซิร์ฟเวอร์(Leave a copy of message on the server)

ในปัจจุบันโปรโตคอลมีออกมาหลายเวอร์ชัน แต่ที่นิยมกันคือ POP3 ซึ่งก็ยังมีข้อจำกัดในการใช้ คือขณะรับและส่งอีเมล์ ฝั่งผู้ใช้จะส่งรหัสผ่านของผู้ใช้ในรูปของข้อความไป ทำให้ไม่ปลอดภัย

นักหากมีการแอบดักข้อมูล เพราะฉะนั้น ตอนเซ็ต POP ไคล์แอนท์ เช่น MS outlook หรือโปรแกรมอื่น ๆ ควรจะเลือกใช้งานการเข้าใช้โดยใช้การตรวจสอบรหัสผ่าน (Log on using Secure Password Authentication: SPA) ด้วย แต่ต้องให้เมล์เซิร์ฟเวอร์มีการสนับสนุนการใช้ SPA ถึงจะใช้งานได้

3.1.4.3 IMAP

IMAP หรือ Internet Message Access Protocol เป็นโปรโตคอลที่เกิดหลัง POP เพื่อแก้ไขข้อจำกัดที่เกิดจาก POP นั่นเอง ทั้งนี้เพราะ POP จะใช้วิธีการโหลดเมล์ที่อยู่บนเซิร์ฟเวอร์มาเก็บไว้ยังเครื่องคอมพิวเตอร์ส่วนบุคคลของผู้ใช้แล้วลบเมล์นั้นทิ้ง (แต่ปัจจุบัน POP พัฒนาขึ้น คือสามารถกำหนดที่เมล์ไคล์แอนท์ได้ว่าจะให้ลบเมล์ทิ้งหรือไม่) ทำให้ผู้ใช้นั้นไม่สามารถอ่านเมล์จากเครื่องคอมพิวเตอร์ส่วนบุคคลเครื่องอื่น ๆ ได้อีก ต้องใช้เครื่องเดิมตลอดซึ่งเป็นปัญหาสำหรับผู้ใช้ทีมีเครื่องเครื่องคอมพิวเตอร์ส่วนบุคคลทั้งที่บ้านและที่ทำงาน หรือองค์กรที่มีเครื่องให้กับพนักงานไม่ครบทุกคน

การทำงานของ IMAP นั้นจะจัดการเมล์ที่อยู่บนเซิร์ฟเวอร์ เช่น อ่านหรือเขียนเมล์ ซึ่งเมล์เหล่านั้นจะยังคงอยู่บนเซิร์ฟเวอร์ ทำให้ผู้ใช้จะใช้เครื่องคอมพิวเตอร์ส่วนบุคคลเครื่องใดอ่านเมล์ก็ได้ หรือสั่งดาวน์โหลดเมล์ที่ต้องการมาเก็บในเครื่องพีซีของตนเองเหมือนกับการทำงานของ POP นอกจากนี้ยังสามารถกำหนดเมล์บ๊อกซ์หนึ่ง ๆ ให้กับผู้ใช้หลาย ๆ คนได้ โดยที่ผู้ใช้เหล่านั้นสามารถเปิดเมล์บ๊อกซ์อ่านได้พร้อม ๆ กัน สำหรับในกรณีเว็บเมล์เครื่องที่เป็นเว็บเซิร์ฟเวอร์ก็จะมีการติดต่อกับเมล์เซิร์ฟเวอร์โดยผ่านโปรโตคอล IMAP เช่นกัน

3.1.5 คำนิยามของอีเมล์ขยะ

G. Hulten จาก Msn Safety Team และ J. Godman จาก Microsoft Research [2] ได้แสดงตัวเลขของการสำรวจคำนิยามของอีเมล์ขยะพบว่า

- 92% เชื่อว่าอีเมล์ขยะคืออีเมล์ที่ประกอบไปด้วยเนื้อหาที่เกี่ยวข้องกับสื่อลามกอนาจาร
- 89% เชื่อว่าอีเมล์ขยะคืออีเมล์ที่ประกอบไปด้วยเนื้อหาที่เกี่ยวข้องกับข้อเสนอทางการเงินต่างๆ
- 76% เชื่อว่าอีเมล์ขยะคืออีเมล์ที่ประกอบไปด้วยเนื้อหาที่เกี่ยวข้องกับการเมืองหรือศาสนา
- 32% จะพิจารณาว่าอีเมล์เหล่านั้นเป็นอีเมล์ขยะถ้าหากมันมาจากผู้ส่งซึ่งทำธุรกิจอยู่ ส่วนคำนิยามซึ่งเป็นลักษณะทางกฎหมายกล่าวว่า อีเมล์ขยะคืออีเมล์ทางการค้าที่เราไม่ได้ให้ความสนใจจากใครก็ตามที่ไม่มีความสัมพันธ์ทางธุรกิจกันมาก่อน และ [2] ได้ให้คำนิยามว่า อีเมล์ขยะหมายถึง อีเมล์ใดๆ ก็ตามที่ผู้ใช้ในระบบรายงานให้ทราบ โดยผู้ใช้เหล่านั้นต้องได้รับการทดสอบอย่างใดอย่างหนึ่งแล้วว่าไม่ใช่สแปมเมอร์

3.2 องค์ประกอบของอีเมล์

อีเมล์ประกอบด้วย ส่วนหัวเรื่อง (Header) และส่วนเนื้อเรื่อง (Body) ดังรูปที่ 3.4

Received: from bright.net (unknown [83.134.197.119])
by zak.futurequest.net ([69.5.6.152])
with SMTP via TCP; 01 May 2005 14:44:44 -0000
From: <curluq@bright.net>
To: <rait@bruce-guenter.dyndns.org>
Subject: 3.25% intere/st ravj
Date: Sun, 1 May 2005 10:05:32 -0500
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Length: 463
Lines: 29

Hello
I have tried contacting you before in regards to the promised low interest mortga(ge rate. The good news is that you did qualify for the loan, for a rate of 4.2% ! You could receive over u380,000 for as little as i500 per month. Ba(d credit? Doesn't matter, low rates are fixed no matter what! To get a fr!ee, no obli,gation consultation click here http://goodwayz.com/x/loan.php?id=somb

Have a great day!
Ellen Horvan

รูปที่ 3.4 แสดงตัวอย่างของอีเมล์ขยะ

3.2.1 ส่วนหัวเรื่อง

จากรูปเราพบว่าส่วนหัวเรื่องประกอบไปด้วยส่วนต่างๆ ดังนี้

1. ส่วน “ได้รับ” (Received) ในส่วนนี้ประกอบไปด้วย รายละเอียดของผู้ส่ง วิธีการที่ใช้ส่ง
2. ส่วน “จาก” (From) ส่วนนี้จะแสดงที่อยู่อีเมล์ของผู้ส่ง
3. ส่วน“ถึง” (To) ส่วนนี้จะแสดงให้เห็นว่า ผู้ส่งได้ทำการส่งข้อความนี้ไปถึงใครบ้าง
4. ส่วน “วันที่” (Date) แสดงวันที่ที่ส่ง
5. ส่วน “ชื่อเรื่อง” (Subject) ส่วนชื่อเรื่องคือส่วนที่จะแสดงให้เห็นเป็นส่วนแรกก่อนที่ผู้ใช้คลิกเข้าไปจึงจะพบกับส่วนเนื้อหาซึ่งเป็นใจความหลักของอีเมล์ ชื่อเรื่องจึงเป็นเหมือนส่วนที่ใช้

สำหรับแสดงเนื้อหาโดยสรุปของอีเมล์นั่นเอง มีผู้นำส่วนชื่อเรื่องมาเป็นส่วนร่วมในการวิเคราะห์ว่าอีเมล์เป็นอีเมล์ขยะหรืออีเมล์ดีเป็นจำนวนมาก เนื่องจากชื่อเรื่องก็อาจสื่อความหมายไปถึงเนื้อความในอีเมล์ได้เช่นกัน

6. ส่วน "รหัสข้อความ" (Message ID) แสดงรหัสของข้อความหรืออีเมล์นั้นๆ

7. ส่วน "เวอร์ชันเอ็มไอเอ็มอี" (MIME-Version) ชื่อเต็มว่าMultipurpose Internet Mail Extensions เป็นมาตราฐานของรูปแบบของอีเมล์

8. ส่วน "ชนิดเนื้อหา" (Content-Type) เป็นส่วนของข้อมูลที่บอกว่าข้อความนี้ถูกแสดงให้เห็นด้วยวิธีใด โดยทั่วๆ ไปแล้วจะเป็นชนิดเอ็มไอเอ็มอี

ในส่วนหัวเรื่องนี้มักถูกนำไปใช้ในการวิเคราะห์อีเมล์ขยะทางฝั่งเซิร์ฟเวอร์เป็นส่วนใหญ่ เนื่องจากค่อนข้างยุ่งยากและไม่คุ้มหากจะใช้เป็นการส่วนตัวหรือฝั่งลูกข่าย แต่สำหรับการวิเคราะห์ที่ต้องการความถูกต้องและปลอดภัยสูง การกรองที่ฝั่งเซิร์ฟเวอร์จึงต้องเข้ามงวดกว่าจึงจำเป็นต้องรับรู้ข้อมูลที่บ่งบอกถึงที่มาที่ไปของอีเมล์เพิ่มเติมในการพิจารณาด้วย

3.2.2 ส่วนเนื้อเรื่อง

ส่วนเนื้อเรื่อง คือส่วนของข้อความที่ผู้ใช้งานอีเมล์ใช้ติดต่อสื่อสารกัน และเป็นส่วนที่นักวิจัยมักนำมาใช้พิจารณาเนื่องจากง่ายต่อตัวกรองในการเรียนรู้ เนื่องด้วยเหตุนี้อีเมล์บางส่วนจึงจำเป็นต้องทำให้เนื้อหาของอีเมล์นั้นคลุมเครือเพื่อหลบหนีบรรดาตัวกรองซึ่งมีอยู่มากมาย[2] ได้จำแนกวิธีต่างๆ ที่ใช้ทำให้อีเมล์คลุมเครือ เช่นการทำส่วนเนื้อหาของอีเมล์ให้เป็นรูปภาพ (Content in Images) การเพิ่มคำดีดีที่ไม่ปะติดปะต่อเข้าไปในอีเมล์ (Good Word Chaff) สุ่มเพิ่มประโยคที่นำมาจากเมล์ดี (Content Chaff) เพราะฉะนั้นในบางครั้งการกรองอีเมล์โดยพิจารณาจากตัวอักษรอย่างเดียวอาจไม่พอสำหรับอีเมล์ขยะที่มีหน้าตาแปลกดังที่ได้กล่าวข้างต้น

3.3 ผลกระทบของอีเมล์ขยะ

ผลกระทบต่อของอีเมล์ขยะนั้นมีหลายอย่างด้วยกัน โดยที่เห็นได้ชัดมีดังต่อไปนี้

1. สิ้นเปลืองเวลาของผู้ใช้ในการจัดการคัดแยกและลบอีเมล์ขยะทิ้ง ผลของอีเมล์ขยะอาจดูไม่ได้ร้ายแรงอะไรถ้าคุณเป็นผู้ใช้ตามบ้าน เพียงแค่ถ้ามีอีเมล์ขยะมาคุณก็แค่ลบเมล์นั้นทิ้งก็ไม่มีอะไรเกิดขึ้น แต่ถ้าเป็นองค์กรธุรกิจ การที่พนักงานต้องคอยมานั่งลบอีเมล์ขยะคงไม่ใช่เรื่องที่น่าชื่นชมนัก มีผลทำให้การทำงานขององค์กรล่าช้า และส่งผลเสียแก่องค์กรไม่น้อย ตัวอย่างเช่น ถ้าบริษัทหนึ่งให้บริการออนไลน์ แล้วมีการส่งอีเมล์ขยะมาสักวันละ 100,000 ฉบับ ซึ่งอาจทำให้เมล์บ๊อกซ์ของบริษัทเต็มส่งผลให้พนักงานแต่ละคนต้องเสียเวลาในการลบเมล์ขยะเมล์ละ 2 วินาที ซึ่งเมื่อมาคิดแล้วเวลารวมที่พนักงานในบริษัทต้องเสียไปกับการลบอีเมล์ขยะถึง 555 ชั่วโมงในการลบหมด

2. สิ้นเปลืองแบนด์วิธทำให้เมล์เซิร์ฟเวอร์ต้องเสียแบนด์วิธไปเป็นจำนวนมากพอๆกับขนาดของอีเมล์ขยะนั้น หากแบนด์วิธมีอยู่อย่างจำกัดก็จะส่งให้อีเมล์อื่นที่เข้ามาจะต้องใช้เวลานานกว่าปกติหรืออาจจะไม่ได้รับในที่สุด

3. สิ้นเปลืองการประมวลผลของหน่วยประมวลผลกลางที่เมล์เซิร์ฟเวอร์

4. สิ้นเปลืองเนื้อที่ในเมล์บ็อกซ์ ซึ่งจะมีผลกระทบมากหากเมล์บ็อกซ์มีการจำกัดเนื้อที่ของผู้ใช้ หากผู้ใช้ทิ้งไว้ไม่ได้มาตรวจสอบบ่อยๆก็จะทำให้เนื้อที่หมดไปได้ หากเป็นอีเมล์ทางธุรกิจที่สำคัญก็จะทำให้เสียหายต่อธุรกิจได้

5. ส่งผลต่อผู้ให้บริการเซิร์ฟเวอร์ที่มีการตั้งค่าในการรีเลย์ไว้ไม่จำกัดกลุ่มที่แน่นอน คือจะทำให้ สแปมเมอร์สามารถใช้เซิร์ฟเวอร์นั้นทำการส่งอีเมล์ขยะออกไป ซึ่งเมื่อมีการสืบต้นตอของตัวแทนขนส่งเมล์ (MTA) ก็จะทำให้เซิร์ฟเวอร์นั้นถูกบล็อกทำให้ไม่สามารถส่งอีเมล์ได้

3.4 ความรู้พื้นฐานเกี่ยวกับทฤษฎีเบย์เซียน (Bayesian Theorem)

ในอดีตอีเมล์ขยะยังไม่ค่อยจะมีผลกระทบต่อระบบคอมพิวเตอร์มากนัก การกรองอีเมล์ขยะในอดีตจึงมีการใช้กฎเกณฑ์อย่างง่ายๆ ในการจำแนก วิธีการที่ใช้เช่น การตั้งกฎในการจำแนกรูปแบบของอีเมล์ขยะ (Pattern-Matching) โดยมีการตั้งกฎที่ระบุว่าอีเมล์ลักษณะไหนที่เป็นอีเมล์ขยะ แต่เนื่องจากในปัจจุบันผู้ส่งอีเมล์ขยะ สามารถส่งอีเมล์ขยะที่มีความสามารถในการหลบเลี่ยงตัวกรองจึงทำให้วิธีการจำแนกรูปแบบของอีเมล์ขยะเริ่มที่จะไม่มีประสิทธิภาพ เพราะวิธีการนี้เป็นวิธีที่มีการตั้งกฎเกณฑ์ที่ตายตัว ซึ่งเมื่อผู้ส่งอีเมล์ขยะเปลี่ยนแปลงรูปแบบในการส่งก็จำเป็นต้องเพิ่มกฎใหม่เข้าไปเรื่อยๆ ทำให้ยากต่อการเปลี่ยนแปลงตามความสามารถในการหลบเลี่ยงที่เปลี่ยนไปของผู้ส่งอีเมล์ขยะ จึงมีการคิดค้นวิธีการในการกรองอีเมล์ขยะที่มีความสามารถเพิ่มขึ้น โดยมีนำทฤษฎีต่างๆ เข้ามาเพื่อใช้ในการพัฒนาตัวกรองอีเมล์ขยะเพื่อให้มีความสามารถที่เพิ่มขึ้น

หนึ่งในวิธีที่ใช้ในปัจจุบันและมีประสิทธิภาพคือการนำการวิเคราะห์ทางสถิติเข้ามามีส่วนช่วยในการเรียนรู้ของอีเมล์ ซึ่งเป็นวิธีการที่มีประสิทธิภาพในการจำแนกประเภทของอีเมล์ขยะคือการจำแนกโดยใช้ทฤษฎีเบย์เซียน ตัวกรองอีเมล์ขยะที่พัฒนาขึ้นโดยการใช้ทฤษฎีเบย์เซียนจะมีพื้นฐานของการนำความน่าจะเป็นของอีเมล์มาคำนวณ ซึ่งเป็นที่วิธีการในการคัดเลือกคีย์เวิร์ดเพื่อนำมาใช้ในการคำนวณ โดยการคัดเลือกคีย์เวิร์ดจะขึ้นกับหลักการในการจำแนกคำ รวมทั้งการตั้งกฎของการจำแนกคำโดยผู้พัฒนาเป็นผู้จัดการ การใช้หลักการทางสถิติมาใช้ในการสร้างตัวกรองเป็นสิ่งที่ค่อนข้างสมเหตุสมผล เพราะว่าความแตกต่างกันระหว่างอีเมล์ทั่วไปกับอีเมล์ขยะ ค่อนข้างจะแยกออกจากกันได้ยาก ซึ่งอีเมล์ขยะทั่วไปจะมีรูปแบบหรือว่าหลักการที่คล้ายคลึงกันกับอีเมล์ทั่วไป ซึ่งสุดท้ายต้องขึ้นกับผู้ใช้เป็นผู้ระบุเองว่าอีเมล์ฉบับไหนที่เป็นอีเมล์ขยะ หรือว่าอีเมล์ทั่วๆไป และผู้ใช้แต่ละคนอาจจะระบุได้แตกต่างกัน ซึ่งอีเมล์ฉบับหนึ่งเป็นอีเมล์ขยะของผู้ใช้คนหนึ่ง แต่ในขณะที่ผู้ใช้อีกคนอาจจะบอกว่าเป็นอีเมล์ปกติ

จากการที่ผู้ใช้เป็นผู้กำหนดนิยามของอีเมล์ขยะเองทำให้สังเกตได้ว่าสิ่งหนึ่งที่แตกต่างกันระหว่างอีเมล์ทั่วไปกับอีเมล์ขยะก็คือ เนื้อหาที่อยู่ภายในอีเมล์ วิธีการทางด้านสถิติจะมีข้อเสียก็คือการใช้งานตัวกรองที่สร้างขึ้นจะต้องให้อีเมล์มีการเรียนรู้ในรูปแบบต่างๆ ซึ่งต้องใช้ระยะเวลาหนึ่งการสอนอีเมล์ให้แก่ระบบตัวกรอง ซึ่งทฤษฎีเบย์เซียนจะมีรายละเอียดดังต่อไปนี้

3.4.1 ทฤษฎีเบย์เซียน

ทฤษฎีเบย์จะเกี่ยวข้องกับเงื่อนไขและสถิติของเหตุการณ์ A และ B ดังสมการที่ (3.1)

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)} \tag{3.1}$$

เมื่อ

$\Pr(A)$ คือความน่าจะเป็นที่มาก่อน (Prior Probability) หรือค่าความน่าจะเป็นซึ่งมีความสำคัญน้อย (Marginal Probability) ของ A

$\Pr(A \mid B)$ คือความน่าจะเป็นแบบมีเงื่อนไข (Conditional Probability) ของ A เมื่อให้ B หรือเรียกว่าความน่าจะเป็นที่มาที่หลัง (Posterior Probability) เนื่องจากถูกแปลงหรือเป็นอิสระจากค่าของ B ที่ได้กำหนดไว้

$\Pr(B \mid A)$ คือความน่าจะเป็นแบบมีเงื่อนไขของ B เมื่อให้ A

$\Pr(B)$ คือความน่าจะเป็นที่มาก่อน (Prior Probability) หรือค่าความน่าจะเป็นซึ่งมีความสำคัญน้อย (Marginal Probability) ของ B ซึ่งกระทำตัวเป็นนอร์มัลไลซิ่งคอนสแตนท์ (Normalizing Constant) ถอดความทฤษฎีใหม่ได้ดังสมการที่ (3.2)

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Normalizing Constant}} \tag{3.2}$$

นั่นคือความน่าจะเป็นที่เกิดทีหลังคือสัดส่วนของจำนวนครั้งที่คล้ายกับความน่าจะเป็นที่เกิดขึ้นก่อน ในทางตรงกันข้าม อัตราส่วน $P(B \mid A) / P(B)$ ในบางครั้งเราจะเรียกว่าความคล้ายมาตราฐาน (Standardized Likelihood) ดังนั้นทฤษฎีนี้จึงถูกถอดความใหม่ได้ดังสมการที่ (3.3)

$$\text{Posterior} = \text{Standardized Likelihood} \times \text{Prior} \tag{3.3}$$

จากนั้นเราจะดีไรฟ์ (Derive) ทฤษฎีโดยเริ่มจากนิยามของความน่าจะเป็นแบบมีเงื่อนไข ความน่าจะเป็นของเหตุการณ์ A เมื่อให้เหตุการณ์ B เป็นไปตามสมการที่ (3.4) คือ

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{3.4}$$

ในทำนองเดียวกัน ความน่าจะเป็นของเหตุการณ์ B เมื่อให้เหตุการณ์ A เป็นไปตามสมการที่ 3.5 คือ

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{3.5}$$

เมื่อนำ 2 สมการนี้มาจัดเรียงใหม่จะเป็นไปตามสมการที่ (3.6) คือ

$$P(A \mid B)P(B) = P(A \cap B) = P(B \mid A)P(A) \tag{3.6}$$

หารทั้งสองข้างด้วย $Pr(B)$ จะได้ทฤษฎีของเบย์จะเป็นไปตามสมการที่ (3.7) คือ

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{3.7}$$

3.4.2 การนำทฤษฎีเบย์เซียนมาใช้ในการกรองอีเมล์ขยะ

ในการนำทฤษฎีของเบย์มาประยุกต์ใช้กับตัวกรองอีเมล์ขยะนั้น นิยมใช้ทฤษฎีเบย์อย่างง่าย (Naïve Bayes)[2] ซึ่งบางครั้งเราอาจเรียกว่าเบย์เซียนเทคนิค โดยทฤษฎีเบย์อย่างง่ายนั้นจะมองว่าคำแต่ละคำเป็นอิสระจากกัน

ถ้าต้องการ P(spam|mail) ซึ่งหมายถึงความน่าจะเป็นที่อีเมล์จะเป็นขยะ สามารถหาได้โดยใช้กฎของเบย์ซึ่งคำนวณตามสมการที่ 3.8 โดยที่ค่า P(ham) หรือความน่าจะเป็นของอีเมล์ดี, P(spam) หรือความน่าจะเป็นของอีเมล์ขยะ, P(mail) หรือความน่าจะเป็นของอีเมล์ หาได้จาก สมการที่ (3.9)-(3.11) ตามลำดับ

$$P(spam \mid mail) = \frac{P(mail \mid spam) \times P(spam)}{P(mail)} \tag{3.8}$$

$$P(ham) = \frac{nMailHam}{nMailTotal} \tag{3.9}$$

$$P(spam) = \frac{nMailSpam}{nMailTotal} \tag{3.10}$$

$$P(mail) = P(mail \mid spam) \times P(spam) + P(mail \mid ham) \times P(ham) \qquad (3.11)$$

จากสมมติฐานความเป็นอิสระที่กล่าวว่า ความน่าจะเป็นของแต่ละตัวเป็นอิสระจากกัน (ข้อสมมติฐานที่ผิด) ทำให้ได้สมการที่ (3.12) และ (3.13)

$$P(mail \mid spam) \approx P(word_1 \mid spam) \times P(word_2 \mid spam) \times ... \times P(word_n \mid spam) \qquad (3.12)$$

$$P(mail \mid ham) \approx P(word_1 \mid ham) \times P(word_2 \mid ham) \times ... \times P(word_n \mid ham) \qquad (3.13)$$

อีเมล์ประกอบด้วยคำหลายๆคำ(Words) เมื่อต้องการทราบว่าอีเมล์ฉบับใดเป็นอีเมล์ขยะ จะพิจารณาจากความน่าจะเป็นที่จะเป็นอีเมล์ขยะของคำแต่ละคำในอีเมล์แล้วจึงหาความน่าจะเป็นรวมของอีเมล์ทั้งฉบับโดยการนำความน่าจะเป็นของคำขยะที่ได้มาคูณกัน เช่นเดียวกัน ถ้าต้องการทราบว่าอีเมล์ฉบับใดเป็นอีเมล์ดีให้พิจารณาจากความน่าจะเป็นที่จะเป็นอีเมล์ดีของคำดีในอีเมล์แล้วทำเช่นเดียวกันกับการหาความน่าจะเป็นของอีเมล์ขยะ

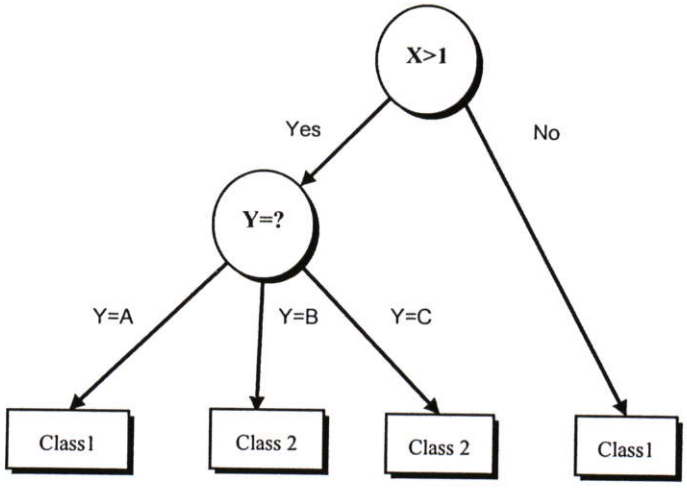
เมื่อได้ความน่าจะเป็นจะเป็นของอีเมล์ที่จะเป็นอีเมล์ขยะ $P(mail \mid spam)$ และความน่าจะเป็นของอีเมล์ที่จะเป็นอีเมล์ดี $P(mail \mid ham)$ ให้นำมาเปรียบเทียบกันถ้าความน่าจะเป็นของฝั่งอีเมล์ขยะมีค่ามากกว่าให้ตอบว่าเป็นอีเมล์ที่นำมาทดสอบเป็นอีเมล์ขยะ ถ้าไม่ก็ตอบว่าเป็นอีเมล์ดี

3.5 ความรู้พื้นฐานเกี่ยวกับแขนงการตัดสินใจ (Decision Tree)

แขนงการตัดสินใจคือโมเดลสำหรับการทำนายซึ่งใช้ Top-Down Strategy มาใช้ในการค้นหาคำตอบในเส้นทางของคำตอบที่สารถเป็นไปได้ องค์ประกอบของแขนงการตัดสินใจประกอบด้วย

1. โหนด (Node) คือสิ่งที่ถูกทดสอบ หรือเงื่อนไข
2. แขนงหรือกิ่ง (Branches) คือ ผลลัพธ์ที่เป็นไปได้ทั้งหมดของโหนดนั้น

ตัวอย่างของแขนงการตัดสินใจแสดงดังรูปที่ 3.5



รูปที่ 3.5 แสดงตัวอย่างของแขนงการตัดสินใจ

แขนงการตัดสินใจสามารถนำมาประยุกต์เป็นโครงสร้างของกฎได้ เช่น

IF (Condition) THEN (Action)

IF (Condition1) AND (Condition2)THEN (Action)

3.6 ความรู้พื้นฐานเกี่ยวกับเจเนติกอัลกอริทึม

ในปี ค.ศ. 1975 จอห์น ฮอลแลนด์และคณะ [15] จากมหาวิทยาลัยมิชิแกนได้ตีพิมพ์หนังสือการปรับตัวในสิ่งแวดล้อมและระบบเทียม (Adaptation in Natural and Artificial System) ขึ้น ซึ่งเป็นครั้งแรกที่นำแนวความคิดของการเลียนแบบวิวัฒนาการธรรมชาติมากมายมารวมไว้ด้วยกันบนคอมพิวเตอร์ ซึ่งเตรียมสิ่งสำคัญต่าง ๆ ในการวิเคราะห์ปัญหาทางคณิตศาสตร์ เสมือนกับการเลียนแบบการวิวัฒนาการบนคอมพิวเตอร์ และเมื่อไม่นานมานี้งานมากมายที่ได้มาจากการเลียนแบบการวิวัฒนาการบนคอมพิวเตอร์ก็ถูกสร้าง จุดประสงค์เดียวเพื่อความเข้าใจในพันธุ์กรรม (Genetic) และวิวัฒนาการ (Evolution) หนังสือของฮอลแลนด์แสดง เคร้าโครงว่ากระบวนการสามารถใช้ในการแก้ปัญหาในโลกที่แท้จริงได้อย่างไรด้วยเทคนิคของการวิวัฒนาการ ฮอลแลนด์ได้ให้ความหมายของเจเนติกอัลกอริทึมว่า เป็นอัลกอริทึมสำหรับการค้นหาข้อมูลและการค้นหาคำตอบที่ดีที่สุด (Optimization) ซึ่งได้รับแนวคิดมาจากกลไกการคัดเลือกสายพันธุ์ตามธรรมชาติ (Natural Selection) และธรรมชาติทางพันธุกรรม (Natural Genetic) คือสิ่งมีชีวิตใดที่มีความแข็งแรงกว่าย่อมมีโอกาสอยู่รอดได้มากกว่านั่นหมายถึงการมีสายพันธุ์ที่ดีย่อมมีโอกาสจะได้รับกา

คัดเลือกนำมาเป็นต้นแบบ เพื่อถ่ายทอดลักษณะดี ๆ ของสายพันธุ์เหล่านั้นไปยังรุ่นต่อๆ ไป มากกว่า มีโอกาสในการอยู่รอดสูง ส่วนสายพันธุ์ที่ไม่ดีก็จะไม่ได้รับการคัดเลือก หรือได้รับการคัดเลือกน้อยกว่า และจะค่อยๆ สูญพันธุ์ไปในที่สุดเจเนติกอัลกอริทึมได้นำกระบวนการวิวัฒนาการของสิ่งมีชีวิตมาประยุกต์ใช้กับงานด้านปัญญาประดิษฐ์ เพื่อค้นหาคำตอบของปัญหาต่าง ๆ โดยเจเนติกอัลกอริทึมเป็นรูปแบบของเทคนิคการค้นหาซึ่งใช้ในการค้นหาคำตอบจากจำนวนคำตอบที่เป็นไปได้ทั้งหมดของการแก้ปัญหานั้นๆ เพื่อให้ได้คำตอบที่เหมาะสมกับปัญหาโดยอาศัยข้อมูลในการช่วยค้นหา ซึ่งข้อมูลหรือวิธีการที่ใช้นี้จำลองมาจากกฎเกณฑ์การคัดเลือกสายพันธุ์ตามธรรมชาตินั่นเอง

เจเนติกอัลกอริทึมมีองค์ประกอบที่สำคัญ 5 องค์ประกอบ ได้แก่

1. นำเสนอปัญหาด้วยรูปแบบโครโมโซม และทางเลือกที่เป็นไปได้ของแต่ละปัญหา
2. วิธีการสร้างประชากรต้นกำเนิด (Initial population) ของทางเลือกที่เป็นไปได้
3. ฟังก์ชันความเหมาะสม (Fitness function) เพื่อให้คะแนนแต่ละทางเลือก
4. เจเนติกโอเปอเรเตอร์ (Genetic Operator) ซึ่งใช้ปรับเปลี่ยนองค์ประกอบของข้อมูลตลอดกระบวนการ
5. ค่าพารามิเตอร์ต่างๆ ซึ่งต้องใช้ในเจเนติกอัลกอริทึม เช่น ขนาดของประชากร ความน่าจะเป็นของการใช้เจเนติกโอเปอเรเตอร์ เป็นต้น

เจเนติกอัลกอริทึมแตกต่างจากวิธีการโดยทั่วไป คือ

1. เป็นวิธีการที่ค้นหาคำตอบภายใต้โครงสร้างของปัญหาอันเกิดจากการเข้ารหัส (Encoding) รูปแบบปัญหา
2. โครงสร้างจากกลุ่มตัวแปรต่างๆ ของปัญหานั้น ไม่ใช่การค้นหาคำตอบจากค่าของกลุ่มตัวแปรนั้นโดยตรง
3. ทำการค้นหาคำตอบจากกลุ่มประชากรคำตอบ(Population)แทนการหาคำตอบใดคำตอบหนึ่ง
4. ทำการค้นหาคำตอบจากผลลัพธ์ของกลุ่มค่าตัวแปรที่เป็นฟังก์ชันเป้าหมาย (Objective Function) ไม่สนใจข้อมูลข่าวสารแวดล้อมอื่นๆ
5. ใช้ความน่าจะเป็น (Probability) ในการค้นหาคำตอบ

3.6.1 พันธุศาสตร์ทางชีววิทยากับเจเนติกอัลกอริทึม

ตามธรรมชาติ สิ่งมีชีวิตแต่ละชนิดจะมีโครงสร้างและพฤติกรรมที่แตกต่างกัน อันเนื่องมาจากสภาพแวดล้อมการดำรงชีวิตที่แตกต่างกัน ลักษณะที่แตกต่างกันนี้มีผลต่ออัตราการมีชีวิตรอดและอัตราการสืบพันธุ์ โดยสิ่งมีชีวิตมีแนวโน้มจะถ่ายทอดคุณลักษณะพิเศษให้กับประชากรรุ่นลูกหลาน (Offspring) และให้กำเนิดสิ่งมีชีวิตที่มีลักษณะพิเศษแตกต่างไปจากเดิมที่มีคุณสมบัติ

เหมาะสม เพื่อให้สามารถดำรงอยู่รอดได้ต่อไปในสภาพแวดล้อมของสิ่งมีชีวิตนั้นๆ ประชากรจะมีแนวโน้มที่จะมีคุณลักษณะที่เหมาะสมต่อการดำรงชีวิตมากกว่ารุ่นบรรพบุรุษ เมื่อเวลาผ่านไปหลายๆ รุ่น (Generation) ของวิวัฒนาการ สิ่งมีชีวิตนั้นก็จะได้สายพันธุ์ใหม่ที่เหมาะสมกับสภาพแวดล้อมมากยิ่งขึ้น ตัวอย่างของวิวัฒนาการเหล่านี้ เช่น มีการสันนิษฐานว่ายีราฟในสมัยโบราณอาจจะมีลำคอไม่ยาวเท่ากับยีราฟในยุคปัจจุบัน ยีราฟดำรงชีวิตด้วยการกินใบไม้ตามยอดไม้ เมื่อจำนวนประชากรยีราฟมีมากขึ้น การแก่งแย่งอาหารเพื่อความอยู่รอดจึงสูงขึ้นตาม ยีราฟตัวที่สามารถกินยอดไม้สูงๆ เท่านั้นจึงจะมีชีวิตอยู่ต่อไป คุณสมบัติที่ดีในการดำรงชีวิตนี้จึงถูกคัดเลือกและถ่ายทอดมายังยีราฟรุ่นลูกหลาน นั่นหมายถึงยีราฟที่คอยาวเท่านั้นที่จะมีโอกาสหาอาหาร และรอดชีวิตสูงกว่ายีราฟคอสั้น

จากที่กล่าวมาข้างต้นว่าเจเนติกอัลกอริทึมเป็นวิธีการที่เลียนแบบมาจากหลักการทางชีววิทยา จึงมีการนำศัพท์ต่างๆ ในด้านพันธุศาสตร์มาประยุกต์ใช้ในกระบวนการเจเนติกอัลกอริทึม ดังต่อไปนี้

3.6.1.1 พันธุศาสตร์ทางชีววิทยา

ในแต่ละเซลล์ (Cell) ของสิ่งมีชีวิตจะประกอบไปด้วยหน่วยย่อยที่มีความสำคัญมากอยู่ในนิวเคลียส (Nucleus) ของเซลล์ นั่นคือโครโมโซม (Chromosome) แต่ละโครโมโซมจะประกอบไปด้วยยีนส์ (Genes) ซึ่งเป็นหน่วยเก็บลักษณะต่างๆ ของสิ่งมีชีวิต ภายในยีนส์จะมีค่าแสดงลักษณะต่างๆ หรือแอลลี (Allele) โดยตำแหน่งของยีนตาละยีนในโครโมโซมจะเรียกว่าโลคัส (Locus) รูปแบบของยีนส์ที่แตกต่างกันเรียกว่าจีโนไทป์ (Genotype) ส่วนลักษณะภายนอกที่ปรากฏเรียกว่าฟีโนไทป์ (Phenotype) [16]

3.6.1.2 พันธุศาสตร์ทางเจเนติกอัลกอริทึม

สำหรับเจเนติกอัลกอริทึม ตัวแปรของปัญหาจะถูกแปลงให้อยู่ในรูปของสตริง (String) ว่าโครโมโซม ภายในโครโมโซมจะประกอบไปด้วยอักขระ (Character) หรือบิท (Bit) แต่ละตำแหน่งของบิทจะเก็บค่าอักขระ (Character value) หรือค่าของบิท (Bit value) ที่แสดงโครงสร้างของแต่ละโครโมโซมของปัญหาที่แตกต่างกัน สรุปความหมายของคำสำคัญต่างๆ ได้ดังตารางที่ 1

ตารางที่ 3.1 คำสำคัญที่ใช้ในกระบวนการเจเนติกอัลกอริทึม

Natural Genetic Terms	Genetic Algorithm Terms
Chromosome	String
Gene	Feature, Character, Bit
Allele	Feature value, Character value, Bit value
Locus	String position
Genotype	Structure
Phenotype	Decoded structure, Alternative solution

3.6.2 ขั้นตอนการทำงานของเจเนติกอัลกอริทึม

3.6.2.1 การกำหนดรูปแบบโครโมโซม (Chromosome Representation)

การกำหนดปัญหาโดยใช้เจเนติกอัลกอริทึมนั้น จะต้องมีการนำปัญหามาเข้ารหัสข้อมูลให้อยู่ในรูปแบบโครโมโซมที่เหมาะสม เช่น อาจนำเสนอในรูปแบบของเลขฐานสอง เลขจำนวนจริง ตัวอักษร) การสลับลำดับกันในพีชคณิต และแบบทรี (Tree) เป็นต้น สำหรับเจเนติกอัลกอริทึมจะใช้กระบวนการเจเนติกอัลกอริทึมแบบง่าย (Simple Genetic Algorithm) [17] ได้แสดงตัวอย่างการเข้ารหัสแบบไบนารี การเข้ารหัสแบบสลับลำดับ (Permutation Encoding) การเข้ารหัสแบบค่า (Value Encoding) และการเข้ารหัสแบบทรี (Tree Encoding) ดังแสดงในรูปที่ 3.6-3.9 ตามลำดับ

Chromosome A	10110010110010101011100101
Chromosome B	11111110000011000001111

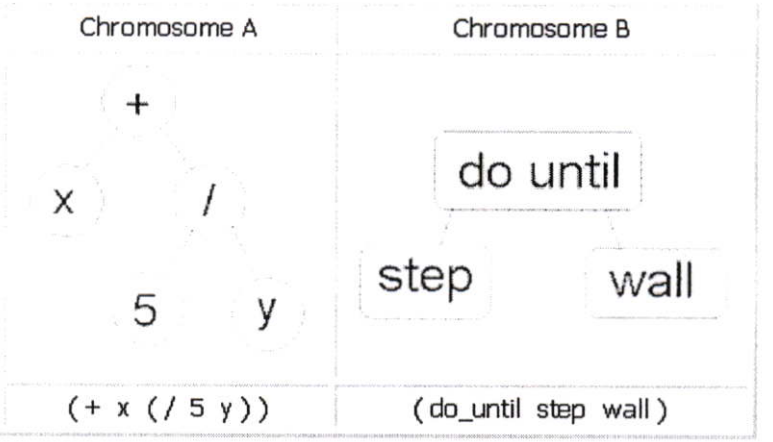
รูปที่ 3.6 แสดงการเข้ารหัสแบบไบนารี

Chromosome A	1 5 3 2 6 4 7 9 8
Chromosome B	8 5 6 7 2 3 1 4 9

รูปที่ 3.7 แสดงการเข้ารหัสแบบสลับลำดับ

Chromosome A	1.2324 5.3243 0.4556 2.3293 2.4545
Chromosome B	ABDJEIFJDHDIERJFDLDFLFEGT
Chromosome C	(back), (back), (right), (forward), (left)

รูปที่ 3.8 แสดงการเข้ารหัสแบบค่า



รูปที่ 3.9 แสดงการเข้ารหัสแบบทรี

3.6.2.2 ประชากร (Population)

ประชากรในกระบวนการเจเนติกอัลกอริทึมจะแบ่งออกเป็น 2 กลุ่มคือ ประชากรรุ่นเก่า (Old Population) และประชากรรุ่นใหม่ (New Population) ประชากรรุ่นเก่าจะถูกสร้างขึ้นมาเพื่อที่จะคัดเลือกไปเป็นประชากรรุ่นใหม่ และสำหรับประชากรต้นกำเนิด (Initial Population) ซึ่งเป็นประชากรรุ่นแรกในกระบวนการสามารถทำได้โดยการสุ่มสร้างค่าที่เป็นไปได้ของแต่ละบิทของแต่ละโครโมโซมตามที่ต้องการ

3.6.2.3 กำหนดฟังก์ชันความเหมาะสม (Fitness Function)

การกำหนดฟังก์ชันความเหมาะสมคือ การสร้างฟังก์ชันเพื่อคำนวณหาความเหมาะสมของประชากรว่าเหมาะสมที่จะถูกคัดเลือกมาเพื่อสร้างประชากรรุ่นต่อไปมากน้อยเพียงใด อาจเป็นการวัดจากค่าความเหมาะสมที่สูงสุด (Max) หรือเป็นค่าความเหมาะสมที่ต่ำสุด (Min) ก็ได้ โดยฟังก์ชันความเหมาะสมนั้นจะแตกต่างกันออกไปสำหรับแต่ละปัญหา

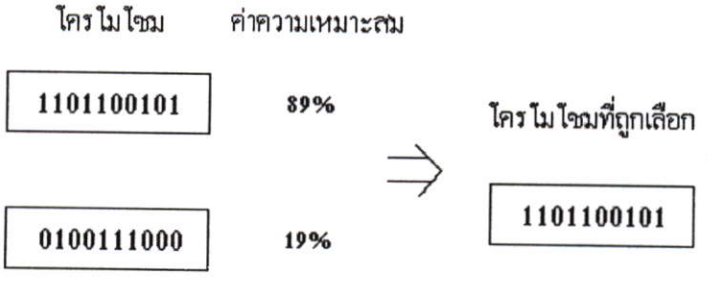
3.6.2.4 การวิเคราะห์ค่าความเหมาะสม (Fitness Evaluation)

เป็นขั้นตอนการถอดรหัสโครโมโซม โดยนำค่าที่ได้จากการถอดรหัสนี้ไปแทนค่าในฟังก์ชันความเหมาะสมของปัญหา ผลลัพธ์ที่ได้จากการคำนวณนี้เรียกว่า ค่าความเหมาะสม ซึ่งค่านี้จะเป็นสิ่งที่แสดงว่า แต่ละโครโมโซมมีความเหมาะสมที่จะนำมาใช้แก้ปัญหามากน้อยเพียงใด หรืออาจเปรียบเทียบได้กับค่าความสามารถในการอยู่รอดของแต่ละโครโมโซม และเป็นการกำหนดโอกาสหรือสัดส่วนที่แต่ละโครโมโซมจะถูกคัดเลือกมาเป็นต้นแบบในการให้กำเนิดประชากรรุ่นต่อไป

3.6.2.5 การคัดเลือก (Selection)

เป็นขั้นตอนที่จำลองแบบการคัดเลือกประชากรตามธรรมชาติ เพื่อคัดเลือกโครโมโซมรุ่นเก่าให้เป็นโครโมโซมต้นแบบหรือโครโมโซมพ่อแม่ เพื่อใช้ในการสร้างประชากรรุ่นลูกต่อไป

สำหรับการคัดเลือกทำได้โดยการวัดค่าความเหมาะสมของแต่ละโครโมโซมโดยวิธีใดวิธีหนึ่ง แล้วคัดเลือกโครโมโซมจำนวนหนึ่งที่มีค่าความเหมาะสมเป็นที่น่าพอใจเก็บไว้ ดังตัวอย่างในรูปที่ 3.10 ดยอาจจะคัดเลือกเอาเฉพาะโครโมโซมที่มีค่าความเหมาะสมสูงสุดหรือในบางครั้งอาจเลือกโครโมโซมที่มีค่าความเหมาะสมปานกลางและต่ำบางส่วนเข้ามาด้วย เพราะบางกรณีการนำสายพันธุ์ที่มีค่าปานกลางหรือต่ำมาผสมกันจะสามารถทำให้เกิดสายพันธุ์ที่ดีในรุ่นต่อไปได้



รูปที่ 3.10แสดงการเลือกโครโมโซมตามค่าความเหมาะสม

การคัดเลือกข้อมูลมีลักษณะเป็นไปตามหลักการที่ว่า การอยู่รอดของสิ่งที่เหมาะสมที่สุด (Survival of the Fittest) [18][19] ถ้าเป็นการวัดค่าความเหมาะสมจากค่าสูงสุด (Maximized Value) ความน่าจะเป็นของของแต่ละโครโมโซม ที่จะได้รับการสุ่มเลือกแต่ละครั้ง (Probability of Value Selection: P_{Si}) จะเป็นไปดังสมการที่ (3.11)

$$P_{Si} = \frac{f_i}{\sum_{i=1}^{n} f_i} \tag{3.11}$$

เมื่อ ค่าความเหมาะสมของแต่ละทางเลือก (f_i) เทียบกับผลรวมค่าความเหมาะสมทั้งหมด หากเป็นการวัดค่าความเหมาะสมจากค่าต่ำสุด (Minimized Value) ความน่าจะเป็นของแต่ละโครโมโซมที่จะได้รับการสุ่มเลือกแต่ละครั้ง จะเป็นไปดังสมการที่ (3.12)

$$P_{Si} = 1 - \frac{f_i}{\sum_{i=1}^{n} f_i} \tag{3.12}$$

ดังนั้นสามารถคำนวณค่าความคาดหวังที่จะสุ่มได้ (Expected Value : E) ของแต่ละโครโมโซม เป็นไปดังสมการที่ (3.13)

$$E = P_{Si} \times Popsize \tag{3.13}$$

เมื่อ Popsize คือขนาดของประชากรทั้งหมด

ในการสุ่มโครโมโซมของเจเนติกอัลกอริทึมแบบง่าย จะใช้แบบจำลองการหมุนวงล้อถ่วงน้ำหนัก (Roulette Wheel) [19] ซึ่งจะกำหนดขนาดของช่องวงล้อตามความน่าจะเป็นที่จะสุ่มได้ในแต่ละครั้ง ของแต่ละโครโมโซมมีวิธีการดังนี้

1. หาค่าความเหมาะสมของแต่ละโครโมโซม
2. หาความน่าจะเป็นที่จะสุ่มได้ในแต่ละครั้งของแต่ละโครโมโซม
3. หาความถี่สะสม (q_i) ของความน่าจะเป็นของแต่ละโครโมโซม ดังสมการที่ (3.14)

$$q_i = \sum_{i=1}^{j} P_{Si} \tag{3.14}$$

4. สร้างเลขสุ่มจำนวนจริง (r) ที่มีค่าอยู่ในช่วง [0.0, 1.0]
5. เลือกโครโมโซมลำดับที่ r ซึ่ง r มีค่าอยู่ระหว่าง q_{i-1} และ q_i

จากวิธีดังกล่าวจะเห็นได้ว่า โครโมโซมใดที่มีความน่าจะเป็นที่จะถูกเลือกน้อยๆ จะมีโอกาสถูกเลือกขึ้นมาน้อยเพราะช่องว่างระหว่าง q_{i-1} และ q_i จะแคบมาก ทำให้โอกาสที่ r จะตกช่องนั้นมีน้อย ในทางตรงกันข้าม โครโมโซมใดที่มีความน่าจะเป็นสูง ก็จะมีโอกาสถูกเลือกมากเนื่องจากช่องว่างระหว่าง q_{i-1} และ q_i จะกว้าง ซึ่งสอดคล้องกับที่ได้กล่าวไปแล้วว่า ถ้าความน่าจะเป็นที่จะถูกเลือกมีค่ามาก ก็จะมีโอกาสที่จะถูกเลือกไปเป็นประชากรรุ่นใหม่สูงตามไปด้วย ดังแสดงในตารางที่ 2

ตารางที่ 3.2 ตัวอย่างการใช้แบบจำลองวงล้อถ่วงน้ำหนัก

โครโมโซม	**1**	**2**	**3**	**4**	**5**
ค่าความเหมาะสม	8	2	17	7	2
ค่าความน่าจะเป็นที่จะสุ่มได้แต่ละครั้ง (P_{Si})	0.22	0.06	0.47	0.19	0.06
ความถี่สะสมค่าความน่าจะเป็น (q_i)	0.22	0.28	0.75	0.94	1.00
สร้างเลขสุ่มจากการหมุนวงล้อแต่ละครั้ง (r)	0.33	0.84	0.45	0.12	0.28
โครโมโซมที่ถูกเลือก	3	4	3	2	1

นอกเหนือจากการคัดเลือกโดยใช้แบบจำลองการหมุนวงล้อถ่วงน้ำหนัก [20] ได้รวบรวมวิธีการอื่นๆ ที่ใช้ในกระบวนการคัดเลือกได้แก่ การคัดเลือกแบบจัดลำดับ (Ranking Selection) การคัดเลือกตามสถานะที่แน่นนอน (Steady-State Selection) และการคัดเลือกแบบมีอภิสิทธิ์ (Elitism Selection) ซึ่งเราสามารถเลือกใช้ได้ตามความเหมาะสมในแต่ละงาน

3.6.2.6 การครอสโอเวอร์ (Crossover)

การครอสโอเวอร์ คือการนำโครโมโซม 2 โครโมโซม มาทำการตามขั้นตอนต่างๆ ซึ่งจะให้ค่าโครโมโซมใหม่ที่จะนำไปใช้ในการคัดเลือกครั้งต่อไป หรือหมายถึงการนำโครโมโซมสองโครโมโซมมาผสมกันเพื่อให้ได้ค่าโครโมโซมขึ้นมาใหม่นั่นเอง ในขั้นตอนนี้จะพยายามสร้างทางเลือกใหม่เพื่อเป็นคำตอบให้กับปัญหา และปรับปรุงทางเลือกให้ดีขึ้นโดยการครอสโอเวอร์ ซึ่งเจเนติกอัลกอริทึมจะพยายามสร้างทางเลือกที่ดีขึ้นโดยการรวมลักษณะที่ดีของแต่ละโครโมโซมเข้าด้วยกัน โครโมโซมที่มีค่าความเหมาะสมสูงกว่ามักจะถูกเลือกมาทำการครอสโอเวอร์บ่อยครั้งกว่า ส่งผลให้มีโอกาสในการอยู่รอดไปยังรุ่นต่อไปสูงกว่า การครอสโอเวอร์สามารถทำได้หลายวิธี เช่น การครอสโอเวอร์หนึ่งจุด (One Point Crossover) การครอสโอเวอร์สองจุด (Two Point Crossover) และการครอสโอเวอร์หลายจุด (Multiple Point Crossover) ซึ่งมีวิธีการโดยทั่วไปดังนี้

1. ประชากรทั้งหมดจะถูกนำมาจับคู่โดยการสุ่ม ซึ่งจะได้ผลการจับคู่ออกมาทั้งหมด N/2 คู่ เมื่อ N คือจำนวนประชากรทั้งหมดในรุ่นนั้นๆ
2. สร้างเลขสุ่มจำนวนจริง (r) ซึ่งมีค่าอยู่ในช่วง [0.0, 1.0] โดยถ้าจำนวนจริงที่สุ่มได้มีค่าน้อยกว่าค่าความน่าจะเป็นในการครอสโอเวอร์ (Probability of Crossover : P_c) แล้วโครโมโซมคู่นั้นจึงจะเกิดการครอสโอเวอร์
3. ครอสโอเวอร์โดยแลกเปลี่ยนส่วนของคู่โครโมโซมพ่อแม่นั้น โดย
   - สุ่มเลือกตำแหน่งที่จะทำการครอสโอเวอร์
   - แลกเปลี่ยนค่าในแต่ละบิทของคู่โครโมโซมพ่อแม่ ตั้งแต่ตำแหน่งที่สุ่มได้จนหมด ซึ่งทำให้เกิดโครโมโซมรุ่นลูกใหม่จำนวน 2 โครโมโซม

การครอสโอเวอร์ในแต่ละรุ่นสามารถทำได้มากกว่า 1 คู่ ขึ้นอยู่กับอัตราค่าความน่าจะเป็นในการครอสโอเวอร์ ซึ่งจำนวนของการครอสโอเวอร์สามารถคำนวณได้ตามสมการที่ 3.15

$$\text{จำนวนของการครอสโอเวอร์} = P_c \times (\text{Popsize} / 2) \qquad (3.15)$$

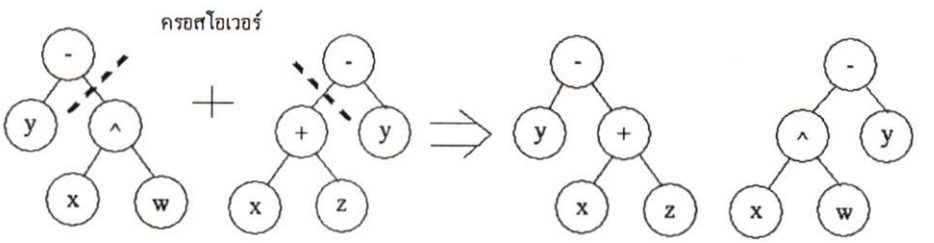
เมื่อ P_c คือความน่าจะเป็นในการครอสโอเวอร์ และ Popsize คือขนาดของประชากรในรุ่นนั้นๆ ตัวอย่างการครอสโอเวอร์แบบไบนารี เช่น ถ้าสุ่มตำแหน่งที่จะทำการครอสโอเวอร์ได้เป็นตำแหน่งที่ 4 การแลกเปลี่ยนส่วนของโครโมโซมจะเกิดขึ้นหลังตำแหน่งที่ 5 เรื่อยไปจนถึงตำแหน่งสุดท้าย เกิดโครโมโซมใหม่ขึ้นมา 2 โครโมโซม ดังแสดงในรูปที่ 3.11

รูปที่ 3.11 แสดงกระบวนการครอสโอเวอร์แบบไบนารี

ตัวอย่างการครอสโอเวอร์โครโมโซมแบบตัวอักษรหลังตำแหน่งที่ 3 แสดงดังรูปที่ 3.12

รูปที่ 3.12 แสดงกระบวนการครอสโอเวอร์แบบตัวอักษร

ตัวอย่างการครอสโอเวอร์แบบทรี แสดงดังรูปที่ 3.13



รูปที่ 3.13 การครอสโอเวอร์ของโครโมโซมแบบทรี

3.6.2.7 การมิวเตชัน (Mutation)

การมิวเตชันหรือการผ่าเหล่า เป็นลักษณะของการนำโครโมโซมเก่ามาสุ่มแก้ไขค่าบางค่า เช่น ทำให้ค่าของบางตำแหน่งเปลี่ยนไป โดยทำการกลับบิทเป็นค่าใหม่ในตำแหน่งบิทที่สุ่มได้ ตามค่าความน่าจะเป็นของการมิวเตชันในแต่ละบิท (Probability of Mutation : P_m) ที่กำหนด โดยการมิวเตชันจะทำการสุ่มค่า r ของแต่ละตำแหน่งบิทในแต่ละโครโมโซม โดยถ้าค่า r ณ ตำแหน่งของบิทใดเป็นไปดังสมการที่ (3.16)

$$r \leq P_m \tag{3.16}$$

ค่าของบิท ณ ตำแหน่งนั้นก็จะถูกทำมิวเตชัน จำนวนบิทที่จะถูกทำการมิวเตชันนั้นสามารถคำนวณได้ดังสมการที่ (3.17)

$$\text{จำนวนของการมิวเตชัน} = P_m \text{ x } Popsize \text{ x } L \qquad (3.17)$$

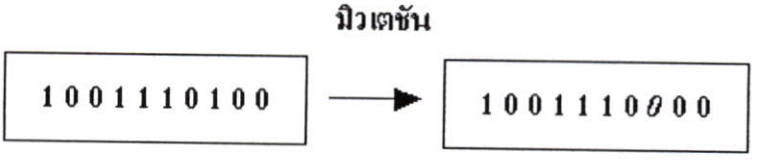
เมื่อ P_m คือ ความน่าจะเป็นของการมิวเตชัน

Popsize คือ ขนาดของประชากรในรุ่นนั้นๆ

L คือ ความยาวของโครโมโซม

ผลจากการมิวเตชัน ทำให้ได้โครโมโซมใหม่ที่มีรูปแบบของโครโมโซมแตกต่างจากเดิม ซึ่งมีโอกาสที่จะเป็นโครโมโซมที่ดีขึ้นหรือเลวลงก็ได้ หากโครโมโซมที่ได้ใหม่เป็นโครโมโซมที่เลวลง คือ มีค่าความเหมาะสมต่ำลง โครโมโซมที่ได้นี้ก็จะถูกคัดออกไปในขั้นตอนการคัดเลือก (Selection) นั่นเอง

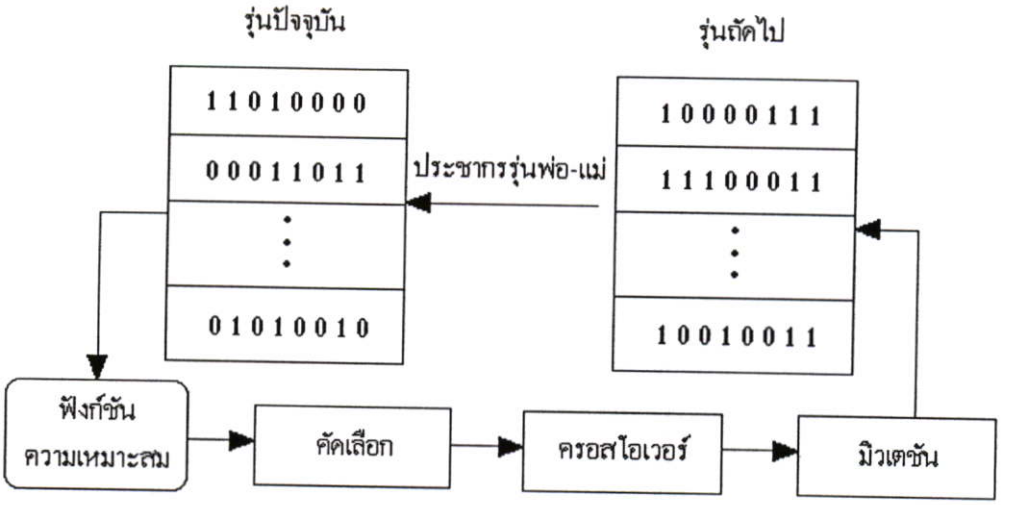
วัตถุประสงค์ของการมิวเตชันคือ เพื่อประกันการสูญหายของข้อมูล และเพื่อความหลากหลายของข้อมูล สำหรับไบนารีมิวเตชัน เป็นการปรับเปลี่ยนข้อมูล ณ ตำแหน่งที่กำหนดนั้น โดยเปลี่ยนข้อมูลจาก 0 เป็น 1 หรือกลับกัน ตัวอย่างของการทำมิวเตชันแสดงดังรูปที่ 3.14 งเป็นการสุ่มเลือกทำมิวเตชัน ณ ตำแหน่งที่ 8



รูปที่ 3.14 แสดงกระบวนการไบนารีมิวเตชัน

3.6.2.8 การสร้างประชากรรุ่นใหม่

ประชากรรุ่นใหม่เป็นกลุ่มโครโมโซมรุ่นลูกที่เกิดจากกระบวนการของวิวัฒนาการต่างๆ ทั้งหมด เริ่มตั้งแต่การวัดค่าความเหมาะสม ทำการคัดเลือก สุ่มเลือกเพื่อนำมาทำครอสโอเวอร์และมิวเตชันตามค่าความน่าจะเป็นที่ได้กำหนดไว้ ซึ่งเมื่อชุดโครโมโซมรุ่นลูกผ่านวิวัฒนาการต่างๆ ดังที่ได้กล่าวไปแล้วนั้นก็จะทำให้เกิดประชากรรุ่นใหม่ โดยที่ประชากรรุ่นใหม่นี้จะถูกถ่ายทอดกลายเป็นประชากรรุ่นเก่า สำหรับวิวัฒนาการรุ่นถัดไปเช่นเดียวกัน ซึ่งจะเรียกวิวัฒนาการนี้ว่า การถ่ายทอดแบบทั่วไป หรือรีโพรดักชันแบบทั่วไป (General Reproduction) ขบวนการทั้งหมดในการสร้างประชากรสามารถแสดงได้ดังรูปที่ 3.15



รูปที่ 3.15 แสดงกระบวนการสร้างประชากรในรุ่นถัดไป

3.6.2.9 การกำหนดค่าตัวแปรต่างๆ

การกำหนดค่าตัวแปรต่างๆ ในกระบวนการเจเนติกอัลกอริทึมนั้น สามารถแบ่งออกได้เป็น 2 กลุ่ม ได้แก่ ตัวแปรที่ใช้ควบคุมการทำงานและเงื่อนไขการสิ้นสุดการทำงาน

a) ตัวแปรที่ใช้ควบคุมการทำงาน ตัวแปรในกลุ่มนี้ได้แก่

1. การกำหนดขนาดของประชากร (Population Size) จำนวนขนาดของประชากรมีผลกระทบต่อประสิทธิภาพ ความเร็วในการค้นหาคำตอบ และการใช้ทรัพยากรของระบบ ถ้าจำนวนประชากรน้อยเกินไป อาจทำให้ได้คำตอบที่ขาดประสิทธิภาพในการแก้ปัญหานั้นๆได้ แต่หากจำนวนประชากรมีมากเกินไปแล้วจะส่งผลให้การทำงานเพื่อค้นหาคำตอบจะต้องใช้เวลาและทรัพยากรมากขึ้น ดังนั้นการกำหนดจำนวนขนาดของประชากรจะต้องมีความเหมาะสม [21]

2. การกำหนดค่าความน่าจะเป็นในการคัดเลือก (Probability of Selection) การสุ่มเลขเพื่อเข้าสู่การคัดเลือกโดยใช้แบบจำลองการหมุนวงล้อถ่วงน้ำหนัก (Roulette Wheel) ถ้าตัวเลขที่สุ่มได้ทำให้เกิดช่วงค่าที่แคบเกินไป หรือกว้างเกินไป อาจทำให้โครโมโซมที่ดีไม่ถูกคัดเลือก หรือทำให้เกิดการคัดเลือกโครโมโซมบางตัวซ้ำๆ แม้ว่าในรุ่นนั้นจะมีประชากรโครโมโซมอื่นๆ อีกก็ตาม

3. การกำหนดค่าความน่าจะเป็นในการครอสโอเวอร์ (Crossover Probability) ที่เหมาะสมเพื่อผลิตโครโมโซมที่มีความหลากหลายในประชากรรุ่นต่อไป การกำหนดค่าความน่าจะเป็นในการครอสโอเวอร์ที่เหมาะสมมีส่วนในการเพิ่มประสิทธิภาพในการค้นหาคำตอบ

4. การกำหนดค่าความน่าจะเป็นในการมิวเตชัน (Mutation Probability) ที่เหมาะสม ซึ่งแต่ละปัญหาก็จะต้องการค่าความน่าจะเป็นในการครอสโอเวอร์และมิวเตชันที่แตกต่างกันไป ดังนั้นจึงควรเลือกใช้ให้เหมาะสมกับแต่ละปัญหา เพื่อให้การค้นหาคำตอบมีประสิทธิภาพมากที่สุด

5. จำนวนรุ่น (Number of Generation) ในการค้นหาคำตอบ

b) เงื่อนไขการสิ้นสุดการทำงาน

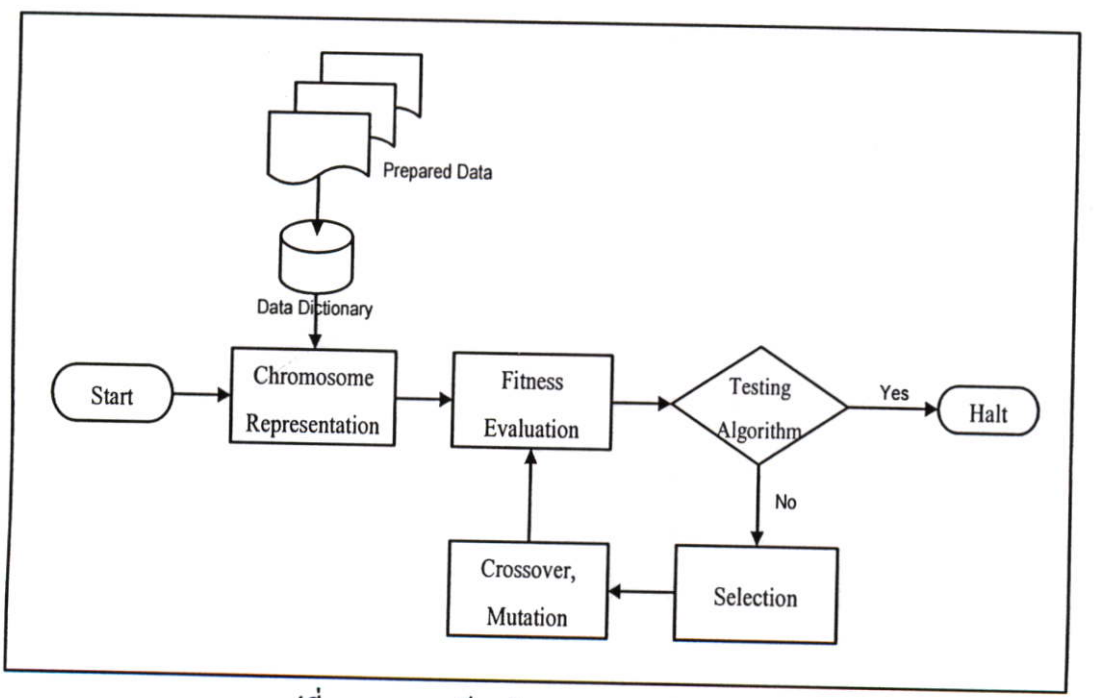
โดยปกติการแก้ไขปัญหาจะเสร็จสิ้นเมื่อได้คำตอบที่ดีที่สุด คือได้โครโมโซมที่มีค่าความเหมาะสมสูงสุด สามารถแก้ไขปัญหาหรือเป็นคำตอบที่ดีที่สุดของปัญหานั้นๆ หรือในแนวทางหนึ่ง สามารถกำหนดให้กระบวนการเจเนติกอัลกอริทึมสิ้นสุดการทำงานเมื่อถึงรุ่นสูงสุด (Max Generation) ที่กำหนดไว้ แล้วนำโครโมโซมที่มีค่าความเหมาะสมสูงสุดมาเป็นคำตอบที่ใกล้เคียง

กล่าวโดยสรุป เจเนติกอัลกอริทึมเป็นเทคนิคที่ใช้ในการค้นหาคำตอบ ซึ่งเลียนแบบมาจากกระบวนการวิวัฒนาการทางธรรมชาติที่นำมาประยุกต์ใช้กับคอมพิวเตอร์ เพื่อช่วยแก้ปัญหาในการหาคำตอบต่างๆ ซึ่งพื้นฐานการทำงานเบื้องต้นเป็นเจเนติกอัลกอริทึมแบบง่าย มีรูปแบบโครโมโซมเป็นแบบไบนารี การคัดเลือกใช้แบบจำลองการหมุนวงล้อถ่วงน้ำหนัก การครอสโอเวอร์เป็นการครอสโอเวอร์แบบหนึ่งจุด และมิวเตชันแบบไบนารี ซึ่งสามารถช่วยแก้ปัญหาในการค้นหาคำตอบให้แก่ระบบได้ ในการประยุกต์ใช้เจเนติกอัลกอริทึมกับปัญหาต่างๆ นั้นจะต้องมีการปรับปรุง เปลี่ยนแปลงในบางส่วน เช่น รูปแบบของโครโมโซม ฟังก์ชันความเหมาะสม หรือค่าตัวแปรต่างๆ เพื่อให้เข้ากับรูปแบบของปัญหา และเพื่อให้สามารถค้นหาคำตอบที่ดีที่สุดให้แก่ปัญหานั้น

บทที่ 4
ตัวกรองอีเมล์ขยะด้วยทฤษฎีเจเนติกอัลกอริทึม

ตัวกรองอีเมล์ขยะโดยใช้เจเนติกอัลกอริทึม มีวัตถุประสงค์เพื่อพัฒนาระบบกำจัดอีเมล์ขยะให้มีประสิทธิภาพมากยิ่งขึ้น โดยใช้กระบวนการเจเนติกอัลกอริทึมซึ่งเป็นกลไกที่เลียนแบบวิวัฒนาการของสิ่งมีชีวิตในธรรมชาติ ตัวกรองอีเมล์ขยะที่สร้างขึ้นนี้จะใช้ตัวดำเนินการทางเจเนติก เช่น การคัดเลือก การครอสโอเวอร์ และการมิวเตชัน เพื่อสร้างรูปแบบของอีเมล์ที่หลากหลายขึ้นจากอีเมล์ที่มีอยู่เดิม ซึ่งจะช่วยให้เจเนติกอัลกอริทึมสามารถเรียนรู้ได้ว่ารูปแบบของอีเมล์แบบใดเป็นอีเมล์ขยะหรืออีเมล์ดี ในวิทยานิพนธ์นี้ เริ่มต้นจากการออกแบบระบบตัวกรองอีเมล์ขยะ การเตรียมอีเมล์ดีและอีเมล์ขยะเพื่อให้อัลกอริทึมเรียนรู้ ซึ่งประกอบไปด้วยขั้นตอนย่อยคือการตัดคำ การจัดกลุ่มคำ และการเตรียมฐานข้อมูลของคำ เมื่อเตรียมข้อมูลเรียบร้อยแล้วจะเข้าสู่กระบวนการเจเนติกอัลกอริทึมเพื่อที่จะเรียนรู้รูปแบบของอีเมล์และดำเนินการจำแนกอีเมล์ต่อไป

ขั้นตอนการสร้างตัวกรองอีเมล์ขยะมี 3 ขั้นตอนหลัก คือ การเตรียมข้อมูล (Data Preparation) การสร้างฐานข้อมูลคำ (Creating Data Dictionary) และ การประยุกต์ใช้เจเนติกอัลกอริทึม (Adopting Genetic Algorithm) ซึ่งแสดงดังรูป 4.1



รูปที่ 4.1 แสดงบล็อกไดอะแกรมของระบบ

4.1 การเตรียมข้อมูล (Data Preparation)

1. อีเมล์ที่นำมาใช้ประกอบไปด้วยอีเมล์ขยะจำนวน 900 ฉบับจาก [3], [4] และอีก 100 ฉบับผู้เขียนได้รวบรวมเองจาก Hotmail และ Yahoo Mail และอีเมล์ดีจำนวน 1,000 ฉบับจาก [4]
2. แบ่งอีเมล์ออกเป็นชุดข้อมูลสำหรับเรียนรู้ (Training Set) จำนวน 1,600 ฉบับ และชุดข้อมูลสำหรับทดสอบ (Testing Set) จำนวน 400 ฉบับ
3. ตัดคำและนับความถี่ของคำ อาศัยหลักการใช้ช่องว่างระหว่างคำเป็นตัวแบ่งแยกคำออกจากกัน ในวิทยานิพนธ์นี้ได้ใช้เครื่องมือช่วยในการตัดคำและนับความถี่ของคำจาก [6] เมื่อตัดคำแล้ว จะคัดคำที่เป็น URL, คำที่อยู่ในส่วนหัวเรื่อง (Header) ของอีเมล์, คำที่ไม่มีความหมาย และตัวเลข ทิ้งไป โดยคำอื่นที่เหลือทั้งหมดต่อไปนี้จะถูกเรียกว่าเป็นคำสำคัญ (Keywords)

4.2 การสร้างฐานข้อมูลของคำ (Creating Data dictionary)

ฐานข้อมูลคำประกอบไปด้วย

1. คำสำคัญ (Keywords)และกลุ่มของคำสำคัญ
   เนื่องจากเราต้องการสร้างโครโมโซมโดยกำหนดให้แต่ละยีนของโครโมโซมเก็บคำสำคัญในอีเมล์ซึ่งแยกเป็นกลุ่มๆ เอาไว้ ดังนั้นจึงต้องมีการจัดกลุ่มให้แก่คำสำคัญทั้งหมด โดยการแบ่งกลุ่มมีหลักการคือ จะรวบรวมคำสำคัญที่มีความหมายใกล้เคียงกัน หรือมาจากรากศัพท์เดียวกัน ให้อยู่ตระกูลเดียวกัน[7] โดยเกณฑ์ในการแบ่งกลุ่มนั้น แบ่งตามการสำรวจกลุ่มของอีเมล์โดยไมโครซอฟต์ [2] และจากการระบุกลุ่มของอีเมล์จากคลังอีเมล์ [4] รวมกับการพิจารณาเนื้อหาในอีเมล์เพิ่มเติม ทำให้สามารถแบ่งคำสำคัญในอีเมล์ออกเป็น 8 กลุ่มหลัก และตัวอย่างของคำสำคัญในกลุ่ม แสดงดังตารางที่ 4.1
2. ความน่าจะเป็นของคำสำคัญ
   การหาความน่าจะเป็นของคำสำคัญทำเพื่อเตรียมความน่าจะเป็นเพื่อนำไปใช้ในการสร้างรูปแบบของโครโมโซม โดยจะอธิบายรายละเอียดของการคำนวณในหัวข้อถัดไป

ตารางที่ 4.1 แสดงตัวอย่างของคำสำคัญในแต่ละกลุ่ม

กลุ่ม	ชื่อกลุ่ม	ตัวอย่างของคำสำคัญในกลุ่ม
G1	**Adult** - กลุ่มคำสำคัญที่เกี่ยวกับสื่อลามกอนาจาร	sex, adult, viagra, hardcore, webcam, teen, girl, nude, gay, xxx, erection, etc.
G2	**Business and Financial** - กลุ่มคำสำคัญที่เกี่ยวกับธุรกิจ และการเงิน	enterprise, share, holder, investor, strategy, mortgage, obligate, fund, refund, loan, etc.
G3	**Commercial** - กลุ่มของคำสำคัญที่เกี่ยวกับการค้าขาย และข้อเสนอชวนเชื่อ	free, special, retail, resell, inexpensive, cartier, louis, buy, etc.
G4	**Medicine, Diet and Beauty** - กลุ่มของคำสำคัญที่เกี่ยวกับการขายยา และการบริการเพื่อรูปร่างและความสวยงาม	diet, fat, herb, weight, lose, age, medicine, nature, health, prescription, etc.
G5	**Traveling and Gambling** - กลุ่มของคำสำคัญที่เกี่ยวกับการท่องเที่ยว การพักผ่อนหย่อนใจ การพนัน และการเสี่ยงดวง	hotel, reserve, travel, trip, holiday, win, bonus, casino, extra, gambling, etc.
G6	**Internet and Home-Based Business** - กลุ่มของคำสำคัญที่เกี่ยวกับธุรกิจทางอินเตอร์เน็ต และการทำธุรกิจผ่านอินเตอร์เน็ตที่บ้าน	internet, home, based, email, earn, subscriber, home based etc.
G7	**Political, Social and Religion** - กลุ่มของคำสำคัญที่เกี่ยวกับการเมือง สังคม ศาสนา และความเชื่อ	policy, political, challenge, campaign, public, church, etc.
G8	**Common** - กลุ่มของคำสำคัญที่มีการใช้ร่วมกันอยู่ในหลายกลุ่ม	now, tell, today, year, monday, texas etc.

4.3 การประยุกต์ใช้เจเนติกอัลกอริทึม (Adopting Genetic Algorithm)

4.3.1 การกำหนดรูปแบบของโครโมโซม (Representation of a Chromosome)

การสร้างโครโมโซมแม่แบบคือการกำหนดรูปแบบของโครโมโซมจากการเลียนแบบรูปแบบของทั้งอีเมล์ขยะ และอีเมล์ดี โดยตัวดำเนินการทางเจเนติกจะทำการสร้างรูปแบบใหม่ๆ ขึ้นเพื่อให้อัลกอริทึมได้เรียนรู้ว่ารูปแบบใด (Template) มีความคล้ายคลึงกับอีเมล์แบบใดรูปแบบของโครโมโซมหรือโครโมโซมแม่แบบเกิดจากการประกอบกันของค่าความน่าจะเป็นเฉลี่ยในแต่ละกลุ่มซึ่งอยู่ในรูปของเลขฐานสองดังรูป 4.2

Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	Gene8
$P_{avg}(G1)$ = 0001100110	$P_{avg}(G2)$ = 0000000000	$P_{avg}(G3)$ = 0000110101	$P_{avg}(G4)$ = 0000000000	$P_{avg}(G5)$ = 0000000000	$P_{avg}(G6)$ = 0000000000	$P_{avg}(G7)$ = 0000000000	$P_{avg}(G8)$ = 0000000000

รูปที่ 4.2 แสดงรูปแบบของโครโมโซม

ค่าความน่าจะเป็นเฉลี่ยคำนวณได้จากความน่าจะเป็นของคำสำคัญใดๆในกลุ่มหารด้วยจำนวนคำในกลุ่มดังสมการ (4.1)

$$P_{avg}(G) = \sum_{i=1}^{\text{total words}} \frac{\text{จำนวนครั้งที่พบคำสำคัญ } i \times \text{ความน่าจะเป็นของคำสำคัญ } i}{\text{จำนวนคำสำคัญทั้งหมดในกลุ่ม (total words)}} \quad (4.1)$$

โดยที่ $P(word_i)$ คำนวณได้จากสมการ (4.2)

$$P(word_i) = \frac{\text{จำนวนครั้งที่พบคำสำคัญ}}{\text{จำนวนคำสำคัญทั้งหมด}} \quad (4.2)$$

ตัวอย่างเช่น คำสำคัญว่า “sex” มีจำนวนครั้งที่พบ 457 ครั้ง จากความถี่ที่ปรากฏของคำทุกคำรวมกัน 1,273 ครั้ง จะได้ว่า P(sex) = 457/1,273 = 0.359 เป็นต้น

และถ้ากำหนดให้อีเมล์ 1 ฉบับ เมื่อผ่านขั้นตอนการตัดคำแล้ว พบว่ามีคำสำคัญว่า “sex”, “adult” ซึ่งเป็นคำสำคัญในกลุ่ม G1 และคำว่า “free”, “special” ซึ่งคำสำคัญในกลุ่ม G3 โดยคำสำคัญ “sex”, “adult”, “free” และ “special” มีค่าความน่าจะเป็น 0.395, 0.005, 0.230 และ 0.090 ตามลำดับ การสร้างโครโมโซมสำหรับอีเมล์ฉบับนี้ ทำได้โดยการหาความน่าจะเป็นของคำสำคัญแต่ละคำแล้วคำนวณความน่าจะเป็นเฉลี่ยของกลุ่ม สำหรับอีเมล์ฉบับนี้มีผลการคำนวณเป็นไปดังตารางที่ 4.2

ตารางที่ 4.2 แสดงการคิดค่าเฉลี่ยความน่าจะเป็นของคำสำคัญในอีเมล์ตัวอย่าง

Group	Word	$P(word_i)$	Probability of Gene
G1	sex	0.395	((1x0.395)+(1x0.005)/2) = 0.200
G1	adult	0.005	
G3	free	0.230	((1x0.230)+(1x0.090)/2) = 0.160
G3	special	0.090	

เนื่องจากค่าความน่าจะเป็นของคำสำคัญที่ต่ำสุดในฐานข้อมูลคือ 0.002 และสูงสุดคือ 1.000 ซึ่งในวิทยานิพนธ์นี้ เราพิจารณาใช้ค่าความแตกต่างด้วยค่าจุดทศนิยม 3 ตำแหน่ง ค่าในช่วง 0.004 – 1.000 จะพบว่ามีค่าที่เป็นจำนวนเต็มทั้งหมด 998 ค่า ซึ่งจะต้องนำมาแปลงเป็นเลขฐานสองเพื่อจะนำไปใช้ในกระบวนการเจเนติกอัลกอริทึม ทำให้เราจึงต้องใช้เลขฐานสองจำนวน 10 บิท ซึ่งมี 1,024 ค่าในการแสดงค่า แต่ก่อนอื่นเราต้องทำการปรับช่วงของค่า 998 ค่า เพื่อแสดงด้วยเลขฐานสอง 10 บิทเสียก่อน ซึ่งสามารถทำได้ดังสมการที่ (4.3)

$$\text{ความน่าจะเป็นที่ปรับแล้ว} = \left[\left(\frac{\text{ความน่าจะเป็นที่นำมาปรับ} - \text{ความน่าจะเป็นต่ำสุด}}{\text{ความน่าจะเป็นสูงสุด} - \text{ความน่าจะเป็นต่ำสุด}}\right) \times 1{,}024\right] + 1 \qquad (4.3)$$

ตัวอย่างเช่น ค่าความน่าจะเป็นที่นำมาปรับคือ 0.200 เมื่อคำนวณตามสมการ (4.3) จะได้ว่า

$$\text{ความน่าจะเป็นที่ปรับแล้ว} = [(0.200\text{-}0.004/1.000\text{-}0.004)*1024]+1$$
$$= 202_{10}$$
$$= (0001100110)_2$$

เมื่อนำทุกยีนมาคำนวณตามสมการที่ 4.2 จะได้ค่าของยีนที่แปลงจากเลขฐานสิบเป็นเลขฐานสองดังแสดงในตารางที่ 4.3

ตารางที่ 4.3 แสดงการแปลงค่าความน่าจะเป็นก่อนปรับและหลังปรับและค่าความน่าจะเป็นหลังปรับเมื่อแปลงเป็นเลขฐานสอง

	ค่าความน่าจะเป็นก่อนปรับ	ค่าความน่าจะเป็นหลังปรับ	ค่าความน่าจะเป็นหลังปรับ(ฐานสอง)
Gene1	0.200	202	0001100110
Gene2	0	0	0000000000
Gene3	0.106	105	0000110101
Gene4	0	0	0000000000
Gene5	0	0	0000000000
Gene6	0	0	0000000000
Gene7	0	0	0000000000
Gene8	0	0	0000000000

4.3.2 การประเมินค่าความเหมาะสม (Fitness Evaluation)

คือการหาชุดของโครโมโซมที่เหมาะสมเพื่อนำไปใช้ในกระบวนการจำแนกอีเมล์ โดยเริ่มจากการกำหนดฟังก์ชันความเหมาะสม (Defining Fitness Function) โดยมีแนวคิดที่ว่าฟังก์ชันที่กำหนดขึ้นมานั้นต้องมีความสามารถในการคัดเลือกโครโมโซมที่ดี และก่อให้เกิดประสิทธิภาพแก่ระบบให้มากที่สุด การกำหนดฟังก์ชันความเหมาะสม (Fitness Function) สำหรับงานวิทยานิพนธ์นี้ ได้กำหนดให้ค่าความเหมาะสมของแต่ละโครโมโซมแม่แบบคือค่าที่บ่งบอกความแม่นยำในการจำแนกอีเมล์ทดสอบ โดยนำชุดข้อมูลสำหรับเรียนรู้มาเป็นตัวทดสอบกับโครโมโซมแม่แบบ (Template) ว่าโครโมโซมแม่แบบแต่ละตัวมีความแม่นยำมากน้อยเท่าไร ซึ่งมีขั้นตอนการทำงานตามชุดคำสั่งเทียม (Pseudo Code) ที่แสดงได้ดังรูปที่ 4.3

```
แต่ละ Template_i  (i = 1 ~ จำนวนของ Template)
  แต่ละ Training_j  (j= 1~ จำนวนของ Training set)
      ถ้า Class ของ Template_i = Spam แล้ว /* Template_i ทำนายว่า Training_j เป็น Spam*/
        ถ้า Training_j = Spam แล้ว
          สถิติการ ClassifySpamAsSpam ของ Template_i ++
        ถ้า Training_j = Ham แล้ว
          สถิติการ ClassifyHamAsSpam ของ Template_i ++
      ถ้า Class ของ Template_i = Ham แล้ว /* Template_i ทำนายว่า Training_j เป็น Ham*/
        ถ้า Training_j = Spam แล้ว
          สถิติการ ClassifySpamAsHam ของ Template_i ++
        ถ้า Training_j = Ham แล้ว
          สถิติการ ClassifyHamAsHam ของ Template_i ++
```

รูปที่ 4.3 แสดงการเก็บผลการจำแนกของแต่ละแม่แบบ (Template) สำหรับใช้คำนวณหาค่าความเหมาะสม

ค่าที่ได้จากการนับข้างต้นจะนำมาใช้ในการคำนวณค่าความเหมาะสม (Fitness Value) ซึ่งเป็นไปดังสมการที่ (4.4)

$$\text{ค่าความเหมาะสม} = \left(N_{S\to S} + N_{H\to H}\right) - \left(N_{S\to H} + N_{H\to S}\right) \qquad (4.4)$$

เมื่อ $N_{S\to S}$ คือ จำนวนการทายอีเมล์ขยะเป็นอีเมล์ขยะ

$N_{H\to H}$ คือ จำนวนการทายอีเมล์ดีเป็นอีเมล์ดี

$N_{S\to H}$ คือ จำนวนการทายอีเมล์ขยะเป็นอีเมล์ดี

$N_{H\to S}$ คือ จำนวนการทายอีเมล์ดีเป็นเมล์ขยะ

ค่าความเหมาะสมของโครโมโซมแม่แบบทั้งหมดหลังจากคำนวณเสร็จแล้วแสดงดังตารางที่4.4

ตารางที่ 4.4 แสดงตัวอย่างโครโมโซมแม่แบบ และค่าความเหมาะสมของโครโมโซม

ลำดับที่	โครโมโซม	ค่าความเหมาะสม
1	[0100101100,1100101000,0000000000,0000000000, 1110010000, 0000000000, 1010001111,0000000000]	15
2	[0001011010, 0011001100, 0000000000, 000000000, 1010111001, 0000000000, 1010100000, 0000000000]	70
3	[0000000000, 000000010, 1001001011, 1000000000, 0000000000,0100110011, 0001000100, 1000100100]	2
4	[0000000000,1010001111, 1100001100,0000000000, 0000111100, 0000000000, 0011000100, 0100011000]	40
5	[0010000000, 1000000000, 0000000000,1111000000, 0011001110, 0101000000,1100110011, 0000000000]	83
…	…	…
1,600	[0000000000, 0011000011,0000000000,0101001111, 0000000000,0000000000,0000000000,0000000000]	25

4.3.3 การคัดเลือกประชากร (Selection of Population)

คือการคัดเลือกโครโมโซมที่เหมาะสมไว้สำหรับใช้เป็น โครโมโซมแม่แบบ(Template) ในรุ่นถัดไป โดยในรอบแรกจะต้องกำหนดประชากรตั้งต้นขึ้นมาจำนวนหนึ่ง ซึ่งได้จากการนำชุดข้อมูลสำหรับเรียนรู้ มาสร้างเป็นโครโมโซมแม่แบบ(Template) โดยในวิทยานิพนธ์นี้ได้กำหนดโครโมโซมสำหรับเป็นประชากรตั้งต้น 1,600 ตัว จากนั้นกำหนดจำนวน และเลือกโครโมโซมพ่อแม่ที่จะใช้ในการครอสโอเวอร์ขึ้นมา โดยสามารถกำหนดจำนวนโครโมโซมพ่อแม่เป็นเปอร์เซ็นต์เมื่อคิดจากประชากรตั้งต้นได้ เช่น กำหนดโครโมโซมพ่อแม่ 10% หมายความว่า จากโครโมโซมตั้งต้น 1,600 ตัว จะใช้โครโมโซมพ่อแม่ 160 คู่ ในการครอสโอเวอร์ในรอบถัดไป ภายหลังจากการครอสโอเวอร์ จะได้โครโมโซมใหม่ขึ้นมาจำนวน 160 คู่ หรือ 320 ตัว เมื่อรวมกับโครโมโซมตั้งต้นจะได้ 1,920 ตัว จากนั้นจะต้องทำการคัดเลือก

โครโมโซมที่เหมาะสมกว่าเอาไว้เพียง 1,600 ตัว เท่ากับจำนวนโครโมโซมตั้งต้นโดยการคัดเลือกนั้น จะใช้วงล้อถ่วงน้ำหนักเป็นตัวคัดเลือกว่าโครโมโซมไหนจะผ่านเข้าสู่กระบวนการเจเนติกในรุ่นถัดไป (อ่านรายละเอียดของการคัดเลือกแบบวงล้อถ่วงน้ำหนักได้ในบทที่ 3)

4.3.4 การครอสโอเวอร์และการมิวเตชัน (Crossover and Mutation)

การครอสโอเวอร์ทำขึ้นเพื่อสร้างประชากรรุ่นใหม่ให้มีความหลากหลายขึ้นในระบบ ในวิทยานิพนธ์นี้ ใช้การครอสโอเวอร์แบบหลายจุด (Multiple-Point Crossover) ซึ่งเราสามารถกำหนดจำนวนบิทที่ต้องการให้เกิดการครอสโอเวอร์เป็นเปอร์เซ็นต์ หรือจะทำการสุ่มเลือกจำนวนบิทที่จะนำมาครอสโอเวอร์กันก็ได้ โดยการครอสจะเกิดขึ้นเฉพาะยีนในกลุ่มเดียวกันเท่านั้น ไม่สามารถกระทำข้ามกลุ่มได้ ส่วนการมิวเตชันทำเพื่อประกันการสูญหาย และเพื่อความหลากหลายของข้อมูล สำหรับไบนารีมิวเตชัน เป็นการปรับเปลี่ยนข้อมูล ณ ตำแหน่งที่กำหนดนั้น โดยเปลี่ยนข้อมูลจาก 0 เป็น 1 สำหรับวิทยานิพนธ์นี้ เราสามารถกำหนดได้ว่าจะให้มีการมิวเตชันเกิดขึ้นหลังจากผ่านกระบวนการทางเจเนติกไปกี่รุ่น และจะมิวเตชันกี่เปอร์เซ็นต์ของบิททั้งหมดในโครโมโซม โดยจะสุ่มเลือกโครโมโซมที่จะมิวเตชัน จากนั้นจึงสุ่มเลือกบิตที่จะทำมิวเตชัน โดยจะทำการกลับบิตเพียงบิตเดียวเพื่อให้ไม่เกิดการกลายพันธุ์ของโครโมโซมมากเกินไป

เมื่อสิ้นสุดการดำเนินการทางเจเนติกแล้ว จะได้โครโมโซมที่ผ่านการคัดเลือกในรอบนี้มาจำนวน 1,000 โครโมโซม ซึ่งโครโมเหล่านี้จะถูกเรียกว่า โครโมโซมแม่แบบ (Template) ซึ่งจะมี 1,000 แ ม่แบบ แม่แบบอีเมล์ขยะเหล่านี้เปรียบเสมือนชุดของกฎ (Rule set) ตัวอย่างของโครโมโซมแม่แบบที่ได้แสดงดังรูปที่ 4.3 ซึ่งจะนำไปใช้ในขบวนการทดสอบเพื่อการจำแนกอีเมล์ต่อไป

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	Gene8
Template 1	0100101100	1100101000	0000000000	0000000000	1110010000	0000000000	1010001111	0000000000

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	Gene8
Template 2	0001011010	0011001100	0000000000	0000000000	1010111001	0000000000	1010100000	0000000000

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	Gene8
Template 3	0000000000	0000000010	1001001011	1000000000	0000000000	0100110011	0001000100	1000100100

.
.
.

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	Gene8
Template 1,600	0000000000	0011000011	0000000000	0101001111	0000000000	0000000000	0000000000	0000000000

รูปที่ 4.4 แสดงแม่แบบของอีเมล์ทั้งหมดที่ได้จากกระบวนการเจเนติกอัลกอริทึมในรุ่นนี้

จากรูปที่ 4.4 จะเห็นว่า แม่แบบจะมีค่าในแต่ละยีนแตกต่างกันเนื่องจากแม่แบบแต่ละตัวมาจากลักษณะของอีเมล์ที่ต่างกัน ซึ่งลักษณะเหล่านี้สามารถนำไปใช้ในการบ่งชี้รูปแบบ (Pattern) ของอีเมล์ที่มีลักษณะแตกต่างกันได้

4.3.5 การทดสอบประสิทธิภาพของระบบ (Testing Algorithm)

การทดสอบระบบว่าเป็นที่พอใจหรือไม่นั้นวัดได้จากความถูกต้องของการจำแนกอีเมล์ว่าเป็นอีเมล์ใดๆ เป็นอีเมล์ขยะหรืออีเมล์ดี โดยใช้ชุดทดสอบที่แยกไว้จำนวน 400 อีเมล์ โดยนำอีเมล์ดังกล่าวมาสร้างรูปแบบของโครโมโซมตามหัวข้อ 4.1.2 จากนั้นนำค่าความน่าจะเป็นเฉลี่ยของแต่ละยีนของโครโมโซมที่แปลงจากอีเมล์ดังกล่าวมาเปรียบเทียบกับค่าความน่าจะเป็นเฉลี่ยในยีนของโครโมโซมแม่แบบทีละโครโมโซม โดยแต่ละโครโมโซมแม่แบบจะทำการเปรียบเทียบไปทีละยีน ถ้าตรงตามเงื่อนไข ยีนนั้นจะได้ 1 คะแนน โดยจะทำการเปรียบเทียบเช่นนี้ไปจนครบ 8 ยีน เมื่อรวมคะแนนแล้วพบว่าจำนวนยีนที่มีเงื่อนไขตรงกัน มากกว่าหรือเท่ากับค่า Gene Threshold แสดงว่าอีเมล์ที่เข้ามาจำแนกนี้ มีรูปแบบที่คล้ายคลึงกันกับโครโมโซมแม่แบบนี้ เพราะฉะนั้น โครโมโซมแม่แบบนี้จะให้คะแนนอีเมล์ฉบับนี้ว่ามีชนิดเดียวกับโครโมโซมแม่แบบนี้ แต่ถ้าไม่ตรงตามเงื่อนไขก็จะให้คะแนนอีเมล์ว่ามีชนิดตรงกันข้ามกับโครโมโซมแม่แบบนี้ ซึ่งมีขั้นตอนการทำงานตามชุดคำสั่งเทียม (Pseudo Code) ที่แสดงได้ ดังรูปที่ 4.5

```
แต่ละ Template_i  (i = 1 ~ จำนวนของ Template)
  แต่ละ Testing_j  (j= 1~ จำนวนของ Testing set)
    แต่ละ Gene_k  (k = 1~ 8)
      ถ้า (ค่าของ Gene_k ของ Testing_j ) ≥ (ค่าของ Gene_k ของ Template_i ) แล้ว
        (GeneMatchCount ของ Testing_j ) ++
      ถ้า (GeneMatchCount ของ Testing_j ) ≥ Gene Threshold แล้ว
        ถ้า Class ของ Template_i = Spam แล้ว /* Template_i ทำนายว่า Testing_j เป็น Spam*/
          ถ้า Testing_j = Spam แล้ว
            สถิติการ ClassifySpamAsSpam ของ Template_i ++
          ถ้า Testing_j = Ham แล้ว
            สถิติการ ClassifyHamAsSpam ของ Template_i ++
        ถ้า Class ของ Template_i = Ham แล้ว /* Template_i ทำนายว่า Testing_j เป็น Ham*/
          ถ้า Testing_j = Spam แล้ว
            สถิติการ ClassifySpamAsHam ของ Template_i ++
          ถ้า Testing_j = Ham แล้ว
            สถิติการ ClassifyHamAsHam ของ Template_i ++
```

รูปที่ 4.5 แสดงการจำแนกอีเมล์โดยพิจารณาเปรียบเทียบจากค่าความน่าจะเป็นเฉลี่ยในยีน

เมื่อ $Template_i$ คือ Template ที่กำลังพิจารณา

$Testing_j$ คือ อีเมล์ที่กำลังพิจารณา

Gene Threshold คือจำนวนขั้นต่ำที่กำหนดไว้ให้ยีนของอีเมล์ที่นำมาจำแนกตรงกับยีนของโครโมโซมแม่แบบ เช่น ถ้า Gene Threshold =3 หมายความว่า ความน่าจะเป็นเฉลี่ยของยีนในอีเมล์ที่นำมาจำแนกต้องตรงตามเงื่อนไขกับค่าความน่าจะเป็นเฉลี่ยของยีนในโครโมโซมแม่แบบ 3 ยีนขึ้นไปโดยโครโมโซมแม่แบบนี้จะให้คะแนนอีเมล์ที่นำมาจำแนกนี้ว่ามีความคล้ายคลึงกับโครโมโซมแม่แบบ แต่ถ้าไม่ตรงตามเงื่อนไข ก็จะให้คะแนนที่ตรงกันข้ามแก่อีเมล์นี้แทน

จากนั้นนำอีเมล์ที่ต้องการจำแนกไปเปรียบเทียบกับโครโมโซมแม่แบบทั้งหมด 1,600 แม่แบบ แล้วรวมคะแนนการของการจำแนก เปรียบเทียบคะแนนระหว่างคะแนนอีเมล์ขยะและคะแนนอีเมล์ดี ก็จะสามารถทำนายว่าอีเมล์ที่นำมาจำแนกเป็นอีเมล์ชนิดใด

ถ้าผลลัพธ์ของการจำแนกอีเมล์ใน Generation นี้มีความเหมาะสมและเป็นที่พอใจ ซึ่งวัดจากค่า Accuracy, Recall และ Precision จึงทำการหยุดกระบวนการเจเนติกอัลกอริทึม ถ้ายังไม่

เป็นที่พอใจก็ดำเนินการทางเจเนติก เช่นการครอสโอเวอร์ การมิวเตชันใน Generation ถัดไป จนกระทั่งได้ผลลัพธ์ที่เหมาะสมและเป็นที่พอใจ

บทที่ 5
ผลการทดลองและการวิเคราะห์

5.1 ขั้นตอนการทดลองตัวกรองอีเมล์ขยะโดยใช้เจเนติกอัลกอริทึม

ขั้นตอนในการกรองอีเมล์ขยะโดยใช้เจติกอัลกอริทึมมีส่วนหลักอยู่ 3 ส่วนคือ การเตรียมข้อมูล (Data Preparation) การสร้างฐานข้อมูลคำ (Creating Data Dictionary) และ การประยุกต์ใช้เจเนติกอัลกอริทึม (Adopting Genetic Algorithm)

1. การเตรียมอีเมล์สำหรับงานวิทยานิพนธ์นี้ ใช้อีเมล์ขยะจำนวน 1,000 ฉบับ และใช้อีเมล์ดีจำนวน 1,000 ฉบับ และ รวมทั้งสิ้น 2,000 ฉบับ โดยทำการทดลองแบบสลับชุดข้อมูล 5 ส่วน (5-folds Cross Validation) โดยแบ่งชุดข้อมูลออกเป็น 5 ส่วนเท่าๆกัน แล้วทำการทดลองทั้งสิ้น 5 ครั้ง แต่ละครั้งจะใช้ชุดข้อมูล 4 ส่วนเป็นชุดข้อมูลฝึกหัดและใช้ชุดข้อมูล 1 ส่วนเป็นชุดข้อมูลทดสอบ จากนั้นอีเมล์จะถูกนำไปผ่านกระบวนการตัดคำ (ดูรายละเอียดของการตัดคำได้ในบทที่ 4) ก่อนเข้าสู่กระบวนการเรียนรู้ของตัวกรองต่อไป
2. กระบวนการเรียนรู้ของตัวกรอง ประกอบไปด้วย การแปลงอีเมล์ให้เป็นโครโมโซมการหาค่าความน่าจะเป็นให้ยีนในโครโมโซม การดำเนินการทางเจเนติกอัลกอริทึม ซึ่งเป็นการทำเพื่อให้ได้รูปแบบของอีเมล์ที่หลากหลายขึ้นมา แล้วคัดเลือกโครโมโซมที่เหมาะสมมาสร้างเป็นโครโมโซมแม่แบบเพื่อใช้จำแนกอีเมล์ต่อไป (สามารถดูรายละเอียดของแต่ละส่วนได้ในบทที่ 4)
3. กระบวนการจำแนกอีเมล์ ประกอบไปด้วย การแปลงอีเมล์ทดสอบให้เป็นโครโมโซมแล้วหาค่าความน่าจะเป็นให้ยีนในโครโมโซม เพื่อที่จะนำค่าความน่าจะเป็นของยีนในโครโซมที่แปลงมาจากอีเมล์ทดสอบไปเปรียบเทียบกับค่าความน่าจะเป็นของยีนในโครโมโซมซึ่งเป็นแม่แบบซึ่งได้จากกระบวนการเรียนรู้ของตัวกรอง (สามารถดูรายละเอียดของการจำแนกได้ในบทที่ 4)

5.2 การปรับพารามิเตอร์ต่างๆ ในงานวิทยานิพนธ์

เนื่องจากตัวดำเนินการในเจเนติกอัลกอริทึมประกอบไปด้วย การคัดเลือก การครอสโอเวอร์ และการมิวเตชัน ซึ่งแต่ละตัวดำเนินการก็สามารถทำหลายวิธี เช่น การคัดเลือก มีแบบการคัดเลือกแบบจัดลำดับ (Ranking) การคัดเลือกแบบวงล้อถ่วงน้ำหนัก (Roulette Wheel) หรือในครอสโอเวอร์ มีเปอร์เซ็นต์ของการครอสเป็นต้น สิ่งเหล่านี้เป็นพารามิเตอร์ที่เราจะต้องหาค่าที่เหมาะสมเพื่อให้ได้ผลลัพธ์ที่ดีที่สุด โดยผลลัพธ์ที่ดีที่สุดสำหรับวิทยานิพนธ์นี้จะเปรียบเทียบจากค่า Accuracy ซึ่งเป็นค่าที่บอกความถูกต้องของการจำแนกอีเมล์ขยะและอีเมล์ดี

Recall คือค่าที่บอกความถูกต้องในการจำแนกอีเมล์ขยะ และ Precision คือค่าที่บอกความถูกต้องในการจำแนกอีเมล์ดี ซึ่งทั้งสามค่าสามารถคิดได้จากสมการ (5.1), (5.2) และ (5.3) [22]

$$\text{Accuracy} = \frac{N_{S\to S} + N_{H\to H}}{N_{S\to S} + N_{H\to H} + N_{S\to H} + N_{H\to S}} \tag{5.1}$$

$$\text{Precision} = \frac{N_{S\to S}}{N_{S\to S} + N_{H\to S}} \tag{5.2}$$

$$\text{Recall} = \frac{N_{S\to S}}{N_{S\to S} + N_{S\to H}} \tag{5.3}$$

เมื่อ $N_{S\to S}$ คือ จำนวนของอีเมล์ขยะที่ถูกทายว่าเป็นอีเมล์ขยะ

$N_{H\to H}$ คือ จำนวนของอีเมล์ดีที่ถูกทายว่าเป็นอีเมล์ดี

$N_{S\to H}$ คือ จำนวนของอีเมล์ขยะที่ถูกทายเป็นเมล์ดี

$N_{H\to S}$ คือ จำนวนของอีเมล์ดีที่ถูกทายเป็นเมล์ขยะ

ในงานวิทยานิพนธ์ได้ทำการทดลองเพื่อหาค่าพารามิเตอร์ที่เหมาะสมของแต่ละตัวดำเนินการ โดยมีพารามิเตอร์กลางที่ได้จากการทดลองอย่างคร่าวๆ ดังนี้

- จำนวนคู่โครโมโซมพ่อแม่ 10%
- วิธีการคัดเลือกแบบวงล้อถ่วงน้ำหนัก
- การครอสโอเวอร์แบบสุ่ม
- ไม่ใช้มิวเตชันพารามิเตอร์
- ยีนเธร์โชว์ 5 ยีน

จากนั้นนำพารามิเตอร์กลางเหล่านี้ไปใช้ในการหาพารามิเตอร์ที่เหมาะสมทีละตัว โดย Fix พารามิเตอร์ตัวอื่นๆ เอาไว้ โดยมีรายละเอียดของการหาพารามเตอร์ที่เหมาะสมแต่ละตัวดังต่อไปนี้

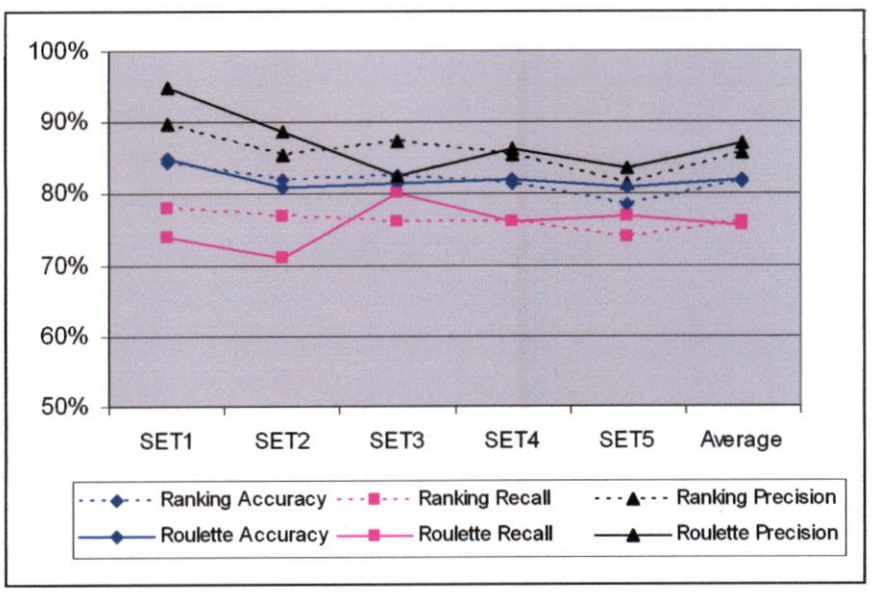
5.2.1 การเปรียบเทียบวิธีการคัดเลือกโครโมโซมระหว่างการคัดเลือกแบบวงล้อถ่วงน้ำหนัก (Roulette Wheel) และการคัดเลือกแบบจัดลำดับ (Ranking)

งานวิทยานิพนธ์นี้ ได้ทำการเปรียบเทียบการคัดเลือกแบบวงล้อถ่วงน้ำหนักและการจัดลำดับกับชุดข้อมูล 5 ชุด ซึ่งได้ผลการทดลองดังแสดงในตารางที่ 5.1

ตารางที่ 5.1 แสดงการเปรียบเทียบค่า Accuracy, Recall และ Precision ระหว่างการคัดเลือกแบบวงล้อถ่วงน้ำหนัก (Roulette Wheel) และการคัดเลือกแบบจัดลำดับ (Ranking)

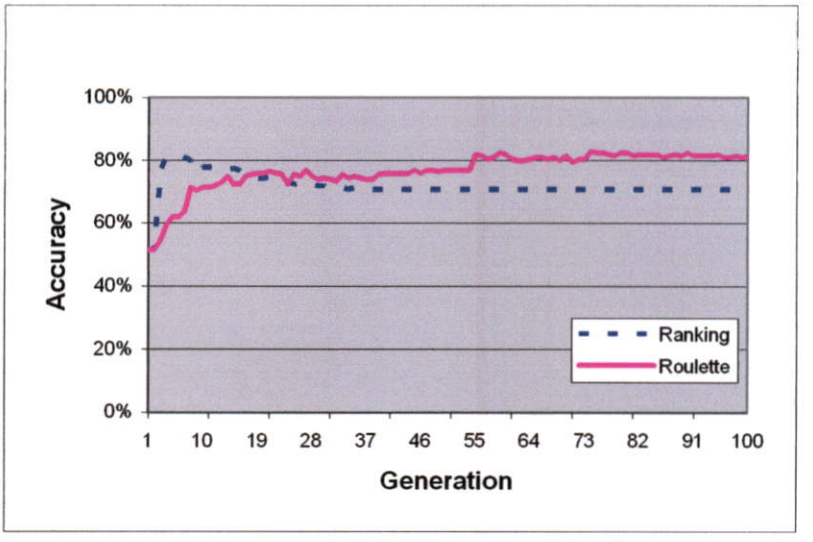
	Ranking			Roulette Wheel		
	Accuracy	**Recall**	**Precision**	**Accuracy**	**Recall**	**Precision**
SET1	84.50%	78.00%	89.66%	85.00%	74.00%	94.87%
SET2	82.00%	77.00%	85.56%	81.00%	71.00%	88.75%
SET3	82.50%	76.00%	87.36%	81.50%	80.00%	82.47%
SET4	81.50%	76.00%	85.39%	82.00%	76.00%	86.36%
SET5	78.50%	74.00%	81.32%	81.00%	77.00%	83.70%
Average	**81.80%**	**76.20%**	**85.86%**	**82.10%**	**75.60%**	**87.23%**

จากตารางที่ 5.1 พบว่าค่าเฉลี่ยของ Accuracy, Recall และ Precision ของวิธีการคัดเลือกแบบวงล้อถ่วงน้ำหนักและแบบจัดลำดับมีค่าใกล้เคียงกันและเมื่อนำค่าAccuracy, Recall และ Precision ของวิธีการคัดเลือกแบบวงล้อถ่วงน้ำหนักและแบบจัดลำดับมาสร้างเป็นกราฟ พบว่าทั้งสองวิธีมีค่าใกล้เคียงกัน โดยค่าเฉลี่ยของวิธีวิธีการคัดเลือกแบบวงล้อถ่วงน้ำหนักมีค่าดีกว่าเล็กน้อยดังรูป 5.1

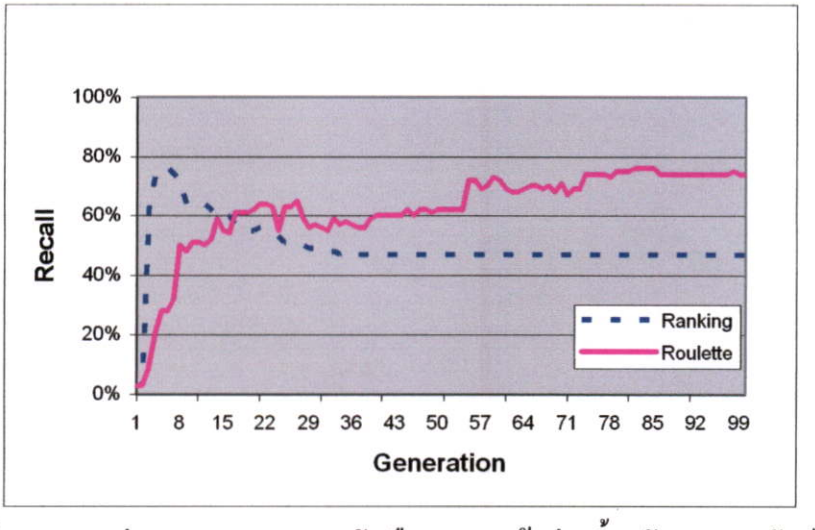


รูปที่ 5.1 แสดงการเปรียบเทียบค่า Accuracy, Recall และ Precision ระหว่างการคัดเลือกแบบวงล้อถ่วงน้ำหนัก (Roulette Wheel) และการคัดเลือกแบบจัดลำดับ (Ranking)

รูปที่ 5.2 ถึง 5.4 เป็นการแสดงค่า Accuracy, Recall และ Precision ของการคัดเลือกแบบวงล้อถ่วงน้ำหนัก โดยใช้เลือกชุดข้อมูลที่ใกล้เคียงกับค่าเฉลี่ย มาสร้างเป็นกราฟเพื่อให้เห็นภาพรวมของผลของวิธีการคัดเลือกทั้งสองแบบ

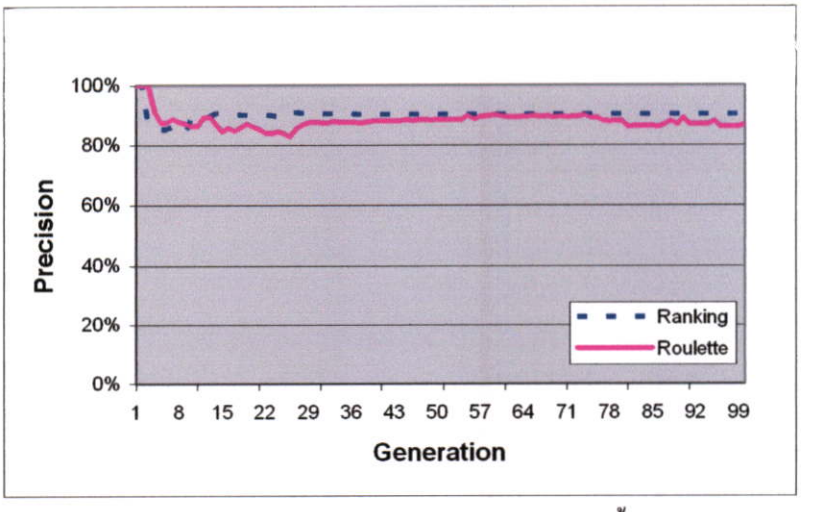


รูปที่ 5.2 แสดงค่า Accuracy ของการการคัดเลือกแบบวงล้อถ่วงน้ำหนัก และการคัดเลือกแบบจัดลำดับ



รูปที่ 5.3 แสดงค่า Recall ของการการคัดเลือกแบบวงล้อถ่วงน้ำหนัก และการคัดเลือกแบบจัดลำดับ

จากรูปที่ 5.2 และจากรูปที่ 5.3 จะเห็นว่าการคัดเลือกแบบจัดลำดับนั้นจะมีค่า Accuracy และ Recall ดีกว่าการคัดเลือกแบบวงล้อถ่วงน้ำหนักในรุ่นแรกๆ และจะลดต่ำลงเรื่อยๆ และคงที่ในที่สุด ส่วนค่า Precision ของการคัดเลือกทั้งสองแบบจะมีค่าใกล้เคียงกันดังรูปที่ 5.4

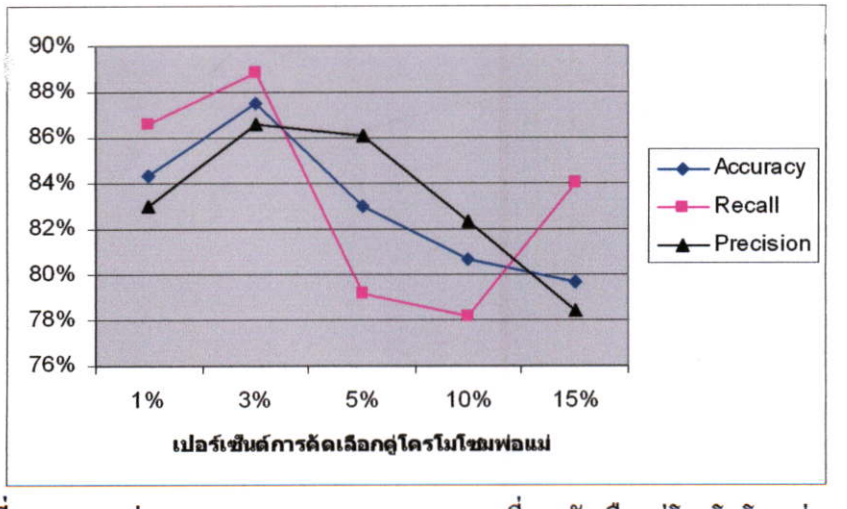


รูปที่ 5.4 แสดงค่า Precision ของการการคัดเลือกแบบวงล้อถ่วงน้ำหนัก และการคัดเลือกแบบจัดลำดับ

การคัดเลือกแบบจัดลำดับนั้นอาจทำให้เกิดโครโมโซมที่มีลักษณะคล้ายคลึงกันมากเกินไปได้ (Over Crowding) เนื่องจากการพยายามเลือกแต่โครโมโซมที่มีค่าความเหมาะสมสูงๆ เอาไว้ ส่งผลให้การไม่เกิดโครโมโซมที่มีความหลากหลาย ดังนั้น สำหรับการคัดเลือกในงานในวิทยานิพนธ์นี้จะใช้การคัดเลือกแบบวงล้อถ่วงน้ำหนักซึ่งสามารถแก้ปัญหาข้างต้นได้

5.2.2 การปรับเปอร์เซ็นต์การเลือกคู่โครโมโซมพ่อแม่ (Parent Pair)

การปรับเปอร์เซ็นต์การเลือกคู่โครโมโซมพ่อแม่ ได้ทำการทดลองกับข้อมูล 5 ชุด โดยในแต่ละชุดมีจำนวนโครโมโซมเป็น 1%, 3%, 5%, 10%, 15% ได้ผลการทดลองดังรูปที่ 5.5 และตารางที่ 5.2



รูปที่ 5.5 แสดงค่า Accuracy, Recall และ Precision ที่การคัดเลือกคู่โครโมโซมพ่อแม่ที่เปอร์เซ็นต์ต่างๆ

ตารางที่ 5.2 แสดงค่า Accuracy, Recall และ Precision ที่การคัดเลือกคู่โครโมโซมพ่อแม่ที่เปอร์เซ็นต์ต่างๆ

Parent Pair	Percent	Accuracy	Recall	Precision
SET 1	1%	85.00%	89.00%	82.41%
	3%	87.50%	87.00%	87.88%
	5%	86.50%	87.00%	86.14%
	10%	85.00%	87.00%	83.65%
	15%	84.00%	84.00%	84.00%
SET 2	1%	81.00%	76.00%	84.44%
	3%	88.50%	90.00%	87.38%
	5%	78.00%	64.00%	88.89%
	10%	79.00%	71.00%	84.52%
	15%	79.50%	81.00%	78.64%
SET 3	1%	83.50%	93.00%	78.15%
	3%	87.50%	91.00%	85.05%
	5%	82.00%	75.00%	87.21%
	10%	82.50%	84.00%	81.55%
	15%	79.50%	98.00%	71.53%
SET 4	1%	85.00%	89.00%	82.41%
	3%	86.50%	90.00%	84.11%
	5%	84.00%	80.00%	86.96%
	10%	78.00%	76.00%	79.17%
	15%	82.50%	75.00%	88.24%
SET 5	1%	87.00%	86.00%	87.76%
	3%	87.50%	86.00%	88.66%
	5%	84.50%	90.00%	81.08%
	10%	79.00%	73.00%	82.95%
	15%	73.00%	82.00%	69.49%
Average	1%	84.30%	86.60%	83.03%
	3%	**87.50%**	**88.80%**	**86.62%**
	5%	83.00%	79.20%	86.06%
	10%	80.70%	78.20%	82.37%
	15%	79.70%	84.00%	78.38%

จากการทดลองพบว่า ค่าเฉลี่ย Accuracy, Recall และ Precision ของเปอร์เซ็นต์การคัดเลือกคู่โครโมโซมพ่อแม่ที่ดีที่สุดคือ ที่การเลือกโครโมโซมพ่อแม่ที่ 3% ดังตารางที่ 5.2 ทั้งนี้เนื่องจากการเลือกคู่โครโมโซมพ่อแม่จำนวนน้อยๆ เพื่อมาทำการตามกระบวนการเจเนติกนั้นจะทำให้เกิดการเปลี่ยนแปลงอย่างค่อยเป็นค่อยไป จึงมีโอกาสพบคำตอบของปัญหาได้หลากหลายกว่าการเลือกคู่โครโมโซมพ่อแม่ที่เปอร์เซ็นต์สูงๆ และเมื่อนำมาสร้างเป็นกราฟจะเห็นได้อย่างชัดเจนว่า การเลือกคู่โครโมโซมพ่อแม่ที่ 3% นั้นให้ผลลัพธ์ที่ดีมากกว่าที่เปอร์เซ็นต์อื่นๆ ดังรูป 5.5

5.2.3 การปรับค่ามิวเตชันพารามิเตอร์

การปรับค่ามิวเตชันพารามิเตอร์ในงานวิทยานิพนธ์นี้มี 2 พารามิเตอร์ด้วยกันคือ จำนวน Generation ที่ผ่านไปก่อนทำการมิวเตชัน และ เปอร์เซ็นต์ของโครโมโซมที่จะถูกมิวเตชัน (Chromosome) การปรับค่าพารามิเตอร์ทั้งสอง เพื่อหาค่าที่เหมาะสมสำหรับการมิวเตชันนั้นแสดงดังตารางที่ 5.3 และ 5.4

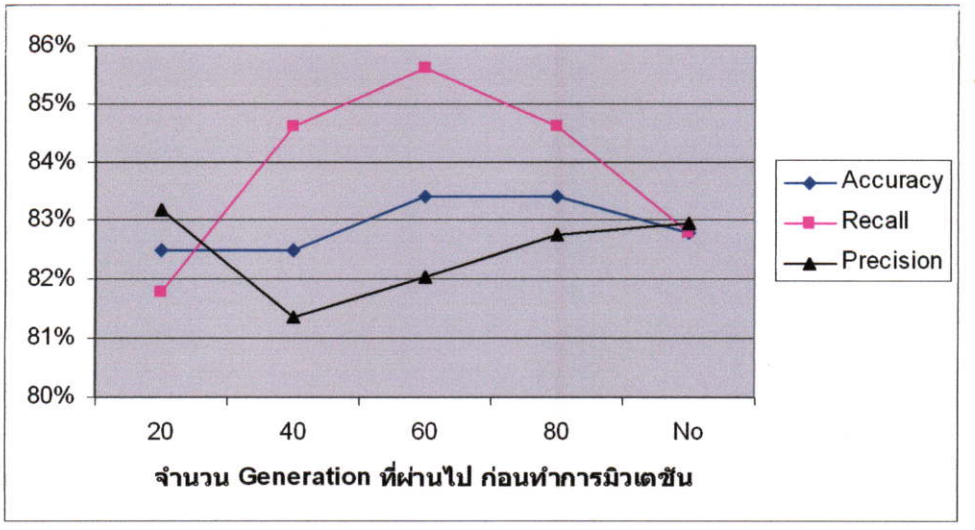
ตารางที่ 5.3 แสดงค่า Accuracy, Recall และ Precision ของการมิวเตชันเมื่อผ่านไป 20, 40, 60, 80 Generation และเมื่อไม่ทำการมิวเตชันเลย (No)

Mutation	Generation	Accuracy	Recall	Precision
SET 1	20	82.50%	84.00%	81.55%
	40	84.00%	87.00%	82.08%
	60	84.00%	90.00%	80.36%
	80	83.50%	85.00%	82.52%
	No	84.50%	90.00%	81.08%
SET 2	20	81.00%	74.00%	86.05%
	40	80.00%	74.00%	84.09%
	60	84.00%	84.00%	84.00%
	80	82.00%	77.00%	85.56%
	No	80.50%	75.00%	84.27%
SET 3	20	83.50%	89.00%	80.18%
	40	83.00%	88.00%	80.00%
	60	83.00%	85.00%	81.73%
	80	84.00%	86.00%	82.69%
	No	80.50%	75.00%	84.27%
SET 4	20	84.00%	86.00%	82.69%
	40	83.00%	84.00%	82.35%
	60	83.00%	85.00%	81.73%
	80	84.50%	86.00%	83.50%
	No	84.50%	85.00%	84.16%
SET 5	20	81.50%	76.00%	85.39%
	40	82.50%	90.00%	78.26%
	60	83.00%	84.00%	82.35%
	80	83.00%	89.00%	79.46%
	No	84.00%	89.00%	80.91%
Average	20	82.50%	81.80%	83.17%
	40	82.50%	84.60%	81.36%
	60	**83.40%**	**85.60%**	**82.03%**
	80	83.40%	84.60%	82.75%
	No	82.80%	82.80%	82.94%

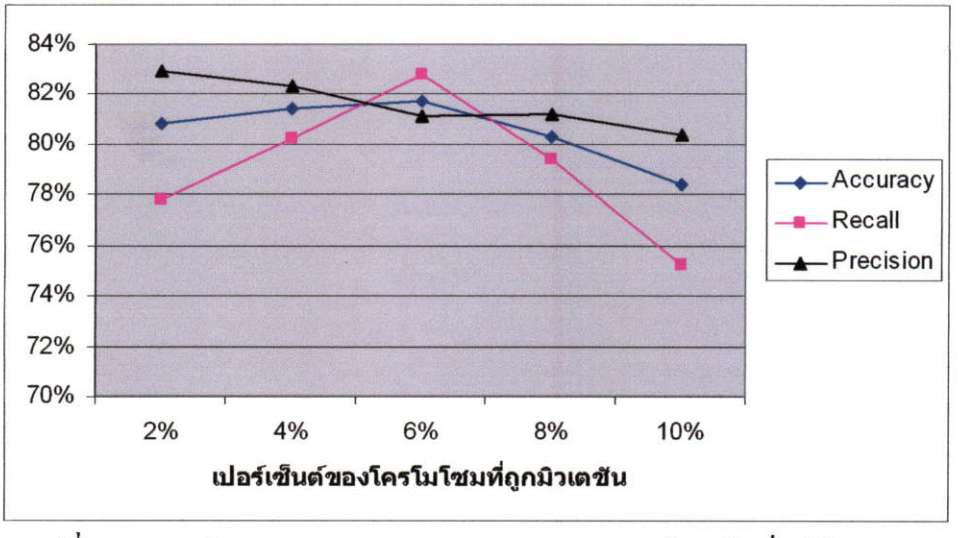
ตารางที่ 5.4 แสดงค่า Accuracy, Recall และ Precision ของการมิวเตชันเมื่อปรับจำนวนโครโมโซมที่เปอร์เซ็นต์ต่างๆ

Mutation	Chromosome	Accuracy	Recall	Precision
SET 1	2%	82.00%	85.00%	80.19%
	4%	83.50%	84.00%	83.17%
	6%	80.00%	74.00%	84.09%
	8%	79.00%	68.00%	87.18%
	10%	78.00%	74.00%	80.43%
SET 2	2%	80.00%	73.00%	84.88%
	4%	82.00%	85.00%	80.19%
	6%	81.50%	86.00%	78.90%
	8%	79.50%	79.00%	79.80%
	10%	76.50%	70.00%	80.46%
SET 3	2%	80.00%	74.00%	84.09%
	4%	79.50%	72.00%	84.71%
	6%	83.00%	86.00%	81.13%
	8%	80.00%	81.00%	79.41%
	10%	78.50%	77.00%	79.38%
SET 4	2%	79.50%	73.00%	83.91%
	4%	79.50%	74.00%	83.15%
	6%	82.50%	84.00%	81.55%
	8%	82.00%	86.00%	79.63%
	10%	77.00%	70.00%	81.40%
SET 5	2%	82.50%	84.00%	81.55%
	4%	82.50%	86.00%	80.37%
	6%	81.50%	84.00%	80.00%
	8%	81.00%	83.00%	79.81%
	10%	82.00%	85.00%	80.19%
Average	2%	80.80%	77.80%	82.92%
	4%	81.40%	80.20%	82.32%
	6%	**81.70%**	**82.80%**	**81.14%**
	8%	80.30%	79.40%	81.17%
	10%	78.40%	75.20%	80.37%

จากตารางที่ 5.3 ถึงตารางที่ 5.4 และรูป 5.6 ถึงรูปที่ 5.7 แสดงให้เห็นว่า ค่าเฉลี่ยค่า Accuracy, Recall และ Precision ที่ดีคือ ทำการมิวเตชันเมื่อผ่านการดำเนินการทางเจเนติกไป 60 Generation และเปอร์เซ็นต์ของโครโมโซมที่ถูกมิวเตชันซึ่งให้ค่าเฉลี่ยของค่า Accuracy, Recall และ Precision ที่ดีคือ 6 % แต่อย่างไรก็ตาม ค่าเฉลี่ยของ ค่า Accuracy, Recall และ Precision เมื่อทำการมิวเตชันไม่สามารถกำหนดแนวโน้มที่แน่นอนได้และไม่ได้ทำให้ผลการทดลองดีขึ้นเสมอไป ในบางกรณีอาจทำให้ผลการทดลองแย่ลง เนื่องจากการมิวเตชันนั่นเอง ดังนั้นในงานวิทยานิพนธ์นี้จึงเลือกที่จะไม่ใช้มิวเตชันพารามิเตอร์



รูปที่ 5.6 แสดงค่า Accuracy, Recall และ Precision ของการมิวเตชันเมื่อผ่านไป 20, 40, 60, 80 Generation และเมื่อไม่ทำการมิวเตชันเลย (No)



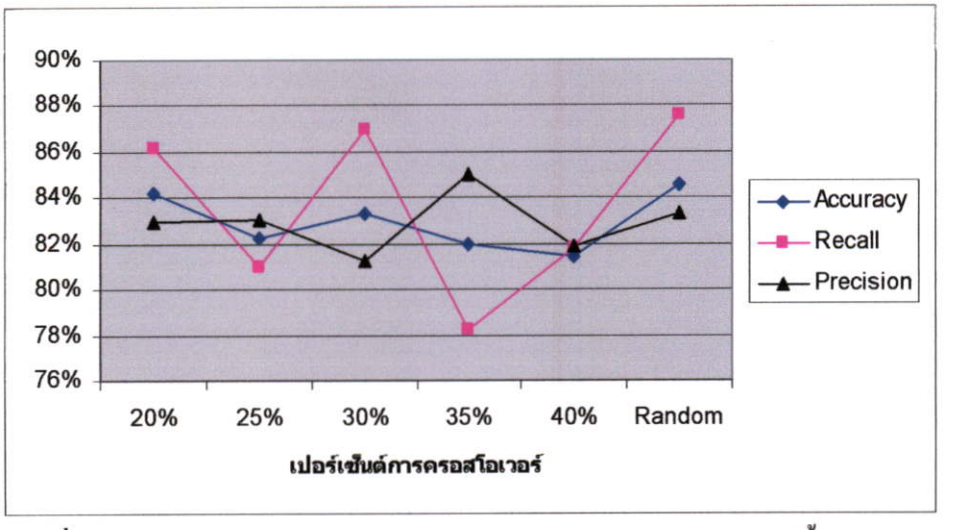
รูปที่ 5.7 แสดงค่า Accuracy, Recall และ Precision ของการมิวเตชันเมื่อปรับจำนวนโครโมโซมที่เปอร์เซ็นต์ต่างๆ

5.2.4 การปรับครอสโอเวอร์พารามิเตอร์

การปรับพารามิเตอร์ของการครอสโอเวอร์ทำได้ 2 แบบ คือการครอสโอเวอร์แบบสุ่ม (Random) และการครอสโอเวอร์แบบกำหนดเปอร์เซ็นต์คงที่ ซึ่งในวิทยานิพนธ์นี้ได้กำหนดเปอร์เซ็นต์ที่ 20%, 25%, 30%, 35%, 40% และได้ผลการทดลองดังแสดงในตารางที่ 5.5

ตารางที่ 5.5 แสดงค่า Accuracy, Recall และ Precision ของการครอสโอเวอร์ทั้งสองแบบ

Crossover	Percent	Accuracy	Recall	Precision
SET 1	20%	85.50%	91.00%	81.98%
	25%	87.50%	88.00%	87.13%
	30%	88.00%	88.00%	88.00%
	35%	84.00%	77.00%	90.00%
	40%	80.50%	96.00%	73.28%
	Random	80.00%	95.00%	73.08%
SET 2	20%	83.50%	84.00%	83.17%
	25%	81.50%	84.00%	80.00%
	30%	83.00%	85.00%	81.73%
	35%	80.50%	80.00%	80.81%
	40%	79.00%	78.00%	79.59%
	Random	84.00%	86.00%	82.69%
SET 3	20%	83.00%	88.00%	80.00%
	25%	79.00%	71.00%	84.52%
	30%	85.00%	96.00%	78.69%
	35%	83.00%	76.00%	88.37%
	40%	83.50%	81.00%	85.26%
	Random	87.50%	90.00%	85.71%
SET 4	20%	84.50%	82.00%	86.32%
	25%	84.00%	83.00%	84.69%
	30%	82.00%	82.00%	82.00%
	35%	81.50%	84.00%	80.00%
	40%	83.00%	81.00%	84.38%
	Random	85.50%	82.00%	88.17%
SET 5	20%	84.50%	86.00%	83.50%
	25%	79.00%	79.00%	79.00%
	30%	78.50%	84.00%	75.68%
	35%	81.00%	74.00%	86.05%
	40%	81.00%	73.00%	86.90%
	Random	86.00%	85.00%	86.73%
Average	20%	84.20%	86.20%	82.99%
	25%	82.20%	81.00%	83.07%
	30%	83.30%	87.00%	81.22%
	35%	82.00%	78.20%	85.05%
	40%	81.40%	81.80%	81.88%
	Random	**84.60%**	**87.60%**	**83.28%**



รูปที่ 5.8 แสดงค่า Accuracy, Recall และ Precision ของการครอสโอเวอร์ทั้งสองแบบ

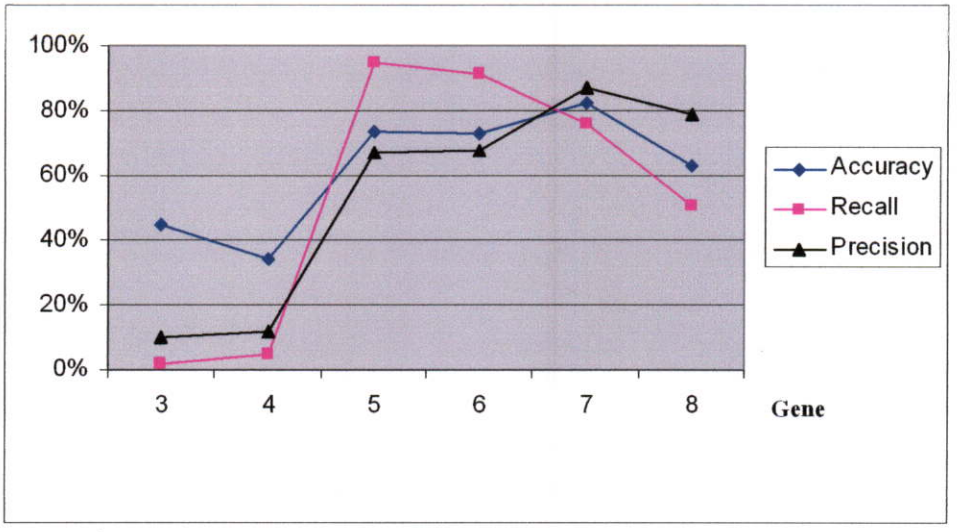
เมื่อทำการทดลองกับชุดข้อมูลทั้ง 5 เซตพบว่า ค่าเฉลี่ย Accuracy, Recall และ Precision ของการครอสโอเวอร์แบบกำหนดเปอร์เซ็นต์คงที่ที่เปอร์เซ็นต์ต่างๆ กัน ไม่สามารถระบุแนวโน้มที่แน่นอนได้และมีค่าเฉลี่ยดังกล่าวน้อยกว่าการครอสโอเวอร์แบบสุ่มเล็กน้อย ดังตารางที่ 5.5 และรูปที่ 5.8 สำหรับในงานวิทยานิพนธ์นี้จึงเลือกการครอสโอเวอร์แบบสุ่ม (Random)

5.2.5 การปรับค่ายีนเธร็โชว์

การปรับค่ายีนเธร็โชว์คือการปรับพารามิเตอร์จำนวนของยีนที่ต้องตรงตามเงื่อนไขระหว่างยีนในโครโมโซมแม่แบบและยีนในโครโมโซมของอีเมล์ทดสอบ ซึ่งสามารถปรับได้ตั้งแต่ 1 ถึง 8 (1 โครโมโซมมี 8 ยีน) เมื่อทำการทดลองกับชุดข้อมูลทั้ง 5 เซตพบว่า ค่าเฉลี่ยของการปรับค่ายีนเท่ากับ 7 ให้ผลลัพธ์ของการกรองที่ดี ดังนั้นจึงเลือกปรับค่ายีนเธร็โชว์เท่ากับ 7 เนื่องจากให้ผลลัพธ์ที่ดีกว่าค่าอื่นๆ ดังแสดงในตารางที่ 5.6

ตารางที่ 5.6 แสดงค่า Accuracy, Recall และ Precision ของยีนเธร์โชว์ที่แตกต่างกัน

Gene Threshold	Gene	Accuracy	Recall	Precision
SET 1	3	45.00%	1.00%	8.33%
	4	36.00%	2.00%	6.25%
	5	66.50%	99.00%	60.00%
	6	79.00%	95.00%	71.97%
	7	85.00%	74.00%	94.87%
	8	38.50%	76.00%	43.43%
SET 2	3	43.50%	3.00%	15.79%
	4	34.50%	6.00%	13.95%
	5	78.50%	88.00%	73.95%
	6	67.50%	91.00%	61.90%
	7	81.00%	71.00%	88.75%
	8	66.00%	35.00%	92.11%
SET 3	3	49.00%	0.00%	0.00%
	4	18.00%	8.00%	10.00%
	5	78.00%	99.00%	69.72%
	6	85.00%	93.00%	80.17%
	7	81.50%	80.00%	82.47%
	8	69.50%	43.00%	91.49%
SET 4	3	46.00%	2.00%	16.67%
	4	37.00%	5.00%	13.89%
	5	73.50%	93.00%	66.91%
	6	70.50%	84.00%	66.14%
	7	82.00%	76.00%	86.36%
	8	74.50%	56.00%	88.89%
SET 5	3	39.50%	2.00%	8.00%
	4	45.00%	2.00%	14.29%
	5	72.50%	95.00%	65.52%
	6	63.00%	94.00%	58.02%
	7	81.00%	77.00%	83.70%
	8	66.00%	43.00%	79.63%
Average	3	44.60%	1.60%	9.76%
	4	34.10%	4.60%	11.68%
	5	73.80%	94.80%	67.22%
	6	73.00%	91.40%	67.64%
	7	**82.10%**	**75.60%**	**87.23%**
	8	62.90%	50.60%	79.11%



รูปที่ 5.9 แสดงค่า Accuracy, Recall และ Precision ของยีนเธร์โชว์ที่แตกต่างกัน

การปรับค่ายีนเธร์โชว์น้อยเกินไป เช่น 3 ยีนหรือ 4 ยีน จะทำให้ค่าเฉลี่ยของ Accuracy, Recall และ Precision ต่ำมากๆ เนื่องจากมีความเป็นได้สูงที่ยีนของโครโมโซมที่เป็นอีเมล์ทดสอบจะตรงกับกับยีนในโครโมโซมแม่แบบเพียง 3 ถึง 4 โครโมโซม ซึ่งยังไม่มากเพียงพอที่จะระบุประเภทของอีเมล์ได้ ทำให้ผลลัพธ์ต่ำมาก การปรับค่ายีนเธร์โชว์ที่ 5 หรือ 6 นั้นพบว่าผลลัพธ์ที่ดีขึ้นมาก แต่จะดีที่สุดเมื่อปรับค่ายีนเธร์โชว์ที่ 7 ยีน และผลจะต่ำลงเมื่อปรับค่ายีนเธร์โชว์ที่ 8 ซึ่งมากเกินไป เพราะเป็นไปได้ยากที่โครโมโซมที่เป็นอีเมล์ทดสอบจะตรงกับกับยีนในโครโมโซมแม่แบบถึง 8 ยีนดังตารางที่ 5.6 และรูปที่ 5.9

จากผลการทดลองการปรับพารามิเตอร์ข้างต้น จะเห็นว่า การปรับวิธีการคัดเลือก, มิวเตชันพารามิเตอร์และครอสโอเวอร์พารามิเตอร์นั้น ไม่ทำให้เห็นความแตกต่างของผลลัพธ์มากนัก ส่วนพารามิเตอร์ที่มีผลต่อการกรองอย่างมากคือ จำนวนคู่โครโมโซมพ่อแม่และจำนวนยีนเธร์โชว์ ดังนั้น เราจะทำการปรับพารามิเตอร์ทั้งสองตัวอย่างละเอียดอีกครั้ง โดยจะทำการ Fix พารามิเตอร์ตัวอื่นๆ ที่เหลือ โดยจำนวนคู่โครโมโซมพ่อแม่ประกอบด้วย 3, 5, 10, 15 และ 30 เปอร์เซ็นต์ ส่วนค่ายีนเธร์โชว์ประกอบด้วย 5, 6 และ 7 ยีน ซึ่งเมื่อทำการทดลองกับข้อมูล 5 ชุดได้ค่า Accuracy, Recall และ Precision ดังตารางที่ 5.7

ตารางที่ 5.7 แสดงค่า Accuracy, Recall และ Precision ของการปรับจำนวนคู่โครโมโซมพ่อแม่และยีนเธร์โชว์ที่ค่าต่างๆ

	Parent Pair	Gene Threshold	Accuracy	Recall	Precision
SET 1	3	5	85.00%	87.50%	83.33%
	5	5	85.25%	86.50%	84.39%
	10	5	85.50%	85.50%	85.50%
	15	5	82.00%	88.00%	78.57%
	30	5	84.00%	88.00%	81.48%
	3	6	87.75%	88.00%	87.56%
	5	6	87.00%	86.00%	87.76%
	10	6	81.75%	83.50%	80.68%
	15	6	83.50%	81.50%	84.90%
	30	6	78.50%	96.00%	71.11%
	3	7	78.50%	64.00%	90.14%
	5	7	85.25%	81.00%	88.52%
	10	7	85.75%	82.00%	88.65%
	15	7	86.25%	82.00%	89.62%
	30	7	86.75%	89.00%	85.17%
SET 2	3	5	88.00%	87.50%	88.38%
	5	5	87.00%	86.00%	87.76%
	10	5	87.00%	87.00%	87.00%
	15	5	78.00%	91.00%	72.22%
	30	5	65.50%	91.50%	60.20%
	3	6	90.50%	91.50%	89.71%
	5	6	83.75%	96.00%	77.11%
	10	6	87.75%	87.50%	87.94%
	15	6	82.75%	86.50%	80.47%
	30	6	71.75%	100.00%	63.90%
	3	7	84.75%	76.00%	92.12%
	5	7	87.25%	87.50%	87.06%
	10	7	84.50%	83.00%	85.57%
	15	7	87.00%	89.50%	85.24%
	30	7	83.75%	83.00%	84.26%
SET 3	3	5	87.00%	86.50%	87.37%
	5	5	86.25%	85.00%	87.18%
	10	5	86.50%	85.00%	87.63%
	15	5	84.75%	85.50%	84.24%
	30	5	64.50%	100.00%	58.48%
	3	6	87.50%	85.00%	89.47%
	5	6	83.25%	77.50%	87.57%
	10	6	85.75%	84.50%	86.67%
	15	6	87.00%	86.50%	87.37%
	30	6	84.00%	85.00%	83.33%
	3	7	85.50%	82.00%	88.17%
	5	7	82.75%	73.50%	90.18%
	10	7	85.25%	83.00%	86.91%
	15	7	85.25%	80.50%	88.95%
	30	7	83.75%	83.00%	84.26%

ตาราง 5.7 (ต่อ)

SET 4	3	5	86.00%	87.00%	85.29%
	5	5	84.75%	85.00%	84.58%
	10	5	86.50%	89.50%	84.43%
	15	5	84.75%	84.50%	84.92%
	30	5	86.00%	84.50%	87.11%
	3	6	86.50%	84.50%	88.02%
	5	6	88.00%	88.50%	87.62%
	10	6	74.50%	99.50%	66.33%
	15	6	78.00%	94.00%	71.21%
	30	6	73.00%	88.00%	67.69%
	3	7	82.75%	77.50%	86.59%
	5	7	83.50%	84.00%	83.17%
	10	7	84.25%	79.00%	88.27%
	15	7	83.25%	77.00%	88.00%
	30	7	70.75%	51.00%	84.30%
SET 5	3	5	88.00%	91.00%	85.85%
	5	5	86.75%	84.50%	88.48%
	10	5	87.00%	88.50%	85.92%
	15	5	86.50%	87.00%	86.14%
	30	5	85.25%	85.50%	85.07%
	3	6	87.25%	87.00%	87.44%
	5	6	88.25%	88.50%	88.06%
	10	6	86.25%	83.50%	88.36%
	15	6	87.75%	87.50%	87.94%
	30	6	86.25%	83.50%	88.36%
	3	7	80.25%	68.50%	89.54%
	5	7	86.25%	86.00%	86.43%
	10	7	84.75%	81.00%	87.57%
	15	7	72.75%	51.50%	89.57%
	30	7	82.75%	80.00%	84.66%
Average	3	5	86.80%	87.90%	86.04%
	5	5	86.00%	85.40%	86.48%
	10	5	86.50%	87.10%	86.10%
	15	5	83.20%	87.20%	81.22%
	30	5	77.05%	89.90%	74.47%
	3	**6**	**87.90%**	**87.20%**	**88.44%**
	5	6	86.05%	87.30%	85.62%
	10	6	83.20%	87.70%	82.00%
	15	6	83.80%	87.20%	82.38%
	30	6	78.70%	90.50%	74.88%
	3	7	82.35%	73.60%	89.31%
	5	7	85.00%	82.40%	87.07%
	10	7	84.90%	81.60%	87.39%
	15	7	82.90%	76.10%	88.28%
	30	7	81.55%	77.20%	84.53%

จากตารางที่ 5.7 จะเห็นว่าจากค่าเฉลี่ยของชุดข้อมูลทั้ง 5 ชุด พบว่าจะให้ผลลัพธ์ที่ดีที่สุดเมื่อปรับจำนวนคู่โครโมโซมพ่อแม่ 3% และเมื่อปรับค่ายีนเธร์โชว์ที่ 6 ยีน

5.3 ผลการทดลองการกรองอีเมล์ขยะโดยใช้เจเนติกอัลกอริทึม

จากการทดลองในหัวข้อ 5.2 จะได้ค่าพารามิเตอร์ที่เหมาะสมที่จะนำมาใช้ในการวิจัย ซึ่งได้แก่

- วิธีการคัดเลือกใช้แบบวงล้อถ่วงน้ำหนัก
- การครอสโอเวอร์ เลือกแบบสุ่ม
- การมิวเตชัน ไม่ใช้พารามิเตอร์นี้
- จำนวนคู่โครโมโซมพ่อแม่ 3%
- ยีนเธร์โชว์ 6 ยีน

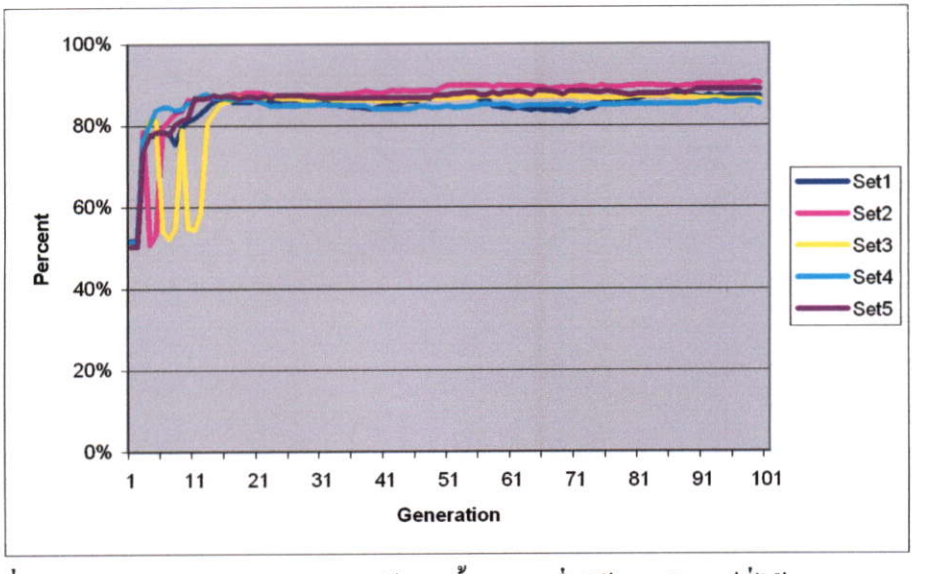
พารามิเตอร์เหล่านี้จะถูกนำมาใช้ในการทดลองประสิทธิภาพของตัวกรอง ซึ่งผลที่ได้เป็นดังตาราง 5.8

ตาราง 5.8 แสดงค่าเฉลี่ย Accuracy, Recall และ Precision ของการทดลองในแต่ละเซตและค่าเฉลี่ย Accuracy, Recall และ Precision ของทุกๆ เซต

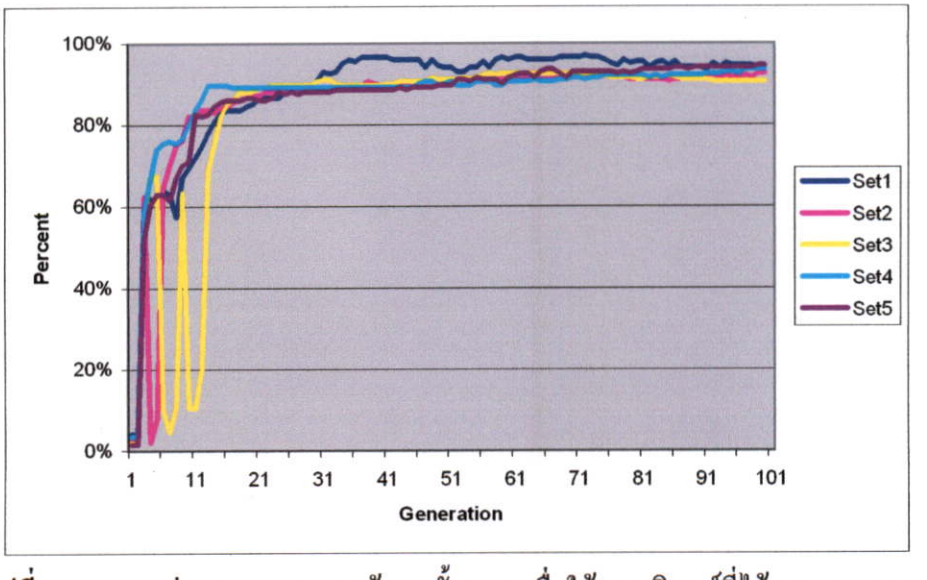
		Acc	Recall	Pre	Average of Each Set		
รอบที่ 1	SET1	86.50%	86.50%	86.50%	87.70%	88.20%	87.33%
	SET2	90.25%	92.50%	88.52%			
	SET3	87.00%	88.00%	86.27%			
	SET4	87.25%	86.00%	88.21%			
	SET5	87.50%	88.00%	87.13%			
รอบที่ 2	SET1	87.25%	87.50%	87.06%	87.15%	87.90%	86.60%
	SET2	88.25%	88.50%	88.06%			
	SET3	86.50%	87.50%	85.78%			
	SET4	85.50%	86.00%	85.15%			
	SET5	88.25%	90.00%	86.96%			
รอบที่ 3	SET1	88.50%	89.50%	87.75%	87.55%	87.50%	87.66%
	SET2	87.25%	84.50%	89.42%			
	SET3	86.75%	84.50%	88.48%			
	SET4	86.75%	86.00%	87.31%			
	SET5	88.50%	93.00%	85.32%			
รอบที่ 4	SET1	86.50%	82.00%	90.11%	84.75%	90.20%	82.46%
	SET2	78.75%	96.50%	71.22%			
	SET3	83.25%	96.50%	76.28%			
	SET4	87.75%	88.50%	87.19%			
	SET5	87.50%	87.50%	87.50%			
รอบที่ 5	SET1	88.25%	89.50%	87.32%	88.10%	88.70%	87.68%
	SET2	90.25%	94.50%	87.10%			
	SET3	87.00%	88.00%	86.27%			
	SET4	86.75%	84.00%	88.89%			
	SET5	88.25%	87.50%	88.83%			
Average of All Set					**87.05%**	**88.50%**	**86.35%**

จากตารางที่ 5.8 จะเห็นว่าเมื่อทดลองด้วยพารามิเตอร์ที่เหมาะสมที่ได้จากการทดลองเมื่อปรับค่าอย่างละเอียดขึ้น พบว่าค่าเฉลี่ยAccuracy, Recall และ Precision ของทั้ง 5 เซตให้ผลลัพธ์ที่ดีขึ้น

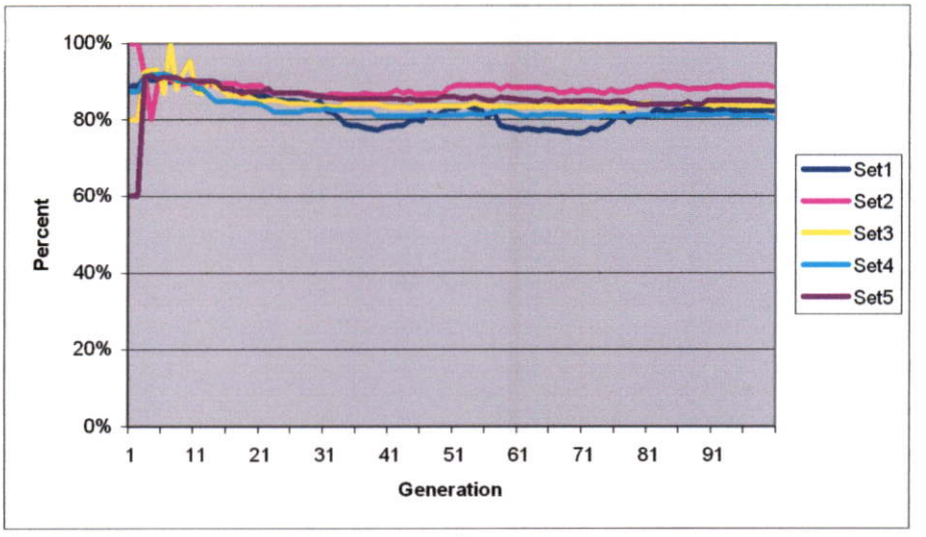
รูปที่ 5.10 ถึงรูปที่ 5.12 เป็นการแสดงค่า Accuracy, Recall และ Precision ของเซตข้อมูลทั้ง 5 เซตในรอบที่ 1 เมื่อใช้พารามิเตอร์ที่เหมาะสมที่ได้จากการทดลองในหัวข้อ 5.2 เพื่อให้เห็นภาพรวมของผลลัพธ์ที่ได้ในแต่ละเซตข้อมูล



รูปที่ 5.10 แสดงค่า Accuracy ของเซตข้อมูลทั้ง 5 เซตเมื่อใช้พารามิเตอร์ที่ได้จากการทดลอง



รูปที่ 5.11 แสดงค่า Recall ของเซตข้อมูลทั้ง 5 เซตเมื่อใช้พารามิเตอร์ที่ได้จากการทดลอง



รูปที่ 5.12 แสดงค่า Precision ของเซตข้อมูลทั้ง 5 เซตเมื่อใช้พารามิเตอร์ที่ได้จากการทดลอง

จากรูปที่ 5.10 ถึงรูปที่ 5.12 จะเห็นว่าค่า Accuracy, Recall และ Precision ของเซตข้อมูลทั้ง 5 จะมีการเปลี่ยนแปลงอย่างมากในช่วง Generation แรกๆ และมีการเปลี่ยนแปลงน้อยมากในช่วง Generation หลังๆ ซึ่งทำให้ค่า Accuracy, Recall และ Precision ของชุดข้อมูลแต่ละเซตจะเริ่มคงที่มากขึ้น ซึ่งเป็นรูปแบบทั่วไปของกระบวนการเจเนติกอัลกอริทึม และให้ค่า Accuracy, Recall และ Precision ที่ดี

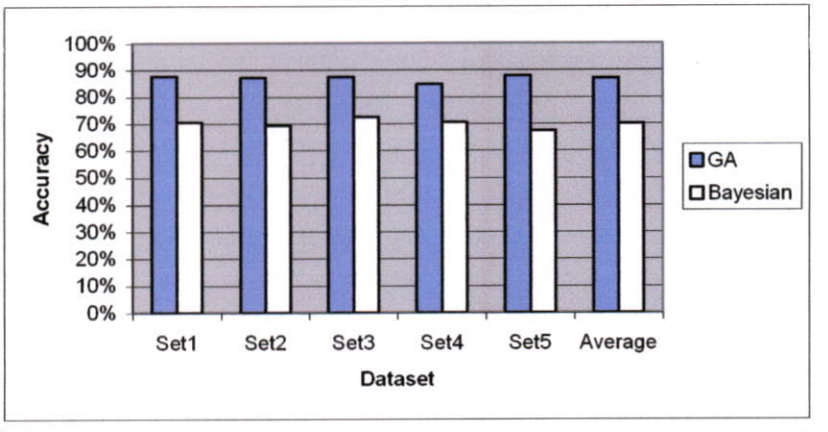
5.4 เปรียบเทียบตัวกรองอีเมล์ขยะที่นำเสนอกับตัวกรองอีเมล์ขยะซึ่งประยุกต์ใช้ทฤษฎีเบย์เซียน

การเปรียบเทียบค่า Accuracy, Recall และ Precision ที่ได้ระหว่างตัวกรองอีเมล์ขยะโดยใช้เจเนติกอัลกอริทึมและผลการทดลองที่ได้ระหว่างตัวกรองอีเมล์ขยะโดยใช้การประยุกต์ใช้ทฤษฎีเบย์เซียนมาเปรียบเทียบกันดังแสดงในตารางที่ 5.9

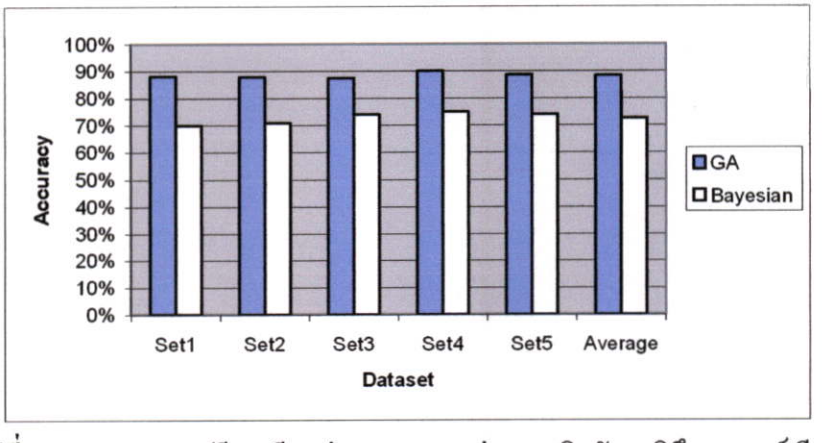
ตารางที่ 5.9 แสดงค่า Accuracy, Recall และ Precision ของเซตข้อมูลทั้ง 5 เซต

	Genetic Algorithm			Bayesian		
	Accuracy	**Recall**	**Precision**	**Accuracy**	**Recall**	**Precision**
Set1	87.70%	88.20%	87.33%	70.50%	70.00%	70.71%
Set2	87.15%	87.90%	86.60%	69.50%	71.00%	68.93%
Set3	87.55%	87.50%	87.66%	72.50%	74.00%	71.84%
Set4	84.75%	90.20%	82.46%	70.50%	75.00%	68.81%
Set5	88.10%	88.70%	87.68%	67.50%	74.00%	65.49%
Average	**87.05%**	**88.50%**	**86.35%**	**70.10%**	**72.80%**	**69.16%**

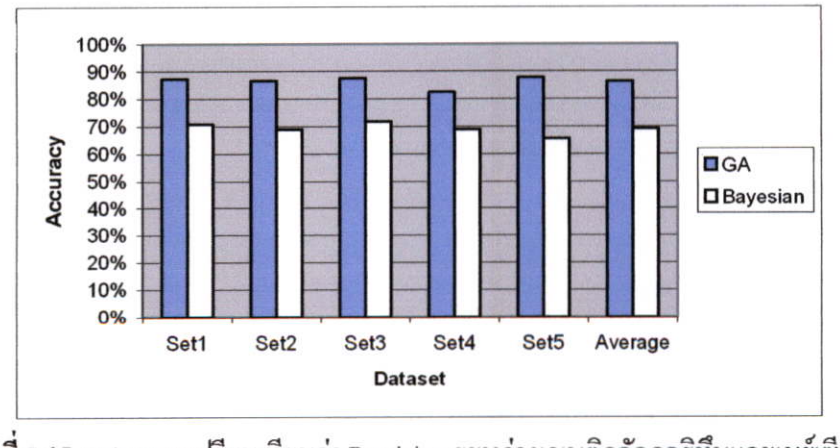
จากตารางที่ 5.9 พบว่าเมื่อนำค่า Accuracy, Recall และ Precision ของวิธีการที่นำเสนอกับวิธีการกรองอีเมล์ขยะโดยใช้การประยุกต์ใช้ทฤษฎีเบย์เซียนมาเปรียบเทียบกันทีละค่าจะได้กราฟดังรูปที่ 5.13 ถึงรูปที่ 5.15



รูปที่ 5.13 แสดงการเปรียบเทียบค่า Accuracy ระหว่างเจเนติกอัลกอริทึมและเบย์เซียน



รูปที่ 5.14 แสดงการเปรียบเทียบค่า Recall ระหว่างเจเนติกอัลกอริทึมและเบย์เซียน



รูปที่ 5.15 แสดงการเปรียบเทียบค่า Precision ระหว่างเจเนติกอัลกอริทึมและเบย์เซียน

จากรูปที่ 5.13 ถึงรูปที่ 5.15 พบว่าค่า Accuracy, Recall และ Precision ของเจเนติกอัลกอริทึมดีกว่าค่า Accuracy, Recall และ Precision ของเบย์เซียนในทุกเซตข้อมูลที่นำมาทดลอง ทั้งนี้เนื่องจากตัวกรองอีเมล์ขยะซึ่งประยุกต์ใช้ทฤษฎีเบย์เซียนนั้นจะเก็บสถิติของทั้งคำดีและคำขยะไว้ในฐานข้อมูลคำ เมื่อมีอีเมล์ทดสอบเข้ามา ตัวกรองจะนำคำในอีเมล์มาแยกเป็นคำดีและคำขยะ จากนั้นเปรียบเทียบระหว่างความน่าจะเป็นของคำที่เป็นคำขยะและความน่าจะเป็นของคำที่เป็นคำดี ถ้าความน่าจะเป็นของคำประเภทใดมากกว่า ก็จะทำนายว่าอีเมล์ทดสอบเป็นอีเมล์ประเภทนั้น ซึ่งจะเห็นว่าเป็นการเปรียบเทียบเพียงแค่ครั้งเดียว ในขณะที่ตัวกรองอีเมล์ขยะที่สร้างจากเจเนติกอัลกอริทึมจะมีการดำเนินการทางเจเนติกไปหลายๆรุ่น เพื่อหารูปแบบที่เหมาะสมเพื่อจะนำรูปแบบเหล่านี้มาใช้ในการกรองอีเมล์ขยะ ทำให้ประสิทธิภาพของการกรองโดยใช้ตัวกรองอีเมล์ขยะโดยใช้เจเนติกอัลกอริทึมดีกว่าประสิทธิภาพของตัวกรองอีเมล์ขยะตามทฤษฎีเบย์เซียน

บทที่ 6
สรุปผลการทดลองและข้อเสนอแนะ

วิทยานิพนธ์นี้ ได้นำเสนอการสร้างตัวกรองอีเมล์ขยะโดยใช้เจเนติกอัลกอริทึม ในการสร้างแม่แบบที่หลากหลายของอีเมล์ เพื่อนำไปสร้างกฎสำหรับกรองอีเมล์ขยะ เนื่องจากตัวดำเนินการในเจเนติกอัลกอริทึมประกอบไปด้วย การคัดเลือก การครอสโอเวอร์ และการมิวเตชัน ซึ่งแต่ละตัวดำเนินการก็สามารถทำหลายวิธี เช่น การคัดเลือก มีแบบ การคัดเลือกแบบจัดลำดับ การคัดเลือกแบบวงล้อถ่วงน้ำหนัก หรือในครอสโอเวอร์ มีเปอร์เซ็นต์ของการครอสเป็นต้นจึงต้องมีการปรับค่าพารามิเตอร์ต่างๆ เพื่อหาค่าพารามิเตอร์ที่เหมาะสมให้กับระบบ และพบว่าเมื่อนำพารามิเตอร์เหล่านี้ไปทดสอบได้ผลลัพธ์ที่ดี จากนั้นได้ประเมินประสิทธิภาพของอัลกอริทึมที่นำเสนอเปรียบเทียบกับตัวกรองอีเมล์ขยะตามทฤษฎีเบย์เซียน พบว่า ค่า Accuracy, Recall และ Precision ของวิธีการที่นำเสนอ ดีกว่าตัวกรองที่สร้างขึ้นด้วยทฤษฎีเบย์เซียน

6.1 สรุปและวิเคราะห์ผลการดำเนินงานวิจัย

จากการทดลองปรับพารามิเตอร์ในบทที่ 5 เพื่อหาค่าพารามิเตอร์ที่เหมาะสม ได้ผลลัพธ์ดังนี้

- การคัดเลือกคู่โครโมโซมพ่อแม่เพื่อที่จะนำไปทำตามกระบวนเจเนติกใช้การคัดเลือกแบบใช้วงล้อถ่วงน้ำหนักเพื่อลดการเกิดโครโมโซมที่มีลักษณะคล้ายคลึงกันมากเกินไป (Over Crowding) ที่เกิดในวิธีการคัดเลือกแบบจัดลำดับซึ่งมีผลให้ตัวกรองไม่สามารถกรองอีเมล์หน้าตาแปลกๆ ได้ ทำให้ผลลัพธ์โดยรวมของระบบต่ำลง
- เลือกเปอร์เซ็นต์ของคู่พ่อแม่ที่จะนำมาครอสโอเวอร์ 3 % เนื่องจากการเลือกคู่พ่อแม่ที่จะนำมาครอสโอเวอร์ในปริมาณน้อยๆ จะทำให้เกิดการเปลี่ยนแปลงอย่างค่อยเป็นค่อยไป ทำให้มีโอกาสพบคำตอบของปัญหาได้หลากหลายกว่าการเลือกคู่โครโมโซมพ่อแม่มาครอสโอเวอร์กันในปริมาณมากๆ
- การครอสโอเวอร์ เลือกแบบสุ่ม เพราะถึงแม้ว่าการครอสโอเวอร์แบบสุ่ม (Random) และการครอสโอเวอร์แบบกำหนดเปอร์เซ็นต์ตายตัวจะมีค่าใกล้เคียงกัน แต่การสุ่มเป็นกระบวนการพื้นฐานของเจเนติกอัลกอริทึม ในกรณีที่ผลการทดลองมีค่าใกล้เคียงกันจึงควรปล่อยให้เกิดกระบวนการครอสโอเวอร์แบบสุ่ม เพื่อเป็นการลดพารามิเตอร์ที่ต้องปรับด้วย และการครอสโอเวอร์การสุ่มก็ยืดหยุ่นกว่าการครอสโอเวอร์แบบกำหนดเปอร์เซ็นต์คงที่
- การมิวเตชัน จะให้ผลลัพธ์ที่ดีที่สุดทำการมิวเตชันเมื่อดำเนินการทางเจเนติกผ่านไป 60 รุ่น และเลือกจำนวนโครโมโซมที่จะมิวเตชัน 6% ของจำนวนโครโมโซมทั้งหมด แต่อย่างไรก็ตามไม่ว่า

จะปรับค่าของการมิวเตชันทั้งสองพารามิเตอร์อย่างไร ค่า Accuracy, Recall และ Precision ก็มีค่าใกล้เคียงกัน ทำให้ไม่สามารถระบุแนวโน้มที่แน่นอนได้ เนื่องจากทำการมิวเตชันไม่ได้ทำให้ผลการทดลองดีขึ้นเสมอไปและในบางกรณีอาจทำให้ผลการทดลองแย่ลง ดังนั้นในงานวิทยานิพนธ์นี้จึงจะไม่ใช้มิวเตชันมาพิจารณาร่วมด้วย

- เลือกจำนวนของยีนเธร์โชว์ที่ 6 ยีน เนื่องจากเป็นค่าที่ทำให้ค่าเฉลี่ยของ Accuracy, Recall และ Precision ดีที่สุด เนื่องจากผลการทดลองในบทที่ 5 แสดงให้เห็นว่า การปรับค่ายีนเธร์โชว์น้อยเกินไป เช่น 3 ถึง 5 ยีน จะทำให้ค่าเฉลี่ยของ Accuracy, Recall และ Precision ต่ำมาก เนื่องจากมีความเป็นได้สูงที่ยีนของโครโมโซมที่เป็นอีเมล์ทดสอบจะตรงกับกับยีนในโครโมโซมที่เป็นกฎเพียง 3 ถึง 5 โครโมโซม ซึ่งในความเป็นจริงจะเห็นว่ายังไม่มากเพียงพอที่จะระบุประเภทของอีเมล์ได้ ทำให้ผลลัพธ์ที่ได้ต่ำ ส่วนการปรับค่ายีนเธร์โชว์ที่ 7 ยีนหรือ 8 ยีน ก็จะทำให้ผลลัพธ์ต่ำลง เนื่องจากการให้โครโมโซมที่เป็นอีเมล์ทดสอบจะตรงกับยีนในโครโมโซมที่เป็นกฎเกือบทั้งหมดทุกยีนจึงระบุประเภทของอีเมล์ได้นั้นจะเกิดขึ้นยากมาก

จากพารามิเตอร์ที่ดีที่สุดข้างต้น เมื่อนำมาการทดลองประสิทธิภาพของตัวกรอง โดยใช้เซตข้อมูลทั้ง 5 เซตได้ ค่าเฉลี่ยของ ค่า Accuracy 87.05% ค่า Recall 88.50% และค่า Precision 86.35%

สำหรับการประเมินประสิทธิภาพของอัลกอริทึมที่นำเสนอ ได้ทำการทดลองเปรียบเทียบประสิทธิภาพกับตัวกรองอีเมล์ขยะตามทฤษฎีเบย์เซียน เมื่อเปรียบเทียบประสิทธิภาพทางด้าน Accuracy, Recall และ Precision พบว่าตัวกรองที่นำเสนอมีประสิทธิภาพดีกว่า

6.2 ปัญหาที่พบในวิทยานิพนธ์

6.2.1 ในปัจจุบันมีอีเมล์ขยะจำนวนมากซึ่งพยายามหลีกเลี่ยงการกรอง โดยจะพยายามเลียนแบบรูปแบบของเมล์ดี เช่น ใช้การใช้คำดีเป็นจำนวนมากเข้ามาปนในเนื้อหาของอีเมล์ขยะ(Good Word Chaff) [2] ซึ่งมีผลลัพธ์ต่ำลง

6.2.2 การเลือกชุดประชากรที่จะมาใช้ในการจำแนกอีเมล์นั้น จะเลือกโดยการปล่อยให้ตัวกรองทำตามกระบวนการเจเนติกไปจนครบ Generation ที่ตั้งไว้ จึงค่อยดูผลว่า Generation ไหน ให้ผลของการกรองที่ดีที่สุด แล้วจึงเลือกประชากรชุดนั้นไว้เพื่อมาใช้เพื่อเป็นแม่แบบในการจำแนกอีเมล์ต่อไป ทำให้เสียเวลาค่อนข้างมาก

6.2.3 เนื่องจากงานที่นำเสนอใช้เจเนติกอัลกอริทึมซึ่งต้องใช้เวลาในการประมวลผลค่อนข้างนานและมีพารามิเตอร์ค่อนข้างเยอะ ทำให้เสียเวลาในการปรับพารามิเตอร์

6.3 ข้อเสนอแนะและแนวทางในการพัฒนาต่อ

6.3.1 การแก้ปัญหาการใช้คำดีเป็นจำนวนมากเข้ามาปนในเนื้อหาของอีเมล์ขยะนั้น ทำได้ค่อนข้างยากสำหรับการกรองอีเมล์ขยะที่ใช้ส่วนเนื้อหาในการวิเคราะห์ ดังนั้นเพื่อจะแก้ปัญหาในกรณีนี้อาจจำเป็นต้องใช้เทคนิคอื่นมาช่วยในการพิจารณาร่วมด้วย เช่น ให้ความสำคัญในเรื่องของเวลา เช่น อีเมล์ที่ได้รับในช่วงกลางคืนมีแนวโน้มที่จะเป็นอีเมล์ขยะมากกว่าอีเมล์ที่ได้รับในช่วงกลางวัน

6.3.2 ควรปล่อยให้ระบบหยุดกระบวนการเอง โดยจะหยุดเมื่อค่าความเหมาะสม(Fitness) ของระบบเริ่มลดลงหรือคงที่

6.3.3 ตัวกรองอีเมล์ขยะในวิทยานิพนธ์นี้ ได้เรียนรู้จากอีเมล์ขยะและอีเมล์ดีในปริมาณหนึ่ง และได้ปรับพารามิเตอร์เพื่อให้เหมาะสมกับชุดการทดลองนี้ ดังนั้นหากมีการปรับเปลี่ยนชุดข้อมูลสำหรับเรียนรู้ จึงอาจต้องปรับค่าพารามิเตอร์ใหม่ให้เหมาะสมกับชุดการทดลอง

เอกสารอ้างอิง

[1] Pelletier L., Almhana J. and Choulakian V. "Adaptive Filtering of SPAM" in Proceedings of the Second Annual Conference on Communication Network and Services Research (CNSR'04) , (2004)

[2] Hulten G. and Goodman J. Tutorial on Junk Mail Filtering [Online]. http://research.microsoft.com/~joshuago/icmltutorialannounce.htm

[3] Junk Email Project [Online]. http://clg.wlv.ac.uk/projects/junk-email/

[4] Enron Email Dataset [Online]. http://www.cs.cmu.edu/%7Eenron/

[5] Bryan K. and Yiming Y. Introducing the Enron Corpus [Online]. http://www.ceas.cc/papers-2004/168.pdf

[6] Gawk Program [Online]. http://www.icewalkers.com/Linux/Software/514530/Gawk.html

[7] Frakes, W., "Stemming Algorithms," in *Information Retrieval and Data Structures*, pp. 131-160, Englewood Cliffs, New Jersey: Prentice Hall, 1992.

[8] Open Relay [Online]. http://www.webopedia.com/TERM/O/open_relay.html

[9] Owen, D. 2006. **An Application Agnostic Review of Current Spam Filtering Technique.** [Online]. Available : http://www.danielowen.com

[10] Open Relay [Online]. http://en.wikipedia.org/wiki/Open_proxy

[11] Sahami, M. Dumais, S. Hackerman, D.and Horvitz, E. **A Bayesian Approach to Filtering Junk E-mail.**

[12] Graham, P. 2002. **A Plan for Spam.** [Online]. http://paulgraham.com/spam.html

[13] Yerazuni, W.S. 2004. **The Spam-Filtering Accuracy Plateau at 99.9% Accuracy and How to Get Past It.** *MIT Spam Conference,2004,*]

[14] **ระบบเมล์.** (2006). [Online]. http://www.itwizard.info/technology/general/Mail%20System.htm

[15] Holland J.H. **Genetic Algorithm**. Scientific American. 1992.

[16] ไพศาล เหล่าสุวรรณ. **พันธุศาสตร์**. กรุงเทพฯ : ไทยวัฒนาพานิช. 2535.

[17] http://cs.payap.ac.th/chainat/cs431/reports/g1/ht/history.html

[18] Lawrence David. Handbook of Genetic Algorithm. New York : Van Nostrand Reinhold. 1991.

[19] David E. Goldberg. **Genetic Algorithm in Search, Optimization, and Machine Learning**. Addison-Wesley publishing Company, Inc. 1989. pp. 102-108.

[20] Introduction to Genetic Algorithm [Online]. http://cs.felk.cvut.cz/~xobitko/ga/

[21] Melanie M. **An Introduction to Genetic Algorithm**. MA : The MIT Press. 1998.

[22] Kei Wei, A Naïve Bayes Spam Filter. 2003 [Online]. http://www.eecs.berkeley.edu/~kwei/courses/cs281a/cs281a.pdf

ภาคผนวก ก
ผลงานวิจัยที่ได้รับการตีพิมพ์

Adaptive Spam Mail Filtering Using Genetic Algorithm

Usarat Sanpakdee, Aranya Walairacht and Somsak Walairacht
Department of Computer Engineering, Faculty of Engineering,
King Mongkut's Institute of Technology Ladkrabang
Ladkrabang, Bangkok, Thailand
som_usarat@yahoo.com, {kwaranya, kwsomsak}@kmitl.ac.th

Abstract — **In this paper, we propose a mechanism for filtering incoming spam mails by generating spam mail prototypes using genetic algorithm. Firstly, words from e-mails are extracted and are categorized by their relating meaning into 7 groups. Then, we compose a string of chromosome having 7 genes, i.e., groups of words. Each gene, represented words in each group, is encoded into binary value. The genetic algorithm and its operations are applied to create varieties of spam mail prototypes which inherit from old spam mails. It saves time for preparing training sets and need no large training set for learning like other methods. The spam mail prototypes are the result of this learning mechanism. The experimental results show that the proposed system has efficiency. When testing with both spams and hams, the accuracy is about 85% in average.**

Keywords — **spam mail, spam prototype, spam filtering, genetic algorithm**

1. Introduction

Over the last decade, e-mail or electronic mail has become a popular method of communication because it has more convenient and cheaper than normal postal mail. On the other hand, e-mail has become a victim of abuse. Some business use e-mail for their benefit such as work at home business group, commercial business group. These businesses send a lot of e-mails which context concern about convince, offer prize for something, give assistant, mortgage, medicine, and etc. This kind of e-mail is called, "spam" mail. [1]

There are several meanings of spam, such as, unwanted, junk e-mail message, unsolicited commercial e-mail (UCE), unsolicited bulk e-mail (UBE), and so on. Spam mail does not only annoy e-mail users, it also increases the load of e-mail server and waste of bandwidth. Thus, Internet Service Providers (ISP) must pay more cost for bandwidth and storage. And, e-mail users feel uncomfortable, lose more time because of slower internet and need to pay more cost to use internet too. [2]

Moreover there are a lot of viruses which disguise in spam mail. When the users read their e-mails, they will receive virus attentively. Virus may be a little disturb system but it continually uses victim's computer for distribute spam. Furthermore, spam may create a way for offender to attack and rob benefit from the system. [3]

There are several techniques to classify spam such as header analysis, address list, keyword list, signature analysis, content statistical analysis. But the popular technique is machine learning. Thus, in this paper, we propose a technique for filtering incoming spam mails by generating spam mail prototypes using genetic algorithm. Taking the advantage of evolution mechanism of genetic algorithm, the system can adaptively generate spam prototype for filtering automatically.

2. Spam Mail Prototype

Our system consists of 2 major processes as shown in Figure1. An input e-mail is passed into a process of keyword extraction. Within the process of genetic algorithm, a chromosome represented that e-mail is constructed from the extracted words. The evolution mechanism creates spam mail prototypes as the output of the system.

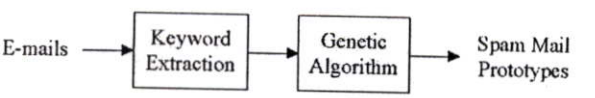


Figure 1. System Block Diagram

The collection of e-mails that considered as spam is called corpus. Spam mails from corpus are encoded to chromosomes and undergo with the genetic operators, i.e., crossover and mutation, and are evaluated by a fitness function. Resulting from the genetic algorithm, rules set or spam mail prototypes are obtained. Figure2 shows a flowchart of spam mail prototypes construction.

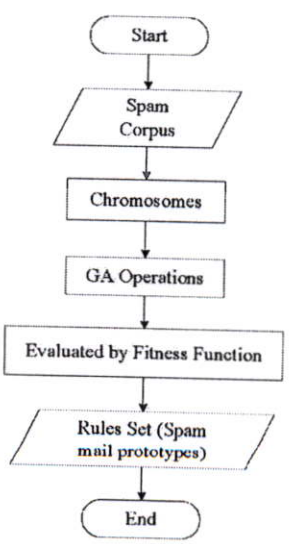


Figure 2. Constructing rules set (spam mail prototypes)

3. Genetic Algorithm

A genetic algorithm (GA) is an algorithm used to find approximate solutions to difficult-to-solve problems through application of the principles of evolutionary biology to computer science. Genetic algorithms use biologically-derived techniques such as inheritance, mutation, natural selection, and recombination (or crossover). Genetic algorithms are typically implemented as a computer simulation in which a population of abstract representations (called chromosomes) of candidate solutions (called individuals) to an optimization problem evolves toward better solutions. Traditionally, solutions are represented in binary as strings of 0s and 1s, but different encodings are also possible. The evolution starts from a population of completely random individuals and happens in generations. In each generation, multiple individuals are stochastically selected from the current population, modified (mutated or recombined) to form a new population, which becomes current in the next iteration of the algorithm.

3.1 Building Representation

An e-mail consists of header, subject and body, for our proposed technique, we extract words within the body part of e-mail only. The extracted words in the form of prepositional, article, and number are discarded. There are 416 words which have been categorized into 7 groups by their relating meaning. Table1 shows all words in each group.

A spam mail is encoded to a binary string. One spam mail represents one chromosome with 7 genes. The binary representation of each gene is computed from words extracted from a spam mail. Figure3 shows a pattern of spam mail chromosome.

	G1	G2	G3	...	G7
Chromosome	weight of G1	weight of G2	weight of G3	...	weight of G7

Figure 3. A pattern of spam mail chromosome

Let's show an example how a chromosome of a spam mail can be constructed. Let an e-mail contains 4 words, namely, "sex", "adult" belong to group G1, and "free", "special" belong to group G2, as shown in Figure 4. We look up table of the weight of these words in data dictionary. The aforementioned weight of word can be calculated by accumulating the frequency of all words and divide by the total number of words in data dictionary. In this case, we have the minimum weight of keyword equal to 0.013 and the maximum weight of keyword equal to 3.31. Then, the minimum and maximum range after normalization becomes 0.004 to 1, or 0000000100 to 1111111000 in binary. Table2 shows computation of the weight of words for this example.

Group	Word	Frequency	$\frac{\text{Frequency}}{\text{Total word(416)}}$	Weight of word	Weight of group
G1	sex	157	0.377	0.114	0.075
G1	adult	49	0.118	0.036	
G3	free	800	1.923	0.581	0.364
G3	special	201	0.483	0.146	

Table 2. Example of calculating weight of word in an e-mail

G1	G2	G3	...	G7
Sex, adult	---	Free, special	...	---

Figure 4. A chromosome before represented to binary string

After we calculate weight of group which average from weight of words that found in the same group. A chromosome, shown in Figure 4, shows weight values in each gene as illustrated in Figure5.

G1	G2	G3	...	G7
0.075	0	0.364	...	0

Figure 5. Weight of each gene in chromosome

The weight of each gene can be encoded into binary string in the following patterns.

Binary 0000000000 represent weight 0.000
Binary 0000000001 represent weight 0.001
Binary 0000000010 represent weight 0.002

.
.

Binary 1111100111 represent weight 0.999
Binary 1111111000 represent weight 1.000

Therefore, weight of G1 gene, 0.075, can be represented by binary value as 0001001011. In the same manner, weight of G3 gene, 0.364, can be represented by binary value as 0101101100. While the rest of genes which has no weight, are represented by binary value 0000000000, as shown in Figure6.

G1	G2	G3	...	G7
0001001011	0000000000	0101101100	...	0000000000

Figure 6. A chromosome representation in binary string

3.2 Genetic Operations

3.2.1 Crossover

For our proposed system, the crossover is allowed for bits of gene within the same group only. We use multiple-point crossover and randomly select the position to cross. In each generation, 15 percent of chromosomes are crossed.

3.2.2 Mutation

Mutation is doing for guarantee that some data will be not disappear. Mutation is done by changing bit in the position which gets from random. In each generation, 2 percent of chromosomes are mutated.

3.3 Evaluation

After e-mails from corpus had been encoded to chromosomes and underwent the operations of genetic algorithm, they are evaluated by the fitness function. The fitness value obtained and used for ranking spam prototypes can be computed from Eq. 1.

$$FitnessFunction = \sum_{i=1}^{i=n} \frac{\text{number of keyword i} \times W_i}{\text{total keywords in an e-mail (n)}} \qquad (1)$$

Where the training weight (W_i) is the summation of weight of any word (w_i) found in each spam mail divided by total e-mails in corpus which we use for training. And w_i is calculated by count number of any word i in each e-mail and divided by total words in that e-mail.

3.4 Selection

After all chromosomes had been evaluated by fitness function, the system selects appropriate chromosomes for filtering incoming e-mails. The selection method used is roulette wheel technique.

4. Rules set for classifying e-mails

The weight of words of gene in testing mail and the weight of words of gene in spam mail prototypes are compared to find match gene. In this proposed system, we specify that if the number of matched gene is greater or equal to 3 then that spam mail prototype will receive one spam score point. The mentioned classification process for spam mails is show in Figure7.

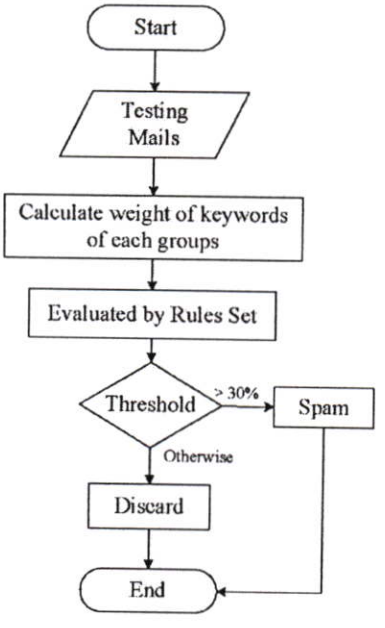


Figure 7. Show the classification process for emails

By comparing testing mails with all spam mail prototypes and the sum of spam score point of all prototypes, if the percentage of spam score point is greater than the percentage of the threshold, then this testing mail is defined to be spam mail. In the experiments, we set the threshold value at 30% in which this threshold value can also be manually adjusted to the appropriate value for optimal result.

4. Experimental Results

We collected spam corpus from [4, 5] and ham corpus from [6]. In the experiments, we use 1,097 of spam mails and 300 of ham (not spam mail).

4.1 Experiment 1

We divided e-mails for training by 80% and for testing by 20% of the total e-mails. The experimental result shows that the average accuracy obtained is 85.53%, with the average values of precision 89.83% and recall of 75.71%. Figure 8 shows the accuracy of training set and validation set in each generation.

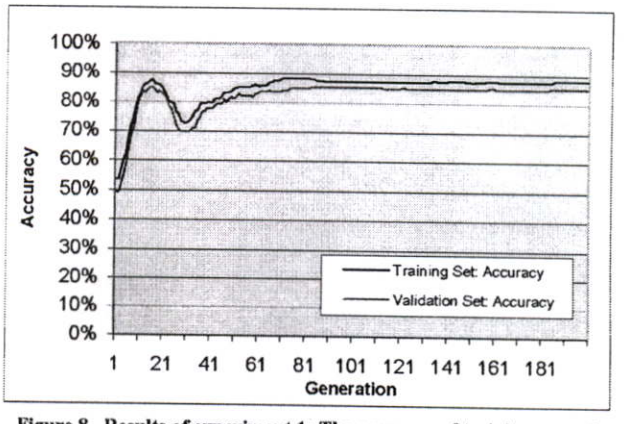


Figure 8. Results of experiment 1: The accuracy of training set and validation set in each generation

4.2 Experiment 2

We divided e-mails for training by 60% and for testing by 40% of the total e-mails. The result shows that the average values for the accuracy is 84.77%, the precision is 90.74% and the recall is 73.13%. Figure 9 shows results of this experiment.

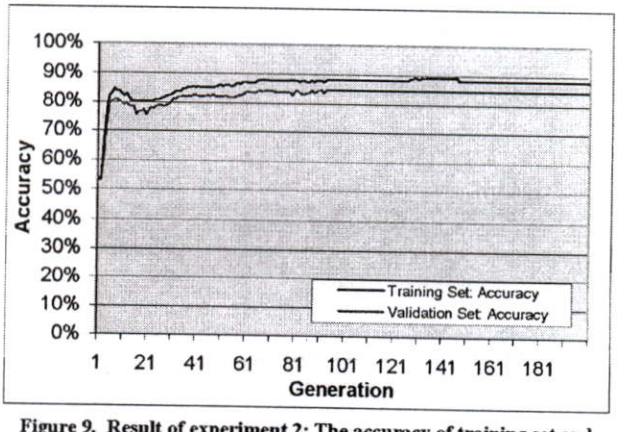


Figure 9. Result of experiment 2: The accuracy of training set and validation set in each generation

4.3 Experiment 3

When we use 100% of spam for testing, the average values of the accuracy obtained is 84.77%, the precision is 90.74% and recall is 73.13%. The results are shown in Figure10.

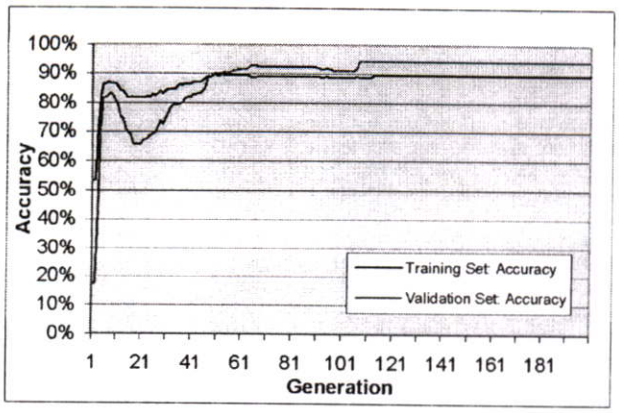


Figure 10. Results of experiment 3: The accuracy of training set and validation set in each generation

4.4 Experiment 4

When we used 50% of spam mails and 50% of ham for testing, the average values of the accuracy is 85.14%, the precision is 90.91% and the recall is 73.86%. The accuracy of training set and validation set in each generation are shown in Figure11.

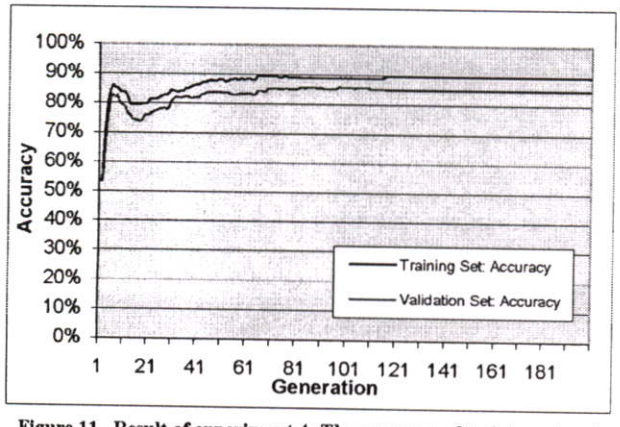


Figure 11. Result of experiment 4: The accuracy of training set and validation set in each generation

5. Conclusions

In this paper, we proposed an adaptive spam mail filtering which uses genetic algorithm and its operations, i.e., crossover and mutation, to create new varieties of spam mail prototypes. The experiments show that proposed adaptive spam mail filtering performs efficient results. In additional, the system allows the threshold value for matching rules set can be manually adjusted to the appropriate level of filtering. In the future, we plan to test our system with more different sets of corpus. The categories and keywords, that represent spam mail from the real situation usages, are to be included to cover broader filtering performance of the system.

REFERENCES

[1] L. Pelletier, J. Almhana, V. Choulakian, 'Adaptive Filtering of SPAM' in Proceedings of the Second Annual Conference on Communication Network and Services Research (CNSR'04), (2004)
[2] A. Boonyu, http://www.dss.go.th/dssweb/st-articles/files/sti_6_2546_spam_mail.pdf
[3] Mehmed Kantardzic(2003). Data Mining: Concept, Model, Medthod, and Algorithms, 'Genetic Algorithms', IEEE Press, 221-245 (2003)
[4] Ramesh Krishnamurthy, Constantin Orasan, A linguistic investigation of the junk emails, http://elg.wlv.ac.uk/projects/junk-email/
[5] Spam Assassin, http://spamassassin.org/publiccorpus/.
[6] Enron Email Data Set, http://www.cs.cmu.edu/~enron/

Table 1. Show words extracted from spam mails categorized in each group

Group	Content	Example of keywords in each group
G1	Adult	adult, aphrodisiac, big, cam, climax, company, cum, desire, erotic, fantasy, fuck, gay, girl, greate, guy, hard, hardcore, heaven, hot, huge, long, man, max, maxlength, nude, orgasm, penis, performance, pheromone, pill, porn, powerful, pussy, satisfy, sex, stamina, sweet, teen, viagra, webcam, x, xxx, xxx-porn, young
G2	Financial	Account, accountant, alert, analyst, attorney, bank, bankruptcy, benefit, bill, billing, broker, budget, building, cash, cheque, commission, consolidate, court, credit, creditor, currency, customer, debt, deposit, discover, economy, entrepreneur, estate, exchange, fee, finance, freedom, fund, help, high-risk, insurance, invest, investor, judgment, legal, legitimate, lender, loan, mastercard, mortgage, obligate, pay, payable, payable, paycheck, promote, purchase, rate, refinance, refund, rent, revenue, risk, service, statement, stock, support, tax, transaction, vat, visa, wealth, worth
G3	Commercial	agency, agent, arrival, bargain, better, brand, buy, camera, cdrom, celeb, chance, cheap, Christmas, collect, college, commerce, computer, cost, deliver, discount, especial, expensive, express, fantastic, free, furnishing, furniture, game, get, gif, gift, great, guarantee, inexpensive, invite, item, just, keyboard, license, lifetime, magazine, maintenance, mall, market, material, materials, mobile, motherboard, mouse, offer, online, only, order, palm, pamphlet, percent, premium, price, produce, product, program, recommend, refill, release, resell, reseller, retail, sale, save, save, sell, ship, shipping, shop, shopping, special, subscribe, supply, surprise, trade, trademark, upgrade, voucher, whole, wholesale, within
G4	Beauty & Diet	after, age, amaze, anti-aging, appetite, beauty, become, before, believe, blood, body, botanic, breast, build, burn, calorie, capsule, card, cell, change, chemical, cholesterol, confirm, course, diet, difference, dose, drug, effect, effective, eliminate, energy, enhance, exercise, eye, face, fast, fat, firm, fit, fitness, flexible, gary, grow, grown, growth, hair, health, healthcare, heart, height, herb, herbal, hormone, improve, inche, incredible, kidney, large, laser, life-changing, light, lose, loss, low, magic, medicine, metabolism, micro-cap, miracle, modem, move, muscle, nature, nutrient, old , over, overweight, permanent, plain, potential, pound, power, protect, reduce, remanufacture, repair, restore, retain, reverse, safe, satisfaction, secret, size, step, strength, strong, tablet, therapy, thin, toxin, treatment, under, virginia, vitamin, weight, woman, wonderful, wrinkle
G5	Traveling	book, deluxe, excite, guide, holiday, honest, hotel, luxury, meal, package, plan, problem, relax, relief, reserve, resort, summer, temple, ticket, tour, train, travel, traveler, trip, vacation,
G6	Home-Based Business	address, astonishment, base, broadcast, bulk, business, comfort, connect, demo, domain, downline, download, earn, email, emailing, ethernet, facemail, fresh, home, homebased, homeworker, host, income, interest, international, internet, investigate, job, list, lucrative, mail, mailbox, mailer, mailing, make, marketing, message, million, money-making, opportunity, part-time, people, private, profit, reach, receive, recipient, require, re-register, return, server, software, subscriber, success, teach, unsubscribe, user, visit, website, work, work-at-home, worker, working
G7	Gambling	action, award, bet, bonus, casino, challenge, extra, gambling, gold, hunt, las, lucky, millionaire, player, poker, prize, reward, rich, vegas, win

ประวัติผู้เขียน

น.ส. อุษารัตน์ แสนปากดี เกิดเมื่อวันที่ 4 พฤศจิกายน พ.ศ. 2524 ที่จ.สกลนคร สำเร็จการศึกษาวิศวกรรมศาสตรบัณฑิต (วิศวกรรมคอมพิวเตอร์) จากมหาวิทยาลัยเทคโนโลยีสุรนารี ปีการศึกษา 2546 จากนั้นในปีการศึกษา 2547 ได้เข้าศึกษาต่อระดับปริญญาโท หลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง