ระบบคัดกรองเมลขยะสำหรับเว็บเบสอีเมล

SPAM MAIL FILTERING FOR WEB BASED E-MAIL

โดย

นันทชัย สมัญญาภรณ์

NANTACHAI SAMANYAPORN

อาจารย์ที่ปรึกษา

ผศ. ดร. จันทร์บูรณ์ สถิตวิริยวงศ์

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 2 ปีการศึกษา 2549

SPAM MAIL FILTERING FOR WEB BASED E-MAIL

NANTACHAI SAMANYAPORN

A SYSTEM DEVELOPMENT PROJECT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2/ 2006

ชื่อหัวข้อ	ระบบคัดกรองเมลขยะสำหรับเว็บเบสอีเมล
นักศึกษา	นายนันทชัย สมัญญาภรณ์
รหัสประจำตัว	47066225
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2549
อาจารย์ที่ปรึกษา	ผศ. ดร. จันทร์บูรณ์ สถิตวิริยวงศ์

บทคัดย่อ

ในปัจจุบันอีเมลเป็นบริการที่ได้รับความนิยมกันอย่างแพร่หลาย และมีการให้บริการผ่านเว็บเบสไคลเอนต์เพื่อช่วยให้สามารถตรวจสอบอีเมลจากที่ใดๆ ก็ได้ผ่านทางเว็บบราวเซอร์ แต่ต้องยอมรับว่ากว่า 80% ของอีเมลที่รับในปัจจุบันเป็นอีเมลที่ผู้รับไม่พึงประสงค์ที่จะรับ ซึ่งได้สร้างปัญหาสำหรับผู้ใช้งานอีเมลต้องพิจารณาคัดแยกอีเมลที่ใช้งานจริงทำให้เสียเวลามาก การนำทฤษฎีความน่าจะเป็นของเบย์ (Bayes) มาใช้พิจารณาอีเมลโดยอ้างจากข้อมูลอีเมลที่ทำการสอนให้ระบบรู้จำมาประมวลผลโดยอัลกอริทึม Bayesian ว่าอีเมลนั้นมีความน่าจะเป็นสแปมมากน้อยเพียงใด เพื่อช่วยลดเวลาในการพิจารณาสแปมของผู้ใช้ได้ดีขึ้น โดยในโครงนี้จะเป็นการพัฒนาปลั๊กอินสำหรับเว็บเบสไคลเอนต์โดยใช้อัลกอริทึม Bayesian

Title	Spam Mail Filtering For Web Based E-Mail
Student	Mr. Nantachai Samanyaporn
Student ID	47066225
Degree	Master of Science
Programme	Information Science
Academic Year	2006
Advisor	Asst. Prof. Dr. Chanboon Sathitwiriyawong

ABSTRACT

In present, Electronic Mail (E-mail) is popular service in Internet. It has web based client for checking e-mail at any where with web browser. But above 80% of e-mails income are spam mail, it makes many problems to user and use long time to classification. Bayesian e-mail filters take advantage of Bayes' theorem to classify emails into categories. This project tries to use Bayesian algorithm to solve false positive problem and reduce time of classification by user better. This project develops plug-in for web base e-mail client to filtering the spam e-mails by using Bayesian algorithm.

กิตติกรรมประกาศ

โครงการพัฒนาระบบงานนี้สำเร็จลุล่วงได้ด้วยการได้รับความช่วยเหลือและความกรุณาจากบุคคลต่าง ๆ เหล่านี้

1. ขอขอบพระคุณ บิดา มารดา ที่ให้โอกาสในการศึกษาเล่าเรียนอย่างเต็มที่
2. ขอขอบพระคุณ ผศ.ดร. จันทร์บูรณ์ สถิตวิริยวงศ์ อาจารย์ที่ปรึกษา ที่ได้กรุณาให้คำปรึกษา แนะนำ สละเวลา ให้การดูแลเอาใจใส่ ช่วยเหลือ ชี้นำ และแก้ไขในสิ่งบกพร่องต่างๆ สำหรับโครงการพัฒนาระบบงานนี้เป็นอย่างมาก
3. ขอขอบคุณ นายธนรัฐ โชติพันธ์ ที่ได้ให้คำปรึกษา แนะนำข้อเสนอต่างๆ ที่เป็นประโยชน์ต่อโครงการพัฒนาระบบงานนี้เป็นอย่างมาก
4. ขอขอบพระคุณคณาจารย์ทุกท่านที่ให้ความรู้มากมาย เพื่อนำความรู้มาใช้ประกอบในโครงการพัฒนาระบบงานนี้
5. ขอบคุณเจ้าหน้าที่ และคณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าคุณทหารลาดกระบัง ที่อำนวยความสะดวกในสถานที่ในการศึกษา ค้นคว้า และปฏิบัติงาน โดยสะดวก และครบถ้วน
6. ขอขอบคุณเพื่อนๆ ทุกคนที่ให้คำปรึกษา ให้กำลังใจและช่วยเหลือในโครงการพัฒนาระบบงานนี้

นายนันทชัย สมัญญาภรณ์

สารบัญ

หน้า

บทคัดย่อภาษาไทย..... I

บทคัดย่อภาษาอังกฤษ..... II

กิตติกรรมประกาศ..... III

สารบัญ.....IV

สารบัญตาราง.....VI

สารบัญรูป.....VII

บทที่ 1 บทนำ.....1

1.1 ความเป็นมาและความสำคัญของปัญหา..... 1

1.2 วัตถุประสงค์.....2

1.3 แนวคิดที่ใช้ในการพัฒนาระบบ.....2

1.4 ขั้นตอนการดำเนินงาน.....2

1.5 ประโยชน์ที่คาดว่าจะได้รับ.....3

บทที่ 2 แนวคิดและทฤษฎีที่เกี่ยวข้อง.....4

2.1 ความรู้พื้นฐานเกี่ยวกับการทำงานของจดหมายอิเล็กทรอนิกส์.....4

2.2 Mail Access Protocol..... 6

2.3 การควบคุมสแปม..... 8

2.4 การกรองสแปม.....10

2.5 การควบคุมสแปมเมล์ในลักษณะอื่นๆ.....13

2.6 กฎของเบย์.....14

บทที่ 3 การออกแบบและพัฒนาระบบ.....18

3.1 การออกแบบระบบ..... 18

3.2 ข้อกำหนดอื่นๆ ที่ใช่ร่วมในระบบ.....23

3.3 ระบบที่ใช้ในการพัฒนา และสภาพแวดล้อมในการพัฒนา.....24

3.4 ความต้องการของฮาร์ดแวร์ในการติดตั้งระบบ.....25

3.5 การติดตั้งระบบ.....25

3.6 การเรียกใช้งาน..... 26

สารบัญ (ต่อ)

หน้า

3.7 การพัฒนาระบบเพื่อเป็นปลั๊กอินให้กับ SquirrelMail.....28

บทที่ 4 การทดลองและผลการดำเนินการ.....29

4.1 วัตถุประสงค์การทดลอง.....29

4.2 เงื่อนไขการทดลอง.....29

4.3 วิธีการทดลอง.....29

4.4 สภาพแวดล้อมในการทดลอง.....30

4.5 ผลการทดลอง.....30

4.6 สรุปผลการทดลอง.....31

บทที่ 5 บทสรุปและข้อเสนอแนะ.....32

5.1 บทสรุป.....32

5.2 ข้อเสนอแนะ.....32

บรรณานุกรม.....36

ประวัติผู้เขียน.....37

สารบัญตาราง

ตารางที่ หน้า

3.1 แสดงรายละเอียดของ Use Case ที่ 1.....20

3.2 แสดงรายละเอียดของ Use Case ที่ 2.....20

3.3 แสดงรายละเอียดของ Use Case ที่ 3.....21

3.4 แสดงรายละเอียดของ Use Case ที่ 4.....21

3.5 แสดงการแยกขอบเขตของอีเมล์ที่นำมาใช้ทดสอบ.....24

สารบัญรูป

รูปที่ หน้า

2.1 แสดงตัวอย่างการส่งต่ออีเมล์จาก SMTP client ไปยัง SMTP server.....5

2.2 แสดงการติดต่อระหว่าง user agent กับ POP3 server 7

2.3 แสดงตัวอย่างคำสั่งในขั้นตอน transaction และ update 7

2.4 แสดงตำแหน่งทีมกนิยมตรวจสอบสแปมเมล.....9

2.5 แสดงหลักการกรองสแปมเมล์ด้วยเทคนิคการทำบัญชีดำ.....11

2.6 แสดงหลักการกรองสแปมเมล์ด้วยเทคนิคการทำบัญชีขาว.....12

3.1 แสดงโครงสร้างของระบบที่ใช้ในการทดลอง.....18

3.2 Use Case Diagram ของระบบงาน.....19

3.3 Activities Diagram ของ Spam Filter System.....22

3.4 แสดงถึงการเข้าใช้งาน SquirrelMail.....26

3.5 หน้า Options เพื่อเข้าถึง PanMail.....27

3.6 แสดงส่วนการเปิดการใช้งาน PanMail.....27

5.1 แสดงตัวอย่างอีเมลที่ไฟล์แนบและจัดเก็บตามมาตรฐาน MIME.....34

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในชีวิตประจำวันเราจะมีการติดต่อสื่อสารกันอยู่ตลอดเวลา ซึ่งในการสื่อสารนั้นช่วยทำให้เราสามารถทำงานร่วมกันได้อย่างราบรื่น และมีประสิทธิภาพ ปัจจุบันระบบอิเล็กทรอนิกส์เข้ามามีบทบาทกับการสื่อสารเป็นอย่างมาก ทำให้เราสามารถใช้การสื่อสารได้อย่างมีประสิทธิภาพ มีความรวดเร็วในการสื่อสาร ทำให้งานที่ทำบรรลุเป้าหมายได้เร็วมากขึ้น ซึ่งการที่เราสามารถรับส่งข่าวสารได้รวดเร็ว จะทำให้เราได้เปรียบในการได้รับรู้ข่าวสาร เพื่อใช้ในการกิจการต่างๆ ได้ทันท่วงที การรับส่งจดหมายอิเล็กทรอนิกส์ (E-Mail) เป็นการสื่อสารแบบอิเล็กทรอนิกส์รูปแบบหนึ่งที่ได้รับความนิยมมาก และมีการใช้งานอย่างแพร่หลาย เนื่องจากเป็นวิธีการติดต่อสื่อสารที่มีค่าใช้จ่ายถูกมาก เมื่อเทียบกับระยะทางการสื่อสารระหว่างผู้ส่งและผู้รับ โดยใช้เวลาเพียงไม่กี่นาทีเท่านั้น ทำให้ในปัจจุบันแทบทุกองค์กรจะมีการใช้งานอีเมลกัน

เมื่อการสื่อสารด้วยวิธีดังกล่าวได้รับความนิยมมากขึ้น อีเมลก็เป็นอีกช่องทางหนึ่งในการประชาสัมพันธ์สินค้าและบริการขององค์กร เนื่องจากว่ามันสามารถเข้าถึงกลุ่มคนได้เป็นจำนวนมาก โดยวิธีการกระจายอีเมลเพื่อหวังผลทางการค้า หรือ กระจายอีเมลเพียงเพื่อรบกวนการทำงานของผู้ใช้ ซึ่งอีเมลเหล่านี้มักจะเป็นอีเมลที่ผู้รับไม่ได้ต้องการที่จะรับแต่อย่างใด และเป็นการทำให้ผู้รับต้องรับภาระในการกำจัด ซึ่งการส่งอีเมลในลักษณะนี้เราเรียกว่า อีเมลขยะ หรือ สแปมเมล (Spam Mail) นั่นเอง

ในการส่งสแปมเมลนั้น จะทำการส่งอีเมลไปหาผู้ใช้อีเมลจำนวนมากเพื่อหวังว่าจะมีผู้ใช้อีเมลจำนวนหนึ่งเปิดอ่านและยอมซื้อหรือใช้บริการสินค้า หรือเชื่อถือในเนื้อหาสาระในอีเมลนั้น ซึ่งเพียงแค่นั้นก็ถือว่าผู้ที่ส่งสแปมเมลได้บรรลุผลที่ต้องการแล้ว เพราะผลที่ได้รับนั้นมันมีความคุ้มค่ามากกว่าค่าใช้จ่ายที่เสียไปในการส่งสแปมเมลในแต่ละครั้ง จึงไม่น่าแปลกใจหากจะเห็นว่าจำนวนของสแปมเมลมีปริมาณมากขึ้นเรื่อยๆ ซึ่งนับวันก็ยิ่งส่งผลกระทบต่อผู้ใช้อีเมล ในการใช้เวลาในการกำจัดสแปมเมลเหล่านี้ และทำให้แบนด์วิดธ์ของระบบเครือข่ายเน็ตเวิร์กต้องเสียไปโดยเปล่าประโยชน์จากการต้องรับส่งเมลเหล่านั้น นั่นหมายถึง ทำให้ผู้ให้บริการอินเตอร์เน็ตต้องมีค่าใช้จ่ายมากขึ้น เพื่อขยายขนาดของแบนด์วิดธ์ เพื่อรองรับการใช้อื่นๆ ซึ่งทำให้ต้องขึ้นค่าบริการที่ต้องเรียกเก็บจากผู้ใช้มากขึ้นด้วย

ในการใช้งานอีเมลในปัจจุบันนั้น นอกจากเราจะใช้งานผ่านทางโปรแกรมเมลไคลเอนต์แล้ว (เช่น Microsoft Outlook, Outlook Express, Mozilla Thunderbird, Endora เป็นต้น) ก็ยังมีการ

ใช้งานผ่านทางเว็บเบสอีเมลอีกด้วย ซึ่งมีข้อดีคือ เราสามารถตรวจสอบอีเมลจากที่ใดก็ได้ โดยผ่านทางการเรียกโปรแกรมเว็บบราวเซอร์เท่านั้น ทำให้ในหลายๆ องค์กรจึงมีการออกแบบให้มีการใช้งานเว็บเบสอีเมลกันมาก ด้วยเหตุนี้จึงได้คิดให้มีการเพิ่มความสามารถในการกรองสแปมเมลผ่านทางระบบเว็บเบสอีเมล เพื่อช่วยอำนวยความสะดวกให้ผู้ลดภาระในการตรวจสอบเมลได้

1.2 วัตถุประสงค์

- ศึกษาการทำงานระบบเน็ตเวิร์ก
- ศึกษาการทำงานระบบอีเมล
- ศึกษาวิธีการคัดแยกสแปมเมล
- สร้างโปรแกรมในส่วนคัดแยกสแปมเมลบนเว็บเบสอีเมล
- ทดสอบประสิทธิภาพในการทำงาน

1.3 แนวคิดที่ใช้ในการพัฒนาระบบ

ในการพัฒนาระบบนี้ ได้มีการนำเอาโครงการพัฒนาระบบงาน "การคัดแยกสแปมเมลโดยอัลกอริทึมเบย์เซียน" มาดัดแปลงเพื่อใช้ในโครงงานนี้ โดยจะนำแนวคิดในโครงงานดังกล่าวเข้ามาร่วมใช้งาน โดยในโครงงานดังกล่าว ได้นำเอาอัลกอริทึมเบย์เซียน ซึ่งเป็นอัลกอริทึมที่เกี่ยวกับการคำนวณทางด้านความเป็นไปได้และสถิติ เข้ามาใช้ในการคำนวณความเป็นไปได้ของอีเมลแต่ละฉบับว่า เป็นสแปมเมลหรือไม่

ในการพัฒนาระบบนั้น ทางผู้พัฒนาได้เลือกใช้ระบบ SquirrelMail เป็นเว็บเบสอีเมล ในการพัฒนาระบบเพิ่มเติมส่วนของการคัดแยกสแปมเมล เนื่องจากตัวโปรแกรมเป็นซอฟท์แวร์แบบเปิดเผยโค้ด (Open Source Software) ทำให้สามารถที่จะนำมาพัฒนาต่อและเพิ่มเติมความสามารถในการทำงานได้ โดยในการพัฒนาระบบ ทางผู้พัฒนาได้เลือกพัฒนาระบบภายใต้สภาพแวดล้อมบนระบบปฏิบัติลีนุกซ์ ซึ่งเป็นระบบปฏิบัติการที่มีความยืดหยุ่นและมีความสามารถในการทำงานสูง อีกทั้งยังมีเครื่องมือในการพัฒนาซอฟท์แวร์เป็นจำนวนมาก อีกทั้งในปัจจุบันทั้งลีนุกซ์ และ SquirrelMail เองก็มีผู้นำไปใช้งานอย่างกว้างขวาง จึงหมาะสมในการนำมาใช้พัฒนาโครงงาน

1.4 ขั้นตอนการดำเนินงาน

- ศึกษาโครงสร้างอีเมล
- ศึกษาโครงงาน "การคัดกรองสแปมเมลโดยอัลกอริทึมเบย์เซียน"

- ออกแบบระบบและทดลองการใช้งาน
- ปรับปรุงระบบให้มีเสถียรภาพ และสามารถใช้งานได้จริง
- จัดทำเอกสารเพื่อนำเสนอผลงาน

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ช่วยอำนวยความสะดวกแก่ผู้ใช้งานอีเมล โดยสามารถช่วยลดปริมาณของอีเมลที่ต้องอ่านและตรวจสอบในแต่ละวันลงได้ โดยในส่วนของเมลที่ถูกระบุว่าเป็นสแปมเมลจะถูกย้ายไปใส่ในอีกไดเร็กทอรี่หนึ่ง เพื่อให้สามารถกลับไปตรวจสอบอีเมลบางฉบับที่อาจไม่ได้เป็นสแปมเมล แต่ตัวระบบระบุว่าเป็นสแปมเมลได้

บทที่ 2
แนวคิดและทฤษฎีที่เกี่ยวข้อง

2.1 ความรู้พื้นฐานเกี่ยวกับการทำงานของจดหมายอิเล็กทรอนิกส์

จดหมายอิเล็กทรอนิกส์ (E-mail หรืออีเมล) จัดเป็นโปรแกรมประยุกต์บนอินเทอร์เน็ตที่นิยมใช้ในการสื่อสาร มีลักษณะคล้ายกับการส่งจดหมายโดยทั่วไป มีความแตกต่างตรงที่มีความรวดเร็วในการส่งไปถึงผู้รับ มีค่าใช้จ่ายต่ำ และมีความง่ายในการใช้งาน นอกจากจะสามารถส่งข้อความที่มีลักษณะในรูปแบบเหมือนจดหมายได้แล้ว อีเมลยังสามารถที่จะแนบส่งไฟล์รูปภาพ ไฟล์เสียง รวมทั้งภาพเคลื่อนไหว ได้อีกด้วย

ระบบอีเมลประกอบด้วย 3 ส่วนหลักคือ User Agents, Mail Server และ Simple Mail Transfer Protocol (SMTP) โดย

- User Agent ซึ่งบางครั้งเรียกว่า mail reader เป็นโปรแกรมที่ ผู้ใช้ใช้ในการอ่านและตอบจดหมายได้
- Mail Server ใช้ในการจัดเก็บ, บริหารจัดการ และเก็บรักษาอีเมล และจัดการเกี่ยวกับกล่องรับจดหมายอิเล็กทรอนิกส์ (Mail Box) ของผู้ใช้แต่ละคน
- Simple Transfer Protocol (SMTP) เป็นโพรโทคอลที่ทำหน้าที่ในการส่งต่ออีเมลระหว่างเครื่องต่างๆ ที่ทำหน้าที่เป็น Mail Server โดยโพรโทคอลนี้จะทำงานอยู่บนชั้น Application Layer ใน OSI Model

ในระบบบนเครือข่ายอินเทอร์เน็ตนั้น การรับส่งอีเมลระหว่างกัน จะทำการส่งผ่านการให้บริการของโพรโทคอล SMTP โดยโพรโทคอลนี้เลือกใช้บริการของ TCP ในการรับส่งอีเมลจาก Mail Server ของผู้ส่งไปยัง Mail Server ของผู้รับ โดย Mail Server ของผู้ส่งจะทำการส่งอีเมลไปยัง Mail Server อื่นๆ ที่ทำหน้าที่เป็นผู้รับ และทำหน้าที่เป็น SMTP client และเมื่อ Mail Server ที่เป็นผู้รับได้รับอีเมลมาแล้ว ก็จะทำหน้าที่เป็น SMTP Server เพื่อที่จะทำการส่งอีเมลนั้น ไปให้ถึง Mail Server ที่เป็นจุดหมายปลายทาง

2.1.1 SMTP

SMTP เป็นส่วนสำคัญในการใช้งานอีเมลโดยจะทำหน้าที่ส่งอีเมลจาก Mail Server ฝั่งผู้ส่งไปยัง Mail Server ฝั่งผู้รับ โดยขั้นแรก SMTP Client จะทำการสร้าง TCP connection บน port 25 กับ SMTP server หลังจากนั้น SMTP client จะทำการบอก Mail Address ของผู้ส่งและผู้รับต่อ

SMTP server เมื่อ SMTP client และ server ดำเนินการในการเชื่อมต่อเสร็จแล้ว SMTP Client จะทำการส่งอีเมลไปยัง SMTP server โดย SMTP จะใช้บริการของ TCP เพื่อส่งอีเมลไปยัง Mail Server และถ้ามีการส่งอีเมลอื่นๆ ไปยัง Mail Server อีกฝั่ง client ก็จะทำกระบวนการนี้ซ้ำในการเชื่อมต่อ TCP เดียวกันนี้ แต่ถ้าไม่มีการส่งอีเมลอื่นอีก จะทำการสั่งให้ TCP ทำการปิดการเชื่อมต่อ

จากรูปที่ 2.1 ต่อไปนี้ กำหนดให้โฮสต์ของ client คือ crepes.fr และโฮสต์ ของ server คือ hamburger.edu

```
S : 220 hamburger.edu
C : HELO crepes.fr
S : 250 Hello crepes.fr, pleased to meet you
C : MAIL FROM : alice@crepes.fr
S : 250 alice@crepes.fr...Sender OK
C : RCPT TO : bob@hamburger.edu
S : 250 bob@hamburger.edu ... Recipent OK
C : DATA
S : 354 Enter mail, end with "." On a line by itself
C : Do you like ketchup?
C : How about pickles?
C : .
S : 250 Message accepted for delivery
C : QUIT
S : 221 hamburger.edu closing connection
```

รูปที่ 2.1 แสดงตัวอย่างการส่งต่ออีเมลจาก SMTP client ไปยัง SMTP server

จากตัวอย่าง client (C:) ส่งอีเมล ("Do you like ketchup? How about pickles?") จาก Mail Server crespes.fr ไปยัง Mail Server (S:) hamburger.edu โดย client จะทำการส่งคำสั่ง 5 คำสั่งคือ HELO, MAIL FROM, RCPT TO, DATA และ QUIT ไปยัง server และจะตอบกลับแต่ละคำสั่งด้วย reply code และคำอธิบาย ในกรณีผู้ส่งต้องการส่งตั้งแต่ 2 อีเมลขึ้นไป ส่งไปยัง Mail Server ฝั่งผู้รับเดียวกัน จะสามารถส่งอีเมลทั้งหมดบนการเชื่อมต่อบน TCP เดียวกันได้

2.1.2 Mail Message Format

ตามรูปแบบโครงสร้างของอีเมลจะประกอบไปด้วยส่วนสำคัญ 2 ส่วนหลัก คือส่วนหัว (Header) ใช้สำหรับเก็บข้อมูลของผู้รับ, ผู้ส่ง, เส้นทางการเดินทางของอีเมล, รวมทั้งประเภทของข้อมูลที่อยู่ในส่วนเนื้อหาด้วย อีกส่วนคือส่วนเนื้อหา (Body) เก็บข้อมูลที่เป็นสาระสำคัญของอีเมลนั้นในรูปแบบต่างๆ เช่น ข้อความ เสียง รูปภาพ หรือภาพเคลื่อนไหว เป็นต้น

ส่วนหัวของอีเมลนั้นประกอบด้วยลำดับของ Header Line อยู่หลายบรรทัด โดยที่ส่วนของ Header Line และส่วนของ Body ของอีเมลจะแยกกันด้วยบรรทัดว่าง โดยทุกอีเมลจะต้องมีส่วนหัวที่เป็น Header line ดังนี้ From : Header line, TO : Header line และ Subject : Header line

อีเมลที่จะส่งไปในการเชื่อมต่อ TCP ประกอบด้วย ส่วนที่เป็น Header ของอีเมล บรรทัดว่าง และ Body โดยบรรทัดสุดท้ายจะเป็นจุด 1 จุด เพื่อบอกว่าจบอีเมลแล้ว

2.1.3 THE MIME Extension for Non-ASCII Data

เนื่องจากส่วนหัวของอีเมลที่กำหนดอยู่ใน RFC 822 สำหรับส่งข้อมูลแบบรหัส ASCII นั้นไม่สามารถทำการส่งข้อมูลประเภท Multimedia ต่างๆ เช่น ไฟล์รูปภาพ ไฟล์เสียง และ ไฟล์วีดีโอได้ ทำให้ต้องทำการเพิ่มส่วนหัวของอีเมลเป็น พิเศษขึ้นมา ซึ่งเรียกว่า MIME (Multimedia Mail Extension) เพื่อใช้ส่งข้อมูลประเภท Multimedia โดยต้องเพิ่มบรรทัดในส่วนหัวของอีเมลเพื่อประกาศรูปแบบเนื้อหา MIME

ประเภทของ MIME มีดังนี้

- **Text** ใช้สำหรับชี้ให้ User Agent ฝั่งผู้รับรู้ว่าส่วนของ Body นั้นประกอบด้วยข้อมูลแบบ text เช่น text/plain
- **Image** ใช้สำหรับชี้ให้ User Agent ฝั่งผู้รับรู้ว่าส่วนของ Body นั้นประกอบด้วยข้อมูลแบบรูปภาพ เช่น image/gif และ image/jpeg
- **Application** ใช้สำหรับชี้ให้ User Agent ฝั่งผู้รับรู้ว่าให้ทำการใดๆ กับส่วนของ Body ด้วยแอพพลิเคชันที่กำหนด เช่น application/msword

2.2 Mail Access Protocol

โพรโทคอลที่ใช้ในการรับอีเมลจาก Mail Server ที่นิยมใช้มีอยู่ด้วยกัน 2 โพรโทคอล ได้แต่ POP3 (Post Office Protocol – Version 3) และ IMAP (Internet Mail Access Protocol) โดย SMTP เป็นโพรโทคอลที่ใช้สำหรับส่งอีเมลระหว่าง User Agent กับ Mail Server หรือระหว่าง Mail Server ด้วยกันเอง ส่วน POP3 และ IMAP จะใช้สำหรับรับอีเมลจาก Mail Server จากฝั่งผู้รับไปยัง user agent ฝั่งผู้รับ

2.2.1 โพรโทคอล POP3

POP3 เป็นโพรโทคอลที่ใช้ในการรับเมลอย่างง่าย โดยการทำงานจะเริ่มจาก User Agent ทำการเปิดการเชื่อมต่อแบบ TCP ไปยัง Mail Server โดยใช้ port 110 เมื่อการเชื่อมต่อแบบ TCP ถูก

สร้างขึ้น POP3 จะมีขั้นตอนการทำงาน 3 ขั้นตอนด้วยกันได้แก่ Authorization, Transaction และ Update โดยจะมีขั้นตอนดังนี้

ขั้นตอนแรก User Agent จะส่งชื่อผู้ใช้และรหัสผ่านของผู้ใช้เพื่อทำการตรวจสอบสิทธิ์ในการเข้าใช้งาน (Authorization) จากนั้นในขั้นตอนที่ 2 User Agent จะทำการสำเนาอีเมลจาก Mail Server และสามารถเลือกอีเมลบน Mail Server ที่จะทำการลบได้ ในขั้นตอนที่ 3 จะเกิดหลังจาก client ส่งคำสั่ง quit เพื่อจบการทำงานของ POP3 ในเวลานั้น Mail Server จะลบอีเมลที่ทำการเลือกไว้ทิ้ง ดูได้จากตัวอย่างในรูปที่ 2.3

ในการทำงานของ POP3 นั้น User Agent จะทำการส่งคำสั่งต่างๆ ออกไป และทาง Server จะตอบกลับแต่ละคำสั่งนั้นๆ ด้วย +OK เพื่อยืนยันว่าได้รับข้อมูลจาก client เป็นที่เรียบร้อยแล้ว หรือ –ERR เพื่อบอกว่าก่อนหน้านี้มีข้อผิดพลาดเกิดขึ้นตามรูปที่ 2.2

```
telnet mailServer 110
+OK POP3 server ready
user bob
+OK
pass hungry
+OK user successfully logged on
```

รูปที่ 2.2 แสดงการติดต่อระหว่าง user agent กับ POP3 server

```
C : list
S : 1 498
S : 2 912
S : .
C : retr 1
S : (blah blah ...
S : .....
S : .....blah)
S : .
C : dele 1
C : retr 2
S : (blah blah ...
S : .....
S : .....blah)
S : .
C : dele 2
C : quit
S : +OK POP3 Server signing off
```

รูปที่ 2.3 แสดงตัวอย่างคำสังในขั้นตอน transaction และ update

หลังจากที่มีการส่งคำสั่ง quit แล้ว POP3 Server จะทำการอัพเดตกล่องจดหมายของผู้ใช้โดยการทำการลบอีเมลที่เลือกเอาไว้แล้ว จะเห็นว่าการทำงานแบบนี้เป็นลักษณะของการทำงาน

แบบ download-and-delete mode ซึ่งก็คือ เมื่อทำการดาวน์โหลดอีเมลลงมาแล้วจะทำการลบอีเมลนั้นออกจากกล่องจดหมายของผู้ใช้ทันที ซึ่งปัญหาที่ตามมาจากการทำงานของ mode นี้จะส่งผลต่อบางกรณีเช่น เมื่อผู้ใช้ทำการอ่านอีเมลจากที่บ้านแล้ว จะไม่สามารถทำการอ่านอีเมลเดิมนั้นได้อีกในที่อื่นๆ นอกจากที่บ้าน แต่ถ้ามีการทำงานที่เป็น download-and-keep mode นั้น จะทำให้สามารถอ่านอีเมลนั้นจากหลายๆ ที่ได้ เพราะไม่มีการลบอีเมลนั้นออกจากกล่องจดหมายของผู้ใช้

2.2.2 โพรโทคอล IMAP

เมื่อผู้รับทำการดาวน์โหลด มายังเครื่องของตนเองโดยใช้โพรโทคอล POP3 ผู้รับสามารถจัดการกับอีเมลที่เครื่องของตนเองเช่น สร้างโฟลเดอร์ที่ใช้เก็บ, ลบ และเคลื่อนย้ายอีเมลระหว่างโฟลเดอร์ได้ แต่ผู้รับไม่สามารถทำการจัดการในลักษณะที่กล่าวมานั้นได้ที่ Mail Server ดังนั้นเพื่อจัดการกับปัญหาตรงนี้โพรโทคอล IMAP (Internet Mail Access Protocol) จึงถูกสร้างขึ้นมาซึ่งมีคุณสมบัติต่างๆ มากกว่า POP3 แต่ก็มีความซับซ้อนมากกว่า POP3 ด้วย

โพรโทคอล IMAP ถูกออกแบบมาเพื่อให้ผู้ใช้สามารถจัดการกับ Remote Mailbox ได้โดยในการทำงานนั้น IMAP Server จะต้องทำการเก็บข้อมูลสถานะของโฟลเดอร์ของผู้ใช้แต่ละคนไว้ซึ่งจะตรงกันข้ามกับ POP3 คือจะไม่เก็บสถานะเกี่ยวกับผู้ใช้เลย โดยมาก IMAP จะถูกนำมาใช้งานในองค์กรธุรกิจ เนื่องจากอีเมลถูกจัดเก็บไว้ที่เดียวเพื่อช่วยในการจัดการเรื่องความมั่นคงของข้อมูลในอีเมล หากมีผู้ใดกระทำการใดๆ กับอีเมลนั้นๆ ก็สามารถกระทำได้เพียงแต่ผู้ที่มีสิทธิเท่านั้น ผู้ใช้ที่รีโมตเข้ามาเพื่อใช้งานก็จะสามารถสืบค้นอีเมลได้ตามลักษณะการจัดเก็บที่คุ้นเคย ทำให้ง่ายต่อการใช้งานไม่เกิดการสับสนในการค้นหาเมล

2.3 การควบคุมสแปมเมล์

มาตราการในการควบคุมสแปมเมลของแต่ละหน่วยงานนั้น โดยปกติแล้วจะมีการออกกฎระเบียบ และการควบคุมสแปมเมลที่แตกต่างกันออกไป แต่ว่าก็ยังสามารถจำแนกแนวทางออกมาได้ 3 ลักษณะหลักดังนี้

- การออกข้อบังคับภายในองค์กร
- การออกมาตรการทางเศรษฐกิจโดยเพิ่มค่าใช้จ่ายในการรับและส่งอีเมล
- การแก้ปัญหาด้วยวิธีทางเทคนิค

2.3.1 การออกข้อบังคับในองค์กร

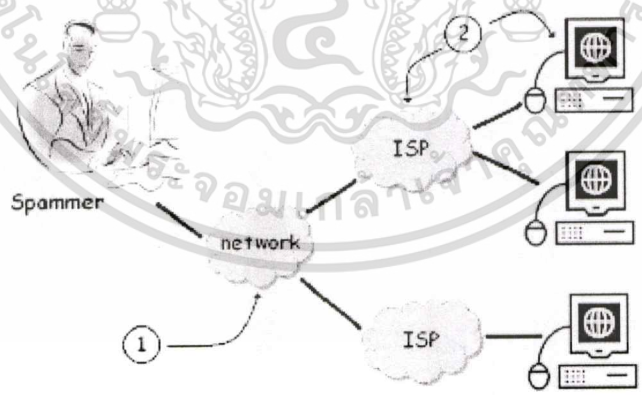
เป็นการแก้ปัญหาในเชิงบริหาร โดยออกนโยบายในองค์กร เพื่อเป็นแนวทางในการปฏิบัติงานต่อการรับและส่งอีเมล หรือให้บริการเป็น Relay Mail โดยส่วนมาก มักจะกล่าวในลักษณะ จะไม่ให้บริการใดๆ แก่ผู้ที่ส่งสแปมเมล เป็นต้น

2.3.2 ออกมาตรการทางเศรษฐกิจ

เนื่องจากการรับและส่งอีเมลมีค่าใช้จ่ายที่ถูกมากต่อการรับส่งอีเมลในหนึ่งครั้ง ทำให้ผู้ใช้ไม่ให้ความสำคัญในการแก้ปัญหาสแปมเมลที่เกิดขึ้นกับตนเองมากนัก จึงมีแนวคิดในการที่จะทำให้ผู้ใช้ดังกล่าวเกิดความตระหนักในการรับและส่งอีเมลในแต่ละครั้ง และหมั่นตรวจสอบกล่องรับจดหมายของตนให้มีสแปมเมลน้อยที่สุด โดยการเพิ่มค่าใช้จ่ายในการรับและส่งอีเมลในแต่ละครั้ง

2.3.3 การแก้ปัญหาโดยวิธีทางเทคนิค

เป็นการแก้ปัญหาทางเทคนิคและเป็นที่นิยมใช้มากที่สุดในการแก้ปัญหาสแปมเมล ตามที่ได้สรุปการทำงานของสแปมเมลตามลักษณะดังกล่าว เราสามารถป้องกันสแปมเมลได้ 2 วิธีหลักคือ ขัดขวางไม่ให้ซอฟต์แวร์ส่งสแปมเมลสามารถทำการส่งผ่านบนระบบเครือข่ายได้ หรือควบคุมการเข้าถึงเมลเซิร์ฟเวอร์โดยสแปมเมอร์ ในกระบวนการนี้จะมีเป้าหมายเพื่อตรวจจับและกำจัดสแปมเมล



รูปที่ 2.4 แสดงตำแหน่งที่นิยมตรวจสอบสแปมเมล

จากรูปที่ 2.4 แสดงถึงตำแหน่งที่ตรวจสอบอีเมลว่าเข้าขายสแปมเมลหรือไม่ โดยจุดที่ 1 ในรูป มักมีลักษณะขัดขวางไม่ให้อีเมลที่เข้าข่ายว่าเป็นสแปมเมลส่งต่อผ่านระบบเครือข่ายไปยังถึงมือผู้รับได้ ส่วนจุดที่ 2 จะเป็นการตรวจสอบซ้อนอีกครั้งจะมีรายละเอียดในการตรวจสอบมากขึ้น มี

รูปแบบในการตรวจสอบเนื้อหาสาระสำคัญของอีเมลว่าจัดอยู่ในข่ายสแปมเมลหรือไม่ ถ้าใช่ก็จะทำการกรองออก หรือรับไว้ทั้งหมดแต่แยกประเภทไว้ให้

2.3.3.1 การขัดขวางสแปมเมล

เป็นวิธีการจัดการกับสแปมเมลได้ตรงจุดที่สุด แต่ตามหลักความเป็นจริงทำได้ยาก คือปิด Relay ทุกเมลเซิร์ฟเวอร์ที่อยู่บนระบบเครือข่ายอินเทอร์เน็ต กำหนดให้โพรโทคอล SMTP ไม่อนุญาตให้มีการส่งต่อเมลที่มีเฮดเดอร์ปลอม และกำหนดให้มีการพิสูจน์ตัวจริงของผู้ส่ง เพื่อง่ายต่อการตรวจสอบผู้ส่งด้วย แต่ในแนวการปฏิบัติจริงๆ คงไม่สามารถกระทำได้ เนื่องจากเราคงไม่สามารถควบคุมเมลเซิร์ฟเวอร์ทุกเครื่องได้ ถึงแม้สามารถทำได้ สแปมเมอร์คงต้องสร้างเส้นทางใหม่ขึ้นมาเองก็ได้โดยอาจแฮคเครื่องคอมพิวเตอร์ของผู้ใช้คนอื่นเพื่อใช้ส่งสแปมเมลอีกทอดก็ได้

2.3.3.2 การพิจารณาอีเมล

ในการขัดขวางสแปมไม่ให้เข้าผ่านระบบนั้นสามารถกระทำได้เพียงในระดับหนึ่งเท่านั้นเนื่องจากพิจารณาเพียงที่มาที่ไปของอีเมลเท่านั้น การจำแนกสแปมด้วยรายละเอียดที่มากขึ้น เช่น พิจารณาจากทั้งเนื้อความของอีเมล สามารถจำแนกว่าอีเมลนั้นเป็นสแปมหรือไม่เป็นได้ดีกว่า ซึ่งจะใช้เทคนิคในการจำแนกกลุ่มของอีเมลปกติ และกลุ่มที่คาดว่าจะเป็นสแปมเมลออกจากกันเพื่อลดปริมาณของสแปมเมลที่จะเข้าถึงผู้ใช้งาน หรือเพื่อลดเวลาในการพิจารณาเลือกอ่านอีเมลของผู้รับได้ดียิ่งขึ้น

2.4 การกรองสแปมเมล์

ในการแก้ปัญหาทางเทคนิคที่กล่าวถึงในข้อ 2.3.3 นั้นสามารถนำเทคนิคในการจำแนกประเภท เพื่อใช้จำแนกสแปมเมลออกจากอีเมลปกติซึ่งมีอยู่หลายวิธีที่นิยมใช้ แต่ที่มักกล่าวถึงสองลักษณะวิธี คือ

- Heuristic Filtering เป็นเรียนรู้ลักษณะของอีเมลและคาดเดาว่าเมลเหล่านั้นเป็นสแปมเมลหรือไม่
- Cooperative Filtering เป็นการร่วมมือกันของกลุ่มผู้ใช้ กำหนดรูปแบบในการสื่อสารร่วมกันเพื่อแยกแยะอีเมลที่ส่งระหว่างกันเป็นอีเมลปกติ ในบางเทคนิคในกลุ่มผู้ใช้จะส่งข้อมูลของสแปมเมลในแต่ละสมาชิกเพื่อทำความเข้าใจตรงกันว่า หากมีอีเมลลักษณะเข้าข่ายผิดปกติตามที่ได้เข้าใจร่วมกันแล้วให้ถือว่าเป็นสแปมเมล

ในส่วนของโครงงาน "การคัดแยกสแปมเมลโดยอัลกอริทึมแบบเบย์เซียน" นั้นจะเป็นกล่าวถึง Heuristic Filtering เนื่องจากทฤษฎีของ Bayes จัดอยู่ในการกรองกลุ่มนี้

2.4.1 Heuristic Filtering

Heuristic Filtering จำแนกออกเป็นสองลักษณะ ดังนี้

- Origin Filtering
- Message Filtering

2.4.1.1 Origin Filtering

เป็นการตรวจสอบก่อนที่จะทำการรับอีเมล โดยใช้ IP address, Domain name ในการตรวจสอบ เป็นการกรองโดยพิจารณาจากที่มาที่ไป และบางส่วนของอีเมล โดยกำหนดเงื่อนไขตั้งไว้ก่อน ว่าตรงกับที่กำหนดไว้หรือไม่ ถ้าตรงก็จะไม่ยอมให้อีเมลนั้นส่งต่อไปถึงมือผู้รับได้ เรียกว่าการทำบัญชีดำ (Blacklist) เป็นวิธีการที่ใช้ได้ตั้งแต่เมลเซิร์ฟเวอร์อื่นๆ ที่ทำหน้าที่เป็น รีเลย์เมลเซิร์ฟเวอร์ จนถึงเมลเซิร์ฟเวอร์ปลายทาง ซึ่งสามารถแยกการเชื่อมต่อทาง IP หรือ TCP ที่เป็นที่มาของสแปมเมลได้ ดังที่แสดงในรูปที่ 2.5 การทำบัญชีดำนิยมตรวจสอบรายชื่อผู้ส่ง หรือชื่อเรื่องของอีเมล แต่วิธีนี้ค่อนข้างจะมีปัญหาอยู่ตรงที่ สแปมเมอร์สามารถที่จะสุ่มชื่อผู้ส่งหรือชื่อเรื่องของอีเมลไม่ใช้ตรงกับบัญชีดำได้

รูปที่ 2.5 แสดงหลักการกรองสแปมเมล์ด้วยเทคนิคการทำบัญชีดำ

อีกลักษณะหนึ่งที่จัดอยู่ในการกรองลักษณะนี้คือการกำหนดให้เซิร์ฟเวอร์ SMTP ทำการตรวจสอบกลับเพื่อเปรียบเทียบค่า IP ของอีเมลที่ส่งมา กับ IP ที่ทำการเชื่อมต่อกับเซิร์ฟเวอร์ มาตรวจสอบว่าตรงกันหรือไม่ ถ้าไม่ตรงกันก็จะทำการแยกออก ซึ่งลักษณะการทำงานนี้สามารถใช้กับรีเลย์เมลในกรณี ที่มีการสุ่มส่งสแปมจำนวนมากเพื่อให้สแปมนั้นอยู่นอกเหนือจากเงื่อนไขที่เมล

เซิร์ฟเวอร์ได้ตั้งไว้ เพื่อจะได้หลุดไปถึงผู้ใช้ ผู้ที่ดูแลเมลเซิร์ฟเวอร์คงต้องกำหนดเงื่อนไขใหม่ๆ เป็นจำนวนมากเพื่อให้มีความสามารถเพียงพอในการกรองสแปม

การทำบัญชีขาว (Whitelist) ก็เป็นอีกทางหนึ่งที่สามารถกรองอีเมลหรือสแปมเมลโดยทำหน้าที่กลับกันกับวิธีการทำบัญชีดำ คืออนุญาตให้ผ่านได้ก็ต่อเมื่ออีเมลมีคุณสมบัติตรงกับเงื่อนไขที่ระบุไว้ เช่น กรองจากอีเมลแอดเดรสของผู้ส่งหรือจากโดเมนที่ไม่รู้จัก เมื่อผู้ส่งหรือโดเมนดังกล่าวไม่มีอยู่ในรายการในบัญชีขาวจะไม่ให้ส่งผ่านไปได้ ข้อเสียของระบบนี้คือค่อนข้างยุ่งยากในการตรวจสอบอีเมลที่ไม่อยู่ในรายการ อีกทั้งยังมีการตรวจสอบผู้ส่งรายใหม่เพื่อปรับปรุงในระบบด้วย และถ้าสแปมเมอร์สามารถที่จะสุ่มชื่อโดเมนให้ตรงกับที่มีอยู่ในระบบก็สามารถเข้าถึงเป้าหมายได้เช่นกัน แสดงตามรูปที่ 2.6

รูปที่ 2.6 แสดงหลักการกรองสแปมเมลด้วยเทคนิคการทำบัญชีขาว

2.4.1.2 Message Filtering

เป็นการตรวจสอบอีเมลที่รับมาแล้วว่าเป็นสแปมเมลหรือไม่ โดยใช้บางส่วนของอีเมล เช่น IP address, Domain name, คำแต่ละคำ เป็นต้น ใช้เปรียบเทียบข้อความในอีเมลเพื่อจำแนกประเภทการพิจารณาจากตัวของอีเมล เทคนิคนี้มีความจำเป็นต้องรู้ว่าสแปมเมอร์มักจะใช้ข้อความหรือคำใด (Keyword) ในการตั้งชื่อเรื่องอีเมล หรือเขียนข้อความอีเมล คุณลักษณะอื่นๆของสแปมเมลนำมาใช้ร่วมกับเทคนิคในการวิเคราะห์ความน่าจะเป็นว่าอีเมลนั้นเป็นสแปมเมลหรือไม่ นิยมใช้เทคนิค Naïve Bayesian filtering เพื่อหาค่าความน่าจะเป็นแล้วนำค่านั้นไปพิจารณาว่าเป็น สแปมเมลหรือไม่ ก่อนที่จะใช้งานเทคนิคเราต้องสอนให้ Agent (Preprocessing) ของเทคนิคนี้รู้จักเซตของคำที่ อยู่ในข่ายสแปม และรู้จักเซตของคำที่ไม่ใช่สแปมเพื่อใช้ในการแยกประเภท

2.4.2 Cooperative Filtering

2.4.2.1 Content labeling

เป็นการเพิ่มส่วนพิเศษ (Label) รวมไว้ในอีเมล ซึ่งลาเบลนี้จะแสดงถึงข้อมูลใดๆ ที่สามารถระบุหรือรับรองได้ว่ามาจากผู้นั้นจริงๆ ข้อมูลในลาเบลจะสามารถให้ผู้รับหรือเครื่องคอมพิวเตอร์ฝั่งผู้รับสามารถแจกแจงได้ทันทีว่า นั่นเป็นอีเมลปกติหรือเป็นสแปมเมล

เนื้อหาสาระของลาเบลอาจจะเป็นการบ่งบอกถึงช่องทางในการติดต่อกัน ระหว่างผู้รับและผู้ส่ง ยกตัวอย่างเช่นอีเมลที่ส่งจากผู้ส่งสแปมเมลจะได้โพรโทคอล SMTP TCP ที่ port 25 แต่การส่งจากผู้ใช้ในกลุ่มที่ตกลงกันอาจเพิ่ม label ที่ระบุหมายเลข port อื่นที่ผู้ส่งสแปมเมลไม่ได้รู้ด้วย

เทคนิคนี้ยังไม่สามารถลดจำนวนสแปมเมลไปได้เท่าที่ควรยังคงต้องใช้ควบคู่กับเทคนิคในการกรองเมลประเภทอื่นๆ นอกจากนี้ยังมีเทคนิคที่นำมาใช้ในการกรองสแปมเมลอีกหลายเทคนิค เช่น genetic algorithms, neural network และอื่นๆ

2.5 การควบคุมสแปมเมลในลักษณะอื่นๆ

2.5.1 Accountability

การสื่อสารอีเมลผ่านเครือข่ายอินเทอร์เน็ตผู้ใช้มีความจำเป็นต้องติดต่อสื่อสาร กับบุคคลที่เชื่อถือได้ มีตัวตนอยู่จริง และไม่สามารถปฏิเสธความรับผิดชอบต่อการรับและส่งอีเมลนั้นได้ จึงมีการนำกลไกใบรับรองดิจิทอล (Digital Signature) เช่น S/MIME และ PGP/MIME มาใช้งาน โดยจำแนกออกเป็น 2 ลักษณะดังนี้

- End to End authentication
- First-Hop accountability

2.5.1.1 End to End authentication

เป็นลักษณะการพิสูจน์ตัวจริงผู้ส่งอีเมลโดยพิจารณาจากใบรับรองดิจิทอล ที่แนบมากับอีเมลนั้นที่รับรองโดยหน่วยงานออกใบรับรองที่น่าไว้วางใจแห่งหนึ่ง ใบรับรองดิจิทัลจะแสดงรายละเอียดที่บ่งบอกถึงตัวผู้ส่ง จะถือได้ว่าอีเมลฉบับนั้นมีความน่าเชื่อถือได้เท่ากับที่ผู้รับชื่อถือที่รับรองใบรับรองดิจิทัลนั้น หน่วยงานที่ออกใบรับรองดิจิทอล (Certificate Authorities: CA) ได้แก่ธนาคาร หน่วยงานรัฐบาล หรือองค์กรเอกชน เป็นต้น

2.5.1.2 First-Hop accountability

นำเอาแนวคิด End to End Authentication มาปรับปรุงใช้ในการพิสูจน์ตัวจริงของเซิร์ฟเวอร์ SMTP ในเบื้องต้นระบบจะทำการปฏิเสธอีเมลที่จากแหล่งที่มาที่ไม่รู้จัก ซึ่งเป็นการทำงานที่เกี่ยวข้องกับการติดต่อกันระหว่างสองส่วนคือ ระหว่าง First-Hop เซิร์ฟเวอร์ของผู้ให้บริการกับลูกค้า และระหว่างเซิร์ฟเวอร์ที่ทำหน้าที่เป็นรีเลย์

CA จะสร้างใบรับรองดิจิทอลที่เป็นที่น่าเชื่อถือต่อเซิร์ฟเวอร์ตัวแรก (First-Hop) ให้แก่โฮสต์ที่ทำการส่ง เพื่อเป็นการแสดงว่าผู้ส่งนั้นเป็นผู้ส่งที่ถูกรับรองโดย CA โดยใบรับรองดิจิทัลที่ถูกสร้างขึ้นมานี้จะมีอายุที่สั้นกว่าใบรับรองที่ออกให้กับผู้ใช้ (Client Certificate) ซึ่งถ้ามีการแก้ไขหรือทำการลอกเลียนแบบจะทำให้ระบบสามารถตรวจจับได้อย่างรวดเร็ว

ในการควบคุมในลักษณะนี้มุ่งเน้นที่จะตรวจสอบว่าผู้ส่งคือใครและมีอยู่จริงหรือไม่ ซึ่งก็มีหลายวิธีเช่น การตรวจสอบจากข้อมูลการเชื่อมต่อกับระบบว่ามาจากโฮสต์ใด ตรวจสอบจากชื่อโดเมนที่จดทะเบียนว่าเป็นของผู้ใด หรือเรียกร้องให้ต้องมีการล็อกอินต่อเซิร์ฟเวอร์ SMTP

2.6 กฎของเบย์

2.6.1 การนำกฎของเบย์ (Bayesian) มาใช้

การพิจารณาคัดแยกสแปมเมลออกจากอีเมลปกติ หรืออีเมลที่ไม่ใช่สแปมด้วยวิธีการทำบัญชีดำ เป็นวิธีที่พบเห็นและนิยมใช้ทั่วไปในปัจจุบัน เนื่องจากมีค่าใช้จ่ายต่ำ และนำไปใช้งานง่าย แต่มีปัญหาในการคัดแยกว่าคัดแยกผิดเสียเป็นส่วนใหญ่ เนื่องจากการทำบัญชีดำ จะพิจารณาอีเมลตามเงื่อนไขที่โปรแกรมหรือระบบตั้งไว้ตามลำดับ อีเมลฉบับเดียวกันอาจจะผ่านเงื่อนไขบางเงื่อนไข และไม่ผ่านเงื่อนไขบางเงื่อนไขได้ ทำให้อาจเกิดความผิดพลาด เพราะมีผลขึ้นอยู่กับลำดับของเงื่อนไขที่ตั้งไว้ รวมถึงข้อความหรือคำบางส่วนที่นำมาพิจารณา เช่น ชื่อเรื่องจดหมาย (Subject) อาจมีโอกาสเป็นได้ทั้ง SPAM และ HAM ก็ได้จึงสร้างปัญหาในการตั้งเงื่อนไข อีกทั้งบางส่วน หรือบางคำของข้อมูล อาจไม่สามารถนำมาตั้งเงื่อนไขได้ เนื่องจากข้อจำกัดของตัวซอร์ฟแวร์เอง ทำให้ไม่ยุติธรรมต่อการคัดแยก มีผลทำให้การคัดแยกผิดพลาดได้ ดังนั้นในการคัดแยกสแปมเมลนั้น จึงควรนำเอาส่วนต่างๆ ของเมลมาประกอบการพิจารณาคัดแยก เพื่อทำให้เกิดความถูกต้องในการพิจารณามากขึ้น

2.6.2 ทฤษฎีของเบย์ (Bayesian)

เนื่องด้วยทฤษฎีของเบย์ หรือ Bayesian ได้กล่าวถึงความน่าจะเป็นของการเกิดเหตุการณ์ใดๆ เมื่อรู้เหตุการณ์อื่นไว้พอสังเขปดังนี้

กำหนดให้ A และ B เป็นเหตุการณ์ใดๆ ความน่าจะเป็นของ A เมื่อรู้ B (ความน่าจะเป็นที่จะเกิดเหตุการณ์ A โดยมีเงื่อนไขว่าเหตุการณ์ B ได้เกิดขึ้นแล้ว) เขียนแทนด้วย P(A|B) สามารถคำนวณได้ด้วยทฤษฎีของเบย์ดังนี้

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

กล่าวคือความน่าจะเป็นของ A เมื่อรู้ B สามารถคำนวณได้จากผลคูณของความน่าจะเป็นของ B เมื่อรู้ A กับความน่าจะเป็นของ A หารด้วยความน่าจะเป็นของ B เราเรียก P(A) ว่าเป็นความน่าจะเป็นเบื้องต้น (prior probability) และเรียก P(A|B) เป็นความน่าจะเป็นภายหลัง (posterior probability)

ความน่าจะเป็นเบื้องต้นเป็นค่าที่ได้จากข้อมูลเบื้องต้น ส่วนความน่าจะเป็นภายหลังเป็นค่าความน่าจะเป็นก่อนที่ถูกปรับด้วยข้อมูลที่เพิ่มขึ้น

2.6.3 การนำทฤษฎีของเบย์มาใช้ในการคัดแยกสแปมเมล

จากทฤษฎีในข้อ 2.6.2 เราสามารถนำไปใช้ในการเรียนรู้ เพื่อคัดแยกสแปมเมลโดยเขียนแทนการคำนวณได้ดังนี้

$$P(spam|words) = \frac{P(words|spam)P(spam)}{P(words)} \tag{2.2}$$

แปลว่าความน่าจะเป็นแบบสุ่มที่คำที่เป็น SPAM จะอยู่ในเซตของเอกสารเท่ากับ ความน่าจะเป็นของคำที่ปรากฏอยู่ในเซตของคำที่เป็น SPAM กับความน่าจะเป็น SPAM หารด้วยความน่าจะเป็นของคำนั้น และเมื่อความน่าจะเป็นของคำเท่ากับ ความน่าจะเป็นของคำที่ไปปรากฏอยู่ในเซตของ SPAM กับความน่าจะเป็น SPAM บวกด้วยความน่าจะเป็นของคำที่ปรากฏอยู่ในเซตของคำที่ไม่ใช่ SPAM (HAM) กับความน่าจะเป็นของคำที่ไม่ใช่ SPAM และเมื่อ $P(words)$ เขียนให้อยู่ในรูปอื่นได้ดังนี้

$$P(words) = P(words|spam)P(spam) + P(word|\overline{spam})P(\overline{spam}) \tag{2.3}$$

เราสามารถเขียนใหม่ในรูปต่อไปนี้

$$P(spam \mid words) = \frac{P(words \mid spam)P(spam)}{P(words \mid spam)P(spam) + P(word \mid \overline{spam})P(\overline{spam})} \quad (2.4)$$

สามารถเขียนเป็นสมการในรูปได้ว่า

$$P(spam_i \mid words) = \frac{P(words \mid spam_i)P(spam_i)}{\sum_j P(words \mid spam_j)P(spam_j)} \quad (2.5)$$

ในการนำไปให้ยกตัวอย่างในกรณีที่เรารู้คำ 2 คำ แทนด้วย a,b จะถูกแทนค่าด้วย

$$P(a \mid b) = \frac{ab}{ab + (1-a)(1-b)} \quad (2.6)$$

และเมื่อเรารู้คำ 3 คำ แทนด้วย a,b,c จะถูกแทนด้วย

$$P(a \mid b \mid c) = \frac{abc}{abc + (1-a)(1-b)(1-c)} \quad (2.7)$$

ค่าความน่าจะเป็นที่ได้จะมีค่าตั้งแต่ 0 - 1

- เมื่อ 0 แสดงว่ามีความน่าจะเป็น SPAM 0%
- เมื่อ 1 แสดงว่ามีความน่าจะเป็น SPAM 100%

จากกฎของเบย์ข้างต้น บอกถึงแนวทางในการนำกฎนี้ไปใช้ได้ว่า เมื่อรับอีเมลเข้ามาเราต้องแยกข้อความในอีเมลนั้นออกเป็นคำๆ ก่อน เพื่อที่จะนำคำเหล่านั้นไปหาความน่าจะเป็นในเงื่อนไขที่ว่า ถ้ามีคำเหล่านี้ปรากฏในเอกสาร มีความน่าจะเป็นมากน้อยเพียงใดที่อีเมลฉบับนี้จะเป็น SPAM แต่เนื่องจากคำเหล่านี้ ก่อนจะนำไปใช้กับกฎดังที่กล่าวมาจำเป็นต้องมีค่าตัวเลขที่แสดงถึงน้ำหนักของคำแต่ละคำนั้นว่า แต่ละคำนั้นมีน้ำหนักโอนเอียงไปทาง SPAM หรือ HAM โดยค่าดังกล่าวเราเรียกว่าค่า Weight ในระบบเราจะเรียกค่า Weight ว่าค่า SPAMCITY และค่าSPAMCITY ของแต่ละคำ หาได้จากสมการดังนี้

$$\text{SPAMCITY}_k = \frac{\text{Spam Probability}_k}{\text{Ham Probability}_k + \text{Spam Probability}_k} \quad (2.8)$$

โดย $\text{Ham Probability}_k = \frac{tf_k}{N_h}$ เมื่อ

Ham Probability_k = ความน่าจะเป็น HAM คำในลำดับที่ k (Term ที่ k)

tf_k = ความถี่ของคำที่ k ที่ปรากฏในเอกสารที่นำมาทดสอบทั้งหมด (N_h)

N_h = จำนวนเอกสารที่นำมาทำ Preprocessing ที่อยู่ใน HAM set

โดย $\text{Spam Probability}_k = \frac{tf_k}{N_s}$ เมื่อ

$\text{Spam Probability}_k$ = ความน่าจะเป็น SPAM คำในลำดับที่ k (Term ที่ k)

tf_k = ความถี่ของคำที่ k ที่ปรากฏในเอกสารที่นำมาทดสอบทั้งหมด(N_s)

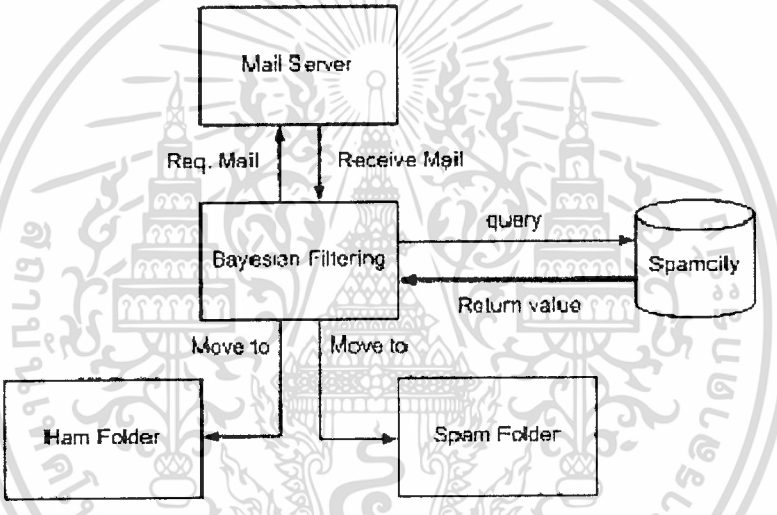
N_s = จำนวนเอกสารที่นำมาทำ Preprocessing ที่อยู่ใน SPAM set

บทที่ 3
การออกแบบและพัฒนาระบบ

3.1 การออกแบบระบบ

3.1.1 โครงสร้างระบบ

ในการพัฒนาระบบได้ออกแบบโครงสร้างของระบบโดยรวมเพื่อให้เข้าใจ การทำงานโดยรวมดังรูปที่ 3.1



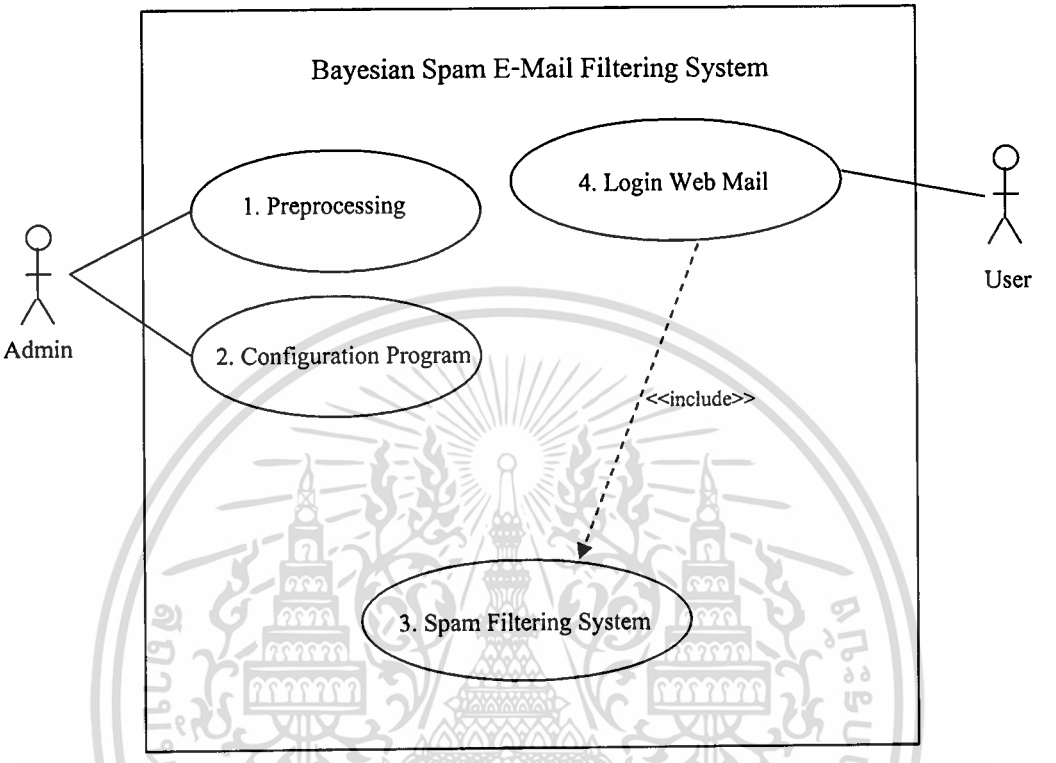
รูปที่ 3.1 แสดงโครงสร้างของระบบที่ใช้ในการทดลอง

โดยระบบจะถูกแทนโดย Bayesian Filtering (ซึ่งจะนำมาจากโครงงาน "การคัดแยกสแปมเมลโดยอะลกอริทึมเบย์เซียน") ทำหน้าที่เข้าไปรับอีเมลจาก Mail Server ด้วย account ที่กำหนดไว้ก่อน เมื่อรับเมลสำเนาไว้แล้วจะทำการแยกออกเป็นคำที่ไม่ซ้ำกัน ใช้เป็นคำหลักที่นำไปสืบค้นเพื่อนำค่าของ SPAMCITY ที่ทำการหาไว้แล้วมาใช้ คำนวณตามอัลกอริทึมเบย์เซียนเพื่อหาความน่าจะเป็นว่าอีเมลแต่ละฉบับมีความน่าจะเป็น SPAM หรือไม่

3.1.2 Use Case Diagram

จากการศึกษาลักษณะการคัดแยกประเภทของอีเมล นำมาออกแบบระบบและเขียนเป็นUse Case Diagram ได้ดังรูปที่ 3.2 โดยระบบจะแยกออกเป็นสามส่วนหลัก คือส่วนที่หนึ่งทำการเตรียมข้อมูล (Preprocessing) เพื่อหาค่า SPAMCITY จากอีเมลที่ใช้นำมาทดลอง ส่วนที่สองเป็นการตั้งค่า

mail account เพื่อใช้ในการรับเมลมาจากเซิร์ฟเวอร์ และอีกส่วนคือส่วนที่ใช้ในการคำนวณความใกล้เคียงของอีเมลกับ SPAM เมล ซึ่งเขียนไว้ใน Use Case ที่ 1-3 ตามลำดับ



รูปที่ 3.2 Use Case Diagram ของระบบงาน

จากรูปที่ 3.2 แสดงการทำงานโดยผู้ใช้ต้องทำการปรับแต่งค่าต่างๆ เช่น การระบุที่อยู่ของเมลเซิร์ฟเวอร์ การระบุ mail account และ รหัสผ่าน เมื่อปรับแต่งค่าเหล่านี้เสร็จแล้วสามารถสั่งให้ระบบทำงาน จากนั้นระบบจะทำงานเองคือ ติดต่อกับเมลเซิร์ฟเวอร์เพื่อรับอีเมล ทำการคัดแยกประเภทของอีเมลว่าเป็น SPAM หรือ HAM และทำการส่งไปอีเมลไปยังโฟลเดอร์ HAM และ SPAM ที่กำหนดไว้

3.1.3 Use Case Description

จาก Use Case Diagram สามารถเขียนอธิบายแต่ละ Use Case ดังตารางที่ 3.1 – 3.7

ตารางที่ 3.1 แสดงรายละเอียดของ Use Case ที่ 1

Use Case	1. Preprocessing
Brief Description	เตรียมค่า SPAMCITY เพื่อนำไปใช้ในการคัดแยก
Actor	Admin
Trigger	-
Pre-condition	-
Post-condition	ระบบมีค่าชื่อเริ่มต้นสำหรับใช้งาน
Primary scenario	1. ทำการติดตั้งระบบดาต้าเบส MySQL 2. ทำการสร้างดาต้าเบส 3. ทำการนำเข้าข้อมูลเพื่อทำเป็น Preprocessing
Alternatives	-

ตารางที่ 3.2 แสดงรายละเอียดของ Use Case ที่ 2

Use Case	2. Configuration System
Brief Description	ตั้งค่าเริ่มต้นระบบก่อนให้ระบบและ User เริ่มทำงาน
Actor	Admin
Trigger	ผู้ใช้แก้ไขไฟล์ ทำการแก้ไขรายละเอียดผ่านทางหน้า Options
Pre-condition	-
Post-condition	ระบบมีค่าชื่อเริ่มต้นสำหรับใช้งาน
Primary scenario	1. ทำการติดตั้งระบบเมลเซิร์ฟเวอร์ และติดตั้ง SquirrelMail 2. ทำการปรับแต่งค่าที่ SquirrelMail เพื่อให้สามารถใช้งาน PanMail ได้ 3. ทำการแก้ไข PanMail ในส่วนของการกำหนดค่าในการติดต่อกับระบบดาต้าเบส MySQL
Alternatives	-

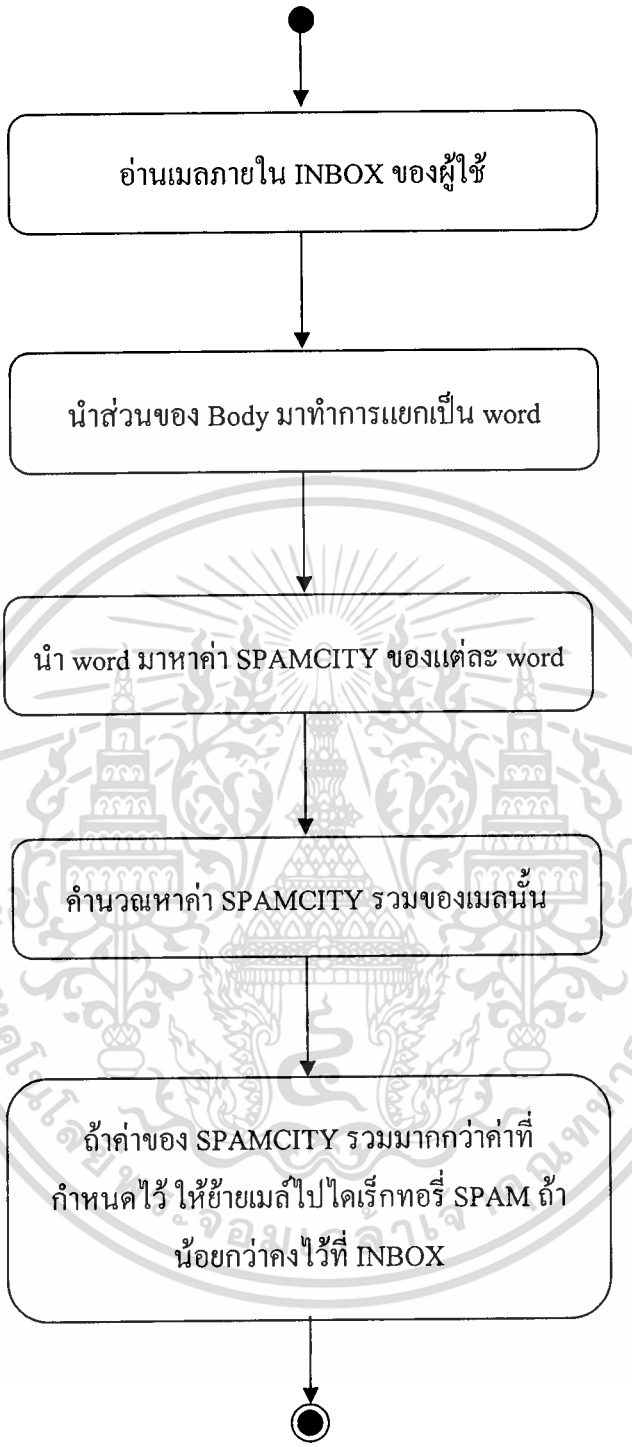
ตารางที่ 3.3 แสดงรายละเอียดของ Use Case ที่ 3

Use Case	3. Spam Filter System
Brief Description	สั่งให้ระบบเริ่มทำงาน
Actor	Login Web Mail
Trigger	-
Pre-condition	-
Post-condition	รับอีเมลจากเมลเซิร์ฟเวอร์ คัดแยกประเภทอีเมล และจัดส่งสู่ผู้รับ
Primary scenario	-
Alternatives	-

ตารางที่ 3.4 แสดงรายละเอียดของ Use Case ที่ 4

Use Case	4. Login Web Mail
Brief Description	ทำการล็อกอินเว็บเบสอีเมล์เพื่อเข้าใช้งาน
Actor	User
Trigger	-
Pre-condition	-
Post-condition	เข้าใช้งานระบบได้
Primary scenario	1. ทำการล็อกอินด้วย username และ password ที่ได้รับจาก Admin
Alternatives	-

ในส่วนการทำงานของ Use Case ที่ 3 ที่เป็นการทำงานของระบบ Spam Filter System นั้น เราสามารถที่จะเขียน Activities Diagram ของ Use Case ดังกล่าวได้ดังรูปที่



รูปที่ 3.3 Activities Diagram ของ Spam Filter System

3.2 ข้อกำหนดอื่นๆ ที่ใช้ร่วมในระบบ

3.2.1 ขอบเขตของเนื้อหาที่ใช้พิจารณาและการแบ่งคำ

ขอบเขตการพิจารณาเนื้อความจดหมายอิเล็กทรอนิกส์เลือกพิจารณาครอบคลุมส่วนหัวของอีเมล และเนื้อความของจดหมายที่เป็นตัวอักษรเท่านั้น ไม่ได้พิจารณาไฟล์ที่แนบมาด้วยทุกชนิด การพิจารณาแยกคีย์เวิร์ดจากอีเมล แยกตามคีย์เวิร์ดจากเครื่องหมายต่อไปนี้

(	เครื่องหมายวงเล็บเปิด
)	เครื่องหมายวงเล็บปิด
:	เครื่องหมายทวิภาค
@	เครื่องหมายแอต
<	เครื่องหมายน้อยกว่า
>	เครื่องหมายมากกว่า
\n	ตัวอักษรแบบ white space - ขึ้นบรรทัดใหม่
\r	ตัวอักษรแบบ white space - เครื่องหมาย Return
\t	ตัวอักษรแบบ white space - เครื่องหมาย Tab
"	เครื่องหมายอัญประกาศคู่
'	เครื่องหมายอัญประกาศเดี่ยว
,	เครื่องหมายจุลภาค
!	เครื่องหมายอัศเจรีย์
?	เครื่องหมายปรัศนีย์
#	เครื่องหมายนัมเบอร์ หรือ เครื่องหมายชาร์ป
$	เครื่องหมายดอลล่า
&	เครื่องหมายแอมเพอร์แซนด์
*	เครื่องหมายดอกจัน
+	เครื่องหมายบวก
/	เครื่องหมายหาร
{	วงเล็บปีกกาเปิด
}	วงเล็บปีกกาปิด
[	วงเล็บก้ามปูเปิด
]	วงเล็บก้ามปูปิด
=	เครื่องหมายเท่ากับ
;	เครื่องหมายอัฒภาค

กล่าวคือเมื่อโปรแกรมเปิดอ่านเอกสารและพบเครื่องหมายดังรายการที่กำหนด จะทำการตัดคำโดยใช้เครื่องหมายดังกล่าวตัวแยก

3.2.2 ขอบเขตของข้อมูลที่นำมาพิจารณา

จากการแบ่งคำ ในอีเมลที่มีความยาวมาก จะพบว่ามีคำที่แบ่งเป็นจำนวนมากซึ่งมีผลกับเวลาในการคำนวณ ในการทดลองมีการกำหนดขอบเขตของข้อมูลในตัวอีเมลเพื่อใช้ในการพิจารณาเปรียบเทียบกันระหว่างความเร็วในการพิจารณาคัดแยก และความถูกต้อง โดยแบ่งออกเป็น 4 ลักษณะดังตารางที่ 3.5

ตารางที่ 3.5 แสดงการแยกขอบเขตของอีเมลที่นำมาใช้ทดสอบ

ขอบเขตที่	ชื่อขอบเขตบนโปรแกรม	ความหมาย
1	Full	พิจารณาข้อความแบบ Text ทั้งหมด
2	Some Content	พิจารณาบางส่วนของ Header เฉพาะบางคีย์ คือ Received, To, Bcc, Cc, From และ Subject ร่วมกับส่วนของ Body ที่มีข้อมูลเป็นข้อความ Text
3	Body	พิจารณาส่วนของ Body ที่มีข้อมูลเป็นข้อความ Text
4	Header	พิจารณาบางส่วนของ Header เฉพาะบางคีย์ คือ Received, To, Bcc, Cc, From และ Subject

ซึ่งในตารางที่ผ่านมานั้น จะเห็นได้ว่าในแต่ละขอบเขตข้อมูลนั้น ส่วนมากแล้วจะมีส่วนของ Body รวมอยู่ด้วยในการพิจารณาถึงความเป็นไปได้ที่จะเป็นสแปมเมล ซึ่งหมายความว่าส่วนของ Body นั้นน่าจะมีค่าน้ำหนักมากที่สุดในการพิจารณาวิเคราะห์ถึงความเป็นไปได้ว่าเมลนั้นน่าจะเป็นสแปมหรือไม่ ในโครงงานนี้จึงได้พัฒนาระบบให้พิจารณาเพียงแค่ Body ในแต่ละเมลเท่านั้น

3.3 ระบบที่ใช้ในการพัฒนา และสภาพแวดล้อมในการพัฒนา

ในการพัฒนาระบบนี้ ได้ทำการสร้างสภาพแวดล้อมในการพัฒนาไว้ดังต่อไปนี้

- เครื่องคอมพิวเตอร์โน๊ตบุ๊ค ใช้ Intel Pentium M 1.5 กิ๊กกะเฮิร์ตส หน่วยความจำ 768 เมกกะไบต์ ติดตั้งระบบปฏิบัติการ Microsoft Windows XP Professional SP2

- ติดตั้งระบบปฏิบัติการ Red Hat Enterprise Linux ES 4 Update 4 บนโปรแกรมจำลองเครื่องคอมพิวเตอร์เสมือน VMWare Workstation 5.5.3 build-34685 โดยทำการกำหนดค่า Virtual PC ให้ใช้หน่วยความจำ 192 เมกกะไบต์ พื้นที่ฮาร์ดดิสก์เสมือน 4 กิ๊กกะไบต์ จำนวน 2 ตัว
- ใช้ภาษา Python เวอร์ชั่น 2.5 ในการศึกษาการทำงานของโครงงาน "การคัดแยกสแปมเมลโดยอัลกอริทึมเบย์เซียน" บนระบบปฏิบัติการวินโดวส์
- ใช้ภาษา PHP เวอร์ชั่น 4.3 ในการพัฒนาปลั๊กอินเพิ่มเติมความสามารถส่วนของการคัดแยกสแปมเมลให้กับ SquirrelMail บนระบบปฏิบัติการลีนุกซ์
- ใช้ Sendmail เวอร์ชั่น 8.13 เป็นเมลเซิร์ฟเวอร์ทำงานบนระบบปฏิบัติการลีนุกซ์
- ใช้ Apache เวอร์ชั่น 2.0 เป็นเว็บเซิร์ฟเวอร์ทำงานบนระบบปฏิบัติการลีนุกซ์
- ใช้ SquirrelMail เวอร์ชั่น 1.4 เป็นเว็บเบสอีเมลทำงานบนระบบปฏิบัติการลีนุกซ์
- ใช้ Dovecot เวอร์ชั่น 0.99 เป็นเซิร์ฟเวอร์ไอแมพ (IMAP) สำหรับติดต่อกับเมลเซิร์ฟเวอร์เพื่อดึงเมลมาให้ SquirrelMail ทำงานบนระบบปฏิบัติการลีนุกซ์
- ใช้ Firefox เวอร์ชั่น 1.5 เป็นเว็บบราวเซอร์ในการเข้าใช้งาน SquirrelMail

3.4 ความต้องการของฮาร์ดแวร์ในการติดตั้งระบบ

เนื่องจากว่าในการที่จะนำเอาระบบที่พัฒนาขึ้นมาเข้าไปใช้งานได้นั้น จำเป็นต้องต้องมีการติดตั้งระบบเป็นเมลเซิร์ฟเวอร์ที่สามารถใช้งานได้จริง ดังนั้นความต้องการของระบบที่จะกำหนดต่อไปนี้ จึงเป็นระบบขั้นต่ำ ที่จะทำให้สามารถใช้งานได้อย่างมีประสิทธิภาพและเพียงพอ โดยควรจะมีสเปกคร่าวๆ ดังนี้

- เครื่องที่มี CPU ความเร็วในการประมวลผลที่ 1.5 GHz. ขึ้นไป
- หน่วยความจำหลัก 512 MB. ขึ้นไป
- ฮาร์ดดิสก์มีขนาดเพียงพอ สามารถรองรับการใช้งานของผู้ใช้ได้

3.5 การติดตั้งระบบ

ในการติดตั้งระบบนั้น จำเป็นต้องมีการลงโปรแกรมต่างๆ แบบเดียวกับระบบที่ใช้ในการพัฒนา โดยข้อกำหนดของเวอร์ชันต่างๆ ของแต่ละโปรแกรมนั้น ให้ยึดขั้นต่ำตามที่ได้ระบุไว้ในหัวข้อที่ 3.3 โดยในการติดตั้งโปรแกรมของโครงงาน มีขั้นตอนดังนี้

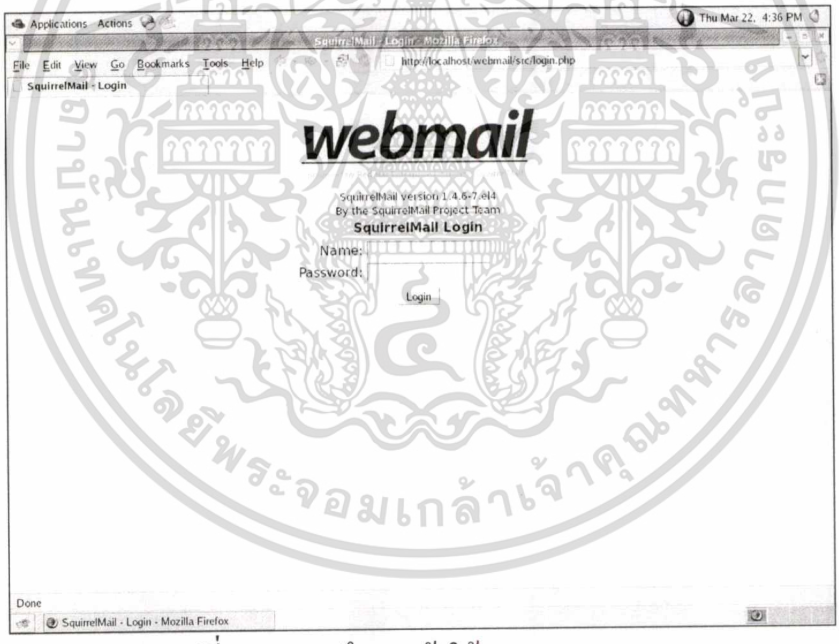
- ทำการคัดลอกไฟล์โปรแกรมซึ่งเป็นไดเร็กทอรี่ชื่อ panmail ไปยังไดเร็กทอรี่

/usr/share/squirrelmail/plugins (โฟลเดอร์นี้ยึดตามโครงสร้างของ Red Hat Enterprise Linux เป็นหลัก ดังนั้นหากมีการใช้งานลีนุกซ์ค่ายอื่นๆ ควรตรวจสอบอีกครั้ง)

- ทำการเพิ่มค่า plugins[]=panmail เข้าไปในไฟล์ /etc/squirrelmail/config.php
- ทำการสร้างดาต้าเบสขึ้นมาภายใต้ระบบดาต้าเบส MySQL แล้วทำการอิมพอร์ตไฟล์ SQL เข้าไปในดาต้าเบสนั้น เพื่อทำการสร้างข้อมูลในส่วนของ preprocessing
- ทำการแก้ไขค่าของ username และ password ในไฟล์ panmail_functions.php ในส่วนของฟังก์ชั่น filter เพื่อให้สามารถติดต่อกับดาต้าเบสได้

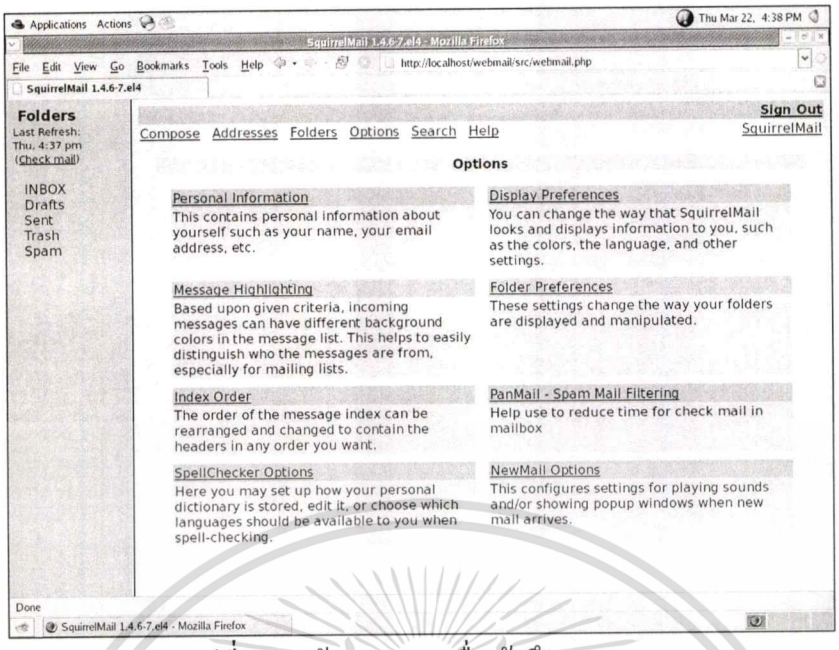
3.6 การเรียกใช้งาน

ในการเรียกใช้งานนั้น ให้ผู้ใช้ทำการเปิดโปรแกรมเว็บบราวเซอร์ จากนั้นทำการพิมพ์ที่อยู่ที่ชี้ไปยัง SquirrelMail จากนั้นทำการล็อกอินด้วย username และ password ที่ได้รับจากผู้ดูแลระบบ

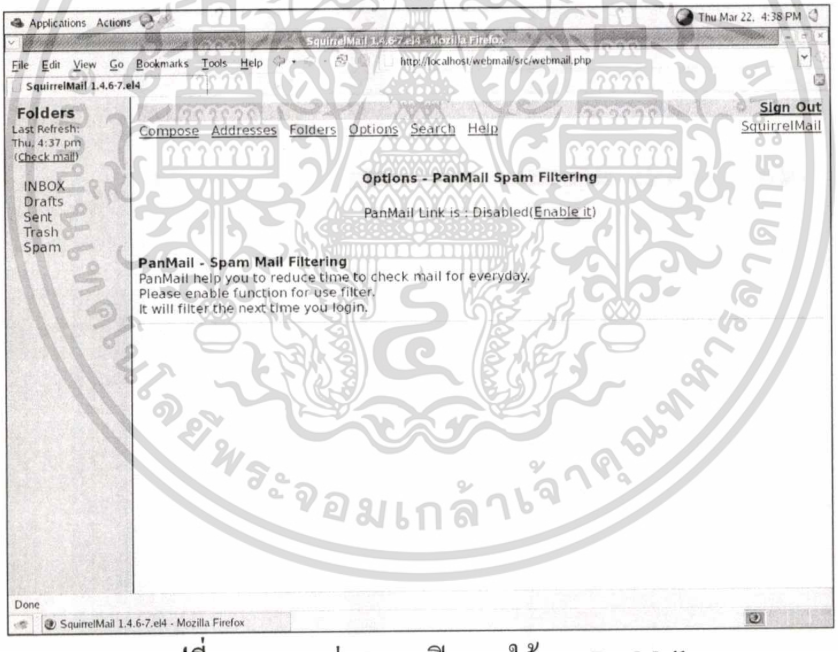


รูปที่ 3.4 แสดงถึงการเข้าใช้งาน SquirrelMail

เมื่อทำการล็อกอินแล้วให้ทำการคลิกที่เมนู Options ที่อยู่ส่วนบนของเว็บ จากนั้นให้คลิกที่ลิงก์ PanMail เพื่อเข้าไปทำการเปิดการใช้งานของระบบ



รูปที่ 3.5 หน้า Options เพื่อเข้าถึง PanMail



รูปที่ 3.6 แสดงส่วนการเปิดการใช้งาน PanMail

เมื่อเข้าไปยังในส่วนของการปรับแต่ง PanMail แล้วให้ทำการคลิกที่ลิงก์ Enable ซึ่งเป็นการเปิดการใช้งานโปรแกรมแล้ว ซึ่งหลังจากที่คลิกแล้ว ให้ทำการล็อกเอาท์ และเมื่อทำการล็อกอินเข้าไปใหม่ โปรแกรมก็จะเริ่มการทำงานโดยอัตโนมัติ

3.7 การพัฒนาระบบเพื่อเป็นปลั๊กอินให้กับ SquirrelMail

ในการพัฒนาระบบเพื่อเป็นปลั๊กอินให้กับ SquirrelMail นั้นมีรายละเอียดค่อนข้างมาก แต่จะขอสรุปเป็นวิธีการคร่าวๆ เพื่อเป็นแนวทางในการศึกษาต่อเพิ่ม

ในการพัฒนาปลั๊กอินให้กับ SquirrelMail นั้น เราจำเป็นจะต้องรู้ว่าไดเร็กทอรี่ที่ทำการเก็บตัวโปรแกรม SquirrelMail อยู่ที่ไหน ซึ่งในที่นี้ขอยกตัวอย่างว่าอยู่ภายใต้ /usr/share/squirrelmail โดยภายใต้ไดเร็กทอรี่นี้จะมีไดเร็กทอรี่ย่อยชื่อ plugins โดยให้เราทำการสร้างไดเร็กทอรี่ชื่อของปลั๊กอินที่เราได้ตั้งไว้ ภายใต้ไดเร็กทอรี่ plugins โดยขอสมมุติว่าชื่อ demo

ในการทำงานของ SquirrelMail นั้นถ้าเราต้องการให้ SquirrelMail เรียกปลั๊กอินที่เราพัฒนาขึ้นมาใช้งาน เราจำเป็นต้องบอกให้ SquirrelMail รับรู้ก่อนว่ามีปลั๊กอินนี้อยู่ในระบบของมัน โดยการเปิดไฟล์คอนฟิกของ SquirrelMail โดยในที่นี้สมมุติว่าให้อยู่ที่ /etc/squirrelmail/config.php โดยใน Section Plugins ให้เราทำการเพิ่มบรรทัดคล้ายดังนี้คือ plugins[ ] = demo จากนั้นทำการบันทึกไฟล์ ซึ่งวิธีการนี้เป็นเหมือนการที่เราทำการลงทะเบียนให้ SquirrelMail รับรู้ว่ามีปลั๊กอินเราให้ใช้งานแล้ว โดย SquirrelMail จะทำการเข้าไปที่ไดเร็กทอรี่ของเราและทำการอ่านไฟล์ setup.php ซึ่งเป็นไฟล์เริ่มต้นการทำงานของปลั๊กอิน ทุกปลั๊กอิน ที่ทำงานร่วมกับ SquirrelMail ซึ่งรายละเอียดการเขียนปลั๊กอินเพิ่ม สามารถหาอ่านได้จากเว็บไซต์ SquirrelMail

บทที่ 4
การทดลองและผลการดำเนินการ

4.1 วัตถุประสงค์การทดลอง

- เพื่อทดลองการคัดแยกอีเมลด้วยอัลกอริทึม Bayesian ว่าจะสามารถลดปัญหาการคัดแยกผิด โดยคัดแยกถูกต้อง 90% ขึ้นไปหรือไม่
- เพื่อหาข้อสรุปถึงในเรื่องความถูกต้องในการพิจารณากับความเร็วในการดำเนินการ

4.2 เงื่อนไขในการทดลอง

4.2.1 ข้อมูลที่ใช้ในการทำ Preprocessing

ข้อมูลตัวอย่างที่นำมาใช้ในการทดลองมาจาก SPAMAssassin public mail Corpus มีอีเมลทั้งสิ้นจำนวน 5,000 ฉบับ แบ่งเป็นอีเมลปกติ (HAM) จำนวน 2,000 ฉบับ และเป็นสแปม (SPAM) จำนวน 3,000 ฉบับ

4.2.2 ข้อมูลที่ใช้ในการทดลอง

- เป็นอีเมลส่วนหนึ่งที่ใช้ในการทำ Preprocessing (ข้อ 4.2.1)โดยสุ่มมาใช้ เป็นจำนวน 200 ฉบับ แบ่งเป็น HAM จำนวน 100 ฉบับ และเป็น SPAM จำนวน 100 ฉบับ
- เป็นอีเมลที่จาก untroubled.org เป็นคนละชุดกับข้อ 4.2.1 สุ่มทดสอบจำนวน 100 ฉบับ

4.3 วิธีการทดลอง

- ทดลองคัดแยกสแปมจากข้อมูลในข้อ 4.2.2 ที่ค่า SPAMCITY ในช่วง $0 < N <= 0.1$ และ $0.9 <= N < 1$
- ทดลองคัดแยกสแปมจากข้อมูลในข้อ 4.2.2 ที่ค่า SPAMCITY ในช่วง $0 < N <= 0.2$ และ $0.8 <= N < 1$
- ทดลองคัดแยกสแปมจากข้อมูลในข้อ 4.2.2 ที่ค่า SPAMCITY ในช่วง $0 < N <= 0.3$ และ $0.7 <= N < 1$

4.4 สภาพแวดล้อมในการทดลอง

โปรแกรมทำงานภายใต้สภาพแวดล้อมของระบบปฏิบัติการลีนุกซ์ โดยได้ทำการรันบนโปรแกรมจำลองการทำงานคอมพิวเตอร์เสมือน VMWare Worskstation โดยมีสเปกที่กำหนดในเครื่องคอมพิวเตอร์เสมือนดังนี้

- CPU : Intel Pentium M 1.5 GHz
- RAM 192 MB
- Harddisk จำลองเป็น SCSI Drive ขนาด 8 กิ๊กกะไบต์

4.5 ผลการทดลอง

หลังจากที่ได้ทดลองคัดแยกสแปมจากข้อมูลในข้อ 4.2.2 ได้ผลการทดลองสรุปได้ดังนี้

- ที่ SPAMCITY ในช่วง 0 < N <= 0.1 และ 0.9 <= N < 1 ได้ผลการทดลองว่า ในการคัดแยกสแปมของเมลประเภท SPAM สามารถทำการคัดแยกเป็นสแปมเมลได้ถูกต้องมากกว่า 80% ขึ้นไป โดยจะยังมีเมลเพียงบางส่วนที่ระบบ ไม่สามารถจำแนกให้ระบุลงไปได้ว่า เป็นสแปมเมล ส่วนการแยกสแปมของเมลประเภท HAM นั้น มีความผิดพลาดในการระบุว่าเมลนั้นเป็นสแปมอยู่โดยประมาณ 20% ซึ่งน่าจะเป็นไปได้ว่าคำที่อยู่ HAM นั้นมีเนื้อหาสาระที่ใกล้เคียงกับสแปมเมล จึงทำให้ยังเกิดความผิดพลาดในการคัดกรอง
- ที่ SPAMCITY ในช่วง 0 < N <= 0.2 และ 0.8 <= N < 1 ได้ผลการทดลองว่า ในการคัดแยกสแปมของเมลประเภท SPAM สามารถทำการแยกเป็นสแปมเมลได้ถูกต้องดีขึ้นกว่าช่วงค่าที่ผ่านมาเล็กน้อย และเมื่อทำการแยกสแปมของเมลประเภท HAM นั้นก็มีความผิดพลาดในการระบุว่าเมลนั้นเป็นสแปมเพิ่มขึ้นเล็กน้อยเช่นกัน
- ที่ SPAMCITY ในช่วง 0 < N <= 0.3 และ 0.7 <= N < 1 ได้ผลการทดลองว่า ในการคัดแยกสแปมเมลของเมลประเภท SPAM นั้นสามารถทำการแยกเป็นสแปมเมลได้ถูกต้องมากที่สุด ซึ่งได้มากกว่า 90% ขึ้นไป แต่ด้วยความที่ช่วงขอบเขตนี้ เป็นช่วงขอบเขตที่ค่อนข้างกว้าง จึงพบว่าหากเมลนั้นมีเนื้อหาสาระใกล้เคียงกับเมลที่สแปมมากๆ จะทำให้เมลนั้นถูกพิจารณาว่าเป็นสแปมไปโดยปริยาย ซึ่งผลการทดลองกับเมลประเภท HAM นั้น ระบบได้ทำการตรวจว่าเป็นสแปมมากถึงกว่า 30% เลยทีเดียว

จากการทดลองในช่วงค่าทั้ง 3 ค่าดังกล่าวจึงสามารถสรุปได้ว่า การเลือกช่วงขอบเขตที่มีความแคบของช่วงค่ามากเกินไปนั้น จะทำให้การจำแนกเมล์นั้นไม่ได้ประสิทธิภาพเท่าที่ควร เนื่องจากว่ายังคงมีเมลบางประเภท ที่มีเนื้อหาใกล้เคียงกับความเป็นสแปมมาก จึงทำให้ระบบไม่สามารถจำแนกชนิดของเมลได้ถูกต้องมากนัก ส่วนการที่ใช้ช่วงขอบเขตที่มีค่ามากเกินไป ก็จะทำ

ให้เมลสาระบางประเภทนั้น ถูกพิจารณาเป็นสแปมได้โดยง่ายเกินไป ซึ่งจากการทดลองจึงได้ทำการประยุกต์ให้โปรแกรมที่พัฒนาขึ้นมานั้น ใช้ค่า SPAMCITY ในช่วงขอบเขตที่ $0 < N <= 0.2$ และ $0.8 <= N < 1$ ซึ่งเป็นช่วงที่ให้ประสิทธิภาพเป็นที่น่าพอใจ

4.6 สรุปผลการทดลอง

จากการทดลองสรุปได้ว่าจำนวนคำที่นำมาใช้ในการพิจารณามีผลต่อการคัดแยก การเลือกคำหรือกลุ่มคำที่มีค่าอำนาจจำแนกสูงจะสามารถลดเวลาในการคัดแยก และยังคงประสิทธิภาพไว้ได้หรือลดลงเพียงเล็กน้อยเท่านั้นในการคัดแยกด้วยอัลกอริทึม Bayesian จากการทดลองสามารถคัดแยก SPAM ที่มาจาก Corpus เดียวกันกับที่นำไปเตรียมข้อมูล (Preprocessing) เฉลี่ยถูกต้องกว่า 80% และมีความถูกต้องลดลงเมื่อทดลองคัดแยก SPAM ที่มาจาก Corpus อื่น คาดว่าเราสามารถปรับในเรื่องของประสิทธิภาพการคัดแยกให้เพิ่มขึ้นได้อีก โดยพิจารณาถึงวิธีเลือกคำที่ดีในการคัดแยก

บทที่ 5
บทสรุปและข้อเสนอแนะ

5.1 บทสรุป

การคัดแยกประเภทของอีเมลโดยใช้อัลกอริทึม Bayesian นั้นให้ผลในการลดปัญหา False Positive เป็นที่น่าพอใจคัดแยก SPAM ถูกต้องประมาณ 80% บน Corpus เดียวกัน แต่ถ้าอีเมลที่มาจาก Corpus อื่นคัดแยกถูกต้องลดลงเหลือประมาณ 70% โดยประมาณ แต่ต้องใช้เวลาในการพิจารณามาก เนื่องจากการพิจารณาที่ได้ผลถูกต้องมากที่สุดต้องพิจารณาจากข้อความแบบ Text ทั้งหมดของอีเมลนั้น หากสามารถลดเวลาในการพิจารณาแยกประเภท และสามารถทำงานได้ดีคงที่หรือดียิ่งขึ้น โดยพัฒนากระบวนเลือกคำหรือขอบเขตที่ใช้ในพิจารณาที่เหมาะสมของอีเมล ให้มีความรวดเร็วและถูกต้องมากขึ้น จะได้รับความนิยมเพิ่มขึ้นอย่างแน่นอน

5.2 ข้อเสนอแนะ

5.2.1 แนวทางการประยุกต์ใช้งาน

โปรแกรมที่จัดทำได้จัดทำในลักษณะเป็น local mail server ที่สามารถคัดแยก SPAM เมลภายในโฟลเดอร์ของผู้รับ (mail box) โดยใช้อัลกอริทึม Bayesian โดยจัดทำเป็นปลั๊กอินเข้ากับเว็บเบสอีเมลไคลเอ็นต์สำหรับ SquirrelMail ทำให้สามารถเข้าใช้งานได้มากกว่าหนึ่งคน

5.2.2 ข้อจำกัดของโครงงานพิเศษนี้

5.2.2.1 ขนาดของอีเมลมีผลต่อความเร็วในการทำงาน

อีเมลที่มีขนาดใหญ่จะใช้เวลาในการพิจารณาแยกคำ การตัดคำที่ซ้ำซ้อนเพื่อใช้ทำเป็นคีย์เวิร์ด และถ้ากรณีที่ใช้เวลานานที่สุดคืออีเมลที่จำนวนคีย์เวิร์ดที่ไม่ซ้ำกันเลย หรือมีจำนวนเข้าใกล้กับขนาดของอีเมล ซึ่งจะทำให้ใช้เวลานานในการนำ คีย์เวิร์ดไปเทียบในฐานข้อมูลเพื่อหาค่า SPAMCITY ทีละตัวจนครบ

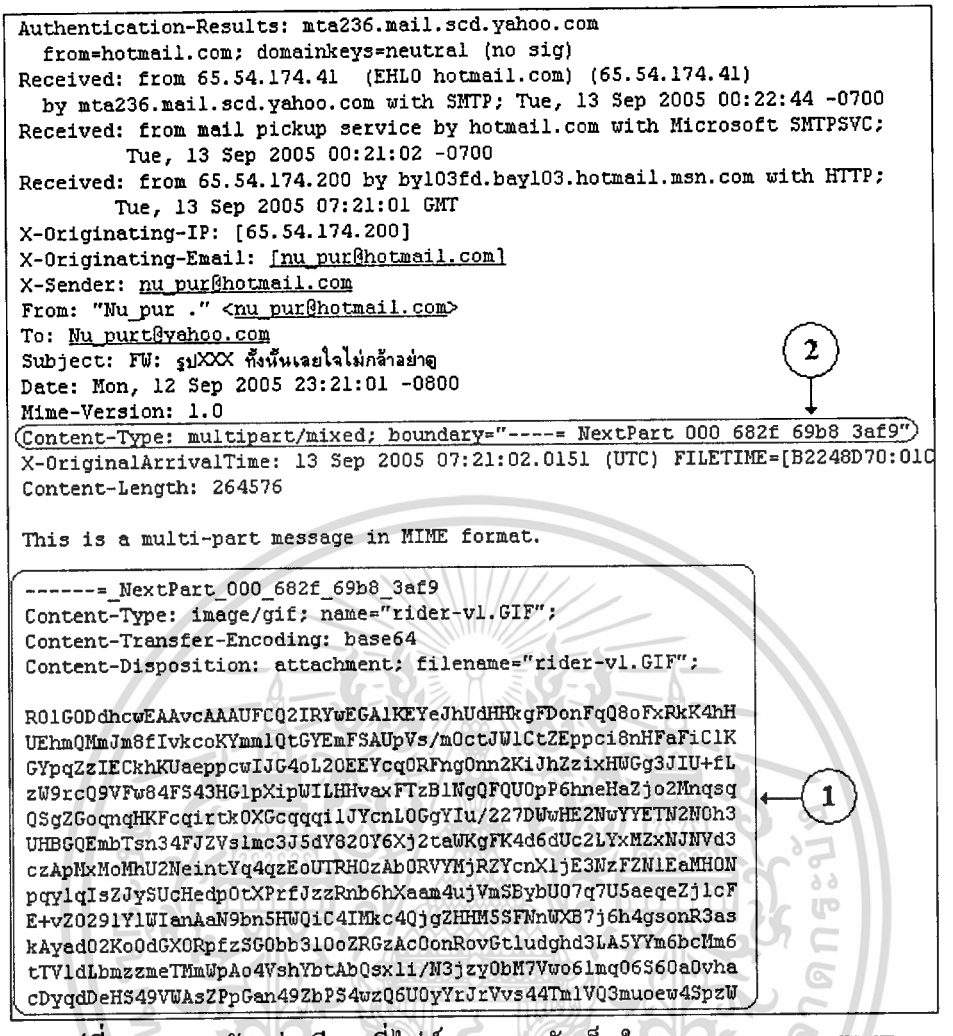
5.2.2.2 การพิจารณาอีเมลที่การ Unicode ด้วยรหัสที่ต่างกัน

อีเมลจะถูกแปลงให้อยู่ในรูปของ text ก่อนรับและส่ง ใน PC แต่ละการแปลงอีเมลให้อยู่ในรูปแบบไฟล์ text จะอ้างอิงกับ Unicode ที่เครื่องนั้นระบุไว้ ในกรณีที่โปรแกรมถูกติดตั้งไว้ที่เซิร์ฟเวอร์ที่มี unicode ไม่ตรงกับอีเมลนั้น จะพบปัญหาไม่สามารถแปลงตัวอักษรได้ถูกต้อง

ตรงกัน ทำให้มีผลต่อการแยกคีย์เวิร์ดและวิเคราะห์ความน่าจะเป็นของเอกสาร ทำให้ส่งผลให้เกิดความคลาดเลื่อนในการวิเคราะห์ได้

5.2.2.3 ปัญหาที่มาจาก Forward อีเมล

ในอีเมลที่มีไฟล์แนบมาด้วยจะ ถูกแปลงไฟล์ที่แนบมานั้นให้อยู่ในรูปแบบของ Text ก่อนตามมาตรฐาน Multipurpose Internet Mail Extensions (MIME) ดังรูปที่ 5.1 แสดงตัวอย่างอีเมลที่มีไฟล์แนบมาด้วยดูได้จากในส่วนที่ 1 ในรูปจะเป็นไฟล์รูปภาพชื่อ rider-V1.GIF ที่แปลงอยู่ในรูปแบบของ Text เรียบร้อยแล้ว ไฟล์ที่แนบมานี้จะถูกหั่นออกเป็นส่วนๆ โดยมี boundary ในส่วนที่ 2 ของรูปเป็นตัวแบ่งในบางอีเมลไคลเอ็นต์ การ forward อีเมล จะไม่ตรวจสอบก่อนว่าอีเมลนั้นมีส่วนใดเป็น Text จริงๆ และส่วนใดเป็น Text ที่อยู่ในรูปแบบของ MIME เมื่อไม่สนใจจึงใส่เนื้อความทั้งหมดในอีเมลในลักษณะของ Text ธรรมดาแล้วส่งต่อ ผู้อ่านอาจจะพบเห็น forward อีเมลในลักษณะนี้อยู่บ้าง กล่าวคือจะพบว่ามีอีเมลที่ส่งต่อมาถึงผู้อ่าน แต่ไม่ได้ถูกแสดงว่าเป็นอีเมลที่มีไฟล์แนบเป็นรูปภาพเห็นเป็นอีเมลที่เต็มไปด้วยข้อความ Text ที่มีส่วนที่ 1 ในรูปที่ 5.1 ปะปนมาด้วย โปรแกรมจะพิจารณาตัดคำเพื่อหาคีย์เวิร์ดนั้นพิจารณาส่วนที่ Forward มาด้วย ทำให้มีจำนวนคีย์เวิร์ดมากขึ้น แต่ไม่มีค่าอำนาจจำแนกพอที่จะพิจารณาได้ว่าอีเมลที่มีคีย์เวิร์ดนี้ปรากฏจะเป็น SPAM หรือไม่



รูปที่ 5.1 แสดงตัวอย่างอีเมลที่ไฟล์แนบและจัดเก็บในตามมาตรฐาน MIME

5.2.2.4 การคัดแยกอีเมลภาษาไทยไม่น่าเชื่อถือ

เนื่องจากรูปแบบการเขียนของภาษาไทยคำแต่ละคำติดกัน เมื่อหมดประโยคจึงเว้นวรรคทำให้ไม่สามารถจะพิจารณาตัดคำได้เหมือนกับในภาษาอังกฤษ เมื่อมีอีเมลภาษาไทยเข้ามาก็จะพิจารณาตัดคำตามเงื่อนไขที่กำหนดไว้ ซึ่งก็จะทำให้ได้คีย์เวิร์ดเช่นกันแต่จะได้คำที่ยาวและมีความถี่ที่ซ้ำกันน้อยมากจนถึงไม่ซ้ำเลย มีผลทำให้การคัดแยกไม่น่าเชื่อถือตามความหมาย SPAM ของผู้ใช้นั้นได้

5.2.3 มุมมองในการพัฒนาต่อ

ในระหว่างการพัฒนาผู้พัฒนาพบปัญหาหลายรูปแบบที่น่าสนใจ เห็นว่ามีความน่าสนใจในการพัฒนา เช่น

- จะทำอย่างไรให้การคัดแยกอีเมลที่ใช้ Unicode ต่างกันให้มีความผิดพลาดน้อยที่สุดในการคัดแยก

- ในส่วนหัวข้ออีเมลจะประกอบไปด้วย Keys ต่างๆ ที่ใช้ในการอ้างถึง values นั้นๆ ตามมาตรฐาน RFC 822 เพื่อการใช้งานของแต่ละเมลเซิร์ฟเวอร์ เช่น Key – "To:" ใช้อ้างถึงvalue ที่ส่งถึงผู้รับ เป็นต้น พบว่าในส่วนหัวของอีเมลนั้นประกอบไปด้วย Keys คำนวณมาก ขึ้นอยู่กับการใช้งานของแต่ละเมลเซิร์ฟเวอร์ ในการพิจารณาคัดแยก SPAM เมลในโปรแกรมนี้นั้นพิจารณาตัว Keys ด้วย ทำให้มีปัญหาว่า Keys เหล่านี้มีผลต่อการคัดแยกมากน้อยเพียงใด มีความความจำเป็นต้องพิจารณาหรือไม่
- ในการคัดแยกในโปรแกรมนี้ได้เลือกคัดแยกเฉพาะส่วนที่เป็น Text นั่นรวมถึงส่วนที่เป็น HTML ด้วย หากการคัดแยกสามารถแยก tag ของ HTML ออกจากการคัดแยกน่าจะส่งผลต่อความแม่นยำในการคัดแยกอีเมลที่มีข้อความที่เป็น HTML ประกอบอยู่ด้วย
- ในบางกรณีเชื่อว่าการคัดแยกอาจไม่สามารถแยก SPAM ออกมาโดยพิจารณาเพียงในส่วนของ Text เท่านั้น หากมีวิธีการพิจารณาในส่วนของไฟล์ที่แนบมาด้วยนั้นน่าจะทำให้คัดแยกได้ดียิ่งขึ้น

บรรณานุกรม

ธนรัฐ โชติพันธ์. 2548. "การคัดแยกสแปมเมลโดยอัลกอริธึมเบย์เซียน". โครงการพัฒนาระบบวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

อรรยา สิงห์สงบ. 2003. **ความพยายามทางกฎหมายกับการแก้ไขปัญหาจดหมายอิเลคทรอนิสก์ขยะ.** [Online]. Available: http://legalaid.bu.ac.th/files/articles/spammail.pdf.

Kurose F James, Keith W. Ross. 2004. **Computer Networking: A Top Down Approach Featuring the Internet.** 2nd edition. Boston: Addison-Wesley.

SquirrelMail. 2007. **SquirrelMail Developer's Manual.** [Online]. Available: http://www.squirrelmail.org/docs/devel/devel.html

Wikipedia. 2005. **Email spam.** [Online]. Available: http://en.wikipedia.org/wiki/Email_spam

ประวัติผู้เขียนโครงงาน

ชื่อผู้จัดทำโครงงาน นายนันทชัย สมัญญาภรณ์

วันเดือนปีเกิด 28 กันยายน พ.ศ. 2523

อิเล็กทรอนิกส์เมล nantachai@gmail.com

ประวัติการศึกษา

ประถมศึกษา โรงเรียนชาญเวทย์ศึกษา

มัธยมศึกษา โรงเรียนวัดสุทธิวราราม

อุดมศึกษา ศึกษาระดับปริญญาตรี วิศวกรรมศาสตรบัณฑิต
สาขาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ในปีการศึกษา 2545