การตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์

MAILING LIST E-MAIL DETECTION

ณัฐติญา ไข่ติยากุล

NATTIYA KHAITIYAKUN

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาเทคโนโลยีสารสนเทศ

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2548

ISBN 974-15-1763-7

MAILING LIST E-MAIL DETECTION

NATTIYA KHAITIYAKUN

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2005

ISBN 974-15-1763-7

หัวข้อวิทยานิพนธ์	การตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์
นักศึกษา	นางสาวณัฐติญา ไข่ติยากุล
รหัสประจำตัว	43067043
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
พ.ศ.	2548
อาจารย์ผู้ควบคุมวิทยานิพนธ์	ผศ. อัครินทร์ คุณกิตติ

บทคัดย่อ

ปัญหาที่เกิดจากการตรวจสอบจดหมายอิเล็กทรอนิกส์จำนวนมากในแต่ละวัน เพื่อตรวจจับเมลลิงลิสต์ โดย Network administrator ซึ่งอาจมีความผิดพลาดเกิดขึ้นเนื่องจากปริมาณของจดหมายและเมลลิงลิสต์ที่เกิดขึ้นใหม่ งานวิจัยนี้ต้องการนำเสนอระบบการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ เพื่อวิเคราะห์หาข้อมูล Mail address ที่ใช้ในการ config ระบบเมลลิงลิสต์ย่อย เริ่มจากการดักจับข้อมูล (Capture process) แล้วเข้าสู่กระบวนการจัดเตรียมข้อมูล (Preprocessing process) เพื่อให้ข้อมูลอยู่ในรูปแบบเดียวกันคือ Detection format ขั้นตอนต่อไปคือการตรวจจับเมลลิงลิสต์ (Detection process) โดยวิเคราะห์หาคุณสมบัติของเมลลิงลิสต์ ซึ่งจดหมายจากเมลลิงลิสต์สามารถแบ่งออกเป็น 2 แบบคือ จดหมายที่เกิดจากการติดต่อกันระหว่างผู้จัดการและสมาชิก และจดหมายที่เกิดจากการสื่อสารระหว่างสมาชิก ในแบบแรกวิเคราะห์ได้จาก Manager Command ส่วนแบบที่สองวิเคราะห์จากความแตกต่างของ Mail Address ของผู้รับและผู้ส่งที่พบใน Rfc822 Header และ Smtp Header จากนั้นกำหนด Initial Confidential Factor (CF_{IN}) และเก็บข้อมูลลงใน CMDB ขั้นตอนสุดท้าย (Postprocessing process) คือการปรับเปลี่ยน Confidential Factor (CF) เมื่อมีข้อมูลอื่นมาสนับสนุนหรือขัดแย้ง ก่อนนำข้อมูลไปจัดเก็บที่ DODB โดยเลือกเฉพาะข้อมูลที่มี CF สูงกว่า Threshold Confidential Factor (CF_{TH}) ผลการทดลองพบว่าเมื่อ CF_{IN} มีค่าเพิ่มสูงขึ้น ค่าความถูกต้อง (%Correctness) มีแนวโน้มลดลงเช่นเดียวกับค่าความผิดพลาดที่ระบบตรวจจับเมลลิงลิสต์ไม่พบ (%Positive Error) แต่ค่าความผิดพลาดที่ระบบตรวจจับเมลลิงลิสต์ไม่ถูกต้อง (%Negative Error) มีแนวโน้มสูงขึ้น ส่วน CF_{TH} เมื่อกำหนดให้ค่าเพิ่มสูงขึ้น %Correctness มีแนวโน้มเพิ่มขึ้นเช่นเดียวกับ %Positive Error แต่ %Negative Error มีแนวโน้มลดลง สำหรับจำนวนจดหมายอิเล็กทรอนิกส์นั้นถ้ามีจำนวนมากขึ้นจะทำให้ค่าความถูกต้องสูงขึ้นและมีค่าความผิดพลาดทั้ง 2 แบบต่ำลง จากการทดลองพบว่าที่ $CF_{IN} = 0.3$ และ $CF_{TH} = 0.8$ เป็นค่าที่เหมาะสมที่สุด

Thesis Title	Mailing List E-mail Detection
Student	Ms Nattiya Khaitiyakun
Student ID.	43067043
Degree	Master of Science
Programme	Information Science
Year	2005
Thesis Advisor	Asst.Prof. Akharin Khunkitti

ABSTRACT

There were many problems obtained from analysis of mailing list characteristic for e-mail processing that handled by network administrators. The troubles cause of new mailing list and tons of mails that pass through network in each day. This article will present ideas and methodology of Mailing List E-mail Detection system that analyzed mailing list e-mail, which can be used in Hierarchical mailing List system [3]. Detection step begins with capture process, which traps e-mail information from transfer channel. Next process is preparing raw data into detection format. The third one is mailing list detection part that compares and analyses e-mail to find out mailing list characteristic. Detected Mailing List may be one of two types; the first one is from mails generated from communications between mailing list manager and subscriber. Another from mails generated from communications between subscribers. The first type can be detected by Manager command. Another can be detected from the different of sender addresses and recipient addresses, which found in RFC822 header and SMTP header. The detection system uses Initial Confidential Factor (CF_{IN}) for detection data and stores results in CMDB. The last process is making decision and determines which data has supported or conflicted with detected data and increased or decreased Confidential Factors (CF). The Confidential Factors for each result finally will be compared to Threshold Confidential Factor (CF_{TH}). If the CFs are greater than CF_{TH}, system will put the final results in DODB. From the experimental results we found that when CF_{IN} has been increased %Correctness and %Positive Error will be decreased but %Negative Error will grow up. However, when CF_{TH} has been increased %Correctness and %Positive Error will grow up but %Negative Error will be decreased. And when the number of mails has been increased, %Correctness will grow up but %Positive Error and %Negative Error will be decreased. From the experiment, $CF_{IN} = 0.3$ and $CF_{TH} = 0.8$ are the optimum values.

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จได้ด้วยความกรุณาจากอาจารย์ ผศ.อัครินทร์ คุณกิตติ ที่ให้ความช่วยเหลือ ให้คำชี้แนะ ช่วยแก้ปัญหาตลอดจนให้ความรู้และประสบการณ์ที่ดีแก่ข้าพเจ้า

ขอขอบคุณเพื่อน ๆ และพี่ ๆ ในห้องปฏิบัติการทุกคนที่ช่วยให้คำปรึกษา ความช่วยเหลือและให้กำลังใจซึ่งกันและกันตลอดมา

สำหรับคุณงามความดีอันใดที่เกิดจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบให้กับบิดามารดา ซึ่งเป็นที่รักและเคารพยิ่ง ตลอดจนครูอาจารย์ที่เคารพทุกท่านที่ได้ประสิทธิ์ประสาทวิชาความรู้และถ่ายทอดประสบการณ์ที่ดีให้แก่ข้าพเจ้า

ณัฐติญา ไข่ติยากุล

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญตาราง (ต่อ)

สารบัญรูป

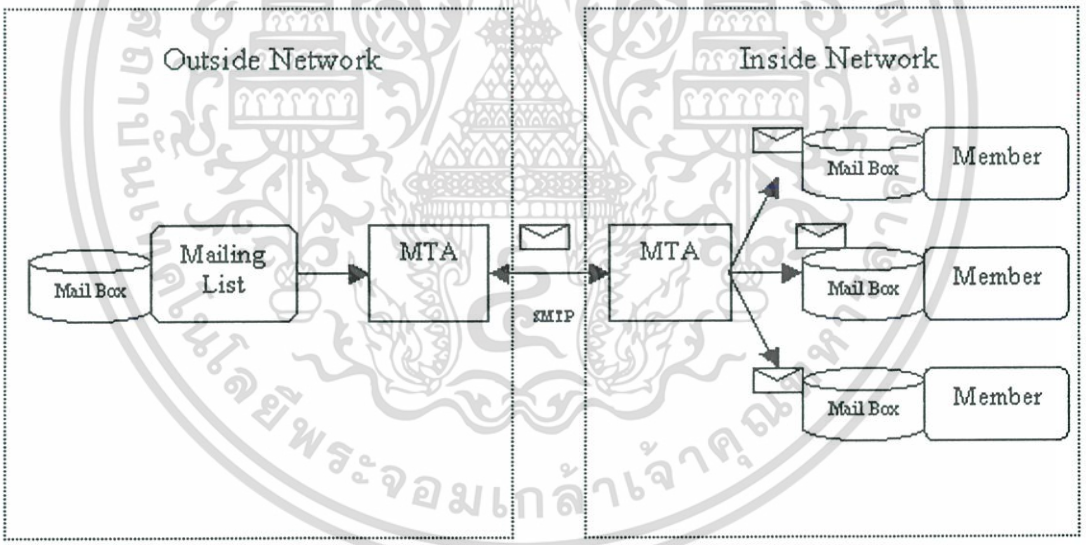
สารบัญรูป (ต่อ)

สารบัญรูป (ต่อ)

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เมลลิงลิสต์ (Mailing List) [1] คือการรวบรวมกลุ่มคนที่มีความสนใจในเรื่องเดียวกันมีความต้องการแลกเปลี่ยนข้อมูลและร่วมแสดงความคิดเห็น ผ่านทางจดหมายอิเล็กทรอนิกส์ [2] โดยมีเมลลิงลิสต์เป็นศูนย์กลาง เมื่อสมาชิกส่งจดหมายอิเล็กทรอนิกส์เข้ามาที่เมลลิงลิสต์ ผู้จัดการเมลลิงลิสต์จะจัดการทำสำเนาของจดหมายอิเล็กทรอนิกส์ฉบับนั้น แล้วส่งต่อให้สมาชิกที่มีอยู่ในบัญชีรายชื่อ จากการทำงานของเมลลิงลิสต์พบว่า มักก่อให้เกิดปัญหาความคับคั่งของการจราจรระหว่างเครือข่าย เนื่องจากเมลลิงลิสต์กระจายจดหมายอิเล็กทรอนิกส์ไปยังสมาชิกทุกคน โดยไม่คำนึงถึงความใกล้เคียงกันของเครือข่าย จึงเกิดความซ้ำซ้อนของข้อมูลจำนวนมากอยู่บนเส้นทางการจราจรระหว่างเครือข่าย นอกจากนั้นยังเกิดความสิ้นเปลืองเนื้อที่ในการเก็บข้อมูลของจดหมายอิเล็กทรอนิกส์ให้กับสมาชิกแต่ละคนอีกด้วย ดังรูปที่ 1.1



รูปที่ 1.1 ภาพแสดงการกระจายจดหมายอิเล็กทรอนิกส์ของเมลลิงลิสต์

แนวทางการแก้ไขคือพัฒนาระบบเมลลิงลิสต์ย่อย (Submailing List) [3] เพื่อลดความซ้ำซ้อนด้วยการรวบรวมกลุ่มสมาชิกของเมลลิงลิสต์ตามความใกล้เคียงกันของเครือข่าย แล้วสร้างระบบเมลลิงลิสต์ย่อยเพื่อเป็นตัวแทนคอยรับจดหมายอิเล็กทรอนิกส์จากเมลลิงลิสต์ ซึ่งเป็นการลดปริมาณความหนาแน่นของข้อมูลบนเส้นทางการจราจรระหว่างเครือข่าย นอกจากนั้นระบบเมลลิงลิสต์ย่อยจะจัดการกับระบบการเข้าถึงข้อมูลของสมาชิก ด้วยการเก็บจดหมายอิเล็กทรอนิกส์ต้น

ฉบับไว้เพียงฉบับเดียว จากนั้นจะส่งจดหมายอิเล็กทรอนิกส์เพื่อแจ้งให้สมาชิกทราบว่ามีข้อมูลใหม่เข้ามาและระยะเวลาในการเก็บข้อมูล ซึ่งสมาชิกสามารถเข้าไปอ่านข้อมูลนั้นๆได้ภายในระยะเวลาที่กำหนด เป็นการลดความซ้ำซ้อนของข้อมูลที่เก็บใน Mail Box ของสมาชิกแต่ละคนภายในเครือข่าย

แต่ปัญหาที่ตามมาก็คือจะทราบได้อย่างไรว่า E-mail Address ใดเป็นสมาชิกของเมลลิงลิสต์เพื่อรวบรวมกลุ่มของสมาชิกและกำหนดค่าต่างๆให้กับระบบเมลลิงลิสต์ย่อย สามารถทำงานได้ถูกต้อง ซึ่งในแบบเดิมนั้นใช้ Network Administrator ช่วยในการตรวจสอบและกำหนดค่าการทำงานต่างๆให้ระบบเมลลิงลิสต์ย่อย ทำให้เกิดความไม่สะดวกเนื่องจากปริมาณของจดหมายอิเล็กทรอนิกส์ที่ผ่านเข้าออกจากเครือข่ายมีปริมาณมาก ทำให้ Network Administrator ไม่สามารถตรวจสอบได้ทั้งหมดและอาจมีเมลลิงลิสต์เกิดขึ้นใหม่เรื่อยๆ อาจเกิดความผิดพลาดในการตรวจสอบ ทำให้ระบบเมลลิงลิสต์ย่อยไม่สามารถแก้ไขปัญหาได้ ส่งผลให้ประสิทธิภาพของการแลกเปลี่ยนข้อมูลระหว่างเครือข่ายลดลงและมีข้อมูลที่ซ้ำซ้อนถูกเก็บอยู่ใน Mail Box ของสมาชิกเป็นการสิ้นเปลืองทรัพยากรของเครือข่าย งานวิจัยนี้จึงนำเสนอแนวความคิดการสร้างระบบการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ หรือ Mailing List E-mail Detection System เพื่อตรวจสอบและรายงานผลให้ Network Administrator นำข้อมูลไปจัดการกำหนดค่าต่างๆให้กับระบบเมลลิงลิสต์ย่อยเป็นการแบ่งเบาภาระให้ Network Administrator และเพิ่มประสิทธิภาพการแก้ไขปัญหาการจราจรระหว่างเครือข่ายและการจัดการทรัพยากรของเครือข่ายอีกด้วย

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

จากปัญหาที่กล่าวมาข้างต้น ความมุ่งหมายและวัตถุประสงค์ของการศึกษาคือ

1. วิเคราะห์หาความแตกต่างระหว่างจดหมายอิเล็กทรอนิกส์จากกลุ่มเมลลิงลิสต์และจดหมายอิเล็กทรอนิกส์ธรรมดาทั่วๆไป
2. วิเคราะห์จดหมายอิเล็กทรอนิกส์เพื่อหาข้อมูลของ E-mail Address ที่เกี่ยวข้องกับเมลลิงลิสต์เพื่อนำไปกำหนดค่าในการทำงานให้กับระบบเมลลิงลิสต์ย่อย เช่น List Address, Member Address, Manager Address และ Manager Type
3. คำนวณค่าความมั่นใจ หรือ Confidential Factor เพื่อใช้ประกอบการตัดสินใจการคัดเลือกข้อมูลที่มีคุณสมบัติของเมลลิงลิสต์
4. รายงานค่าการทำงานต่างๆที่ได้จากการวิเคราะห์ให้ Network Administrator

1.3 สมมุติฐานของการศึกษา

จากความมุ่งหมายและวัตถุประสงค์ของการศึกษาที่กล่าวมาข้างต้น สมมุติฐานของการศึกษาเพื่อให้บรรลุถึงความมุ่งหมายและวัตถุประสงค์ในการแก้ปัญหาคือ

1. คัดลอกข้อมูลที่ผ่านเข้าและออกจากเครือข่าย นำข้อมูลที่ได้เข้าสู่กระบวนการจัดเตรียมข้อมูลจดหมายอิเล็กทรอนิกส์ให้อยู่ในรูปแบบเดียวกัน จากนั้นเข้าสู่กระบวนการตรวจจับเพื่อวิเคราะห์คุณสมบัติและลักษณะเฉพาะของจดหมายอิเล็กทรอนิกส์
2. เมื่อวิเคราะห์ข้อมูลจดหมายอิเล็กทรอนิกส์ พบคุณสมบัติและลักษณะเฉพาะของกลุ่มเมลลิงลิสต์แล้ว นำข้อมูลมาพิจารณาเพื่อคัดเลือก E-mail Address ที่เกี่ยวข้องกับเมลลิงลิสต์ตามเงื่อนไขที่กำหนดไว้
3. คำนวณค่าความมั่นใจ หรือ Confidential Factor จากเงื่อนไขที่ได้จากการวิเคราะห์คุณสมบัติและลักษณะเฉพาะของจดหมายอิเล็กทรอนิกส์จากกลุ่มเมลลิงลิสต์
4. รายงานข้อมูลที่วิเคราะห์ได้ รวมถึงค่าของ E-mail Address ต่างๆ ประกอบด้วย List Address, Member Address, Manager Address และ Manager Type โดยใช้ค่าความมั่นใจ หรือ Confidential Factor ประกอบการตัดสินใจ

1.4 ขอบเขตของการศึกษา

จากสมมติฐานข้างต้น สามารถกำหนดขอบเขตของการศึกษาได้ดังนี้

1. ศึกษาเกี่ยวกับการทำงานของจดหมายอิเล็กทรอนิกส์รูปแบบต่างๆ
2. ศึกษาเกี่ยวกับการทำงานของเมลลิงลิสต์รวมถึงการทำงานของ Software ที่ช่วยในการจัดการเมลลิงลิสต์ หรือ Mailing List Manager (MLM)
3. ศึกษาเกี่ยวกับการทำงานของระบบเมลลิงลิสต์ย่อย เพื่อทราบถึงข้อมูลที่ต้องการนำไปจัดการในระบบ

1.5 ขั้นตอนของการศึกษา

ขั้นตอนของการศึกษาโดยสรุปมีรายละเอียดดังนี้

1. ศึกษาการทำงานในรูปแบบต่างๆของจดหมายอิเล็กทรอนิกส์ทั่วๆไปรวมถึง จดหมายอิเล็กทรอนิกส์ที่เกิดจากการทำงานของเมลลิงลิสต์
2. วิเคราะห์เปรียบเทียบหาความแตกต่างระหว่างจดหมายอิเล็กทรอนิกส์ธรรมดาทั่วๆไปและจดหมายอิเล็กทรอนิกส์ที่มาจากกลุ่มเมลลิงลิสต์
3. ออกแบบระบบที่จะใช้ตรวจจับการทำงานของเมลลิงลิสต์ตามสมมติฐานของการศึกษา เลือกสภาพแวดล้อมที่ระบบนั้นจะทำงานอยู่ภายใต้และเลือกเครื่องมือที่จะใช้ในการพัฒนาระบบ

4. ทำการพัฒนาระบบตามที่ได้ออกแบบมาภายใต้สภาพแวดล้อมและเครื่องมือที่กำหนด
5. นำระบบที่ได้มาทดลองเพื่อนำผลมาพิสูจน์สมมติฐานของการศึกษาที่ตั้งขึ้นมา
6. วิเคราะห์ วิจารณ์ และสรุปผลการทดลองที่ได้

บทที่ 2

จดหมายอิเล็กทรอนิกส์จากเมลลิงลิสต์และระบบเมลลิงลิสต์ย่อย

2.1 จดหมายอิเล็กทรอนิกส์

จดหมายอิเล็กทรอนิกส์ หรือ E-mail คือเครื่องมือที่ใช้ติดต่อสื่อสารระหว่างกลุ่มคนผ่านทางเครือข่ายอินเตอร์เน็ต เป็นที่นิยมกันแพร่หลายเนื่องจากมีความสะดวกรวดเร็ว ขั้นตอนการส่งและรับจดหมายอิเล็กทรอนิกส์ผ่านเครือข่ายอิเตอร์เน็ตนั้นสามารถแบ่งได้เป็น 2 ส่วนคือ Mail User Agents (MUA) และ Mail Transfer Agents (MTA) โดย MUA คือโปรแกรมที่ user ใช้สำหรับอ่าน เขียน ตอบและส่งจดหมายอิเล็กทรอนิกส์ ส่วน MTA คือโปรแกรมที่ทำหน้าควบคุมการส่งจดหมายอิเล็กทรอนิกส์ระหว่างเครือข่ายเปรียบเสมือนที่ทำการไปรษณีย์นั่นเอง ในการส่งและรับจดหมายอิเล็กทรอนิกส์ระหว่าง MTA นั้นใช้โปรโตคอลช่วยในการแลกเปลี่ยนข้อมูลคือ Simple Mail Transfer Protocol (SMTP) [4] โดยรูปแบบ หรือ Format ของจดหมายอิเล็กทรอนิกส์นั้นใช้มาตรฐานของ RFC822 Standard for the Format of ARPA Internet Text Messages ซึ่งประกอบด้วยส่วนสำคัญ 2 ส่วนคือ Header และ Body โดยใน Header ประกอบด้วยข้อมูลและรายละเอียดต่างๆของผู้ส่งและผู้รับ รวมถึงการระบุประเภทของ Software ที่ใช้เพื่อเปิดจดหมายอิเล็กทรอนิกส์ เป็นต้น ดังในรูปที่ 2.1

```
220 killbil.dipac.it.kmitl.ac.th ESMTP Sendmail 8.12.5/8.12.5; Mon, 16 May 2005 12:31:20 +0700
EHLO recognition.dipac.it.kmitl.ac.th
250-killbil.dipac.it.kmitl.ac.th Hello recognition.dipac.it.kmitl.ac.th [192.168.1.13], pleased to meet you
250-ENHANCEDSTATUSCODES
250-PIPELINING
250-8BITMIME
250-SIZE
250-DSN
250-ETRN
250-DELIVERBY
250 HELP
MAIL From:<root@recognition.dipac.it.kmitl.ac.th> SIZE=574
250 2.1.0 <root@recognition.dipac.it.kmitl.ac.th>... Sender ok
RCPT To:<hanu@killbil.dipac.it.kmitl.ac.th>
DATA
250 2.1.5 <hanu@killbil.dipac.it.kmitl.ac.th>... Recipient ok
354 Enter mail, end with "." on a line by itself
Received: from recognition.dipac.it.kmitl.ac.th (localhost.localdomain [127.0.0.1])
        by recognition.dipac.it.kmitl.ac.th (8.12.5/8.12.5) with ESMTP id j4G5teMw001023
        for <hanu@killbil.dipac.it.kmitl.ac.th>; Mon, 16 May 2005 12:55:40 +0700
Received: from localhost (root@localhost)
        by recognition.dipac.it.kmitl.ac.th (8.12.5/8.12.5/Submit) with ESMTP id j4G5teqb001019
        for <hanu@killbil.dipac.it.kmitl.ac.th>; Mon, 16 May 2005 12:55:40 +0700
Date: Mon, 16 May 2005 12:55:39 +0700 (ICT)
From: root <root@recognition.dipac.it.kmitl.ac.th>
To: hanu@killbil.dipac.it.kmitl.ac.th
Subject: what 's going on?
Message-ID: <Pine.LNX.4.44.0505161250310.1006-100000@recognition.dipac.it.kmitl.ac.th>
MIME-Version: 1.0
Content-Type: TEXT/PLAIN; charset=US-ASCII

i don't know what 's happen there? please tell me details.

.
250 2.0.0 j4G5VKEs000770 Message accepted for delivery
QUIT
221 2.0.0 killbil.dipac.it.kmitl.ac.th closing connection
```

รูปที่ 2.1 Header และ Body ของจดหมายอิเล็กทรอนิกส์

มีรูปแบบคือ keyword : value โดย keyword หมายถึงคำสำคัญที่ใช้แสดงค่ารายละเอียดของ Header เช่น Date : , To : , From : เป็นต้น ส่วน value คือค่ารายละเอียดหรือข้อมูลของจดหมายอิเล็กทรอนิกส์ฉบับนั้น เช่น From : max@yahoo.com หมายความว่าจดหมายอิเล็กทรอนิกส์ฉบับนี้ถูกส่งมาจาก E-mail address ชื่อ max@yahoo.com เป็นต้น ส่วน Body คือส่วนของข้อความ หรือ Message ที่ผู้ส่งต้องการสื่อสารกับผู้รับซึ่งในปัจจุบันได้มีการพัฒนา MIME (Multipurpose Internet Mail Extensions) ขึ้นมาเพื่อทำให้ผู้ส่งสามารถแนบเอกสารอย่างอื่น เช่น รูปภาพ เสียง ไปกับจดหมายอิเล็กทรอนิกส์ได้

สำหรับการติดต่อสื่อสารแลกเปลี่ยนความคิดเห็นและข้อมูล ผ่านทางจดหมายอิเล็กทรอนิกส์นั้น ผู้ส่งสามารถกระจายข้อมูลไปให้ผู้รับคราวละหลายๆคนได้ โดยใช้ Cc : หรือ Bcc : Header แต่ในกรณีที่เป็นการร่วมแสดงความคิดเห็นของกลุ่มสมาชิกที่มีความสนใจในเรื่องเดียวกันแล้ว การติดต่อผ่านทางจดหมายอิเล็กทรอนิกส์ถือว่ามีความสะดวกมาก แต่อาจมีปัญหาจากการเปลี่ยนแปลงที่อยู่ของสมาชิก เช่น การสมัครเข้าเป็นสมาชิกใหม่ หรือ การย้ายออกจากการเป็นสมาชิก ทำให้สมาชิกในกลุ่มทุกคนต้องเสียเวลาแก้ไขข้อมูลให้ถูกต้อง ซึ่งอาจเกิดความผิดพลาดทำให้การติดต่อสื่อสารเป็นไปอย่างไม่ถูกต้อง นั่นจึงเป็นที่มาของการนำระบบเมลลิงลิสต์เข้ามาใช้ ดังจะกล่าวในหัวข้อถัดไป

2.2 เมลลิงลิสต์

แนวทางการพัฒนาระบบการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์นั้น มีความจำเป็นที่จะต้องศึกษาการทำงานของเมลลิงลิสต์ในรูปแบบต่างๆ รวมถึงศึกษาลักษณะเฉพาะและค้นหาความแตกต่างของจดหมายอิเล็กทรอนิกส์ที่เกิดจากเมลลิงลิสต์ เปรียบเทียบกับจดหมายอิเล็กทรอนิกส์แบบธรรมดาทั่วๆไปซึ่งสามารถสรุปได้ดังต่อไปนี้

2.2.1 ลักษณะทั่วๆไปของเมลลิงลิสต์

เมลลิงลิสต์คือการรวบรวมกลุ่ม E-mail Address ของสมาชิกเข้าไว้ด้วยกันและมี List Address เป็นตัวแทนของกลุ่ม เมื่อสมาชิกต้องการส่งข้อความหรือข้อมูลถึงสมาชิกคนอื่นๆสามารถทำได้โดยการส่งจดหมายอิเล็กทรอนิกส์จ่าหน้าถึง List Address และมี Mailing List Manager รับหน้าที่ทำสำเนาและกระจายจดหมายอิเล็กทรอนิกส์ให้กับสมาชิกที่อยู่ในบัญชีรายชื่อ หรือ List นอกจากนั้นยังมีหน้าที่จัดการเกี่ยวกับระบบสมาชิก เช่นการเปลี่ยนแปลงที่อยู่ หรือ E-mail Address ของสมาชิก การรับสมัครสมาชิก การย้ายออกของสมาชิก เป็นต้น จากการศึกษาพบว่าเมลลิงลิสต์นั้นสามารถแบ่งประเภทตามลักษณะการจัดการของ Manager หรือ ผู้จัดการของเมลลิงลิสต์ได้ 2 ประเภทคือ

1. Manually Management

พบในเมลลิงลิสต์ที่มีขนาดเล็ก มีจำนวนสมาชิกไม่มากนัก มีการเปลี่ยนแปลงเข้าและออกจากการเป็นสมาชิกไม่บ่อยนัก ส่วนใหญ่มักใช้ Alias Function ของ MTA เช่น Sendmail ช่วยจัดการเกี่ยวกับการกระจายจดหมายอิเล็กทรอนิกส์ Mailing list Manager หรือผู้จัดการเมลลิงลิสต์ มักใช้มนุษย์ในการจัดการดูแล จะพบปัญหาเกิดขึ้นในกรณีที่สมาชิกต้องการเปลี่ยนข้อมูล เช่น การเปลี่ยนแปลง E-mail Address เพื่อรับข้อมูล ผู้จัดการเมลลิงลิสต์อาจไม่สามารถจัดการได้ในทันทีทำให้สมาชิกขาดการติดต่อสื่อสาร ไม่สามารถรับข้อมูลในช่วงเวลาที่รอการทำงานของผู้จัดการเมลลิงลิสต์ และถ้ามีจำนวนสมาชิกเพิ่มมากขึ้นทำให้การดูแลทำได้ไม่ทั่วถึง

2. Software Management

พบในเมลลิงลิสต์ที่มีขนาดใหญ่ มีจำนวนสมาชิกมาก มีการเปลี่ยนแปลงเข้าออกของสมาชิกบ่อย มักจะใช้ Mailing List Software ช่วยในการจัดการ Software ที่นิยมใช้และพบมากในเครือข่ายอินเตอร์เน็ตคือ Majordomo เนื่องจากเป็น Software ที่ให้บริการฟรี ความสามารถของ Software เหล่านี้คือควบคุมระบบสมาชิกของเมลลิงลิสต์ไม่ว่าจะเป็นการสมัครสมาชิกหรือการยกเลิกสมาชิก และยังควบคุมการกระจายจดหมายอิเล็กทรอนิกส์ให้กับสมาชิกอีกด้วย

นอกจากนี้การใช้ Mailing List Software ในการจัดการยังสามารถแบ่งประเภทของเมลลิงลิสต์ตามลักษณะการเข้าถึงข้อมูลหรือสิทธิในการร่วมแสดงความคิดเห็นของสมาชิก เป็นการรักษาความปลอดภัยให้กับสมาชิกและข้อมูลของเมลลิงลิสต์ได้ดังต่อไปนี้

Open and Closed

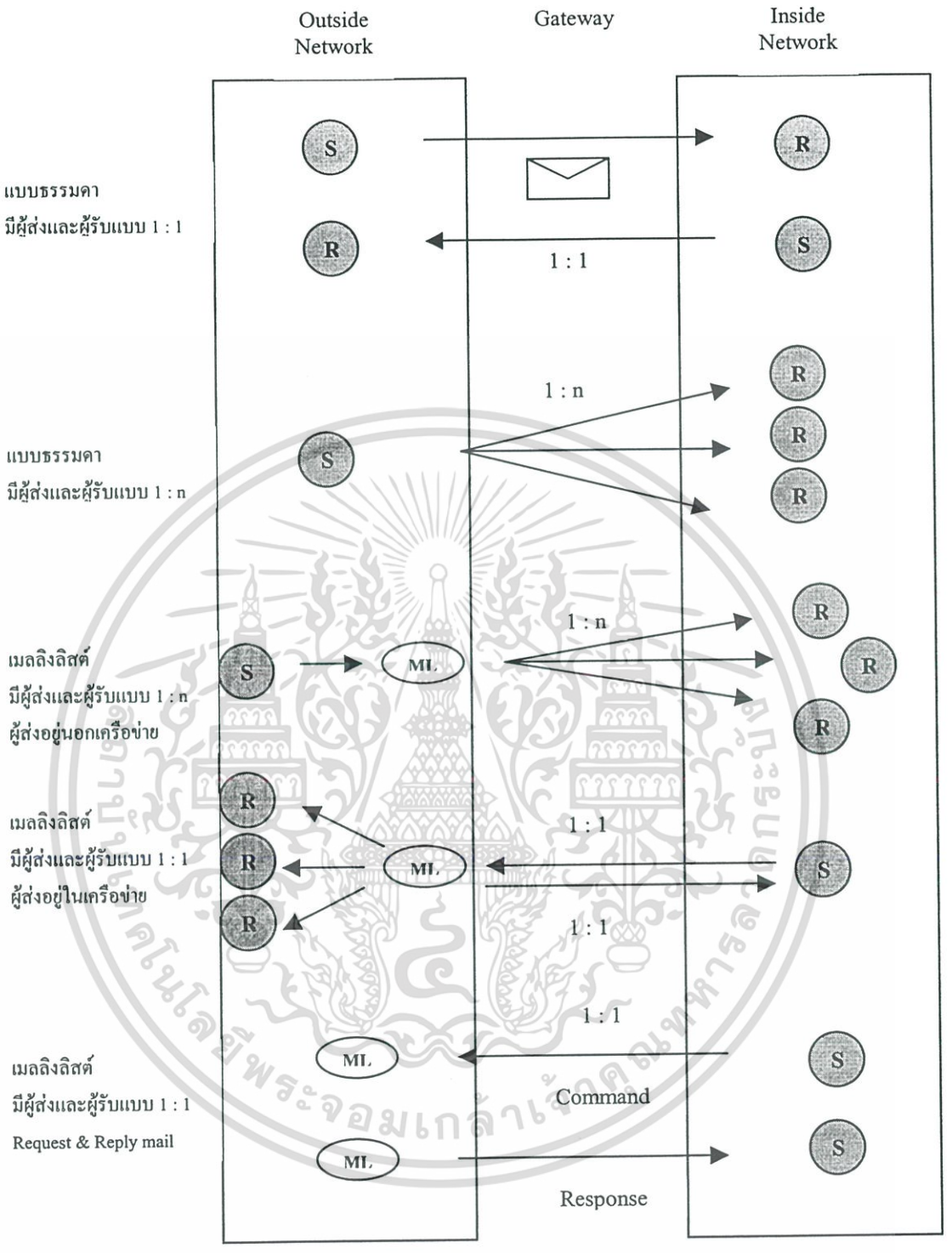
Open Mailing List คือเมลลิงลิสต์ที่อนุญาตให้ทุกๆคนสามารถสมัครสมาชิกได้แสดงให้เห็นข้อดีของการใช้งาน Mailing List Software ในการรับสมัครสมาชิกได้โดยอัตโนมัติแต่ถ้าเป็นแบบ Closed Mailing List การสมัครสมาชิกต้องได้รับอนุญาตจาก List Maintainer หรือ List Owner ก่อน

Public and Private

Public Mailing List คือเมลลิงลิสต์ที่อนุญาตให้บุคคลภายนอกสามารถ post ข้อความ หรือส่งข้อมูลมายังเมลลิงลิสต์ได้ แต่ถ้าเป็น Private Maiing List จะอนุญาตเฉพาะสมาชิก หรือ Subscriber เท่านั้นที่จะกระจายข้อมูลได้ เป็นการรักษาความปลอดภัยให้กับเมลลิงลิสต์และสมาชิกจากบุคคลภายนอกที่อาจสร้างความเสียหายให้แก่เมลลิงลิสต์ได้ เช่น จดหมายอิเล็กทรอนิกส์ที่มี virus เป็นต้น

Moderated

Moderated Mailing List คือเมลลิงลิสต์ที่มี Moderator คอยตรวจสอบข้อมูลหรือจดหมายอิเล็กทรอนิกส์ก่อนที่จะกระจายให้กับสมาชิก เป็นการป้องกันข้อความที่มีความไม่เหมาะสมเป็นการกลั่นกรองข้อมูลเพื่อรักษาความปลอดภัยให้กับสมาชิก



รูปที่ 2.2 ภาพเปรียบเทียบการทำงานของจดหมายอิเล็กทรอนิกส์แบบเมลลิงลิสต์และแบบธรรมดา

2.2.2 ลักษณะเฉพาะของจดหมายอิเล็กทรอนิกส์จากกลุ่มเมลลิงลิสต์

จากการศึกษาจดหมายอิเล็กทรอนิกส์ที่เกิดจากการทำงานของเมลลิงลิสต์ โดยเปรียบเทียบกับจดหมายอิเล็กกทรอนิกส์ธรรมดาทั่วๆไปสามารถสรุปได้ดังต่อไปนี้

1. การส่ง Request ถึงผู้จัดการเมลลิงลิสต์

คือการส่งจดหมายอิเล็กทรอนิกส์เพื่อร้องขอให้ผู้จัดการเมลลิงลิสต์ จัดการเกี่ยวกับระบบสมาชิก เช่น การสมัครสมาชิก การยกเลิกสมาชิก และการเปลี่ยนแปลงข้อมูลของสมาชิก โดยทั่วไปถ้าเป็นการจัดการแบบ Manually Management นั้นสมาชิกสามารถติดต่อสื่อสารกับผู้จัดการเมลลิงลิสต์ได้ โดยใช้ภาษาที่สื่อสารในชีวิตประจำวันได้ แต่ถ้าเป็นการจัดการแบบ Software Management นั้นไม่สามารถใช้คำพูดหรือภาษาที่ใช้ตามปกติได้ เนื่องจากระบบใช้ Software ในการควบคุมจัดการจำเป็นที่จะต้องใช้ Mailing List Command ในการสื่อสาร

2. การป้องกันไม่ให้สมาชิกรับจดหมายผิดประเภท

จดหมายอิเล็กทรอนิกส์ที่สมาชิกได้รับจากเมลลิงลิสต์นั้น ส่วนใหญ่จะเป็นข้อมูลหรือการแสดงความคิดเห็นของสมาชิกที่อยู่ในกลุ่ม จดหมายประเภทอื่นๆนอกเหนือจากนี้ เช่น จดหมายรายงานความผิดพลาดในการกระจายข้อมูลให้กับสมาชิก ผู้จัดการเมลลิงลิสต์มีหน้าที่ป้องกันไม่ให้ถูกส่งกลับไปยังสมาชิกผู้ Post ข้อความ หรือสมาชิกคนอื่นๆ โดยมีการป้องกันคือ การใช้ Header เพื่อรับจดหมายประเภทนี้โดยเฉพาะ เช่น Error-To : Header หรือการเปลี่ยนแปลง Mail Address ใน Header ของผู้ส่งจดหมายให้เป็นของ Manager Address แทน

3. ความแตกต่างของ Email Address ที่ระบุว่าเป็นผู้รับจดหมาย

การตรวจสอบว่าจดหมายอิเล็กทรอนิกส์ที่เข้ามามีใครบ้างเป็นผู้รับนั้น สามารถตรวจสอบได้จาก SMTP Command คือ RCPT TO และใน Mail Header คือ To : และ Cc : Header ถ้า Mail Address ที่พบจากทั้ง 2 ส่วนนี้แตกต่างกันแล้ว อาจเป็นไปได้ว่า Mail Address ใน SMTP Command เกิดจากการกระจายของ Mail Address ที่อยู่ใน Mail Header ซึ่งเป็นคุณสมบัติของเมลลิงลิสต์

4. พบชื่อชนิดของ Mailing List Software ใน E-mail Address ของผู้รับและผู้ส่ง

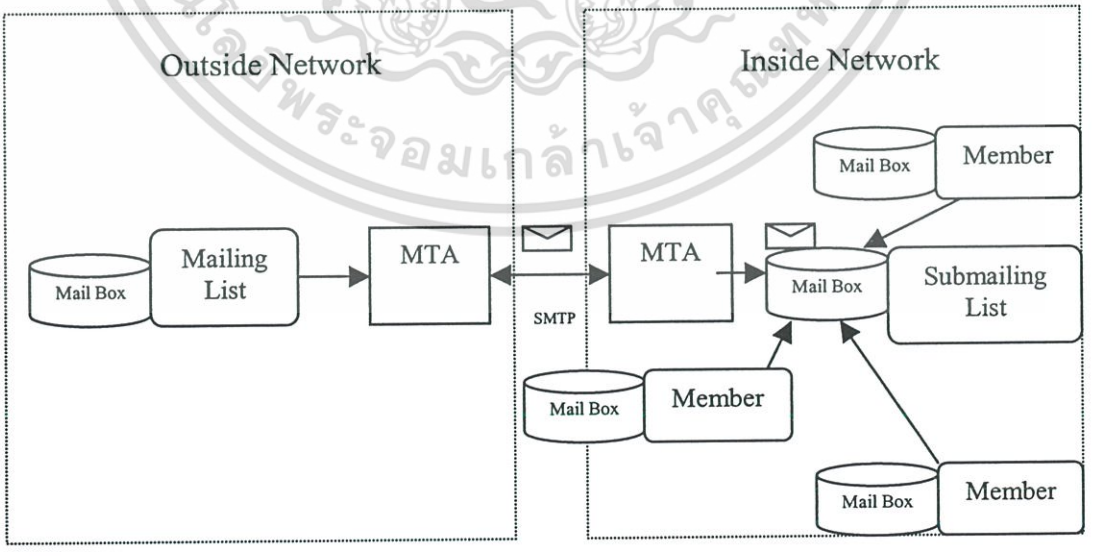
ถ้าเป็นการจัดการแบบ Manually Management พบว่ามักจะใช้ E-mail Address ของผู้จัดการเมลลิงลิสต์เป็น Manager Address หรือ List Address แต่ถ้าเป็นแบบ Software Management มักจะพบว่ามีการนำชื่อของ Software มาตั้งเป็นชื่อของ Manager Address ส่วน List Address นั้นจะตั้งตามชื่อของกลุ่ม

เมื่อได้ข้อสรุปเกี่ยวกับลักษณะเฉพาะของจดหมายอิเล็กทรอนิกส์ที่เกิดจากการทำงานของเมลลิงลิสต์แล้ว ในหัวข้อถัดไปจะกล่าวถึงความต้องการของระบบเมลลิงลิสต์ย่อย เพื่อเป็นแนวทางการสร้างระบบตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ต่อไป

2.3 ระบบเมลลิงลิสต์ย่อย

คือการรวบรวมกลุ่มสมาชิกของเมลลิงลิสต์เข้าตามความใกล้เคียงกันของเครือข่าย ตั้งเป็น Submailing List หรือ ระบบเมลลิงลิสต์ย่อย ขึ้นมาเพื่อรับจดหมายอิเล็กทรอนิกส์จาก Main Mailing List หรือ เมลลิงลิสต์หลัก เพียงฉบับเดียวจากนั้นจึงกระจายต่อให้กับสมาชิกภายในเมลลิงลิสต์ย่อยในภายหลัง โดยการส่งข้อความเพื่อแจ้งให้สมาชิกทราบว่ามีข้อมูลใหม่เข้ามาและระยะเวลาในการเก็บข้อมูล ซึ่งสมาชิกสามารถเข้าไปอ่านข้อมูลเหล่านั้นและคัดลอกข้อมูลที่สำคัญมาเก็บไว้ได้ภายในระยะเวลาที่กำหนดไว้ หลักการทำงานของระบบคือความต้องการลดปริมาณข้อมูลจดหมายอิเล็กทรอนิกส์ที่ซ้ำซ้อนไม่ให้เข้ามาในเครือข่าย เพื่อลดความหนาแน่นของการจราจรระหว่างเครือข่ายเป็นการเพิ่มประสิทธิภาพและความเร็วในการถ่ายโอนข้อมูลระหว่างเครือข่ายได้ดียิ่งขึ้น และยังประหยัดพื้นที่ที่เก็บข้อมูลใน Mail Box ของสมาชิกแต่ละคนอีกด้วย

จากการศึกษาการทำงานของระบบเมลลิงลิสต์ย่อย เพื่อทราบถึงข้อมูลและรายละเอียดที่ระบบต้องการนำไปใช้ พบว่า Mail Address ที่ระบบเมลลิงลิสต์ย่อยต้องการนำไปใช้คือ List Address, Member Address, Manager Address และ Manager Type หลังจากตรวจพบว่าจดหมายอิเล็กทรอนิกส์มาจากกลุ่มเมลลิงลิสต์แล้ว ระบบเมลลิงลิสต์ย่อยต้องการทราบข้อมูลของ Manager Address และ Manager Type เพื่อจัดการสมัครเป็นสมาชิก จากนั้นจึงรวบรวม Member Address เข้าไว้ด้วยกันและเปลี่ยนสถานะจากการเป็นสมาชิกของเมลลิงลิสต์หลักให้มาเป็นสมาชิกของเมลลิงลิสต์ย่อยแทน เมื่อมีจดหมายอิเล็กทรอนิกส์ส่งมาจาก List Address ระบบจะจัดการแจ้งข่าวกระจายให้กับ Member Address ที่รวบรวมไว้ จะเห็นว่านอกจากจะลดปริมาณข้อมูลที่เข้ามายังเครือข่ายและช่วยลดการใช้ทรัพยากรของเครือข่ายแล้ว ระบบเมลลิงลิสต์ย่อยยังช่วยลดภาระการทำงานของเมลลิงลิสต์หลักอีกด้วย



รูปที่ 2.3 การทำงานของระบบเมลลิงลิสต์ย่อย

ในหัวข้อถัดไปจะกล่าวถึงปัญหาที่เกิดจากเมลลิงลิสต์และระบบเมลลิงลิสต์ย่อย รวมถึงที่มาของงานวิจัย

2.4 ปัญหาและที่มาของงานวิจัย

จากการกระจายจดหมายอิเล็กทรอนิกส์ของผู้จัดการเมลลิงลิสต์ พบว่าเมื่อมีสมาชิกส่งข้อความเข้ามาที่ List Address ผู้จัดการเมลลิงลิสต์จะจัดการทำสำเนาของจดหมายอิเล็กทรอนิกส์ฉบับนั้นแล้วจัดส่งให้กับสมาชิกตาม E-mail Address ที่มีอยู่ใน List โดยไม่สนใจความใกล้เคียงกันของเครือข่าย ทำให้เกิดความซ้ำซ้อนของข้อมูลจำนวนที่เดินทางมายังเครือข่าย ทำให้ประสิทธิภาพของการจราจรระหว่างเครือข่ายลดลง และยังสิ้นเปลืองพื้นที่ในการเก็บข้อมูลที่ซ้ำซ้อนกันของสมาชิกแต่ละคน จึงเป็นที่มาของการพัฒนาระบบเมลลิงลิสต์ย่อยเพื่อลดจำนวนและปริมาณข้อมูลที่เข้ามาในเครือข่าย ด้วยการรวบรวมกลุ่มสมาชิกของเมลลิงลิสต์เข้าไว้ด้วยกันแล้วจัดการ ตามที่กล่าวไว้ในหัวข้อ 2.3

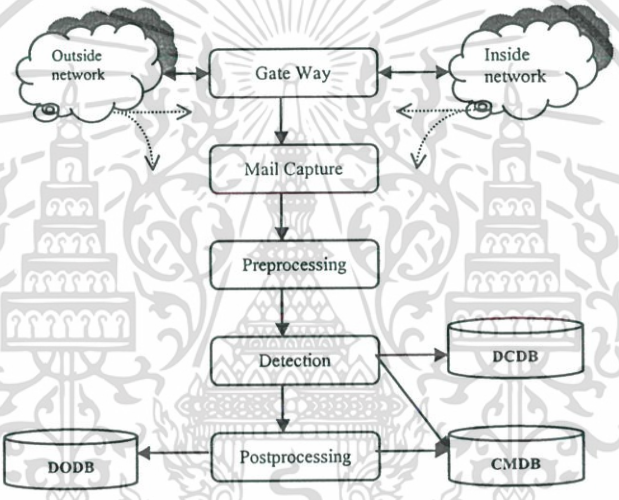
แต่ปัญหาที่ตามมาคือ ระบบเมลลิงลิสต์ย่อยไม่สามารถจำแนกประเภทของจดหมายอิเล็กทรอนิกส์และคัดเลือกข้อมูล Mail Address ที่ต้องการเองได้ยังต้องอาศัยความสามารถของ Network Administrator ช่วยคัดกรองจดหมายอิเล็กทรอนิกส์ที่ผ่านเข้ามาในเครือข่ายและนำข้อมูลไปป้อนให้กับระบบเมลลิงลิสต์ย่อย ซึ่งในวันหนึ่งๆถ้ามีจดหมายอิเล็กทรอนิกส์จำนวนมาก Network Administrator อาจไม่สามารถตรวจสอบได้หมดภายในวันเดียว ทำให้มีจดหมายตกค้างและไม่สามารถตรวจสอบหาจดหมายอิเล็กทรอนิกส์ของกลุ่มเมลลิงลิสต์ได้ทั้งหมด นอกจากนั้นอาจมีเมลลิงลิสต์เกิดขึ้นใหม่เรื่อยๆ อาจเกิดความผิดพลาดในการตรวจสอบทำให้ระบบเมลลิงลิสต์ย่อยไม่สามารถแก้ไขปัญหาได้ ทำให้ประสิทธิภาพของการจราจรระหว่างเครือข่ายลดลงและสิ้นเปลืองทรัพยากรของเครือข่าย

จึงเป็นที่มาของงานวิจัยเพื่อสร้างระบบตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ เพื่อทำหน้าที่ตรวจสอบจดหมายอิเล็กทรอนิกส์และคัดเลือกข้อมูลที่ระบบเมลลิงลิสต์ย่อยต้องการ ดังจะกล่าวในหัวข้อถัดไป

2.5 ระบบการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์

Mailing List E-mail Detection System หรือ ระบบการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ คือระบบที่สร้างขึ้นเพื่อวิเคราะห์คุณสมบัติและลักษณะเฉพาะของจดหมายอิเล็กทรอนิกส์ที่เกิดจากการทำงานของเมลลิงลิสต์ โดยคำนวณค่าความมั่นใจ หรือ Confidential Factor เพื่อใช้ประกอบการตัดสินใจ

การทำงานของระบบเริ่มจากการดักจับข้อมูลของจดหมายอิเล็กทรอนิกส์ที่ผ่านเข้าและออกจากเครือข่าย โดยการคัดลอกข้อมูลของจดหมายอิเล็กทรอนิกส์ เมื่อได้ข้อมูลดิบมาแล้วนำข้อมูลเหล่านั้นเข้าสู่กระบวนการจัดเตรียมข้อมูลให้อยู่ในรูปแบบเดียวกัน จากนั้นผ่านเข้าสู่ขั้นตอนการตรวจจับ ในกระบวนการนี้เป็นขั้นตอนการวิเคราะห์เพื่อหาลักษณะเฉพาะของจดหมายอิเล็กทรอนิกส์จากกลุ่มเมลลิงลิสต์ เมื่อพบข้อมูลที่ตรงกับข้อสันนิษฐานเบื้องต้น ระบบจะคำนวณค่าความมั่นใจให้กับข้อมูล โดยค่าความมั่นใจสามารถแบ่งได้เป็นหลายประเภท คือ ค่าความมั่นใจที่แสดงว่า Mail Address นี้เป็น List Address, ค่าความมั่นใจที่แสดงว่า Mail Address นี้เป็น Member Address, ค่าความมั่นใจที่แสดงว่า Mail Address นี้เป็น Manager Address และ ค่าความมั่นใจที่แสดงว่าเมลลิงลิสต์นี้ใช้ Software ชนิดนี้ (Manager Type) ในการจัดการ การกำหนดค่าความมั่นใจนี้ขึ้นอยู่กับประเภทของข้อมูลที่นำไปใช้ในการกำหนดค่าต่างๆให้กับระบบเมลลิงลิสต์ย่อย



รูปที่ 2.4 Mailing List E-mail Dectection System

ขั้นตอนต่อไปคือการพิจารณาตัดสินใจคัดเลือกข้อมูลเพื่อส่งต่อไปให้ระบบเมลลิงลิสต์ย่อย ในกระบวนการนี้เปรียบเสมือนการกลั่นกรองข้อมูล เมื่อพบข้อมูลอื่นๆ หรือ จดหมายฉบับอื่นที่มีความขัดแย้งกับข้อมูลเดิมที่สรุปได้ ระบบจะลดค่าความมั่นใจลง ในทางกลับกันถ้าพบข้อมูลที่สนับสนุนข้อมูลเดิม ระบบจะเพิ่มค่าความมั่นใจขึ้น สำหรับการตัดสินใจเลือกข้อมูลนั้นจะคัดเลือกโดยพิจารณาจากค่าความมั่นใจเปรียบเทียบกับ Threshold Confidential Factor ซึ่งค่า CF_{TH} นี้นำมาจากการทดลองหลายๆครั้งจนได้ค่าความมั่นใจที่สามารถยืนยันได้ว่าข้อมูลที่มีค่าความมั่นใจมากกว่า หรือ เท่ากับ CF_{TH} เป็นข้อมูลจดหมายอิเล็กทรอนิกส์ที่เกิดจากกลุ่มเมลลิงลิสต์

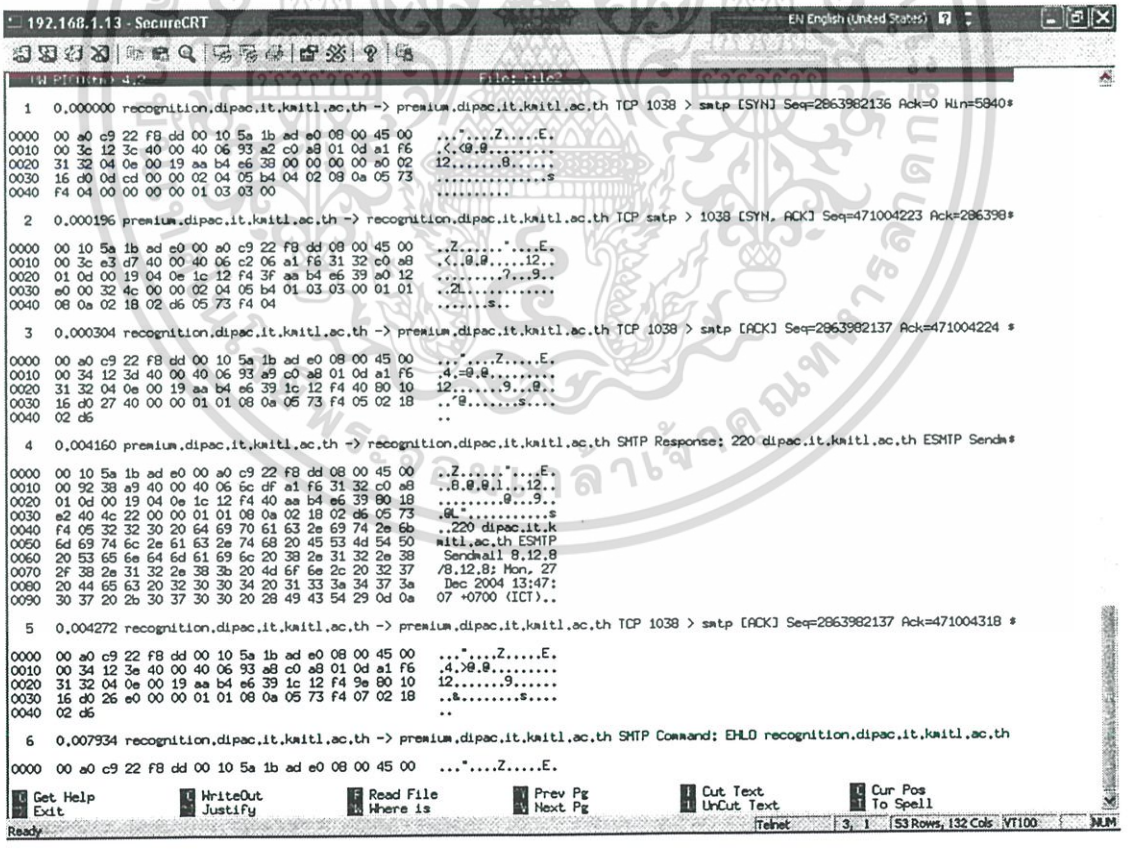
ในหัวข้อถัดไปจะกล่าวถึงขั้นตอนการทำงานในแต่ละกระบวนการ รวมถึงการออกแบบและวิธีการดำเนินงานวิจัยการสร้างระบบตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์

บทที่ 3

Mailing List E-mail Detection System Process

ขั้นตอนการสร้างระบบการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ เริ่มจากการดักจับข้อมูลของจดหมายอิเล็กทรอนิกส์ที่ผ่านเข้าและออกจากเครือข่าย (Capture Process) เมื่อได้ข้อมูลมาแล้วจากนั้นเข้าสู่กระบวนการถัดไปคือ กระบวนการจัดเตรียมข้อมูล (Preprocessing Process) เพื่อจัดเตรียมข้อมูลของจดหมายอิเล็กทรอนิกส์ให้อยู่ในรูปแบบเดียวกัน จากนั้นเข้าสู่กระบวนการตรวจจับ (Detection Process) เพื่อวิเคราะห์หาลักษณะเฉพาะของจดหมายอิเล็กทรอนิกส์จากกลุ่มเมลลิงลิสต์และคำนวณหาค่าความมั่นใจ (Confidential Factor) สุดท้ายเป็นกระบวนการตัดสินใจ (Postprocessing Process) พิจารณาคัดเลือกข้อมูลที่มีคุณลักษณะของเมลลิงลิสต์ แล้วรายงานผลให้ Network Administrator ทราบเพื่อนำข้อมูลที่สรุปได้ไปจัดการ Configuration ระบบเมลลิงลิสต์ย่อยต่อไป

3.1 Capture Process



รูปที่ 3.1 แสดงตัวอย่างข้อมูลจาก Capture Process

เริ่มต้นด้วย Capture Process หรือ กระบวนการดักจับข้อมูลของจดหมายอิเล็กทรอนิกส์ที่ผ่านเข้าและออกจากเครือข่าย จากศึกษาการทำงานของจดหมายอิเล็กทรอนิกส์พบว่าการแลกเปลี่ยนข้อมูลระหว่าง MTA ของผู้รับและผู้ส่งจดหมายอิเล็กทรอนิกส์นั้นใช้โปรโตคอล SMTP ในการดำเนินการ ดังนั้นในขั้นตอนนี้จำเป็นต้องดักจับข้อมูลที่ผ่านเข้าและออกจาก port 25 ซึ่งเป็น port ที่ใช้ในการสื่อสารแลกเปลี่ยนข้อมูลของ SMTP โดยในขั้นตอนนี้ได้ใช้ tool คือ Tethereal 0.9.6 เป็น tool ที่ใช้ดักจับข้อมูลของโปรโตคอลบนเครือข่าย อยู่ในสภาพแวดล้อมคือ Redhat8.0 ซึ่งข้อมูลที่ได้จะอยู่ในรูปของ text format

โดยการดักจับข้อมูลจดหมายอิเล็กทรอนิกส์นั้นเป็นการคัดลอกข้อมูลเท่านั้น ไม่มีการเปลี่ยนแปลงข้อมูลของจดหมาย พบว่าข้อมูลที่ดักจับได้นั้นแสดงการแลกเปลี่ยนข้อมูลของผู้รับและผู้ส่งจดหมายอิเล็กทรอนิกส์โดย MTA ของทั้งสองฝ่าย จากตัวอย่างสามารถแบ่งข้อมูลออกเป็น 2 ส่วน คือ SMTP Envelop หรือ MTA Conversation และ SMTP Content โดยในส่วนแรกนั้นประกอบด้วย SMTP Command และ SMTP Reply Code ดังรูปที่ 3.2

```
50   0.300933 premium.dipac.it.kmitl.ac.th -> detection.dipac.it.kmitl.ac.th SMTP Response: 250 2.1.0 <hanu@detection.dipac

0000  00 06 5b 7d 1a d9 00 a0 c9 22 f8 dd 08 00 45 00   ..[}.....".....E.
0010  00 72 d8 73 40 00 40 06 cd 33 a1 f6 31 32 c0 a8   .r.s@.@..3..12..
0020  01 0e 00 19 80 1c 58 64 a6 62 91 76 78 6d 80 18   ......Xd.b.vxm..
0030  e2 40 e5 4c 00 00 01 01 08 0a 12 88 43 65 00 56   .@.L........Ce.V
0040  21 61 32 35 30 20 32 2e 31 2e 30 20 3c 68 61 6e   !a250 2.1.0 <han
0050  75 40 64 65 74 65 63 74 69 6f 6e 2e 64 69 70 61   u@detection.dipa
0060  63 2e 69 74 2e 6b 6d 69 74 6c 2e 61 63 2e 74 68   c.it.kmitl.ac.th
0070  3e 2e 2e 2e 20 53 65 6e 64 65 72 20 6f 6b 0d 0a   >... Sender ok..

 51   0.302111 detection.dipac.it.kmitl.ac.th -> premium.dipac.it.kmitl.ac.th SMTP Command: RCPT To:<panic1@premium.dipac.it

0000  00 a0 c9 22 f8 dd 00 06 5b 7d 1a d9 08 00 45 00   ...".....[}....E.
0010  00 69 59 55 40 00 40 06 4c 5b c0 a8 01 0e a1 f6   .iYU@.@.L[......
0020  31 32 80 1c 00 19 91 76 78 6d 58 64 a6 a0 80 18   12.....vxmXd....
0030  19 20 24 36 00 00 01 01 08 0a 00 56 21 64 12 88   . $6.......V!d..
0040  43 65 52 43 50 54 20 54 6f 3a 3c 70 61 6e 69 63   CeRCPT To:<panic
0050  31 40 70 72 65 6d 69 75 6d 2e 64 69 70 61 63 2e   1@premium.dipac.
0060  69 74 2e 6b 6d 69 74 6c 2e 61 63 2e 74 68 3e 0d   it.kmitl.ac.th>.
0070  0a 44 41 54 41 0d 0a                              .DATA..

 52   0.305064 premium.dipac.it.kmitl.ac.th -> detection.dipac.it.kmitl.ac.th SMTP Response: 250 2.1.5 <panic1@premium.dipac

0000  00 06 5b 7d 1a d9 00 a0 c9 22 f8 dd 08 00 45 00   ..[}.....".....E.
0010  00 a7 03 39 40 00 40 06 a2 39 a1 f6 31 32 c0 a8   ...9@.@..9..12..
0020  01 0e 00 19 80 1c 58 64 a6 a0 91 76 78 a2 80 18   ......Xd...vx...
0030  e2 40 75 7b 00 00 01 01 08 0a 12 88 43 66 00 56   .@u{........Cf.V
0040  21 64 32 35 30 20 32 2e 31 2e 35 20 3c 70 61 6e   !d250 2.1.5 <pan
0050  69 63 31 40 70 72 65 6d 69 75 6d 2e 64 69 70 61   ic1@premium.dipa
0060  63 2e 69 74 2e 6b 6d 69 74 6c 2e 61 63 2e 74 68   c.it.kmitl.ac.th
0070  3e 2e 2e 2e 20 52 65 63 69 70 69 65 6e 74 20 6f   >... Recipient o
0080  6b 0d 0a 33 35 34 20 45 6e 74 65 72 20 6d 61 69   k..354 Enter mai
0090  6c 2c 20 65 6e 64 20 77 69 74 68 20 22 2e 22 20   l, end with "."
00a0  6f 6e 20 61 20 6c 69 6e 65 20 62 79 20 69 74 73   on a line by its
00b0  65 6c 66 0d 0a                                    elf..

 53   0.307240 detection.dipac.it.kmitl.ac.th -> premium.dipac.it.kmitl.ac.th SMTP Message Body

0000  00 a0 c9 22 f8 dd 00 06 5b 7d 1a d9 08 00 45 00   ...".....[}....E.
0010  03 2c 59 56 40 00 40 06 49 97 c0 a8 01 0e a1 f6   .,YV@.@.I.......
0020  31 32 80 1c 00 19 91 76 78 a2 58 64 a7 13 80 18   12.....vx.Xd....
0030  19 20 5e 64 00 00 01 01 08 0a 00 56 21 66 12 88   . ^d.......V!f..
0040  43 66 52 65 63 65 69 76 65 64 3a 20 66 72 6f 6d   CfReceived: from
```

รูปที่ 3.2 แสดงตัวอย่าง SMTP Command และ SMTP Reply Code

ในส่วนที่สองนั้น คือจดหมายอิเล็กทรอนิกส์ที่ต้องการส่งประกอบด้วย Header และ Body ดังแสดงในรูปที่ 3.3 จากนั้นเริ่มเข้าสู่กระบวนการจัดเตรียมข้อมูลต่อไป

```
 15   0.043506 HML.dipac.it.kmitl.ac.th -> detection.dipac.it.kmitl.ac.th SMTP Message Body

0000  00 06 5b 7d 1a d9 00 e0 29 8f 86 ef 08 00 45 00   ..[}....).....E.
0010  02 37 66 e9 40 00 40 06 4e 73 c0 a8 01 06 c0 a8   .7f.@.@.Ns......
0020  01 0e 06 2c 00 19 92 b3 a3 95 91 7f f9 aa 80 18   ...,............
0030  19 20 2a 07 00 00 01 01 08 0a 07 3a c5 e8 00 56   . *........:...V
0040  20 de 52 65 63 65 69 76 65 64 3a 20 66 72 6f 6d    .Received: from
0050  20 6c 6f 63 61 6c 68 6f 73 74 20 28 68 61 69 72    localhost (hair
0060  32 40 6c 6f 63 61 6c 68 6f 73 74 29 0d 0a 09 62   2@localhost)...b
0070  79 20 48 4d 4c 2e 64 69 70 61 63 2e 69 74 2e 6b   y HML.dipac.it.k
0080  6d 69 74 6c 2e 61 63 2e 74 68 20 28 38 2e 31 31   mitl.ac.th (8.11
0090  2e 36 2f 38 2e 31 31 2e 36 29 20 77 69 74 68 20   .6/8.11.6) with
00a0  45 53 4d 54 50 20 69 64 20 6a 33 4b 37 76 47 4a   ESMTP id j3K7vGJ
00b0  30 33 32 30 35 0d 0a 09 66 6f 72 20 3c 64 65 74   03205...for <det
00c0  65 63 74 69 6f 6e 40 64 65 74 65 63 74 69 6f 6e   ection@detection
00d0  2e 64 69 70 61 63 2e 69 74 2e 6b 6d 69 74 6c 2e   .dipac.it.kmitl.
00e0  61 63 2e 74 68 3e 3b 20 57 65 64 2c 20 32 30 20   ac.th>; Wed, 20
00f0  41 70 72 20 32 30 30 35 20 31 34 3a 35 37 3a 31   Apr 2005 14:57:1
0100  36 20 2b 30 37 30 30 0d 0a 44 61 74 65 3a 20 57   6 +0700..Date: W
0110  65 64 2c 20 32 30 20 41 70 72 20 32 30 30 35 20   ed, 20 Apr 2005
0120  31 34 3a 35 37 3a 31 36 20 2b 30 37 30 30 20 28   14:57:16 +0700 (
0130  49 43 54 29 0d 0a 46 72 6f 6d 3a 20 68 61 69 72   ICT)..From: hair
0140  32 40 48 4d 4c 2e 64 69 70 61 63 2e 69 74 2e 6b   2@HML.dipac.it.k
0150  6d 69 74 6c 2e 61 63 2e 74 68 0d 0a 54 6f 3a 20   mitl.ac.th..To:
0160  64 65 74 65 63 74 69 6f 6e 40 64 65 74 65 63 74   detection@detect
0170  69 6f 6e 2e 64 69 70 61 63 2e 69 74 2e 6b 6d 69   ion.dipac.it.kmi
0180  74 6c 2e 61 63 2e 74 68 0d 0a 53 75 62 6a 65 63   tl.ac.th..Subjec
0190  74 3a 20 62 75 73 69 6e 65 73 73 0d 0a 4d 65 73   t: business..Mes
01a0  73 61 67 65 2d 49 44 3a 20 3c 50 69 6e 65 2e 4c   sage-ID: <Pine.L
01b0  4e 58 2e 34 2e 34 34 2e 30 35 30 34 32 30 31 34   NX.4.44.05042014
01c0  35 36 34 31 30 2e 33 32 30 30 2d 31 30 30 30 30   56410.3200-10000
01d0  30 40 48 4d 4c 2e 64 69 70 61 63 2e 69 74 2e 6b   0@HML.dipac.it.k
01e0  6d 69 74 6c 2e 61 63 2e 74 68 3e 0d 0a 4d 49 4d   mitl.ac.th>..MIM
01f0  45 2d 56 65 72 73 69 6f 6e 3a 20 31 2e 30 0d 0a   E-Version: 1.0..
0200  43 6f 6e 74 65 6e 74 2d 54 79 70 65 3a 20 54 45   Content-Type: TE
0210  58 54 2f 50 4c 41 49 4e 3b 20 63 68 61 72 73 65   XT/PLAIN; charse
0220  74 3d 55 53 2d 41 53 43 49 49 0d 0a 0d 0a 64 65   t=US-ASCII....de
0230  76 65 6c 6f 70 6d 65 6e 74 20 6d 61 6e 61 67 65   velopment manage
0240  72 0d 0a 0d 0a                                    r....

 16   0.083143 detection.dipac.it.kmitl.ac.th -> HML.dipac.it.kmitl.ac.th TCP smtp > 1580 [AC
32 Len=0

0000  00 e0 29 8f 86 ef 00 06 5b 7d 1a d9 08 00 45 00   ..).....[}....E.
0010  00 34 d7 13 40 00 40 06 e0 4b c0 a8 01 0e c0 a8   .4..@.@..K......
0020  01 06 00 19 06 2c 91 7f f9 aa 92 b3 a5 98 80 10   .....,..........
```

รูปที่ 3.3 แสดงตัวอย่าง Header และ Body ของจดหมายอิเล็กทรอนิกส์

3.2 Preprocessing Process

ในขั้นตอนนี้เป็นการจัดเตรียมข้อมูลให้อยู่ในรูปแบบเดียวกันคือ Detection Format เนื่องเครือข่ายอินเตอร์เน็ตมีการใช้งาน MTA หลายรูปแบบแตกต่างกัน ดังนั้นจึงต้องจัดการให้ข้อมูลของจดหมายอิเล็กทรอนิกส์ที่ได้จาก Capture Process อยู่ในรูปแบบเดียวกันก่อนดังรูปที่ 3.4 ซึ่งใน Detection Format นั้นสามารถแบ่งออกเป็น 2 รูปแบบคือ ส่วนที่เป็นการแลกเปลี่ยนข้อมูลระหว่าง MTA หรือ SMTP Envelop เรียกว่า SMTP Variable Format และในส่วนที่สองคือส่วนที่เป็นจดหมายอิเล็กทรอนิกส์ หรือ SMTP Content เรียกว่า Rfc822 Variable Format ตามมาตรฐานของรูปแบบจดหมายอิเล็กทรอนิกส์ สามารถสรุปรายละเอียดได้ดังตารางที่ 3.1

```
<smtp>
        <smtp_envelop>
                <smtp_source_domain>killbil.dipac.it.kmitl.ac.th</smtp_source_domain>
                <smtp_arrive_time>Wed, 20 Apr 2005 20:21:39 +0700</smtp_arrive_time>
                <smtp_destination_domain>detection.dipac.it.kmitl.ac.th</smtp_destination_domain>
                <smtp_mail_from>kangkaroo1@killbil.dipac.it.kmitl.ac.th</smtp_mail_from>
                <smtp_rcpt_to>damage1@detection.dipac.it.kmitl.ac.th</smtp_rcpt_to>
        </smtp_envelop>
        <smtp_content>
                <mail_header>
                        <rfc822_from>kangkaroo1@killbil.dipac.it.kmitl.ac.th</rfc822_from>
                        <rfc822_to>damage1@detection.dipac.it.kmitl.ac.th</rfc822_to>
                        <rfc822_message_id><Pine.LNX.4.44.0504201953530.5342-100000@killbil.dipac.it.kmitl.ac.th></rfc822_message_id
>
                        <rfc822_subject>li & fung</rfc822_subject>
                </mail_header>
                <mail_body>
thaila      nd limited....
                </mail_body>
        </smtp_content>
</smtp>
<smtp>
        <smtp_envelop>
                <smtp_source_domain>killbil.dipac.it.kmitl.ac.th</smtp_source_domain>
                <smtp_arrive_time>Wed, 20 Apr 2005 20:23:27 +0700</smtp_arrive_time>
                <smtp_destination_domain>detection.dipac.it.kmitl.ac.th</smtp_destination_domain>
                <smtp_mail_from>knight1@killbil.dipac.it.kmitl.ac.th</smtp_mail_from>
                <smtp_rcpt_to>day2@detection.dipac.it.kmitl.ac.th<dog2@detection.dipac.it.kmitl.ac.th</smtp_rcpt_to>
        </smtp_envelop>
        <smtp_content>
                <mail_header>
                        <rfc822_from>knight1@killbil.dipac.it.kmitl.ac.th</rfc822_from>
                        <rfc822_to>day2@detection.dipac.it.kmitl.ac.th</rfc822_to>
                        <rfc822_message_id><Pine.LNX.4.44.0504202020340.5381-100000@killbil.dipac.it.kmitl.ac.th></rfc822_message_id
>
                        <rfc822_subject>li & fung limited</rfc822_subject>
                </mail_header>
                <mail_body>
is today one of the premier global consumer products export trading ..companies..                                                ..
                </mail_body>
        </smtp_content>
</smtp>
<smtp>
        <smtp_envelop>
                <smtp_source_domain>detection.dipac.it.kmitl.ac.th</smtp_source_domain>
                <smtp_arrive_time>Wed, 20 Apr 2005 20:23:41 +0700</smtp_arrive_time>
                <smtp_destination_domain>killbil.dipac.it.kmitl.ac.th</smtp_destination_domain>
                <smtp_mail_from>dance2@detection.dipac.it.kmitl.ac.th</smtp_mail_from>
```

รูปที่ 3.4 Detection Format

ตารางที่ 3.1 Detection Variables

Variable	Description	Type
Smtp_source_domain	Domain ได้จาก smtp conversation	Single
Smtp_destination_domain	Domain ได้จาก smtp conversation	Single
Smtp_mail_from	Mail address ได้จาก smtp command 'MAIL FROM'	Single
Smtp_rcpt_to	Mail address ได้จาก smtp command 'RCPT TO'	Multi
Smtp_arrive_time	แสดงเวลาที่ mail เข้ามาในเครือข่าย	Single
Rfc822_from	Mail address ได้จาก From: header	Single
Rfc822_sender	Mail address ได้จาก Sender: header	Single
Rfc822_return_path	Mail address ได้จาก Return-Path: header	Single
Rfc822_reply_to	Mail address ได้จาก Reply-To: header	Single
Rfc822_error_to	Mail address ได้จาก Error-To: header	Single
Rfc822_to	Mail address ได้จาก To: header	Multi
Rfc822_cc	Mail address ได้จาก Cc: header	Multi
Rfc822_message_id	Unique code ได้จาก Message-id: header	Single
Rfc822_in_reply_to	Unique code ได้จาก In-Reply-To: header	Single

ตารางที่ 3.1 Detection Variables (ต่อ)

Variable	Description	Type
Rfc822_references	Unique code ได้จาก Reference: header	Single
Rfc822_subject	ได้จาก Subject: header	Single
Rfc822_mail_body	Mail Body	Single

สำหรับข้อมูล Variable ต่างๆนั้นคือข้อมูลที่ได้จากจดหมายอิเล็กทรอนิกส์ โดยมี Type เป็น 2 ชนิดคือ Single และ Muti หมายความว่า Variable แต่ละมีข้อมูลที่มีค่าเดียวหรือเป็นข้อมูลที่สามารถมีได้หลายค่า เช่น Smtp_mail_from คือ E-mail Address ของผู้ส่งจดหมายอิเล็กทรอนิกส์นั้นสามารถมีได้เพียงคนเดียวต่อจดหมาย 1 ฉบับ นั่นคือมี Type เป็นแบบ Single แต่สำหรับ Smtp_rcpt_to นั้นคือ E-mail Address ของผู้รับจดหมายซึ่งสามารถมีผู้รับได้หลายคนต่อจดหมาย 1 ฉบับจะกำหนด Type เป็นแบบ Multi เป็นต้น เมื่อได้ข้อมูลอยู่ในรูปแบบที่ต้องการแล้ว จากนั้นจึงเข้าสู่กระบวนการถัดไป

3.3 Detection Process

ในขั้นตอนนี้เป็นกระบวนการวิเคราะห์จดหมายอิเล็กทรอนิกส์เพื่อค้นหาลักษณะเฉพาะของจดหมายอิเล็กทรอนิกส์จากกลุ่มเมลลิงลิสต์ และคำนวณหาค่าความมั่นใจ หรือ Confidential Factor (CF) จากนั้นนำข้อมูลที่ได้บันทึกลงในฐานข้อมูล ซึ่งในระบบการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ประกอบด้วยฐานข้อมูลที่สำคัญดังต่อไปนี้

3.3.1 Detection Output Database (DODB)

คือฐานข้อมูลของผลลัพธ์ที่ได้จากกระบวนการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ ข้อมูลที่ได้จะถูกนำไปพิจารณาในขั้นตอนสุดท้ายเพื่อรายงานให้ Network Administrator ทราบและนำข้อมูลไป Config ระบบเมลลิงลิสต์ย่อย ประกอบด้วย

1. List Address Table (LA Table)

คือตารางแสดงผลลัพธ์การวิเคราะห์ว่ามี Mail Address ใดบ้างที่สรุปว่าเป็น List Address ของเมลลิงลิสต์

ตารางที่ 3.2 List Address Table

List address	CF_LA
Listname@Listdomain	Value between 0 to 1

ประกอบด้วย 2 field ดังต่อไปนี้คือ

- List Address หมายถึง Mail Address ที่สรุปว่าเป็น List Address ของเมลลิงลิสต์
- CF_LA หมายถึงค่าความมั่นใจว่า Mail Address ที่พบเป็น List Address ของเมลลิงลิสต์

2. List Member Address Table (LM Table)

คือตารางแสดงผลลัพธ์การวิเคราะห์ว่ามี Mail Address ใดบ้างที่สรุปว่าเป็น Member Address ของเมลลิงลิสต์

ตารางที่ 3.3 List Member Address Table

List address	Member address	CF_LM
Listname@Listdomain	Membername@Member domain	Value between 0 to 1

ประกอบด้วย field ดังต่อไปนี้

- List Address หมายถึง Mail Address ที่สรุปว่าเป็น List Address ของเมลลิงลิสต์
- Member Address หมายถึง Mail Address ที่สรุปว่าเป็นสมาชิกของเมลลิงลิสต์
- CF_LM หมายถึงค่าความมั่นใจว่า Mail Address ที่พบเป็นสมาชิกของเมลลิงลิสต์

3. List Manager Address Table (MA Table)

คือตารางแสดงผลลัพธ์การวิเคราะห์ว่ามี Mail Address ใดบ้างที่สรุปว่าเป็น Manager Address ของเมลลิงลิสต์

ตารางที่ 3.4 List Manager Address Table

List address	Manager address	CF_MA
Listname@Listdomain	Managername@Manager domain	Value between 0 to 1

ประกอบด้วย field ดังต่อไปนี้

- List Address หมายถึง Mail Address ที่สรุปว่าเป็น List Address ของเมลลิงลิสต์
- Manager Address หมายถึง Mail Address ที่สรุปว่าเป็น Manager Address ของเมลลิงลิสต์
- CF_MA หมายถึงค่าความมั่นใจว่า Mail Address ที่พบเป็นของผู้จัดการเมลลิงลิสต์

4. List Manager Type Table (MT Table)

คือตารางแสดงผลลัพธ์การวิเคราะห์ว่าเมลลิงลิสต์ใช้ Mailing List Software ชนิดใดในการจัดการเมลลิงลิสต์

ตารางที่ 3.5 List Manager Type Table

Manager address	Manager Type	CF_MT
Managername@Managerdo main	Manual, Listserv, Majordomo, Listproc, Smartlist or none	Value between 0 to 1

ประกอบด้วย field ดังต่อไปนี้

- Manager Address หมายถึง Mail Address ที่สรุปว่าเป็น Manager Address ของเมลลิงลิสต์
- Manager Type หมายถึงชนิดของ Mailing List Software ที่ใช้จัดการเมลลิงลิสต์
- CF_MT หมายถึงค่าความมั่นใจว่า Manager Address ใช้ Mailing List Software ชนิดนี้ในการจัดการเมลลิงลิสต์

โดยค่าความมั่นใจกำหนดให้มีค่าอยู่ระหว่าง 0 ถึง 1นั่นคือ ถ้าค่าความมั่นใจมีค่าเท่ากับ 0 หมายถึงข้อมูลนี้ไม่ใช่ข้อมูลของจดหมายอิเล็กทรอนิกส์ที่เกิดจากการทำงานของเมลลิงลิสต์แน่นอน และถ้าค่าความมั่นใจมีค่าเท่ากับ 1 หมายถึงข้อมูลนี้เป็นข้อมูลของจดหมายอิเล็กทรอนิกส์ที่เกิดจากการทำงานของเมลลิงลิสต์แน่นอน

นอกจาก Detection output database (DODB) แล้วยังมี Consequent mail database (CMDB) เพื่อเก็บ temporary file ซึ่งต้องเก็บข้อมูลไว้ในกรณีที่รอข้อมูลอื่นเพื่อมาสนับสนุนยืนยันว่ามีลักษณะเฉพาะของจดหมายอิเล็กทรอนิกส์จากกลุ่มเมลลิงลิสต์จริง หรือรอข้อมูลที่ขัดแย้งกับข้อมูลที่สรุปไว้ เพื่อนำมาปรับค่าความมั่นใจให้มีความถูกต้องก่อนการนำไปใช้งาน

3.3.2 Consequent Mail Database (CMDB)

คือฐานข้อมูลสำหรับพักข้อมูล Mail Detection ที่มีผลต่อเนื่องกันหรือ Mail ที่ยังไม่แน่ใจและยังสรุปไม่ได้ต้องรอข้อมูลอื่นมาสนับสนุน ประกอบด้วย

1. Suspected List Address Table (SLA Table)

คือตารางเก็บข้อมูลของ Mail Address ที่ยังไม่แน่ใจหรือยังวิเคราะห์ไม่ได้แต่เป็นไปได้ว่า Mail Address นี้เป็น List Address ของเมลลิงลิสต์

ตารางที่ 3.6 Suspected List Address Table

Message-id	List address	Time stamp	CF_SLA
rfc822_message _id	Listname@listdo main	Arrive time	Value between 0-1

ประกอบด้วย field ดังต่อไปนี้

- Message-id หมายถึง unique code ของจดหมายอิเล็กทรอนิกส์แต่ละฉบับ

- List Address หมายถึง Mail Address ที่เป็นไปได้ว่าจะเป็น List Address ของเมลลิงลิสต์
- CF_SLA หมายถึงค่าความมั่นใจว่า Mail Address ที่พบเป็น List Address ของเมลลิงลิสต์
- Time Stamp หมายถึงเวลาที่เริ่มต้นเก็บข้อมูลนี้

2. Suspected List Member Address Table (SLM Table)

คือตารางเก็บข้อมูลของ Mail Address ที่ยังไม่แน่ใจหรือยังวิเคราะห์ไม่ได้แต่เป็นไปได้ว่า Mail Address นี้เป็น Member Address ของเมลลิงลิสต์

ตารางที่ 3.7 Suspected List Member Address Table

Message-id	Member address	Time stamp	CF_SLM
rfc822_messag e_id	Membername@me mberdomain	Arrive time	Value between 0-1

ประกอบด้วย field ดังต่อไปนี้

- Message-id หมายถึง unique code ของจดหมายอิเล็กทรอนิกส์แต่ละฉบับ
- Member Address หมายถึง Mail Address ที่มีความเป็นไปได้ว่าเป็น Member Address ของเมลลิงลิสต์
- Time Stamp หมายถึงเวลาที่เริ่มต้นเก็บข้อมูลนี้
- CF_SLM หมายถึงค่าความมั่นใจว่า Mail Address ที่พบเป็น Member Address ของเมลลิงลิสต์

3. Suspected List Manager Address (SMA Table)

คือตารางเก็บข้อมูลของ Mail Address ที่ยังไม่แน่ใจหรือยังวิเคราะห์ไม่ได้แต่เป็นไปได้ว่า Mail Address นี้เป็น Manager Address ของเมลลิงลิสต์

ตารางที่ 3.8 Suspected List Manager Address Table

Message-id	Manager address	Status	Time stamp	CF_SMA
rfc822_message_ id	Managername@ managerdomain	subscribe, unsubscribe, general, none	Arrive time	Value between 0-1

ประกอบด้วย field ดังต่อไปนี้

- Message-id หมายถึง unique code ของจดหมายอิเล็กทรอนิกส์แต่ละฉบับ
- Manager Address หมายถึง Mail Address ที่มีความเป็นไปได้ว่าเป็น Manager Address ของเมลลิงลิสต์

- Status หมายถึงสถานะของจดหมายอิเล็กทรอนิกส์ว่าได้จากการวิเคราะห์ข้อมูลของจดหมายประเภทใด แบ่งออกเป็น 4 สถานะด้วยกันคือ

  subscribe หมายถึงข้อมูลได้จากจดหมายอิเล็กทรอนิกส์ที่เกิดจากการสมัครเป็นสมาชิกของเมลลิงลิสต์

  unsubscribe หมายถึงข้อมูลได้จากจดหมายอิเล็กทรอนิกส์ที่เกิดจากการยกเลิกสมาชิกของเมลลิงลิสต์

  general หมายถึงข้อมูลได้จากจดหมายอิเล็กทรอนิกส์ที่เกิดจากการร้องขอให้ผู้จัดการเมลลิงลิสต์จัดการเกี่ยวกับการแก้ไขข้อมูล หรือ ต้องการทราบข้อมูลเกี่ยวกับเมลลิงลิสต์

  none หมายถึงข้อมูลได้จากจดหมายอิเล็กทรอนิกส์ที่เกิดจากการกระจายจดหมายของเมลลิงลิสต์

- Time Stamp หมายถึงเวลาที่เริ่มต้นเก็บข้อมูลนี้
- CF_SMA หมายถึงค่าความมั่นใจว่า Mail Address ที่พบเป็น Manager Address ของเมลลิงลิสต์

4. Suspected List Manager Type Table (SMT Table)

   คือตารางเก็บข้อมูลชนิดของ Mailing List Software ที่คิดว่าน่าจะใช้ในการจัดการเมลลิงลิสต์

ตารางที่ 3.9 Suspected List Manager Type Table

Message-id	Manager address	Manager type	Time stamp	CF_SMT
rfc822_message_id	Managername@managerdomain	Manual, listserv, majordomo, listproc, smartlist, none	Arrive time	Value between 0-1

ประกอบด้วย field ดังต่อไปนี้

- Message-id หมายถึง unique code ของจดหมายอิเล็กทรอนิกส์แต่ละฉบับ
- Manager Address หมายถึง Mail Address ที่มีความเป็นไปได้ว่าเป็น Manager Address ของเมลลิงลิสต์
- Manager Type หมายถึงชนิดของ Mailing List Software ที่มีความเป็นไปได้ว่าใช้จัดการเมลลิงลิสต์
- Time Stamp หมายถึงเวลาที่เริ่มต้นเก็บข้อมูลนี้
- CF_SMT หมายถึงค่าความมั่นใจว่าเมลลิงลิสต์นี้ใช้ Mailing List Software ชนิดนี้ในการช่วยจัดการระบบ

3.3.3 Detection Configuration Database (DCDB)

คือฐานข้อมูลแสดงลักษณะและรายละเอียดต่างๆ เพื่อใช้ประกอบการวิเคราะห์หาลักษณะเฉพาะของจดหมายอิเล็กทรอนิกส์ที่เกิดจากการส่ง request ไปยังผู้จัดการเมลลิงลิสต์ประกอบด้วย

1. List Manager Command Table

คือตารางที่รวบรวมข้อมูลรายละเอียดของ Mailnig List Software ชนิดต่างๆ

ตารางที่ 3.10 List Manager Command Table

Manager Type	Command part flag	Subscribe format ID list <keyword,pattern>	Unsubscribe format ID list <keyword,pattern>	General command format ID list <keyword,pattern>
Majordomo	0	(1,2),(1,3)	(1,2), (1,3), (1,4), (1,5)	(1,1), (2,1), (3,3), (5,3), (4,2)
Listserv	0	(4,4)	(5,4),(4,4)	(9,1),(10,4),(10,3),(11,1)
Listproc	0	(1,4),(2,4)	(1,4),(2,4)	(1,1),(2,1),(6,3),(7,3),(4,1)
Smartlist	1	(1,1),(2,1),(3,1)	(1,1)	(1,1),(5,1),(3,1),(2,1)
Mailbase	0	(2,5)	(3,4)	(12,1),(8,3)
Ecartis	1	(1,2),(1,1)	(1,2),(1,1)	(2,1),(5,3)
Lyris	0	None	None	(1,1),(3,1)
Minimalist	1	(1,2),(1,3)	(1,4),(1,5)	(4,2),(3,3),(5,3),(1,1)
Mailserv	0	(1,2),(1,3)	(1,4),(1,5)	(9,1),(10,1),(11,1)

ประกอบด้วย field ดังต่อไปนี้

- Manager Type หมายถึงชนิดของ Mailing List Software
- Command Part Flag หมายถึง Location หรือ สถานที่ที่พบ Mailing List Command โดย bit 0 หมายความว่าพบคำสั่งที่ Mail Body และ bit 1 หมายความว่าพบคำสั่งที่ Subject
- Subscribe Format ID List หมายถึงรูปแบบ หรือ Format การใช้งานคำสั่งกลุ่ม Subscribe เพื่อสมัครเป็นสมาชิกของเมลลิงลิสต์ เก็บข้อมูลอยู่ในรูปแบบคือ <keyword,pattern> โดยที่ keyword หมายถึง Mailing List Command ส่วน pattern หมายถึงรูปแบบการใช้งานคำสั่งนั้นๆ
- Unsubscribe Format ID List หมายถึงรูปแบบ หรือ Format การใช้งานคำสั่งกลุ่ม Unsubscribe เพื่อยกเลิกการเป็นสมาชิกของเมลลิงลิสต์ ลักษณะการเก็บข้อมูลเหมือนกันกับ Subscribe Format ID List
- General Format ID List หมายถึงรูปแบบ หรือ Format การใช้งานคำสั่งกลุ่ม General เพื่อร้องขอรายละเอียดเกี่ยวกับเมลลิงลิสต์จากผู้จัดการเมลลิงลิสต์ ลักษณะการเก็บข้อมูลเหมือนกันกับ Subscribe Format ID List

โดยใน Subscribe Format ID List, Unsubscribe Format ID List และ General Format ID List ประกอบด้วยตารางย่อยดังต่อไปนี้

2. Subscribe Keyword Table

คือตารางแสดง Mailing List Command ที่อยู่ในกลุ่มของการสมัครสมาชิก

ตารางที่ 3.11 Subscribe Keyword Table

Subscribe keyword ID	Subscribe keyword
1	Subscribe
2	Join
3	Sign on
4	SUBSCRIBE

3. Subscribe Pattern Table

คือตารางแสดง Pattern ของ Keyword ที่อยู่ในกลุ่มของการสมัครสมาชิก

ตารางที่ 3.12 Subscribe Pattern Table

Subscribe pattern ID	Subscribe pattern
1	<keyword>
2	<keyword> listname
3	<keyword> listname address
4	<keyword> listname fullname
5	<keyword> listname firstname lastname

4. Unsubscribe Keyword Table

คือตารางแสดง Mailing List Command ที่อยู่ในกลุ่มของการยกเลิกสมาชิก

ตารางที่ 3.13 Unsubscribe Keyword Table

Unsubscribe keyword ID	Unsubscribe keyword
1	Unsubscribe
2	Sign off
3	Leave
4	UNSUBSCRIBE
5	SIGN OFF

5. Unsubscribe Pattern Table

คือตารางแสดง Pattern ของ Keyword ที่อยู่ในกลุ่มของการยกเลิกสมาชิก

ตารางที่ 3.14 Unsubscribe Pattern Table

Unsubscribe pattern ID	Unsubscribe pattern
1	<keyword>
2	<keyword> *
3	<keyword> * address
4	<keyword> listname
5	<keyword> listname address

6. General Keyword Table

คือตารางแสดง Mailing List Command ที่อยู่ในกลุ่มของการ request เพื่อขอข้อมูลทั่วๆไปเกี่ยวกับเมลลิงลิสต์

ตารางที่ 3.15 General Keyword Table

General command keyword ID	General command keyword
1	Help
2	Lists
3	Info
4	Which
5	Who
6	Information
7	Recipients
8	Review
9	HELP
10	INFO
11	LISTS
12	List me

7. General Pattern Table

คือตารางแสดง Pattern ของ Keyword ที่อยู่ในกลุ่มของการ request เพื่อขอข้อมูลทั่วๆไปเกี่ยวกับเมลลิงลิสต์

ตารางที่ 3.16 General Pattern Table

General command pattern ID	General command pattern
1	<keyword>
2	<keyword> address
3	<keyword> listname
4	<keyword> topic

3.3.4 ขั้นตอนการวิเคราะห์จดหมายอิเล็กทรอนิกส์จากเมลลิงลิสต์

สำหรับขั้นตอนการวิเคราะห์เพื่อหาลักษณะเฉพาะของจดหมายอิเล็กทรอนิกส์จากเมลลิงลิสต์สามารถสรุปเป็น Algorithm ได้ดังต่อไปนี้

```
Detection Algorithm
1. if (rfc822_from or rfc822_to or rfc822_sender or rfc822_return_path or rfc822_reply_to or
rfc822_error_to) equal to Manager type in DCDB then
        type flag equal to 1
2. if (rfc822_subject or rfc822_mail_body) equal to (subscribe keyword or unsubscribe keyword
or general command keyword in DCDB) then
        command flag equal to 1
3. if command flag equal to 1 then /*check command format*/
if  command format equal to (subscribe pattern or unsubscribe pattern or general pattern
in DCDB) then
                cmd_format flag equal to 1
4. if smtp_mail_from not equal to rfc822_from then
        sender flag equal to 1
5. for (i = 1 to M) /*M is number of rfc822_to and rfc822_cc*/
        for (j = 1 to N) /*N is number of smtp_rcpt_to*/
                if smtp_rcpt_tot[j] = (rfc822_to or rfc822_cc[i]) then
                        X = X+1
   if (X = 0) or (X < N) then
        recipient flag equal to 1 /*each smtp_rcpt_to not match rfc822_to or rfc822_cc*/
```

รูปที่ 3.5 Detection Algorithm

Detection Algorithm วิเคราะห์จดหมายอิเล็กทรอนิกส์เพื่อค้นหาลักษณะของเมลลิงลิสต์ซึ่งสามารถสรุปได้คือ

1. ค้นหา Manager Type ใน From: header, To: header, Sender: header, Return-Path: header, Reply-To: header และ Error-To: header
2. ค้นหา Subscribe keyword, Unsubscribe keyword และ General command keyword ใน Subject และ Body ของจดหมายอิเล็กทรอนิกส์
3. จากนั้นตรวจสอบว่า keyword เหล่านั้นใช้งานถูกต้องตาม pattern หรือไม่
4. ตรวจสอบว่าผู้ส่งใน smtp_mail_from ตรงกันกับผู้ส่งใน rfc822_from หรือไม่
5. ตรวจสอบว่าผู้รับใน smtp_rcpt_to พบใน rfc822_to และ rfc822_cc หรือไม่

เมื่อได้ข้อมูลจากการตรวจสอบแล้ว จึงนำข้อมูลเหล่านั้นมาพิจารณาเพื่อเข้าสู่กระบวนการจัดเก็บข้อมูลที่ CMDB ในกรณีที่ข้อมูลจดหมายอิเล็กทรอนิกส์ฉบับนั้นมีคุณสมบัติของเมลลิงลิสต์ดัง algorithm ต่อไปนี้

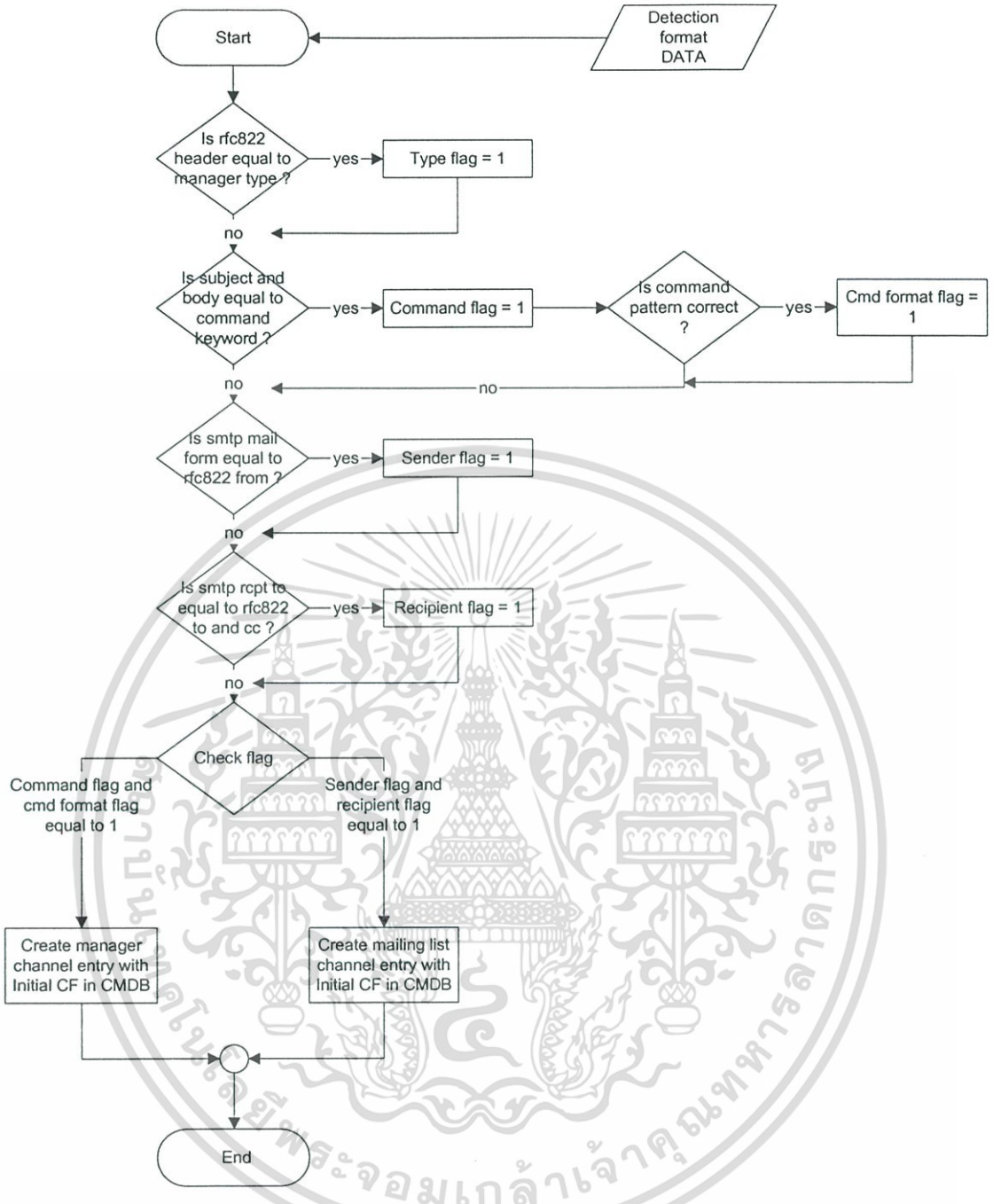
Create Algorithm Segment information in CMDB

1. if command flag and cmd_format flag equal to 1 then
   - Message-id equal to rfc822_message_id
   - List address equal to listname + @ + smtp_destination_domain
   - Time stamp equal to smtp_arrive_time
   - Member address equal to rfc822_from
   - Manager address equal to rfc822_to
   - Manager type equal to Type flag
   - Status equal to group of type flag
   - CF_SLA, CF_SLM, CF_SMA and CF_SMT equal to initial CF or CF_{IN}
2. if sender flag and recipient flag equal to 1 then
   - Message-id equal to rfc822_message_id
   - List address equal to rfc822_to
   - Time stamp equal to smtp_arrive_time
   - Member address equal to smtp_rcpt_to
   - Manager address equal to smtp_mail_from
   - Manager type equal to Type flag
   - Status equal to none
   - CF_SLA, CF_SLM, CF_SMA and CF_SMT equal to initial CF or CF_{IN}

รูปที่ 3.6 Create Algorithm

การจัดเก็บข้อมูลจดหมายอิเล็กทรอนิกส์ที่มีคุณสมบัติของเมลลิงลิสต์ใน CMDB จะกำหนดค่าความมั่นใจเริ่มต้น หรือ Initial Confidential Factor (CF_{IN}) ซึ่งได้จากการทดลองสุ่มค่าเริ่มต้นหลายๆค่าแล้วเลือกค่าความมั่นใจเริ่มต้นที่เหมาะสมที่สุด นอกจากนี้ยังเป็นการจัดเก็บเพื่อรอข้อมูลอื่นที่อาจจะสนับสนุนหรือขัดแย้งกับข้อมูลเดิม ซึ่งต่อไปจะเป็นรูปแสดงภาพรวมของ Algorithm รวมทั้งหมดของขั้นตอนการตรวจจับ

จะเริ่มต้นจากการรับข้อมูลจากกระบวนการจัดเตรียมข้อมูล ซึ่งอยู่ในรูปแบบที่กำหนดคือ Detection Format เพื่อเข้าสู่ขั้นตอนการตรวจจับเมลลิงลิสต์คือ Detection algorithm เพื่อคัดเลือกจดหมายที่มีข้อมูลตามข้อสันนิษฐานจากนั้นจะได้ ข้อมูลคือ Detected Mail เพื่อผ่าน Create algorithm โดยใน algorithm นี้จะคัดเลือกข้อมูลที่มีคุณสมบัติของเมลลิงลิสต์เข้าไปเก็บไว้ใน CMDB เพื่อพักข้อมูลและรอข้อมูลอื่นที่อาจจะมาสนับสนุนหรือขัดแย้ง โดยจะกำหนดค่าความมั่นใจเริ่มต้น หรือ Initial Confidential Factor ให้กับข้อมูล ซึ่งสามารถแบ่งประเภทของข้อมูลออกเป็น 2 ส่วนคือ ข้อมูลที่เกิดจากการติดต่อระหว่างสมาชิกและผู้จัดการเมลลิงลิสต์ และข้อมูลที่เกิดจากการกระจายจดหมายของเมลลิงลิสต์



รูปที่ 3.7 ภาพรวมของ Detection Process

ต่อไปเป็นขั้นตอนสุดท้ายของระบบการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์

3.4 Postprocessing Process

คือกระบวนการสุดท้าย มีหน้าที่ตรวจสอบข้อมูลที่สรุปได้จากกระบวนการตรวจจับใน Consequent Mail Database (CMDB) เพื่อปรับเปลี่ยน Confidential Factor หรือ ค่าความมั่นใจ ใน

กรณีที่มีข้อมูลที่สนับสนุนข้อมูลเดิม หรือในกรณีที่มีข้อมูลที่แสดงความขัดแย้งกับข้อมูลเดิม เพื่อให้ค่าความมั่นใจที่สรุปได้มีความถูกต้องมากที่สุด นอกจากนี้ยังเป็นกระบวนการคัดเลือกข้อมูลจาก CMDB เพื่อนำไปเก็บไว้ที่ Detection Output Database (DODB) ดัง Algorithm ดังต่อไปนี้

Postprocess Algorithm in CMDB

```
1. if command flag and cmd_format flag equal to 1 then
        if listname + @ + smtp_destination_domain equal to List address then
                increase CF_SLA
        else    decrease CF_SLA
        if rfc822_from equal to Member address then
                increase CF_SLM
        else    decrease CF_SLM
        if rfc822_to equal to Manager address then
                increase CF_SMA
                if Type flag equal to Manager type then
                        increase CF_SMT
                else    decrease CF_SMT
        else    decrease CF_SMA

2. if sender flag and recipient flag equal to 1 then
        if rfc822_to equal to List address then
                increase CF_SLA
        else    decrease CF_SLA
        if smtp_rcpt_to equal to Member address then
                increase CF_SLM
        else    decrease CF_SLM
        if smtp_mail_from equal to Manager address then
                increase CF_SMA
                if Type flag equal to Manager type then
                        increase CF_SMT
                else    decrease CF_SMT
        else    decrease CF_SMA

3. if command flag, cmd_format, sender flag and recipient flag equal to 0
        if (rfc822_to or rfc822_from) equal to List address
                decrease CF_SLA
        if (rfc822_to or rfc822_from) equal to Manager address
                decrease CF_SMA
```

รูปที่ 3.8 Postprocess Algorithm

ใน Postprocess algorithm คือกระบวนการปรับเปลี่ยนค่าความมั่นใจ เมื่อมีข้อมูลที่สรุปได้ตรงกับข้อมูลเดิมหรือมาสนับสนุนข้อมูลที่มีอยู่แล้ว ระบบจะเพิ่มค่าความมั่นใจ และเมื่อมีข้อมูลที่สรุปขัดแย้งกับข้อมูลเดิม ระบบจะลดค่าความมั่นใจลง เป็นไปตาม algorithm ดังต่อไปนี้

Confidential Factor Algorithm *in CMDB*

if increase CF
$CF_{NEW} = CF_{OLD} + positive(CF_{OLD})$
if decrease CF
$CF_{NEW} = CF_{OLD} - negative(CF_{OLD})$

Positive(x)
Return((1-x)/2)
Negative(x)
Return(x/2)

รูปที่ 3.9 Confidential Factor Algorithm

นอกจากนี้ใน Detection algorithm นั้นพบว่าเมื่อตรวจจับจดหมายอิเล็กทรอนิกส์ที่เกิดจากการติดต่อสื่อสารระหว่างสมาชิกและผู้จัดการเมลลิงลิสต์ จะมีบางคำสั่งใน General command keyword ซึ่งเมื่อตรวจสอบพบจะไม่สามารถระบุชื่อของเมลลิงลิสต์ที่ Submailing list ต้องการได้ ระบบจึงกำหนดให้ List address ของเมลลิงลิสต์นั้นๆเป็น Listname + @ + destination domain แทนเพื่อรอข้อมูลของจดหมายอิเล็กทรอนิกส์ฉบับอื่นๆ มาสนับสนุนเพื่อค้นหาชื่อของเมลลิงลิสต์นั้นๆ ดัง Listname algorithm ในรูปที่ 3.10

Listname Algorithm *in CMDB*
if sender flag and recipient flag equal to 1 then
if List address equal to listname + @ + list domain then
if smtp_destination_domain *mail* equal to list domain
if smtp_rcpt_to *mail* equal to Member address
if smtp_source_domain *mail* equal to manager domain then
List address equal to rfc822_to

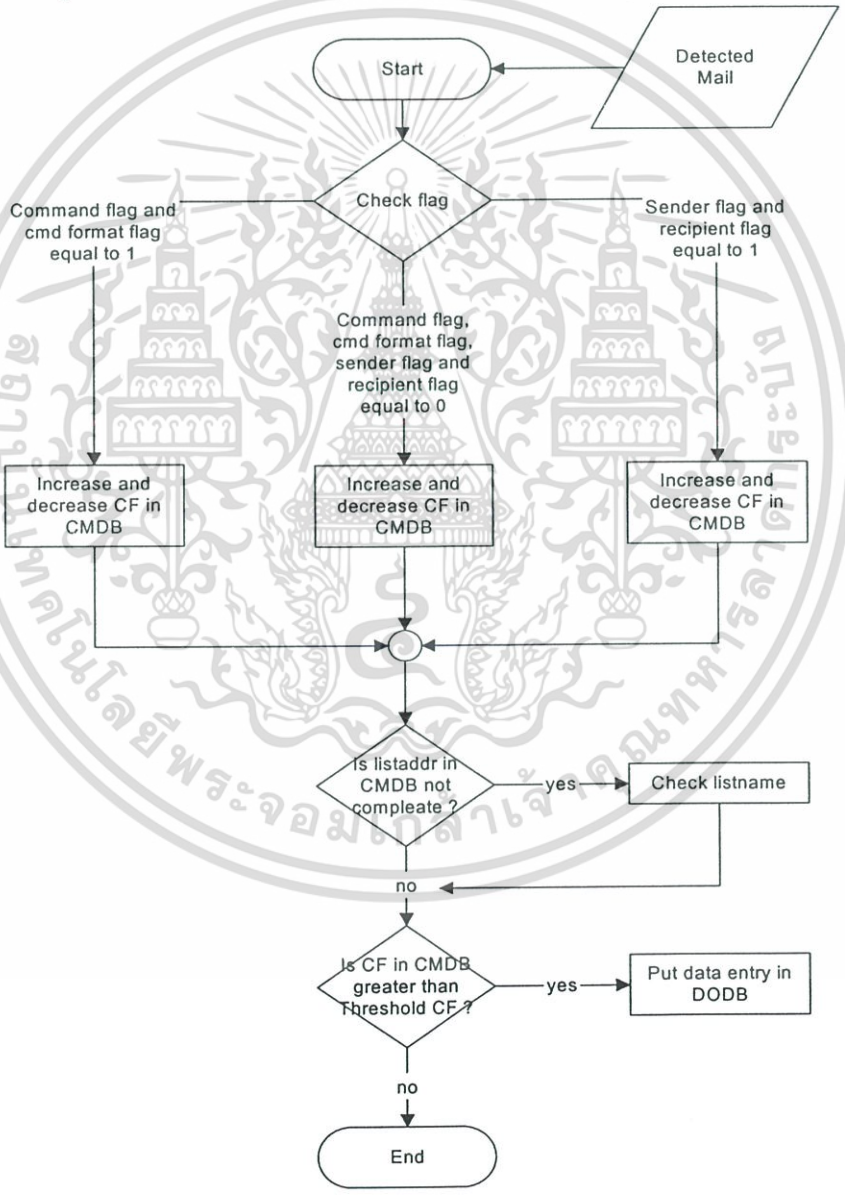
รูปที่ 3.10 Listname Algorithm

ต่อไปเป็น algorithm แสดงวิธีการคัดเลือกข้อมูลเพื่อนำไปเก็บใน DODB เพื่อนำข้อมูลที่ได้ไปจัดการ config ระบบ Submailing list สำหรับค่า Threshold Confidential Factor เป็นเกณฑ์ค่าความมั่นใจที่แน่ใจว่าถ้า ข้อมูลจดหมายอิเล็กทรอนิกส์ฉบับใดมีค่าความมั่นใจมากกว่า CF_{TH} นี้แล้ว จดหมายอิเล็กทรอนิกส์ฉบับนั้นเกิดจากการทำงานของเมลลิงลิสต์ โดยค่าความมั่นใจนี้ได้จากการทดลอง เพื่อหาค่าที่มีความเหมาะสมและถูกต้องมากที่สุด

Threshold Algorithm
if CF_SLA greater than CFT_LA and CF_SLM greater than CFT_LM and CF_SMA greater than CFT_MA and CF_SMT greater than CFT_MT then
copy List address, Member address, Manager address, Message-id and Manager type
CF_LA equal to CF_SLA
CF_LM equal to CF_SLM
CF_MA equal to CF_SMA
CF_MT equal to CF_SMT

รูปที่ 3.11 Threshold Algorithm

ซึ่งต่อไปจะเป็นรูปแสดงภาพรวมของ Algorithm รวมของขั้นตอนสุดท้ายในการตรวจสอบข้อมูล



รูปที่ 3.12 ภาพรวมของ Algorithm ใน Postprocessing Process

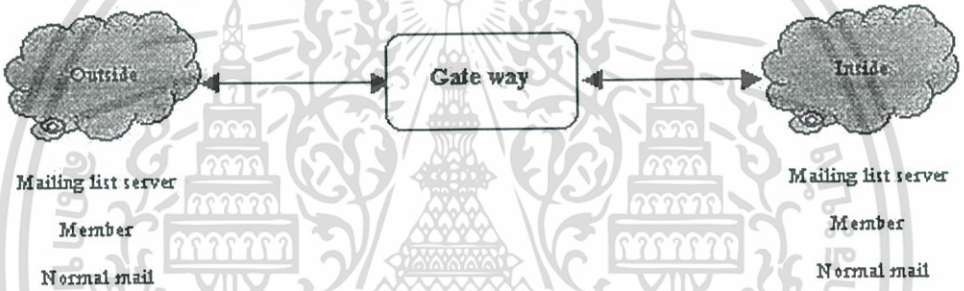
สำหรับภาพรวมของ algorithm ในขั้นตอนสุดท้ายนี้ เริ่มต้นจากการนำข้อมูล Detected Mail ที่ได้จากขั้นตอน Detection Process มาตรวจสอบกับข้อมูลเดิมที่สรุปไว้ใน CMDB เพื่อตรวจสอบว่ามีข้อมูลใดบ้างที่สรุปตรงกันนั่นคือเป็นข้อมูลที่สนับสนุนข้อมูลเดิม ระบบจะเพิ่มค่าความมั่นใจ แต่ในทางตรงกันข้ามเมื่อระบบพบข้อมูลที่ขัดแย้งกับข้อมูลเดิม ระบบจะปรับลดค่าความมั่นใจลง ตามเงื่อนไขใน Confidential Factor algorithm จากนั้นเมื่อผ่านข้อมูลเข้าสู่ Listname algorithm ระบบจะตรวจสอบว่ามีข้อมูลใดใน CMDB ที่มี List address ที่ยังไม่สมบูรณ์และมีข้อมูลอื่นๆสรุปตรงกับข้อมูลเมลลิงลิสต์ที่เข้ามาใหม่ ระบบจะจัดการเปลี่ยนข้อมูล List address ให้ถูกต้อง และใน algorithm สุดท้ายคือขั้นตอนการเปรียบเทียบค่าความมั่นใจกับเกณฑ์ค่าความมั่นใจที่ตั้งไว้ หรือ Threshold Confidential Factor เมื่อค่าความมั่นใจมีค่าสูงกว่าเกณฑ์ที่กำหนด ระบบจะนำข้อมูลไปเก็บไว้ใน DODB แสดงผลลัพธ์ที่ได้ให้ Network Administrator ทราบ

ในบทถัดไปจะกล่าวเกี่ยวการออกแบบการทดลอง การกำหนดสภาพแวดล้อมในการทำการทดลอง รวมถึงผลการทดลองต่างๆ

บทที่ 4
การทดลองและการวิเคราะห์ผล

4.1 การออกแบบการทดลอง

การออกแบบการทดลองเพื่อหาค่า Initial Confidential Factor และ Threshold Confidential Factor ที่เหมาะสมนั้น ทำได้โดยการทดลองสร้างเมลลิงลิสต์ที่ใช้ได้งานจริง 50 เมลลิงลิสต์ กระจายอยู่ในเครือข่ายต่างๆที่จำลองขึ้นมา และมีการจัดการโดยใช้ Mailing List Manager Software จากนั้น ทำการทดลองเพื่อศึกษาผลกระทบของ Initial Confidential Factor, Threshold Confidential Factor และจำนวนจดหมายอิเล็กทรอนิกส์ที่มีผลต่อ %Correctness, %Positive Error และ %Negative Error โดยจำลองสถานการณ์ดังรูปที่ 4.1



รูปที่ 4.1 การจำลองสถานการณ์เพื่อทำการทดลอง

4.2 สภาพแวดล้อมของการทดลอง

ประกอบด้วย

- Computer จำนวน 5 เครื่อง CPU Pentium III 550 Mhz RAM ขนาด 128MB Hard Disk 10 GB
- ระบบปฏิบัติการคือ RedHat Linux 8.0
- ใช้ Sendmail-8.12.5-7 เป็น Mail Transport Agent
- ใช้ Majordomo-1.94.5 เป็น Software ในการจัดการเมลลิงลิสต์
- ใช้ Tethereal 0.9.6 ในการดักจับข้อมูลจดหมายอิเล็กทรอนิกส์

4.3 การทดลองและการวิเคราะห์ผล

สำหรับการทดลอง เริ่มจำลอง Outside Network ขึ้นมา 4 เครือข่ายและ Inside Network 1 เครือข่ายและกำหนดให้แต่ละเครือข่ายมี User เครือข่ายละ 15 คน จัดการสร้าง Mail server ขึ้นมาในแต่ละเครือข่าย จากนั้นสร้างเมลลิงลิสต์ขึ้นมาเครือข่ายละ 10 เมลลิงลิสต์ โดยกำหนดให้แต่ละเมลลิงลิสต์ของ Outside Network มีสมาชิกอยู่ใน Inside Network เมลลิงลิสต์ละ 10 E-mail address ส่วนเมลลิงลิสต์ใน Inside Network กำหนดให้มีสมาชิกอยู่ใน Outside Network เมลลิงลิสต์ละ 10 E-mail address เช่นกัน โดยกำหนดให้ Inside Network เป็นเครือข่ายที่มี Mailing List E-mail Detection System หรือ ระบบตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์อยู่

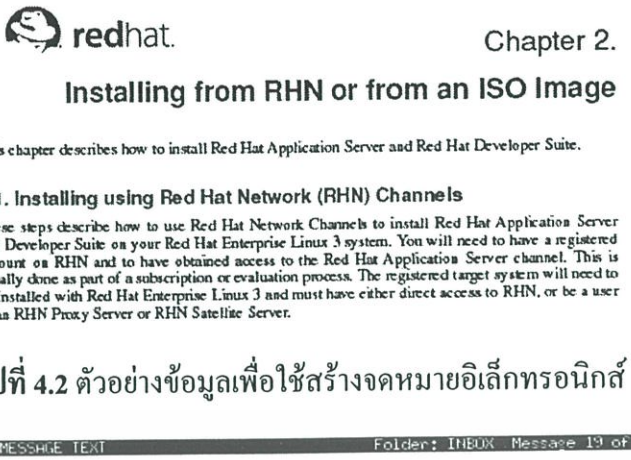
จากนั้นทำการทดลอง จดหมายที่ใช้ในการทดลองแบ่งออกเป็น 2 แบบคือ

- จดหมายอิเล็กทรอนิกส์จากเมลลิงลิสต์แบบแรกคือ จดหมายที่เกิดจากการติดต่อระหว่าง Mailing list manager และสมาชิก เช่น จดหมายที่ต้องการสมัครสมาชิก หรือยกเลิกสมาชิก เป็นต้น แบบที่ 2 คือจดหมายที่เกิดจากการติดต่อสื่อสารแลกเปลี่ยนข้อมูลและความคิดเห็นระหว่างสมาชิกด้วยกันผ่านทางเมลลิงลิสต์ โดยในการทดลองนั้นจะสร้างจดหมายจากเมลลิงลิสต์ทั้งสองแบบจำนวนเท่าๆกัน
- จดหมายอิเล็กทรอนิกส์แบบธรรมดาคือจดหมายที่ User ของแต่ละเครือข่ายส่งข้อความถึงกัน ซึ่งสามารถแบ่งตามจำนวนของผู้ส่ง ต่อ จำนวนของผู้รับที่พบอยู่ใน Rfc822 header คือ To: และ Cc: Header แบ่งได้ 2 แบบคือ การส่งแบบ 1:1 คือผู้ส่ง 1 คนต่อผู้รับ 1 คนและการส่งแบบ 1:n คือผู้ส่ง 1 คนต่อผู้รับหลายคน โดยในการทดลองนั้นจะสร้างจดหมายจากเมลลิงลิสต์ทั้งสองแบบจำนวนเท่าๆกัน นอกจากนั้นจำลองจดหมายที่มีการใช้งาน Bcc: Header ร่วมด้วย

โดยแบ่งการทดลองออกเป็น 3 กรณี คือ

1. กรณีที่จำนวนจดหมายอิเล็กทรอนิกส์จากเมลลิงลิสต์ ใกล้เคียงกันกับจดหมายอิเล็กทรอนิกส์แบบธรรมดา มีสัดส่วนโดยประมาณคือ 50 : 50
2. กรณีที่จำนวนจดหมายอิเล็กทรอนิกส์จากเมลลิงลิสต์ มากกว่าจดหมายอิเล็กทรอนิกส์แบบธรรมดา มีสัดส่วนโดยประมาณคือ 60 : 40
3. กรณีที่จำนวนจดหมายอิเล็กทรอนิกส์จากเมลลิงลิสต์ น้อยกว่าจดหมายอิเล็กทรอนิกส์แบบธรรมดา มีสัดส่วนโดยประมาณคือ 40 : 60

สำหรับข้อมูลจดหมายอิเล็กทรอนิกส์ที่ใช้ทำการทดลองนั้น จำลองขึ้นโดยการคัดลอกข้อมูลจาก Redhat Document [7] มีลักษณะเป็นข้อความเพียงอย่างเดียว ซึ่งอาจมีข้อความบางคำตรงกับ Mailing list command keyword เช่น who, lists และ information เป็นต้น ดังตัวอย่างข้อมูลในรูป 4.2



redhat.

Chapter 2.

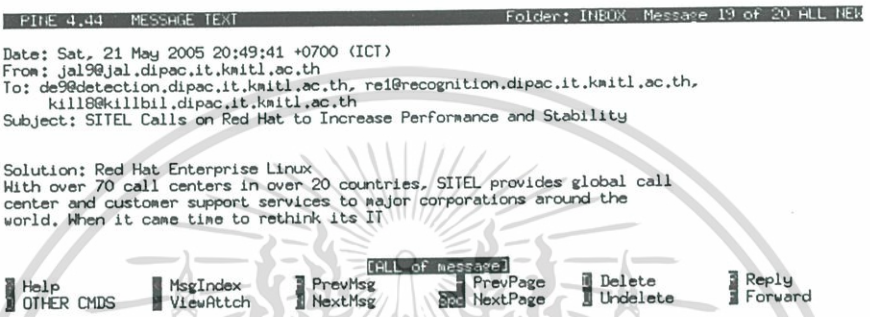
Installing from RHN or from an ISO Image

This chapter describes how to install Red Hat Application Server and Red Hat Developer Suite.

2.1. Installing using Red Hat Network (RHN) Channels

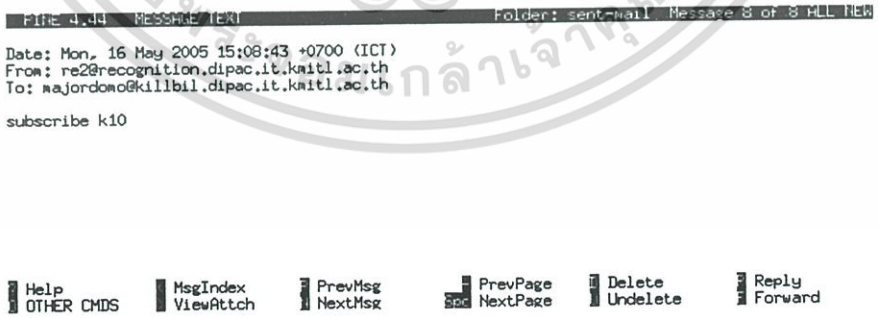
These steps describe how to use Red Hat Network Channels to install Red Hat Application Server and Developer Suite on your Red Hat Enterprise Linux 3 system. You will need to have a registered account on RHN and to have obtained access to the Red Hat Application Server channel. This is usually done as part of a subscription or evaluation process. The registered target system will need to be installed with Red Hat Enterprise Linux 3 and must have either direct access to RHN, or be a user of an RHN Proxy Server or RHN Satellite Server.

รูปที่ 4.2 ตัวอย่างข้อมูลเพื่อใช้สร้างจดหมายอิเล็กทรอนิกส์



รูปที่ 4.3 ตัวอย่างจดหมายอิเล็กทรอนิกส์แบบธรรมดา

รูปที่ 4.4 ตัวอย่างจดหมายอิเล็กทรอนิกส์เกิดจากการกระจายของเมลลิงลิสต์



รูปที่ 4.5 ตัวอย่างจดหมายอิเล็กทรอนิกส์เกิดจากการสมัครสมาชิก

เมื่อได้ข้อมูลที่ใช้สร้างจดหมายอิเล็กทรอนิกส์แล้ว จึงใช้โปรแกรม Pine ซึ่งเป็น MUA ชนิดหนึ่งในการส่งจดหมายอิเล็กทรอนิกส์ไปยังเมลลิงลิสต์ หรือ user คนอื่นๆที่อยู่ในเครือข่าย

สำหรับตัวแปรที่เกี่ยวข้องในการทดลองมีดังต่อไปนี้

ตารางที่ 4.1 ตัวแปรที่เกี่ยวข้อง

parameter	description
Initial Confidential Factor (CF_{IN})	คือค่าความมั่นใจเริ่มต้นที่กำหนดให้กับข้อมูลที่ผ่านการตรวจจับพบลักษณะและคุณสมบัติของเมลลิงลิสต์
Threshold Confidential Factor (CF_{TH})	คือค่าความมั่นใจที่ใช้เป็นเกณฑ์การสรุปข้อมูล Mail Detection ว่าเป็นข้อมูลของเมลลิงลิสต์
Exact ML	คือจำนวนเมลลิงลิสต์ทั้งหมดที่สร้างขึ้นเพื่อทำการทดลอง
Detect ML	คือจำนวนเมลลิงลิสต์ที่ระบบตรวจจับได้ ประกอบด้วย List Address, Member Address, Manager Address และ Manager Type รวมถึงค่าความมั่นใจ CF_LA, CF_LM, CF_MA และ CF_MT
Correct ML	คือจำนวนเมลลิงลิสต์ที่ระบบตรวจจับได้ถูกต้อง โดยวัดค่าความถูกต้องจากข้อมูล List Address, Member Address, Manager Address และ Manager Type ที่ตรวจจับได้ต้องถูกต้องตรงกับข้อมูลเมลลิงลิสต์ที่สร้างขึ้น และค่าความมั่นใจ CF_LA, CF_LM, CF_MA และ CF_MT จะต้องมีค่าสูงกว่า CF_{TH}
Positive ML	คือจำนวนเมลลิงลิสต์ที่ระบบตรวจจับไม่พบ สามารถคำนวณได้ดังสมการ (4-1)
Negative ML	คือจำนวนเมลลิงลิสต์ที่ระบบตรวจจับผิดพลาด โดยความผิดพลาดวัดจากข้อมูลเมลลิงลิสต์ที่สรุปมาไม่ตรงกับเมลลิงลิสต์ที่จัดสร้างขึ้น สามารถคำนวณได้ดังสมการ (4-2)
%Correctness	คือเปอร์เซ็นต์ค่าความถูกต้องของเมลลิงลิสต์ที่ตรวจจับได้ สามารถคำนวณได้ดังสมการ (4-3)
%Positive Error	คือเปอร์เซ็นต์ค่าความผิดพลาดของเมลลิงลิสต์ที่ตรวจจับไม่พบ สามารถคำนวณได้ดังสมการ (4-4)
%Negative Error	คือเปอร์เซ็นต์ค่าความผิดพลาดของเมลลิงลิสต์ที่ตรวจจับได้ สามารถคำนวณได้ดังสมการ (4-5)

สูตรที่ใช้ในการคำนวณ

$$PositiveML = ExactML - CorrectML \tag{4-1}$$

$$NegativeML = DetectML - CorrectML \tag{4-2}$$

$$Correctness = \frac{CorrectML}{DetectML} \times 100\% \tag{4-3}$$

$$PositiveError = \frac{PositiveML}{ExactML} \times 100\% \tag{4-4}$$

$$NegativeError = \frac{NegativeML}{DetectML} \times 100\% \quad (4\text{-}5)$$

4.3.1 การทดลองเพื่อศึกษาผลกระทบของ Initial Confidential Factor ที่มีผลต่อ %Correctness, %Positive Error และ %Negative Error

สำหรับการทดลองนั้นมีความต้องการหาค่า CF_{IN} และ CF_{TH} ที่เหมาะสม โดยการทดลองวัด %Correctness, %Positive Error และ %Negative Error ที่ค่า CF_{IN} และค่า CF_{TH} ต่างๆกัน และนำข้อมูลที่ได้มา สร้างกราฟแสดงความสัมพันธ์

ตารางที่ 4.2 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนใกล้เคียงกันกับจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.1$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	48	48	0	2	100	4	0
0.2	50	50	49	1	1	98	2	2
0.3	50	50	49	1	1	98	2	2
0.4	50	50	49	1	1	98	2	2
0.5	50	53	50	3	0	94.33962	0	5.660377
0.6	50	53	50	3	0	94.33962	0	5.660377
0.7	50	53	50	3	0	94.33962	0	5.660377
0.8	50	53	50	3	0	94.33962	0	5.660377
0.9	50	57	50	7	0	87.7193	0	12.2807

ตารางที่ 4.3 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนใกล้เคียงกันกับจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.2$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	45	45	0	5	100	10	0
0.2	50	45	45	0	5	100	10	0
0.3	50	47	47	0	3	100	6	0
0.4	50	48	47	1	3	97.91667	6	2.083333
0.5	50	48	47	1	3	97.91667	6	2.083333
0.6	50	48	47	1	3	97.91667	6	2.083333
0.7	50	49	48	1	2	97.95918	4	2.040816
0.8	50	49	48	1	2	97.95918	4	2.040816
0.9	50	52	50	2	0	96.15385	0	3.846154

ตารางที่ 4.4 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนใกล้เคียงกันกับจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.3$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	43	43	0	7	100	14	0
0.2	50	43	43	0	7	100	14	0
0.3	50	44	44	0	6	100	12	0
0.4	50	46	45	1	5	97.82609	10	2.173913
0.5	50	46	45	1	5	97.82609	10	2.173913
0.6	50	46	45	1	5	97.82609	10	2.173913
0.7	50	46	45	1	5	97.82609	10	2.173913
0.8	50	46	45	1	5	97.82609	10	2.173913
0.9	50	46	45	1	5	97.82609	10	2.173913

ตารางที่ 4.5 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนใกล้เคียงกันกับจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.4$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	39	39	0	11	100	22	0
0.2	50	39	39	0	11	100	22	0
0.3	50	43	43	0	7	100	14	0
0.4	50	43	43	0	7	100	14	0
0.5	50	45	44	1	6	97.77778	12	2.222222
0.6	50	45	44	1	6	97.77778	12	2.222222
0.7	50	46	45	1	5	97.82609	10	2.173913
0.8	50	46	45	1	5	97.82609	10	2.173913
0.9	50	46	45	1	5	97.82609	10	2.173913

ตารางที่ 4.6 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนใกล้เคียงกันกับจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.5$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	39	39	0	11	100	22	0
0.2	50	39	39	0	11	100	22	0
0.3	50	39	39	0	11	100	22	0
0.4	50	39	39	0	11	100	22	0
0.5	50	39	39	0	11	100	22	0
0.6	50	41	40	1	10	97.56098	20	2.439024
0.7	50	41	40	1	10	97.56098	20	2.439024
0.8	50	41	40	1	10	97.56098	20	2.439024
0.9	50	41	40	1	10	97.56098	20	2.439024

ตารางที่ 4.7 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนใกล้เคียงกันกับจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.6$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	36	36	0	14	100	28	0
0.2	50	36	36	0	14	100	28	0
0.3	50	36	36	0	14	100	28	0
0.4	50	36	36	0	14	100	28	0
0.5	50	37	37	0	13	100	26	0
0.6	50	37	37	0	13	100	26	0
0.7	50	39	38	1	12	97.4359	24	2.564103
0.8	50	39	38	1	12	97.4359	24	2.564103
0.9	50	40	39	1	11	97.5	22	2.5

ตารางที่ 4.8 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนใกล้เคียงกันกับจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.7$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	32	32	0	18	100	36	0
0.2	50	32	32	0	18	100	36	0
0.3	50	32	32	0	18	100	36	0
0.4	50	32	32	0	18	100	36	0
0.5	50	32	32	0	18	100	36	0
0.6	50	32	32	0	18	100	36	0
0.7	50	36	36	0	14	100	28	0
0.8	50	38	37	1	13	97.36842	26	2.631579
0.9	50	39	38	1	12	97.4359	24	2.564103

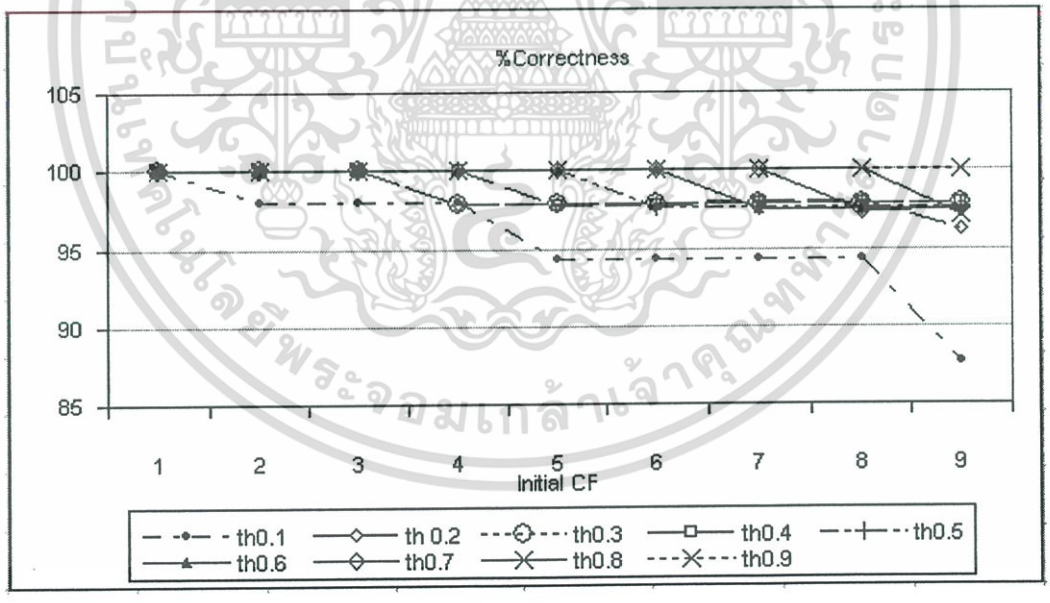
ตารางที่ 4.9 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนใกล้เคียงกันกับจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.8$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	22	22	0	28	100	56	0
0.2	50	22	22	0	28	100	56	0
0.3	50	27	27	0	23	100	46	0
0.4	50	27	27	0	23	100	46	0
0.5	50	31	31	0	19	100	38	0
0.6	50	31	31	0	19	100	38	0
0.7	50	31	31	0	19	100	38	0
0.8	50	31	31	0	19	100	38	0
0.9	50	34	33	1	17	97.05882	34	2.941176

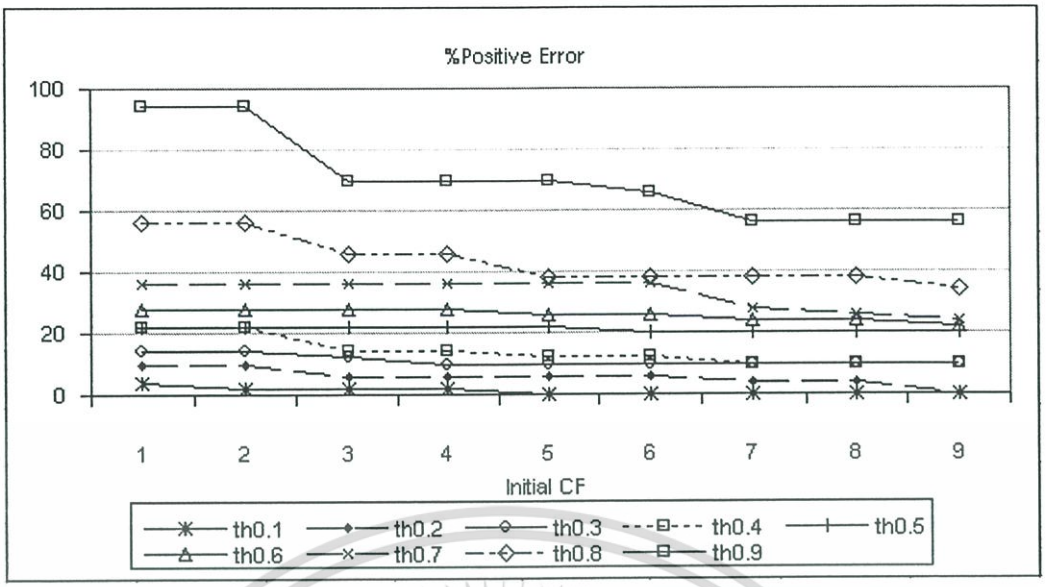
ตารางที่ 4.10 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนใกล้เคียงกันกับจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.9$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	3	3	0	47	100	94	0
0.2	50	3	3	0	47	100	94	0
0.3	50	15	15	0	35	100	70	0
0.4	50	15	15	0	35	100	70	0
0.5	50	15	15	0	35	100	70	0
0.6	50	17	17	0	33	100	66	0
0.7	50	22	22	0	28	100	56	0
0.8	50	22	22	0	28	100	56	0
0.9	50	22	22	0	28	100	56	0

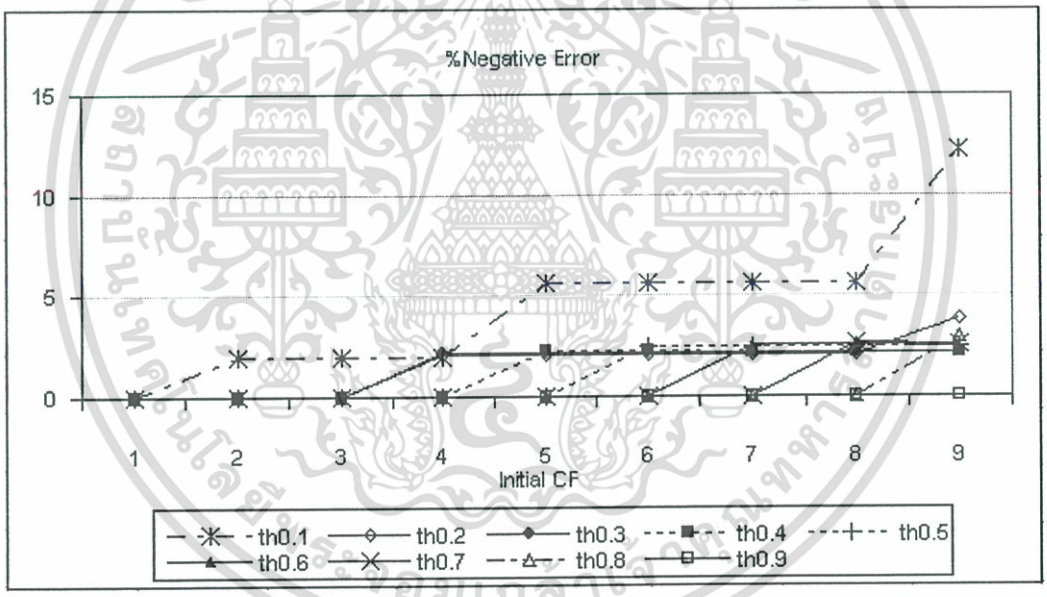
จากตารางที่ 4.2 ถึง ตารางที่ 4.10 คือตารางแสดงผลการทดลองการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ ในกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนใกล้เคียงกันกับจดหมายแบบธรรมดา โดยกำหนดให้ CF_{IN} มีค่าตั้งแต่ 0.1 – 0.9 ที่ CF_{TH} ค่าต่างๆกัน ซึ่งสามารถนำข้อมูลที่ได้มาสร้างกราฟแสดงความสัมพันธ์ระหว่าง %Correctness, %Positive, %Negative Error และ CF_{IN} ที่ CF_{TH} ค่าต่างๆเพื่อวิเคราะห์และเปรียบเทียบหาค่า CF_{IN} และ CF_{TH} ที่เหมาะสม



รูปที่ 4.6 กราฟแสดงความสัมพันธ์ระหว่าง %Correctness และ CF_{IN} ที่ CF_{TH} ค่าต่างๆ



รูปที่ 4.7 กราฟแสดงความสัมพันธ์ระหว่าง %Positive Error และ CF_{IN} ที่ CF_{TH} ค่าต่างๆ



รูปที่ 4.8 กราฟแสดงความสัมพันธ์ระหว่าง %Negative Error และ CF_{IN} ที่ CF_{TH} ค่าต่างๆ

<u>วิเคราะห์กราฟ %Correctness, %Positive Error, %Negative Error และ CF_{IN} เพื่อหาค่า CF_{IN} ที่เหมาะสม (ในกรณีที่จำนวนจดหมายจากเมลลิงลิสต์ใกล้เคียงกับจำนวนจดหมายแบบธรรมดา)</u>

จากกราฟที่ 4.6, 4.7 และ 4.8 พบว่าถ้ากำหนดให้ค่า CF_{IN} มีค่าต่ำ จะทำให้ค่า %Correctness หรือ ค่าความถูกต้องมีค่าสูงที่สุด แต่ถ้าเพิ่มค่า CF_{IN} ให้สูงขึ้นเรื่อยๆ กราฟของค่าความถูกต้องจะมีแนวโน้มลดลง เช่นเดียวกันกับแนวโน้มของกราฟ %Positive Error ที่มีค่าสูงเมื่อกำหนดค่า CF_{IN} ต่ำ

นั่นหมายถึงมีจำนวนเมลลิงลิสต์ที่ตรวจจับได้น้อยกว่าจำนวนเมลลิงลิสต์ที่มีอยู่จริง และเมื่อเพิ่มค่า CF_{IN} ให้สูงขึ้นเรื่อยๆ ค่า %Positive Error จะลดลง คือจำนวนของเมลลิงลิสต์ที่ตรวจจับได้มีจำนวนเพิ่มมากขึ้น

แต่ในทางตรงกันข้ามค่า %Negative Error จะมีค่าน้อยเมื่อกำหนดค่า CF_{IN} ต่ำ และจะมีแนวโน้มเพิ่มขึ้นเมื่อค่า CF_{IN} สูงขึ้นเรื่อยๆ นั่นหมายถึงมีจำนวนเมลลิงลิสต์ที่ตรวจจับผิดพลาดสูงขึ้นถ้ากำหนดให้ค่า CF_{IN} มีค่าสูง

ต่อไปเป็นการแสดงผลการทดลองการตรวจจับเมลลิงลิสต์ในกรณีที่ จำนวนของจดหมายจากเมลลิงลิสต์มากกว่าจำนวนจดหมายแบบธรรมดา

ตารางที่ 4.11 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนมากกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.1$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	50	50	0	0	100	0	0
0.2	50	51	50	1	0	98.03922	0	1.960784
0.3	50	51	50	1	0	98.03922	0	1.960784
0.4	50	51	50	1	0	98.03922	0	1.960784
0.5	50	53	50	3	0	94.33962	0	5.660377
0.6	50	53	50	3	0	94.33962	0	5.660377
0.7	50	53	50	3	0	94.33962	0	5.660377
0.8	50	53	50	3	0	94.33962	0	5.660377
0.9	50	57	50	7	0	87.7193	0	12.2807

ตารางที่ 4.12 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนมากกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.2$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	48	48	0	2	100	4	0
0.2	50	48	48	0	2	100	4	0
0.3	50	49	48	1	2	97.95918	4	2.040816
0.4	50	50	49	1	1	98	2	2
0.5	50	50	49	1	1	98	2	2
0.6	50	50	49	1	1	98	2	2
0.7	50	50	49	1	1	98	2	2
0.8	50	50	49	1	1	98	2	2
0.9	50	52	49	3	1	94.23077	2	5.769231

ตารางที่ 4.13 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนมากกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.3$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	45	45	0	5	100	10	0
0.2	50	45	45	0	5	100	10	0
0.3	50	46	46	0	4	100	8	0
0.4	50	47	46	1	4	97.87234	8	2.12766
0.5	50	47	46	1	4	97.87234	8	2.12766
0.6	50	47	46	1	4	97.87234	8	2.12766
0.7	50	47	46	1	4	97.87234	8	2.12766
0.8	50	47	46	1	4	97.87234	8	2.12766
0.9	50	47	46	1	4	97.87234	8	2.12766

ตารางที่ 4.14 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนมากกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.4$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	39	39	0	11	100	22	0
0.2	50	39	39	0	11	100	22	0
0.3	50	39	39	0	11	100	22	0
0.4	50	39	39	0	11	100	22	0
0.5	50	41	40	1	10	97.56098	20	2.439024
0.6	50	41	40	1	10	97.56098	20	2.439024
0.7	50	42	41	1	9	97.61905	18	2.380952
0.8	50	42	41	1	9	97.61905	18	2.380952
0.9	50	43	42	1	8	97.67442	16	2.325581

ตารางที่ 4.15 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนมากกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.5$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	32	32	0	18	100	36	0
0.2	50	33	33	0	17	100	34	0
0.3	50	33	33	0	17	100	34	0
0.4	50	33	33	0	17	100	34	0
0.5	50	33	33	0	17	100	34	0
0.6	50	34	33	1	17	97.05882	34	2.941176
0.7	50	34	33	1	17	97.05882	34	2.941176
0.8	50	34	33	1	17	97.05882	34	2.941176
0.9	50	34	33	1	17	97.05882	34	2.941176

ตารางที่ 4.16 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนมากกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.6$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	29	29	0	21	100	42	0
0.2	50	29	29	0	21	100	42	0
0.3	50	29	29	0	21	100	42	0
0.4	50	30	30	0	20	100	40	0
0.5	50	30	30	0	20	100	40	0
0.6	50	30	30	0	20	100	40	0
0.7	50	31	30	1	20	96.77419	40	3.225806
0.8	50	31	30	1	20	96.77419	40	3.225806
0.9	50	31	30	1	20	96.77419	40	3.225806

ตารางที่ 4.17 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนมากกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.7$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	28	28	0	22	100	44	0
0.2	50	28	28	0	22	100	44	0
0.3	50	28	28	0	22	100	44	0
0.4	50	28	28	0	22	100	44	0
0.5	50	28	28	0	22	100	44	0
0.6	50	28	28	0	22	100	44	0
0.7	50	28	28	0	22	100	44	0
0.8	50	29	28	1	22	96.55172	44	3.448276
0.9	50	29	28	1	22	96.55172	44	3.448276

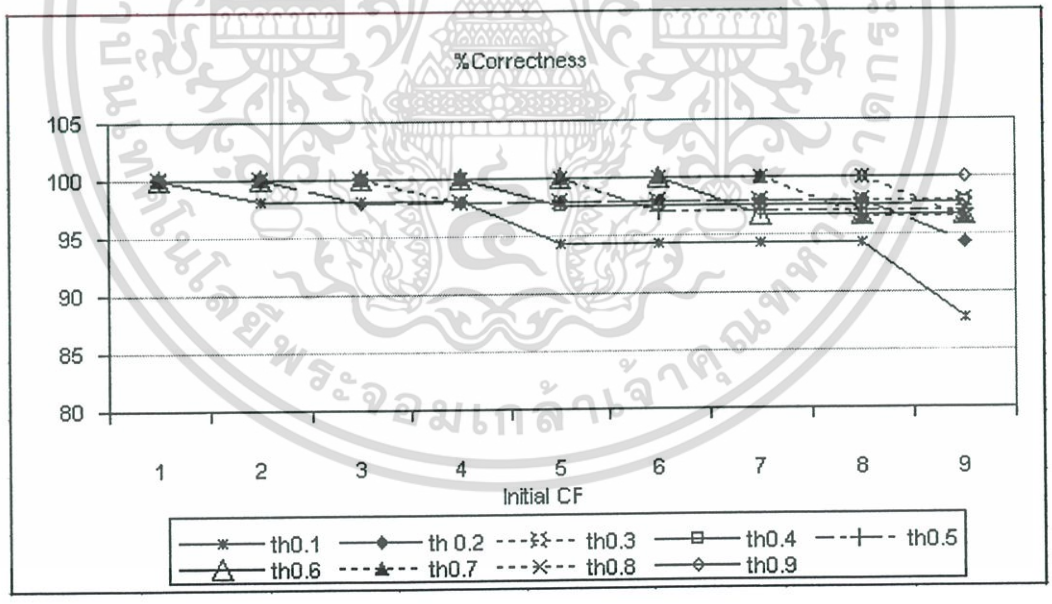
ตารางที่ 4.18 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนมากกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.8$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	27	27	0	23	100	46	0
0.2	50	27	27	0	23	100	46	0
0.3	50	27	27	0	23	100	46	0
0.4	50	27	27	0	23	100	46	0
0.5	50	27	27	0	23	100	46	0
0.6	50	27	27	0	23	100	46	0
0.7	50	27	27	0	23	100	46	0
0.8	50	27	27	0	23	100	46	0
0.9	50	28	27	1	23	96.42857	46	3.571429

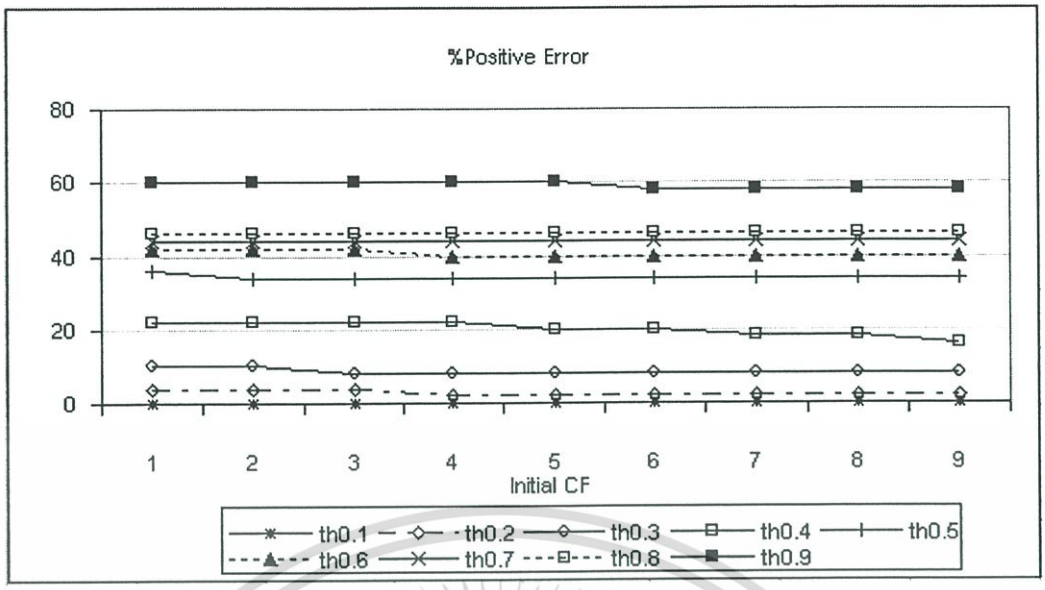
ตารางที่ 4.19 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนมากกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.9$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	20	20	0	30	100	60	0
0.2	50	20	20	0	30	100	60	0
0.3	50	20	20	0	30	100	60	0
0.4	50	20	20	0	30	100	60	0
0.5	50	20	20	0	30	100	60	0
0.6	50	21	21	0	29	100	58	0
0.7	50	21	21	0	29	100	58	0
0.8	50	21	21	0	29	100	58	0
0.9	50	21	21	0	29	100	58	0

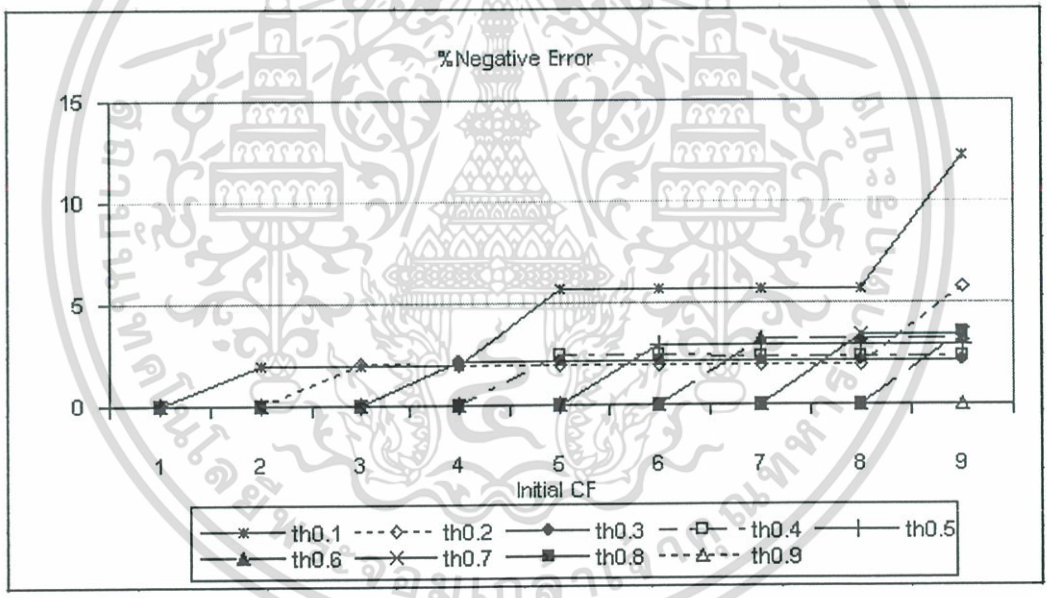
จากตารางที่ 4.11 ถึง ตารางที่ 4.19 คือตารางแสดงผลการทดลองการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ ในกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนมากกว่าจดหมายแบบธรรมดา โดยกำหนดให้ CF_{IN} มีค่าตั้งแต่ 0.1 – 0.9 ที่ CF_{TH} ค่าต่างๆกัน ซึ่งสามารถนำข้อมูลที่ได้มา สร้างกราฟแสดงความสัมพันธ์ระหว่าง %Correctness, %Positive, %Negative Error และ CF_{IN} ที่ CF_{TH} ค่าต่างๆเพื่อวิเคราะห์และเปรียบเทียบหาค่า CF_{IN} และ CF_{TH} ที่เหมาะสม



รูปที่ 4.9 กราฟแสดงความสัมพันธ์ระหว่าง %Correctness และ CF_{IN} ที่ CF_{TH} ค่าต่างๆ



รูปที่ 4.10 กราฟแสดงความสัมพันธ์ระหว่าง %Positive Error และ CF_{IN} ที่ CF_{TH} ค่าต่างๆ



รูปที่ 4.11 กราฟแสดงความสัมพันธ์ระหว่าง %Negative Error และ CF_{IN} ที่ CF_{TH} ค่าต่างๆ

<u>วิเคราะห์กราฟ %Correctness, %Positive Error, %Negative Error และ CF_{IN} เพื่อหาค่า CF_{IN} ที่เหมาะสม (ในกรณีที่จำนวนจดหมายจากเมลลิงลิสต์มากกว่าจำนวนจดหมายแบบธรรมดา)</u>

จากกราฟที่ 4.9, 4.10 และ 4.11 พบว่าถ้ากำหนดให้ค่า CF_{IN} มีค่าต่ำ จะทำให้ค่า %Correctness หรือ ค่าความถูกต้องมีค่าสูงที่สุด แต่ในขณะเดียวกันถ้าเพิ่มค่า CF_{IN} ให้สูงขึ้นเรื่อยๆ กราฟของค่าความถูกต้องจะมีแนวโน้มลดลง เช่นเดียวกันกับแนวโน้มของกราฟ %Positive Error ที่มีค่าสูงเมื่อ

กำหนดค่า CF_{IN} ต่ำ นั่นหมายถึงมีจำนวนเมลลิงลิสต์ที่ตรวจจับได้น้อยกว่าจำนวนเมลลิงลิสต์ที่มีอยู่จริง และเมื่อเพิ่มค่า CF_{IN} ให้สูงขึ้นเรื่อยๆ ค่า %Positive Error จะลดลง คือจำนวนของเมลลิงลิสต์ที่ตรวจจับได้มีจำนวนเพิ่มมากขึ้น

แต่ในทางตรงกันข้ามค่า %Negative Error จะมีค่าน้อยเมื่อกำหนดค่า CF_{IN} ต่ำ และจะมีแนวโน้มเพิ่มขึ้นเมื่อค่า CF_{IN} สูงขึ้นเรื่อยๆ นั่นหมายถึงมีจำนวนเมลลิงลิสต์ที่ตรวจจับผิดพลาดสูงขึ้นถ้ากำหนดให้ค่า CF_{IN} มีค่าสูง

จากการวิเคราะห์ข้างต้นพบว่าแนวโน้มการลดลงของค่าความถูกต้อง และแนวโน้มการลดลงของ %Positive Error รวมถึงแนวโน้มการเพิ่มขึ้นของ %Negative Error เมื่อกำหนดให้ค่า CF_{IN} มีค่าสูงขึ้นเรื่อยๆ ซึ่งสรุปได้เช่นเดียวกับ ในกรณีที่จำนวนจดหมายจากเมลลิงลิสต์ใกล้เคียงกับจำนวนจดหมายแบบธรรมดา ต่อไปเป็นการแสดงผลการทดลองการตรวจจับเมลลิงลิสต์ในกรณีที่ จำนวนของจดหมายจากเมลลิงลิสต์น้อยกว่าจำนวนจดหมายแบบธรรมดา

ตารางที่ 4.20 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนน้อยกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.1$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	48	48	0	2	100	4	0
0.2	50	50	49	1	1	98	2	2
0.3	50	50	49	1	1	98	2	2
0.4	50	50	49	1	1	98	2	2
0.5	50	51	50	1	0	98.03922	0	1.960784
0.6	50	51	50	1	0	98.03922	0	1.960784
0.7	50	51	50	1	0	98.03922	0	1.960784
0.8	50	51	50	1	0	98.03922	0	1.960784
0.9	50	54	50	4	0	92.59259	0	7.407407

ตารางที่ 4.21 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนน้อยกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.2$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	45	45	0	5	100	10	0
0.2	50	45	45	0	5	100	10	0
0.3	50	47	46	1	4	97.87234	8	2.12766
0.4	50	48	47	1	3	97.91667	6	2.083333
0.5	50	48	47	1	3	97.91667	6	2.083333
0.6	50	48	47	1	3	97.91667	6	2.083333
0.7	50	49	48	1	2	97.95918	4	2.040816
0.8	50	49	48	1	2	97.95918	4	2.040816
0.9	50	50	49	1	1	98	2	2

ตารางที่ 4.22 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนน้อยกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.3$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	43	43	0	7	100	14	0
0.2	50	43	43	0	7	100	14	0
0.3	50	44	44	0	6	100	12	0
0.4	50	46	45	1	5	97.82609	10	2.173913
0.5	50	46	45	1	5	97.82609	10	2.173913
0.6	50	46	45	1	5	97.82609	10	2.173913
0.7	50	46	45	1	5	97.82609	10	2.173913
0.8	50	46	45	1	5	97.82609	10	2.173913
0.9	50	46	45	1	5	97.82609	10	2.173913

ตารางที่ 4.23 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนน้อยกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.4$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	39	39	0	11	100	22	0
0.2	50	39	39	0	11	100	22	0
0.3	50	43	43	0	7	100	14	0
0.4	50	43	43	0	7	100	14	0
0.5	50	45	44	1	6	97.77778	12	2.222222
0.6	50	45	44	1	6	97.77778	12	2.222222
0.7	50	46	45	1	5	97.82609	10	2.173913
0.8	50	46	45	1	5	97.82609	10	2.173913
0.9	50	46	45	1	5	97.82609	10	2.173913

ตารางที่ 4.24 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนน้อยกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.5$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	39	39	0	11	100	22	0
0.2	50	39	39	0	11	100	22	0
0.3	50	39	39	0	11	100	22	0
0.4	50	39	39	0	11	100	22	0
0.5	50	39	39	0	11	100	22	0
0.6	50	41	40	1	10	97.56098	20	2.439024
0.7	50	41	40	1	10	97.56098	20	2.439024
0.8	50	41	40	1	10	97.56098	20	2.439024
0.9	50	41	40	1	10	97.56098	20	2.439024

ตารางที่ 4.25 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนน้อยกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.6$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	36	36	0	14	100	28	0
0.2	50	36	36	0	14	100	28	0
0.3	50	36	36	0	14	100	28	0
0.4	50	36	36	0	14	100	28	0
0.5	50	37	37	0	13	100	26	0
0.6	50	37	37	0	13	100	26	0
0.7	50	39	38	1	12	97.4359	24	2.564103
0.8	50	39	38	1	12	97.4359	24	2.564103
0.9	50	40	39	1	11	97.5	22	2.5

ตารางที่ 4.26 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนน้อยกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.7$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	32	32	0	18	100	36	0
0.2	50	32	32	0	18	100	36	0
0.3	50	32	32	0	18	100	36	0
0.4	50	32	32	0	18	100	36	0
0.5	50	32	32	0	18	100	36	0
0.6	50	32	32	0	18	100	36	0
0.7	50	36	36	0	14	100	28	0
0.8	50	38	37	1	13	97.36842	26	2.631579
0.9	50	39	38	1	12	97.4359	24	2.564103

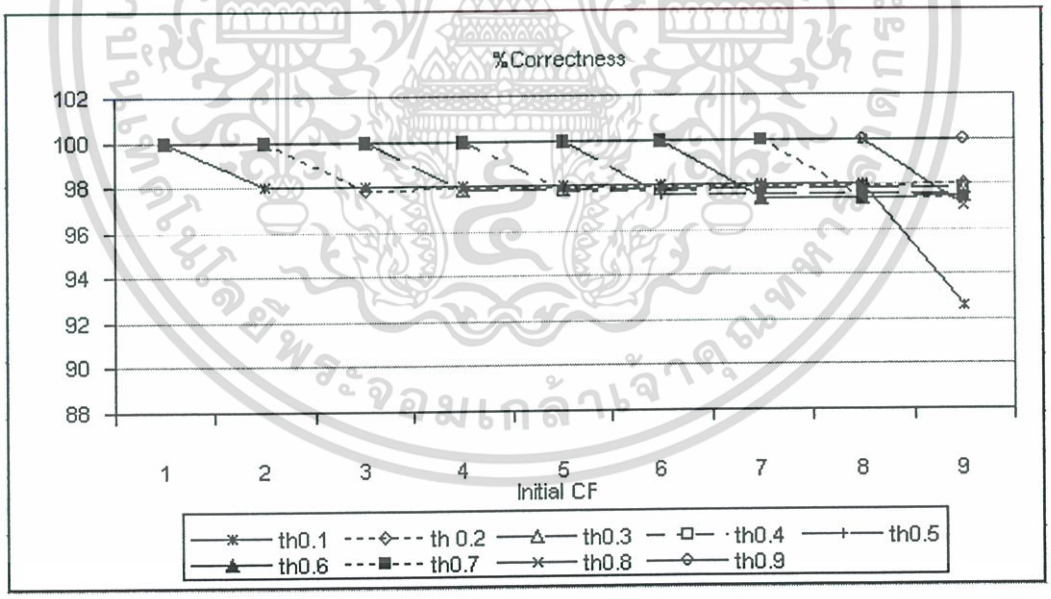
ตารางที่ 4.27 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนน้อยกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.8$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	22	22	0	28	100	56	0
0.2	50	22	22	0	28	100	56	0
0.3	50	27	27	0	23	100	46	0
0.4	50	27	27	0	23	100	46	0
0.5	50	31	31	0	19	100	38	0
0.6	50	31	31	0	19	100	38	0
0.7	50	31	31	0	19	100	38	0
0.8	50	31	31	0	19	100	38	0
0.9	50	34	33	1	17	97.05882	34	2.941176

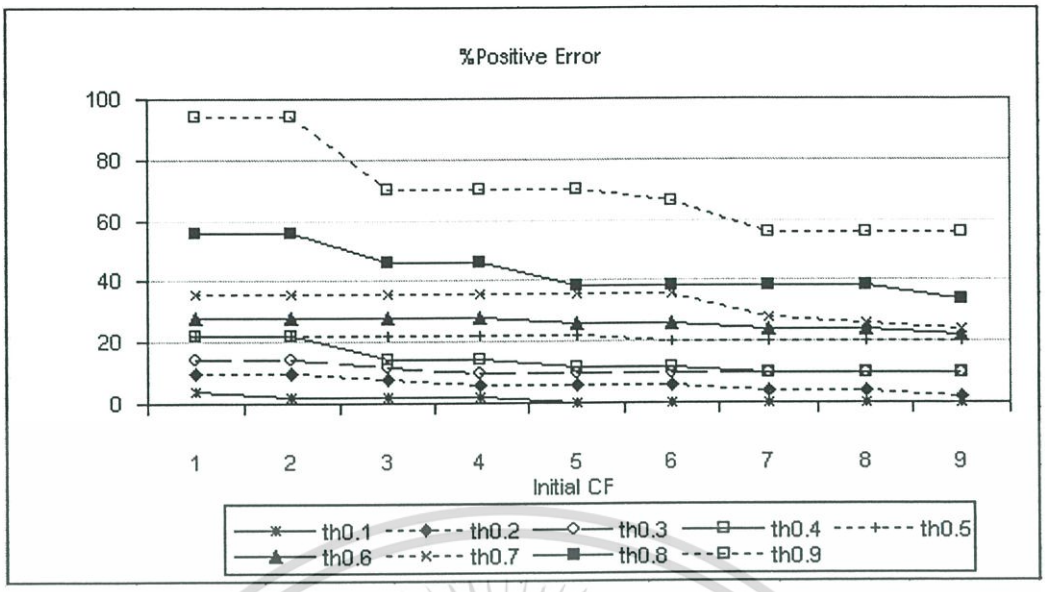
ตารางที่ 4.28 ผลการทดลองกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนน้อยกว่าจดหมายแบบธรรมดา ที่ $CF_{TH} = 0.9$

Initial CF	Exact ML	Detect ML	Correct ML	Negative ML	Positive ML	%Correctness	%Positive Error	%Negative Error
0.1	50	3	3	0	47	100	94	0
0.2	50	3	3	0	47	100	94	0
0.3	50	15	15	0	35	100	70	0
0.4	50	15	15	0	35	100	70	0
0.5	50	15	15	0	35	100	70	0
0.6	50	17	17	0	33	100	66	0
0.7	50	22	22	0	28	100	56	0
0.8	50	22	22	0	28	100	56	0
0.9	50	22	22	0	28	100	56	0

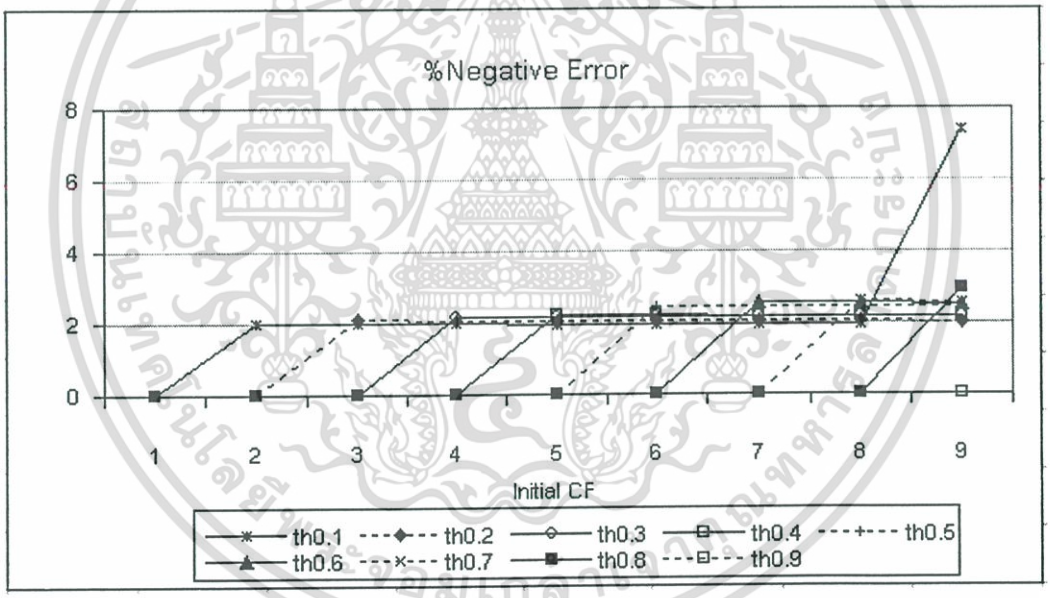
จากตารางที่ 4.20 ถึง ตารางที่ 4.28 คือตารางแสดงผลการทดลองการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ ในกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนน้อยกว่าจดหมายแบบธรรมดา โดยกำหนดให้ CF_{IN} มีค่าตั้งแต่ 0.1 – 0.9 ที่ CF_{TH} ค่าต่างๆกัน ซึ่งสามารถนำข้อมูลที่ได้มา สร้างกราฟแสดงความสัมพันธ์ระหว่าง %Correctness, %Positive, %Negative Error และ CF_{IN} ที่ CF_{TH} ค่าต่างๆเพื่อวิเคราะห์และเปรียบเทียบหาค่า CF_{IN} และ CF_{TH} ที่เหมาะสม



รูปที่ 4.12 กราฟแสดงความสัมพันธ์ระหว่าง %Correctness และ CF_{IN} ที่ CF_{TH} ค่าต่างๆ



รูปที่ 4.13 กราฟแสดงความสัมพันธ์ระหว่าง %Positive Error และ CF_{IN} ที่ CF_{TH} ค่าต่างๆ



รูปที่ 4.14 กราฟแสดงความสัมพันธ์ระหว่าง %Negative Error และ CF_{IN} ที่ CF_{TH} ค่าต่างๆ

<u>วิเคราะห์กราฟ %Correctness, %Positive Error, %Negative Error และ CF_{IN} เพื่อหาค่า CF_{IN} ที่เหมาะสม (ในกรณีที่จำนวนจดหมายจากเมลลิงลิสต์น้อยกว่าจำนวนจดหมายแบบธรรมดา)</u>

จากกราฟที่ 4.12, 4.13 และ 4.14 พบว่าถ้ากำหนดให้ค่า CF_{IN} มีค่าต่ำ จะทำให้ค่า %Correctness หรือ ค่าความถูกต้องมีค่าสูงที่สุด แต่ในขณะเดียวกันถ้าเพิ่มค่า CF_{IN} ให้สูงขึ้นเรื่อยๆ กราฟของค่าความถูกต้องจะมีแนวโน้มลดลง เช่นเดียวกันกับแนวโน้มของกราฟ %Positive Error ที่มีค่าสูงเมื่อ

กำหนดค่า CF_{IN} ต่ำ นั่นหมายถึงมีจำนวนเมลลิงลิสต์ที่ตรวจจับได้น้อยกว่าจำนวนเมลลิงลิสต์ที่มีอยู่จริง และเมื่อเพิ่มค่า CF_{IN} ให้สูงขึ้นเรื่อยๆ ค่า %Positive Error จะลดลง คือจำนวนของเมลลิงลิสต์ที่ตรวจจับได้มีจำนวนเพิ่มมากขึ้น

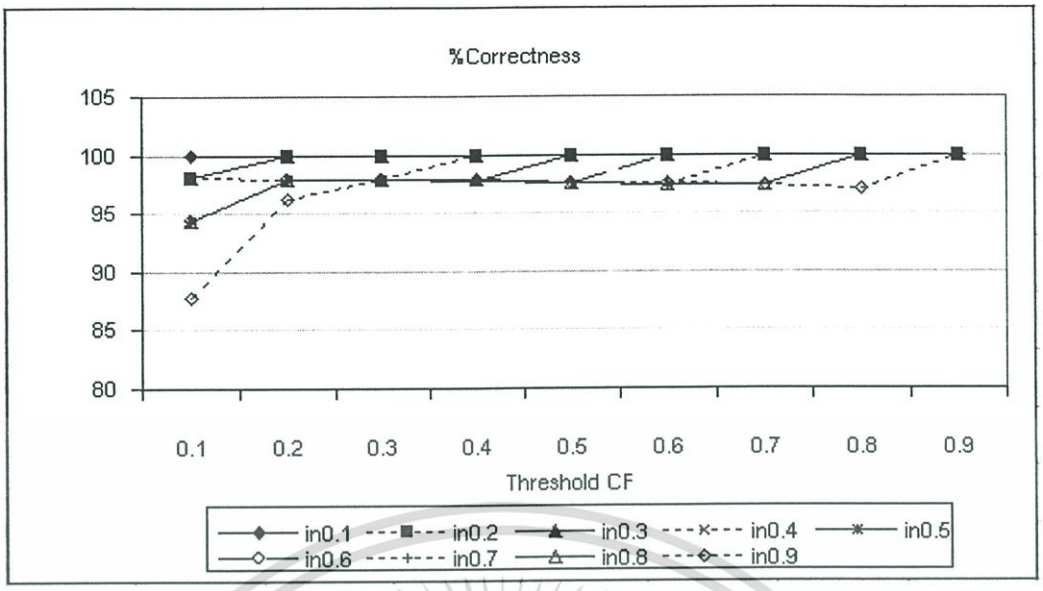
แต่ในทางตรงกันข้ามค่า %Negative Error จะมีค่าน้อยเมื่อกำหนดค่า CF_{IN} ต่ำ และจะมีแนวโน้มเพิ่มขึ้นเมื่อค่า CF_{IN} สูงขึ้นเรื่อยๆ นั่นหมายถึงมีจำนวนเมลลิงลิสต์ที่ตรวจจับผิดพลาดสูงขึ้นถ้ากำหนดให้ค่า CF_{IN} มีค่าสูง

จากการวิเคราะห์ข้างต้นพบว่าแนมโน้มการลดลงของค่าความถูกต้อง และแนวโน้มการลดลงของ %Positive Error รวมถึงแนวโน้มการเพิ่มขึ้นของ %Negative Error เมื่อกำหนดให้ค่า CF_{IN} มีค่าสูงขึ้นเรื่อยๆ ซึ่งสรุปได้เช่นเดียวกับ ในกรณีที่จำนวนจดหมายจากเมลลิงลิสต์ใกล้เคียงกับจำนวนจดหมายแบบธรรมดา และในกรณีที่จำนวนของจดหมายจากเมลลิงลิสต์มากกว่าจำนวนจดหมายแบบธรรมดา

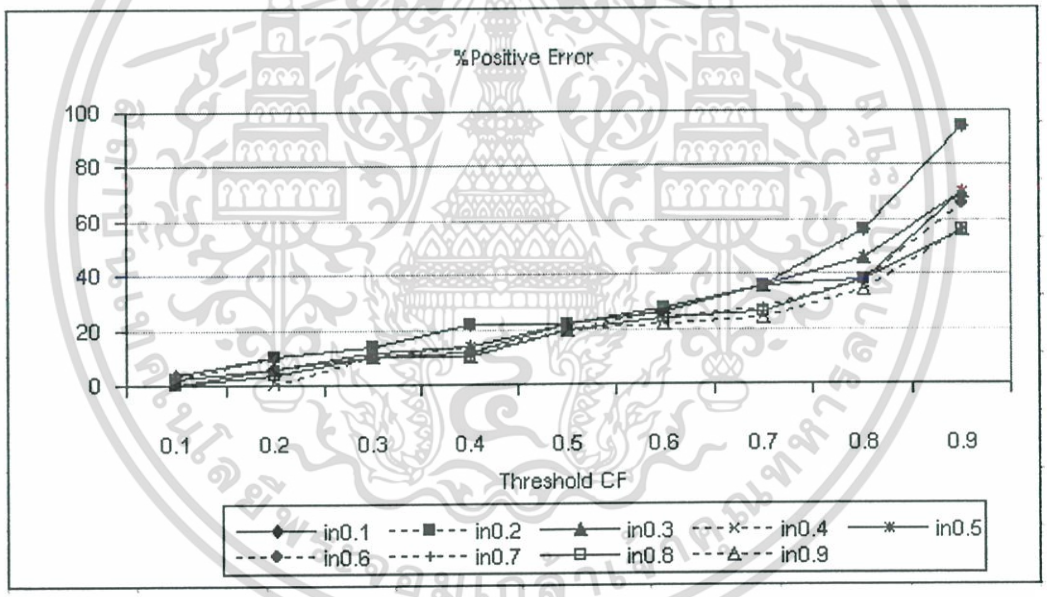
จะเห็นว่าในการกำหนดค่า CF_{IN} นั้นสมควรเริ่มที่ค่าต่ำ เช่น 0.1 หรือ 0.2 หรือ 0.3 เพราะเป็นช่วงที่ทำให้มี %Correctness หรือ ค่าความถูกต้องสูงที่สุดไม่ว่าค่า CF_{TH} จะเป็นเท่าไหร่ก็ตาม นอกจากนั้นยังมีค่า %Negative Error น้อย นั่นคือมีจำนวนเมลลิงลิสต์ที่ตรวจจับผิดพลาดน้อย แต่ที่ค่า CF_{IN} ต่ำนี้จะมีผลทำให้จำนวนเมลลิงลิสต์ที่ตรวจจับได้มีจำนวนน้อยกว่าเมลลิงลิสต์ที่มีอยู่จริง แต่ไม่มีผลกระทบกับ Submailing List ที่ต้องนำข้อมูลไปใช้ ในทางกลับกัน ถ้าต้องการให้ได้จำนวนเมลลิงลิสต์ที่ตรวจจับได้มีจำนวนมากขึ้นแต่สิ่งที่ตามมาก็คือ %Negative Error จะมากขึ้นไปด้วย หมายถึงจำนวนเมลลิงลิสต์ที่ตรวจจับผิดพลาดจะมีจำนวนมากขึ้น ซึ่งเมื่อนำข้อมูลไปใช้กับ Submailing List อาจส่งผลให้เกิดความเสียหายได้

4.3.2 การทดลองเพื่อศึกษาผลกระทบของ Threshold Confidential Factor ที่มีผลต่อ %Correctness, %Positive Error และ %Negative Error

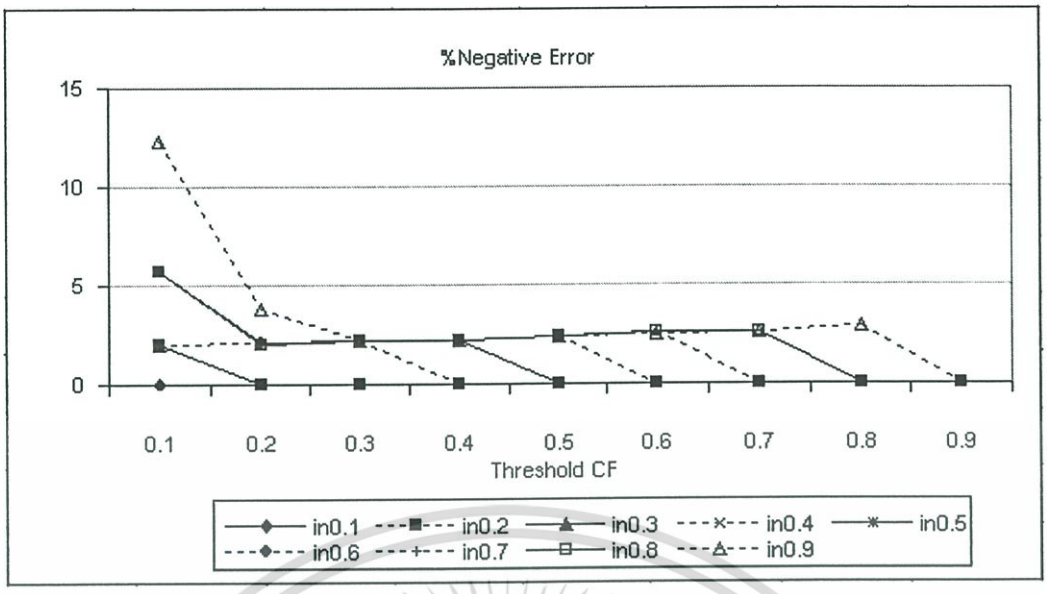
ต่อไปเป็นการนำผลการทดลองที่ได้จาก ตารางที่ 4.2 ถึง 4.10 คือผลการทดลองในกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนใกล้เคียงกันกับจดหมายแบบธรรมดา มาสร้างกราฟเพื่อหาค่า Threshold Confidential Factor ที่เหมาะสม



รูปที่ 4.15 กราฟแสดงความสัมพันธ์ระหว่าง %Correctness และ CF_{TH} ที่ CF_{IN} ค่าต่างๆ



รูปที่ 4.16 กราฟแสดงความสัมพันธ์ระหว่าง %Positive Error และ CF_{TH} ที่ CF_{IN} ค่าต่างๆ



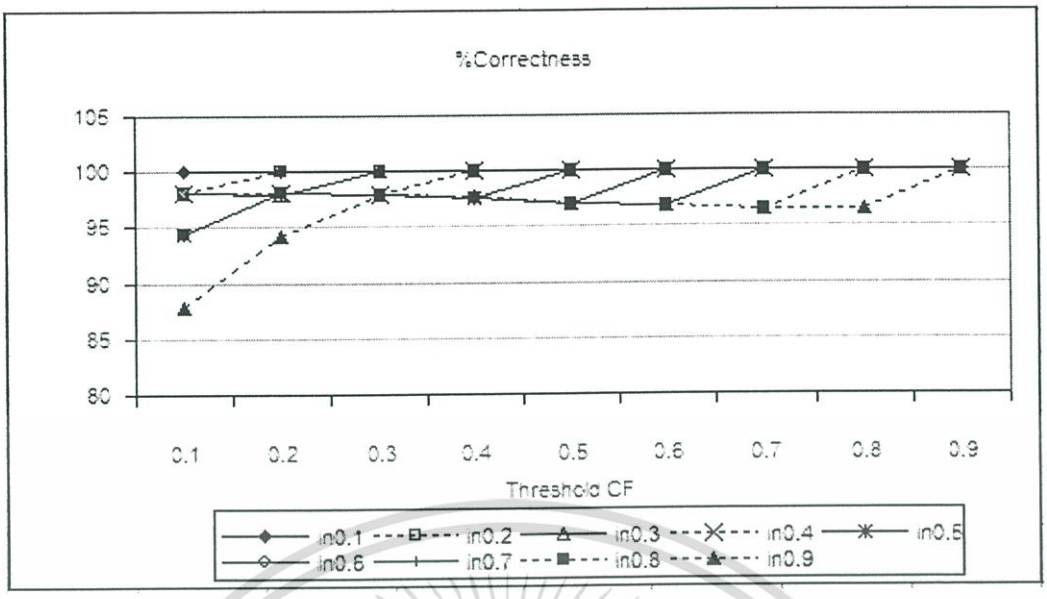
รูปที่ 4.17 กราฟแสดงความสัมพันธ์ระหว่าง %Negative Error และ CF_{TH} ที่ CF_{IN} ค่าต่างๆ

<u>วิเคราะห์กราฟ %Correctness, %Positive Error, %Negative Error และ CF_{TH} เพื่อหาค่า CF_{TH} ที่เหมาะสม (ในกรณีที่จำนวนจดหมายจากเมลลิงลิสต์ใกล้เคียงกับจำนวนจดหมายแบบธรรมดา)</u>

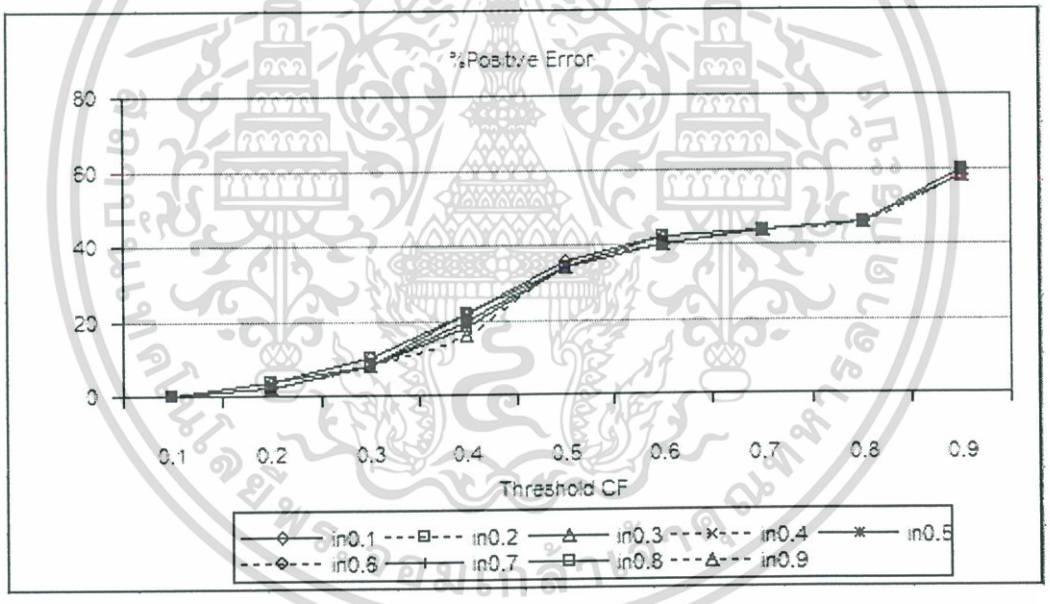
จากการพิจารณากราฟ %Correctness, %Positive Error, %Negative Error และค่า CF_{TH} ที่ค่า CF_{IN} ต่างๆกันพบว่า ถ้ากำหนดให้ CF_{TH} มีค่าสูงแล้ว %Correctness หรือค่าความถูกต้องจะมีค่าสูงตามไปด้วย นอกจากนั้นยังทำให้ค่า %Positive Error สูงที่สุด แต่สำหรับค่า %Negative Error จะมีค่าต่ำที่สุด นั่นหมายถึงจำนวนเมลลิงลิสต์ที่ตรวจจับได้จะมีจำนวนน้อยกว่าในกรณีที่กำหนดให้ค่า CF_{TH} มีค่าต่ำๆ แต่เมลลิงลิสต์ที่ตรวจจับได้จะมีความถูกต้องสูงและมีความผิดพลาดน้อยกว่า

ในทางกลับกันถ้ากำหนดให้ค่า CF_{TH} มีค่าต่ำแล้ว %Correctness หรือค่าความถูกต้องและค่า %Positive Error จะมีค่าต่ำที่สุด แต่ %Negative Error จะมีค่าสูงที่สุด นั่นคือจำนวนเมลลิงลิสต์ที่ตรวจจับได้มีจำนวนมากกว่าในกรณีที่กำหนดให้ค่า CF_{TH} มีค่าสูงๆแต่เมลลิงลิสต์ที่ตรวจจับได้จะมีความถูกต้องน้อยและมีความผิดพลาดมากกว่า

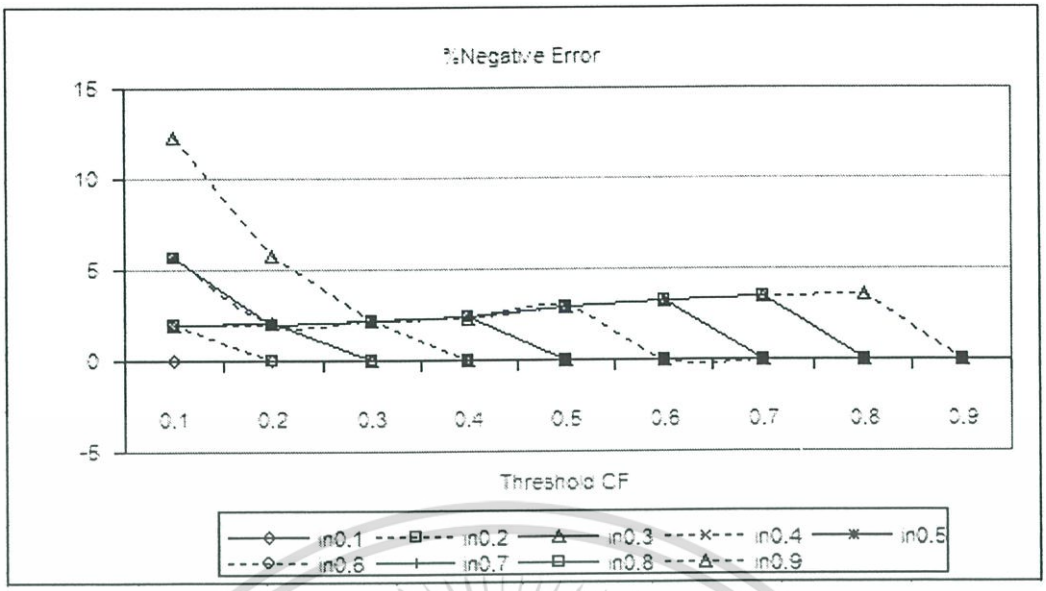
ต่อไปเป็นการนำผลการทดลองที่ได้จาก ตารางที่ 4.11 ถึง 4.19 คือผลการทดลองในกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนมากกว่าจดหมายแบบธรรมดา มาสร้างกราฟเพื่อหาค่า Threshold Confidential Factor ที่เหมาะสม



รูปที่ 4.18 กราฟแสดงความสัมพันธ์ระหว่าง %Correctness และ CF_{TH} ที่ CF_{IN} ค่าต่างๆ



รูปที่ 4.19 กราฟแสดงความสัมพันธ์ระหว่าง %Positive Error และ CF_{TH} ที่ CF_{IN} ค่าต่างๆ



รูปที่ 4.20 กราฟแสดงความสัมพันธ์ระหว่าง %Negative Error และ CF_{TH} ที่ CF_{IN} ค่าต่างๆ

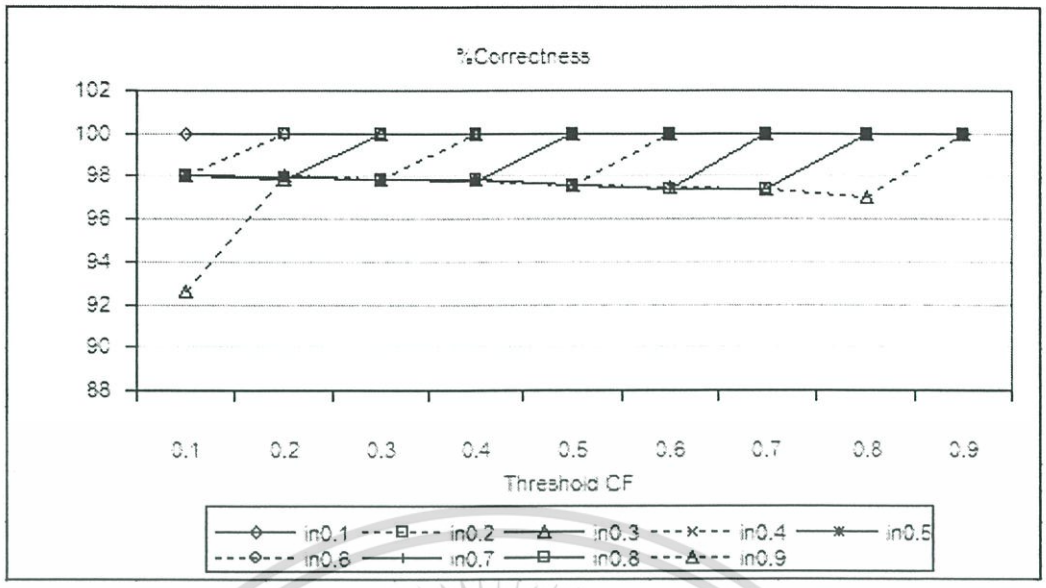
<u>วิเคราะห์กราฟ %Correctness, %Positive Error, %Negative Error และ CF_{TH} เพื่อหาค่า CF_{TH} ที่เหมาะสม (ในกรณีที่จำนวนจดหมายจากเมลลิงลิสต์มากกว่าจำนวนจดหมายแบบธรรมดา)</u>

จากการพิจารณากราฟ %Correctness, %Positive Error, %Negative Error และค่า CF_{TH} ที่ค่า CF_{IN} ต่างๆกันพบว่า ถ้ากำหนดให้ CF_{TH} มีค่าสูงแล้ว %Correctness หรือค่าความถูกต้องจะมีค่าสูงตามไปด้วย นอกจากนั้นยังทำให้ค่า %Positive Error สูงที่สุด แต่สำหรับค่า %Negative Error จะมีค่าต่ำที่สุด นั่นหมายถึงจำนวนเมลลิงลิสต์ที่ตรวจจับได้จะมีจำนวนน้อยกว่าในกรณีที่กำหนดให้ค่า CF_{TH} มีค่าต่ำๆ แต่เมลลิงลิสต์ที่ตรวจจับได้จะมีความถูกต้องสูงและมีความผิดพลาดน้อยกว่า

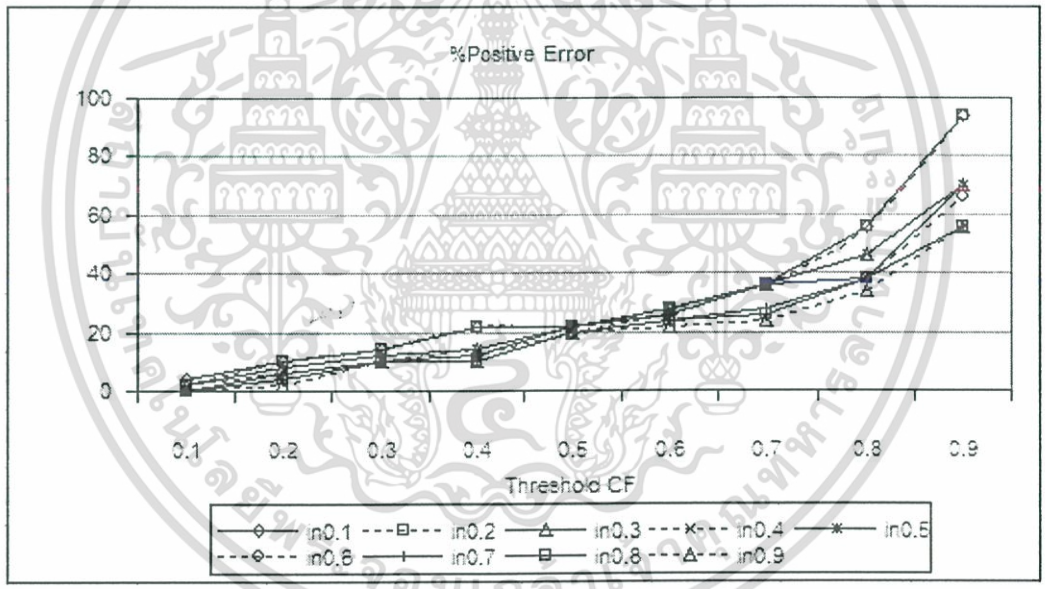
ในทางกลับกันถ้ากำหนดให้ค่า CF_{TH} มีค่าต่ำแล้ว %Correctness หรือค่าความถูกต้องและค่า %Positive Error จะมีค่าต่ำที่สุด แต่ %Negative Error จะมีค่าสูงที่สุด นั่นคือจำนวนเมลลิงลิสต์ที่ตรวจจับได้มีจำนวนมากกว่าในกรณีที่กำหนดให้ค่า CF_{TH} มีค่าสูงแต่เมลลิงลิสต์ที่ตรวจจับได้จะมีความถูกต้องน้อยและมีความผิดพลาดมากกว่า

ซึ่งจากผลที่วิเคราะห์ได้สามารถสรุปได้ตรงกันกับ ในกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนใกล้เคียงกันกับจดหมายแบบธรรมดา นั่นคือเมื่อกำหนดให้ค่า CF_{TH} มีค่าสูงจะเป็นผลให้ %Correctness และ %Positive Error มีค่าสูงกว่า การกำหนดให้ค่า CF_{TH} มีค่าต่ำ ซึ่งตรงกันข้ามกับ %Negative Error

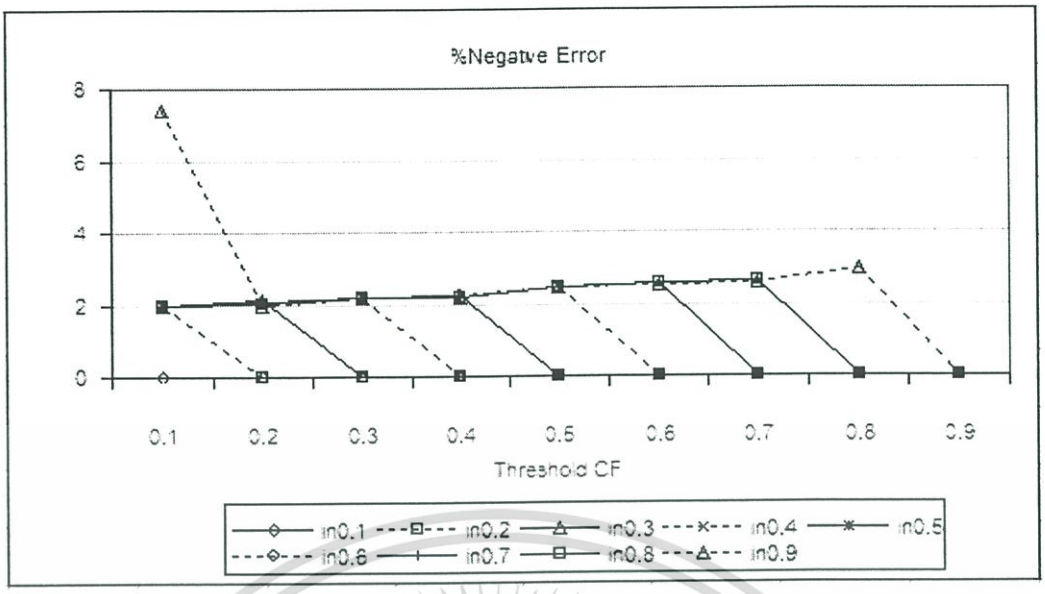
ต่อไปเป็นการนำผลการทดลองที่ได้จาก ตารางที่ 4.20 ถึง 4.28 คือผลการทดลองในกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนน้อยกว่าจดหมายแบบธรรมดา มาสร้างกราฟเพื่อหาค่า Threshold Confidential Factor ที่เหมาะสม



รูปที่ 4.21 กราฟแสดงความสัมพันธ์ระหว่าง %Correctness และ CF_{TH} ที่ CF_{IN} ค่าต่างๆ



รูปที่ 4.22 กราฟแสดงความสัมพันธ์ระหว่าง %Positive Error และ CF_{TH} ที่ CF_{IN} ค่าต่างๆ



รูปที่ 4.23 กราฟแสดงความสัมพันธ์ระหว่าง %Negative Error และ CF_{TH} ที่ CF_{IN} ค่าต่างๆ

<u>วิเคราะห์กราฟ %Correctness, %Positive Error, %Negative Error และ CF_{TH} เพื่อหาค่า CF_{TH} ที่เหมาะสม (ในกรณีที่จำนวนจุดหมายจากเมลลิงลิสต์น้อยกว่าจำนวนจุดหมายแบบธรรมดา)</u>

จากการพิจารณากราฟ %Correctness, %Positive Error, %Negative Error และค่า CF_{TH} ที่ค่า CF_{IN} ต่างๆกันพบว่า ถ้ากำหนดให้ CF_{TH} มีค่าสูงแล้ว %Correctness หรือค่าความถูกต้องจะมีค่าสูงตามไปด้วย นอกจากนั้นยังทำให้ค่า %Positive Error สูงที่สุด แต่สำหรับค่า %Negative Error จะมีค่าต่ำที่สุด นั่นหมายถึงจำนวนเมลลิงลิสต์ที่ตรวจจับได้จะมีจำนวนน้อยกว่าในกรณีที่กำหนดให้ค่า CF_{TH} มีค่าต่ำๆ แต่เมลลิงลิสต์ที่ตรวจจับได้จะมีความถูกต้องสูงและมีความผิดพลาดน้อยกว่า

ในทางกลับกันถ้ากำหนดให้ค่า CF_{TH} มีค่าต่ำแล้ว %Correctness หรือค่าความถูกต้องและค่า %Positive Error จะมีค่าต่ำที่สุด แต่ %Negative Error จะมีค่าสูงที่สุด นั่นคือจำนวนเมลลิงลิสต์ที่ตรวจจับได้มีจำนวนมากกว่าในกรณีที่กำหนดให้ค่า CF_{TH} มีค่าสูงแต่เมลลิงลิสต์ที่ตรวจจับได้จะมีความถูกต้องน้อยและมีความผิดพลาดมากกว่า

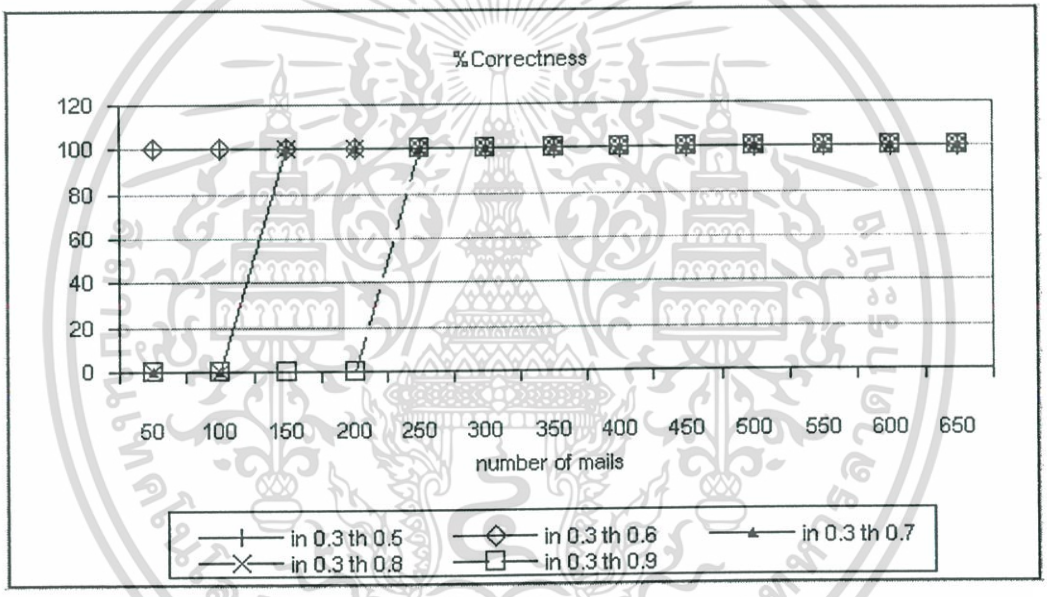
ซึ่งจากผลที่วิเคราะห์ได้สามารถสรุปได้ตรงกันกับ ในกรณีที่จุดหมายจากเมลลิงลิสต์มีจำนวนใกล้เคียงกันกับจุดหมายแบบธรรมดา และในกรณีที่จุดหมายจากเมลลิงลิสต์มีจำนวนมากกว่าจุดหมายแบบธรรมดา นั่นคือเมื่อกำหนดให้ค่า CF_{TH} มีค่าสูงจะเป็นผลให้ %Correctness และ %Positive Error มีค่าสูงกว่า การกำหนดให้ค่า CF_{TH} มีค่าต่ำ ซึ่งตรงกันข้ามกับ %Negative Error

จึงสามารถสรุปได้ว่าเมื่อกำหนดให้ค่า CF_{TH} มีค่าสูงจะเป็นผลให้ค่าความถูกต้อง หรือ %Correctness และ %Positive Error มีค่าสูงตามไปด้วย แต่ในทางกลับกันจะเป็นผลให้ %Negative Error มีค่าต่ำ ซึ่งหมายความว่าที่ค่า CF_{TH} สูงจะได้ข้อมูลที่มีความถูกต้องและมีความผิดพลาดน้อย แต่จำนวนของเมลลิงลิสต์ที่ตรวจจับได้จะมีจำนวนน้อยกว่า การกำหนดให้ค่า CF_{TH} ต่ำ

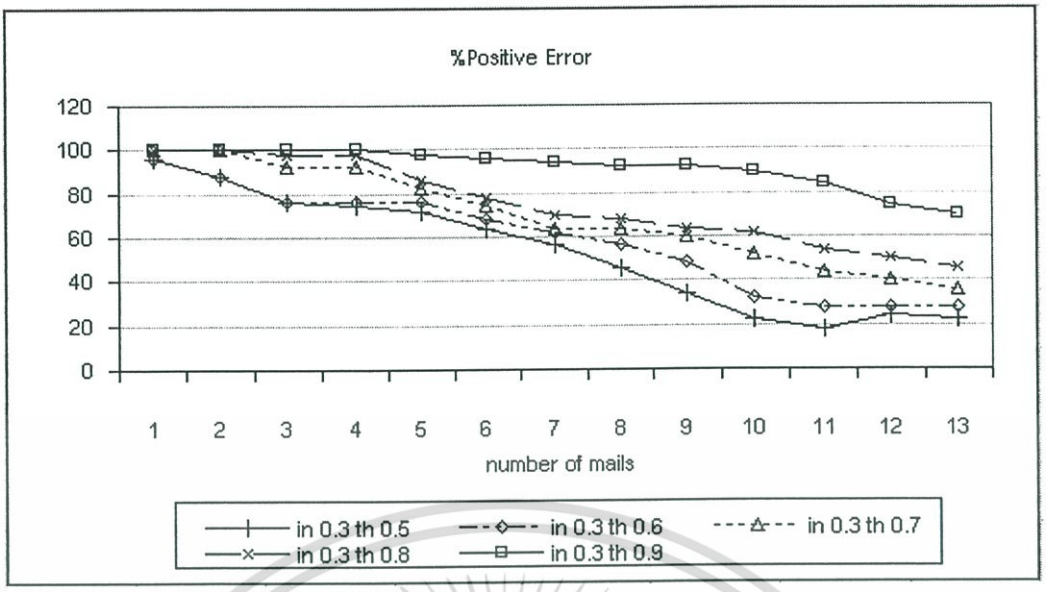
4.3.3 การทดลองเพื่อศึกษาผลกระทบของจำนวนจดหมายอิเล็กทรอนิกส์ ที่มีผลต่อ %Correctness, %Positive Error และ %Negative Error เพื่อกำหนดค่า CF_{IN} และ CF_{TH} ที่เหมาะสม

ต่อไปเป็นกราฟแสดงความสัมพันธ์ระหว่าง %Correctness, %Positive Error, %Negative Error และ ปริมาณของข้อมูล หรือ จำนวนจดหมายอิเล็กทรอนิกส์ที่ผ่านระบบตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ โดยผลจากการสรุปจากกราฟด้านบนควรกำหนดให้ CF_{IN} มีค่าอยู่ในช่วงต่ำคือมีค่าอยู่ระหว่าง 0.1 – 0.3 และผลสรุปจากกราฟด้านบนเช่นกันว่าควรกำหนดให้ CF_{TH} มีค่าอยู่ในช่วงสูงคือมีค่าอยู่ระหว่าง 0.7 – 0.9

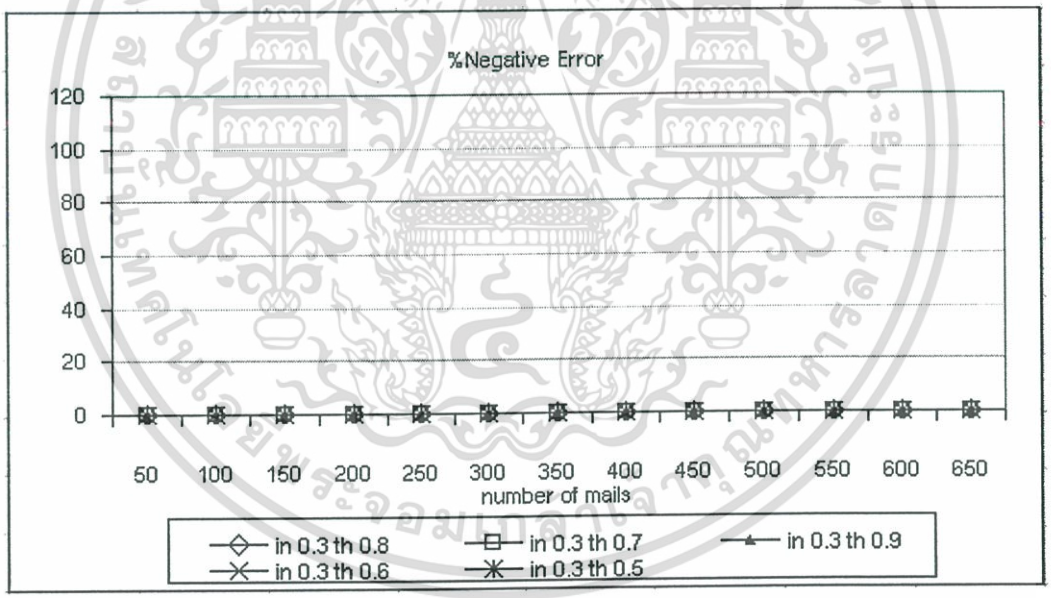
ในกราฟ 4.24, 4.25 และ 4.26 ได้จากผลการทดลองการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ ในกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนใกล้เคียงกับจดหมายแบบธรรมดา



รูปที่ 4.24 กราฟแสดงความสัมพันธ์ระหว่าง %Correctness และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{IN} = 0.3$ และ CF_{TH} ค่าต่างๆ

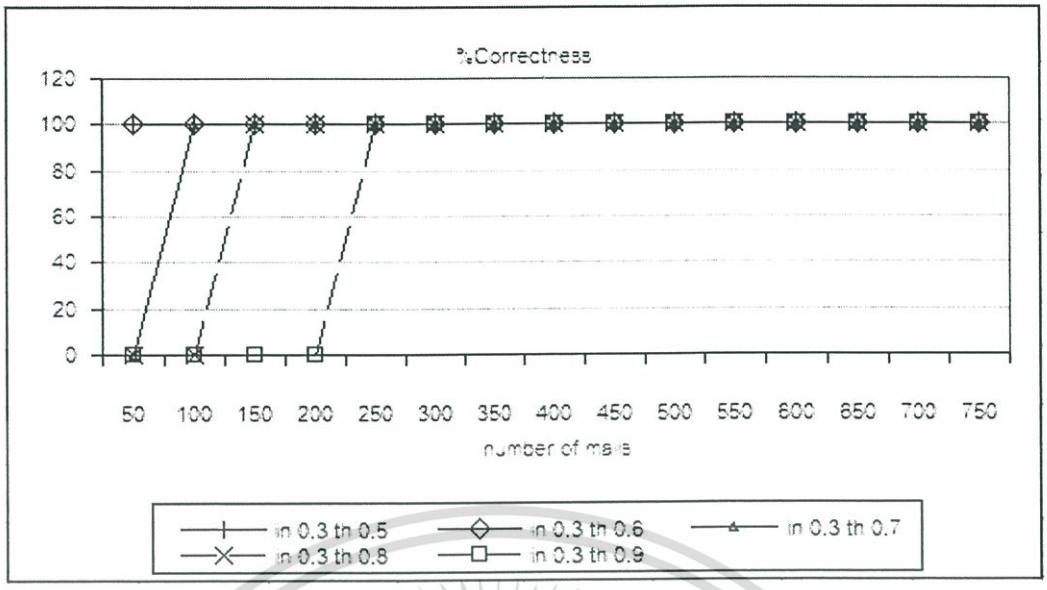


รูปที่ 4.25 กราฟแสดงความสัมพันธ์ระหว่าง %Positive Error และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{IN} = 0.3$ และ CF_{TH} ค่าต่างๆ

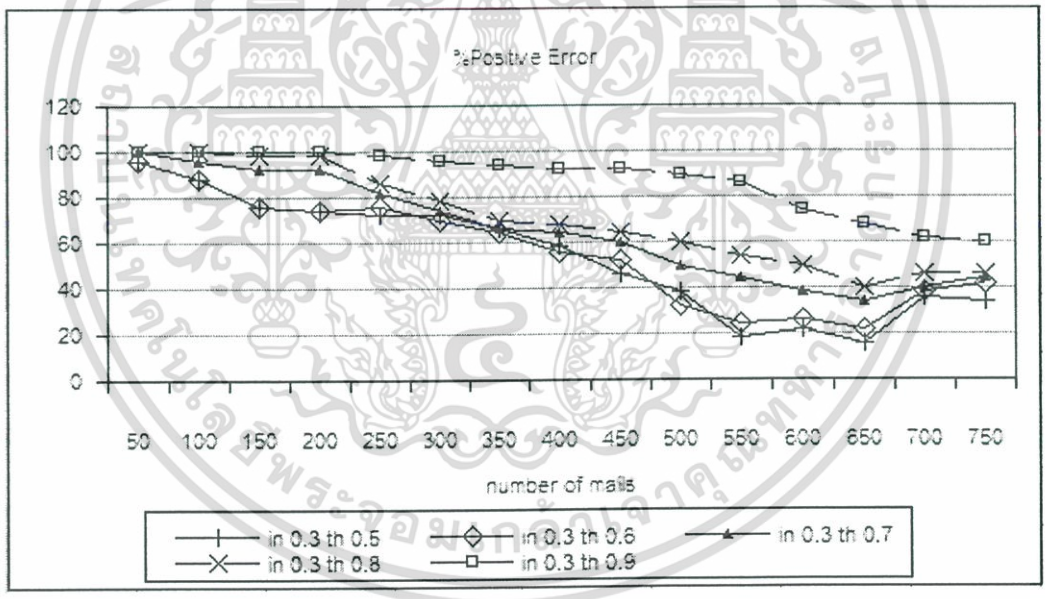


รูปที่ 4.26 กราฟแสดงความสัมพันธ์ระหว่าง %Negative Error และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{IN} = 0.3$ และ CF_{TH} ค่าต่างๆ

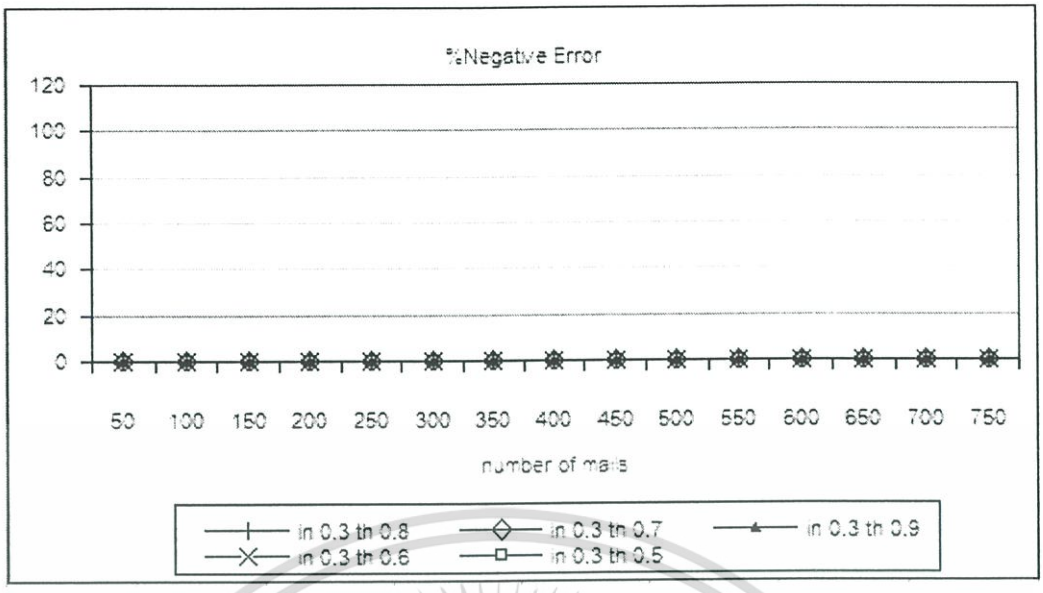
ในกราฟ 4.27, 4.28 และ 4.29 ได้จากผลการทดลองการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ ในกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนมากกว่าจดหมายแบบธรรมดา



รูปที่ 4.27 กราฟแสดงความสัมพันธ์ระหว่าง %Correctness และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{IN} = 0.3$ และ CF_{TH} ค่าต่างๆ

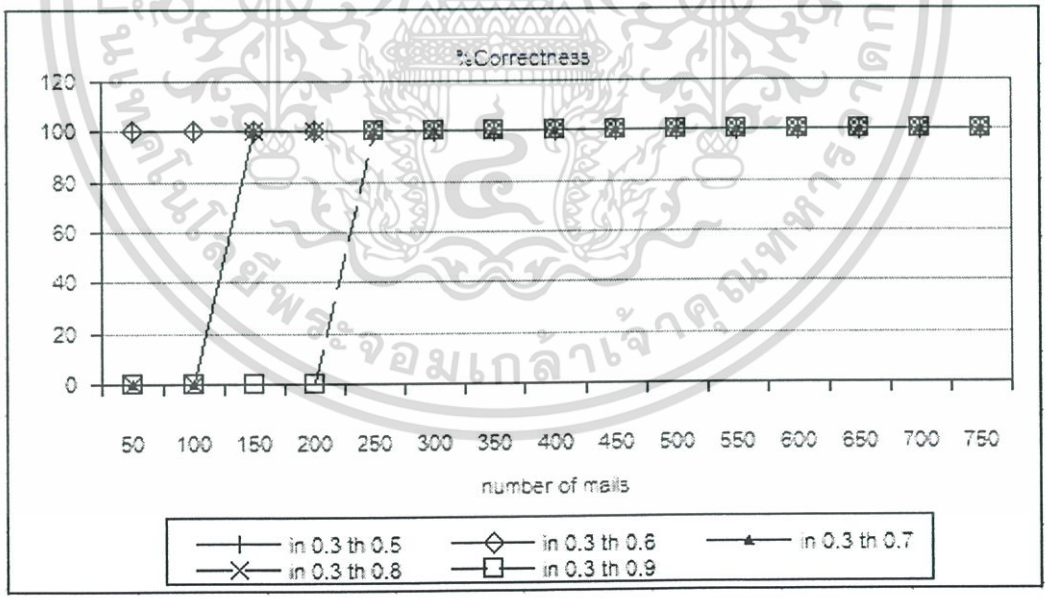


รูปที่ 4.28 กราฟแสดงความสัมพันธ์ระหว่าง %Positive Error และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{IN} = 0.3$ และ CF_{TH} ค่าต่างๆ

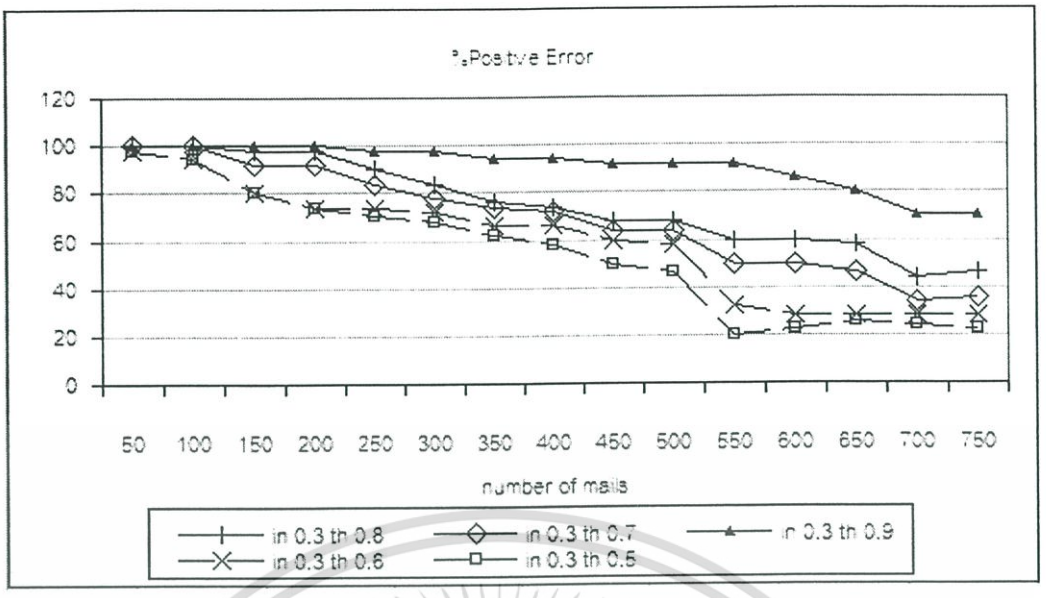


รูปที่ 4.29 กราฟแสดงความสัมพันธ์ระหว่าง %Negative Error และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{IN} = 0.3$ และ CF_{TH} ค่าต่างๆ

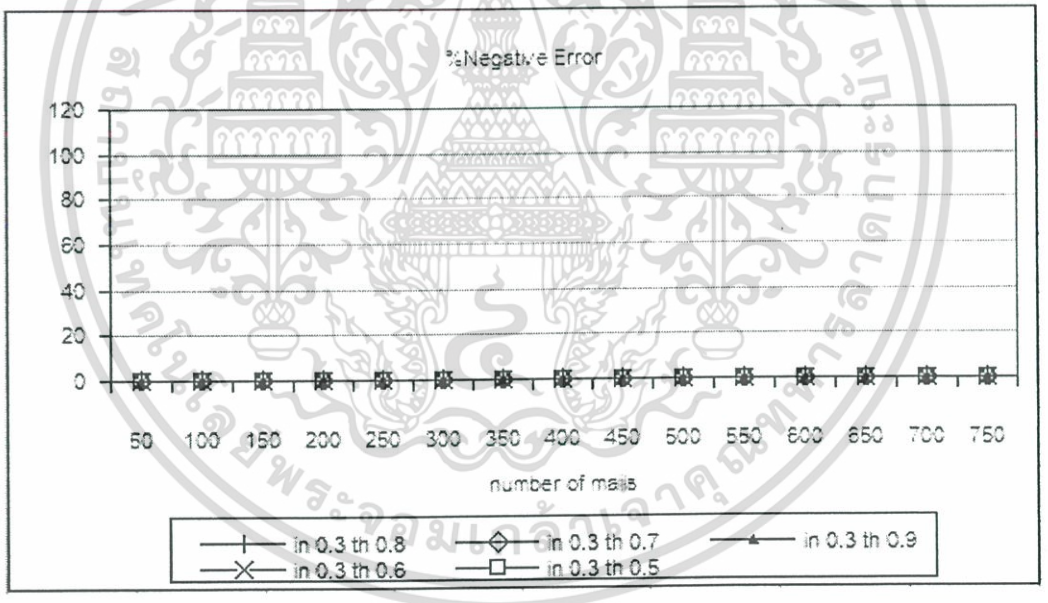
ในกราฟ 4.30, 4.31 และ 4.32 ได้จากผลการทดลองการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ ในกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนน้อยกว่าจดหมายแบบธรรมดา



รูปที่ 4.30 กราฟแสดงความสัมพันธ์ระหว่าง %Correctness และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{IN} = 0.3$ และ CF_{TH} ค่าต่างๆ



รูปที่ 4.31 กราฟแสดงความสัมพันธ์ระหว่าง %Positive Error และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{IN} = 0.3$ และ CF_{TH} ค่าต่างๆ



รูปที่ 4.32 กราฟแสดงความสัมพันธ์ระหว่าง %Negative Error และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{IN} = 0.3$ และ CF_{TH} ค่าต่างๆ

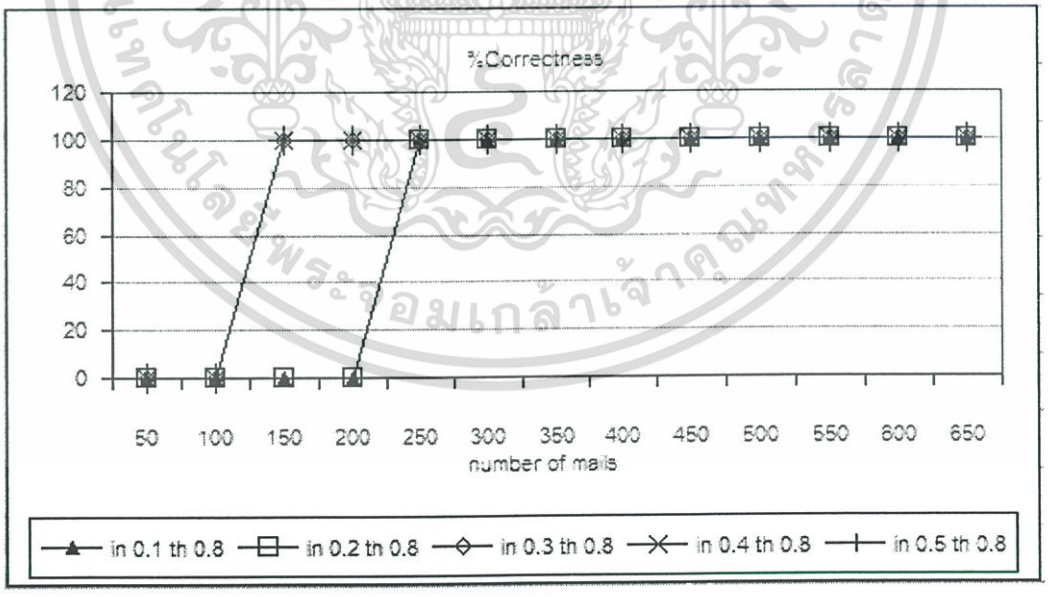
<u>วิเคราะห์กราฟความสัมพันธ์ระหว่าง %Correctness, %Positive Error, %Negative Error และจำนวนของจดหมายอิเล็กทรอนิกส์เพื่อหาค่า CF_{TH} ที่เหมาะสม</u>

จากกราฟพบว่า เมื่อเพิ่มจำนวนของจดหมายอิเล็กทรอนิกส์ที่ใช้ในการตรวจจับเมลลิงลิสต์พบว่า %Correctness มีแนวโน้มเพิ่มสูงขึ้น ส่วน %Positive Error มีแนวโน้มต่ำลง นั่นหมายถึงมีจำนวนเมลลิงลิสต์ที่ตรวจจับได้เพิ่มมากขึ้น และค่า %Negative Error มีค่าเป็น 0 หรือมีค่าน้อยมาก แสดงว่าถ้ามีปริมาณของจดหมายอิเล็กทรอนิกส์ที่ผ่านเข้าออกจากระบบ จำนวนมาก จะทำให้ระบบสามารถตรวจจับเมลลิงลิสต์ได้แม่นยำมากยิ่งขึ้น

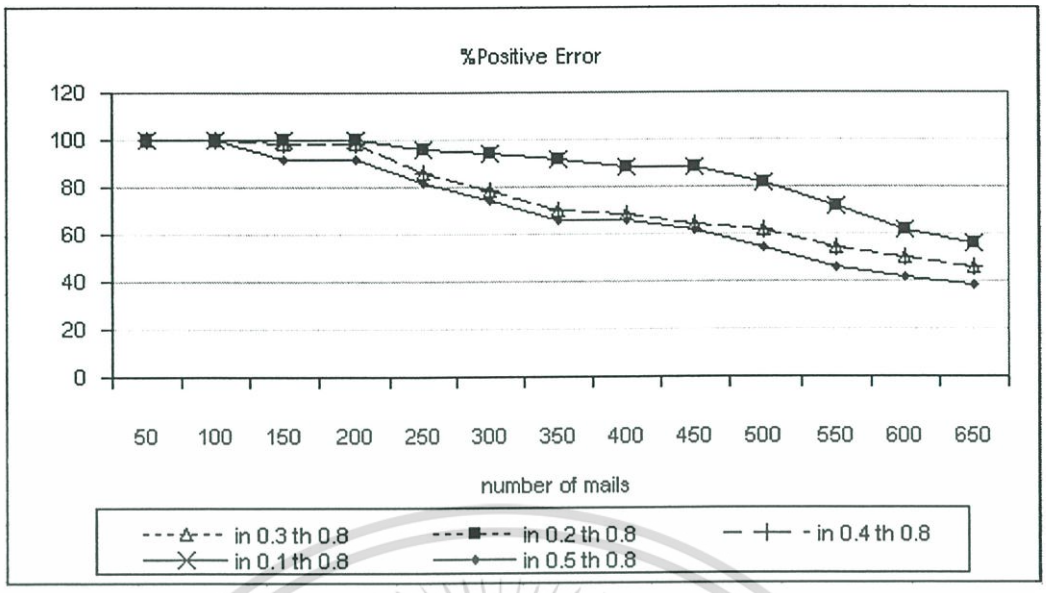
นอกจากนี้เมื่อพิจารณาที่ค่า CF_{TH} ต่างๆพบว่า ที่ CF_{TH} มีค่าต่ำจะใช้จำนวนจดหมายอิเล็กทรอนิกส์ในการตรวจจับน้อยกว่าการกำหนดให้ค่า CF_{TH} มีค่าสูง เมื่อนำข้อมูลที่วิเคราะห์ได้ประกอบกับการวิเคราะห์ในหัวข้อที่ 4.3.2 ซึ่งสรุปไว้ว่า ที่ CF_{TH} มีค่าสูงจะมีค่าความถูกต้องสูงและให้ค่าความผิดพลาดต่ำ จะสามารถสรุปได้ว่าควรเลือกค่า $CF_{TH} = 0.8$ เพราะทำให้ข้อมูลเมลลิงลิสต์ที่ตรวจจับได้มีความถูกต้องสูงและมีความผิดพลาดน้อย นอกจากนั้นยังใช้จำนวนจดหมายในการตรวจจับน้อย

ต่อไปเป็นกราฟแสดงความสัมพันธ์ระหว่าง %Correctness, %Positive Error, %Negative Error และ ปริมาณของข้อมูล ที่ $CF_{TH} = 0.8$ และค่า CF_{IN} ต่างๆเพื่อหาค่าที่เหมาะสม

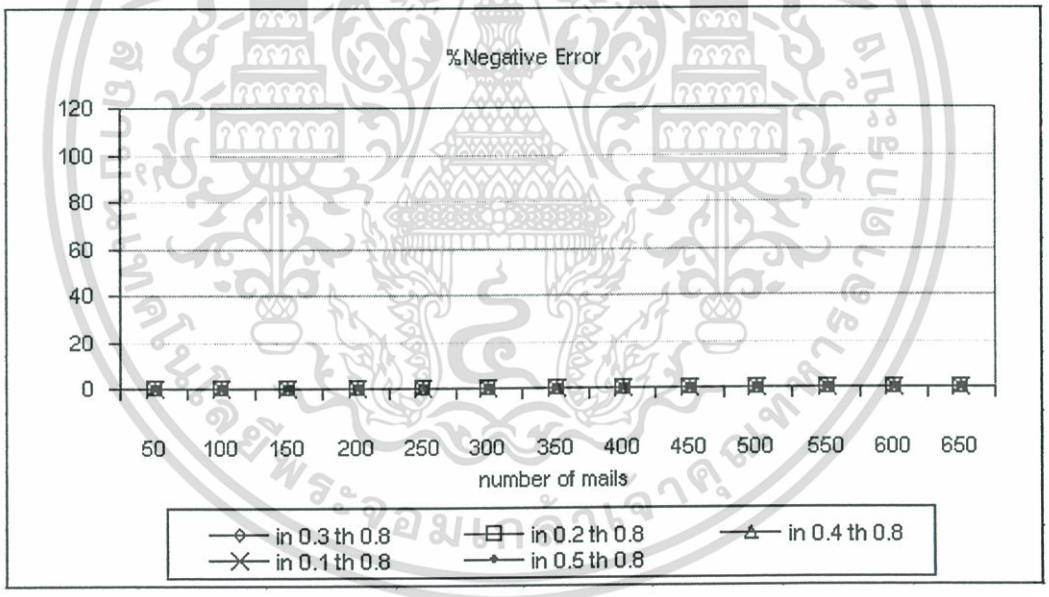
ในกราฟ 4.33, 4.34 และ 4.35 ได้จากผลการทดลองการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ ในกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนใกล้เคียงกับจดหมายแบบธรรมดา



รูปที่ 4.33 กราฟแสดงความสัมพันธ์ระหว่าง %Correctness และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{TH} = 0.8$ และ CF_{IN} ค่าต่างๆ

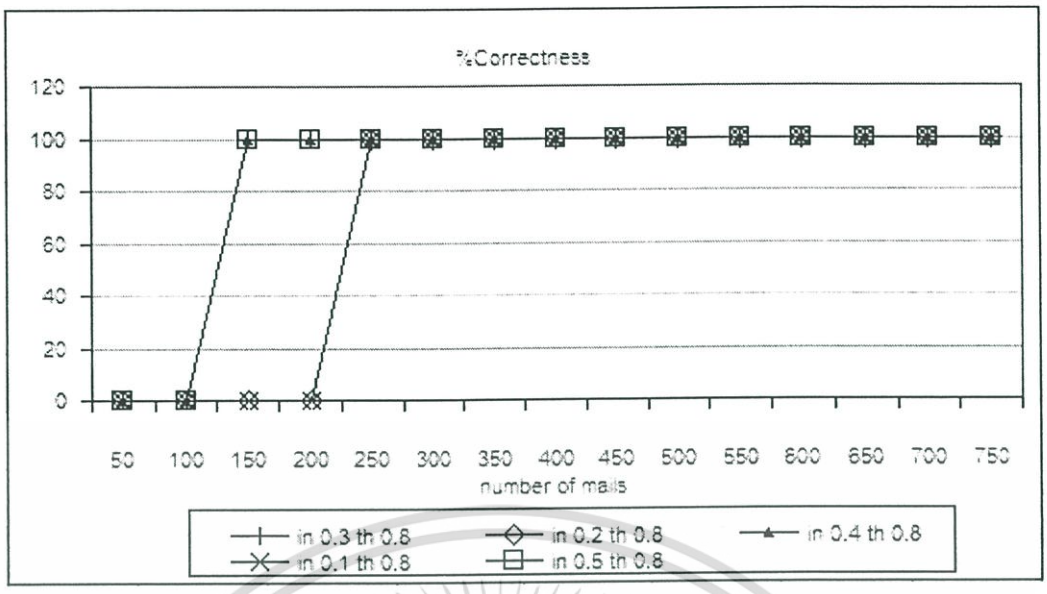


รูปที่ 4.34 กราฟแสดงความสัมพันธ์ระหว่าง %Positive Error และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{TH} = 0.8$ และ CF_{IN} ค่าต่างๆ

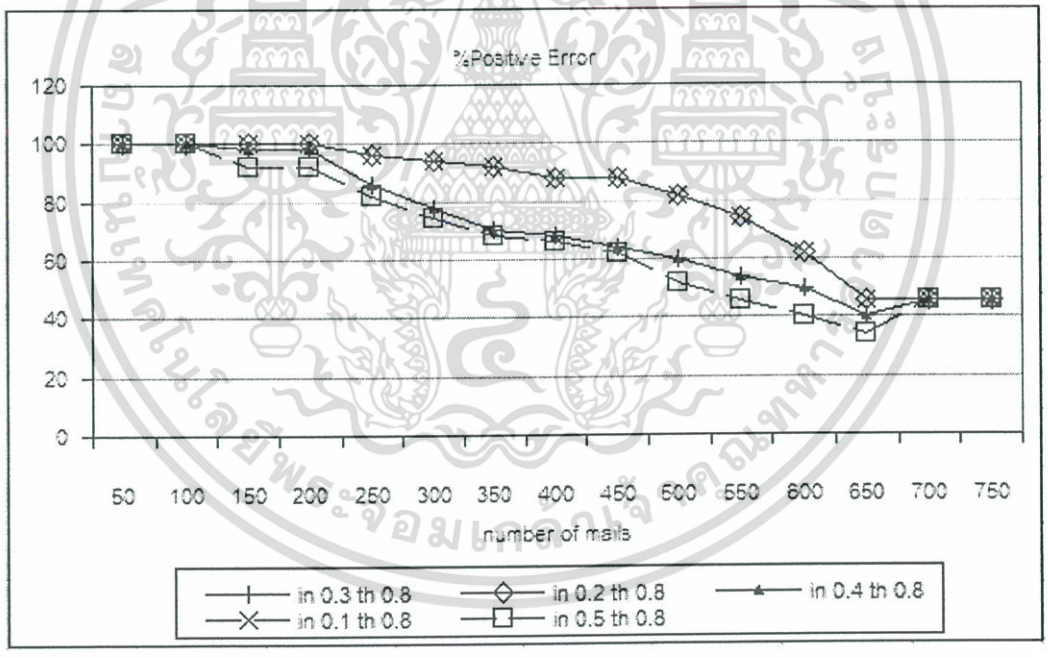


รูปที่ 4.35 กราฟแสดงความสัมพันธ์ระหว่าง %Negative Error และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{TH} = 0.8$ และ CF_{IN} ค่าต่างๆ

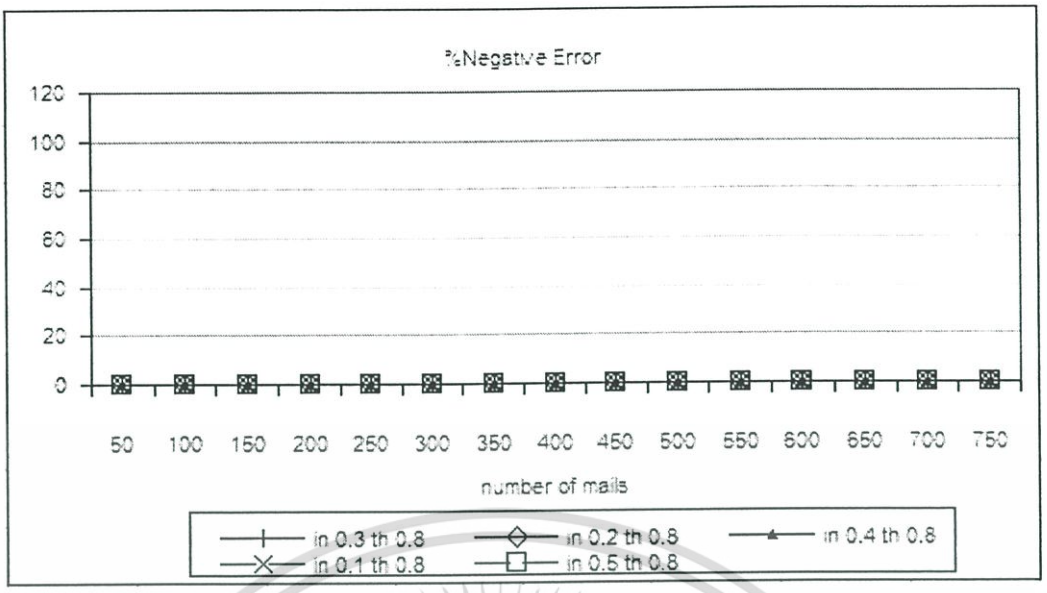
ในกราฟ 4.36, 4.37 และ 4.38 ได้จากผลการทดลองการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ ในกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนมากกว่าจดหมายแบบธรรมดา



รูปที่ 4.36 กราฟแสดงความสัมพันธ์ระหว่าง %Correctness และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{TH} = 0.8$ และ CF_{IN} ค่าต่างๆ

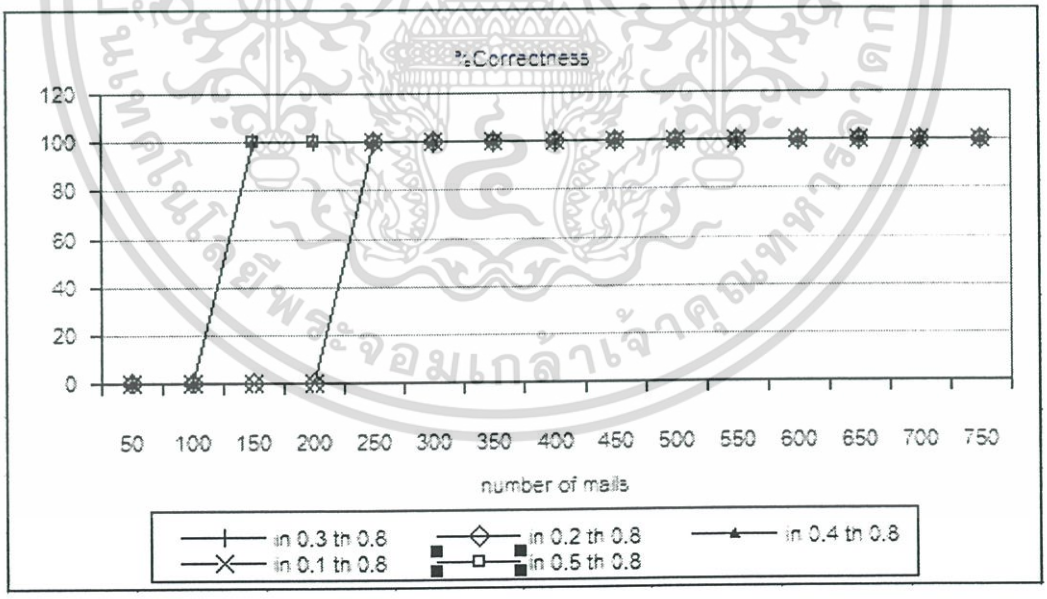


รูปที่ 4.37 กราฟแสดงความสัมพันธ์ระหว่าง %Positive Error และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{TH} = 0.8$ และ CF_{IN} ค่าต่างๆ

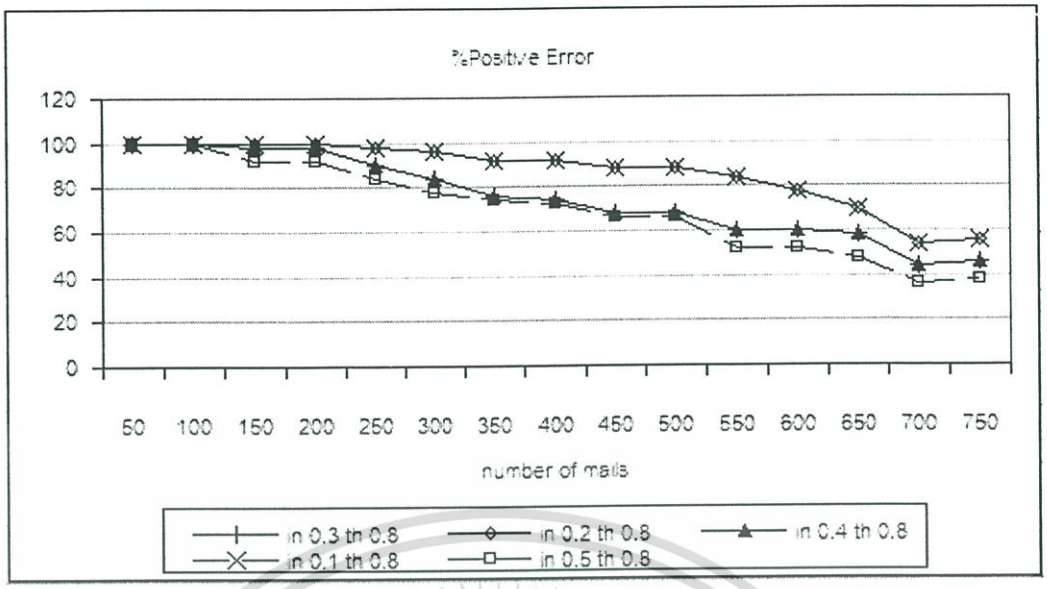


รูปที่ 4.38 กราฟแสดงความสัมพันธ์ระหว่าง %Negative Error และจำนวนจดหมายอิเล็กทรอนิกส์ ที่ $CF_{TH} = 0.8$ และ CF_{IN} ค่าต่างๆ

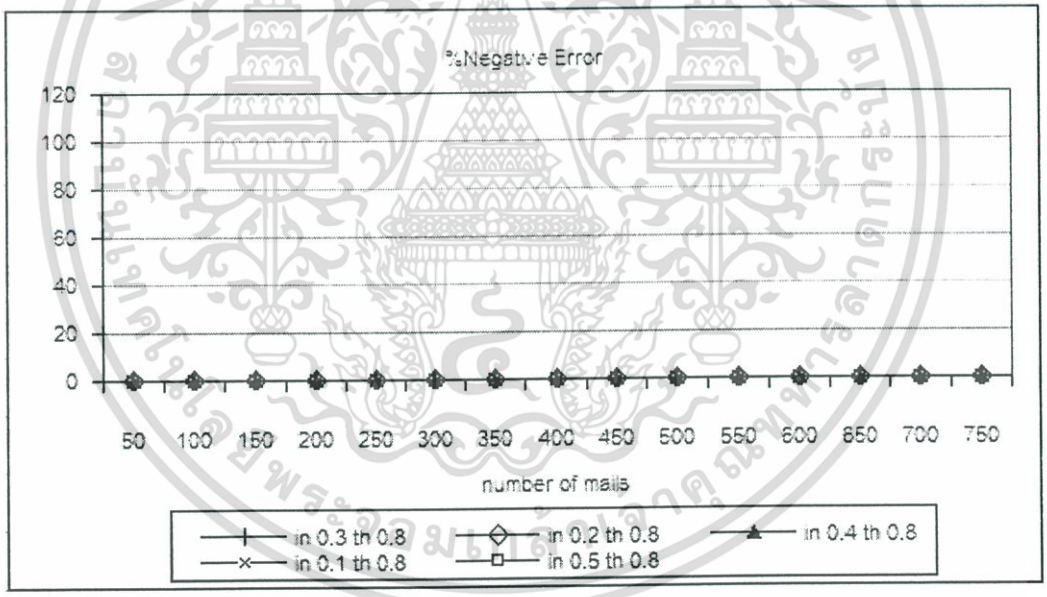
ในกราฟ 4.39, 4.40 และ 4.41 ได้จากผลการทดลองการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ ในกรณีที่จดหมายจากเมลลิงลิสต์มีจำนวนน้อยกว่าจดหมายแบบธรรมดา



รูปที่ 4.39 กราฟแสดงความสัมพันธ์ระหว่าง %Correctness และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{TH} = 0.8$ และ CF_{IN} ค่าต่างๆ



รูปที่ 4.40 กราฟแสดงความสัมพันธ์ระหว่าง %Positive Error และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{TH} = 0.8$ และ CF_{IN} ค่าต่างๆ



รูปที่ 4.41 กราฟแสดงความสัมพันธ์ระหว่าง %Negative Error และจำนวนจดหมายอิเล็กทรอนิกส์ที่ $CF_{TH} = 0.8$ และ CF_{IN} ค่าต่างๆ

<u>วิเคราะห์กราฟความสัมพันธ์ระหว่าง %Correctness, %Positive Error, %Negative Error และจำนวนของจดหมายอิเล็กทรอนิกส์เพื่อหาค่า CF_{IN} ที่เหมาะสม</u>

จากกราฟพบว่า เมื่อมีจำนวนของจดหมายอิเล็กทรอกนิกส์ที่ผ่านระบบการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์เพิ่มขึ้นเรื่อยๆ ทำให้ %Correctness หรือ ค่าความถูกต้องมีแนวโน้มสูง

ขึ้นเรื่อยๆเช่นกัน นอกจากนั้นค่า %Positive Error ยังมีแนวโน้มลดลงเรื่อยๆ นั่นคือจำนวนของเมลลิงลิสต์ที่ตรวจจับได้มีจำนวนเพิ่มขึ้นเรื่อยๆ ส่วนค่า %Negative Error มีค่าเป็น 0 หรือมีค่าน้อยมาก แสดงว่า ถ้ามีปริมาณของจดหมายอิเล็กทรอนิกส์ที่ผ่านเข้าออกจากระบบเป็นจำนวนมาก จะทำให้ระบบสามารถตรวจจับเมลลิงลิสต์ได้แม่นยำมากยิ่งขึ้น

นอกจากนี้เมื่อพิจารณาที่ค่า CF_{IN} ต่างๆพบว่า ที่ CF_{IN} มีค่าต่ำจะใช้จำนวนจดหมายอิเล็กทรอนิกส์ในการตรวจจับมากกว่าการกำหนดให้ค่า CF_{IN} มีค่าสูง เมื่อนำข้อมูลที่วิเคราะห์ได้ประกอบกับการวิเคราะห์ในหัวข้อที่ 4.3.1 ซึ่งสรุปไว้ว่า ที่ CF_{IN} มีค่าต่ำจะมีค่าความถูกต้องสูงและให้ค่าความผิดพลาดต่ำ จะสามารถสรุปได้ว่าควรเลือกค่า $CF_{IN} = 0.3$ เพราะทำให้ข้อมูลเมลลิงลิสต์ที่ตรวจจับได้มีความถูกต้องสูงและมีความผิดพลาดน้อย นอกจากนั้นยังใช้จำนวนจดหมายในการตรวจจับน้อย

จะเห็นว่าระบบตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์นั้น จะต้องใช้ปริมาณข้อมูลจดหมายหลายฉบับขึ้นไปจึงจะสามารถตรวจจับพบเมลลิงลิสต์ได้ เป็นเพราะว่าระบบตรวจจับเมลลิงลิสต์มีขั้นตอนในการปรับเปลี่ยนค่าความมั่นใจซึ่งต้องใช้ข้อมูลหลายๆข้อมูล ไม่สามารถสรุปได้จากข้อมูลของจดหมายอิเล็กทรอนิกส์ที่ผ่านเข้ามาเพียงฉบับเดียว นอกจากนี้ยังอาจเป็นเพราะว่า ในช่วงเวลาที่ดักจับข้อมูลนั้นมีปริมาณของจดหมายแบบอื่นๆเข้ามาปะปนเป็นจำนวนมาก

4.4 สรุปผลการทดลอง

สำหรับความผิดพลาดที่เกิดขึ้นจากระบบไม่ว่าจะเป็น %Positive Error หรือ %Negative Error นั้นเกิดขึ้นจาก ข้อมูลของจดหมายที่นำมาทดลองนั้นเป็นเหตุทำให้เกิดความผิดพลาดขึ้น พบว่าค่า %Negative Error หรือความผิดพลาดจากที่ระบบตรวจจับเมลลิงลิสต์ไม่ถูกต้อง นั่นคือ Mail address ที่สรุปว่าเป็นเมลลิงลิสต์นั้นแท้ที่จริงแล้วเป็น Mail address ของ User ที่อยู่ในเครือข่าย ที่ระบบวิเคราะห์ผิดพลาดอาจเป็นเพราะจดหมายแบบธรรมดานั้น ระบบจะไม่สามารถตรวจจับ Mail address ของผู้รับที่เกิดจากการส่งโดยใช้ Bcc: Header ได้เนื่องจากการทำงานของ Bcc: Header นั้นจะไม่แสดง Mail address ของผู้รับดังนั้นใน Rfc822 Header จะไม่ปรากฏ Mail address ที่เกิดจาก Bcc: Header อยู่แต่ใน Smtp Header นั้นจะระบุ Mail address ที่เป็นของผู้รับทกคน ทำให้จดหมายฉบับนี้มีคุณสมบัติตรงกับจดหมายที่เกิดจากการกระจายของเมลลิงลิสต์ หมายความว่าถ้าจดหมายฉบับใดมีการใช้งาน Bcc: Header ระบบจะสรุปว่าเป็นการทำงานของเมลลิงลิสต์ซึ่งเป็นข้อผลิดพลาดซึ่งอาจจะเกิดขึ้นได้

นอกจากนั้นสำหรับ %Positive Error หรือความผิดพลาดจากที่ระบบตรวจจับเมลลิงลิสต์ไม่พบ นั่นคือ Mail address นั้นเป็นของเมลลิงลิสต์แต่ระบบกลับวิเคราะห์ว่าเป็น Mail address ของ User ทั่วไป สาเหตุที่ระบบวิเคราะห์ผิดพลาดอาจเป็นเพราะมีจดหมายที่สมาชิกส่งถึงเมลลิงลิสต์ปะปนอยู่หลายฉบับซึ่งจดหมายประเภทนี้จะไม่พบความแตกต่างระหว่างผู้รับที่อยู่ใน Smtp Header และ

Rfc822 Header เพราะเป็นแบบ 1:1 ดังนั้นระบบจะวิเคราะห์ว่าเป็นจดหมายแบบธรรมดาเมื่อมีการจดหมายที่กระจายจากเมลลิงลิสต์กลับมา ทำให้ระบบลดค่าความมั่นใจลงเนื่องจากผลการวิเคราะห์ขัดแย้งกัน ทำให้ผลลัพธ์สุดท้ายที่ได้เกิดความผิดพลาดขึ้น

จากผลการทดลอง โดยการจำลองสถานการณ์เพื่อศึกษาผลกระทบต่างๆที่เกิดขึ้นกับการตั้งค่าความมั่นใจเริ่มต้น หรือ Initial Confidential Factor และ เกณฑ์ค่าความมั่นใจของระบบ หรือ Threshold Confidential Factor ของระบบตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ สามารถวัดค่าของผลกระทบในเรื่องของ %Correctness หรือ ค่าความถูกต้อง รวมถึง ค่าความผิดพลาดซึ่งสามารถบ่งออกเป็น 2 ชนิดคือ %Positive Error และ %Negative Error โดยค่าความผิดพลาดชนิดแรกคือความผิดพลาดที่ระบบไม่สามารถตรวจจับเมลลิงลิสต์ได้ ส่วนชนิดที่สองนั้นคือความผิดพลาดที่ระบบตรวจจับไม่ถูกต้อง ซึ่งจากผลกระทบที่เกิดขึ้นสามารถสรุปได้ดังต่อไปนี้

1. สำหรับ CF_{IN} นั้นพบว่าเมื่อกำหนดให้มีค่าในช่วงต่ำแล้วผลที่เกิดขึ้น คือ %Correctness จะมีค่าสูงที่สุดและมีค่าความผิดพลาด คือ %Negative Error น้อยที่สุด แต่ %Positive Error จะมีค่าสูง หมายถึงข้อมูลเมลลิงลิสต์ที่ตรวจจับได้จะมีค่าความถูกต้องสูง มีข้อมูลเมลลิงลิสต์ที่ตรวจจับไม่ถูกต้องน้อย และจำนวนเมลลิงลิสต์ที่ตรวจจับไม่พบจะมีจำนวนมาก

2. ถ้ากำหนดให้ CF_{TH} มีค่าสูงแล้ว %Correctness หรือค่าความถูกต้องจะมีค่าสูงตามไปด้วย นอกจากนั้นยังทำให้ค่า %Positive Error สูงที่สุด แต่สำหรับค่า %Negative Error จะมีค่าต่ำที่สุด นั่นหมายถึงจำนวนเมลลิงลิสต์ที่ตรวจจับได้จะมีจำนวนน้อยกว่าในกรณีที่กำหนดให้ค่า CF_{TH} มีค่าต่ำๆ แต่เมลลิงลิสต์ที่ตรวจจับได้จะมีความถูกต้องสูงและมีความผิดพลาดน้อยกว่า

3. สำหรับจำนวนจดหมายอิเล็กทรอนิกส์ที่ใช้ในการตรวจจับนั้น ถ้ามีจำนวนเพิ่มมากขึ้นจะเป็นผลให้ %Correctness หรือค่าความถูกต้องมีค่าสูงเพิ่มขึ้น ส่วน %Negative Error หรือความผิดพลาดที่ระบบตรวจจับไม่ถูกต้องนั้นจะมีค่าลดลง และ %Positive Error หรือค่าความผิดพลาดที่ระบบตรวจจับเมลลิงลิสต์ไม่พบนั้น จะมีแนวโน้มลดลงนั่นคือจำนวนเมลลิงลิสต์ที่ตรวจจับพบมีจำนวนมากขึ้น

จากข้อสรุปด้านบนพบว่าค่าที่เหมาะสมคือ CF_{IN} ควรอยู่ในช่วง 0.1 – 0.3 และ CF_{TH} ควรอยู่ในช่วง 0.7 – 0.9 แต่เมื่อพิจารณาร่วมกับผลกระทบที่เกิดจากจำนวนจดหมายอิเล็กทรอนิกส์ที่ใช้ในการตรวจจับแล้วพบว่า ที่ค่า CF_{IN} ต่ำจะใช้จำนวนจดหมายมากกว่าค่าสูง ดังนั้นค่าที่เหมาะสมคือ CF_{IN} = 0.3 ซึ่งจะใช้จำนวนจดหมายน้อยกว่า 0.1 และ 0.2 ส่วนที่ค่า CF_{TH} ต่ำจะใช้จำนวนจดหมายน้อยกว่าค่าสูง ดังนั้นค่าที่เหมาะสมคือ CF_{TH} = 0.8 ซึ่งจะใช้จำนวนจดหมายน้อยกว่า 0.9 แต่จะเท่ากันกับ 0.7 ซึ่งค่า CF_{TH} = 0.8 จะเหมาะสมกว่าเพราะเมื่อพิจารณาค่าความถูกต้องจะมีสูงกว่า และมีความผิดพลาดน้อยกว่า

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

จากปัญหาที่เกิดขึ้นในการจัดการแยกแยะและแจกแจงประเภท ของจดหมายอิเล็กทรอนิกส์ที่ผ่านเข้าออกเครือข่ายจำนวนมากในแต่ละวันของผู้ดูแลเครือข่าย เพื่อตรวจสอบข้อมูลของจดหมายที่เกิดจากการทำงานของเมลลิงลิสต์ให้กับระบบเมลลิงลิสต์ย่อย หรือ Submailing List เพื่อจัดการแก้ปัญหาความซ้ำซ้อนของข้อมูลที่จัดเก็บอยู่ใน Mail Box ของ User ภายในเครือข่ายนั้น

งานวิจัยนี้ จึงมีความต้องการนำเสนอระบบการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์เพื่อแก้ปัญหานี้ โดยเริ่มจากการศึกษาลักษณะการทำงานของจดหมายจากเมลลิงลิสต์และค้นหาความแตกต่างระหว่างจดหมายจากเมลลิงลิสต์และจดหมายแบบธรรมดาทั่วๆไป เพื่อนำไปสู่การพัฒนาระบบการตรวจจับเมลลิงลิสต์ โดยระบบจะเริ่มจากการ Capture ข้อมูลจดหมายอิเล็กทรอนิกส์ที่ผ่านเข้าและออกจากเครือข่าย จากนั้นจัดการให้ข้อมูลอยู่ใน Detection Format เพื่อเข้าสู่ระบบการตรวจจับข้อมูล เมื่อระบบตรวจพบจดหมายที่มีลักษณะเฉพาะของเมลลิงลิสต์แล้ว จะกำหนดค่า Initial Confidential Factor เพื่อกำหนดค่าความมั่นใจเริ่มต้นให้กับข้อมูล จากนั้นผ่านเข้าสู่กระบวนการปรับเปลี่ยนค่าความมั่นใจ โดยระบบจะทำการปรับเปลี่ยนค่าความมั่นใจเพิ่มขึ้นเมื่อมีข้อมูลอื่นๆมาสนับสนุน และจะปรับเปลี่ยนค่าความมั่นใจลดลงถ้ามีข้อมูลอื่นๆมาขัดแย้ง เมื่อข้อมูลใดๆที่ตรวจจับได้มีค่าความมั่นใจมากกว่า Threshold Confidential Factor ซึ่งคือเกณฑ์ค่าความมั่นใจในการคัดเลือกข้อมูลเมลลิงลิสต์เพื่อจะรายงานให้ผู้ดูแลระบบทราบและนำไปจัดการ config ระบบเมลลิงลิสต์ย่อย โดยข้อมูลที่สรุปได้ประกอบด้วย Mail address ของเมลลิงลิสต์ ผู้จัดการเมลลิงลิสต์ รวมถึงสมาชิกของเมลลิงลิสต์ที่อยู่ภายในเครือข่าย และชนิดของ Mailing List Software ที่ใช้ในการจัดการเมลลิงลิสต์นั้นๆ

จากการทดลองนั้นมีการวัดค่าตัวแปร 3 ตัวแปรด้วยกันคือ %Correctness, %Positive Error และ %Negative Error โดยค่า %Correctness คือเปอร์เซ็นต์ค่าความถูกต้องที่ระบบตรวจจับเมลลิงลิสต์ได้ ส่วน %Positive Error คือเปอร์เซ็นต์ค่าความผิดพลาดที่ระบบตรวจจับเมลลิงลิสต์ไม่พบ และสุดท้าย %Negative Error คือเปอร์เซ็นต์ค่าความผิดพลาดที่ระบบตรวจจับผิดพลาด ซึ่งสามารถสรุปได้ว่า ถ้ากำหนดค่า Initial Confidential Factor สูง และค่า Threshold Confidential Factor ที่มีค่าต่ำ พบว่า %Positive Error มีค่าต่ำหมายถึงจำนวนเมลลิงลิสต์ที่ตรวจจับได้มีจำนวนสูงกว่า การกำหนดค่า Initial Confidential Factor ต่ำ และค่า Threshold Confidential Factor ที่มีค่าสูง อย่างไรก็ตามในระบบการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์นั้น ต้องการข้อมูลที่มีความถูกต้องแม่นยำในการตรวจจับเมลลิงลิสต์ จึงเป็นผลให้ต้องเลือกค่า Initial Confidential Factor ต่ำ และค่า

Threshold Confidential Factor ที่มีค่าสูงเพื่อให้มีความถูกต้องหรือ %Correctness สูงที่สุด และมีความผิดพลาด หรือ %Negative Error น้อย จากการทดลองจึงสรุปว่าค่า Initial Confidential Factor ควรมีค่าเท่ากับ 0.3 และค่า Threshold Confidential Factor ควรมีค่าเท่ากับ 0.8 ซึ่งสามารถนำไปใช้ในกรณีที่เมลลิงลิสต์มีการจัดการ โดยใช้ Majordomo

5.2 ประโยชน์ของงานวิจัยและข้อเสนอแนะ

เนื่องจากในวันหนึ่งๆ ในแต่ละเครือข่ายจะมีปริมาณจดหมายอิเล็กทรอนิกส์จำนวนมาก ถ้าปล่อยให้เป็นหน้าที่ของผู้ดูแลเครือข่าย หรือ Network Administrator ในการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ จะทำให้เสียเวลาในการจัดการดูแลรักษาเครือข่ายและอาจทำให้เกิดความผิดพลาดเนื่องจากการเกิดขึ้นใหม่ของเมลลิงลิสต์ เมื่อนำระบบการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ หรือ Mailing List Detection from Electronic Mail System เข้ามาใช้ ทำให้สามารถตรวจจับเมลลิงลิสต์และแจกแจงรายละเอียดเช่น List address, Manager Address และ Member address รวมถึง Mailing List Software ที่ใช้ในการจัดการเมลลิงลิสต์ ได้ออกมามีความถูกต้องและแม่นยำ อีกทั้งยังสามารถตรวจจับเมลลิงลิสต์ที่เกิดขึ้นใหม่ได้ ทำให้ประสิทธิภาพในการแก้ปัญหาความซ้ำซ้อนของข้อมูลจดหมายที่จัดเก็บอยู่ใน Mail Box ของ User ที่เป็นสมาชิกของเมลลิงลิสต์มีความสะดวกและรวดเร็วมากกว่าเดิม

จากการทำการทดลองพบว่าระบบการตรวจจับเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ นั้นสามารถตรวจจับเมลลิงลิสต์ได้จริง และมีความถูกต้องแม่นยำ แต่มีการใช้ Mailing List Software ในการทดลองคือ Majordomo เพียงตัวเดียว เนื่องจากเป็น Free Software และใช้งานง่ายทั้งยังเป็นที่นิยมใช้ในการจัดการเมลลิงลิสต์ทั่วๆไป จึงเป็นข้อเสนอแนะในการทำการทดลองระบบกับเมลลิงลิสต์ที่ใช้ Software ชนิดอื่นในการจัดการ ซึ่งเมื่อทำการทดลองกับ Software อีกหลายๆชนิดแล้วอาจจะได้ค่า Initial Confidential Factor และค่า Threshold Confidential Factor ที่มีความเหมาะสมมากกว่าที่สรุปไว้ข้างต้น ซึ่งวิธีการสรุปค่าที่เหมาะสมคือ สมควรเลือกค่าที่ทำให้ระบบสามารถตรวจจับเมลลิงลิสต์ได้ถูกต้องมากที่สุด และมีความผิดพลาดที่เกิดจากระบบตรวจจับไม่ถูกต้องต่ำที่สุด นั่นคือ มีค่า %Correctness สูง และ %Negative Error ต่ำ ส่วน %Positive Error นั้นควรเลือกที่มีค่าต่ำ นั่นหมายถึงจำนวนเมลลิงลิสต์มีจำนวนมาก แต่ต้องพิจารณาค่าความถูกต้องเป็นหลัก นอกจากนี้ยังต้องพิจารณาร่วมกับจำนวนจดหมายที่ใช้ตรวจจับด้วย นั่นคือควรเลือกค่าที่ใช้จำนวนจดหมายในการตรวจจับน้อย หมายถึงใช้เวลาน้อยนั่นเอง

นอกจากนี้ระบบควรมีการปรับปรุง เกี่ยวกับขั้นตอนการปรับเปลี่ยนค่าความมั่นใจ (Confidential Factor Algorithm) ในกรณีที่พบข้อมูลที่มีความเป็นไปได้สูง เช่น ข้อมูลจดหมายสมัครสมาชิก เป็นต้น ควรมีการปรับเปลี่ยนค่าความมั่นใจที่สูงกว่าข้อมูลแบบอื่นๆ อาจแก้ไขด้วย

การสร้างเป็นเงื่อนไขของขั้นตอนนี้ เพื่อให้ได้ข้อมูลที่ถูกต้องได้เร็วมากขึ้น หรือใช้จำนวนจดหมายในการตรวจจับน้อยลง เพื่อปรับปรุงและพัฒนาระบบให้มีประสิทธิภาพมากขึ้น

เอกสารอ้างอิง

[1] A. Schwartz. "**Managing Mailing List**" 1st O'reilly & Associates. Inc March 1996.

[2] D. H. Crocker. "**Standard for The Format of Arpa Internet Text Message**" Request for Comments 822. Dept. of Engineering University of Delaware. August 13, 1982.

[3] S. Chomjan and A. Khunkitti. "**A New Hierachical Based Approach Mailing List System**" Proceedings of The 2003 Internation Conference on Informatioin and Communication Technologies (ICT2003) Bang na Campus, Assumption University. April 8-10, 2003.

[4] J. B. Postel. "**Simple Mail Transfer Protocol**" Request for Comments 821. Information Sciences Institute, Southern California University. August 1982.

[5] G. Neufeld. "**The Use of URLs as Meta-Syntax for Core Mail List Commands and their Transport through Message Header Fields**" Request for Comments 2369. Nisto. July 1998.

[6] C. Liu, J. Peek, R. Jones, B. Buus and A. Nye. "**Managing Internet Information Services**" 1st O'reilly & Associates. Inc December 1994.

[7] Red Hat, Inc. "**Red Hat Network 3.7 Reference Guide**", 2005
[Online].Available : http://www.redhat.com/docs/manuals/RHNetwork

[8] V. D. Skahan. "**Majordomo Frequency Ask Question**", October 2001
[Online].Available : http://www.greatcircle.com/majordomo/majordomo-faq.html

[9] The C++ Resource Network. "**Reference**", 2000-2003
[Online].Available : http://www.cplusplus.com/Reference

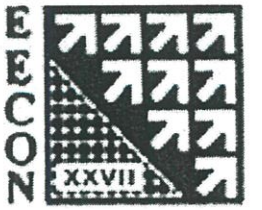
ภาคผนวก ก.
บทความทางวิชาการที่ได้รับการตีพิมพ์

1. ณัฐติญา ไข่ติยากุล และ อัครินทร์ คุณกิตติ. "การรู้จำคุณลักษณะของเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ (Mailing List Characteristic Recognition from E-mail)." การประชุมวิชาการทางวิศวกรรมไฟฟ้า ครั้งที่ 27 (EECON-27) 11-12 พฤศจิกายน 2547 มหาวิทยาลัยขอนแก่น

2. N. Khaitiyakun and A. Khunkitti. "Mailing List Characteristic from Electronic Mail" International Conference on Control, Automation and System (ICCAS2004). August 25-27, 2004 The Shangri-La Hotel, Bangkok, Thailand

การประชุมวิชาการทางวิศวกรรมไฟฟ้า ครั้งที่ 27

27th Electrical Engineering Conference (EECON 27)



บทความทางวิศวกรทางวิศวกรรมไฟฟ้า

ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น

Back to content

Back to main

การรู้จำคุณลักษณะของเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์

Mailing List Characteristic Recognition from E-mail

ณัฐพิญา ไข่เดียวกุล และ อัครินทร์ คุณสิทศิ
คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
3 หมู่ 2 ถนนฉลองกรุง แขวงลำปะทิว เขตลาดกระบัง กรุงเทพมหานคร 10520
โทร.0-1830-3114 โทรสาร 0-2956-9479 E-mail: hanuhana@yahoo.com,akharin@it.kmitl.ac.th

บทคัดย่อ

ปัญหาส่วนใหญ่ที่พบในการจราจรบนเครือข่ายอินเตอร์เน็ตนั้นเกิดจากความหนาแน่นของปริมาณข้อมูลจดหมายอิเล็กทรอนิกส์ ส่วนหนึ่งมาจากการทำงานของเมลลิงลิสต์จากปัญหาดังกล่าวระบบเมลลิงลิสต์ย่อยจึงถูกสร้างขึ้นเพื่อลดจำนวนของจดหมายที่ถูกส่งมายังเครือข่าย โดย network administrator มีหน้าที่ในการแจกแจงประเภทของจดหมายซึ่งอาจมีความผิดพลาดเกิดขึ้นเนื่องจากปริมาณของจดหมายที่เข้ามาในแต่ละวัน บทความนี้ต้องการนำเสนอระบบการรู้จำคุณลักษณะและพฤติกรรมของเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ เพื่อวิเคราะห์หาข้อมูลที่ใช้ในการ config ระบบเมลลิงลิสต์ย่อยได้โดยอัตโนมัติ เริ่มจากการดักจับข้อมูล (capture process) จากนั้นเข้าสู่กระบวนการจัดเตรียมข้อมูล (preprocessing process) เพื่อให้ข้อมูลอยู่ในรูปแบบเดียวกัน (recognized format) ขั้นตอนต่อไปคือกระบวนการรู้จำ (recognition process) โดยการเปรียบเทียบกับลักษณะของจดหมายทั่วๆไป และคำนวณหาค่าความมั่นใจ (confidential factor) เพื่อใช้พิจารณาตัดสินใจความถูกต้องของข้อมูล ในขั้นตอนสุดท้าย (postprocessing process) ก่อนรายงานผลให้เมลลิงลิสต์ย่อยดำเนินการจัดการต่อไป

คำสำคัญ : เมลลิงลิสต์, เมลลิงลิสต์ย่อย, ระบบการรู้จำ, ค่าความมั่นใจ

Abstract

The most of network traffic problems came from mailing list mail. That could solve by submailing list system, which reduce number of mails transfer between networks. Analyses of mailing list characteristic in electronic mail are a feature of submailing list system, which handle by network administrators. In day time there were many mails in network traffic so network administrator could not review every mails in due time. This article will present ideas and recognize methodology of recognition system for automatic working in submailing list system. Recognize step begin with capture process, which use to trap e-mail information from transfer channel. Next process is preparing raw data into recognized format. Then the third one is recognize part and find out confidential factor. The last process is make decision and determine which electronic mail has properties of mailing list characteristic. Afterward deliver result to submailing list for carry on.

Keyword : Mailing list, Submailing list, Recognition system, Confidential factor

CP08

1.บทนำ

เมลลิงลิสต์ [1] คือการรวบรวมกลุ่มคนที่มีความสนใจในเรื่องเดียวกันต้องการแลกเปลี่ยนข้อมูลและความคิดเห็นผ่านทางจดหมายอิเล็กทรอนิกส์ [2] เมื่อมีสมาชิกส่งข้อมูลมาที่ list address จากนั้น mailing list manager จะจัดการทำสำเนาของจดหมายฉบับนั้นแล้วกระจายให้สมาชิก (subscriber) ที่มีอยู่ในบัญชีรายชื่อโดยไม่สนใจว่าสมาชิกจะอยู่ในเครือข่ายเดียวกันหรือใกล้เคียงกัน จึงเกิดความซ้ำซ้อนของข้อมูลจำนวนมากอยู่บนเส้นทางการจราจรระหว่างเครือข่าย

ปัญหาที่เกิดขึ้นจากการทำงานของเมลลิงลิสต์นั้นแนวทางการแก้ไขคือพัฒนาระบบเมลลิงลิสต์ย่อย [3] เพื่อลดความซ้ำซ้อนด้วยการรวบรวมกลุ่มสมาชิกของเมลลิงลิสต์หลักตามความใกล้เคียงกันของเครือข่าย จากนั้นสร้างระบบเมลลิงลิสต์ย่อยเพื่อเป็นตัวแทนรับจดหมายจากเมลลิงลิสต์หลัก เป็นการลดปริมาณความหนาแน่นของข้อมูลซึ่งปัญหาที่ตามมาคือจะทราบได้อย่างไรว่า e-mail address ใดเป็นของเมลลิงลิสต์เพื่อที่จะกำหนดค่าให้กับระบบเมลลิงลิสต์ย่อยทำงานได้ถูกต้องซึ่งในแบบเดิมต้องใช้ network administrator ในการตรวจสอบและกำหนดค่าต่างๆทำให้เกิดความไม่สะดวกเนื่องจากปริมาณของจดหมายที่มีเป็นจำนวนมากและมีเมลลิงลิสต์เกิดขึ้นใหม่เรื่อยๆทำให้อาจเกิดความผิดพลาดในการตรวจสอบ เป็นสาเหตุให้เมลลิงลิสต์ย่อยไม่สามารถแก้ไขปัญหาได้ บทความนี้จึงนำเสนอแนวความคิดการสร้างระบบการรู้จำคุณลักษณะของเมลลิงลิสต์จากจดหมายอิเล็กทรอนิกส์ (Mailing list recognition system) เพื่อตรวจสอบและรายงานผลให้เมลลิงลิสต์ย่อยได้โดยอัตโนมัติเป็นการแบ่งเบาภาระให้ network administrator

จุดมุ่งหมายของงานวิจัยนี้ต้องการวิเคราะห์จดหมายอิเล็กทรอนิกส์เพื่อหา mail address ที่เป็นของเมลลิงลิสต์และสมาชิก จากนั้นส่งค่าการ

ทำงานต่างๆที่เมลลิงลิสต์ย่อยต้องการในการ config ระบบโดยอัตโนมัติ ซึ่งการทำงานของ Mailing list recognition system นี้ใช้การพิจารณา confidential factor ในการตัดสินใจว่า mail address ใดมีคุณสมบัติของเมลลิงลิสต์โดย confidential factor ได้จากการคำนวณในกระบวนการรู้จำจดหมายอิเล็กทรอนิกส์

ในหัวข้อถัดไปจะกล่าวถึงลักษณะการทำงานและพฤติกรรมของกลุ่มเมลลิงลิสต์ เพื่อเปรียบเทียบกับลักษณะของจดหมายอิเล็กทรอนิกส์ทั่วไป จากนั้นจะเป็นแนวคิดของงานวิจัยและขั้นตอนการทำงานของระบบการรู้จำคุณลักษณะของเมลลิงลิสต์ แนวทางการวิจัยในขั้นตอนต่อไป สุดท้ายสรุปเกี่ยวกับการทำงานของระบบ Mailing list recognition ประโยชน์ที่ได้และงานอื่นๆที่สามารถนำหลักการทำงานของระบบไปพัฒนาปรับปรุงต่อไปได้

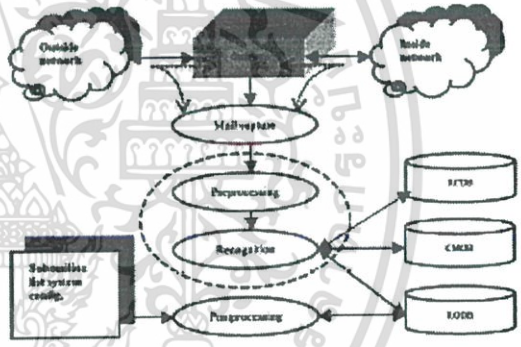
2.เมลลิงลิสต์

เมลลิงลิสต์สามารถแบ่งตามลักษณะการจัดการได้ 2 ประเภท คือ manually management และ software management ในแบบแรกนั้นเหมาะสำหรับเมลลิงลิสต์ที่มีขนาดเล็กมีจำนวนสมาชิกไม่มาก มีการเปลี่ยนแปลงเข้าและออกจากการเป็นสมาชิกไม่บ่อยนัก ส่วนแบบที่สองนั้นพบมากในเมลลิงลิสต์ที่มีขนาดใหญ่มีจำนวนสมาชิกมาก มีการเปลี่ยนแปลงข้อมูลของสมาชิกบ่อย software ที่นิยมใช้คือ Majordomo, Listserv, Listproc ความสามารถของ software เหล่านี้คือใช้ควบคุมระบบสมาชิก เช่น การสมัครสมาชิก การยกเลิกสมาชิก นอกจากนี้ยังควบคุมการกระจายของจดหมายอิเล็กทรอนิกส์ให้กับสมาชิกอีกด้วย จากการสังเกตพบว่าการจัดการเมลลิงลิสต์โดยใช้ software นั้นจะก่อให้เกิดปัญหาสูงกว่าแบบแรกเพราะมีจำนวนสมาชิกมากทำให้มีปริมาณการแลกเปลี่ยนข้อมูลและการแสดงความคิดเห็นสูงตามไปด้วย นอกจากนี้พฤติกรรมของจดหมายอิเล็กทรอนิกส์จากเมลลิงลิสต์ที่แตกต่างจากจดหมายอิเล็กทรอนิกส์ทั่วๆไปคือ

1. มีการใช้งาน mailing list command เพื่อสมัครสมาชิก ยกเลิกสมาชิก หรือเปลี่ยนแปลงข้อมูลอื่นๆ ในกรณีของ software manager นั้นไม่สามารถสื่อสารโดยใช้ภาษาที่ใช้ในชีวิตประจำวันได้ จำเป็นต้องใช้ mailing list command เพื่อออกคำสั่งให้ manager จัดการตามคำร้องขอของสมาชิก เช่น subscriber command เป็นคำสั่งเพื่อการสมัครสมาชิกเป็นต้น
2. มีการป้องกันสมาชิกจากจดหมายอิเล็กทรอนิกส์ที่ผิดประเภท เช่น จดหมายรายงานความผิดพลาดจากการสื่อสาร เมื่อสมาชิกส่งข้อความมาที่ list address จากนั้น mailing list manager จะกระจายจดหมายฉบับนั้นไปยังสมาชิกทุกๆคนโดยเพิ่มเติมและเปลี่ยนแปลง header ของจดหมาย เช่น Return-Path, Reply-To, Sender ให้ใส่ e-mail address ของเมลลิงลิสต์เอง ดังนั้นเมื่อมีความผิดพลาดเกิดขึ้น รายงานดังกล่าวจะถูกส่งไปยังเมลลิงลิสต์เท่านั้น
3. มีการใช้งาน mailing list header เช่น List-Post, List-Subscribe เป็นต้น มีไว้เพื่อระบุรายละเอียดเกี่ยวกับเมลลิงลิสต์นั้นๆ
4. เมื่อตรวจสอบพบว่า e-mail address ใน To, Cc header นั้นสามารถกระจายออกเป็น e-mail address อื่นๆได้อีก เช่น การใช้งาน aliases function ใน sendmail เป็นต้น โดยตรวจสอบจาก address ของผู้รับที่พบใน SMTP command ว่าตรงกับ address ของผู้รับที่อยู่ใน mail header หรือไม่

เมื่อได้ข้อสรุปเกี่ยวกับพฤติกรรมของจดหมายอิเล็กทรอนิกส์แบบเมลลิงลิสต์แล้วในหัวข้อถัดไปจะกล่าวถึงแนวทางการสร้างระบบการรู้จำคุณลักษณะจดหมายอิเล็กทรอนิกส์ของเมลลิงลิสต์ โดยเริ่มจากขั้นตอนการตรวจสอบจดหมายที่ผ่านเข้าและออกจากเครือข่าย รวมถึงขั้นตอนการวิเคราะห์เพื่อหาพฤติกรรมของเมลลิงลิสต์ จากนั้นแสดงวิธีการคำนวณหาค่าความมั่นใจ หรือ confidential factor เพื่อใช้ประกอบการตัดสินใจ และในกระบวนการสุดท้ายคือการคัดเลือกข้อมูลเพื่อรายงานผลให้ระบบเมลลิงลิสต์ย่อย

3.แนวคิดและขั้นตอนการทำงาน



รูปที่1 Mailing list recognition system

ขั้นตอนการสร้างระบบการรู้จำเริ่มจากการดักจับข้อมูลของจดหมายอิเล็กทรอนิกส์ที่ผ่านเข้าและออกจากเครือข่าย (Capture process) เมื่อได้ข้อมูลมาแล้วจากนั้นเข้าสู่กระบวนการถัดไปคือกระบวนการจัดเตรียมข้อมูล (Preprocessing process) เพื่อจัดเตรียมข้อมูลของจดหมายอิเล็กทรอนิกส์ให้อยู่ในรูปแบบเดียวกัน สาเหตุที่ไม่สามารถนำข้อมูลที่ได้ไปวิเคราะห์ได้เลยเป็นเพราะว่า ในเครือข่ายอินเตอร์เน็ตนั้นมีการใช้งาน MTA อยู่หลายชนิดซึ่งอาจมีข้อมูลบางส่วนที่แตกต่างกันจึงมีความจำเป็นต้องจัดการให้อยู่ในรูปแบบที่ระบบต้องการเสียก่อน จากนั้นเข้าสู่กระบวนการรู้จำ (Recognition process) เพื่อวิเคราะห์หาพฤติกรรมของเมลลิงลิสต์ และคำนวณหาค่าความมั่นใจ (Confidential factor) สุดท้ายเป็นกระบวนการตัดสินใจ (Postprocessing process) พิจารณาคัดเลือกข้อมูลที่มีคุณลักษณะของเมลลิงลิสต์ แล้วรายงานผลให้เมลลิงลิสต์ย่อยดำเนินการจัดการต่อไป

3.1 Capture process

เริ่มต้นด้วยกระบวนการดักจับข้อมูลของจดหมายอิเล็กทรอนิกส์ที่ผ่านเข้าและออกจากเครือข่ายพบว่าการแลกเปลี่ยนข้อมูลของจดหมายอิเล็กทรอนิกส์นั้นใช้โปรโตคอล SMTP ในการส่งและรับข้อมูลดังนั้นจึงต้องดักจับข้อมูลที่พอร์ต 25 ซึ่งเป็นพอร์ตในการสื่อสารของโปรโตคอล SMTP ในการดักจับข้อมูลนั้นเป็นการคัดลอกจดหมายอิเล็กทรอนิกส์ที่ผ่านเข้าออกเครือข่ายโดยไม่มีการเปลี่ยนแปลงข้อมูลของจดหมาย พบว่าข้อมูลที่ดักจับได้นั้นแสดงการแลกเปลี่ยนข้อมูลของผู้รับและผู้ส่งจดหมายอิเล็กทรอนิกส์โดย MTA ของทั้ง 2 ฝั่ง สามารถแบ่งข้อมูลออกเป็น 2 ส่วนคือในส่วนแรก smtp envelop คือ MTA conversation ประกอบด้วย smtp command และ smtp reply code ในส่วนที่ 2 smtp content คือ จดหมายอิเล็กทรอนิกส์ที่ต้องการส่งประกอบด้วย e-mail header และ mail body ข้อมูลที่ดักจับได้นี้มีรูปแบบเป็น text format จากนั้นเริ่มเข้าสู่กระบวนการจัดเตรียมข้อมูลต่อไป

3.2 Preprocessing process

ในขั้นตอนนี้เป็นการจัดเตรียมข้อมูลให้อยู่ในรูปแบบเดียวกันคือ recognized format มีรูปแบบดังตารางต่อไปนี้

ตารางที่1 Recognized variable format

Variable	Format	Type
Smtp-mail-from	<name@domain>	Single
Smtp-recipient	<name@domain>	Multi
Rfc822-from	<name@domain>	Single
Rfc822-sender	<name@domain>	Single
Rfc822-reply-to	<name@domain>	Multi
Rfc822-return-path	<name@domain>	Single
Rfc822-recipient	<name@domain>	Multi
Rfc822-error-to	<name@domain>	Single
Rfc822-subject	<text>	Multi
Rfc822-body	<text>	Multi

โดยที่ recognized variable เหล่านี้ประกอบด้วยข้อมูลที่นำมาใช้ในขั้นตอนการรู้จำและข้อมูลที่จำเป็นในการตั้งค่าของเมลลิงลิสต์ย่อย ส่วนใหญ่ข้อมูลที่นำมาใช้จะเป็น mail address ซึ่งจัดอยู่ในรูปแบบดังตารางที่1 ส่วน Type หมายถึงจำนวนสมาชิกของ variable ที่สามารถเก็บไว้ได้

3.3 Recognition process

ก่อนที่จะเข้าสู่กระบวนการการรู้จำนั้นต้องทราบว่า output หรือข้อมูลที่ระบบเมลลิงลิสต์ย่อยต้องนำไปใช้ในการจัดการระบบนั้นประกอบด้วยข้อมูลใดบ้างซึ่งสามารถสรุปได้ดังต่อไปนี้

Recognized output database (RODB) เป็นฐานข้อมูลของผลลัพธ์ที่ได้จากการรู้จำ

ตารางที่2 List address table (LA table)

List address	CF_LA
Listname@listdomain	ค่าคงที่มีค่าระหว่าง 0-1

ตารางที่3 List Member Address table (LM table)

List address	Member address	CF_LM
Listname@listdomain	Membername@memberdomain	ค่าคงที่มีค่าระหว่าง 0-1

ตารางที่4 List Manager Address table (MA table)

List address	Manager address	CF_MA
Listname@listdomain	Managername@managerdomain	ค่าคงที่มีค่าระหว่าง 0-1

ตารางที่5 List Manager Type table (MT table)

Manager address	Manager type	CF_MT
Managername@managerdomain	Manual, listserv, majordomo, listproc, smartlist, N/A	ค่าคงที่มีค่าระหว่าง 0-1

นอกจาก Recognized output database (RODB) แล้วยังมี Consequent mail database (CMDB) ก็คือฐานข้อมูลสำหรับหักข้อมูล mail recognition ที่มีผลต่อเนื่องกันหรือ mail ที่ยังไม่แน่ใจต้องใช้ข้อมูลอื่นมายืนยัน เช่น จดหมายที่ตรวจสอบพบ mailing list command ซึ่งต้องการ confirm mail เพื่อยืนยันความต้องการของ subscriber หรือ reply mail เพื่อตอบรับความต้องการของสมาชิก เป็นต้น ประกอบด้วย SLA, SLM, SMA และ SMT table โดยแต่ละตารางประกอบด้วยค่าความมั่นใจคือ CF_SLA, CF_SLM, CF_SMA, CF_SMT เมื่อค่าความมั่นใจมีค่ามากกว่า $CF_{THRESHOLD}$ จึงจัดการย้ายข้อมูลไปเก็บที่ RODB เพื่อรายงานให้เมลลิงลิสต์ย่อยทราบต่อไป

ตารางที่6 List Manager Command table

Type ID	Type name	Command format ID list	Command part
Sequence ID	Listserv, majordomo, listproc, manual, N/A	(<keyword>, <pattern>)	Bit 0 – Rfc822-body Bit 1 – Rfc822-subject

สุดท้ายตารางที่ 6 คือ Recognition configuration database (RCDB) คือฐานข้อมูลแสดงลักษณะและข้อมูลต่างๆของ mailing list manager เพื่อการวิเคราะห์ข้อมูลในขั้นตอนนี้

ในรูปที่ 2 recognition algorithm แสดงการทำงานของกระบวนการรู้จำ ประกอบด้วยเงื่อนไขการเพิ่มข้อมูลในตารางต่างๆรวมถึงการแก้ไข CF value และการลบ entry ใน RODB และ CMDB ซึ่งจะมีค่าความมั่นใจเริ่มต้นคือ $CF_{INITIAL}$ เป็นค่าคงที่ที่กำหนดขึ้น ส่วนใน modify confidential factor algorithm นั้นแบ่งออกเป็น 2 ส่วน คือ Positive CF algorithm และ Negative CF algorithm ใช้ในการเปลี่ยนแปลงค่าของ CF ในกรณีแรกนั้นเมื่อมีข้อมูลเข้ามาสนับสนุนข้อมูลเดิมค่าความมั่นใจจึงเพิ่มขึ้น ส่วนในกรณีที่ 2 เมื่อไม่มีข้อมูลเข้ามาสนับสนุนหรือมีข้อมูลอื่นที่พิสูจน์ได้ว่าข้อมูลเดิมไม่เป็นจริงค่าความมั่นใจจึงลดลง

3.4 Postprocessing process

ในขั้นตอนสุดท้ายเป็นขั้นตอนการตัดสินใจคัดเลือกข้อมูลจาก CMDB ไปเก็บไว้ที่ RODB เพื่อรายงานให้เมลลิงลิสต์ย่อยทราบและใช้ในการตั้งค่าระบบเพื่อแก้ไขปัญหาที่เกิดขึ้น โดยการเปรียบเทียบค่าความมั่นใจของข้อมูลกับเกณฑ์ค่าความมั่นใจที่กำหนดค่าไว้ $CF_{THRESHOLD}$ ซึ่งได้จากการทดลองหาค่าความมั่นใจที่มีความถูกต้องมากที่สุด

CP08

Back to content

Back to main

```
if Rfc822-mailing-list-header not equal to null then
        if Rfc822-recipient in SLA table then
            increase CF_SLA and Mcount = Mcount+1
            if Rfc822-from in SLM table then
                increase CF_SLM
            else
                add member address by Rfc822-from in SLM table
                increase CF_SLM
        else
        create entry in LA table by list address = Rfc822-recipient and
        CF_LA = initial CF
        create entry in LM table by member address = Rfc822-from and
        CF_LM = initial CF
else
        if Rfc822-recipient in SLA table then
            decrease CF_SLA
if Rfc822-from not equal to (Smtp-mail-from or Rfc822-sender or Rfc822-
return-path or Rfc822-error-to or Rfc822-reply-to) then
        for (i = 1 to M) /*M is number of Rfc822-recipient*/
            for (j = 1 to N) /*N is number of Smtp-recipient*/
                if Smtp-recipient[j] == Rfc822-recipient[i] then
                    x = x+1
    if (x==0) or (x<N) then /*each Smtp-recipient not match Rfc822-
    recipient*/
            if Rfc822-recipient in SLA table then
                increase CF_SLA and Mcount = Mcount+1
                if Rfc822-from in SLM table then
                    increase CF_SLM
                else
                    add member address by Rfc822-from in SLM
    table
                    increase CF_SLM
            else
        create entry in LA table by list address = Rfc822-recipient and
        CF_LA = initial CF
        create entry in LM table by member address = Rfc822-from and
        CF_LM = initial CF
else
        if Rfc822-recipient in SLA then
            decrease CF_SLA
if (Rfc822-subject or Rfc822-body) in list manager command table then
        if (Rfc822-recipient or Rfc822-from) in SMA table then
            if Rfc822-from in SLM table then
                increase CF_SLM and CF_SMA
        else
        create entry in LA table by list address =
        listname+'@'+managerdomainbn and CF_LA = initial CF
        create entry in MA table by manager address = Rfc822-recipient
        and CF_MA = initial CF
        create entry in MT table by manager type = Rfc822-reicipient name
        or manager type and CF_MT = initial MT
else
        if (Rfc822-recipient or Rfc822-from) in SMA table then
            if Rfc822-from in SLM table then
                decrease CF_SLM and CF_SMA
```

รูปที่2 Recognition algorithm

```
if increase CF                  /*Modify confidential factor algorithm*/
        CF_x = CF_x + positive(CF_x)
If decrease CF
        CF_X = CF_x - negative(CF_x)

if receive CF_x then            /*Positive CF algorithm*/
        positive CF = (1 – CF_x)/2
        return positive CF
if receive CF_x then            /*Negative CF algorithm*/
        negative CF = CF_x / 2
        return negative CF
```

รูปที่3 Modify confidential factor algorithm

การทดลองเบื้องต้นโดยสร้างระบบ Mail server และ Mailing list เพื่อทดลองและศึกษาพฤติกรรมการทำงานของเมตลิงลิสต์ได้ผลการทดลองดังรูปที่ 4 ผลที่ได้มีความถูกต้องเพราะข้อมูลดิบเกิดจากการกำหนดค่าซึ่งในความเป็นจริงข้อมูลอาจมีความซับซ้อนและมีรูปแบบที่แตกต่างไปขึ้นอยู่กับประเภทของ MLM สำหรับค่า $CF_{INITIAL}$ เกิดจากการกำหนดค่ามื่อทำการทดลองหลายๆครั้งจึงจะสามารถสรุปได้ว่าค่าที่เหมาะสมคือค่าใด

```
<MID>001</MID>
<List_addr>dipac@premium.dipac.it.kmitl.ac.th</List_addr>
<CF_SLA>x</CF_SLA>
<Time_stamp>Wed, 19 May 2004 16:07:12</Time_stamp>
<Member_addr>testuser@recognition.dipac.it.kmitl.ac.th</Member_addr>
<CF_SLM>x</CF_SLM>
<Manager_addr> majordomo@premium.dipac.it.kmitl.ac.th
</Manager_addr>
<Manager_type> majordomo</Manager_type>
<CF_SMT>x</CF_SMT>
<Status>subscribe</Status>
<Mcount>1</Mcount>
<CF_SMA>x</CF_SMA>
```

รูปที่ 4 ผลการทดลองการรู้จำจดหมายสมัครสมาชิกเมลลิงลิสต์

4.สรุป

ปัญหาที่เกิดขึ้นจากจดหมายอิเล็กทรอนิกส์ของกลุ่มเมลลิงลิสต์บนเส้นทางการจราจรระหว่างเครือข่ายนั้นมีการค้นคว้าหาทางแก้ไข โดย MLM (Mailing list manager software) บางตัวได้พยายามรวบรวมกลุ่มสมาชิกที่มีที่อยู่ใกล้เคียงกันหรือมีเครือข่ายเดียวกันเพื่อลดปริมาณของจดหมายที่ต้องส่งออกไป แต่ในเครือข่ายอินเตอร์เน็ตนั้นมีมากมายหลายเครือข่ายซึ่งบางครั้งอาจทำให้การแก้ไขปัญหาทำได้ไม่ทั่วถึง ดังนั้นการแก้ไขปัญหาโดยใช้ระบบเมลลิงลิสต์อ่อยนั้นจะทำได้ทั่วถึงมากกว่าอีกทั้งยังเป็นการจัดการที่ใช้ได้กับ MLM ทุกๆตัว

โดยส่วนใหญ่แล้วพบว่าการทำงานของระบบเมลลิงลิสต์อ่อยนั้นสามารถแบ่งได้เป็น 2 ส่วนคือการจำแนกประเภทของจดหมายอิเล็กทรอนิกส์และการตั้งค่าข้อมูลในระบบ ซึ่งหน้าที่การจำแนกประเภทจดหมายนั้นตกอยู่กับผู้จัดการเครือข่ายทำให้มีภาระนอกเหนือจากการดูแลเครือข่ายเพิ่มขึ้น ดังนั้นถ้าสามารถสร้างระบบการรู้จำจำแนกประเภทของจดหมายอิเล็กทรอนิกส์และส่งต่อข้อมูลให้ระบบเมลลิงลิสต์อ่อยได้โดยอัตโนมัติจะลดภาระของ network administrator ทั้งยังเพิ่มประสิทธิภาพในการแก้ไขปัญหาที่เกิดขึ้นเนื่องจากสามารถลดความผิดพลาดที่อาจเกิดจากการตรวจสอบไม่ทั่วถึงของผู้จัดการเครือข่าย นอกจากนั้นการสร้างระบบการรู้จำนี้ยังสามารถพัฒนาปรับปรุงให้รู้จำคุณลักษณะของจดหมายอิเล็กทรอนิกส์ที่ก่อให้เกิดปัญหาบนเส้นทางการจราจรระหว่างเครือข่ายประเภทอื่นๆ เช่น ปัญหาที่เกิดจากเมลล์ขยะ เป็นต้น

เอกสารอ้างอิง

[1] A. Schwartz, Managing Mailing List 1st ed, O'reilly & Associates. Inc, March 1996.

[2] P. Jacob, Electronic Mail, Artech House, Boston, 1995.

[3] S. Chomjan and A. Khunkitti, "A New Hierarchical Based Approach Mailing List System", Proceedings of The 2003 International Conference on Information and Communication Technologies (ICT2003), Bang na Campus, Assumption University, April 8-10, 2003.

ICCAS '04

ICCAS 2004

2004 International Conference on Control, Automation and Systems

August 25-27, 2004

The Shangri-La Hotel, Bangkok, Thailand

- Welcome Message
- Conference Organization
- Conference Information
- Sponsors
- Table of Contents
- Author Index
- Search This CD-ROM
- Exit

ICASE

http://www.kmitl.ac.th

http://www.iccas.org

Mailing List Characteristic from Electronic Mail

N. Khaitiyakun and A. Khunkitti

Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Ladkrabang , Bangkok 10520 , Thailland
Email: hanuhana@yahoo.com and akharin@it.kmitl.ac.th

Abstract: Principle of mailing list was distributed messages to all subscribers in one time. But mailing list operation has constructed a network traffic problem. Because mailing list manager distributed mails without concentrate on subscriber network. If our network has many of subscribers, there will be redundant data in traffic channel. Submailing list has purpose to reduce problems. Analyses of mailing list characteristic in electronic mail were a feature of submailing list system, which manage by human hand (Network Administrator). That will cause trouble for network traffic if Network Administrator could not seek for mailing list characteristic from e-mails in due time. This article will present ideas and recognize methodology for automatic working in submailing list system. Recognize step begin with capture process, which use to trap e-mail information from transfer channel. Next process is preparing raw data into recognition format. Then the third one is recognize part and find out confidential factor. The last process is make decision and determine which electronic mail has properties of mailing list characteristic. Afterward deliver result to submailing list for carry on.

Keywords: Mailing list system, Recognized system, Confidential Factor

1. INTRODUCTION

Mailing list [1] is a group of e-mail addresses that can all be reached by sending a single message to one address, List address. Subscribers can have discussion by sending message to the list address. Each message will be distribute to all list subscribers. First advantage of mailing list is distributing information from a central source to lots of people in once time and second is discussing a project among several participants. Third one is exchanging questions and answers with other users of a product or service. Disadvantage of mailing list is data redundancies when mailing list distributed message to subscribers who live in the same network or neighbor networks. Submailing list [2] was created to support this trouble. By convert all subscribers to submailing list member and subscribe itself to main mailing list. When main mailing list has to transmit messages, there will be only one copy has sent to our network.

Before start recognition system we have to familiar with mailing list behavior that shown below.

- There are mailing list manager commands in Subject: header or mail body.
- Mailing list manager has to protect subscribers from communication error mails. By create e-mail address for receive error report. At the same time for general e-mail, error report will send to e-mail address found in From: header.
- There are mailing list headers [3].
- Normally we could find e-mail address of receiver from Recipient header in e-mail or RCPT TO command in smtp conversation. In general case, if found e-mail address in RCPT TO command we also found the same address in Recipient header too. But in mailing list case there are irregular between recipient.

All cases above have to consider together and then calculate confidential factor (CF) for use to decide which information was suitable to configured at submailing list system. Next we mention about recognition methodology and explain recognize idea. Follows by implement and display algorithm of recognize process. Last topic is talked about conclusion and future work.

2. METHODOLOGY OF RECOGNITION SYSTEM

Recognition system was created for inspect mailing list characteristic from e-mail [4]. Which can summarize in four processes. First process was trapping e-mail data, Capture process. These are conversation between receiver and sender MTA that based on SMTP (Simple Mail Transfer Protocol) [5]. Then rearrange data to recognize format, Preprocessing process. Next recognized process has to find out mailing list character and calculate confidential factor. Last one Postprocessing process has to make decision and hand up to configure at submailing list system.

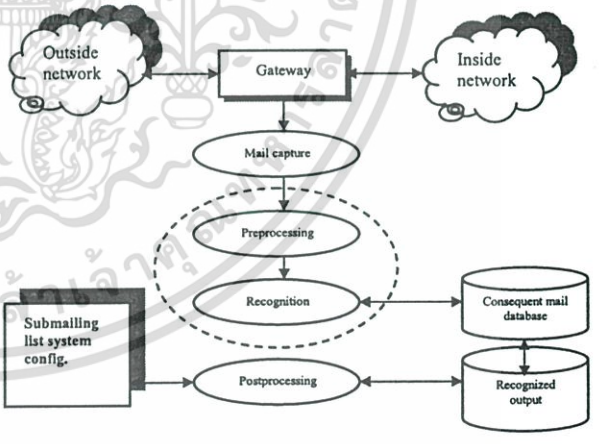


Fig. 1 Mailing list Recognition System

2.1 Capture process

The first process of recognition system is capture process, which used to capture mail information. Start process when there are some data enter SMTP port (port 25). The raw data can separate in 2 type, SMTP envelope and SMTP content. In SMTP envelope consists of SMTP command and SMTP reply code that is exchange between receiver and sender MTA. Next SMTP

content consists of Mail header and Mail body that based on standard of RFC822. All of raw data are in text format.

```
220 dipac.it.kmitl.ac.th ESMTP Sendmail 8.12.8/8.12.8; Mon, 10 Nov 2003 10:34:21 +0700 (ICT)
EHLO system.dipac.it.kmitl.ac.th
250-dipac.it.kmitl.ac.th Hello system [192.168.1.11], pleased to meet you
MAIL From:<root@system.dipac.it.kmitl.ac.th> SIZE=476
250 2.1.0 <root@system.dipac.it.kmitl.ac.th>... Sender ok
RCPT To:<hanu@dipac.it.kmitl.ac.th>
DATA
250 2.1.5 <hanu@dipac.it.kmitl.ac.th>... Recipient ok
354 Enter mail, end with "." on a line by itself
Received: from system.dipac.it.kmitl.ac.th (localhost.localdomain [127.0.0.1])
        by system.dipac.it.kmitl.ac.th (8.12.5/8.12.5) with ESMTP id hAA3Thov001035
        for <hanu@dipac.it.kmitl.ac.th>; Mon, 10 Nov 2003 10:29:44 +0700
Received: from localhost (root@localhost)
        by system.dipac.it.kmitl.ac.th (8.12.5/8.12.5/Submit) with ESMTP id hAA3ThoO001031
        for <hanu@dipac.it.kmitl.ac.th>; Mon, 10 Nov 2003 10:29:43 +0700
Date: Mon, 10 Nov 2003 10:29:43 +0700 (ICT)
From: root <root@system.dipac.it.kmitl.ac.th>
To: hanu@dipac.it.kmitl.ac.th
Subject: test
Message-ID: <Pine.LNX.4.44.0311101029240.1030-100000@system.dipac.it.kmitl.ac.th>
MIME-Version: 1.0
Content-Type: TEXT/PLAIN; charset=US-ASCII

test

.
250 2.0.0 hAA3YLmm068646 Message accepted for delivery
QUIT
221 2.0.0 dipac.it.kmitl.ac.th closing connection
```

Fig. 2 An example of capturing data from Ethereal

2.2 Preprocessing process

This step was data prepared process. Before system started recognition process, the raw material from capturing process should be rearranged in recognize format. Because there are many MTA in cyber world so data that pass capturing process may be have different layout. That was a reason we should make it in the same pattern.

Table 1 Recognize variables discover in Recognized process

Variable	Format	Type
Smtp-mail-from	<name@domain>	single
Smtp-recipient	<name@domain>	multi
Rfc822-from	<name@domain>	single
Rfc822-sender	<name@domain>	single
Rfc822-reply-to	<name@domain>	multi
Rfc822-return-path	<name@domain>	single
Rfc822-recipient	<name@domain>	multi
Rfc822-error-to	<name@domain>	single
Rfc822-subject	<text>	multi
Rfc822-body	<text>	Multi

From table 1, recognize variables could clarify in 2 types, Smtp-variable and Rfc822-variable. Smtp-variable obtains from smtp command and smtp reply code (smtp envelope). The smtp information showed about sender address and receiver address, which came from MTA communication. Rfc822-variable obtains from mail header and mail body (smtp content). The Rfc822 information presented e-mail header and mail body which came from human communication. Such as Rfc822-from which store e-mail address found in From header. Each variable, establish in variables format with angle bracket. And each of them contains data that could be single value or multi value. There are another variables that appear in preprocessing process. That use to configure in submailing list system. After this step the data was sent to next process.

2.3 Recognized process

Recognize processing is main process in mailing list recognition system. Here the system had to recognize electronic mail. That means the system find out mailing list characteristic by use observations found above for recognize part and calculates confidential factor (CF).

Before we mention about recognized algorithm, we should early present all database that use to keep information. First one was recognized output database (RODB) that use to keep result from recognized process. Recognized output database consists of 4 tables below.

Table 2 List Address table (LA table)

List address	CF_LA
Listname@listdomain	Value between 0-1

Table 3 List Member Address table (LM table)

List address	Member address	CF_LM
Listname@listdomain	Membername@memberdomain	Value between 0-1

Table 4 List Manager Address table (MA table)

List address	Manager address	CF_MA
Listname@listdomain	Managername@managerdomain	Value between 0-1

Table 5 List Manager Type table (MT table)

Manager address	Manager type	CF_MT
Managername@managerdomain	Manual, listserv, majordomo, listproc, smartlist, N/A	Value between 0-1

Confidential factor (CF) presents possibility of mailing list characteristic in each data record. There has value between zero and one. The most CF came from measured the experiment several times.

Second one was consequent mail database (CMDB). In this database has function like temporary file, which keep mailing list data that could not conclude. Because this mailing list data need more than information to prove and confirm their situation. Each record in CMDB also has CF, which shown possibility value. If we found another data that contain mailing list characteristic as same as record in CMDB, the CF will increase. On the other hand if we didn't have new information that agrees with our data for along time, the CF will decrease. Data record in CMDB could transfer to RODB by compare CMDB confidential factor with threshold value ($CF_{threshold}$).

Besides recognized output database and consequent mail database, there was recognized configuration database (RCDB). That contains information about mailing list manager characteristic such as mailing list manager type, mailing list manager command, mailing list command format etc. The recognized configuration database consists of 3 tables; list command manager table, command keyword table and command pattern table. Next process we will talk about determination process.

Table 6 Suspected List Address table (SLA table)

List address	CF_SLA	Time stamp	Status	Mcount	MID
Listname@list domain	Value between 0-1	Arrive time	<subscribe>, <unsubscribe>, <general>	Counter	Mail sequence ID

Table 7 Suspected List Member Address table (SLM table)

List address	Member address	CF_SLM	Time stamp	Status	Mcount	MID
Listname@listdomain	Membername @memberdomain	Value between 0-1	Arrive time	<subscribe>, <unsubscribe>, <general>	Counter	Mail sequence ID

Table 8 Suspected List Manager Address table (SMA table)

List address	Manager address	CF_SMA	Time stamp	Status	Mcount	MID
Listname@listdomain	Managername @managerdomain	Value between 0-1	Arrive time	<subscribe>, <unsubscribe>, <general>	Counter	Mail sequence ID

Table 9 Suspected List Manager Type table (SMT table)

Manager address	Manager type	CF_SMT	Time stamp	Status	Mcount	MID
Managername @managerdomain	Manual, listserv, majordomo, listproc, smartlist, N/A	Value between 0-1	Arrive time	<subscribe>, <unsubscribe>, <general>	Counter	Mail sequence ID

Table 10 List Manager Command table

Type ID	Manager type	Command part flag	Subscribe format ID list	Unsubscribe format ID list	General format ID list
Sequence ID	Manual, listserv, majordomo, listproc, smartlist, N/A	Bit 0 = Rfc822-body Bit 1 = Rfc822-subject	(<keyword>,<pattern>)	(<keyword>,<pattern>)	(<keyword>,<pattern>)

2.4 Postprocessing process

In last process is decided and consider step. Which information suitable for submailing list system. By reconsider confidential factor and accurate mailing list characteristic in RODB again. Then send data record to configure at submailing list system. If found some mistake remove record from RODB to CMDB for recheck.

All of processes have to decrease responsibility of network administer. That duty is checking for find mailing list mail and gives information to configure submailing list system. But this is a heavy job because there are many mails come to network. If network administers cannot work in time that will cause problem in our traffic to Internet. So mailing list recognition system has to create for support this situation as present before.

3. Algorithm of Recognized Process

In recognized process could divide mailing list characteristic in 2 types, 'manager channel characteristic' and 'list channel characteristic'. For manager channel characteristic means e-mail that communicate between mailing list manager and user (subscriber). The recognized process has to use information from RCDB for recognized. For list channel characteristic means e-mail that communicate among subscriber. The recognized process use mailing list behavior to analyze. In figure 3 was presented recognized process algorithm.

Anyway there are other algorithms that corporate with recognized process such as confidential factor algorithm. Which use for calculate CF value, that consist of 2 part positive and negative algorithm. In positive algorithm use to increase CF value when the system found support reason. In the

opposite way negative algorithm use to decrease CF value when the system found another reasons to disprove.

```
/* Recognized algorithm */
if Rfc822-mailing-list-header not equal to null then
        if Rfc822-recipient in SLA table then
            increase CF_SLA and Mcount = Mcount+1
            if Rfc822-from in SLM table then
                increase CF_SLM
            else
                add member address by Rfc822-from in SLM table
                increase CF_SLM
        else
        create entry in LA table by list address = Rfc822-recipient
        and CF_LA = initial CF
        create entry in LM table by member address = Rfc822-from
        and CF_LM = initial CF
else
        if Rfc822-recipient in SLA table then
            decrease CF_SLA

if Rfc822-from not equal to (Smtp-mail-from or Rfc822-sender or
Rfc822-return-path or Rfc822-error-to or Rfc822-reply-to) then
        for (i = 1 to M) /*M is number of Rfc822-recipient*/
            for (j = 1 to N) /*N is number of Smtp-recipient*/
                if Smtp-recipient[j] == Rfc822-recipient[i] then
                    x = x+1
    if (x==0) or (x<N) then /*each Smtp-recipient not match
    Rfc822-recipient*/
            if Rfc822-recipient in SLA table then
                increase CF_SLA and Mcount = Mcount+1
                if Rfc822-from in SLM table then
                    increase CF_SLM
                else
                    add member address by Rfc822-from in SLM
    table
                    increase CF_SLM
            else
            create entry in LA table by list address = Rfc822-recipient
            and CF_LA = initial CF
            create entry in LM table by member address = Rfc822-
            from and CF_LM = initial CF
else
        if Rfc822-recipient in SLA then
            decrease CF_SLA

if (Rfc822-subject or Rfc822-body) in list manager command table
then
        if (Rfc822-recipient or Rfc822-from) in SMA table then
            if Rfc822-from in SLM table then
                increase CF_SLM and CF_SMA
        else
        create entry in LA table by list address =
        listname+@+managerdomaibn and CF_LA = initial CF
        create entry in MA table by manager address = Rfc822-
        recipient and CF_MA = initial CF
        create entry in MT table by manager type = Rfc822-
        reicipient name or manager type and CF_MT = initial MT
else
        if (Rfc822-recipient or Rfc822-from) in SMA table then
            if Rfc822-from in SLM table then
                decrease CF_SLM and CF_SMA
```

Fig. 3 Algorithm of recognized process

```
/*Confidential factor algorithm*/

if increase CF
        CF_x = CF_x + positive(CF_x)
If decrease CF
        CF_X = CF_x + negative(CF_x)

/*Positive CF algorithm*/
if receive CF_x then
        positive CF = (1 – CF_x)/2
        return positive CF

/*Negative CF algorithm*/
if receive CF_x then
        negative CF = CF-x / 2
        return negative CF
```

Fig. 4 Algorithm of confidential factor

4. Conclusion and Future Work

Truly mailing list manager (MLM) software perceives the network traffic troubles from mailing list operation. Some of MLM try to solve the problems but mailing list was consisting of many network systems so they could not wholly correct. Submailing list was another way to decrease data redundancies in network traffic. However if submailing list system could automatically work, there will make higher performance for network traffic.

From above concept, that would like to make submailing list system automatically learning mailing list characteristic and separate which e-mail come from mailing list system. This paper would like to present mailing list recognition system for recognize mail and give useful information to submailing list system. Begin with trap electronic mail from transfer channel and make it in recognize format. Then use assumption found in mailing list behavior to recognize message. And also find out confidential factor for indicate mailing list characteristic. Confidential factor could be able to swap by discover positive reason to support or negative case to conflict old data. When finishes there are also having revise process for make decision and decide which data suitable for submailing list operational. In addition we can solve problem about Spam mail, which extremely found in network traffic problems, by improve recognize process to focus on Spam mail behavior.

Future work we have to find out constant value by testing in real situation. Bring the result to improve and modify mailing list recognition system for high efficient and consistency of the system.

REFERENCES

[1] A. Schwartz, *Managing Mailing List 1st ed*, O'reilly & Associates. Inc, March 1996.

[2] S. Chomjan and A. Khunkitti, "A New Hierarchical Based Approach Mailing List System", *Proceedings of The 2003 International Conference on Information and Communication Technologies (ICT2003)*, Bang na Campus, Assumption University, April 8-10, 2003.

[3] G. Neufeld, "The Use of URLs as Meta-Syntax for Core Mail List Commands and their Transport through Message Header Fields", *Request for Comments 2369*, Nisto, July 1998.
[4] D. H. Crocker, "Standard for The Format of Arpa Internet Text Messages", *Request for Comments 822*, Dept. of Engineering University of Delaware, August 13, 1982.
[5] J. B. Postel, "Simple Mail Transfer Protocol", *Request for Comments 821*, Information Sciences Institute, Southern California University, August 1982.

ประวัติผู้เขียน

ชื่อ-นามสกุล	นางสาวณัฐติญา ไข่ติยากุล
วัน เดือน ปีเกิด	13 มิถุนายน 2520
ที่อยู่	530 หมู่บ้านฉัตรแก้ว แขวงคลองจั่น เขตบางกะปิ กรุงเทพมหานคร 10240
วุฒิการศึกษา	วิศวกรรมศาสตรบัณฑิต สาขาไฟฟ้ากำลัง
สถานที่สำเร็จการศึกษา	สถาบันเทคโนโลยีนานาชาติ สิรินธร มหาวิทยาลัยธรรมศาสตร์
ปีที่สำเร็จการศึกษา	ปีการศึกษา 2541