การประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะโดยเจ้าของอีเมล

นายอรรถกร องค์ศิริพร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2552

PERFORMANCE EVALUATION OF SPAM FILTER BY E-MAIL OWNERS

Mr. Athakorn Ongsiriporn

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2009

หัวข้อวิทยานิพนธ์	การประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะโดยเจ้าของอีเมล
โดย	นายอรรถกร องค์ศิริพร
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	อาจารย์ ดร.ยรรยง เต็งอำนวย

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(อาจารย์ จารุมาตร ปิ่นทอง)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(อาจารย์ ดร.ยรรยง เต็งอำนวย)

..... กรรมการ
(อาจารย์ ดร.เกริก ภิรมย์โสภา)

..... กรรมการภายนอกมหาวิทยาลัย
(ดร.โกเมน พิบูลย์โรจน์)

อรรถกร องค์ศิริพร : การประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะโดยเจ้าของอีเมล. (PERFORMANCE EVALUATION OF SPAM FILTER BY E-MAIL OWNERS) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : อ.ดร.ยรรยง เต็งอำนวย, 41 หน้า.

การประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะเป็นกระบวนการที่สำคัญ โดยผลการประเมินใช้สำหรับเปรียบเทียบเพื่อเลือกใช้ และเป็นแนวทางในการพัฒนาประสิทธิภาพของระบบคัดกรองอีเมลขยะ ปัจจุบันการประเมินมิได้ทำในสภาพแวดล้อมที่ใกล้เคียงกับความเป็นจริง ผู้ใช้งานไม่มีส่วนร่วมในกระบวนการประเมิน อีกทั้งมีปัญหาความเป็นส่วนตัว ทำให้ผลการประเมินไม่สอดคล้องกับความเป็นจริง งานวิจัยนี้นำเสนอกระบวนวิธีสำหรับประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะ ในสภาพแวดล้อมที่ใกล้เคียงกับความเป็นจริง ใช้อีเมลจริงจากองค์กรเป้าหมาย และผู้ใช้งานจะเป็นอาสาสมัครในกระบวนการประเมินเพื่อหลีกเลี่ยงการละเมิดความเป็นส่วนตัว กระบวนวิธีนี้ใช้การสุ่มตัวอย่างอีเมลเพื่อลดงานของอาสาสมัคร มุ่งเน้นให้ใช้งานได้ง่ายและมีค่าใช้จ่ายต่ำ โดยมีสมมติฐานว่าอาสาสมัครเป็นระบบคัดกรองในอุดมคติซึ่งคัดกรองอีเมลได้ถูกต้องเสมอ ผลการทดลองบ่งชี้ว่ากระบวนวิธีนี้สามารถประเมินประสิทธิภาพของระบบคัดกรองอีเมลขยะได้เป็นอย่างดี และอาสาสมัครตอบรับเข้าร่วมกระบวนการประเมินอย่างรวดเร็ว

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2552

5071458921 : MAJOR COMPUTER SCIENCE

KEYWORDS : SPAM / FILTER / EVALUATION / MANUAL / PRIVACY

ATHAKORN ONGSIRIPORN: PERFORMANCE EVALUATION OF SPAM FILTER BY E-MAIL OWNERS. THESIS ADVISOR: YUNYONG TENG-AMNUAY, Ph.D., 41 pp.

Performance evaluation of spam filter is an important process for choosing and improvement of spam filter. Nowadays, the evaluation does not based on realistic environment. Users do not participate in the process and privacy problems are always encountered. In this research, we propose a framework which evaluate in realistic environment. The framework uses a corpus of real e-mails. Users are volunteers who evaluate their e-mails without privacy problems. This is an ease of use and cheap framework. It uses a sampling method for reducing volunteers' jobs. Volunteers in this framework are an ideal spam filter which always gives correct classification. The result show a capability of this framework for spam filters evaluation and volunteer have quick response to our process.

Department : Computer Engineering Student's Signature .....

Field of Study : Computer Science Advisor's Signature .....

Academic Year : 2009

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยดี เนื่องมาจากความช่วยเหลืออย่างดียิ่งของอาจารย์ ดร.ยรรยง เต็งอำนวย อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้สละเวลาให้คำปรึกษาแนะนำแนวทางและการแก้ปัญหาต่างๆ เกี่ยวกับงานวิจัยอย่างดีตลอดมาจนเสร็จสมบูรณ์ ตลอดจนคำปรึกษาคำแนะนำข้อคิดสำหรับการดำเนินชีวิตต่อไปในอนาคต และผู้วิจัยขอกราบขอบพระคุณ อาจารย์จารุมาตร ปิ่นทอง อาจารย์ ดร.เกริก ภิรมย์โสภา และดร.โกเมน พิบูลย์โรจน์ คณะกรรมการสอบวิทยานิพนธ์ทุกท่านที่ได้ให้คำแนะนำ ข้อคิดเห็น ข้อเสนอแนะ และแนวทางในการพัฒนางานวิจัยนี้ให้สมบูรณ์ยิ่งขึ้น และขอขอบพระคุณคณาจารย์ทุกท่านที่กรุณาประสิทธิประสาทวิชาความรู้อันมีคุณค่ายิ่งแก่ผู้วิจัย

ขอขอบพระคุณ คุณชยา ลิมจิตติ คุณปณิตา บุญมา แห่งสำนักงานเทคโนโลยีสารสนเทศ จุฬาลงกรณ์มหาวิทยาลัย ที่ให้ความช่วยเหลือในการเก็บรวบรวมข้อมูลในการทำวิจัยและอำนวยความสะดวกในการทำวิจัยในครั้งนี้

ขอขอบคุณเพื่อนๆ และพี่ๆ ทุกคนที่ให้คำปรึกษา ให้กำลังใจ ข้อคิดเห็น และแนวคิดที่ดีทุกๆ ด้าน โดยเฉพาะอย่างยิ่ง คุณอรจิรา จริงจิตร ที่ให้กำลังใจ และช่วยเหลือในการทำงานวิจัยนี้

ท้ายนี้ขอขอบพระคุณ คุณพ่อ คุณแม่ ที่ให้การสนับสนุนและเอาใจใส่ดูแลผู้วิจัยเป็นอย่างดี ตลอดจนเป็นกำลังใจให้สามารถทำงานวิจัยนี้สำเร็จลุลวงได้ด้วยดี

สารบัญ

บทที่ หน้า

สารบัญตาราง

สารบัญภาพ

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

อีเมลขยะเป็นปัญหาที่สำคัญของระบบเครือข่ายคอมพิวเตอร์ เนื่องจากสัดส่วนของอีเมลขยะคิดเป็นร้อยละ 85.9 [1] ของปริมาณอีเมลทั้งหมด และจะเพิ่มเป็นร้อยละ 95 ในปี 2015 [2] ทำให้สิ้นเปลืองทรัพยากรของระบบเครือข่ายจำนวนมากและสร้างปัญหาแก่ผู้ใช้ การป้องกันอีเมลขยะโดยทั่วไปจะใช้วิธีคัดกรอง (filter) ทั้งการคัดกรองที่เครื่องแม่ข่าย และเครื่องคอมพิวเตอร์ส่วนบุคคล โดยถูกปรับปรุงและพัฒนาอย่างต่อเนื่องเพื่อให้สามารถคัดกรองอีเมลขยะได้อย่างมีประสิทธิภาพและทันต่อการคัดกรองอีเมลขยะรูปแบบใหม่ๆ [3]

การประเมินประสิทธิภาพระบบคัดกรองมีความสำคัญสำหรับกำหนดแนวทางในการปรับปรุงและพัฒนา แต่ในปัจจุบันวิธีการประเมินยังห่างไกลจากสภาพความเป็นจริง [3] [4] [5] [6] [7] [8] เนื่องจากคลังอีเมลที่ใช้ในการทดสอบมิได้มาจากระบบเครือข่ายที่ใช้งานจริงขององค์กรเป้าหมาย รูปแบบของอีเมลที่ผ่านเข้ามายังระบบเป้าหมายอาจมีความแตกต่างจากคลังอีเมลจากแหล่งอื่นที่นำมาทดสอบ และผู้ใช้งานในระบบก็มิได้มีส่วนในการประเมิน รวมถึงมิได้มีความสัมพันธ์กับคลังอีเมลที่ใช้ทดสอบ ผลการประเมินที่ได้จะไม่ใช่ผลที่ได้จากสถานการณ์การทำงานจริงของระบบคัดกรองนั้นๆ

งานวิจัยนี้จึงเสนอกระบวนวิธีสำหรับประเมินระบบคัดกรองอีเมลขยะ โดยใช้กลุ่มอีเมลจริงจากองค์กรที่เป็นสถาบันการศึกษาขนาดใหญ่ มีผู้ใช้จำนวนมาก และให้เจ้าของอีเมลมีส่วนร่วมในขั้นตอนการประเมิน เพื่อที่จะไม่ละเมิดความเป็นส่วนตัวของผู้ใช้งาน

1.2 วัตถุประสงค์ของการวิจัย

เพื่อออกแบบกระบวนวิธีสำหรับประเมินประสิทธิภาพของระบบคัดกรองอีเมลขยะที่เครื่องแม่ข่าย โดยอาศัยเจ้าของอีเมลที่ใช้งานจริงในระบบโดยไม่ละเมิดความเป็นส่วนตัว

1.3 ขอบเขตของการวิจัย

1.3.1 ออกแบบกระบวนวิธีสำหรับประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะ

1.3.2 สามารถประเมินประสิทธิภาพได้แม้มีระบบคัดกรองเพียงระบบเดียว

1.3.3 ประเมินเฉพาะระบบคัดกรองอีเมลขยะที่ทำงานบนเครื่องแม่ข่าย

1.3.4 ใช้ข้อมูลจากมหาวิทยาลัยในประเทศ โดยมี นิสิต นักศึกษา และบุคลากร ไม่ต่ำกว่า 30,000 คน เช่น จุฬาลงกรณ์มหาวิทยาลัย

1.3.5 กลุ่มอาสาสมัครเป็นกลุ่มบุคลากรในจุฬาลงกรณ์มหาวิทยาลัย

1.3.6 แยกแยะกลุ่มอาสาสมัครตามประเภทของผู้ใช้ เช่น คณาจารย์ เจ้าหน้าที่ นิสิตระดับปริญญาตรี ปริญญาโท และปริญญาเอก

1.3.7 ใช้วิธีของ ทาโร ยามาเน่ [9] กำหนดขนาดตัวอย่างอีเมล

1.3.8 การประเมินอีเมลด้วยตาของอาสาสมัครไม่มีความผิดพลาด และยอมรับในวิจารณญาณของการประเมินด้วยตาของอาสาสมัคร

1.3.9 กระบวนการคัดกรองอีเมลแต่ละฉบับด้วยตานั้นให้สามารถเปิดอ่านได้เฉพาะเจ้าของเท่านั้น

1.4 ขั้นตอนของการวิจัย

1.4.1 ศึกษาวิธีการประเมินระบบคัดกรองอีเมลขยะ

1.4.2 ศึกษาวิธีคัดกรองอีเมลขยะจากงานวิจัย และวิธีที่ใช้งานอย่างแพร่หลายในปัจจุบัน

1.4.3 ออกแบบกระบวนวิธีประเมินระบบคัดกรองอีเมลขยะที่ใช้กลุ่มอีเมลของกลุ่มอาสาสมัครที่ใช้งานระบบเป้าหมาย

1.4.4 สร้างระบบคัดกรองอีเมลขยะอ้างอิง

1.4.5 สร้างระบบประเมินระบบคัดกรองอีเมลขยะโดยใช้กลุ่มอีเมลของกลุ่มอาสาสมัครที่ใช้งานระบบเป้าหมาย

1.4.6 ทดสอบวิธีการประเมินระบบคัดกรองอีเมลขยะกับระบบคัดกรองที่ใช้งานปัจจุบัน ระบบคัดกรองเก่า และระบบคัดกรองอ้างอิง

1.4.7 วิเคราะห์ผลการประเมิน

1.4.8 สรุปผลและเรียบเรียงวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 เป็นเครื่องมือสำหรับผู้ดูแลระบบ และองค์กร ในการประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะ

1.5.2 สร้างความมั่นใจให้กับผลการประเมินระบบคัดกรองอีเมลขยะ เนื่องจากใช้ข้อมูลจริงและผู้ใช้งานระบบมีส่วนร่วมในการประเมิน

1.5.3 ลดค่าใช้จ่าย และความเสี่ยงให้กับองค์กรในการตัดสินใจเลือกซื้อ และเลือกใช้งานระบบคัดกรองอีเมลขยะ

1.5.4 เป็นระบบที่ใช้งานได้ง่าย มีค่าใช้จ่ายที่ต่ำ และไม่ต้องใช้เจ้าหน้าที่เทคนิค ทำให้สามารถประเมินได้บ่อยครั้งตามต้องการ

1.5.5 รักษาความเป็นส่วนตัวของผู้ใช้

1.5.6 การประเมินได้บ่อยทำให้สามารถปรับตามความต้องการและความสนใจของผู้ใช้งานที่เปลี่ยนไปได้

1.5.7 ผลการประเมินใช้เป็นแนวทางสำหรับผู้ดูแลระบบในการปรับปรุง พัฒนาระบบคัดกรอง และสามารถใช้เป็นดัชนีชี้วัดระดับของวิจารณญาณในเรื่องอีเมลขยะของผู้ใช้งานในองค์กรได้ด้วย

1.6 โครงสร้างของวิทยานิพนธ์

เนื้อหาของวิทยานิพนธ์ฉบับนี้แบ่งออกเป็น 5 บทดังนี้ คือ บทที่ 1 เป็นบทนำของงานวิจัย บทที่ 2 กล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องกับงานวิจัยชิ้นนี้ บทที่ 3 กล่าวถึงวิธีการดำเนินงานวิจัยในแต่ละขั้นตอนอย่างละเอียด บทที่ 4 เป็นการอธิบายผลการทดลอง และบทที่ 5 เป็นการสรุปผลการทดลองและข้อเสนอแนะจากงานวิจัย ซึ่งอาจเป็นประโยชน์กับการวิจัยเพิ่มเติมในอนาคต

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 อีเมลขยะ

สันนิษฐานว่า คำว่า “สแปม (Spam)” มาจากละครสั้นทางโทรทัศน์ ชื่อ Monty Python’s Flying Circus ซึ่งมีการร้องประสานเสียงคำว่า S-P-A-M ซ้ำไปซ้ำมาและร้องดังทำให้ไม่สามารถได้ยินบทสนทนาอื่นในละคร บางกลุ่มสันนิษฐานว่ามาจากเนื้อกระป๋องที่รับประทานในมื้อกลางวันซึ่งส่วนใหญ่ประกอบด้วยเนื้อเทียม บางกลุ่มกล่าวว่า มาจากพฤติกรรมของสมาชิกชมรม MUSH (Multi User Shared Hallucination) ซึ่งมักพิมพ์ คำว่า S-P-A-M เพื่อทำให้ผู้ใช้รายอื่นไม่สามารถเข้าร่วมสนทนาออนไลน์ของกลุ่มได้ [10]

ยังไม่มีคำจำกัดความที่ชัดเจน และเป็นที่ยอมรับ นายกุยโด เซอร์เยน [10] กล่าวว่า คำจำกัดความที่เหมาะสมของอีเมลขยะต้องครอบคลุมคุณสมบัติต่างๆ คือ พฤติกรรมเกี่ยวกับการโฆษณา จิตวิทยาของผู้รับ บริบททางกฎหมาย ข้อคิดเห็นทางเศรษฐศาสตร์ และปัญหาทางเทคโนโลยี OECD [10] ให้คำจำกัดความว่า เป็นอีเมลที่ผู้รับไม่ต้องการ ถูกส่งไปยังผู้รับครั้งละมากๆ ซึ่งอาจไม่รู้จักผู้รับมาก่อน มีการส่งซ้ำหลายครั้ง โดยหวังผลเกี่ยวกับการค้า มีเนื้อหาโกหกหรือทำให้ไม่พอใจ [10] อีเมลขยะถูกเรียกว่า Unsolicited Bulk E-mail (UBE) [11] หรือ Unsolicited Commercial E-mail (UCE)

อีเมลขยะถูกนำเสนอเป็น Internet Request for Comments (RFC) ในปี ค.ศ. 1975 และนำเสนอในวารสาร “Communications of the ACM” ต่อมาในปี ค.ศ. 1982 อีเมลขยะฉบับแรกส่งจาก DEC marketing ไปยังอีเมลทั้งหมดใน Arpanet แต่คำว่าอีเมลขยะถูกใช้เมื่อนายแคนเตอร์ และนายซีเกล ได้กระจายข่าวเกี่ยวกับความไม่ยุติธรรมของการชิงโชค U.S. Green Card [10]

อีเมลขยะแบ่งออกเป็นประเภทต่างๆ ตามเป้าหมายของผู้ส่ง เช่น การโฆษณาสินค้า การประกาศต่างๆ การปล่อยข่าวหลอกลวง (fraud) การแอบอ้างชื่อผู้อื่น (phishing) การเตือน การขอความเมตตา (hoaxes) จดหมายลูกโซ่ (chain e-mail) เนื้อหาว่าร้ายผู้อื่น (joe jobs) กระจายซอฟต์แวร์ที่เป็นอันตรายเช่น ไวรัส มัลแวร์ อีเมลตอบรับ หรือยืนยันบริการต่างๆ (bounce message)

2.2 การป้องกันอีเมลขยะ

มีงานวิจัย องค์กร และหน่วยงานจำนวนมากค้นคว้าหาวิธีป้องกันอีเมลขยะ โดยนำเสนอวิธีป้องกันอีเมลขยะหลายวิธี เช่น การออกกฎหมายเอาผิดกับผู้ส่งอีเมลขยะ [2] [10] การจัดตั้งองค์กรเพื่อความร่วมมือในการต่อต้านอีเมลขยะ (SpamCop) [12] การป้องกันมิให้ผู้ส่งอีเมลขยะดักจับที่อยู่อีเมล [10]

เนื่องจากอีเมลขยะมีปริมาณมาก จึงนิยมใช้เทคโนโลยีทางคอมพิวเตอร์ในการป้องกัน ประกอบด้วยวิธีดังต่อไปนี้

2.2.1 ระงับการติดต่อกับเครื่องคอมพิวเตอร์ หรือเครือข่ายที่ส่งอีเมลขยะ (IP Blocking)

2.2.2 กำหนดบัญชีดำ (Blacklisting) ระบุที่อยู่อีเมล หมายเลขไอพี หรือช่วงหมายเลขไอพี ชื่อโดเมน ที่ส่งอีเมลขยะเป็นประจำ เพื่องดให้บริการรับ-ส่งอีเมลกับข้อมูลในรายการดังกล่าว ตัวอย่างบัญชีดำเช่น Domain Name System Blacklists (DNSBLs), Exploits Block List (XBL), Spamhaus block list (SBL) เป็นต้น [2] [10]

2.2.3 กำหนดบัญชีขาว (Whitelisting) ทำงานในลักษณะเดียวกับบัญชีดำ แต่เป็นรายการที่อนุญาตให้ใช้บริการรับส่งอีเมลได้ ตัวอย่างบัญชีขาวได้แก่ Domain Name System Whitelists (DNSWLs) [10]

2.2.4 บัญชีเทา (Graylisting) จะใช้พฤติกรรมการไม่ส่งอีเมลซ้ำเมื่อไม่สามารถส่งอีเมลสำเร็จในการทำงาน เมื่อมีการใช้บริการจากที่อยู่อีเมล หมายเลขไอพี หรือช่วงหมายเลขไอพี ชื่อโดเมน ที่อยู่นอกบัญชีขาว จะงดให้บริการในครั้งแรกแต่จะ ให้บริการเมื่อมีการขอใช้บริการซ้ำเข้ามา [2]

2.2.5 การคัดกรอง (Filtering) เป็นวิธีอัตโนมัติที่ใช้งานสะดวก การคัดกรองจะตรวจสอบเนื้อหาทั้งหมดของอีเมล หรือส่วนใดส่วนหนึ่งเช่น หัวเรื่อง เป็นต้น การคัดกรองโดยวิธีทางคณิตศาสตร์เช่น วิธี Bayesian การคัดกรองโดยใช้วิธีทางปัญญาประดิษฐ์เช่น Support Vector Machine และ Neural Networks [10]

2.2.6 วิธีอื่นๆ เช่น การไม่อนุญาตให้ใช้บริการเครื่องแม่ข่ายอีเมลที่อยู่ภายนอก (TCP Blocking) การเข้าสู่ระบบก่อนการใช้บริการ (authentication) การพิสูจน์ผู้ส่ง (verification) จำกัดการใช้หน่วยประมวลผลกลาง หน่วยความจำหลัก หรือต้องจ่ายเงินเมื่อต้องการส่งอีเมล จำกัดปริมาณการส่งอีเมล การเสนอที่อยู่อีเมลรูปแบบที่ยากแก่การเข้าใจโดยคอมพิวเตอร์ การรับเฉพาะอีเมลจากผู้ส่งที่น่าเชื่อถือ

2.3 **ความเป็นส่วนตัวของอีเมล**

ข้อมูลที่ส่งทางอีเมลบางครั้งเป็นความลับระหว่างผู้ส่งและผู้รับไม่ว่าในระดับองค์กรหรือส่วนบุคคล วัตถุประสงค์หลักของการรักษาความเป็นส่วนตัวคือ อนุญาตให้เฉพาะเจ้าของข้อมูลเท่านั้นสามารถเข้าถึงข้อมูลดังกล่าวได้ การถูกละเมิดความเป็นส่วนตัวของผู้ใช้งานอีเมลมาจากหลายสาเหตุเช่น การโดนโจมตีจากอีเมลขยะ นโยบายการตรวจสอบอีเมลของบุคลากรในหน่วยงาน การถูกดักจับข้อมูลระหว่างการส่งทั้งทางจากผู้ไม่ประสงค์ดีหรือหน่วยงานของรัฐ การละเมิดความเป็นส่วนตัวโดยผู้ดูแลระบบ การรักษาความเป็นส่วนตัวนั้นมิใช่ผลิตภัณฑ์ แต่เป็นกระบวนการ [13] เช่น การตั้งรหัสผ่านเพื่อเข้าระบบอีเมล การเชื่อมต่อผ่านช่องทางที่ปลอดภัย การเข้ารหัสข้อมูลของอีเมลก่อนส่ง เป็นต้น

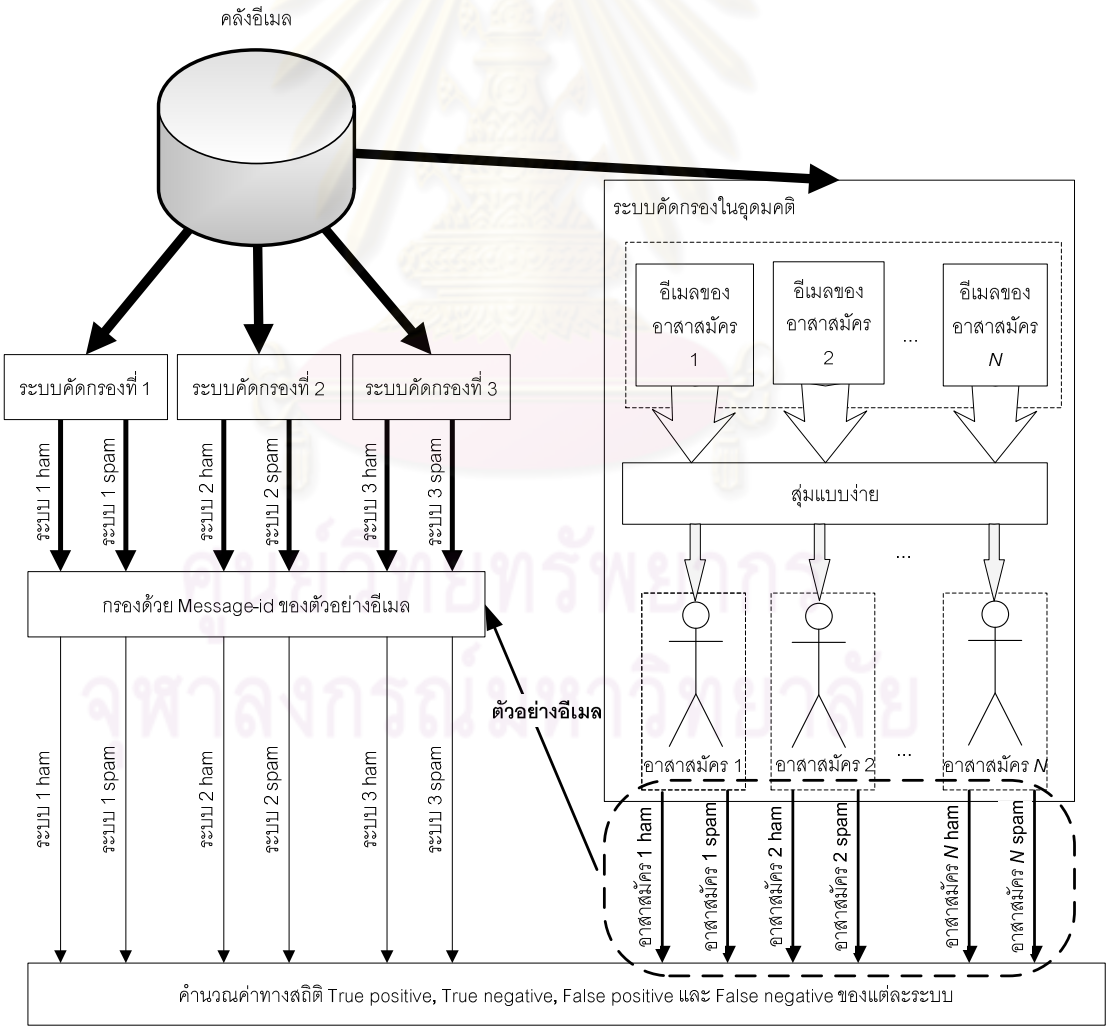
ในประเทศที่มีกฎหมายการรับรองเรื่องการรักษาความลับและมีกฎหมายคุ้มครองเกี่ยวข้องกับการติดต่อสื่อสารทางจดหมายได้ให้การคุ้มครองการติดต่อสื่อสารทางอีเมลในลักษณะเดียวกัน โดยคุ้มครองการละเมิดความเป็นส่วนตัวทุกกรณี

สิทธิการเข้าถึงข้อมูลอีเมลนั้นไม่อนุญาตให้ผู้อื่นนอกจากเจ้าของสามารถเข้าถึงข้อมูลดังกล่าวได้ แม้ว่าอีเมลจะถูกเก็บรักษาไว้ที่ผู้ให้บริการอินเทอร์เน็ต แต่ผู้ให้บริการอินเทอร์เน็ตนั้นไม่มีสิทธิในการเข้าถึงเนื้อความของอีเมลเหล่านั้น นอกเหนือไปจากการบริหารจัดการตามสิทธิอำนาจ เช่น การสำรองข้อมูล เป็นต้น [14]

บทที่ 3

วิธีดำเนินการวิจัย

จากปัญหาที่เกิดขึ้นของกระบวนวิธีสำหรับประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะในปัจจุบันที่ได้กล่าวไว้ข้างต้น ซึ่งวิธีการประเมินนั้นยังห่างไกลจากสภาพความเป็นจริง โดยไม่ได้ใช้อีเมลจริงขององค์กรเป้าหมายในการประเมิน และผู้ใช้งานไม่มีส่วนร่วมในการประเมิน งานวิจัยนี้จึงมีจุดมุ่งหมายเพื่อนำเสนอกระบวนวิธีในการประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะโดยเจ้าของอีเมล ใช้คลังอีเมลที่รวบรวมจากอีเมลจริงขององค์กรเป้าหมายในการประเมิน และกระบวนการประเมินนี้จะไม่ก่อให้เกิดปัญหาความเป็นส่วนตัว โดยภาพรวมของระบบแสดงดังรูปที่ 3.1



รูปที่ 3.1 ภาพรวมของระบบประเมินประสิทธิภาพ

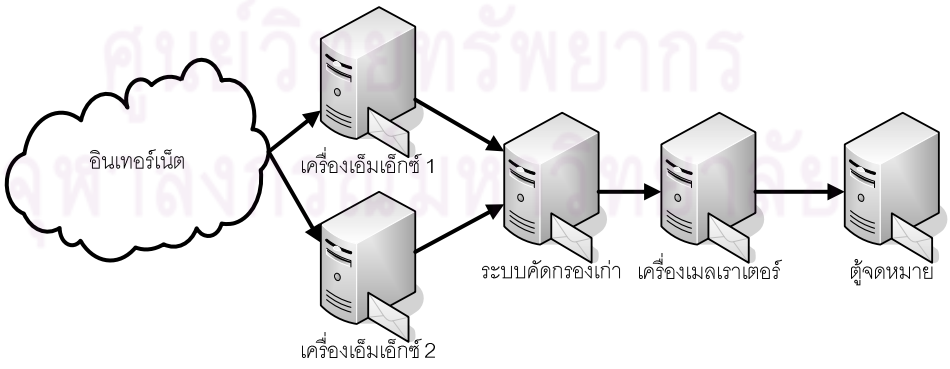
จากรูปที่ 3.1 คลังอีเมลจะถูกผ่านไปยังระบบคัดกรองทั่วไปและระบบในอุดมคติ กรองอีเมลที่ถูกคัดกรองด้วยระบบทั่วไปด้วย Message-id ของอีเมลที่ถูกคัดกรองโดยระบบในอุดมคติ เปรียบเทียบผลการคัดกรองของระบบทั่วไปกับระบบในอุดมคติ และคำนวณหาค่าทางสถิติเพื่อเปรียบเทียบประสิทธิภาพ จากภาพรวมของระบบสามารถออกแบบวิธีดำเนินการวิจัยได้ดังต่อไปนี้

1. ศึกษาระบบอีเมลปัจจุบันขององค์กรเป้าหมาย
2. รวบรวมและวิเคราะห์คลังอีเมล
3. คัดเลือกอาสาสมัคร
4. เตรียมระบบคัดกรองที่ต้องการประเมินประสิทธิภาพ
5. สร้างระบบคัดกรองอ้างอิงจากงานวิจัย
6. ป้อนคลังอีเมลผ่านระบบคัดกรองที่ต้องการประเมินประสิทธิภาพ
7. ติดต่อกลุ่มอาสาสมัคร
8. สุ่มอีเมลของอาสาสมัครแต่ละคนและให้ประเมินอีเมลของตนด้วยตา
9. วิเคราะห์และสรุปผลการประเมินประสิทธิภาพ

3.1 ศึกษาระบบอีเมลปัจจุบันขององค์กรเป้าหมาย

3.1.1 ระบบอีเมลของจุฬาลงกรณ์มหาวิทยาลัย

การทำวิจัยครั้งนี้ผู้วิจัยกำหนดให้จุฬาลงกรณ์มหาวิทยาลัยเป็นองค์เป้าหมาย ซึ่งมีบุคลากรและนิสิตรวมกันประมาณ 30,000 คน มากพอสำหรับทำการวิจัย และสามารถใช้งานทรัพยากรของระบบเครือข่ายคอมพิวเตอร์ได้พอสมควร

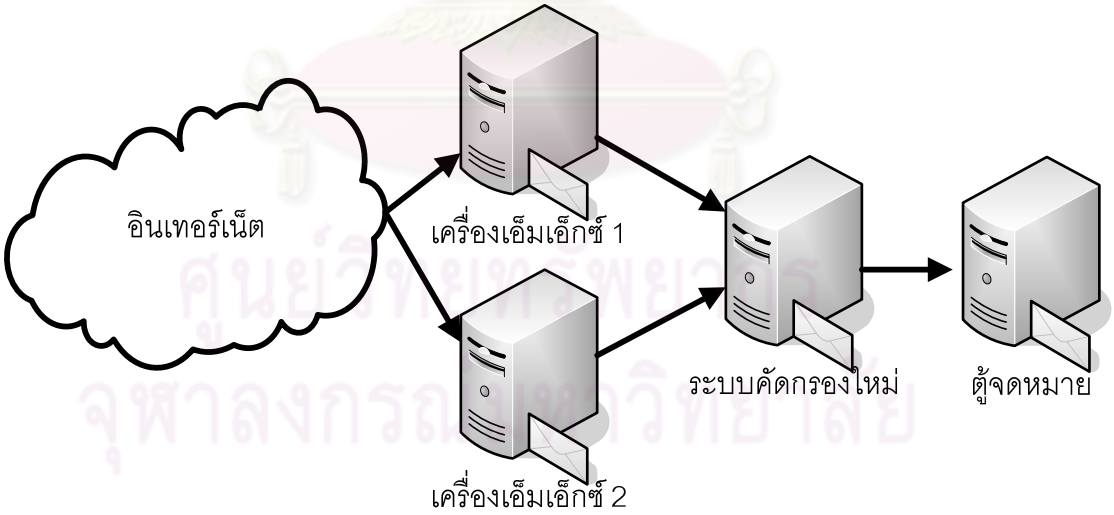


รูปที่ 3.2 ระบบอีเมลขณะเริ่มวิจัยของจุฬาลงกรณ์มหาวิทยาลัย

ในขณะเริ่มต้นวิจัยระบบอีเมลของจุฬาลงกรณ์มหาวิทยาลัยมีโครงสร้าง และองค์ประกอบดังรูปที่ 3.2 เมื่ออีเมลจากภายนอกส่งมายังผู้รับในมหาวิทยาลัย เครื่องคอมพิวเตอร์

เอ็มเอ็กซ์ 1 และ เอ็มเอ็กซ์ 2 ซึ่งทำหน้าที่เป็นเมลเอ็กเซนเจอร์รับอีเมลที่ส่งถึงผู้รับของมหาวิทยาลัย ระบบจะตรวสอบต้นทางของอีเมลว่าเป็นแหล่งส่งอีเมลขยะหรือไม่ โดยตรวจสอบจากฐานข้อมูลอาร์บีแอล (RBL) ได้แก่ abuseat.org, dsbl.org และ spamhaus.org และตรวจรายชื่ออีเมลของผู้รับว่ามีอยู่ในระบบหรือไม่ โดยตรวจสอบจากฐานข้อมูลแบบเบริกเลห์ ดีบี (Berkley DB) ผ่านโปรโตคอลแอลแดป (LDAP) หลังจากนั้นจะส่งอีเมลต่อไปยังเครื่อง SPAM Filter ซึ่งทำหน้าที่คัดกรองอีเมลที่ได้รับว่าเป็นอีเมลขยะหรือแนบไวรัสมาด้วยหรือไม่ ติดตั้งโปรแกรมไอเอ็มเอสเอส รุ่น 7.0 (IMSS 7.0) เป็นโปรแกรมคัดกรองอีเมลขยะและอีเมลไวรัส หากพบว่าเป็นอีเมลขยะให้แทรกข้อความ “SPAM” ในส่วนหัวข้อของอีเมล อีเมลจะถูกส่งต่อไปยังเครื่องเมลเราเตอร์ ซึ่งมีที่ค้นหาและส่งอีเมลไปยังเครื่องคอมพิวเตอร์ที่มีตู้จดหมายของผู้รับอยู่ ซึ่งมีหน้าที่เก็บอีเมลของผู้ใช้งานทั้งหมด และให้บริการอ่านอีเมลผ่านโปรโตคอล ป๊อบ รุ่นที่ 3 (POP3) และ ไอแมพ รุ่นที่ 4 (IMAP4) เครื่องที่มีตู้จดหมายจะมีกระจายไปยังหน่วยงานต่างๆ หรือสำหรับแต่ละโดเมนย่อยในมหาวิทยาลัยเพื่อรับอีเมลที่ส่งมายังโดเมนย่อยนั้น

สำนักเทคโนโลยีสารสนเทศ จุฬาลงกรณ์มหาวิทยาลัย มีโครงการที่จะเปลี่ยนระบบคัดกรองอีเมลขยะตามรูปที่ 3.2 โดยหลังจากมีการเปลี่ยนแปลงระบบคัดกรองอีเมลขยะแล้วทำให้ระบบอีเมลของจุฬาลงกรณ์มหาวิทยาลัย เป็นดังรูปที่ 3.3



รูปที่ 3.3 ระบบอีเมลใหม่ของจุฬาลงกรณ์มหาวิทยาลัย

จากรูปที่ 3.3 เมื่อเปรียบเทียบกับรูปที่ 3.2 พบกว่าได้มีการรวมหน้าที่ของเครื่องที่เป็นระบบคัดกรองอีเมลขยะและเครื่องเมลเราเตอร์ เนื่องจากระบบคัดกรองอีเมลขยะใหม่สามารถทำหน้าที่ดังกล่าวได้ในตัวเอง

หลังจากติดตั้งเครื่องคอมพิวเตอร์เครื่องใหม่ให้แก่ระบบอีเมลแล้ว ระบบเดิมก็ยังคงทำงานอยู่แต่แบ่งหน้าที่รับผิดชอบโดย ระบบใหม่จะรับผิดชอบในการคัดกรองอีเมลของอาจารย์และเจ้าหน้าที่ และระบบเก่าจะรับผิดชอบในการคัดกรองอีเมลของนิสิต

3.1.2 อีเมลแอดเดรสของผู้ใช้งาน

ผู้ใช้งานระบบอีเมลของจุฬาลงกรณ์มหาวิทยาลัยแต่ละคนจะมีอีเมลแอดเดรสอย่างน้อย 1 ประเภทซึ่งประกอบด้วย

3.1.2.1 อีเมลหลัก คืออีเมลที่องค์กรสร้างให้เพื่อบริการแก่ผู้ใช้งานแต่ละคน มีการใช้งานอยู่เป็นประจำ เช่น First.L@Chula.ac.th

3.1.2.2 อีเมลกลุ่ม คืออีเมลแอดเดรสที่สร้างขึ้นเพื่อนประโยชน์บางประการ เช่น การกระจายข่าวของกลุ่มผู้ดูแลระบบ หรือกลุ่มอาจารย์เป็นต้น อีเมลกลุ่มจะประกอบด้วยอีเมลสมาชิก เมื่อมีอีเมลมาถึงอีเมลกลุ่ม ระบบอีเมลจะกระจายส่งอีเมลนั้นไปยังอีเมลสมาชิกทุกอีเมล เช่น noc@it.chula.ac.th เป็นต้น

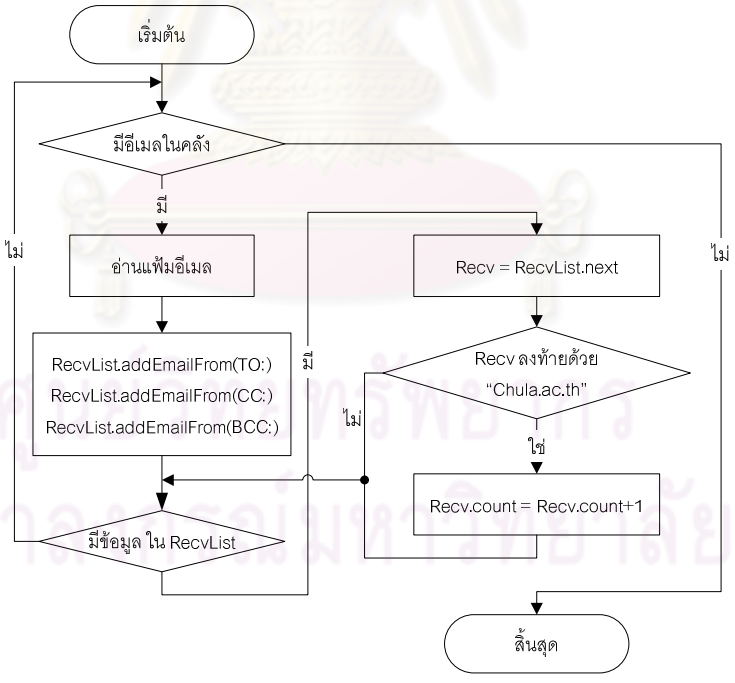
3.2 รวบรวมและวิเคราะห์คลังอีเมล

จุดมุ่งหมายและวัตถุประสงค์ของงานวิจัยนี้ คือการประเมินประสิทธิภาพของระบบคัดกรองอีเมลขยะในสถานการณ์ที่ใกล้เคียงกับความเป็นจริง ให้ผู้ใช้งานมีส่วนร่วมในกระบวนการประเมิน และใช้อีเมลจริงในการประเมินโดยไม่ก่อให้เกิดปัญหาความเป็นส่วนตัว ดังนั้นคุณสมบัติของคลังอีเมลที่ใช้ในการวิจัยครั้งนี้ต้องเป็นอีเมลจริงประกอบด้วยอีเมลดีและอีเมลขยะ ซึ่งส่งถึงผู้ใช้งานทุกคนและผ่านระบบอีเมลของจุฬาลงกรณ์มหาวิทยาลัย อีเมลทั้งหมดต้องไม่ผ่านการคัดกรองด้วยระบบคัดกรองที่ใช้ทดสอบตัวใดมาก่อน หากใช้คลังอีเมลซึ่งผ่านการคัดกรองมาแล้วด้วยระบบคัดกรองใด คลังอีเมลทั้งหมดจะไม่ถูกคัดกรองหรือตัดสินว่าเป็นอีเมลขยะอีกเมื่อถูกคัดกรองซ้ำด้วยระบบคัดกรองนั้นอีก เป็นผลทำให้ผลการคัดกรองและผลการประเมินประสิทธิภาพผิดพลาด จากข้อมูลทางสถิติพบว่านิสิตมีสถิติการใช้งานอีเมลของมหาวิทยาลัยต่ำมากจึงไม่เหมาะที่จะใช้นิสิตเป็นกลุ่มอาสาสมัครในการวิจัย อีกทั้งการเก็บรวบรวมอีเมลจากระบบอีเมลของนิสิตยังทำได้ยากลำบาก เนื่องจากโปรแกรมไอเอ็มเอสเอส ไม่สามารถทำสำเนาอีเมลทั้งหมดไปยังตู้จดหมายอื่นได้ ทางผู้วิจัยจึงเลือกอีเมลของคณาจารย์ และเจ้าหน้าของมหาวิทยาลัยซึ่งมีการใช้งานอย่างต่อเนื่องเป็นคลังอีเมลของงานวิจัยนี้

จากคุณสมบัติของคลังอีเมลที่ต้องการนำไปสู่วิธีการเก็บรวบรวมคลังอีเมลจากระบบอีเมลของจุฬาลงกรณ์มหาวิทยาลัย แต่เนื่องจากขณะเริ่มต้นรวบรวมระบบอีเมลได้เปลี่ยนแปลงไป

ดังแสดงในรูป 3.3 ซึ่งส่งผลให้การเก็บรวบรวมอีเมลของกลุ่มเป้าหมายที่ต้องการทำได้สะดวกขึ้น เนื่องจากระบบคัดกรองใหม่คัดกรองเฉพาะอีเมลของอาจารย์และเจ้าหน้าที่เท่านั้น หลังติดตั้งระบบคัดกรองใหม่ อีเมลขาเข้าของอาจารย์และเจ้าหน้าที่ทั้งหมดจะต้องผ่านระบบคัดกรองใหม่นี้ก่อนที่จะถูกส่งไปยังตู้จดหมายของผู้รับ ดังนั้นการรวบรวมคลังอีกเมลจะต้องสำเนาอีเมลที่ผ่านระบบคัดกรองใหม่ โดยปรับตั้งให้ระบบคัดกรองใหม่ให้ทำสำเนาอีเมลที่เป็นอีเมลขยะและอีเมลดีมาที่ตู้รับจดหมายของผู้วิจัย

เนื่องจากระบบคัดกรองใหม่มีระบบเซนต์เดอร์เบส (SenderBase) [15] ซึ่งคอยตรวจสอบความน่าเชื่อถือของอีเมลโดยอาศัยคุณสมบัติต่างๆ เช่น ตรวจสอบหมายเลขไอพีของเครื่องผู้ส่งว่าถูกขึ้นบัญชีดำหรือไม่หรือถูกรายงานจากผู้รับว่าผู้ส่งอีเมลมักส่งอีเมลขยะเป็นต้น ระบบเซนต์เดอร์เบสสามารถกำจัดอีเมลได้มากถึงร้อยละ 85 ก่อนที่จะเข้ามายังระบบอีเมลของจุฬาลงกรณ์มหาวิทยาลัย ทำให้มีอีเมลผ่านเข้ามาเพียงร้อยละ 15 เท่านั้น อย่างไรก็ดีปริมาณดังกล่าวมากพอที่จะนำมาใช้ในงานวิจัย และเมื่อรวบรวมคลังอีเมลได้แล้วนำคลังอีเมลมาวิเคราะห์เพื่อนับจำนวนอีเมลที่ส่งถึงผู้รับแต่ละคนต่อไป



รูปที่ 3.4 การนับจำนวนอีเมลของผู้ใช้แต่ละคน

จากรูปที่ 3.4 การวิเคราะห์คลังอีเมลเพื่อนับจำนวนอีเมลขาเข้าของผู้ใช้งานแต่ละคน จะอ่านอีเมลครั้งละฉบับจากคลังอีเมล อ่านรายชื่อผู้รับอีเมลจากอีเมลนั้นประกอบด้วย ผู้รับ (TO)

คาร์บอนก๊อบปี้ (CC) และไบรด์คาร์บอนก๊อบปี้ (BCC) อีเมลแอดเดรสต้องลงท้ายด้วยโดเมนเนม “Chula.ac.th” เท่านั้น บวกเพิ่มจำนวนอีเมลขาเข้าของแต่ละผู้รับ แล้วอ่านอีเมลฉบับถัดไปจนหมดคลังอีเมล

3.3 **คัดเลือกอาสาสมัคร**

จากสถิติการใช้งานอีเมลของจุฬาลงกรณ์มหาวิทยาลัยพบว่า กลุ่มผู้ใช้ที่มีการใช้งานอย่างต่อเนื่องคือกลุ่มคณาจารย์และเจ้าหน้าที่มหาวิทยาลัย ส่วนนิสิตมีสถิติการใช้งานต่ำเนื่องจากนิสิตนิยมใช้งานฟรีอีเมล ดังนั้นกลุ่มอาสาสมัครของงานวิจัยนี้เลือกจากเหล่าคณาจารย์และเจ้าหน้าที่มหาวิทยาลัย การคัดเลือกอาสาสมัครอาศัยสถิติการใช้งาน ซึ่งได้จากปริมาณอีเมลของผู้ใช้งานแต่ละคนในคลังอีเมลเป็นเกณฑ์ในการคัดเลือกกลุ่มอาสาสมัคร กลุ่มอาสาสมัครต้องมีส่วนรวมในการประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะที่จะถูกนำมาใช้งานในองค์กรของกลุ่มอาสาสมัครนั้น เพื่อให้ได้ระบบคัดกรองที่เหมาะสมกับสถานการณ์ในการใช้งานจริงและกลุ่มผู้ใช้งานในองค์กร เพราะความถูกต้องของการคัดกรอง อีเมลขึ้นอยู่กับเจ้าของอีเมลเป็นหลัก โดยงานวิจัยนี้มีสมมติฐานว่าอาสาสมัครเป็นระบบคัดกรองในอุดมคติตัดสินใจได้ถูกต้องเสมอโดยไม่มีข้อผิดพลาดใดๆ มีหน้าที่ 2 ประการคือ

3.3.1 **ตรวจสอบผลการคัดกรองของระบบคัดกรองที่ต้องการประเมินประสิทธิภาพ**

การตัดสินว่าอีเมลฉบับใดเป็นอีเมลขยะหรือไม่ ผู้ที่มีสิทธิ์ตัดสินคือเจ้าของอีเมลแต่เพียงผู้เดียวเท่านั้น เนื่องจากอีเมลเป็นข้อมูลส่วนตัว ดังนั้นการตรวจสอบผลการคัดกรองของระบบคัดกรอง จำเป็นต้องใช้เจ้าของอีเมลฉบับนั้นมาตรวจสอบ ระบบคัดกรองอีเมลขยะจะแยกอีเมลออกเป็น 2 กลุ่มคือ อีเมลดีและอีเมลขยะ เจ้าของจะต้องตรวจสอบอีเมลทั้ง 2 กลุ่มว่ามีผลลบลวง หรือผลบวกลวงหรือไม่ เพื่อนำค่าที่ได้มาเปรียบเทียบประสิทธิภาพของระบบคัดกรองแต่ละระบบต่อไป

3.3.2 **คัดกรองกลุ่มตัวอย่างอีเมลที่สุ่มจากคลังอีเมล**

กลุ่มอาสาสมัครต้องประเมินอีเมลของตนในกลุ่มอีเมลตัวอย่างที่ถูกสุ่มออกมาจากคลัง อีเมล เพื่อคำนวณหาปริมาณอีเมลขยะที่แท้จริงของคลังอีเมล และนำไปเปรียบเทียบกับปริมาณ อีเมลขยะที่ระบบคัดกรอง สามารถคัดกรองได้

3.4 เตรียมระบบคัดกรองที่ต้องการประเมินประสิทธิภาพ

งานวิจัยนี้สร้างกระบวนวิธีสำหรับประเมินประสิทธิภาพของระบบคัดกรองอีเมลขยะ จะต้องมีระบบคัดกรองอีเมลขยะเข้าร่วมทั้งหมด 3 ระบบ ดังนี้

3.4.1 ระบบเก่า

คือระบบคัดกรองอีเมลขยะที่ใช้งานก่อนการเปลี่ยนระบบคัดกรองอีเมลขยะของสำนักเทคโนโลยีสารสนเทศ จุฬาลงกรณ์มหาวิทยาลัย ในเดือน กันยายน 2551 โดยระบบคัดกรองนี้ใช้โปรแกรมไอเอ็มเอสเอส รุ่นที่ 7 เป็นระบบคัดกรองอีเมลขยะ ให้บริการคัดกรองอีเมลของนิสิต

3.4.2 ระบบปัจจุบัน

คือระบบคัดกรองอีเมลขยะที่ สำนักเทคโนโลยีสารสนเทศ จุฬาลงกรณ์มหาวิทยาลัย เปลี่ยนและใช้งานตั้งแต่เดือน กันยายน 2551 เป็นต้นมา ซึ่งเป็นเครื่องคอมพิวเตอร์เฉพาะสำหรับจัดการระบบอีเมล และคัดกรองอีเมลขยะ ให้บริการแก่คณาจารย์และเจ้าหน้าที่มหาวิทยาลัย

3.4.3 ระบบอ้างอิง

คือระบบที่สร้างขึ้นโดยนำวิธีการคัดกรองอีเมลขยะจากงานวิจัย และวิธีที่ใช้กันอย่างแพร่หลายในปัจจุบัน โดยรายละเอียดแสดงในหัวข้อ 3.5

เนื่องจากการเปรียบเทียบ และวิเคราะห์ผลการประเมินระบบคัดกรองอีเมลขยะจำเป็นต้องมีระบบคัดกรองอีเมลขยะซึ่งเป็นที่ยอมรับ เพื่ออ้างอิงผลลัพธ์ และประสิทธิภาพในการคัดกรอง วิธีการคัดกรองอีเมลขยะมีการวิจัย และพัฒนาอย่างต่อเนื่อง ดังนั้นระบบอ้างอิงจากงานวิจัย ควรพัฒนาให้ทันสมัย เช่น การสร้างหรือปรับปรุงระบบอ้างอิงรุ่นต่อๆ มาให้ใช้วิธีการคัดกรองที่ทันสมัยที่สุดเป็นต้น เพื่อเพิ่มความถูกต้องของผลการประเมิน โดยกระบวนการสร้างระบบอ้างอิงจะกล่าวถึงในลำดับถัดไป

การเปรียบเทียบผลการคัดกรองระหว่าง ระบบเก่า และระบบปัจจุบัน กับระบบอ้างอิง เพื่อเปรียบเทียบผลการตัดกรองระหว่างระบบคัดกรองที่ขายในท้องตลาด กับระบบคัดกรองโอเพนซอร์สซึ่งไม่มีค่าใช้จ่าย อีกทั้งเป็นการนำเสนอทางเลือกใหม่สำหรับระบบคัดกรองอีเมลขยะ

อีเมลที่ถูกคัดกรองโดยระบบคัดกรองทั้ง 3 ประเภทจะถูกตรวจสอบความถูกต้องด้วยตาของกลุ่มอาสาสมัคร

3.5 สร้างระบบคัดกรองอ้างอิงจากงานวิจัย

กระบวนการสร้างระบบคัดกรองอ้างอิงจากงานวิจัยนั้นมีกระบวนการดังต่อไปนี้

3.5.1 ศึกษางานวิจัยเกี่ยวกับอีเมลขยะ

ในปี ค.ศ. 2007 คอร์แม็ค และคณะ [4] ทำการประเมินประสิทธิภาพของโปรแกรมโอเพนซอร์สสำหรับคัดกรองอีเมลขยะและใช้งานกันอย่างแพร่หลาย 6 โปรแกรม ซึ่งให้ความสำคัญกับโปรแกรม สแปมแอสแซสซิน (Spamassassin) เป็นหลัก โดยเปรียบเทียบประสิทธิภาพการคัดกรองของวิธีการคัดกรองต่างๆ ที่โปรแกรมสแปมแอสแซสซินรองรับ

ดังนั้นงานวิจัยนี้จึงนำสแปมแอสแซสซินมาเป็นระบบอ้างอิงของงานวิจัยเพื่อใช้เปรียบเทียบกับผลการคัดกรองของ ระบบเก่า และระบบปัจจุบัน โดยปรับตั้งให้โปรแกรมใช้การค่ามาตรฐานในการทำงาน

3.5.2 สร้างระบบคัดกรองอ้างอิง

การสร้างระบบคัดกรองอ้างอิงใช้คอมพิวเตอร์ส่วนบุคคลติดตั้งโปรแกรม ดังรายละเอียดตามตารางที่ 3.1

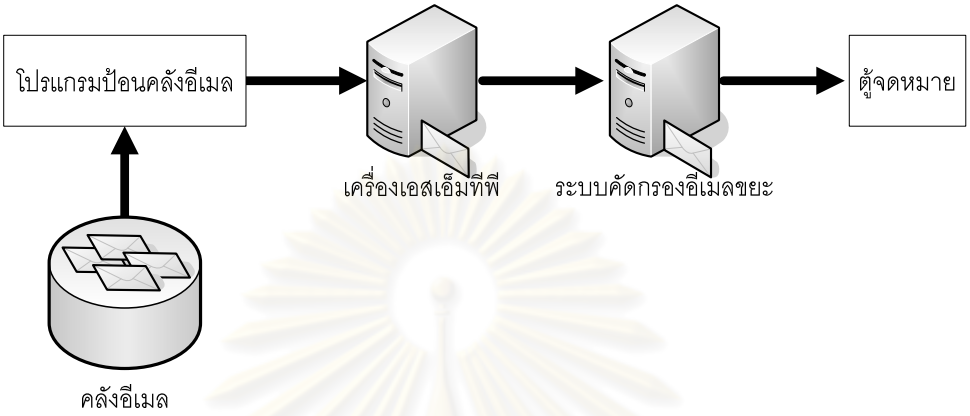
ตารางที่ 3.1 รายละเอียดซอร์ฟแวร์ของระบบคัดกรองอ้างอิง

ประเภท	รายละเอียด
ระบบปฏิบัติการ	ฟรีบีเอสดี รุ่น 7.0 (FreeBSD 7.0)
โปรแกรมแม่ข่ายอีเมล	โพสฟิก รุ่น 2.3 (Postfix 2.3)
โปรแกรมคัดกรองอีเมลขยะ	สแปมแอสแซสซิน รุ่น 3.2.5 (Spamassassin 3.2.5)

หลังจากติดตั้งระบบปฏิบัติการเรียบร้อยแล้ว ติดตั้งโปรแกรมโพสฟิก และโปรแกรมสแปมแอสแซสซิน ใช้โปรแกรมสแปมแอสแซสซินรุ่นที่ 3.2.5 ที่บนระบบปฏิบัติการฟรีบีเอสดี รุ่นที่ 7 การปรับตั้งค่าของโปรแกรมสแปมแอสแซสซินจะใช้ค่ามาตรฐานของโปรแกรม และปรับตั้งให้ระบุผลการคัดกรองไว้ที่หัวข้อของอีเมล โดยรายละเอียดของค่ามาตรฐานแสดงในภาคผนวก จากการปรับตั้งค่าข้างต้นเมื่อพบว่า อีเมลฉบับใดเป็นอีเมลขยะ โปรแกรมจะเพิ่มข้อความ “****SPAM****” ด้านหน้าหัวข้ออีเมลนั้นเพื่อแยกอีเมลดีออกจากอีเมลขยะ

3.6 ป้อนคลังอีเมลผ่านระบบคัดกรองที่ต้องการประเมินประสิทธิภาพ

การป้อนคลังอีเมลผ่านระบบคัดกรองที่ต้องการประเมินประสิทธิภาพ เพื่อให้ระบบคัดกรองวิเคราะห์คลังอีเมล แยกอีเมลดี และอีเมลขยะออกจากกัน ดังมีรายละเอียดดังนี้



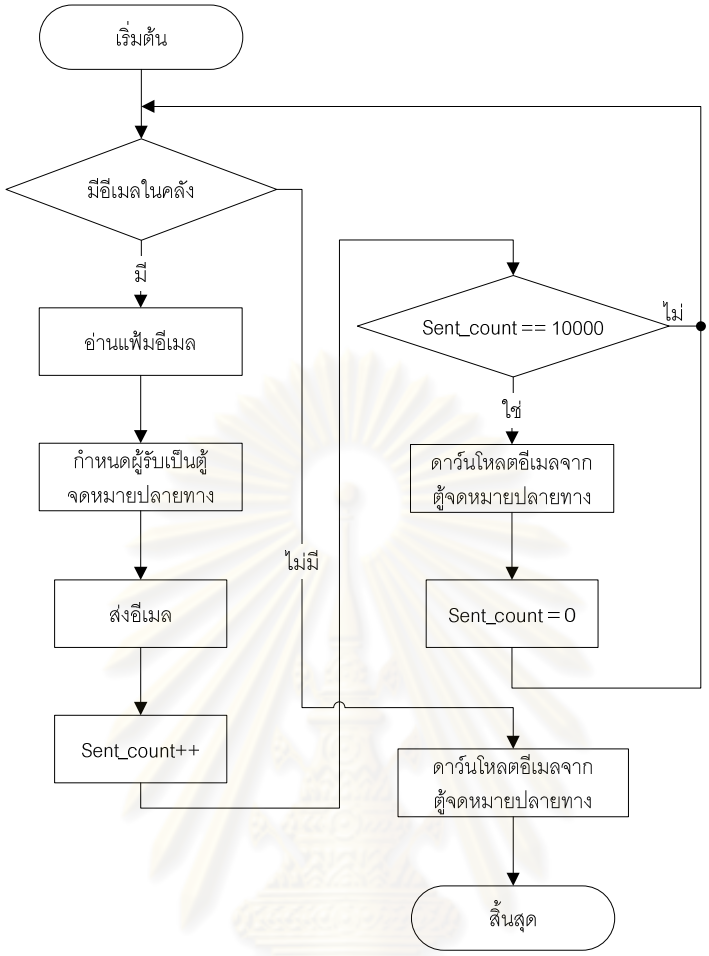
รูปที่ 3.5 การเชื่อมต่อของคอมพิวเตอร์เพื่อป้อนคลังอีเมลผ่านระบบคัดกรอง

จากรูปที่ 3.5 การป้อนคลังอีเมลผ่านระบบคัดกรองอีเมลขยะนั้น งานวิจัยนี้กำหนดให้สามารถป้อนคลังอีเมลผ่านระบบคัดกรองได้เพียงครั้งละ 1 ระบบคัดกรองเท่านั้น จากการศึกษาระบบอีเมลของจุฬาลงกรณ์มหาวิทยาลัยดังกล่าวในตอนต้น การส่งอีเมลผ่านระบบคัดกรองระบบเก่าและระบบปัจจุบัน สามารถทำได้โดยส่งอีเมลไปยังตู้จดหมายของนิสิตเมื่อต้องการป้อนคลังอีเมลผ่านระบบเก่า และส่งอีเมลไปยังตู้จดหมายของคณาจารย์หรือเจ้าหน้าที่เมื่อต้องการป้อนคลังอีเมลผ่านระบบปัจจุบัน ผู้วิจัยได้แจ้งสำนักเทคโนโลยีสารสนเทศ จุฬาลงกรณ์มหาวิทยาลัยให้สร้างตู้จดหมายให้แก่ผู้วิจัยโดยมีสิทธิเทียบเท่าตู้จดหมายของคณาจารย์เพื่อให้รับอีเมลหลังจากผ่านการคัดกรองโดยระบบปัจจุบัน และใช้ตู้จดหมายของผู้วิจัยรับอีเมลซึ่งผ่านการคัดกรองโดยระบบเก่า

สำหรับระบบคัดกรองอ้างอิงนั้นผู้วิจัยได้สร้างระบบคัดกรองอ้างอิงขึ้น โดยติดตั้งภายในห้องทดลอง ทำการป้อนคลังอีเมลผ่านระบบดังกล่าว และสร้างตู้จดหมายเพื่อรับอีเมลจากระบบคัดกรองนั้น ตู้จดหมายสำหรับระบบคัดกรองแต่ละประเภทแสดงดังตารางที่ 3.2

ตารางที่ 3.2 ตู้จดหมายปลายทางที่ผ่านการระบบคัดกรองประเภทต่างๆ

ประเภทของระบบคัดกรอง	ตู้จดหมายปลายทาง
ระบบคัดกรองเก่า	Athakorn.O@Student.Chula.ac.th
ระบบคัดกรองปัจจุบัน	Athakorn.O@Chula.ac.th
ระบบคัดกรองอ้างอิง	Athakorn.O@[หมายเลขไอพีภายใน]



รูปที่ 3.6 การป้อนคลังอีเมลผ่านระบบคัดกรองที่มาทดสอบ

จากรูปที่ 3.6 การป้อนคลังอีเมลผ่านระบบคัดกรองแต่ละประเภทเริ่มต้นโดยอ่านอีเมลจากคลังอีเมลครั้งละฉบับ กำหนดชื่อผู้รับตามตารางที่ 3.2 โดยพิจารณาว่ากำลังทดสอบระบบคัดกรองใด จากนั้นส่งอีเมลดังกล่าวไปยังปลายทาง เมื่อส่งอีเมลครบ 10,000 ฉบับแล้ว ให้ดึงอีเมลจากตู้รับจดหมายปลายทางมาวิเคราะห์ว่าเป็นอีเมลขยะหรือไม่เพื่อป้องกันตู้จดหมายเต็มเป็นผลให้ไม่สามารถส่งอีเมลเข้าไปได้ และเมื่อส่งอีเมลหมดคลังแล้วให้ทำการดึงอีเมลจากตู้รับจดหมายอีกครั้งเพื่อให้แน่ใจว่าอีเมลทั้งหมดในคลังถูกคัดกรองและบันทึกผลการคัดกรองเรียบร้อยแล้ว

3.7 ติดต่อกลุ่มอาสาสมัคร

การติดต่อกลุ่มอาสาสมัครจะทำการติดต่อผ่านทางอีเมลของอาสาสมัครมีการใช้งานอยู่เป็นประจำ โดยส่งอีเมลเพื่อข้อความร่วมมือไปยังกลุ่มอาสาสมัคร

เรื่อง ขอความร่วมมือเพื่อเข้าร่วมงานวิจัยเรื่อง "การประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะโดยผู้ใช้งาน"

เรียน ท่านอาสาสมัคร เจ้าของอีเมล Athakorn.O@Chula.ac.th

ข้าพเจ้านายอรรถกร องค์ศิริพร นิสิตคณะวิศวกรรมศาสตร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ หลักสูตรวิทยาศาสตรมหาบัณฑิตสาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ (ภาคนอกเวลาราชการ) รหัสประจำตัว 5071458921 กำลังทำวิทยานิพนธ์ ในหัวข้อ "การประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะโดยผู้ใช้งาน" โดย มี อ.ดร.ยรรยง เต็งอำนวย (Yunyong.T@Chula.ac.th) เป็นอาจารย์ที่ปรึกษา และได้รับการสนับสนุนจากสำนักเทคโนโลยีสารสนเทศ จุฬาลงกรณ์มหาวิทยาลัย

เนื่องด้วยท่านอาสาสมัครมีการใช้งานอีเมลของ จุฬาลงกรณ์มหาวิทยาลัย อย่างต่อเนื่อง และมีปริมาณเพียงพอ ข้าพเจ้าจึงใคร่ขอความกรุณานำอีเมลของท่านเป็นข้อมูลในการวิจัยและขอความ ร่วมมือท่านอาสาสมัคร เข้าร่วมการทำวิจัยในครั้งนี้ อีเมลของท่านจะถูกเก็บไว้เป็นความลับโดยมีเพียงท่านอาสาสมัครเท่านั้น ที่สามารถดูอีเมลของท่านได้

การร่วมทำวิจัยท่านอาสาสมัครต้องพิจารณาอีเมลของท่านเองด้วยตา เป็นจำนวน 20 ฉบับ และตัดสินว่าเป็นอีเมลขยะ (SPAM) หรือไม่ ผู้วิจัยขออนุญาตเก็บข้อมูลส่วนตัวของท่านอันประกอบด้วย เพศ อายุ สถานภาพ เพื่อประกอบการวิเคราะห์ โดยจะดำเนินการในช่วงเดือน ธันวาคม 2552 และการทำวิจัยในครั้งนี้จะไม่สร้างภาระให้แก่ท่านอาสาสมัครจนเกินไป

ผู้จัดทำจะประสานงานกับท่านอาสาสมัครผ่านทางอีเมล การเข้าร่วมการวิจัยจะทำผ่านทางเว็บไซต์ isel.cp.eng.chula.ac.th/SpamFilterEvaluationWeb

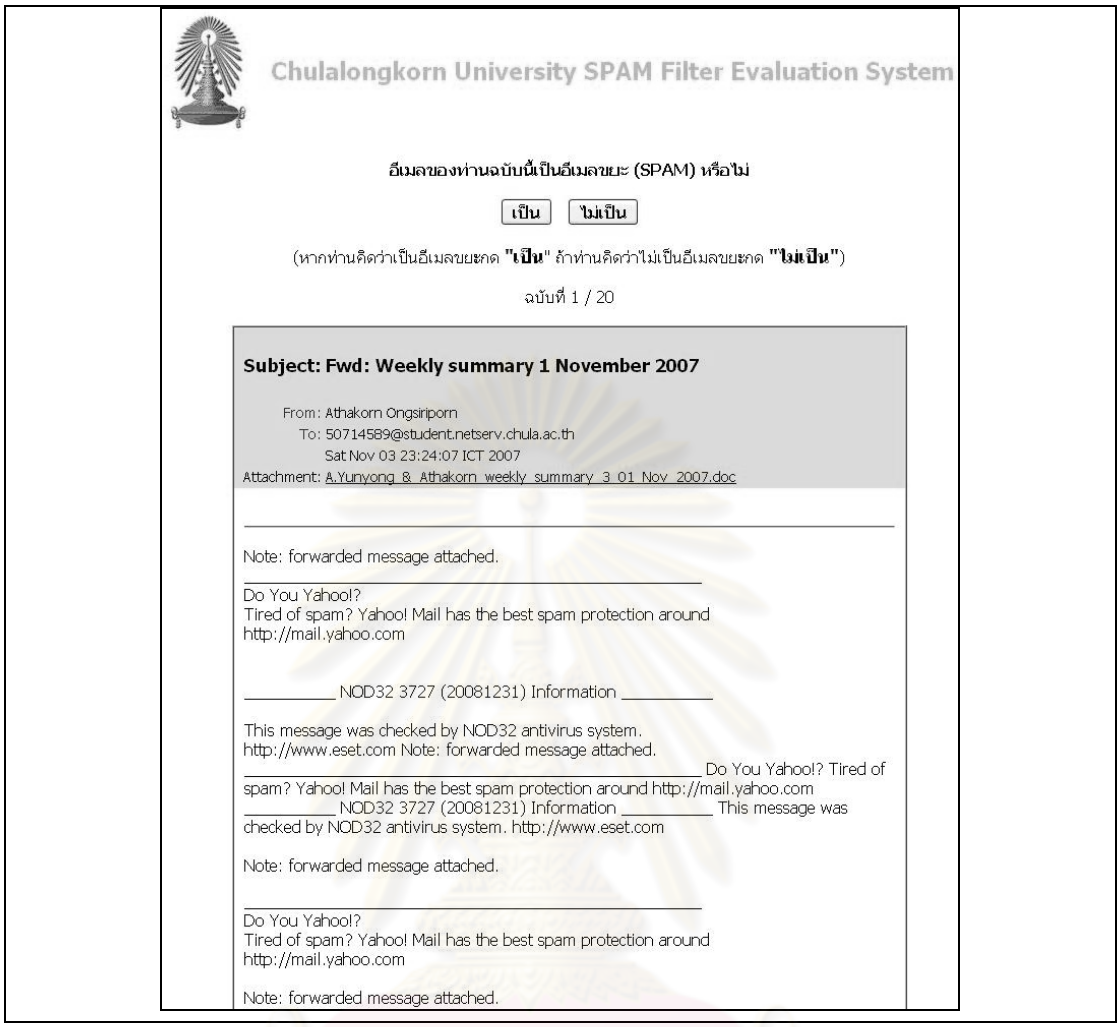
ผู้จัดทำขอขอบพระคุณในความร่วมมือของท่าน

ยินดีเข้าร่วม - กรุณาตอบอีเมล (Reply) ฉบับนี้

ขอแสดงความนับถืออย่างสูง

(นายอรรถกร องค์ศิริพร)

<u>ตัวอย่างหน้าจอสำหรับประเมินอีเมลของตนเองด้วยตา</u>



รูปที่ 3.7 ตัวอย่างจดหมายขอความร่วมมือ

จากรูปที่ 3.7 กลุ่มอาสาสมัครจะได้รับจดหมายขอความร่วมมือจากผู้ทำการวิจัย เนื้อหาของจดหมายจะชี้แจงถึงรายละเอียดของการเข้าร่วมงานวิจัยในครั้งนี้ประกอบด้วย ข้อมูลส่วนตัวของผู้วิจัย เหตุผลที่อาสาสมัครท่านนั้นถูกเลือกให้เข้าร่วมการวิจัย รายละเอียดของการวิจัย และช่วงเวลาในการวิจัย

ระบบจะส่งอีเมลไปยังอาสาสมัครที่ยินดีเข้าร่วมงานวิจัยอีกครั้งเพื่อแจ้งวิธีการประเมินระบบแก่อาสาสมัคร

เรียน ท่านอาสาสมัคร เจ้าของอีเมล Athakorn.O@Chula.ac.th

ขอบพระคุณสำหรับการร่วมงานวิจัย "การประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะโดยผู้ใช้งาน"

เริ่มต้นการประเมินอีเมลของท่านที่

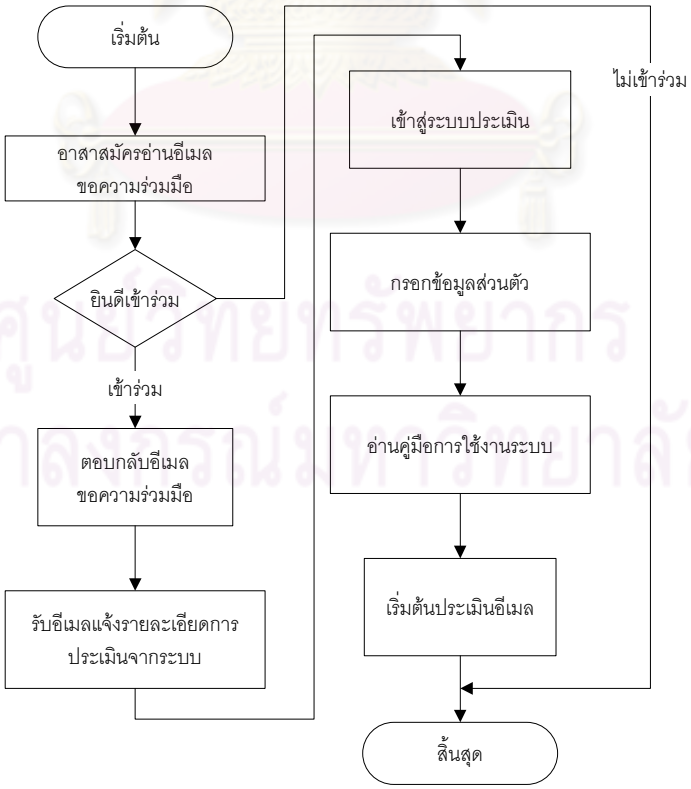
http://isel.cp.eng.chula.ac.th/SpamFilterEvaluationWeb?acceptedEmail=QXRoYWtvcm4uT0BTdHVkZW50LmNodWxhLmFjLnRo

ขอแสดงความนับถืออย่างสูง

(นายอรรถกร องค์ศิริพร)

รูปที่ 3.8 ตัวอย่างจดหมายตอบกลับจากระบบ

รูปที่ 3.8 แสดงจดหมายตอบกลับจากผู้ทำการวิจัย เนื้อหาของจดหมายแจ้งที่อยู่เว็บไซต์ที่อาสาสมัครจะใช้ประเมินอีเมลของตนด้วยตา



รูปที่ 3.9 การเข้าร่วมงานวิจัยของอาสาสมัคร

จากรูปที่ 3.9 เมื่ออาสาสมัครได้รับจดหมายขอความร่วมมือเรียบร้อยแล้ว หากอาสาสมัครยินดีเข้าร่วมงานวิจัย อาสาสมัครจะต้องตอบกลับอีเมลขอความร่วมมือมายังผู้ส่ง และได้รับอีเมลแจ้งที่อยู่เว็บไซต์สำหรับประเมินอีเมลของตนด้วยตา

เมื่ออาสาสมัครเข้าไปในเว็บไซต์ต้องแสดงตัวโดยการเข้าสู่ระบบ และใช้ชื่อผู้ใช้และรหัสผ่านสำหรับการตรวจสอบอีเมลของจุฬาลงกรมหาวิทยาลัยในการเข้าสู่ระบบ เมื่อทำการเข้าสู่ระบบเรียบร้อยอาสาสมัครต้องกรอกข้อมูลส่วนตัวเพื่อประกอบในการวิเคราะห์ผลการวิจัย ประกอบด้วย เพศ อายุ และสถานภาพ เมื่อกรอกข้อมูลเรียบร้อย ระบบจะแสดงคู่มือการใช้งานแก่อาสาสมัคร เมื่ออาสาสมัครอ่านคู่มือเรียบร้อยระบบจึงเริ่มต้นแสดงอีเมลให้อาสาสมัครประเมินทันที

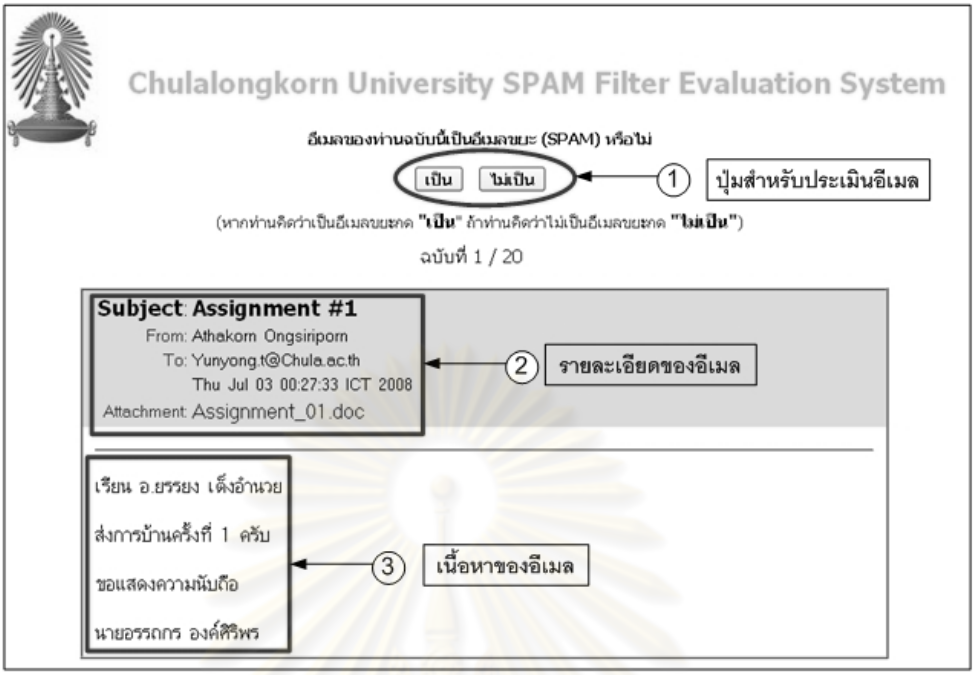
3.8 สุ่มอีเมลของอาสาสมัครแต่ละคนและประเมินอีเมลของตนด้วยตา

งานวิจัยนี้ใช้อาสาสมัครที่เป็นผู้ใช้งานอีเมลขององค์กรเป้าหมายในกระบวนการประเมินประสิทธิภาพ โดยมีหน้าที่ดังกล่าวไปแล้วในหัวข้อ 3.3.1 และ 3.3.2 เมื่ออาสาสมัครผ่านขั้นตอนต่างๆ ดังแสดงในผังงานรูปที่ 3.9 เรียบร้อยแล้ว ระบบจะสุ่มอีเมลของอาสาสมัครคนนั้นขึ้นมาให้ประเมินว่าอีเมลฉบับดังกล่าวเป็นอีเมลขยะหรือไม่

จากผลการวิเคราะห์ระบบปัจจุบันตามหัวข้อ 3.1 แสดงให้เห็นว่าอาสาสมัครแต่ละคนสามารถมีอีเมลแอดเดรสได้มากกว่า 1 ตัว ดังนั้นการสุ่มอีเมลให้อาสาสมัครแต่ละคนประเมินด้วยตาต้องสุ่มจากทุกอีเมลแอดเดรสที่อาสาสมัครคนนั้นเป็นเจ้าของ

อาสาสมัครแต่ละคนสามารถได้รับจดหมายขอความร่วมมือมากกว่า 1 ฉบับจากผู้วิจัย โดยแต่ละฉบับจะระบุถึงอีเมลที่อาสาสมัครผู้นั้นเป็นเจ้าของ เมื่ออาสาสมัครตอบรับยินดีเข้าร่วมการวิจัยแล้ว การสุ่มอีเมลให้อาสาสมัครพิจารณาด้วยตาต้องสุ่มจากอีเมลทั้งหมดที่อาสาสมัครเป็นเจ้าของหรือเป็นสมาชิกอีเมลกลุ่มนั้น ระบบจะแสดงอีเมลที่ถูกสุ่ม ให้อาสาสมัครดูด้วยตา และตัดสินว่าอีเมลฉบับนั้นเป็น อีเมลขยะหรือไม่

การสุ่มอีเมลจากคลังอีเมลนั้นใช้วิธีการสุ่มตัวอย่างแบบง่าย (Simple Random Sampling) ซึ่งถือว่าอีเมลทุกฉบับของอาสาสมัครมีความน่าจะเป็นเท่ากันที่จะถูกเลือกให้อาสาสมัครประเมินในแต่ละครั้ง อีเมลที่ถูกประเมินไปแล้วจะไม่ถูกสุ่มขึ้นมาซ้ำอีก และระบบจะสุ่มอีเมลเฉพาะที่อาสาสมัครคนนั้นๆ เป็นเจ้าของเพื่อป้องกันปัญหาความเป็นส่วนตัว



รูปที่ 3.10 ตัวอย่างหน้าจอสำหรับประเมินอีเมลของตนด้วยตา

จากรูปที่ 3.10 หน้าจอสำหรับประเมินอีเมลของตนด้วยตา ถูกออกแบบให้ใช้งานได้ง่าย และไม่สร้างภาระให้แก่อาสาสมัครมากจนเกินไป ซึ่งประกอบด้วย 3 ส่วนสำคัญคือ

1. ปุ่มสำหรับประเมินอีเมล ประกอบด้วย ปุ่ม “เป็น” และ “ไม่เป็น” อาสาสมัครจะกดปุ่มเพื่อตัดสินว่าอีเมลที่ถูกแสดงอยู่นั้น เป็น หรือไม่เป็น อีเมลขยะ

2. รายละเอียดของอีเมล แสดงรายละเอียดของอีเมลที่ถูกแสดงอยู่ในขณะนั้น ประกอบด้วย

- หัวข้อของอีเมล (Subject)
- ผู้ส่ง (From)
- ผู้รับ (To)
- วันที่
- แฟ้มแนบ (Attachment)

3. เนื้อหาของอีเมล แสดงเนื้อหาของอีเมลที่ถูกสุ่มขึ้นมาให้ประเมินในขณะนั้น

อาสาสมัครแต่ละคนต้องทำการวิเคราะห์อีเมลของตนเป็นจำนวน 20 ฉบับ แต่สามารถประเมินต่อไปได้หากอาสาสมัครต้องการ ทั้งนี้การไม่สร้างภาระให้แก่อาสาสมัครมากจนเกินไปทำให้สามารถประเมินได้บ่อยครั้งตามต้องการ

3.9 วิเคราะห์และสรุปผลการประเมินประสิทธิภาพ

หลังจากอีเมลถูกคัดกรองโดยระบบคัดกรองแล้วต้องนำมาเปรียบเทียบกับผลการประเมินด้วยตาโดยอาสาสมัคร การวิเคราะห์ผลการประเมินจะต้องกำหนดความหมายของผลการคัดกรองโดยระบบคัดกรองและการประเมินด้วยอาสาสมัคร โดยกำหนดความหมายของค่า True positive (TP), False positive (FP), True negative (TN) และ False negative (FN) ดังแสดงในตารางที่ 3.3 และความหมายทางกายภาพดังแสดงในตารางที่ 3.4

ตารางที่ 3.3 ตารางคอนฟิวชั่นของผลการคัดกรอง

		ประเมินด้วยอาสาสมัคร	
		อีเมลขยะ	**อีเมลดี**
คัดกรองโดยระบบคัดกรอง	**อีเมลขยะ**	TP (%)	FP (%)
	อีเมลดี	FN (%)	TN (%)

ตารางที่ 3.4 ความหมายทางกายภาพของผลการคัดกรอง

ผลการคัดกรอง	**ความหมายทางกายภาพ**
TP	สามารถระบุอีเมลขยะได้
TN	สามารถระบุอีเมลดีได้
FP	ระบุอีเมลดีว่าเป็นอีเมลขยะ (อีเมลดีสูญหาย)
FN	ระบุอีเมลขยะว่าเป็นอีเมลดี (อีเมลขยะหลุดลอด)

งานวิจัยนี้วิเคราะห์และสรุปผลการประเมินประสิทธิภาพของระบบคัดกรองอีเมลขยะในประเด็นต่างๆ ดังต่อไปนี้

3.9.1 ประสิทธิภาพโดยรวมของระบบคัดกรอง

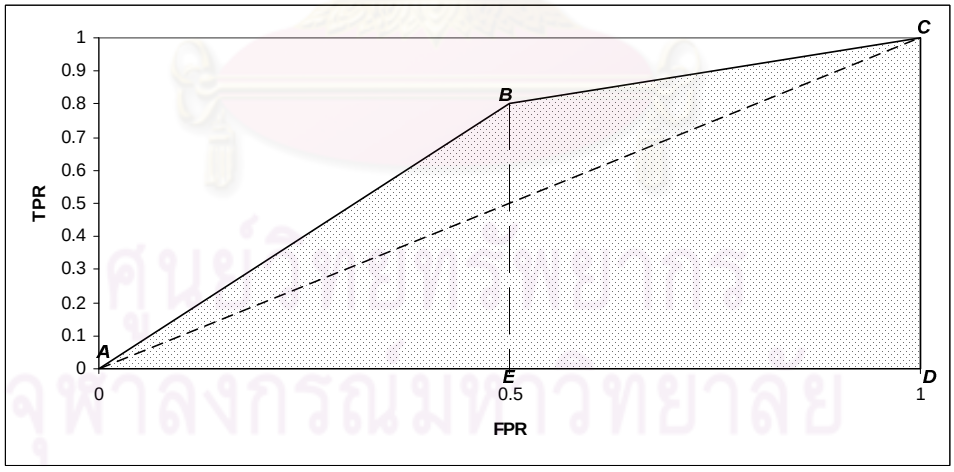
ใช้วิธีทางสถิติวิเคราะห์ประสิทธิภาพของระบบคัดกรองอีเมลขยะ โดยใช้แผนภูมิ Receiver operating characteristic (ROC) [16] และคำนวณพื้นที่ใต้กราฟ (Area under an ROC curve, AUC) เปรียบเทียบค่า AUC ของแต่ละระบบ ระบบใดมีค่า AUC สูงกว่าหมายถึงมีประสิทธิภาพในการคัดกรองอีเมลขยะที่ดีกว่า

แผนภูมิ ROC วาดค่า False positive rate (FPR) บนแกน X และค่า True positive rate (TPR) บนแกน Y ค่า FPR และ TPR คำนวนได้จากสมการที่ (1) และ (2) ตามลำดับ

$$\text{False positive rate (FPR)} = \frac{FP}{(FP+TN)} \quad (1)$$

$$\text{True positive rate (TPR)} = \frac{TP}{(TP+FN)} \quad (2)$$

การวิเคราะห์ด้วยตาของอาสาสมัครนั้น จัดเป็นตัวคัดกรองแบบดีสครีต (Discrete classifier) [16] ซึ่งมีตารางคอนฟิวชั่นเพียงตารางเดียว จึงมีเพียง 1 จุดบนแผนภูมิ ROC



รูปที่ 3.11 พื้นที่ใต้แผนภูมิ ROC

จากรูปที่ 3.11 แสดงการพื้นที่ใต้แผนภูมิ ROC สามารถคำนวณพื้นที่ใต้กราฟได้จากสมการที่ (3)

AUC = พื้นที่ ABE + พื้นที่ BCDE (3)

3.9.2 การวิเคราะห์จุดแข็งและจุดด้อยของระบบคัดกรอง

งานวิจัยนี้จะวิเคราะห์จุดแข็งและจุดด้อยของระบบคัดกรองแต่ละตัวโดยวิเคราะห์จากค่า TPR และ FPR โดยมีรายละเอียดดังนี้

3.9.2.1 ระบบคัดกรองที่มีอัตราอีเมลดีสูญหายน้อยที่สุด

พิจารณาได้จากค่า FPR ของแต่ละระบบคัดกรอง คำนวณได้จากสมการที่ (1) ระบบใดมีค่า FPR ต่ำกว่าหมายถึงมีอัตราอีเมลดีสูญหายน้อยกว่า แต่ปริมาณอีเมลขยะอาจเพิ่มขึ้น

3.9.2.2 ระบบคัดกรองที่มีอัตราการถูกรบกวนจากอีเมลขยะน้อยที่สุด

พิจารณาได้จากค่า TPR ของแต่ละระบบคัดกรอง คำนวณได้จากสมการที่ (2) ระบบใดมีค่า TPR สูงกว่าหมายถึงมีอัตราการคัดกรองอีเมลที่ดีกว่า แต่อาจมีอีเมลหายบางส่วน

การวิเคราะห์ดังกล่าวจะเป็นทางเลือกให้แก่องค์กรเพื่อเลือกระบบคัดกรองที่เหมาะสมกับความต้องการขององค์กร

บทที่ 4

ผลการวิจัย

จากวิธีดำเนินงานวิจัย ผลการวิจัยมีรายละเอียดประกอบด้วย

1. ผลการรวบรวมและวิเคราะห์คลังอีเมลและการคัดเลือกอาสาสมัคร
2. ผลการคัดกรองโดยระบบคัดกรอง
3. ผลการประเมินอีเมลของตนด้วยตาของอาสาสมัคร
4. ผลการวิเคราะห์และเปรียบเทียบประสิทธิภาพระบบคัดกรองอีเมลขยะ

4.1 ผลการรวบรวมและวิเคราะห์คลังอีเมลและการคัดเลือกอาสาสมัคร

การรวบรวมคลังอีเมลจากระบบเครือข่ายขององค์กรเป้าหมายนั้นต้องรวบรวมอีเมลขาเข้าทั้งอีเมลดีและอีเมลขยะ ระบบอีเมลที่ใช้งานอยู่ในปัจจุบันไม่สามารถสำเนาอีเมลขาเข้าทั้งหมดได้ทันทีเนื่องจากข้อจำกัดของระบบและการรักษาความปลอดภัย อย่างไรก็ดีระบบสามารถสำเนาอีเมลที่ถูกคัดกรองว่าเป็นอีเมลขยะและอีเมลดีหลังจากทำการคัดกรองแล้วออกมาได้ ทำให้สามารถรวบรวมได้ทั้งอีเมลดีและอีเมลขยะ ดังนั้นจึงตั้งค่าของอีเมลเกตเวย์ให้สำเนามาที่ตู้จดหมายปลายทางเพื่อนำมาใช้เป็นคลังอีเมล

อีเมลที่ถูกคัดลอกจากตู้จดหมายถูกแยกออกเป็น 1 แฟ้มต่ออีเมล 1 ฉบับ วิธีการคัดลอกนั้นสามารถทำได้หลายวิธี เช่น คัดลอกจากตู้จดหมายโดยตรงบนเครื่องแม่ข่าย หรือใช้โปรแกรมอีเมลไคลแอนต์ดาวน์โหลดอีเมลทั้งหมด

ในงานวิจัยนี้ตู้จดหมายที่รวบรวมคลังอีเมลจัดเก็บอีเมล 1 แฟ้มต่อ 1 อีเมลแต่การคัดลอกโดยตรงทำได้ลำบากเนื่องจากต้องใช้สิทธิของผู้ดูแลระบบเท่านั้นในการคัดลอก ผู้วิจัยจึงใช้โปรแกรมอีเมลไคลแอนต์ดาวน์โหลดอีเมลทั้งหมดมาที่เครื่องคอมพิวเตอร์ส่วนบุคคล แต่โปรแกรมอีเมลไคลแอนต์นั้นเก็บอีเมลทุกฉบับในแฟ้มเดียวด้วยรูปแบบของแฟ้มเอ็มบ๊อก (mbox) การแยกออกเป็น 1 แฟ้มต่อ 1 อีเมลใช้โปรแกรม git-mailsplit บนระบบปฏิบัติการฟรีบีเอสดีแยกอีเมล โดยตั้งชื่อแฟ้มของอีเมลแต่ละฉบับเป็นลำดับตัวเลขคือ 1, 2, 3 ตามลำดับ โดยมีรายละเอียดของการรวบรวมอีเมลจากองค์กรเป้าหมายดังแสดงในตารางที่ 4.1

ตารางที่ 4.1 รายละเอียดของคลังอีเมล

หัวข้อ	ข้อมูล
เริ่มต้นรวบรวม	14 กรกฎาคม 2552
รวบรวมถึง	5 สิงหาคม 2552
รวมระยะเวลา	22 วัน
ขนาดคลังอีเมล	134,167 ฉบับ
จำนวนอีเมลถึงตู้จดหมายผู้รับ	165, 085 ฉบับ (ฉบับเดียวถึงหลายคน)

หลังจากรวบรวมคลังอีเมลแล้วจึงทำการวิเคราะห์เพื่อหาอีเมลแอดเดรสที่มีปริมาณอีเมลขาเข้ามากที่สุด โดยใช้วิธีในข้อ 3.2 และใช้ตารางแจกแจงความถี่วิเคราะห์หาอีเมลแอดเดรสที่มีปริมาณอีเมลที่เหมาะสมสำหรับเป็นกลุ่มอาสาสมัคร

ตารางที่ 4.2 ตารางแจกแจงความถี่ของปริมาณอีเมลขาเข้าแต่ละอีเมลแอดเดรส

ชั้นที่	จำนวนอีเมล	อีเมลแอดเดรส	ชั้นที่	จำนวนอีเมล	อีเมลแอดเดรส
1	1-50	7,517	26	1,251-1,300	1
2	51-100	653	27	1,301-1,350	2
3	101-150	238	28	1,351-1,400	0
4	151-200	96	29	1,401-1,450	0
5	201-250	50	30	1,451-1,500	0
6	251-300	27	31	1,501-1,550	0
7	301-350	25	32	1,551-1,600	0
8	351-400	12	33	1,601-1,650	2
9	401-450	7	34	1,651-1,700	1
10	451-500	1	35	1,701-1,750	0
11	501-550	1	36	1,751-1,800	0
12	551-600	2	37	1,801-1,850	0
13	601-650	0	38	1,851-1,900	0
14	651-700	2	39	1,901-1,950	2
15	701-750	3	40	1,951-2,000	0
16	751-800	0	41	2,001-2,050	0
17	801-850	0	42	2,051-2,100	0

ตารางที่ 4.2 (ต่อ)

ชั้นที่	จำนวนอีเมล	อีเมลแอดเดรส	ชั้นที่	จำนวนอีเมล	อีเมลแอดเดรส
18	851-900	0	43	2,101-2,150	0
19	901-950	1	44	2,151-2,200	0
20	951-1,000	1	45	2,201-2,250	0
21	1,001-1,050	0	46	2,251-2,300	1
22	1,051-1,100	0	47	2,301-2,350	0
23	1,101-1,150	0	48	2,351-2,400	2
24	1,151-1,200	0	49	2,401-2,450	0
25	1,201-1,250	0	50	2,451-2,500	1

จากตารางที่ 4.2 อีเมลส่วนใหญ่มีปริมาณอีเมลขาเข้า 1 ถึง 50 ฉบับในระยะเวลา 22 วัน ซึ่งคิดเป็น 2.27 ฉบับต่อวัน และเมื่อดูจากข้อมูลดิบแล้วพบว่าส่วนมากมีจำนวน 1 ฉบับเท่านั้น แสดงถึงปริมาณการใช้งานที่ต่ำเกินไป ดังนั้นจึงใช้เจ้าของอีเมลแอดเดรสที่มีปริมาณอีเมลขาเข้ามากกว่า 50 ฉบับ เป็นกลุ่มอาสาสมัครในการวิจัยนี้โดยมีจำนวนทั้งสิ้น 849 อีเมลแอดเดรส

4.2 ผลการคัดกรองโดยระบบคัดกรอง

งานวิจัยนี้ทำการทดสอบระบบคัดกรอง 3 ประเภท คือระบบเก่า ระบบปัจจุบัน และระบบอ้างอิง ขณะทำการทดลองเป็นระบบคัดกรองเก่าและระบบปัจจุบัน กำลังใช้งานอยู่ในองค์กรเป้าหมายดังนั้นการทดลองต้องคำนึงถึงผลกระทบต่อผู้ที่กำลังใช้งาน โดยไม่ป้อนคลังอีเมลผ่านระบบคัดกรองดังกล่าวในชั่วโมงทำงาน

ระบบปัจจุบันที่ใช้งานในองค์กรเป้าหมายนั้นเป็นระบบคัดกรองใหม่ที่นำมาใช้งานจึงมีความทันสมัยและมีระบบรักษาความปลอดภัยที่ดี การป้อนอีเมลผ่านระบบดังกล่าวจำเป็นต้องป้อนครั้งละน้อยๆ การป้อนอีเมลปริมาณมากจะทำให้ระบบคัดกรองบันทึกหมายเลขไอพีของเครื่องเอสเอ็มทีพีในบัญชีดำเนื่องจากระบบพบว่าเครื่องเอสเอ็มทีพีมีแนวโน้มในการส่งอีเมลขยะเข้ามาทำให้ไม่สามารถป้อนอีเมลผ่านไปได้ สำหรับระบบเก่าสามารถป้อนคลังอีเมลได้อย่างสะดวกไม่มีปัญหาดังเช่นระบบคัดกรองปัจจุบัน ด้านระบบอ้างอิงเป็นระบบที่สร้างขึ้นภายในห้องทดลองดังนั้นจึงสามารถควบคุมและตั้งค่าต่างๆ ได้ การป้อนคลังอีเมลผ่านจึงทำได้ง่ายที่สุด ผลการคัดกรองแสดงดังตารางที่ 4.3

ตารางที่ 4.3 ผลการคัดกรองของระบบคัดกรองอีเมลขยะ

ระบบคัดกรอง	อีเมลขยะ	ร้อยละอีเมลขยะ	อีเมลดี	ร้อยละอีเมลดี	รวม (ฉบับ)
ระบบเก่า	85,417	51.74%	79,668	48.26%	165,085
ระบบปัจจุบัน	67,262	40.74%	97,823	59.26%	165,085
ระบบอ้างอิง	82,572	50.02%	82,513	49.98%	165,085

จากตารางที่ 4.3 ระบบคัดกรองอ้างอิงและระบบคัดกรองเก่าให้ผลการคัดกรองที่ใกล้เคียงกัน แต่ผลการคัดกรองข้างต้นจะต้องนำไปเปรียบเทียบกับการประเมินด้วยตาของอาสาสมัครเพื่อตรวจสอบความถูกต้องของผลการคัดกรอง เพื่อใช้ประเมินประสิทธิภาพของแต่ละระบบ

ในช่วงเริ่มต้นทดลองการป้อนคลังอีเมลบางครั้งเกิดปัญหา เช่น โปรแกรมสำหรับป้อนคลังอีเมลยังมีข้อผิดพลาดบางส่วน เป็นผลให้ทำงานผิดพลาดและจัดการกับข้อผิดพลาดไม่เหมาะสมทำให้ต้องทดลองซ้ำบ่อยครั้ง แต่เมื่อทำการทดลองมาเป็นระยะเวลาหนึ่งโปรแกรมมีเสถียรภาพมากขึ้น และข้อผิดพลาดถูกแก้ไข ทำให้ปัญหาในการทดลองลดน้อยลง

4.3 ผลการประเมินอีเมลของตนด้วยตาของอาสาสมัคร

จากผลการวิเคราะห์ในหัวข้อ 4.1 อาสาสมัครของงานวิจัยนี้คือเจ้าของอีเมลที่มีปริมาณอีเมลขาเข้ามากกว่า 50 ฉบับ จากหัวข้อ 3.7 เมื่อส่งอีเมลขอความร่วมมือและขออนุญาตไปยังอาสาสมัครที่มีปริมาณอีเมลขาเข้าตามจำนวนที่กำหนด อาสาสมัครบางส่วนตอบรับเข้าร่วมในทันที ซึ่งรายละเอียดการติดต่อไปยังอาสาสมัครแสดงดังตารางที่ 4.4

ตารางที่ 4.4 ผลการติดต่ออาสาสมัคร

หัวข้อ	จำนวน
จำนวนอาสาสมัครที่ติดต่อ	849 คน
จำนวนอาสาสมัครที่ตอบรับ	60 คน
ร้อยละของอาสาสมัครที่ตอบรับ	7.06

มีอาสาสมัครตอบรับเข้าร่วมงานวิจัยทั้งสิ้น 60 คน คิดเป็นร้อยละ 7.06 ของกลุ่มอาสาสมัครที่ปริมาณอีเมลตามเกณฑ์ อาสาสมัครบางกลุ่มทำการประเมินด้วยตาในทันทีหลังจากตอบรับเข้าร่วมงานวิจัย ซึ่งรายละเอียดของการประเมินและผลการประเมินแสดงดังตารางที่ 4.5

ตารางที่ 4.5 รายละเอียดของการประเมินด้วยตาของอาสาสมัคร

หัวข้อ	จำนวน
จำนวนอีเมลที่ถูกประเมิน	2,783 ฉบับ
ร้อยละของจำนวนอีเมลที่ถูกประเมิน	1.69
เฉลี่ยประเมินคนละ (กำหนดไว้ที่ 20 ฉบับ)	46.38 ฉบับ
ระยะเวลา	14 วัน

จากตารางที่ 4.5 อาสาสมัครที่เข้าร่วมงานวิจัยทั้ง 60 คน ประเมินอีเมลรวมกันทั้งหมด 2,783 ฉบับ คิดเป็นร้อยละ 1.69 ของปริมาณอีเมลในคลัง อาสาสมัครแต่ละคนทำการประเมินโดยเฉลี่ย 46.38 ฉบับ จากข้อมูลข้างต้นอาสาสมัครให้การตอบรับอย่างดีต่อกระบวนการประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะขององค์กรเป้าหมาย เนื่องจากอาสาสมัครมีอัตราการประเมินอีเมลที่สูง การประเมินโดยอาสาสมัครใช้เวลาทั้งสิ้น 14 วัน ตั้งแต่เริ่มต้นติดต่อไปยังอาสาสมัคร

จากหัวข้อ 3.3 หน้าที่ของอาสาสมัครในการประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะมี 2 ประการ จากหัวข้อ 3.3.2 ผลการคัดกรองกลุ่มตัวอย่างอีเมลที่สุ่มจากคลังอีเมลแสดงดังตารางที่ 4.6

ตารางที่ 4.6 ผลการคัดกรองตัวอย่างอีเมลของอาสาสมัคร

อีเมลขยะ (ฉบับ)	ร้อยละอีเมลขยะ	อีเมลดี (ฉบับ)	ร้อยละอีเมลดี	รวม (ฉบับ)
1,604	57.64%	1,179	42.36%	2,783

จากตารางที่ 4.6 จำนวนอีเมลที่ถูกประเมินคือ 2,783 ฉบับ เป็นอีเมลขยะ 1,604 ฉบับ คิดเป็นร้อยละ 57.64 ของอีเมลที่ถูกประเมิน และเป็นอีเมลดี 1,179 ฉบับ คิดเป็นร้อยละ 42.36 ของอีเมลที่ถูกประเมิน เมื่อเปรียบเทียบข้อมูลจากตารางที่ 4.3 และตารางที่ 4.6 พบว่า ผลการคัดกรองของระบบคัดกรองเก่าและระบบคัดกรองอ้างอิงนั้นใกล้เคียงกับผลการคัดกรองของอาสาสมัคร อย่างไรก็ดีการวิเคราะห์ถึงประสิทธิภาพของระบบคัดกรองทั้งสาม จำต้องวิเคราะห์ในรายละเอียดดังได้อธิบายไว้ในข้อ 3.9

จากวิธีการกำหนดขนาดกลุ่มตัวอย่างของ ทาโร่ ยามาเน่ [9] อีเมลที่ถูกประเมินด้วยตาของอาสาสมัครมีปริมาณมากพอเพื่อใช้เป็นกลุ่มตัวอย่างของคลังอีเมลในงานวิจัยนี้ที่ระดับความ

เชื่อมั่น 95% ±2% ดังนั้นสามารถใช้เป็นตัวแทนอีเมลทั้งหมดในคลังอีเมลได้ และจะนำมาเปรียบเทียบกับผลการคัดกรองของระบบคัดกรองอีเมลขยะต่อไป

4.4 ผลการวิเคราะห์และเปรียบเทียบประสิทธิภาพระบบคัดกรองอีเมลขยะ

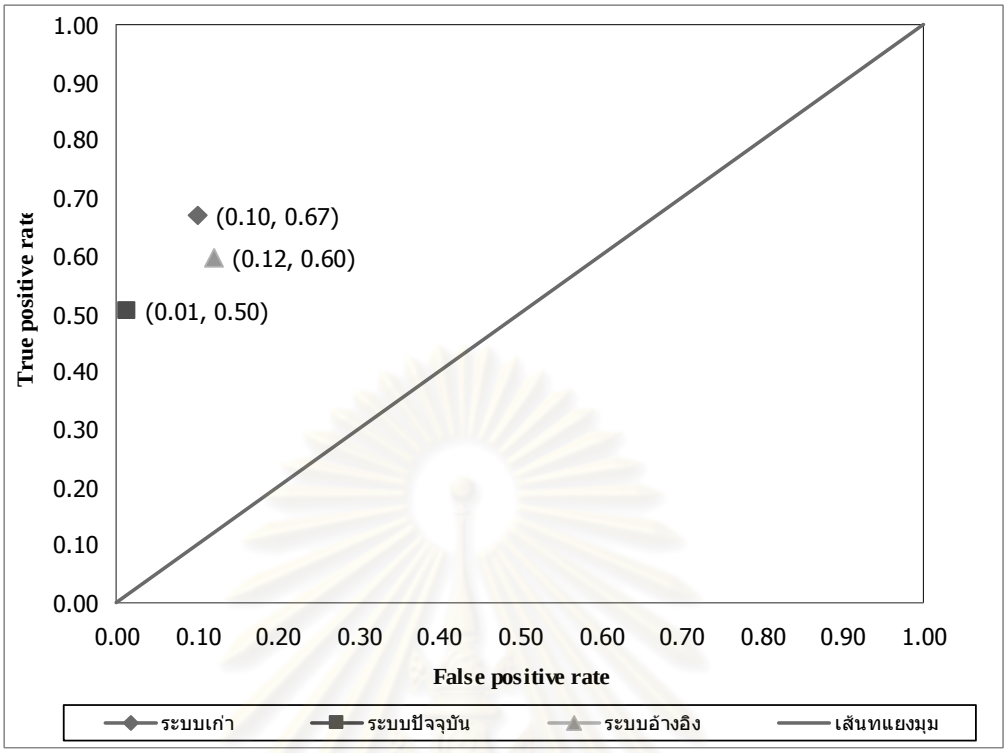
การวิเคราะห์ประสิทธิภาพของระบบคัดกรองใช้วิธีตามหัวข้อ 3.9 โดยคำนวณค่าตามตารางคอนติงเจนซีของแต่ละระบบเพื่อวิเคราะห์ประสิทธิภาพของแต่ละระบบคัดกรอง ผลการคำนวณแสดงตามตารางที่ 4.7 นำค่าจากตารางที่ 4.7 มาคำนวณค่า TPR และ FPR เพื่อวาดแผนภูมิ ROC ตารางที่ 4.8 แสดงค่าทั้งสองของแต่ละระบบคัดกรอง และวาดแผนภูมิ ROC ดังรูปที่ 4.1

ตารางที่ 4.7 ตารางคอนติงเจนซีของแต่ละระบบคัดกรองอีเมลขยะ

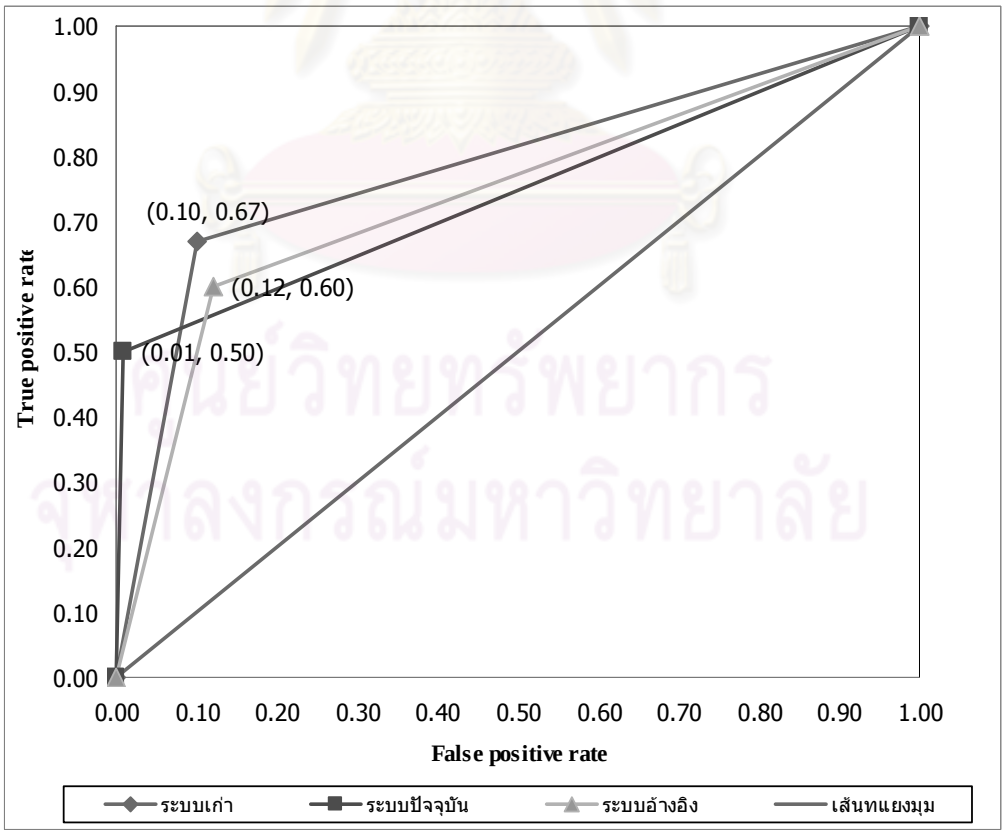
ระบบคัดกรอง	TP	FP	FN	TN
ระบบเก่า	38.52%	4.31%	19.12%	38.05%
ระบบปัจจุบัน	29.03%	0.54%	28.60%	41.83%
ระบบอ้างอิง	34.35%	5.17%	23.28%	37.19%

ตารางที่ 4.8 ค่า FPR และ TPR ของแต่ละระบบคัดกรอง

ระบบคัดกรอง	FPR	TPR
ระบบเก่า	0.10	0.67
ระบบปัจจุบัน	0.01	0.50
ระบบอ้างอิง	0.12	0.60



รูปที่ 4.1 แผนภูมิ ROC ของการประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะ



รูปที่ 4.2 แผนภูมิ ROC ของการประเมินสำหรับคำนวณค่า AUC

จากรูปที่ 4.1 ลากเส้นจากจุด (0,0) ผ่านจุด (FPR,TPR) ไปยังจุด (1,1) ได้ดังภาพที่ 4.2 คำนวณหาค่า AUC จากรูปที่ 4.2 เพื่อวิเคราะห์ประสิทธิภาพของระบบคัดกรองอีเมลขยะดังแสดงในตารางที่ 4.9

ตารางที่ 4.9 ค่า AUC ของแต่ละระบบคัดกรอง

ระบบคัดกรอง	AUC
ระบบเก่า	0.79
ระบบปัจจุบัน	0.75
ระบบอ้างอิง	0.74

จากผลการทดลองที่กล่าวมาสามารถสรุปและวิเคราะห์ผลการทดลองได้ดังนี้

4.4.1 ประสิทธิภาพโดยรวมของระบบคัดกรอง

จากค่า AUC ตามตารางที่ 4.9 ระบบคัดกรองอีเมลขยะที่มีประสิทธิภาพโดยรวมสูงสุดในการคัดกรองอีเมลขยะคือระบบคัดกรองเก่า ซึ่งมีค่า AUC สูงสุด โดยมีประสิทธิภาพในการคัดกรองอีเมลขยะได้ดีที่สุด

4.4.2 การวิเคราะห์จุดแข็งและจุดด้อยของระบบคัดกรอง

จากผลการการทดลองจากตารางที่ 4.8 สามารถวิเคราะห์ถึงจุดเด่นและจุดด้อยของแต่ละระบบซึ่งผลการทดลองจะชี้ให้เห็นข้อดีและข้อเสียของแต่ละระบบคัดกรอง ในงานวิจัยนี้จะวิเคราะห์ 2 ด้าน ซึ่งกำหนดไว้ในข้อ 3.9 ดังต่อไปนี้

4.4.2.1 ระบบคัดกรองที่มีอัตราอีเมลดีสูญหายน้อยที่สุด

จากตารางที่ 4.8 ค่า FPR ของระบบปัจจุบันมีค่าต่ำที่สุด แสดงว่ามีอัตราอีเมลดีถูกคัดกรองเป็นอีเมลขยะน้อยที่สุดหรือมีอัตราที่อีเมลสำคัญจะหายน้อยที่สุด แต่จะมีปริมาณอีเมลขยะเพิ่มขึ้นมาในตู้จดหมายมากขึ้น หากองค์กรเป้าหมายต้องการระบบคัดกรองที่ปลอดภัยคือไม่ทำอีเมลสำคัญหายระบบคัดกรองปัจจุบันเป็นตัวเลือกที่เหมาะสมที่สุด

4.4.2.2 ระบบคัดกรองที่มีอัตราการถูกรบกวนจากอีเมลขยะน้อยที่สุด

จากตารางที่ 4.8 ค่า TPR ของระบบคัดกรองเก่ามีค่าสูงที่สุด แสดงว่ามีอัตราการคัดกรองอีเมลขยะได้ดีที่สุด ทำให้อีมเมลขยะเข้ามายังตู้จดหมายได้น้อย อย่างไรก็ดีอาจ

มีอีเมลสำคัญหายไปบางส่วน หากองค์กรเป้าหมายต้องการระบบคัดกรองที่สามารถคัดกรองอีเมลขยะได้ดีระบบคัดกรองเก่าจะเป็นตัวเลือกที่เหมาะสมที่สุด

ผลการประเมินประสิทธิภาพสามารถสรุปได้ดังตารางที่ 4.10

ตารางที่ 4.10 สรุปผลการประเมินประสิทธิภาพ

หัวข้อ	**ระบบคัดกรอง**
ระบบที่มีประสิทธิภาพสูงสุด	ระบบเก่า
ระบบที่อีเมลดีสูญหายน้อยที่สุด	ระบบปัจจุบัน
ระบบที่อีเมลขยะรบกวนน้อยที่สุด	ระบบเก่า

การเลือกระบบคัดกรองให้เหมาะสมกับความต้องการขององค์กร เมื่ออ้างอิงจากผลการทดลองข้างต้น ทำให้ประสิทธิภาพการทำงานของระบบคัดกรองดียิ่งขึ้นและถูกต้องตรงกับความต้องการของผู้ใช้งานในองค์กรนั้นๆ เนื่องจากกระบวนการทั้งหมดใช้อีเมลจริงและอาสาสมัครซึ่งเป็นผู้ใช้งานจริงในการประเมินเพื่อให้ใกล้เคียงกับสภาพความเป็นจริงมากที่สุด อย่างไรก็ดี ประเด็นสำคัญที่งานวิจัยนี้เน้นเป็นอย่างยิ่งคือปัญหาความเป็นส่วนตัวของผู้ใช้ กระบวนการต่างๆ อาศัยจำกัดสิทธิของอาสาสมัครแต่ละคนให้สามารถเห็นและประเมินได้เฉพาะอีเมลของตนเท่านั้น

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้นำเสนอกระบวนวิธีสำหรับประเมินประสิทธิภาพของระบบคัดกรองอีเมลขยะโดยเจ้าของอีเมล ซึ่งประเมินในสภาพใกล้เคียงกับความเป็นจริง ใช้อีเมลจริงเป็นคลังอีเมลและให้ผู้ใช้งานจริงจากองค์กรเป้าหมายเป็นอาสาสมัครเพื่อเข้าร่วมในกระบวนการประเมิน โดยกระบวนวิธีนี้จะใช้เป็นเครื่องมือสำหรับใช้เปรียบเทียบประสิทธิภาพของระบบคัดกรองอีเมลขยะ เพื่อใช้ผลการประเมินประกอบการตัดสินใจเลือกใช้ จัดซื้อ หรือวางแผนในการพัฒนาประสิทธิภาพของระบบคัดกรอง กระบวนการประเมินจะเปรียบเทียบการคัดกรองอีเมลขยะระหว่างระบบคัดกรองทั่วไปกับการคัดกรองโดยอาสาสมัครซึ่งเป็นระบบคัดกรองในอุดมคติ สามารถคัดกรองได้อย่างไม่มีข้อผิดพลาด อาสาสมัครเข้าร่วมการประเมินผ่านเว็บไซต์ที่ง่ายต่อการใช้งานและไม่เป็นภาระแก่อาสาสมัครจนเกินไป กระบวนการประเมินจะไม่ก่อให้เกิดปัญหาความเป็นส่วนต่ออาสาสมัคร อาสาสมัครสามารถมองเห็นและประเมินได้เฉพาะอีเมลของตนเท่านั้น เนื่องจากกระบวนการประเมินของอาสาสมัครไม่เป็นภาระแก่อาสาสมัครจนเกินไปทำให้สามารถประเมินได้บ่อยครั้งตามต้องการ ซึ่งเป็นประโยชน์ต่อการปรับปรุงและพัฒนาประสิทธิภาพของระบบคัดกรองอีเมลขยะ อาสาสมัครที่เคยเข้าร่วมกระบวนการสามารถกำหนดให้เป็นตัวแทนขององค์กรเป้าหมายสำหรับการประเมินในครั้งต่อๆ ไปได้

ในส่วนของการทดลอง งานวิจัยนี้นำกระบวนวิธีที่ออกแบบมาสร้างเป็นระบบจริงเพื่อเปรียบเทียบระบบคัดกรองเก่า ระบบปัจจุบัน และระบบอ้างอิง กับการประเมินโดยอาสาสมัครโดยมีสมมติฐานว่าอาสาสมัครเป็นระบบคัดกรองในอุดมคติ ระบบทำการป้อนคลังอีเมลผ่านระบบคัดกรองทั้งสามระบบ สุ่มตัวอย่างอีเมลของอาสาสมัครแต่ละคนจากคลังอีเมลให้เจ้าของอีเมลประเมินด้วยตา ใช้วิธีการกำหนดตัวของ ทาโร ยามาเน่ [9] กำหนดขนาดตัวอย่างที่เหมาะสมกับขนาดของคลังอีเมล อาสาสมัครสามารถเลือกประเมินอีเมลของตนได้มากกว่าหรือเท่ากับขนาดตัวอย่างที่กำหนดให้ จากนั้นจึงจะใช้วิธีทางสถิติวิเคราะห์ประสิทธิภาพโดยรวมของระบบคัดกรอง วิเคราะห์ข้อดีและข้อด้อยของแต่ละระบบ และสรุปผลการประเมินประสิทธิภาพ

ผลการเปรียบเทียบประสิทธิภาพของระบบเก่า และระบบปัจจุบัน กับระบบอ้างอิงพบว่าประสิทธิภาพของระบบคัดกรองอ้างอิงใกล้เคียงกับระบบทั้งสอง ซึ่งระบบคัดกรองอ้างอิงก็เป็นอีก

ทางเลือกสำหรับองค์กร หากพัฒนาระบบคัดกรองอ้างอิงอย่างต่อเนื่องอาจทำให้ประสิทธิภาพของการคัดกรองใกล้เคียงกับระบบคัดกรองที่ขายในท้องตลาด

จากผลการวิจัยและทดสอบกระบวนวิธีสามารถสรุปได้ดังนี้

5.1.1 กระบวนวิธีที่ออกแบบและผลการทดลองในงานวิจัยนี้สามารถประเมินประสิทธิภาพของระบบคัดกรองอีเมลขยะได้ โดยสามารถวิเคราะห์ค่า TP, TN, FP และ FN สำหรับคำนวณค่า AUC TPR และ FPR เพื่อใช้ในการวิเคราะห์ เปรียบเทียบประสิทธิภาพของระบบคัดกรอง และสรุปผลการประเมินได้

5.1.2 คลังอีเมลที่ใช้ในการวิจัยมีปริมาณอีเมลเพียงพอและประกอบด้วยอีเมลของผู้ใช้งานจำนวนมากพอที่จะใช้ในงานวิจัย ทำให้สามารถวิเคราะห์และประเมินประสิทธิภาพของระบบคัดกรองอีเมลขยะได้

5.1.3 จากกระบวนวิธีที่ออกแบบ กลุ่มอาสาสมัครจะประเมินผ่านเครื่องมือที่ใช้งานง่าย อีกทั้งการสุ่มอีเมลช่วยลดภาระของอาสาสมัคร และมีการรักษาความเป็นส่วนตัวอย่างดี เป็นผลให้อาสาสมัครยินดีเข้าร่วมกระบวนการประเมิน และปริมาณอีเมลที่ถูกประเมินสูงเกินขนาดตัวอย่างที่กำหนด ทำให้สามารถสรุปผลการทดลองได้เร็ว

5.1.4 จากผลการทดลองระบบคัดกรองที่มีประสิทธิภาพสูงสุดคือ ระบบที่มีค่า AUC สูงที่สุด ระบบที่ปริมาณอีเมลดีสูญหายน้อยที่สุดคือ ระบบที่มีค่า FPR ต่ำที่สุด และ ระบบที่มีปริมาณอีเมลขยะรบกวนน้อยที่สุดคือที่มีค่า TPR สูงที่สุด

5.1.5 กระบวนวิธีที่ออกแบบสามารถประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะได้แม้มีเพียงระบบเดียว โดยเป็นการเปรียบเทียบกับระบบในอุดมคติ

5.2 ปัญหาและข้อเสนอแนะ

5.2.1 ในขณะเริ่มต้นวิจัยองค์กรเป้าหมายมีการเปลี่ยนแปลงระบบอีเมลทำให้การดำเนินการทดลองเป็นไปอย่างยากลำบาก

5.2.2 คลังอีเมลที่ใช้ในการทดลองไม่ทันสมัยมากนักเนื่องจากถูกรวบรวมก่อนการประเมินโดยอาสาสมัครเป็นเวลาประมาณ 4 เดือน อาจทำให้เนื้อหาในอีเมลนั้นล้าสมัยไป เป็นผลให้การประเมินผิดพลาด ควรใช้คลังอีเมลที่รวบรวมไว้ไม่นานเกินไปในกระบวนการทดลอง

5.2.3 เนื่องจากองค์กรเป้าหมายมีระบบรักษาความปลอดภัยสูง การป้อนอีเมลผ่านระบบคัดกรองปัจจุบันทำได้ยากลำบาก โดยต้องแบ่งป้อนคลังอีเมลครั้งละ 10,000 ฉบับ เพื่อป้องกันระบบคัดกรองตัดสินว่ากระบวนการประเมินกำลังส่งอีเมลขยะเข้ามายังองค์กรเป้าหมาย ต้องพัฒนาต่อไปให้ระบบประเมินประสิทธิภาพเป็นส่วนหนึ่งของอีเมลเพื่อจะสามารถใช้งานได้สะดวกขึ้น

5.2.4 การสุ่มตัวอย่างอีเมลในการทดลองนี้ใช้การสุ่มอีเมลแบบง่าย ควรศึกษาวิธีการสุ่มตัวอย่างเพิ่มเติม เพื่อให้สามารถกำหนดกลุ่มตัวอย่างได้ดีขึ้น

5.2.5 จากตารางที่ 4.2 อาสาสมัครให้ความร่วมมือเป็นอย่างดี หากพัฒนากระบวนการประเมินและการติดต่อประสานงานกับอาสาสมัครให้ดียิ่งขึ้น อาจทำให้ผู้ใช้งานทุกคนยินดีเข้าร่วมกระบวนการประเมิน ทำให้ผลการประเมินมีความถูกต้องมากยิ่งขึ้น

5.2.6 ไม่สามารถปิดระบบ SenderBase ที่องค์กรเป้าหมายใช้งานได้เนื่องจากมีผลกระทบต่อองค์กรพอสมควร ทำให้ปริมาณอีเมลขยะที่รวบรวมเป็นคลังอีเมลลดลง ควรปิดระบบ SenderBase เพื่อรวบรวมคลังอีเมลเพื่อให้คลังอีเมลใกล้เคียงกับสภาพการณ์จริงมากยิ่งขึ้น

5.3 งานวิจัยในอนาคต

5.3.1 การเปรียบเทียบระบบคัดกรองอีเมลขยะราคาแพงกับระบบคัดกรองจากโปรแกรมแจกฟรี

5.3.2 โปรแกรมโอเพนซอร์สสำหรับประเมินประสิทธิภาพระบบคัดกรองอีเมลขยะโดยเจ้าของอีเมล

5.3.3 พัฒนาโมดูลสำหรับประเมินประสิทธิภาพให้เป็นส่วนหนึ่งในระบบอีเมลขององค์กรเป้าหมาย

5.3.4 การวิเคราะห์ความเบี่ยงเบนของการคัดกรองจากผลการคัดกรองของอาสาสมัคร (ระบบในอุดมคติ) ซึ่งนำไปใช้เป็นแนวทางในการปรับปรุงระบบคัดกรอง ทั้งในการใช้งานและอ้างอิงในเชิงวิจัย รวมถึงประโยชน์ในการพัฒนาโปรแกรมแจกฟรี หรือโอเพนซอร์ส

5.3.5 การเพิ่มจำนวนอาสาสมัครเพื่อนการประเมินประสิทธิภาพของระบบคัดกรองอีเมลขยะโดยเจ้าของอีเมล

รายการอ้างอิง

[1] The SenderBase Network. Global SPAM Volume. Available from: http://www.senderbase.org/home/detail_spam_volume. [2010, January 25]

[2] Hoanca, B. How good are our weapons in the spam wars? Technology and Society Magazine, IEEE 25(1): 22-30, 2006.

[3] Fumera, G., Pillai, I., and Roli, F. Spam filtering based on the analysis of text information embedded into images Journal of Machine Learning Research, 2699-2720, 2006.

[4] Blanzieri, E., and Bryl, A. A survey of anti-spam techniques. Technical report. 2006, 2008.

[5] Cormack, G. V., and Lynam, T. R. Online supervised spam filter evaluation. ACM Trans. 2007.

[6] Lai, C., and Tsai, M. An empirical performance comparison of machine learning methods for spam e-mail categorization. Hybrid Intelligent Systems. HIS '04. Fourth International Conference on, 5-8 Dec. 2004, 44-48, 2004.

[7] Tran, M., and Armitage, G. End-users' resource consumption of spam and a 3D anti-spam evaluation framework. TENCON 2005 IEEE Region 10, 2005.

[8] Yeh, C., Wu, C., and Doong, S. Effective spam classification based on meta-heuristics Systems, Man and Cybernetics. 2005 IEEE International Conference, Vol. 4. 3872-3877, 2005.

[9] Glenn, D. I. Determining Sample Size. Agricultural education and communication department, Florida cooperative extension service, Institute of food and agricultural sciences, University of Florida, 2003.

[10] Schryen, G. Anti – Spam Measures analysis and design. New York: Springer. 2005.

[11] Spamhaus, Spamhaus Statistics: The Top 10, Available from: http://www.spamhaus.org/statistics/countries.lasso, 2006.

[12] IronPort Systems. Spam Cop. Available from: http://www.spamcop.net/ [2008, June 21]

[13] Diesner, J., and Carley, K. M., Exploration of Communication Networks from the Enron Email Corpus. Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining. Newport Beach, CA, 2005, 3-14.

[14] Cauce [Online]. Available from: http://www.cauce.org/ [2008, July 30].

[15] The SenderBase Network. Cisco IronPort SenderBase Security Network. Available from: http://www.senderbase.org/. [2010, January 25]

[16] F. Tom, “An intorduction to ROC analysis,” Pattern recognition letters 27, 861-874, 2006

[17] Garcia, F. D., Hoepman, J. H., and Van Nieuwenhuizen, J. Spam Filter Analysis. Security and Protection in Information Processing Systems. IFIP TC11 19th International Information Security Conference (SEC2004), 395-410, Toulouse, France: Kluwer Academic Publishers, 2004.

ภาคผนวก

ค่ามาตรฐานของโปรแกรมสแปมแอสแซสซิน

```
# This is the right place to customize your installation of SpamAssassin.
#
# See 'perldoc Mail::SpamAssassin::Conf' for details of what can be
# tweaked.
#
# Only a small subset of options are listed below
#
###########################################################################

#   Add *****SPAM***** to the Subject header of spam e-mails
#
rewrite_header Subject *****SPAM*****

#   Save spam messages as a message/rfc822 MIME attachment instead of
#   modifying the original message (0: off, 2: use text/plain instead)
#
report_safe 0

#   Set which networks or hosts are considered 'trusted' by your mail
#   server (i.e. not spammers)
#
# trusted_networks 212.17.35.

#   Set file-locking method (flock is not safe over NFS, but is faster)
#
# lock_method flock

#   Set the threshold at which a message is considered spam (default: 5.0)
#
# required_score 5.0

#   Use Bayesian classifier (default: 1)
#
use_bayes 1

#   Bayesian classifier auto-learning (default: 1)
#
bayes_auto_learn 1

#   Set headers which may provide inappropriate cues to the Bayesian
#   classifier
#
bayes_ignore_header X-Bogosity
bayes_ignore_header X-Spam-Flag
bayes_ignore_header X-Spam-Status
```

รูปที่ 1 แสดงค่ามาตรฐานของโปรแกรมสแปมแอสแซสซิน

ประวัติผู้เขียนวิทยานิพนธ์

นายอรรถกร องค์ศิริพร เกิดเมื่อวันที่ 20 ธันวาคม พ.ศ. 2525 ที่จังหวัดกรุงเทพมหานคร สำเร็จการศึกษาหลักสูตรวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยบูรพา เมื่อปีการศึกษา 2547 และเข้าศึกษาต่อหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ปัจจุบันทำงานอยู่ที่ บริษัท ดีเอสที เวิลด์วายด์ เซอร์วิส ประเทศไทย จำกัด