บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

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

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

งานวิจัยนี้ ได้ศึกษาทำการศึกษาโดยมีจุดมุ่งหมายและวัตถุประสงค์สำคัญดังนี้

1.2.1. ศึกษาตัวกรองอีเมล์ขยะโดยเทคนิคที่นิยมใช้กันทั่วไปได้แก่ การประยุกต์ใช้ทฤษฎีเบย์เซียน

1.2.2. ศึกษาและพัฒนาตัวกรองอีเมล์ขยะ (Spam Mail Filter) โดยประยุกต์ใช้ทฤษฎีเจเนติกอัลกอริทึม

1.2.3. ทดสอบประสิทธิภาพของการกรองอีเมล์ขยะโดยการประยุกต์ใช้เจเนติกอัลกอริทึม รวมทั้งวิเคราะห์ผลการการกรองอีเมล์ขยะที่ได้จากการใช้พารามิเตอร์ต่างๆ ในกระบวนการทางเจเนติก

1.2.4. ศึกษาผลลัพธ์ของการกรองอีเมล์ขยะโดยประยุกต์ใช้เจเนติกอัลกอริทึมเปรียบเทียบกับการประยุกต์ใช้ทฤษฎีเบย์เซียน

1.3 สมมติฐานของการศึกษา

การพัฒนาตัวกรองอีเมล์ขยะในงานนี้เป็นการนำเจเนติกอัลกอริทึมที่มีตัวดำเนินการทางเจเนติกมาสร้างรูปแบบของอีเมล์ให้มีความหลากหลายเพิ่มขึ้นมา โดยแต่ละรูปแบบของอีเมล์ที่ถูกสร้างขึ้นมาใหม่นั้นจะถูกนำไปเป็นกฎที่ใช้สำหรับกรองอีเมล์ต่อไป ซึ่งจะทำให้การกรองมีประสิทธิภาพมากขึ้น โดยที่ไม่ต้องแก้ไขปรับเปลี่ยนพารามิเตอร์ต่างๆของระบบทั้งหมด หรือต้องคอยเสียเวลาปรับปรุงชุดข้อมูลฝึกหัดอย่างสม่ำเสมอ ดังเช่นงานวิจัยที่ผ่านมา

1.4 ทฤษฎีหรือแนวความคิดที่ใช้ในการวิจัย

สำหรับการกรองอีเมล์ขยะโดยใช้ทฤษฎีเจเนติกอัลกอริทึมนี้ จะต้องอาศัยหลักการและทฤษฎีดังต่อไปนี้

1.4.1 ทฤษฎีเจเนติกอัลกอริทึม (Genetic Algorithm)

1.4.2 ระบบแขนงการตัดสินใจ (Decision Tree)

1.4.3 ทฤษฎีเบย์เซียน (Bayesian Theorem)

1.5 ขอบเขตการวิจัย

งานวิจัยนี้มีจุดประสงค์เพื่อที่จะสร้างตัวกรองอีเมล์ขยะซึ่งใช้อีเมล์ขยะและอีเมล์ดีในการให้ระบบเรียนรู้ ซึ่งมีขอบเขตการวิจัย ดังนี้

1.5.1 เป็นตัวกรองอีเมล์ขยะ ซึ่งใช้ทั้งอีเมล์ขยะและอีเมล์เป็นชุดเรียนรู้

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

1.6 ขั้นตอนของการศึกษา

สำหรับขั้นตอนของการทำการศึกษาวิจัย สามารถแบ่งออกเป็นลำดับได้ดังนี้

1.6.1 ศึกษาค้นคว้าผลงานวิจัยและเอกสารทางวิชาการในหัวข้อที่เกี่ยวข้อง ที่มีผู้ทำวิจัยมาแล้ว

1.6.2 กำหนดหัวข้อ เป้าหมาย วัตถุประสงค์ และขอบเขตของการวิจัย

1.6.3 ศึกษาทฤษฎีและหลักการที่เกี่ยวข้องกับการวิจัย

1.6.4 วิเคราะห์และออกแบบตัวกรองอีเมล์ขยะโดยใช้เจเนติกอัลกอริทึม

1.6.5 พัฒนาโปรแกรม ทำการทดลองพร้อมทั้งบันทึกผลที่ได้จากการทดลองในแต่ละขั้นตอน

1.6.6 ปรับค่าพารามิเตอร์ให้เหมาะสมกับการทดลอง

1.6.7 วิเคราะห์ เปรียบเทียบผลการทดลองที่ได้ และสรุปผลการทดลอง

1.6.8 จัดทำเอกสารประกอบวิทยานิพนธ์

1.7 ข้อจำกัดของการศึกษา

1.7.1 อีเมล์ที่นำมาใช้ทดลองจะพิจารณาเฉพาะส่วนของเนื้อหา (Body) เท่านั้น

1.7.2 อีเมล์ที่นำมาพิจารณาต้องมีคำต่างๆ ปรากฏอยู่ในส่วนของเนื้อหา อีเมล์ที่มีรูปภาพหรือไฟล์แนบ (Attach File) จะไม่ถูกนำมาพิจารณา

1.7.3 คำศัพท์ที่ใช้บ่งบอกความเป็นอีเมล์ขยะทั้งหมดจะถูกจัดเก็บไว้ในฐานข้อมูลคำ โดยจะถูกจัดเป็น 8 กลุ่ม สำหรับคำอื่นๆ ที่นอกเหนือจากนี้จะไม่นำมาใช้ในการพิจารณา

1.7.4 การดำเนินการต่างๆ ของกระบวนการทางเจเนติกอัลกอริทึมทำงานภายในกลุ่มหรือระหว่างกลุ่มที่สอดคล้องกัน

1.8 คำจำกัดความที่ใช้ในการศึกษา

1.8.1 อีเมล์ขยะจะถูกแปลงให้อยู่ในรูปแบบของโครโมโซมซึ่งจะประกอบไปด้วยยีนทั้งหมด 8 ยีน โดยยีนแต่ละยีนจะแสดงค่าความน่าจะเป็นที่จะเป็นขยะตามที่พบในกลุ่มของยีนนั้นๆ โดยเฉลี่ย

1.8.2 โครโมโซมแม่แบบ (Template) หมายถึงชุดของโครโมโซมทั้งหมดที่ได้จากกระบวนการเจเนติกอัลกอริทึม ซึ่งจะถูกนำไปใช้ในการพิจารณาว่าอีเมล์ทดสอบว่าเป็นอีเมล์ประเภทใด

1.9 เครื่องมือและอุปกรณ์ที่ใช้ในงานวิจัย

เครื่องมือและอุปกรณ์ที่ใช้ในการวิจัยในครั้งนี้ ได้แก่

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

1.9.2 ระบบปฏิบัติการ Windows XP Professional

1.9.3 โปรแกรม Microsoft Visual Basic.Net เวอร์ชัน 6.0

1.9.4 โปรแกรม Microsoft Excel

1.9.5 โปรแกรมที่ใช้ในการตัดคำและนับความถี่ของคำ GNU Awk เวอร์ชัน3.1.3 [6]

1.9.6 โปรแกรมที่ใช้ในการหารากศัพท์ของคำ Word Stemming [7]

1.10 โครงสร้างของวิทยานิพนธ์

วิทยานิพนธ์ฉบับนี้แบ่งออกเป็น 6 บท แต่ละบทประกอบด้วยเนื้อหาดังต่อไปนี้

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

บทที่ 2 กล่าวถึงผลงานวิจัยที่เกี่ยวข้องกับกรองอีเมล์ขยะ ได้แก่ เทคนิคต่างๆ ในการกรองอีเมล์ขยะ และการกรองอีเมล์ขยะโดยใช้เครื่องมือเรียนรู้และวิธีการทางสถิติ

บทที่ 3 กล่าวถึงองค์ความรู้ที่เป็นพื้นฐานในงานวิจัยนี้ ซึ่งได้แก่ทฤษฎีต่างๆที่เกี่ยวข้องกับอีเมล์ องค์ประกอบของอีเมล์ ความรู้พื้นฐานเกี่ยวกับเจเนติกอัลกอริทึม ระบบแขนงการตัดสินใจและความรู้พื้นฐานเกี่ยวกับทฤษฎีเบย์เซียน

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

บทที่ 5 กล่าวถึง การทดลองและผลการทดลองการกรองอีเมล์ขยะโดยใช้เจเนติกอัลกอริทึม โดยใช้พารามิเตอร์ที่แตกต่างกัน การปรับหาพารามิเตอร์ที่เหมาะสม และ การกรองอีเมล์ขยะโดยใช้ทฤษฎีเบย์เซียนเพื่อการเปรียบเทียบผลการทดลองและประสิทธิภาพกับวิธีการที่นำเสนอ จากนั้นวิเคราะห์ผลที่ได้จากการศึกษาวิจัยการกรองอีเมล์ขยะโดยใช้เจเนติกอัลกอริทึม

บทที่ 6 กล่าวถึง บทสรุปและบทวิจารณ์ รวมถึงข้อเสนอแนะ และแนวทางในการพัฒนาตัวกรองอีเมล์ขยะต่อไปในอนาคต