

บทที่ 6

สรุปผลการทดลองและข้อเสนอแนะ

วิทยานิพนธ์นี้ได้นำเสนอการสร้างตัวกรองอีเมลขยะโดยใช้เจเนติกอัลกอริทึม ในการสร้างแม่แบบที่หลากหลายของอีเมล เพื่อนำไปสร้างกฎสำหรับกรองอีเมลขยะ เนื่องจากตัวดำเนินการในเจเนติกอัลกอริทึมประกอบไปด้วย การคัดเลือก การครอสโอเวอร์ และการมิวเตชัน ซึ่งแต่ละตัวดำเนินการก็สามารถทำได้หลายวิธี เช่น การคัดเลือก มีแบบ การคัดเลือกแบบจัดลำดับ การคัดเลือกแบบวงล้อถ่วงน้ำหนัก หรือในครอสโอเวอร์ มีเปอร์เซ็นต์ของการครอสเป็นต้นจึงต้องมีการปรับค่าพารามิเตอร์ต่างๆ เพื่อหาค่าพารามิเตอร์ที่เหมาะสมให้กับระบบ และพบว่าเมื่อนำพารามิเตอร์เหล่านี้ไปทดสอบได้ผลลัพธ์ที่ดี จากนั้นได้ประเมินประสิทธิภาพของอัลกอริทึมที่นำเสนอเปรียบเทียบกับตัวกรองอีเมลขยะตามทฤษฎีเบย์เซียน พบว่า ค่า Accuracy, Recall และ Precision ของวิธีการที่นำเสนอ ดีกว่าตัวกรองที่สร้างขึ้นด้วยทฤษฎีเบย์เซียน

6.1 สรุปและวิเคราะห์ผลการดำเนินงานวิจัย

จากการทดลองปรับพารามิเตอร์ในบทที่ 5 เพื่อหาค่าพารามิเตอร์ที่เหมาะสม ได้ผลลัพธ์ดังนี้

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

จะปรับค่าของการมีวเตชันทั้งสองพารามิเตอร์อย่างไร ค่า Accuracy, Recall และ Precision ก็มีค่าใกล้เคียงกัน ทำให้ไม่สามารถระบุแนวโน้มที่แน่นอนได้ เนื่องจากทำการมีวเตชันไม่ได้ทำให้ผลการทดลองดีขึ้นเสมอไปและในบางกรณีอาจทำให้ผลการทดลองแย่ลง ดังนั้นในงานวิทยานิพนธ์นี้จึงจะไม่ใช้มีวเตชันมาพิจารณาด้วย

- เลือกจำนวนของยีนเรอร์โว์ที่ 6 ยีน เนื่องจากเป็นค่าที่ทำให้ค่าเฉลี่ยของ Accuracy, Recall และ Precision ดีที่สุด เนื่องจากผลการทดลองในบทที่ 5 แสดงให้เห็นว่า การปรับค่ายีนเรอร์โว์น้อยเกินไป เช่น 3 ถึง 5 ยีน จะทำให้ค่าเฉลี่ยของ Accuracy, Recall และ Precision ต่ำมาก เนื่องจากมีความเป็นไปได้สูงที่ยีนของโครโมโซมที่เป็นอีเมลต์ทดสอบจะตรงกับยีนในโครโมโซมที่เป็นกฎเพียง 3 ถึง 5 โครโมโซม ซึ่งในความเป็นจริงจะเห็นว่ายังไม่มากเพียงพอที่จะระบุประเภทของอีเมลต์ได้ ทำให้ผลลัพธ์ที่ได้ต่ำ ส่วนการปรับค่ายีนเรอร์โว์ที่ 7 ยีนหรือ 8 ยีน ก็จะทำให้ผลลัพธ์ต่ำลง เนื่องจากการให้โครโมโซมที่เป็นอีเมลต์ทดสอบจะตรงกับยีนในโครโมโซมที่เป็นกฎเกือบทั้งหมดทุกยีนจึงระบุประเภทของอีเมลต์ได้นั้นจะเกิดขึ้นยากมาก

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

สำหรับการประเมินประสิทธิภาพของอัลกอริทึมที่นำเสนอ ได้ทำการทดลองเปรียบเทียบประสิทธิภาพกับตัวกรองอีเมลล์ยะตามทฤษฎีเบย์เซียน เมื่อเปรียบเทียบประสิทธิภาพทางด้าน Accuracy, Recall และ Precision พบว่าตัวกรองที่นำเสนอมีประสิทธิภาพดีกว่า

6.2 ปัญหาที่พบในวิทยานิพนธ์

6.2.1 ในปัจจุบันมีอีเมลล์ยะจำนวนมากซึ่งพยายามหลีกเลี่ยงการกรอง โดยจะพยายามเลียนแบบรูปแบบของเมลล์ดี เช่น ใช้การใช้คำดีเป็นจำนวนมากเข้ามาปนในเนื้อหาของอีเมลล์ยะ(Good Word Chaff) [2] ซึ่งมีผลลัพธ์ต่ำลง

6.2.2 การเลือกชุดประชากรที่จะมาใช้ในการจำแนกอีเมลล์นั้น จะเลือกโดยการปล่อยให้ตัวกรองทำตามกระบวนการเจเนติกไปจนครบ Generation ที่ตั้งไว้ จึงค่อยดูผลว่า Generation ไหน ให้ผลของการกรองที่ดีที่สุด แล้วจึงเลือกประชากรชุดนั้นไว้เพื่อมาใช้เพื่อเป็นแม่แบบในการจำแนกอีเมลล์ต่อไป ทำให้เสียเวลาค่อนข้างมาก

6.2.3 เนื่องจากงานที่นำเสนอใช้เจเนติกอัลกอริทึมซึ่งต้องใช้เวลาในการประมวลผลค่อนข้างนาน และมีพารามิเตอร์ค่อนข้างเยอะ ทำให้เสียเวลาในการปรับพารามิเตอร์

6.3 ข้อเสนอแนะและแนวทางในการพัฒนาต่อ

6.3.1 การแก้ปัญหาการใช้คำดีเป็นจำนวนมากเข้ามาปนในเนื้อหาของอีเมลล์ขณะนั้น ทำได้ค่อนข้างยากสำหรับการกรองอีเมลล์ขยะที่ใช้ส่วนเนื้อหาในการวิเคราะห์ ดังนั้นเพื่อจะแก้ปัญหาในกรณีนี้อาจจำเป็นต้องใช้เทคนิคอื่นมาช่วยในการพิจารณาไปด้วย เช่น ให้ความสำคัญในเรื่องของเวลา เช่น อีเมลล์ที่ได้รับในช่วงกลางคืนมีแนวโน้มที่จะเป็นอีเมลล์ขยะมากกว่าอีเมลล์ที่ได้รับในช่วงกลางวัน

6.3.2 ควรปล่อยให้ระบบหยุดกระบวนการเอง โดยจะหยุดเมื่อค่าความเหมาะสม(Fitness) ของระบบเริ่มลดลงหรือคงที่

6.3.3 ตัวกรองอีเมลล์ขยะในวิทยานิพนธ์นี้ได้เรียนรู้จากอีเมลล์ขยะและอีเมลล์ดีในปริมาณหนึ่ง และได้ปรับพารามิเตอร์เพื่อให้เหมาะสมกับชุดการทดลองนี้ ดังนั้นหากมีการปรับเปลี่ยนชุดข้อมูลสำหรับเรียนรู้ จึงอาจต้องปรับค่าพารามิเตอร์ใหม่ให้เหมาะสมกับชุดการทดลอง