

บทที่ 1

บทนำ

1.1 หลักการและเหตุผล

ปัจจุบันซึ่งเป็นยุคของเทคโนโลยีข้อมูลข่าวสารที่มีจำนวนข้อมูลข่าวสารเพิ่มมากขึ้นอย่างรวดเร็ว ทำให้เกิดปริมาณของข้อมูลข่าวสารที่มากเกินไปเกินความต้องการ โดยเฉพาะข้อมูลข่าวสารที่มีการเปลี่ยนแปลงเนื้อหาบ่อย ๆ เช่น Streaming, News Feed เป็นต้น ส่งผลให้การบริโภคข้อมูลข่าวสารให้ครบทุกรายละเอียดนั้นต้องใช้เวลาในการสรุปใจความสำคัญจากเอกสารต้นฉบับเป็นการสรุปประเด็นหรือเนื้อหาที่สำคัญของเอกสารต้นฉบับเพื่อให้ผู้อ่านได้รับทราบประเด็นหรือเนื้อหาในส่วนที่เป็นใจความสำคัญโดยไม่จำเป็นต้องอ่านเนื้อหาทั้งหมดของเอกสารฉบับนั้น ส่งผลให้ผู้อ่านใช้เวลาในการอ่านเอกสารแต่ละฉบับน้อยลงและทำให้รับทราบข้อมูลข่าวสารต่าง ๆ ได้อย่างครบถ้วนและรวดเร็ว

งานวิจัยที่เกี่ยวข้องกับการสรุปใจความสำคัญสำหรับเอกสารที่เป็นภาษาต่างประเทศจะมีจำนวนค่อนข้างมาก ทั้งนี้เพราะภาษาต่างประเทศจะมีโครงสร้างของประโยคที่ไม่ซับซ้อนมากนัก อีกทั้งยังมีการระบุขอบเขตของประโยคเอาไว้อย่างชัดเจน ส่งผลให้การสรุปใจความสำคัญทำได้ง่าย และข้อสรุปที่ได้ก็มีประสิทธิภาพค่อนข้างดี ซึ่งแตกต่างจากภาษาในทวีปเอเชีย เช่น ภาษาไทย ภาษาจีน เป็นต้น ที่มีโครงสร้างของประโยคที่ค่อนข้างซับซ้อนและไม่มีเครื่องหมายหรือสัญลักษณ์ที่ใช้แสดงถึงจุดสิ้นสุดของประโยคที่ชัดเจนเหมือนกับภาษาต่างประเทศ ทำให้การหาขอบเขตของประโยคทำได้ยาก ส่งผลให้งานวิจัยที่เกี่ยวข้องกับการสรุปใจความสำคัญสำหรับเอกสารที่เป็นภาษาในทวีปเอเชียมีจำนวนน้อยและข้อสรุปที่ได้ก็ไม่ค่อยมีประสิทธิภาพ

เทคนิคที่ใช้สำหรับการสรุปใจความสำคัญมีหลายเทคนิคด้วยกัน เช่น การพิจารณาจากความถี่และตำแหน่งของประโยคในเอกสาร การพิจารณาจากความสัมพันธ์ระหว่างประโยคในเอกสาร เทคนิค Singular Value Decomposition, Fuzzy Clustering Algorithm เป็นต้น สำหรับงานวิจัยนี้ผู้วิจัยเลือกใช้เทคนิคการสรุปแฟร็กทัลเป็นเทคนิคในการสรุปใจความสำคัญจากเอกสาร เนื่องจากมีผู้วิจัยกลุ่มหนึ่ง คือ Yang, C. C., and Wang, F. L. ได้นำเทคนิคนี้มาใช้ในการสรุปใจความสำคัญจากเอกสารภาษาจีนซึ่งเป็นภาษาที่มีโครงสร้างของประโยคคล้ายคลึงกับภาษาไทย โดยพบว่าข้อสรุปที่ได้มีความเที่ยง (Precision) ประมาณ 85%[13] ซึ่งถือได้ว่าข้อสรุปที่ได้มีประสิทธิภาพค่อนข้างดี ดังนั้นผู้วิจัยจึงมีแนวคิดที่จะนำเทคนิคนี้มาใช้ในการสรุปใจความสำคัญจากเอกสารภาษาไทยที่มีการจัดโครงสร้างเพื่อนำเสนอวิธีการสรุปเอกสารภาษาไทยที่มีการจัดโครงสร้างโดยใช้เทคนิคการสรุปแฟร็กทัลและพัฒนาซอฟต์แวร์ต้นแบบขึ้นมาใช้งาน

1.2 วัตถุประสงค์ของการศึกษา

- 1) เพื่อนำเสนอวิธีการสรุปเอกสารภาษาไทยที่มีการจัดโครงสร้างโดยใช้เทคนิคการสรุปแฟร็กทัล
- 2) เพื่อพัฒนาซอฟต์แวร์ต้นแบบสำหรับการสรุปเอกสารภาษาไทยที่มีการจัดโครงสร้างโดยใช้เทคนิคการสรุปแฟร็กทัล

1.3 ขอบเขตของการศึกษา

ขอบเขตของข้อมูลที่น่าสนใจ

- 1) ชนิดของคำภาษาไทย (Thai Part of Speech) จำนวน 47 ชนิด ในคลังข้อมูลออร์คิด (ORCHID Corpus)[1][8]
- 2) ชุดข้อมูลฝึกหัดที่จัดทำขึ้นจากคลังข้อมูลออร์คิด (ORCHID Training Set) จำนวน 7,054 ชุด
- 3) รายงานผลการดำเนินงานประจำปีพุทธศักราช 2550 ของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ จำนวน 8 หัวเรื่อง

ขอบเขตของงานค้นคว้าแบบอิสระ

- 1) ศึกษาหลักการทำงานของเทคนิคการสรุปแฟร็กทัลเพื่อนำมาใช้ในการออกแบบวิธีการสรุปเอกสารภาษาไทยที่มีการจัดโครงสร้างโดยใช้เทคนิคการสรุปแฟร็กทัล
- 2) ออกแบบวิธีการสรุปเอกสารภาษาไทยที่มีการจัดโครงสร้างโดยใช้เทคนิคการสรุปแฟร็กทัล
- 3) ตั้งสมมติฐานสำหรับการพิจารณาคำสำคัญของเอกสารต้นฉบับขึ้นมาว่า “คำสำคัญ คือ คำที่ปรากฏในเอกสารต้นฉบับซึ่งมีจำนวนมากกว่าหรือเท่ากับ 20% ของจำนวนคำทั้งหมดในเอกสารต้นฉบับ”
- 4) จัดทำข้อสรุปตามวิธีการที่ได้ออกแบบไว้โดยใช้รายงานผลการดำเนินงานประจำปีพุทธศักราช 2550 ของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติเป็นเอกสารต้นฉบับ
- 5) ใช้ภาษา Perl ในการพัฒนาซอฟต์แวร์ต้นแบบ โดยมีเพิ่มข้อความในรูปของ LaTeX ที่ถูกจัดทำขึ้นจากเอกสารต้นฉบับซึ่งถูกตัดคำด้วยโปรแกรม Swath[1] เป็นอินพุตของซอฟต์แวร์ และได้เพิ่มข้อความที่มีการระบุขอบเขตของประโยคในเอกสารต้นฉบับเป็นเอาต์พุตของซอฟต์แวร์