บทที่ 3
การอัดข้อความภาษาไทย

เราจะได้รับผลประโยชน์ และประสิทธิภาพที่เพิ่มขึ้นอย่างมากในการประมวลผลข้อมูลที่เกี่ยวกับข้อความ โดยวิธีการปรับเปลี่ยนหรือปรับปรุงวิธีการเก็บข้อมูลข้อความที่เก็บอยู่ตามปกติในการทำงานด้วยเครื่องคอมพิวเตอร์ให้อยู่ในรูปแบบใหม่ ซึ่งสามารถทำให้ประหยัดพื้นที่ของหน่วยเก็บข้อมูล และจะใช้เวลาในการประมวลผลน้อยลงด้วยเนื่องจากเอกสารมีขนาดลดลง นอกจากนี้เรายังได้เอกสารที่เก็บอยู่ในรูปแบบที่เข้ารหัสลับ (Encrypted Form) เพราะเอกสารไม่ได้ใช้รูปแบบของการเก็บอักษรปกติจึงเป็นการรักษาความลับของเอกสารอีกด้วย

ข้อความในเอกสารจะมีความเหลือเฟือรวมอยู่ด้วย ซึ่งจะแตกต่างกันไปในแต่ละภาษา ความเหลือเฟือของข้อมูลข้อความแบ่งออกได้ 2 ลักษณะคือ ความเหลือเฟือของสถิติการใช้งานของอักษรแต่ละตัวที่มีการใช้งานไม่เท่ากัน และความเหลือเฟืออีกลักษณะหนึ่งเป็นความเหลือเฟือที่เกี่ยวข้องกับซีแมนติคเฉพาะภาษาของข้อความนั้น (Salton 1989 : 111-112) วิธีการอัดข้อความอาจพัฒนาจากอัลกอริทึมมาตรฐานที่มีอยู่ เช่น อัลกอริทึมฮัฟแมน หรือพัฒนาเป็นวิธีการเฉพาะจากลักษณะของข้อมูลใช้งาน

3.1 ความเหลือเฟือในข้อความภาษาไทย

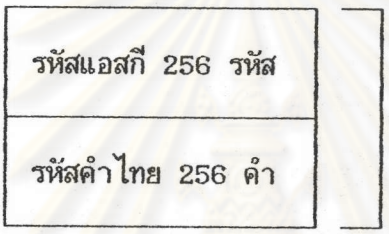
จากงานวิจัยวิเคราะห์ข้อมูลคำไทยเป็นงานวิจัยที่มีประโยชน์อย่างมากสำหรับการนำไปใช้ในงานด้านต่าง ๆ ของภาษาไทยกับคอมพิวเตอร์ ส่วนหนึ่งของงานวิจัยนี้ได้มีการวิจัยเกี่ยวกับการแจกแจงความถี่ของคำที่ใช้ในชีวิตประจำวัน ซึ่งเป็นตารางคำไทยจัดเรียงลำดับตามความถี่ของการใช้งานจากมากไปหาน้อยที่รวบรวมมาโดยการสุ่มตัวอย่างจากหนังสือ และเอกสารต่าง ๆ ได้แก่ หนังสือพิมพ์ วารสาร นิตยสาร รายงาน จดหมายราชการ หนังสืออ่านทั่ว ๆ ไป ยกเว้นหนังสือประเภทวรรณคดี หรือตำราวิชาการที่แปลมาจากหนังสือต่างประเทศ โดยใช้จำนวนคำตัวอย่างทั้งสิ้น 133,568 คำ แสดงไว้ดังตารางที่ 2 ในภาคผนวก

อีกส่วนหนึ่งของงานวิจัยนี้ เป็นการวิจัยความถี่การใช้งานของอักษรทุกตัวในภาษาไทย แสดงไว้ดังตารางที่ 1 ในภาคผนวก

ความถี่ของคำที่ใช้ในชีวิตประจำวัน และความถี่ของอักษรจากงานวิจัยดังกล่าว จัดว่าเป็นความเหลือเฟือของข้อความภาษาไทย ซึ่งสามารถนำไปใช้เป็นประโยชน์ในการอัดข้อความภาษาไทยได้เป็นอย่างดี

3.2 การอัดข้อความภาษาไทยที่ขึ้นกับซีแมนติค

การเก็บข้อความภาษาไทยในเครื่องคอมพิวเตอร์ อักษร 1 ตัวในข้อความจะใช้เนื้อที่ 8 บิตตามค่าของรหัสภาษาไทยในตารางแอสกี แนวความคิดเริ่มต้นในการอัดข้อความภาษาไทยคือ การเปลี่ยนรูปแบบการเข้ารหัสด้วยรหัสใหม่ที่มีขนาด 9 บิต โดยที่รหัสใหม่จะประกอบด้วยรหัสของอักษรจากตารางแอสกี 256 รหัส และรหัสของคำไทยที่ใช้ในชีวิตประจำวันจากงานวิจัยอีก 256 รหัส (คำ) การเข้ารหัสลักษณะนี้เป็นการแทนรหัสจากรหัสเดิมที่มีขนาดไม่คงที่คือ รหัสเดิมที่อาจเป็นอักษรตัวเดียว 1 ตัว หรือกลุ่มอักษรที่เป็นคำที่ตรงกับคำในตารางคำไทยด้วยรหัสใหม่ขนาดคงที่ 9 บิต



จากข้อความตัวอย่าง

"สำหรับตอนที่ 3 นี้จะกล่าวถึง เทคนิคทั่วไปเกี่ยวกับตัวอักษร และรหัสภาษาไทย ที่
1 2 3 4 5 6 7 8 9 10 11 12 13 14
มีผลต่อการพัฒนาทางซอฟแวร์ต่อไป"
15 16 17 18 19 20

ขนาดของข้อความเท่ากับ 107 ตัวอักษร (856 บิต) ประกอบด้วยคำไทยที่ใช้ในชีวิตประจำวัน 20 คำ (62 ตัวอักษร) และอักษรเดี่ยวอีก 45 ตัวอักษร ถ้าเราทดลองเปลี่ยนเป็นการแทนรหัสด้วยรหัสใหม่ขนาด 9 บิต จะได้ผลดังนี้

จำนวนบิตของรหัสคำไทยทั้งหมด = 20 x 9 = 180 บิต
จำนวนบิตของรหัสอื่น ๆ = 45 x 9 = 405 บิต
ขนาดใหม่ของข้อความ = 180 + 405 = 585 บิต
ดังนั้น ประสิทธิภาพการอัด = $100 - \frac{585}{856} \times 100$

= 31.7 เปอร์เซนต์

การอัดข้อมูลโดยหลักการดังกล่าว เป็นตัวอย่างหนึ่งของวิธีการอัดข้อมูลที่ขึ้นกับซีแมนติคภาษาไทย โดยการลดความเหลือเฟือที่เกิดจากซีแมนติคของภาษาไทยออกไป วิธีการนี้จะได้ผลใน

ทางลบทันทีถ้าข้อความนั้นไม่มีคำตรงกับคำไทยในตารางที่ได้จัดเตรียมไว้ หรือผลที่ได้จะลดลงถ้าข้อความนั้นมีชีแมนติคลักษณะอื่นรวมอยู่ด้วย เช่น ภาษาอังกฤษ การตีเส้นตาราง ดังนั้นการนำไปใช้งานจริงจะมีข้อจำกัด ไม่มีความยืดหยุ่น

3.3 การอัดข้อความภาษาไทยที่ไม่ขึ้นกับชีแมนติค

ในบทที่ 2 เราได้ศึกษาวิธีการอัดข้อมูลที่ไม่ขึ้นกับชีแมนติควิธีต่าง ๆ วิธีการที่ได้รับการยอมรับอย่างมากและมีการนำไปใช้อย่างกว้างขวางมีอยู่ 2 วิธี คือ วิธีการของอัลกอริทึมฮัฟแมน เป็นวิธีการอัดข้อมูลที่รับประกันความเหลือเฟือต่ำสุด ถือเป็นวิธีการแม่แบบของการอัดข้อมูลซึ่งมักจะนำไปเปรียบเทียบประสิทธิภาพกับวิธีการอื่น ๆ เสมอ สำหรับอีกวิธีการหนึ่ง คือ วิธีการของอัลกอริทึมแอลแซดดับบิว เป็นวิธีการสามารถลดความเหลือเฟือในลักษณะของการเกิดซ้ำอักษรหรือรูปแบบที่ใช้บ่อย อัลกอริทึมของวิธีการนี้ไม่ซับซ้อน ทำงานได้รวดเร็ว และให้ประสิทธิภาพการอัดที่ดีแต่ประสิทธิภาพจะต่ำในช่วงแรก ดังนั้นวิธีการนี้จึงเหมาะกับข้อมูลขนาดใหญ่

จากลักษณะความเหลือเฟือในข้อความภาษาไทยที่กล่าวถึงในตอนต้น จะเห็นว่าวิธีการอัดข้อมูลโดยอัลกอริทึมฮัฟแมนมีความเหมาะสมสำหรับนำมาใช้ในการอัดข้อความ เพราะสามารถลดความเหลือเฟือของอักษรที่มีความถี่ในการใช้งานไม่เท่ากัน สำหรับวิธีการของอัลกอริทึมแอลแซดดับบิว เป็นอีกวิธีการหนึ่งที่มีความเหมาะสมเช่นกันเพราะสามารถลดความเหลือเฟือของกลุ่มอักษร หรือ การเกิดซ้ำอักษร

3.4 แนวทางการวิจัยการอัดข้อความภาษาไทย

การวิจัยครั้งนี้เป็นการพัฒนาวิธีการอัดข้อมูลภาษาไทย โดยนำวิธีการอัดข้อมูลที่ไม่ขึ้นกับชีแมนติคที่เหมาะสมกับการอัดข้อความภาษาไทย ดังที่กล่าวในหัวข้อ 3.3 มาปรับเปลี่ยนอัลกอริทึมให้เป็นอัลกอริทึมที่มีการพิจารณาชีแมนติคของภาษาไทยด้วย โดยจะนำคำที่ใช้ในชีวิตประจำวันจากตารางที่ 2 จากภาคผนวก สร้างเป็นชีแมนติคของภาษาไทยเพิ่มเข้าไปในอัลกอริทึม

ขั้นตอนแรกของการวิจัย ได้พัฒนาโปรแกรมตามหลักการเดิมของแต่ละวิธีการขึ้นมาก่อนแล้วทดสอบความถูกต้องของการอัดและการขยายข้อมูล ต่อจากนั้นจึงทำการปรับเปลี่ยนอัลกอริทึมใหม่ ซึ่งรายละเอียดของการปรับเปลี่ยนจะกล่าวถึงในบทต่อไป