บทที่ 5

บทสรุปและข้อเสนอแนะ

ระบบสืบค้นข้อมูลภาษาไทยโดยใช้แฟ้มข้อมูลผกผันนั้น สามารถใช้โครงสร้างข้อมูลได้หลายรูปแบบ ซึ่งแต่ละรูปแบบจะมีข้อดีข้อเสียแตกต่างกันไป ส่วนที่สำคัญอีกส่วนหนึ่งคือการหาคำหลัก เพื่อนำคำหลักที่ได้ไปทำดัชนี แต่เนื่องจากคำภาษาไทยมีรูปแบบการเขียนที่ติดกัน จึงทำให้ต้องมีการพัฒนาอัลกอริทึมการจัดทำดัชนีขึ้น

อัลกอริทึมการจัดทำดัชนีที่ใช้ในระบบสืบค้นข้อมูลนั้น จะแตกต่างกับอัลกอริทึมการตัดคำที่ใช้ในโปรแกรมประมวลผลคำ เพราะในโปรแกรมประมวลผลคำจะใช้การตัดคำที่ท้ายบรรทัดเพื่อขึ้นบรรทัดใหม่ ซึ่งก็คือการหาจุดตัดของคำตรงบริเวณท้ายบรรทัดเท่านั้น แต่ในระบบสืบค้นข้อมูลต้องการได้คำหลัก เพื่อใช้ในการทำดัชนี การเขียนข้อความภาษาไทยนั้นเขียนติดกัน หนึ่งข้อความอาจจะมีการตัดคำได้หลายรูปแบบ ดังนั้นอัลกอริทึมการจัดทำดัชนีจะต้องสามารถดึงคำต่างๆ เหล่านั้นออกมาให้ได้ทั้งหมด จึงจะทำให้สามารถค้นหาคำที่ต้องการได้ครบถ้วนทุกคำ

ในการวิจัยครั้งนี้จึงได้เน้นหนักในเรื่องของอัลกอริทึมการจัดทำดัชนีเป็นหลัก เนื่องจากในส่วนอื่นๆ ของระบบสืบค้นข้อมูล สามารถนำสิ่งที่ได้มีการคิดค้นขึ้นมาแล้วนำมาใช้ร่วมกับภาษาไทยได้ แต่อัลกอริทึมการจัดทำดัชนีภาษาไทยที่เหมาะสมกับระบบสืบค้นข้อมูลนั้น ยังไม่ได้มีการวิจัยกันอย่างจริงจัง จึงเป็นเรื่องที่เน้นหนักเป็นพิเศษสำหรับการวิจัยครั้งนี้

สรุปการทำงานของอัลกอริทึมการจัดทำดัชนี

อัลกอริทึมการจัดทำดัชนีที่นำเสนอในงานวิจัยนี้ มีขั้นตอนหลักๆ อยู่ 4 ขั้นตอนดังต่อไปนี้

ขั้นตอนที่ 1 หาคำในพจนานุกรมที่มีขนาดใหญ่ที่สุด ในแต่ละตำแหน่งของข้อความ โดยคำที่ได้ออกมานั้นจะต้องไม่เป็นส่วนหนึ่งของคำที่ได้

ขั้นตอนที่ 2 นำคำที่ได้จากขั้นตอนที่ 1 มาสร้างกราฟการต่อและทับกันของคำ ซึ่งเป็นกราฟแบบมีทิศทางและน้ำหนัก

ขั้นตอนที่ 3 แต่ละส่วนของกราฟ หาเส้นทางที่มีค่าน้ำหนักน้อยที่สุดจากทางซ้ายสุดไปทางขวาสุดของแต่ละส่วนของกราฟ

ขั้นตอนที่ 4 เป็นขั้นตอนสุดท้าย เริ่มต้นด้วยการหากลุ่มของตัวอักษรที่ไม่รู้จัก (ไม่มีอยู่ในพจนานุกรม) ตำกลุ่มของตัวอักษรที่ได้ไปแบ่งพยางค์โดยใช้กฎ ซึ่งในการวิจัยครั้งนี้ใช้อัลกอริทึมการแบ่งคำภาษาไทย ที่ใช้ในโปรแกรมซียูไรต์เตอร์ พยางค์ที่ได้จากการแบ่งพยางค์โดยใช้กฎ จะถูกเก็บไว้ในเซตของคำที่แบ่งได้จากกฎ จากนั้นหาเซตของคำที่มีอยู่ในพจนานุกรมโดยใช้ผลลัพธ์ที่ได้จากขั้นตอนที่ 3 ลบคำที่เป็นส่วนหนึ่งของพยางค์ในเซตของคำที่ไม่อยู่ในพจนานุกรม และเพิ่มคำที่อยู่ในพจนานุกรมบางคำเพื่อให้สมบูรณ์ จะได้เป็นเซตของคำที่อยู่ในพจนานุกรม

ผลลัพธ์ที่ได้จากอัลกอริทึมการจัดทำดัชนีคือ เซตของคำที่มีอยู่ในพจนานุกรม และเซตของคำที่แบ่งได้จากกฎ คำที่อยู่ในทั้งสองเซตคือคำหลักที่จะนำไปทำเป็นดัชนีเพื่อใช้ในการค้นหาข้อมูลในระบบสืบค้นข้อความภาษาไทย

ในการวิจัยครั้งนี้ยังได้นำเสนออัลกอริทึมการจัดทำดัชนีภาษาไทยอีกหนึ่งวิธี โดยผลลัพธ์ที่ได้จากอัลกอริทึมนี้จะได้ เซตของคำที่มีอยู่ในพจนานุกรม และเซตของ sistring ของกลุ่มคำที่ไม่อยู่ในพจนานุกรม โดยมีขั้นตอนการทำงานดังต่อไปนี้

ขั้นตอนที่ 1 หาคำในพจนานุกรมที่มีขนาดใหญ่ที่สุด ในแต่ละตำแหน่งของข้อความ โดยคำที่ได้ออกมานั้นจะต้องไม่เป็นส่วนหนึ่งของคำที่ได้

ขั้นตอนที่ 2 นำคำที่ได้จากขั้นตอนที่ 1 มาสร้างกราฟการต่อและทับกันของคำ ซึ่งเป็นกราฟแบบมีทิศทางและน้ำหนัก

ขั้นตอนที่ 3 แต่ละส่วนของกราฟ หาเส้นทางที่มีค่าน้ำหนักน้อยที่สุดจากทางซ้ายสุดไปทางขวาสุดของแต่ละส่วนของกราฟ

ขั้นตอนที่ 4 เป็นขั้นตอนสุดท้าย เริ่มต้นด้วยการหากลุ่มของตัวอักษรที่ไม่รู้จัก (ไม่มีอยู่ในพจนานุกรม) นำกลุ่มของตัวอักษรที่ได้นำไปรวมกับคำที่อยู่ด้านหน้าของกลุ่มตัวอักษรเหล่านั้น เกิดเป็นกลุ่มตัวอักษรขึ้นมาใหม่ กลุ่มตัวอักษรใดที่ทับกันบางส่วน หรืออยู่ติดกันให้นำมารวมกัน sistring ของกลุ่มตัวอักษรที่ได้เก็บไว้ในเซตของ sistring ของกลุ่มคำที่ไม่อยู่ในพจนานุกรม จากนั้นหาเซตของคำที่มีอยู่ในพจนานุกรมโดยใช้ผลลัพธ์ที่ได้จากขั้นตอนที่ 3 ลบคำที่เป็นส่วนหนึ่งของพยางค์ในเซตของคำที่ไม่อยู่ในพจนานุกรม และเพิ่มคำที่อยู่ในพจนานุกรมบางคำเพื่อให้สมบูรณ์ จะได้เป็นเซตของคำที่อยู่ในพจนานุกรม

เซตของคำที่อยู่ในพจนานุกรมใช้เป็นคำหลักเพื่อในไปทำดัชนีของระบบสืบค้นข้อมูลโดยใช้แฟ้มข้อมูลผกผัน ส่วนเซตของ sistring ของกลุ่มคำที่ไม่อยู่ในพจนานุกรมนั้น ใช้โครงสร้างข้อมูลแบบทรัย ในการจัดเก็บ

ข้อเสนอแนะ

วิทยานิพนธ์นี้ได้ทำให้เกิดแนวความคิดในการวิจัยอื่นๆ ต่อไปอีก เช่น

- การออกแบบระบบสืบค้นข้อมูลภาษาไทยโดยใช้แฟ้มข้อมูลผกผัน ร่วมกับโครงสร้างข้อมูลแบบทรัย โดยที่คำหลักที่อยู่ในพจนานุกรมที่ได้จากการอัลกอริทึมการแบ่งคำนั้น ใช้แฟ้มข้อมูลผกผันในการจัดเก็บ ส่วนกลุ่มตัวอักษรที่ไม่รู้จักใช้โครงสร้างข้อมูลแบบทรัยในการจัดเก็บ
- การคิดค้นอัลกอริทึมการหาคำที่ไม่มีอยู่ในพจนานุกรม
- การสร้าง Search Engine ภาษาไทย เพื่อใช้ใน WWW (World Wide Web)
- สร้างพจนานุกรมภาษาไทย ที่มีจำนวนคำที่เหมาะสม อาจจะมีพจนานุกรมสำหรับชื่อคน คำทับศัพท์ เพื่อใช้ในอัลกอริทึมการแบ่งภาษาไทย โดยคำนึงถึงประสิทธิภาพในการแบ่งคำเป็นหลัก
- พัฒนาอัลกอริทึมการจัดทำดัชนีที่เกิดขึ้นในกรณีที่ 2 แบบที่ 4 และ 5 ซึ่งอัลกอริทึมการจัดทำดัชนีในการวิจัยนี้ไม่สามารถทำได้

วิธีการจัดเก็บคำศัพท์เพื่อใช้ในการตรวจสอบการแบ่งคำนั้น ถือได้ว่าเป็นส่วนสำคัญซึ่งจะมีผลต่อประสิทธิภาพของอัลกอริทึมการแบ่งคำ ถ้าคำศัพท์ทั้งหมดสามารถเก็บอยู่ในหน่วยความจำหลักของเครื่องคอมพิวเตอร์ได้ ก็จะทำให้ความเร็วในการเข้าถึงดีขึ้น วิธีการเข้าถึงคำศัพท์นั้นเป็นสิ่งที่สำคัญเช่นกัน เนื่องจากอัลกอริทึมการแบ่งคำที่คิดค้นขึ้นนั้น จะต้องมีการตรวจสอบกับคำศัพท์อยู่ตลอด ถ้าการเข้าถึงคำศัพท์มีประสิทธิภาพเท่าไร อัลกอริทึมการแบ่งคำก็จะมีประสิทธิภาพดียิ่งขึ้น