บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

มนุษย์เราอาศัยห้องสมุดเป็นแหล่งค้นหาข้อมูลมานานแล้ว เนื่องจากห้องสมุดเป็นศูนย์รวมข้อมูลต่างๆ หลากหลายประเภท จึงทำให้มนุษย์ประสบปัญหาในการค้นหาข้อมูลในห้องสมุดที่มีอยู่อย่างมหาศาล แต่ในปัจจุบันเทคโนโลยีทางคอมพิวเตอร์ได้พัฒนาไปมาก ส่งผลให้มนุษย์สามารถค้นหาข้อมูลได้สะดวกมากขึ้น โดยนักวิจัยได้นำเอาเทคนิคทางโครงสร้างข้อมูลมาสนับสนุนพัฒนาระบบการค้นคืนข้อความ เพื่อเป็นกลไกหรือเครื่องมือให้มนุษย์ใช้ในการค้นหาอย่างมีประสิทธิภาพทั้งในแง่เนื้อที่ที่ใช้จัดเก็บข้อมูลกับความรวดเร็วในการค้นหา

การค้นหาข้อมูลแบ่งได้เป็น 2 ชนิด คือ

1. การค้นหาแบบไม่อาศัยดรรชนี (no indexing)

การค้นหาแบบไม่อาศัยดรรชนี หรือการใช้กราดตรวจแบบฟรีเท็ค (free-text scan) [5] เป็นการค้นหาข้อมูลโดยไม่ได้จัดเตรียมข้อมูล หรือไม่มีการสร้างดรรชนี (indexing) เตรียมล่วงหน้าเอาไว้ก่อน การค้นหาจะนำข้อมูลที่ต้องการค้นหาไปเปรียบเทียบกับข้อมูลทุกๆ ตัวที่มีอยู่ และการค้นหาประสบผลสำเร็จต่อเมื่อพบข้อมูลนั้น

จะเห็นว่าการค้นข้อมูลแบบนี้เหมาะสมกับข้อมูลที่มีขนาดเล็ก แต่สำหรับข้อมูลมากๆ จะมีประสิทธิภาพค่อนข้างช้า บางครั้งจำเป็นต้องอาศัยฮาร์ดแวร์ (hardware) เข้าช่วย

2. การค้นหาแบบอาศัยดรรชนี (indexing)

เป็นการค้นหาข้อมูลโดยอาศัยดรรชนีที่สร้างเตรียมไว้ล่วงหน้าก่อน เพื่อความรวดเร็วในการค้นหาภายหลัง วิธีการสร้างดรรชนีนั้นมีหลากหลายวิธี เช่น การใช้แฟ้มผกผัน (inverted file) [5] เป็นต้น และวิธีเหล่านี้ส่วนใหญ่แล้วมักมีการจัดทำคำหลัก (keyword) เพื่อเป็นคีย์ (key) ในการค้นหาภายหลัง ซึ่งการจัดทำคำหลักนี้ ทำให้เกิดข้อจำกัดบางประการ คือ

1. เนื่องจากจำเป็นต้องนำส่วนหนึ่งของข้อมูลมาเป็นคำหลัก เพื่อใช้ในการค้นคืนภายหลัง ซึ่งเรียกว่า การจัดทำดรรชนีจึงทำให้เกิดข้อจำกัดว่า ควรจะเลือกส่วนใดของข้อมูลมาเป็นคำหลักบ้าง และคำหลักควรยาวเท่าใด เป็นต้น

2. เนื่องจากการค้นคืนอาศัยดรรชนี หรือคือคำหลักเป็นคีย์ในการค้นหาข้อมูล ดังนั้นการค้นหาข้อมูลจึงถูกจำกัดตามคุณสมบัติของคำหลักนั้นๆ ไปโดยปริยาย

จะเห็นว่าประเด็นสำคัญ คือ การสร้างดรรชนีให้กับข้อมูลโดยดรรชนีดังกล่าวส่วนใหญ่ถูกออกแบบให้มีคำหลักเป็นคีย์ในการค้นหาข้อมูล ซึ่งคำหลักจำเป็นต้องมีจุดเริ่มต้นกับจุดสิ้นสุดของคำ และเมื่อพิจารณาถึงโครงสร้างทางภาษาของข้อมูลที่เป็นภาษาอังกฤษ พบว่าการสร้างดรรชนีดังกล่าวมีความเหมาะสม เนื่องจากโครงสร้างภาษาอังกฤษประกอบด้วย ประโยคที่ประกอบด้วยคำหลายๆ คำเรียงต่อๆ กันไป โดยมีช่องว่าง (space) เป็นตัวคั่นคำ เช่น I go to school. ซึ่งการกำหนดจุดเริ่มต้นกับจุดสิ้นสุดของคำหลักทำได้ง่าย แต่โครงสร้างทางภาษาของข้อมูลที่เป็นภาษาไทยมิเป็นเช่นนั้น เนื่องจากโครงสร้างภาษาไทยประกอบด้วยคำหลายๆ คำเรียงต่อๆ กันไป โดยไม่มีตัวคั่นคำที่ชัดเจน หรือคำบางคำยังเป็นคำซ้อน คือคำที่ประกอบด้วยคำอื่นๆ อีกหลายๆ คำ เช่น คำว่า “ตากลม” อาจหมายถึง “ตา-กลม” หรือ “ตาก-ลม” ก็ได้ ทำให้การกำหนดคำหลักทำได้ยากกว่า ดังนั้นการออกแบบดรรชนีที่ใช้คำหลักเป็นคีย์ในการค้นหาข้อมูล อาจไม่เหมาะสมกับข้อมูลภาษาไทยก็ได้ ซึ่งปัญหาเหล่านี้อาจแก้ไขได้ โดยการนำเอาโครงสร้างข้อมูลที่สร้างดรรชนีที่ไม่ใช้คำหลักเป็นคีย์ในการค้นหาข้อมูลมาพัฒนาระบบการค้นคืน ซึ่งโครงสร้างหนึ่ง คือ ต้นไม้แพ็ต (PAT tree) ทั้งนี้ทั้งนั้นโครงสร้างข้อมูลแบบนี้ยังสามารถนำไปใช้กับหลักการสร้างดรรชนีแบบเดิมได้อีกด้วย

1.2 วัตถุประสงค์

จากที่กล่าวมาแล้วข้างต้น วิทยานิพนธ์ฉบับนี้จึงได้กำหนดวัตถุประสงค์หลัก คือ

1. เพื่อศึกษาและพัฒนาโปรแกรมจัดเก็บดรรชนีของข้อความภาษาไทย / อังกฤษโดยอาศัยโครงสร้างข้อมูลต้นไม้แพ็ต

2. เพื่อศึกษาและพัฒนาโปรแกรมค้นคืนข้อความภาษาไทย / อังกฤษโดยอาศัยโครงสร้างข้อมูลต้นไม้แพ็ตที่สร้างขึ้น

1.3 ขอบเขตของการวิจัย

1. โปรแกรมจัดเก็บดรรชนีของข้อความสามารถจัดเก็บเพิ่มเติมข้อความได้
2. โปรแกรมค้นคืนสามารถค้นคืนแบบเติมหน้า (prefix) และแบบบูล (boolean)
3. การพัฒนาโปรแกรมอาศัยภาษา C และทำงานบนระบบปฏิบัติการยูนิกส์ (unix)
4. ส่วนของการทดสอบจะทดสอบกับข้อความประเภทต่างๆ เช่น บทความทั่วไป บทความข่าวสาร ในด้านความรวดเร็วและความถูกต้องในการค้นคืน

1.4 ขั้นตอนและวิธีการดำเนินการวิจัย

1. ศึกษาโครงสร้างข้อมูลแบบต้นไม้แพ็ต
2. ศึกษาโครงสร้างข้อมูลแบบแพ็ตอะเรย์ (PAT array) เพื่อใช้แทนต้นไม้แพ็ต
3. พัฒนาโปรแกรมเพื่อสร้างดรรชนีของข้อมูลโดยอาศัยต้นไม้แพ็ต
4. พัฒนาโปรแกรมเพื่อสร้างดรรชนีของข้อมูลโดยอาศัยแพ็ตอะเรย์
5. พัฒนาโปรแกรมเพื่อค้นข้อมูลโดยอาศัยดรรชนีที่สร้างขึ้น
6. พัฒนาโปรแกรมเพื่อใช้สำหรับการรับข้อมูลต่างๆ จากผู้ใช้
7. พัฒนาโปรแกรมเพื่อใช้แสดงผลลัพธ์และคำอธิบายต่างๆ
8. ทำการทดสอบและปรับปรุงโปรแกรมให้ปฏิบัติงานได้อย่างมีประสิทธิภาพ

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. ผู้สนใจทั่วไปสามารถนำเอาโปรแกรมไปใช้ประโยชน์ต่อได้
2. เพื่อเป็นแนวทางในการส่งเสริมให้มีการพัฒนาระบบการค้นคืนข้อความภาษาไทยมากยิ่งขึ้น
3. สามารถนำไปพัฒนากับข้อมูลที่มีลักษณะเป็นตัวอักษรที่ยาวเรียงต่อๆกันไปโดยไม่มีการแบ่งคำ

1.6 โครงสร้างของวิทยานิพนธ์

วิทยานิพนธ์ฉบับนี้แบ่งออกเป็น 7 บท อันได้แก่

บทที่ 1 กล่าวถึงความเป็นมาของปัญหา วัตถุประสงค์ ขอบเขต และเนื้อหาของวิทยานิพนธ์

บทที่ 2 กล่าวถึงแนวความคิดและทฤษฎีของโครงสร้างข้อมูลที่นำมาประยุกต์ใช้กับระบบการค้นคืนข้อความแบบต่างๆ โดยเปรียบเทียบข้อดี-ข้อเสียของแต่ละแบบด้วย

บทที่ 3 กล่าวถึงการวิเคราะห์ออกแบบระบบการค้นคืนข้อความภาษาไทย โดยเลือกโครงสร้างข้อมูลต้นไม้แพ็ตและแพ็ตอะเรย์ เป็นกลไกในการจัดการข้อมูลภาษาไทย

บทที่ 4 กล่าวถึงรายละเอียด และขั้นตอนการสร้างต้นไม้แพ็ต แพ็ตอะเรย์ รวมถึงการค้นคืนข้อความแบบข้อมูลร่วมกับแบบบูลอีกด้วย

บทที่ 5 กล่าวถึงการทดสอบความสัมพันธ์ของขนาดของคีย์กับขนาดแฟ้มดรรชนี

บทที่ 6 กล่าวถึงผลการทดสอบโปรแกรมระบบการค้นคืนข้อความภาษาไทยโดยใช้ต้นไม้แพ็ต

บทที่ 7 กล่าวถึงบทสรุปและข้อเสนอแนะในการพัฒนาระบบการค้นคืนข้อความภาษาไทยโดยใช้ต้นไม้แพ็ต

นอกจากเนื้อหาดังกล่าวแล้ววิทยานิพนธ์นี้ยังประกอบด้วยภาคผนวกอีก 1 บท ดังนี้

ภาคผนวก ก. แสดงแฟ้มข้อมูลต่างๆ ที่ใช้ในระบบการค้นคืนข้อความภาษาไทยโดยใช้ต้นไม้แพ็ต