บทที่ 1

บทนำ

ความเป็นมาและปัญหา

ระบบสืบค้นข้อมูล (Information Retrieval) ได้ถูกพัฒนาขึ้นเพื่อช่วยจัดการกับตัวอักษรที่มีจำนวนมาก ซึ่งระบบนี้ได้ถูกพัฒนาขึ้นเมื่อปี 1940 ในปัจจุบันหลาย ๆ มหาวิทยาลัย องค์กร ห้องสมุด รวมถึงอินเตอร์เนต (Internet) ได้นำระบบนี้เข้ามาใช้ เพื่อช่วยในการค้นหาข้อมูลต่าง ๆ การค้นหาข้อมูลที่ต้องการจากเอกสารต่าง ๆ จำนวนหลายล้านฉบับนั้นจะต้องใช้เวลาในการค้นหามาก แต่ระบบสืบค้นข้อมูลจะเป็นสิ่งที่ทำให้การค้นหาข้อมูลที่ต้องการได้อย่างรวดเร็วขึ้น สรุปได้ว่าระบบการสืบค้นข้อมูลมีไว้เพื่อให้สามารถค้นหาข้อมูลที่ต้องการได้ง่ายขึ้น

ระบบสืบค้นข้อมูลสามารถสร้างขึ้นได้โดยใช้แฟ้มข้อมูลได้หลายอย่าง หนึ่งในนั้นคือการใช้แฟ้มข้อมูลผกผัน ลักษณะของแฟ้มข้อมูลผกผันจะมีการนำเอาคำจากเอกสารที่ต้องการจะเก็บ นำคำเหล่านั้นมาทำเป็นดัชนีเพื่อใช้ในการค้นหาข้อมูลต่อไป สังเกตได้ว่าจะต้องมีการดึงเอาคำออกจากเอกสาร คำในภาษาอังกฤษนั้นจะมีการเขียนโดยมีช่องว่างแบ่งแยกระหว่างคำ จึงไม่มีปัญหาในการแบ่งคำ หรือการดึงคำเพื่อที่จะนำคำเหล่านั้นมาทำดัชนี แต่คำในภาษาไทยนั้นจะมีลักษณะการเขียนซึ่งเป็นเอกลักษณ์เฉพาะตัว ผิดจากภาษาอังกฤษอยู่หลายประการ แต่ลักษณะที่ทำให้เกิดปัญหากับระบบสืบค้นข้อมูลที่ใช้แฟ้มข้อมูลผกผัน คือการเขียนไม่เว้นระยะระหว่างคำ จากที่กล่าวไว้ข้างต้นแฟ้มข้อมูลผกผันจะต้องนำเอาคำจากเอกสารนำมาทำดัชนี แต่คำในภาษาไทยมีการเขียนติดกัน จึงต้องมีการคิดค้นอัลกอริทึมการค้นหาคำเพื่อที่จะนำมาจัดทำดัชนี

ปัจจุบันในการประมวลผลคำภาษาไทยนั้น จะมีการตัดพยางค์ที่ท้ายบรรทัดเพื่อขึ้นบรรทัดใหม่ การตัดพยางค์ในลักษณะนี้ไม่สามารถนำมาใช้กับระบบสืบค้นข้อมูลโดยใช้แฟ้มผกผันได้ เนื่องจากการตัดพยางค์ในลักษณะนี้ จะเป็นการหาจุดที่สามารถตัดพยางค์ได้ ไม่ใช้การดึงออกมาเป็นคำ ๆ จากจุดนี้จึงเป็นที่มาของการทำวิทยานิพนธ์ฉบับนี้

เนื่องด้วยระบบการสืบค้นข้อมูลโดยใช้แฟ้มผกผันนั้นจะต้องนำคำจากเอกสารมาทำดัชนี คำที่จะนำไปทำดัชนีคือคำที่มาจากพจนานุกรม แต่ไม่เสมอไป เพราะคำในภาษาไทยจะมีคำที่เกิดขึ้นมาใหม่เสมอ เช่น คำทับศัพท์ภาษาต่างประเทศ ชื่อคน รวมถึงการเปลี่ยนรูปของคำเพื่อใช้ในโคลงกลอนต่าง ๆ ดังนั้นอัลกอริทึมที่คิดค้นขึ้นจึงยึดพจนานุกรมเป็นหลัก และมีการคิดค้นวิธีการที่จะสามารถจัดการกับคำที่เกิดขึ้นใหม่ หรือคำที่ไม่มีอยู่ในพจนานุกรม

วัตถุประสงค์ของวิทยานิพนธ์

จากที่กล่าวมาแล้วข้างต้น วิทยานิพนธ์ฉบับนี้จึงได้กำหนดวัตถุประสงค์หลัก คือการศึกษาและพัฒนาอัลกอริทึมการจัดทำดัชนีภาษาไทย โดยใช้พจนานุกรมเป็นหลัก เพื่อนำมาใช้ในระบบสืบค้นข้อมูลภาษาไทยโดยใช้แฟ้มผกผัน

ขอบเขตและเงื่อนไขของวิทยานิพนธ์

1. การออกแบบอัลกอริทึมการจัดทำดัชนีภาษาไทย จะใช้พจนานุกรมเป็นหลัก และคิดค้นวิธีการจัดการกับคำที่ไม่มีอยู่ในพจนานุกรม
2. อัลกอริทึมการจัดทำดัชนีภาษาไทย ไม่มีเรื่องไวยากรณ์ของภาษาไทยเข้ามาเกี่ยวข้อง
3. โปรแกรมทำงานภายใต้ระบบปฏิบัติการแบบ 32 บิต
4. การพัฒนาโปรแกรม จะใช้ภาษาระดับสูง (high level language) ที่มีความยืดหยุ่นสูง และพัฒนาโปรแกรมบนเครื่องไมโครคอมพิวเตอร์
5. รหัสภาษาไทยที่ใช้ จะใช้รหัสของสำนักงานมาตรฐานอุตสาหกรรม (ส.ม.อ.)
6. การพิจารณาประสิทธิภาพจะเปรียบเทียบกับอัลกอริทึมการแบ่งคำที่ใช้พจนานุกรมเช่นเดียวกับอัลกอริทึมที่คิดค้นขึ้น

ขั้นตอนการวิจัย

1. ศึกษาอัลกอริทึมการแบ่งคำ การตัดคำภาษาไทย ที่มีใช้อยู่ในปัจจุบัน
2. ศึกษาลักษณะคำในภาษาไทย ถึงผลกระทบของการเขียนคำติดกัน แนวความคิดในการออกแบบอัลกอริทึมที่เหมาะสม

3. ออกแบบและพัฒนาอัลกอริทึมการจัดทำดัชนี เพื่อใช้ในระบบสืบค้นข้อความภาษาไทย โดยใช้แฟ้มข้อมูลผกผัน

4. ศึกษาทางทฤษฎีถึงประสิทธิภาพในแง่เวลาที่ใช้ในอัลกอริทึมการจัดทำดัชนี จำนวนคำที่ได้จากการตัดคำ

5. ทดลองกับเอกสารที่ผ่านอัลกอริทึมการจัดทำดัชนี กับระบบสืบค้นข้อมูลโดยใช้แฟ้มผกผันที่มีใช้อยู่ในปัจจุบัน ว่าใช้งานได้หรือไม่

6. สรุปผลการวิจัยและข้อเสนอแนะ

ประโยชน์ที่คาดว่าจะได้รับ

1. ความรู้และแนวความคิดในการพัฒนาโครงสร้างข้อมูลและอัลกอริทึมที่ใช้ในการสืบค้นข้อมูล โดยใช้แฟ้มข้อมูลแบบผกผัน

2. ความเข้าใจถึงการทำงานของโครงสร้างข้อมูล และอัลกอริทึมในการสืบค้นคำ

3. ใช้เป็นแนวทางในการวิจัยเกี่ยวกับการออกแบบอัลกอริทึมการจัดทำดัชนีในภาษาอื่น ๆ ที่มีลักษณะการเขียนประโยคคล้ายภาษาไทย

4. ผลที่ได้จากการวิจัยสามารถนำไปสร้าง Search Engine สำหรับภาษาไทย เพื่อใช้ในการค้นหาข้อมูล

5. แนวคิดในการออกแบบระบบสืบค้นข้อความภาษาไทย โดยใช้แฟ้มข้อมูลแบบอื่น ๆ

โครงสร้างของวิทยานิพนธ์

วิทยานิพนธ์ฉบับนี้แบ่งออกเป็น 5 บท อันได้แก่

บทที่ 1 กล่าวถึงความเป็นมาของปัญหา วัตถุประสงค์ ขอบเขต และเนื้อหาของวิทยานิพนธ์

บทที่ 2 กล่าวถึงการแบ่งคำที่มีอยู่ในปัจจุบัน การใช้พจนานุกรมในการตัดคำ ภาษาไทยมีลักษณะเป็นเช่นไร และมีกี่รูปแบบ ลักษณะอัลกอริทึมการจัดทำดัชนีที่นำมาใช้กับระบบสืบค้นข้อมูล โดยใช้แฟ้มข้อมูลผกผัน รวมถึงทฤษฎีต่าง ๆ ที่เกี่ยวข้อง

บทที่ 3 กล่าวถึงแนวความคิดในการพัฒนาและออกแบบอัลกอริทึม และอัลกอริทึมที่คิดค้นขึ้นในการวิจัยครั้งนี้

บทที่ 4 รายงานผลการทดสอบอัลกอริทึมที่คิดค้นขึ้น

บทที่ 5 บทสรุปพร้อมทั้งคำแนะนำสำหรับการวิจัยต่อ ๆ ไป

นอกจากเนื้อหาดังกล่าวแล้ววิทยานิพนธ์นี้ยังประกอบด้วยภาคผนวกอีก 2 บท ดังนี้

ภาคผนวก ก. แสดงถึงอัลกอริทึมการตัดคำที่ใช้ในโปรแกรมซียูไรท์เตอร์ ซึ่งนำมาใช้เป็นส่วนหนึ่งของการทำวิทยานิพนธ์ฉบับนี้

ภาคผนวก ข. ตัวอย่างข้อความภาษาไทยที่ใช้ในการทดสอบอัลกอริทึม

ภาคผนวก ค. ความหมายของ sistrings และรูปแบบโครงสร้างข้อมูลแบบ trie