บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

สารสนเทศในรูปแบบของตัวอักษรหรือข้อความเป็นรูปแบบที่มีความสำคัญ เพราะว่ามนุษย์ใช้ตัวอักษรในการสื่อสารกัน ในปัจจุบันสารสนเทศที่เป็นตัวอักษรได้ผลิตออกมามากมาย เช่น หนังสือ วารสาร รายงานการประชุม เอกสารวิชาการ เป็นต้น ปัจจุบันยุคแห่งสารสนเทศได้มาถึงแล้ว และมาพร้อมกับกองสารสนเทศจำนวนมากที่รอให้ทุกคนเลือก เมื่อต้องการค้นหาข้อมูลที่ต้องการ ซึ่งงานเหล่านี้มักจะมีปริมาณข้อมูลสูงมาก ทำให้เป็นการยากที่จะค้นหาข้อมูลแต่ที่ตรงตามต้องการ (Relevant) ของผู้ใช้เท่านั้น จะเห็นว่าการที่จะช่วยผู้ใช้ค้นหาข้อมูลที่ตรงตามตามต้องการจะช่วยลดปริมาณข้อมูลที่ค้นคืนโดยแยกเอกสารที่ไม่ตรงตามต้องการ(Nonrelevant) ออกมาและยังช่วยลดเวลาในการตรวจสอบข้อมูลว่าตรงตามต้องการหรือไม่

ระบบค้นคืนข้อมูลในปัจจุบันจะเกี่ยวข้องกับการหาเอกสารที่ตรงตามต้องการกับรูปแบบความต้องการของผู้ใช้ (Information need) หรือคิวรี (Query) จากฐานข้อมูลของเอกสารจำนวนมาก การแสดงความต้องการของผู้ใช้โดยการใช้คิวรี ซึ่งประกอบไปด้วยคำจำนวนหนึ่ง ระบบค้นคืนข้อมูลจะหาคำที่ใช้ในคิวรี เปรียบเทียบกับคำที่ใช้ในเอกสารของฐานข้อมูล และแสดงผลการค้นคืนเป็นรายการเอกสารที่พบคำเหล่านั้น ซึ่งน่าจะตรงกับความต้องการของผู้ใช้ ปัญหาพื้นฐานของระบบค้นคืนคือ ปกติคิวรีหนึ่งๆ จะมีจำนวนคำที่น้อย และมีใจความที่ไม่สมบูรณ์ได้ ทำให้ไม่สามารถแสดงความต้องการที่แท้จริงของผู้ใช้ได้ นอกจากนี้ยังมีปัญหาที่เกิดจากความแตกต่างของคำที่ใช้แสดงสิ่งเดียวกันของผู้ใช้ระบบกับคำที่ใช้อยู่ในเอกสาร ไม่เหมือนกันเช่น คำว่า “หมู” และ “สุกร” หมายถึงสิ่งเดียวกัน ได้มีความพยายามแก้ไขปัญหาเหล่านี้ได้แก่ การทำสเตมมิ่ง (Stemming) ซึ่งเป็นการลดรูปของคำ เช่น running หรือ runner ลดรูปคำเป็น run (วิธีนี้ไม่สามารถนำมาใช้กับภาษาไทยได้) และการใช้บัญชีคำ (Thesaurus) โดยใช้รวบรวมกลุ่มคำที่มีความสัมพันธ์กันไว้ด้วยกันเช่น คำว่า “ยานพาหนะ” “รถยนต์” และ “รถกระบะ” ถือเป็นกลุ่มคำเดียวกัน

ดังนั้นงานวิจัยนี้ได้เสนอรูปแบบการค้นคืน เพื่อช่วยปรับปรุงผลการค้นคืนให้ดีขึ้นดังนี้

1. การให้น้ำหนักคำ เป็นส่วนที่จำเป็นกับระบบค้นคืนข้อมูล เพราะสามารถช่วยสร้างผลของการค้นคืนที่ดีกว่า เมื่อได้ผลการค้นคืนมาเป็นจำนวนมาก ระบบจะแสดงผลการค้นคืนในแบบรายการเอกสาร ที่ถูกจัดลำดับตามความสำคัญ ที่น่าจะตรงตามความต้องการของผู้ใช้ (Relevance ranking)

2. การค้นคืนย้อนกลับ (Relevance feedback) เป็นวิธีที่ออกแบบมาเพื่อปรับปรุงคิวรีใหม่โดยอาศัยผลการค้นคืนเริ่มต้นที่ผู้ใช้ป้อนกลับให้แก่ระบบ

ทั้งนี้ในการวิจัยนี้จะอาศัยโครงสร้างแถวลำดับแพ็ต (PAT array) เป็นดัชนี (Index)เพื่อใช้ในระบบค้นคืนข้อมูลที่พัฒนาขึ้น

1.2 การวิจัยที่เกี่ยวข้อง

โปรแกรมระบบค้นคืนข้อความ ได้เคยมีผู้จัดทำขึ้นมาบ้างแล้ว เช่นโปรแกรมระบบค้นคืนข้อความโดยใช้แนวความคิดแฟ้มข้อมูลผกผัน (พรทิพย์ บัวสาม, 2535) โปรแกรมระบบค้นคืนข้อความโดยใช้แนวความคิดแบบจำลองปริภูมิเวกเตอร์ (สาโรช เมาลานนท์, 2535) และระบบการค้นคืนข้อความภาษาไทยโดยใช้ต้นไม้แพ็ต (เปรมิน จินดาวิมลเลิศ, 2539) ซึ่งระบบสามารถจัดเก็บและค้นหาข้อความได้ดีพอสมควร

ข้อดีของระบบเดิม

1. สามารถค้นหาข้อมูลได้รวดเร็ว เมื่อเทียบกับปริมาณข้อมูลที่จัดเก็บ
2. สามารถใช้ได้กับข้อความภาษาไทย/อังกฤษ

ข้อเสียของระบบเดิม

1. ระบบติดต่อกับผู้ใช้ไม่สะดวก
2. การประเมินผลคำนึงถึงเพียงความเร็วของการค้นคืนอย่างเดียวโดยไม่คำนึงถึงความสามารถการค้นคืนได้ตรงตามต้องการผู้ใช้
3. ไม่มีระบบช่วยเหลือในการปรับปรุงผลการค้นคืน

1.3 วัตถุประสงค์ของการวิจัย

เพื่อพัฒนาระบบค้นคืนสารสนเทศเพื่อใช้ในการค้นคืนข้อความ โดยวิธีการจัดลำดับ และวิธีการค้นคืนย้อนกลับ บนโครงสร้างข้อมูลแถวลำดับแพ็ต

1.4 ขอบเขตการวิจัย

1. ระบบค้นคืนข้อความสามารถใช้ได้กับเอกสารที่เป็นภาษาไทย/อังกฤษ
2. ระบบค้นคืนข้อความสามารถค้นคืนแบบวิธีการจัดลำดับ และวิธีการค้นคืนย้อนกลับ
3. ระบบค้นคืนข้อความอาศัยแฟ้มดัชนีของข้อความโดยใช้โครงสร้างข้อมูลแถวลำดับแพ็ต
4. ระบบค้นคืนข้อความนี้จะมีขั้นตอนการให้น้ำหนักและการค้นคืนย้อนกลับที่เหมาะสมกับโครงสร้างข้อมูลแถวลำดับแพ็ต
5. ระบบค้นคืนข้อความสามารถค้นคืนแบบตรง(exact match)กับคิวรีเท่านั้น
6. ลักษณะคิวรีที่ป้อนอยู่ในรูปแบบตรรกะหรือ(OR) เท่านั้น
7. การทดสอบ จะทดสอบกับเอกสารประเภทต่างๆ เช่นบทความทั่วไป บทความข่าวสาร ที่มีขนาดประมาณ 1 Megabyte จำนวนประมาณ 100 เอกสารในด้านของความถูกต้อง
8. การพัฒนาโปรแกรมอาศัยเครื่องไมโครคอมพิวเตอร์ และภาษาที่ใช้ในการพัฒนาโปรแกรมจะใช้ภาษาวิชวลเบสิค ทำงานบนระบบปฏิบัติการวินโดว์ 95 รุ่นภาษาไทย

1.5 ขั้นตอนและวิธีดำเนินการวิจัย

1. ศึกษาขั้นตอนวิธีการจัดลำดับ และวิธีการค้นคืนย้อนกลับ
2. ศึกษาระบบค้นคืนสารสนเทศที่ใช้โครงสร้างข้อมูลแถวลำดับแพ็ตเพื่อประยุกต์ตามวิธีการจัดลำดับและวิธีการค้นคืนย้อนกลับ
3. ออกแบบโครงสร้างของข้อมูลและโครงสร้างแฟ้มข้อมูลที่ใช้ตามแนววิธีนี้
4. พัฒนาโปรแกรมเพื่อค้นคืนข้อความที่ใช้ตามแนววิธีนี้
5. เขียนโปรแกรมและทดสอบโปรแกรม
6. สรุป ทดสอบ ประเมินผล และข้อเสนอแนะ

1.6 ประโยชน์ที่คาดว่าจะได้รับ

ทำให้ทราบถึงปัญหาและแนวทางการพัฒนาระบบค้นคืนสารสนเทศเช่น ระบบค้นคืนหนังสือในห้องสมุด ระบบค้นคืนข้อมูลในซีดีรอม และระบบฐานข้อมูลกฎหมาย เป็นต้น ให้มีประสิทธิภาพดีขึ้น ด้วยวิธีการจัดลำดับ และวิธีการค้นคืนย้อนกลับโดยเฉพาะกับข้อมูลภาษาไทย