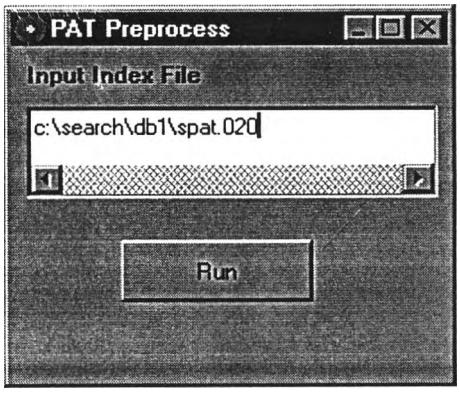
บทที่ 4

การพัฒนาโปรแกรม

ในการพัฒนาโปรแกรม เนื่องจากระบบค้นคืนเอกสารที่พัฒนาขึ้น จำเป็นต้องมีการติดต่อ (Interface) กับผู้ใช้ในส่วนการรับคิวรีจากผู้ใช้ ส่วนการแสดงผลการค้นคืนแบบจัดลำดับ และส่วนการให้ผู้ใช้เลือกเอกสารที่ตรงตามต้องการเพื่อใช้ในการค้นคืนย้อนกลับ ดังนั้นจึงเลือกใช้เครื่องมือที่สามารถพัฒนาโปรแกรมติดต่อประสานกับผู้ใช้แบบกราฟฟิก (Graphic) ซึ่งในการพัฒนาโปรแกรมระบบค้นคืนเอกสารในที่นี้ใช้โปรแกรมภาษาวิชวลเบสิค (Visual Basic) บนเครื่องไมโครคอมพิวเตอร์ โดยใช้ระบบปฏิบัติการวินโดว์ภาษาไทย ระบบค้นคืนเอกสารที่พัฒนาขึ้นประกอบด้วย 2 โปรแกรมคือ

1. โปรแกรมจัดเก็บดัชนีเพื่อใช้เก็บค่าน้ำหนักคำ
2. โปรแกรมค้นคืนเอกสาร

4.1 โปรแกรมจัดเก็บดัชนีเพื่อเก็บค่าน้ำหนักคำ



รูปที่ 4.1 แสดงโปรแกรมสร้างดัชนีเพื่อเก็บค่าน้ำหนักคำ

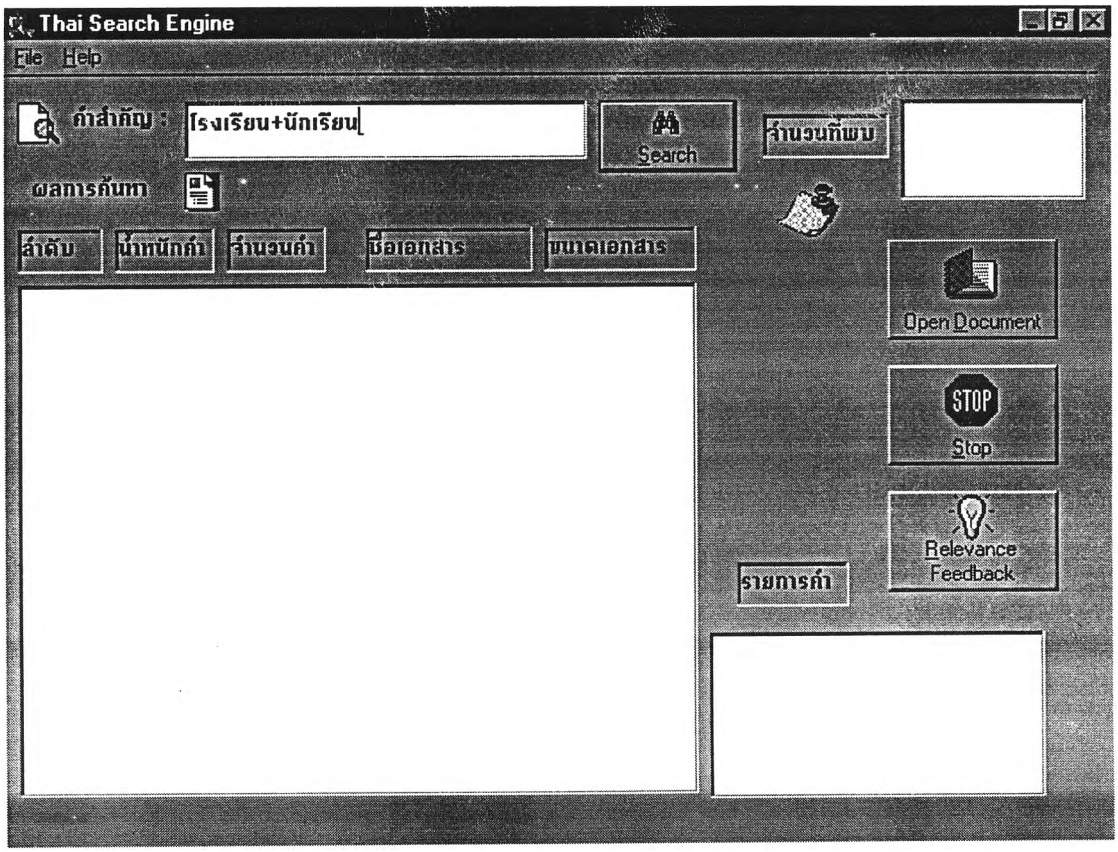
เป็นโปรแกรมเพื่อใช้สร้างดัชนีเพิ่มเติมข้อมูลเกี่ยวกับการหาน้ำหนักคำ ซึ่งในการวิจัยนี้ใช้การเก็บค่าตัวชี้ซิสตริงกับค่าความถี่ของซิสตริงไว้ด้วยกัน โดยมีการสร้างไว้ก่อนทำการค้นคืน เพื่อลดเวลาในการประมวลผลในช่วงการค้นคืนโดยให้ป้อนดัชนีเดิมที่ใช้ โดยในที่นี้ให้ป้อนชื่อแฟ้ม

แถวลำดับแพ็ตแบบสั้นเข้าสู่ระบบดังรูปที่ 4.1 ซึ่งในที่นี้แฟ้มแถวลำดับแพ็ตแบบสั้นซึ่งเป็นดัชนีเดิม ชื่อ "spat.020" เมื่อระบบได้ชื่อแฟ้มแถวลำดับแพ็ตแบบสั้นแล้ว ระบบจะนำข้อมูลในแถวลำดับแพ็ตแบบสั้นมาใช้สร้างดัชนีอันใหม่ โดยอ่านข้อมูลจากแถวลำดับแพ็ตทั้งหมด เพื่อนับจำนวนของทุกๆ ซิสตริงในแถวลำดับแพ็ต แล้วนำข้อมูลที่ได้มาเก็บไว้ในดัชนีใหม่ ซึ่งเก็บค่าความถี่ของแต่ละซิสตริงไว้ โดยใช้ชื่อแฟ้มดัชนีอันใหม่นี้ว่า "merge.wgh"

4.2 โปรแกรมค้นคืนเอกสาร

เป็นโปรแกรมที่พัฒนาขึ้นในการวิจัยนี้เพื่อใช้ในการค้นคืนแบบจัดลำดับ และการค้นคืนย้อนกลับ ประกอบด้วยส่วนประกอบที่สำคัญดังนี้

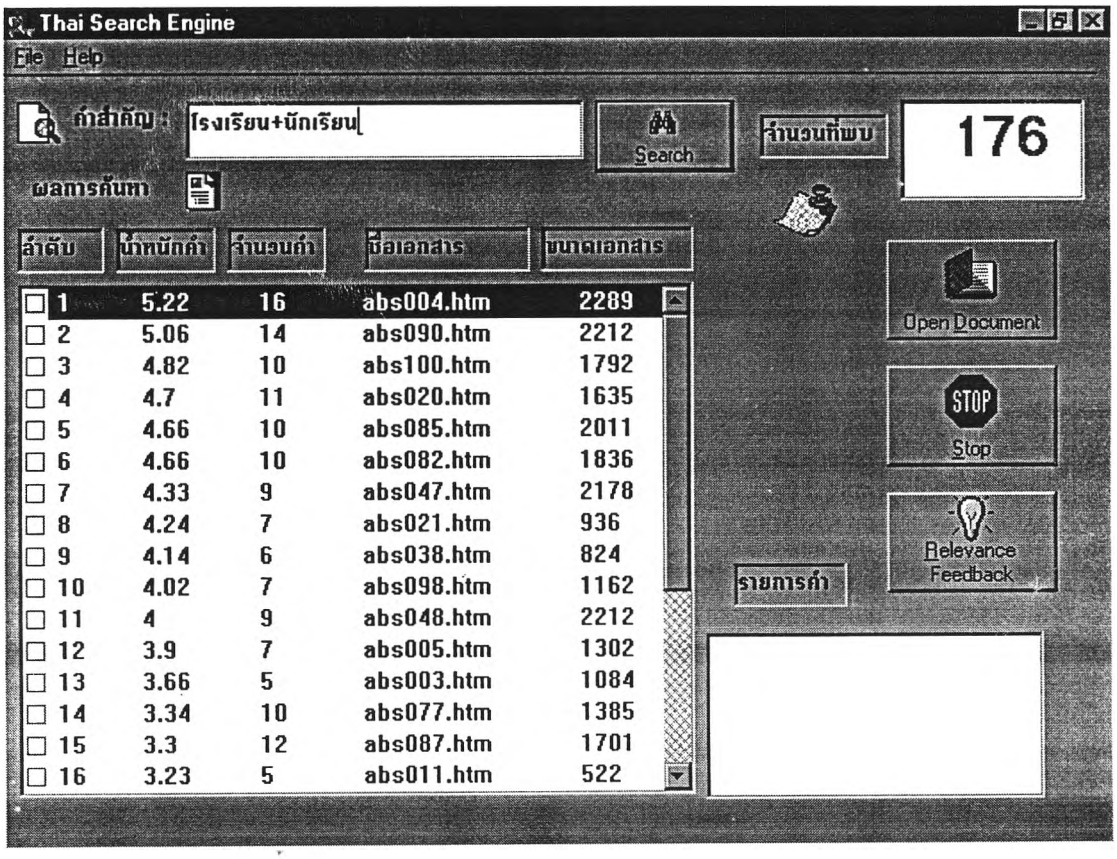
1. ส่วนรับคิวรีจากผู้ใช้ เป็นส่วนที่ให้ผู้ใช้ป้อนความต้องการข้อมูลของผู้ใช้ ในรูปแบบของคำสำคัญ หรือคิวรี ดังรูปที่ 4.2



รูปที่ 4.2 แสดงโปรแกรมส่วนรับคิวรี

โดยคิวรีที่ผู้ใช้ป้อน สามารถเลือกการค้นคืนในรูปแบบของบูลีน (Boolean) ได้โดยสามารถค้นคืนแบบตรรกะแบบ “และ” (AND) โดยใช้เครื่องหมาย “-” และตรรกะแบบ “หรือ” (OR) โดยใช้เครื่องหมาย “+” เชื่อมระหว่างคิวรีที่ใช้ค้นคืน เช่น ถ้าคิวรีเท่ากับคำว่า “โรงเรียน+นักเรียน” หมายถึงต้องการค้นคืนเอกสารที่มีคำว่า “โรงเรียน” หรือมีคำว่า “นักเรียน” ปรากฏอยู่ในเอกสารคำใดคำหนึ่งหรือทั้งสองคำ และถ้าคิวรีเท่ากับคำว่า “โรงเรียน-นักเรียน” หมายถึงต้องการค้นคืนเอกสารที่มีคำว่า“โรงเรียน” และมีคำว่า “นักเรียน” ปรากฏอยู่ในเอกสารทั้งสองคำ ซึ่งในการค้นคืนจะนำผลการค้นคืนที่ได้มาประมวลผลตามตรรกะนั้น (AND, OR) ภายหลังแล้วจึงแสดงผลลัพธ์สุดท้าย ตัวอย่างของการป้อนคิวรีในส่วนรับคิวรีของโปรแกรมดังรูปที่ 4.2 กำหนดให้คิวรีเท่ากับคำว่า “โรงเรียน+นักเรียน” ป้อนลงในช่องคำสำคัญ เมื่อต้องการเริ่มต้นค้นคืนให้กดปุ่มค้นหา (Search)

2. ส่วนแสดงผลการค้นคืนแบบจัดลำดับ เป็นส่วนสำคัญที่ใช้แสดงผลลัพธ์ของการค้นคืนให้แก่ผู้ใช้ เมื่อผู้ใช้ป้อนคิวรีให้แก่ระบบ



รูปที่ 4.3 แสดงโปรแกรมส่วนแสดงผลการค้นคืนแบบจัดลำดับ

จากรูปที่ 4.3 ตัวอย่างผลการค้นคืนคิวรีคำว่า "โรงเรียน+นักเรียน" โดยผู้ใช้สามารถเลือกเปิดเอกสารที่ค้นคืนมาได้โดยการเลือกชื่อแฟ้มที่ต้องการจากผลการค้นคืนแล้วกดปุ่มเปิดเอกสาร (Open Document) ระบบจะเปิดแฟ้มข้อมูลนั้นให้ผู้ใช้ดูข้อมูลในแฟ้มนั้น นอกจากนั้นเมื่อมีผลการค้นคืนมากเกิน การค้นคืนต้องใช้เวลานาน ผู้ใช้สามารถยกเลิกการค้นคืนด้วยการกดปุ่มหยุด (Stop) ระบบจะหยุดการค้นคืนแล้วกลับมารอรับคิวรีใหม่จากผู้ใช้ โดยสามารถเก็บผลการค้นคืนไว้ในแฟ้มชื่อ "output.log" โดยรูปแบบผลลัพธ์การค้นคืนประกอบด้วย

- ลำดับ เป็นลำดับเอกสารที่ค้นคืนแล้วมีคำที่ตรงกับคิวรี
- น้ำหนักคำ เป็นค่าน้ำหนักคำที่คำนวณได้จากสูตรน้ำหนักคำ ซึ่งเป็นค่าที่ใช้แสดงว่าเอกสารนั้น น่าจะตรงตามต้องการของผู้ใช้มากน้อยเพียงใด โดยถ้าเอกสารใดมีค่าน้ำหนักคำมาก จะแสดงว่าเอกสารนั้นน่าจะตรงตามต้องการของผู้ใช้มากด้วย
- จำนวนคำ เป็นค่าที่แสดงจำนวนที่พบคำที่ใช้ในคิวรี ในแต่ละรายการเอกสาร เช่น จำนวนคำเท่ากับ 3 แสดงว่าในเอกสารนั้น มีคำที่ใช้ในคิวรีปรากฏในเอกสาร 3 ครั้ง เป็นต้น
- ชื่อเอกสาร เป็นชื่อแฟ้มที่เก็บเอกสารที่มีคำตรงกับที่ใช้ในคิวรี ซึ่งชื่อเอกสารจะไม่มีซ้ำกัน
- ขนาดเอกสาร เป็นค่าที่แสดงขนาดของแฟ้มเอกสารของรายการนั้นๆ โดยมีหน่วยเป็นไบต์
- จำนวนคำรวม เป็นค่าที่แสดงจำนวนตำแหน่งทั้งหมดที่พบคำที่ใช้ในคิวรีภายในฐานข้อมูล

ตัวอย่างผลของการค้นคืนและการจัดลำดับของ 4 ตัวอย่างเอกสารดังนี้

เอกสารที่ 1 ชื่อ Doc001.htm ประกอบด้วยข้อความ "การค้นคืนสารสนเทศ"

เอกสารที่ 2 ชื่อ Doc002.htm ประกอบด้วยข้อความ "การจัดเก็บและการค้นคืนของสารสนเทศ"

เอกสารที่ 3 ชื่อ Doc003.htm ประกอบด้วยข้อความ "การจัดเก็บฐานข้อมูล"

เอกสารที่ 4 ชื่อ Doc004.htm ประกอบด้วยข้อความ "การค้นคืนสารสนเทศและฐานข้อมูล"

และกำหนดให้คิวรีทดสอบตัวอย่างคือคำว่า "สารสนเทศ+ฐานข้อมูล"

จากนั้นทดสอบค่าน้ำหนักคำตามสูตรวิธีคำนวณทั้งห้าได้ดังนี้

สูตรที่ 1 $w_{ij} = tf_{ij}$

สูตรที่ 2 $w_{ij} = tf_{ij} / n_i$

สูตรที่ 3 $w_{ij} = tf_{ij} * (\log(N/n_i) + 1)$

สูตรที่ 4 $$w_{ij} = \frac{[tf_{ij} * \log(N/n_i) * (K1 + 1)]}{[K1 * ((1-b) + (b * (filesize_j / avgfilesize))) + tf_{ij}]}$$

สูตรที่ 5 $$w_{ij} = \frac{\log(tf_{ij}) + 1}{0.7 + (0.3 * (filesize_i / avgfilesize))}$$

ผลการค้นคืน สำหรับสูตรน้ำหนักคำแบบที่ 1 สำหรับตัวอย่างเอกสารดังตารางที่ 4.1

ลำดับ	น้ำหนักคำ	จำนวนคำ	ชื่อเอกสาร	ขนาดเอกสาร
1	2	2	Doc004.htm	29
2	1	1	Doc001.htm	17
3	1	1	Doc002.htm	33
4	1	1	Doc003.htm	20

ตารางที่ 4.1 แสดงตัวอย่างผลการค้นคืนสำหรับสูตรน้ำหนักคำแบบที่ 1

ผลการค้นคืน สำหรับสูตรน้ำหนักคำแบบที่ 2 สำหรับตัวอย่างเอกสารดังตารางที่ 4.2

ลำดับ	น้ำหนักคำ	จำนวนคำ	ชื่อเอกสาร	ขนาดเอกสาร
1	0.83	2	Doc004.htm	29
2	0.5	1	Doc003.htm	20
3	0.33	1	Doc002.htm	33
4	0.33	1	Doc001.htm	17

ตารางที่ 4.2 แสดงตัวอย่างผลการค้นคืนสำหรับสูตรน้ำหนักคำแบบที่ 2

ผลการค้นคืน สำหรับสูตรน้ำหนักคำแบบที่ 3 สำหรับตัวอย่างเอกสารดังตารางที่ 4.3

ลำดับ	น้ำหนักคำ	จำนวนคำ	ชื่อเอกสาร	ขนาดเอกสาร
1	2.98	2	Doc004.htm	29
2	1.69	1	Doc003.htm	20
3	1.29	1	Doc002.htm	33
4	1.29	1	Doc001.htm	17

ตารางที่ 4.3 แสดงตัวอย่างผลการค้นคืนสำหรับสูตรน้ำหนักคำแบบที่ 3

ผลการค้นคืน สำหรับสูตรน้ำหนักคำแบบที่ 4 สำหรับตัวอย่างเอกสารดังตารางที่ 4.4

ลำดับ	น้ำหนักคำ	จำนวนคำ	ชื่อเอกสาร	ขนาดเอกสาร
1	0.91	2	Doc004.htm	29
2	0.77	1	Doc003.htm	20
3	0.34	1	Doc001.htm	17
4	0.25	1	Doc002.htm	33

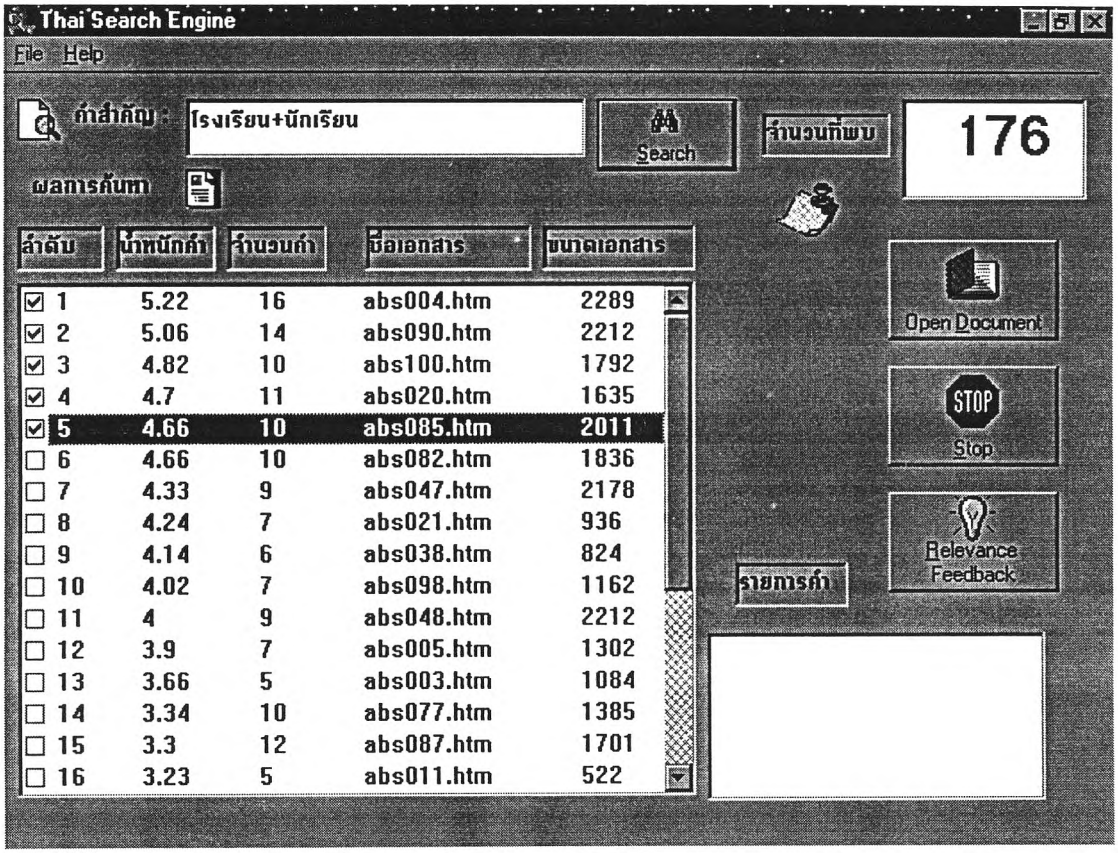
ตารางที่ 4.4 แสดงตัวอย่างผลการค้นคืนสำหรับสูตรน้ำหนักคำแบบที่ 4

ผลการค้นคืน สำหรับสูตรน้ำหนักคำแบบที่ 5 สำหรับตัวอย่างเอกสารดังตารางที่ 4.5

ลำดับ	น้ำหนักคำ	จำนวนคำ	ชื่อเอกสาร	ขนาดเอกสาร
1	1.91	2	Doc004.htm	29
2	1.11	1	Doc001.htm	17
3	1.06	1	Doc003.htm	20
4	0.91	1	Doc002.htm	33

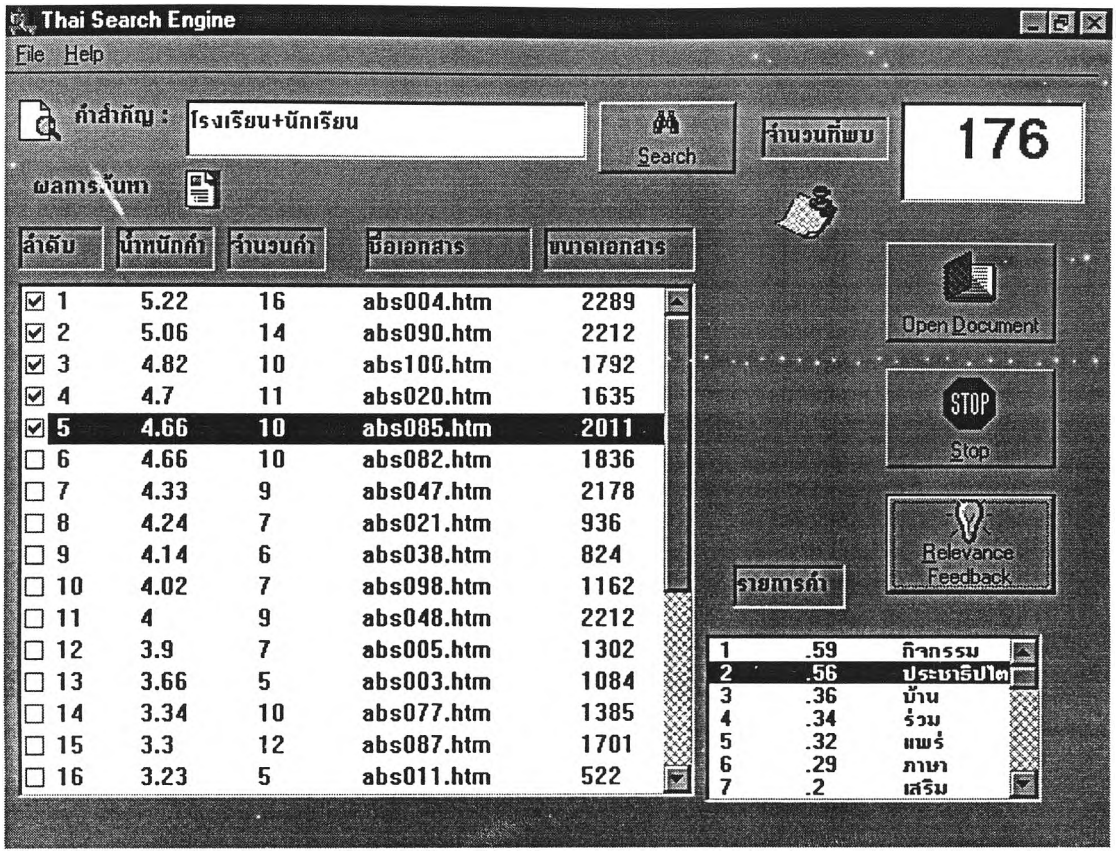
ตารางที่ 4.5 แสดงตัวอย่างผลการค้นคืนสำหรับสูตรน้ำหนักคำแบบที่ 5

3. ส่วนให้ผู้ใช้เลือกรายการเอกสารที่ตรงตามต้องการของผู้ใช้ โดยเมื่อผู้ใช้ได้รับผลการค้นคืนแล้ว ผู้ใช้ยังไม่พอใจผลการค้นคืนที่ได้ ต้องการทำการค้นคืนย้อนกลับ ซึ่งในที่นี้ผู้ใช้ต้องแสดงตัวอย่างเอกสารที่ตรงตามต้องการของผู้ใช้ โดยคลิกหน้ารายการเอกสารที่แสดงในรายการค้นคืน เมื่อเอกสารถูกเลือกจะมีเครื่องหมายถูก แสดงหน้ารายการเอกสารเหล่านั้น ดังรูปที่ 4.4 แสดงตัวอย่างการเลือกห้าเอกสารแรกว่าเป็นเอกสารที่ตรงตามต้องการ



รูปที่ 4.4 แสดงโปรแกรมส่วนเลือกเอกสารที่ตรงตามต้องการ

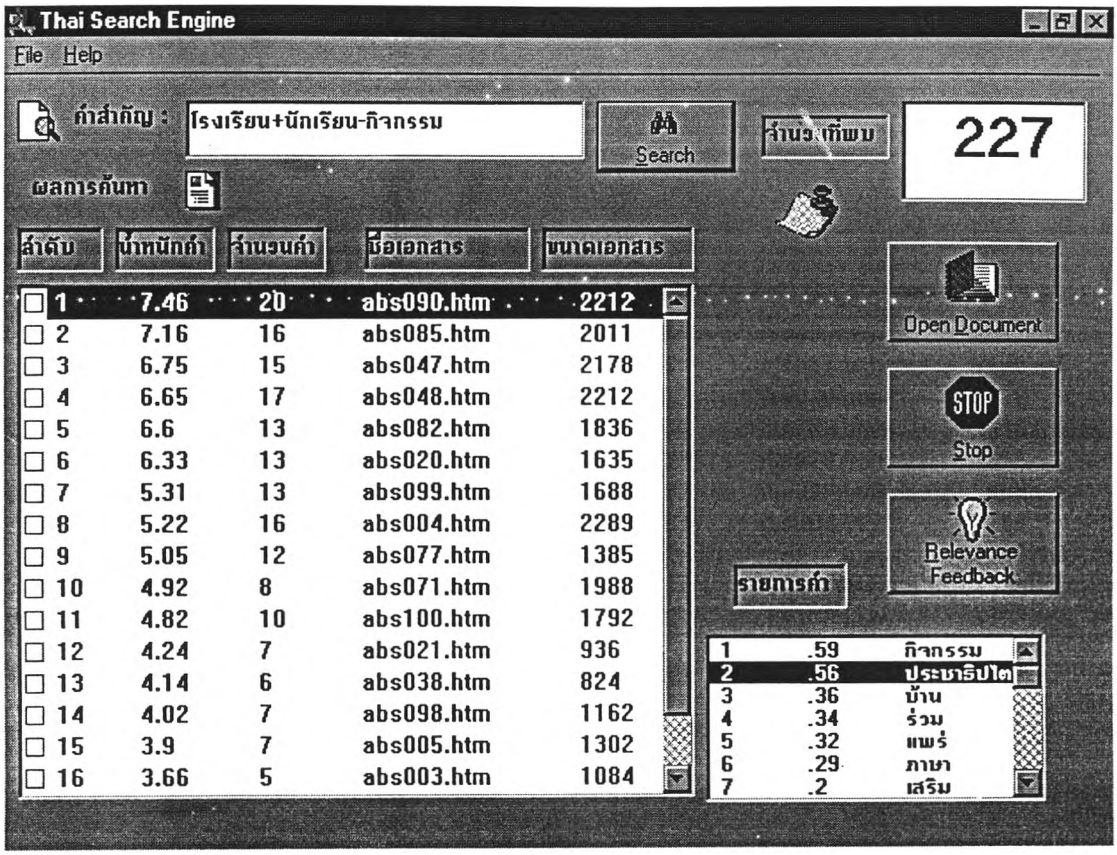
4. ส่วนแสดงคำใหม่ที่ได้จากการค้นคืนย้อนกลับ เมื่อผู้ใช้ได้เลือกเอกสารที่ตรงตามต้องการของผู้ใช้แล้วป้อนกลับสู่ระบบ โดยกดปุ่มค้นคืนย้อนกลับ ระบบจะแสดงคำใหม่ในรูปแบบของรายการคำ ซึ่งเป็นคำที่น่าจะเป็นประโยชน์ต่อผู้ใช้ในการนำไปปรับปรุงคิวรีเดิมของผู้ใช้ที่ยังให้ผลไม่เป็นที่น่าพอใจ ปรับปรุงให้เป็นคิวรีใหม่ที่ให้ผลที่ดีขึ้น ผลลัพธ์ที่ได้เป็นรายการคำใหม่ที่ระบบเสนอให้ผู้ใช้ ตรงตำแหน่งรายการคำ โดยการแสดงผลรายการคำที่ระบบเสนอให้อยู่ในรูปของ ลำดับที่ ค่าน้ำหนักคำและคำ ลำดับคำจะถูกจัดลำดับจากมากไปน้อย ตามค่าน้ำหนักคำที่คำนวณได้ซึ่งถ้าค่าน้ำหนักคำมากย่อมแสดงว่าคำนั้นน่าจะเป็นประโยชน์กับผู้ใช้มากด้วย



รูปที่ 4.5 แสดงโปรแกรมส่วนแสดงคำใหม่ที่ได้จากการค้นคืนย้อนกลับ

ตัวอย่างการค้นคืนย้อนกลับดังรูปที่ 4.5 คำว่า “กิจกรรม” เป็นรายการลำดับแรกดังนั้นน่าจะเป็นประโยชน์กับผู้ใช้มากที่สุดเนื่องจากมีค่าน้ำหนักคำสูงสุดคือ 0.59 ส่วนคำอื่นๆ ที่แสดงในรายการคำจะถูกเรียงตามค่าความสำคัญของคำตามลำดับ จากรายการคำที่ระบบเสนอให้แก่ผู้ใช้ เมื่อผู้ใช้ได้เลือกคำที่เกี่ยวข้องแล้ว ให้ผู้ใช้นำคำหรือกลุ่มคำที่เกี่ยวข้องเหล่านั้นไปใช้ปรับปรุงคิวรีเดิม โดยการเชื่อมคิวรีเดิมกับคำหรือกลุ่มคำใหม่ที่ผู้ใช้ได้เลือกมา ด้วยตรรกะแบบบูลีนในรูปแบบ “และ” “หรือ” ดังที่กล่าวไว้ข้างต้น

ตัวอย่างผลของการค้นคืนย้อนกลับเมื่อได้ปรับปรุงคิวรีแล้ว ดังรูปที่ 4.6 จากคิวรีเดิมคือคำว่า “โรงเรียน+นักเรียน” เมื่อผู้ใช้ทำการค้นคืนย้อนกลับแล้วได้จะรายการคำใหม่ โดยให้ผู้ใช้เลือกคำที่เกี่ยวข้อง ซึ่งในที่นี้เลือกคำว่า “กิจกรรม” ซึ่งเป็นคำที่มีค่าน้ำหนักคำสูงสุดเป็นคำที่เกี่ยวข้องนำมาใช้ปรับปรุงคิวรีเดิม ให้เป็นคิวรีใหม่คือคำว่า “โรงเรียน+นักเรียน-กิจกรรม” หมายถึงต้องการค้นคืนเอกสารที่มีคำว่า “โรงเรียน” หรือมีคำว่า “นักเรียน” และมีคำว่า “กิจกรรม” แล้วทำการค้นคืนใหม่อีกครั้ง



รูปที่ 4.6 แสดงโปรแกรมส่วนแสดงผลการค้นคืนใหม่ที่ได้จากการค้นคืนย้อนกลับ

4.3 การเปรียบเทียบกับระบบเดิม

การเปรียบเทียบโปรแกรมระบบค้นคืนบนโครงสร้างแพ็ตต้นฉบับบนระบบดอสกับโปรแกรมที่พัฒนาขึ้นใหม่บนระบบวินโดว์ โดยระบบใหม่ที่พัฒนาขึ้นสามารถสร้างผลลัพธ์การค้นคืนในรูปเรียงตามลำดับความสำคัญของเอกสารได้อย่างถูกต้องตามหลักภาษาศาสตร์และมีระบบค้นคืนย้อนกลับช่วยสร้างคิวรีใหม่แก่ผู้ใช้ ส่วนระบบเดิมสร้างผลลัพธ์การค้นคืนเรียงตามชื่อแฟ้มและไม่มีระบบช่วยสร้างคิวรีใหม่แก่ผู้ใช้