บทที่ 4

การพัฒนาระบบการค้นคืนข้อความภาษาไทยโดยใช้ต้นไม้แพ็ต

4.1 การพัฒนาระบบการค้นคืนข้อความภาษาไทยโดยใช้ต้นไม้แพ็ต

การพัฒนาระบบการค้นคืนข้อความภาษาไทยโดยใช้ต้นไม้แพ็ต จะแสดงพร้อมอัลกอริทึมการสร้างต้นไม้แพ็ต แพ็ตอะเรย์ การค้นคืนข้อความ ทั้งรูปภาพและตัวอย่างประกอบ เพื่อเพิ่มความเข้าใจมากยิ่งขึ้น

4.2 ระบบการค้นคืนข้อความภาษาไทย

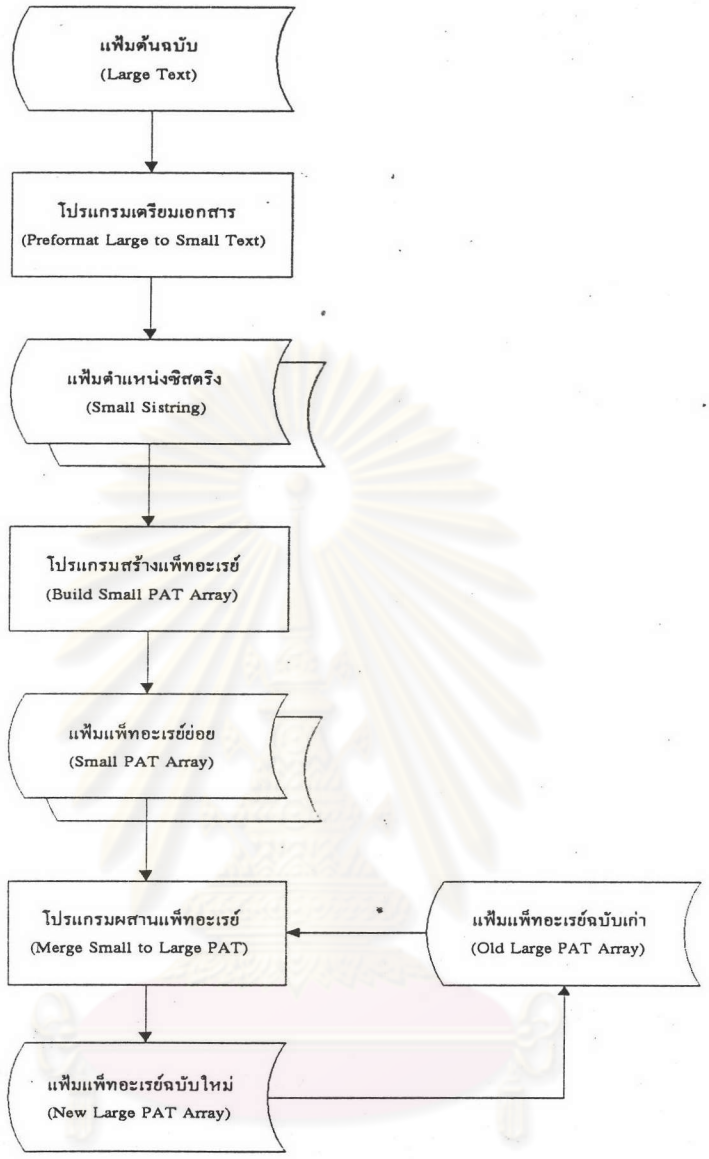
ระบบการค้นคืนข้อความภาษาไทยแบ่งออกเป็นสองส่วนหลักๆ คือ

1. การสร้างแพ็ตอะเรย์ของเอกสารเพื่อใช้ค้นคืนภายหลัง
2. การค้นคืนแบบเติมหน้ากับแบบบูล

4.3 การสร้างแพ็ตอะเรย์

การสร้างแพ็ตอะเรย์เป็นการสร้างดรรชนีให้กับเอกสารวิธีหนึ่งที่มีประสิทธิภาพ โดยแบ่งออกเป็น 3 ขั้นตอน ดังนี้

1. การเตรียมเอกสาร (Preformat Large to Small Text)
2. การสร้างแพ็ตอะเรย์ (Build Small PAT Array)
3. การผสานแพ็ตอะเรย์ (Merge Small to Large PAT Array)



รูปที่ 4.1 การสร้างแพ็ตอะเรย์

รูปที่ 4.1 แสดงขั้นตอนการสร้างแพ็ตอะเรย์ของแฟ้มต้นฉบับ (Large Text) ผ่านโปรแกรมตามขั้นตอนต่างๆ จนได้ผลลัพธ์สุดท้าย คือ แฟ้มแพ็ตอะเรย์ฉบับใหม่สุด (New Large PAT Array) โดยมีหลักการ 3 ขั้นตอน คือ

1. การตัดแบ่งแฟ้มต้นฉบับ (Large Text) ออกเป็น แฟ้มต้นฉบับย่อย (Small Text) หลายๆ ฉบับ พร้อมทั้งค้นหาตำแหน่งซิสตริงของแต่ละฉบับ และบันทึกเฉพาะ ตำแหน่งซิสตริงลงในแฟ้มตำแหน่งซิสตริง (Small Sistring) โดยอาศัยโปรแกรมเตรียมเอกสาร (Preformat Large to Small Text)

2. การสร้างแฟ้มแพ็ตอะเรย์ย่อย (Small PAT Array) ซึ่งประกอบด้วยแฟ้มแพ็ตอะเรย์ (PAT Array) กับแฟ้มแพ็ตอะเรย์ซ้ำ (Duplicate PAT Array) จากแต่ละแฟ้มตำแหน่งซิสตริง (Small Sistring) โดยอาศัยโปรแกรมสร้างแพ็ตอะเรย์ (Build Small PAT Array)

3. การผสานแต่ละแฟ้มแพ็ตอะเรย์ย่อย (Small PAT Array) เข้ากับแฟ้มแพ็ตอะเรย์ฉบับเก่า (Old Large PAT Array) เก็บลงแฟ้มแพ็ตอะเรย์ฉบับใหม่ (New Large PAT Array) โดยอาศัยโปรแกรมผสานแพ็ตอะเรย์ (Merge Small to Large PAT Array)

หลังจากสร้างแพ็ตอะเรย์เสร็จแล้ว นำแฟ้มแพ็ตอะเรย์ฉบับใหม่ไปใช้ในการค้นคืนข้อความต่อไป

4.3.1 <u>การเตรียมเอกสาร</u>

ดังที่ได้กล่าวไว้แล้วเกี่ยวกับลักษณะและปริมาณของเอกสาร พบว่า ส่วนใหญ่แล้วเอกสารที่นำมาจัดเก็บเพื่อค้นคืนภายหลัง มักจะมีปริมาณมากเกินกว่าที่เครื่องคอมพิวเตอร์จะสามารถประมวลผลได้เสร็จสิ้นภายในหนเดียวได้ โดยจะทยอยสร้างแฟ้มข้อมูลเสริมเพื่อการค้นคืนไปเรื่อยๆ ดังนั้นขั้นตอนการเตรียมเอกสารจะจัดแบ่งเอกสารฉบับใหญ่ๆ หรือที่เรียกว่าแฟ้มต้นฉบับออกเป็นฉบับย่อยๆ เช่น 64,000 หรือ 100,000 ไบต์เป็นต้น ซึ่งมีความเหมาะสมกับการประมวลผล ทั้งนี้ขึ้นกับจำนวนหน่วยความจำหลัก ของเครื่องคอมพิวเตอร์นั้นๆ โดยกำหนดเป็นพารามิเตอร์ (Parameter) ตัวหนึ่งของระบบ จากนั้นจะคำนวณหาตำแหน่งซิสตริงของแฟ้มต้นฉบับย่อยแต่ละฉบับ แล้วจัดเก็บเฉพาะ ตำแหน่งซิสตริง ลงแฟ้มตำแหน่งซิสตริง เพื่อใช้ในการ สร้างแพ็ตอะเรย์ต่อไป โดยวิธีการคำนวณอาศัยหลักทางภาษาศาสตร์เบื้องต้นประกอบ คือ

1. เป็นซิสตริงภาษาอังกฤษที่ขึ้นต้นด้วย a-z, A-Z และ 0-9
2. เป็นซิสตริงภาษาไทยที่ขึ้นต้นด้วยพยัญชนะ ก-ฮ, สระ เอ, แอ, โอ, ใอ, ไอ และ ๐-๙

อย่างไรก็ตามสำหรับกฎทางภาษาศาสตร์สามารถปรับปรุงให้มีความถูกต้องและเหมาะสมมากขึ้นในภายหลังได้

4.3.1.1 ขั้นตอนการเตรียมเอกสาร

1. โหลดข้อมูลขนาดเท่ากับพารามิเตอร์กำหนดขนาดแฟ้มต้นฉบับย่อย จากแฟ้มต้นฉบับเก็บไว้ในหน่วยความจำ

2. ตรวจสอบข้อมูลทีละตำแหน่ง (ตัวอักษร) ว่าเป็นซิสตริงที่ถูกต้องตามกฎทางหลักภาษาศาสตร์เบื้องต้นหรือไม่ ถ้าถูกต้องบันทึกเฉพาะ ตำแหน่งซิสตริง ลงแฟ้มตำแหน่งซิสตริง จนกระทั่งหมดข้อมูลตรวจสอบ

3. ปฏิบัติตามขั้นตอนที่ 1 กับ 2 จนหมดแฟ้มต้นฉบับ ทำให้ได้แฟ้มตำแหน่งซิสตริงหลายๆ แฟ้มข้อมูล เพื่อนำไปสร้างต้นไม้แพ็ตในขั้นตอนต่อไป

4.3.1.2 ข้อกำหนดเบื้องต้นของการเตรียมเอกสาร

1. พารามิเตอร์กำหนดชื่อแฟ้มต้นฉบับ จะต้องมีนามสกุลเป็น เอช ที เอ็ม (HTM) เสมอ
2. ต้องระบุ พารามิเตอร์กำหนดขนาดแฟ้มต้นฉบับย่อย
3. ต้องระบุ พารามิเตอร์กำหนดความยาวซิสตริง
4. มีทางเลือกพิมพ์ข้อมูลทางสถิติหรือซิสตริงเพื่อตรวจสอบก่อนการสร้างแพ็ตอะเรย์ ซึ่ง
   แฟ้มข้อมูล <input>.Brr เก็บข้อมูลทางสถิติ โดย
      <input> คือ ชื่อแฟ้มต้นฉบับ
      B คือ ตัวระบุว่าเป็นแฟ้มข้อมูลประเภทเก็บข้อมูลสถิติ
      rr คือ ตัวเลขต่อเนื่องกัน (Running Number)
   แฟ้มข้อมูล <input>.Drr เก็บซิสตริง โดย
      D คือ ตัวระบุว่าเป็นแฟ้มข้อมูลประเภทเก็บซิสตริง
5. แฟ้มตำแหน่งซิสตริงที่ถูกตัดแบ่งจากแฟ้มต้นฉบับ คือ <input>.Irr โดย
   I คือ ตัวระบุว่าเป็นแฟ้มข้อมูลประเภทเก็บตำแหน่งซิสตริง

4.3.1.3 โครงสร้างแฟ้มตำแหน่งซิสตริง

ประกอบด้วย 2 ส่วน ดังนี้

1. ส่วนหัว (Header) มีองค์ประกอบหลักของข้อมูล คือ

ชื่อข้อมูล	รายละเอียดของข้อมูล
1. Large Text	พารามิเตอร์กำหนดชื่อแฟ้มต้นฉบับ
2. Size Small Text	พารามิเตอร์กำหนดขนาดแฟ้มต้นฉบับย่อย
3. Base Large Text	ตำแหน่งเริ่มต้นแฟ้มต้นฉบับ
4. Length Small Text	ความยาวแฟ้มต้นฉบับย่อย
5. Number of Sistrings	จำนวนซิสตริงทั้งหมด
6. Length Sistring	พารามิเตอร์กำหนดความยาวซิสตริง
7. Number of Small Text	จำนวนแฟ้มต้นฉบับย่อย
8. Flag Last Small Text	ตัวระบุแฟ้มต้นฉบับย่อยเป็นฉบับสุดท้าย

ตารางที่ 4.1 ส่วนหัวแฟ้มตำแหน่งซิสตริง

2. ส่วนข้อมูล (Detail) เก็บตำแหน่งซิสตริงที่ใช้ในการอ้างอิงถึงซิสตริงในแฟ้มต้นฉบับ

สำหรับขั้นตอนการเตรียมเอกสาร อาศัยหน่วยความจำขนาดเท่ากับพารามิเตอร์กำหนดขนาดแฟ้มต้นฉบับย่อย และ ประสิทธิภาพของโปรแกรม ขึ้นกับปัจจัย ดังต่อไปนี้

1. เวลาทั้งหมดในการอ่านแฟ้มต้นฉบับจากหน่วยความจำสำรองเก็บในหน่วยความจำ
2. เวลาทั้งหมดในการตรวจสอบข้อมูลทีละตัวอักษร แต่เนื่องจากประมวลผลในหน่วยความจำ จึงใกล้เคียงศูนย์
3. เวลาทั้งหมดในการบันทึกแฟ้มตำแหน่งซิสตริงทั้งหมด

4.3.2 การสร้างแพ็ตอะเรย์

การสร้างแพ็ตอะเรย์ คือ การจัดเรียงลำดับซิสตริงแบบหนึ่ง โดยนำ ตำแหน่งซิสตริง ที่ใช้ในการอ้างอิงถึงซิสตริงในแฟ้มต้นฉบับ ของแฟ้มตำแหน่งซิสตริง ตามที่ได้แบ่งไว้ในขั้นตอน การเตรียมเอกสาร มาสร้างต้นไม้แพ็ต เพื่อหาแพ็ตอะเรย์ ซึ่งแพ็ตอะเรย์ที่ได้จากต้นไม้แพ็ต คือ ตำแหน่งซิสตริงที่ถูกเรียงลำดับตามรหัสข้อมูลของซิสตริง แล้วจัดเก็บแพ็ตอะเรย์ลงแฟ้มแพ็ตอะเรย์ย่อย โดยในกรณีของการสร้างแพ็ตอะเรย์ที่ไม่เกิดการซ้ำกันของซิสตริง จะจัดเก็บ

แพ็ตอะเรย์ลงในส่วนของแฟ้มแพ็ตอะเรย์ ส่วนในกรณีเกิดการซ้ำกันของซิสตริง จะจัดเก็บแพ็ตอะเรย์ซ้ำลงในแฟ้มแพ็ตอะเรย์ซ้ำ และจัดเก็บแพ็ตอะเรย์กับตำแหน่งแพ็ตอะเรย์ซ้ำของแฟ้มแพ็ตอะเรย์ซ้ำ ลงในแฟ้มแพ็ตอะเรย์ด้วย

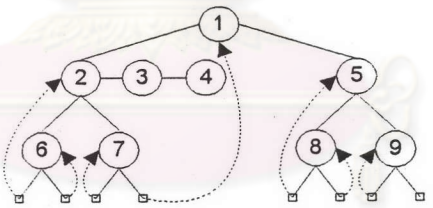
4.3.2.1 ขั้นตอนการสร้างแพ็ตอะเรย์

1. โหลดแฟ้มต้นฉบับย่อยตามข้อมูลในส่วนหัวของแฟ้มตำแหน่งซิสตริงฉบับที่ต้องการสร้างแพ็ตอะเรย์เก็บในหน่วยความจำ และในกรณีเป็นฉบับสุดท้ายจะเพิ่มช่องว่างไปเป็นจำนวนความยาวของซิสตริงหักออกหนึ่ง

2. จองหน่วยความจำเท่ากับจำนวนซิสตริงทั้งหมดในส่วนหัวของแฟ้มตำแหน่งซิสตริงบวกหนึ่งคูณกับขนาดของโหนด (รวมโหนดรากด้วย)

3. โหลดตำแหน่งของซิสตริงทั้งหมด ในแฟ้มตำแหน่งซิสตริง เก็บลงแต่ละโหนดในหน่วยความจำที่จองไว้

4. สร้างต้นไม้แพ็ต ถ้าเกิดซิสตริงซ้ำกันให้เก็บเรียงต่อๆ กันไป เช่น โหนดที่ 2, 3 และ 4 ซ้ำกัน ดังรูปที่ 4.2



รูปที่ 4.2 ต้นไม้แพ็ตที่มีซิสตริงทั้งหมด 9 ซิสตริงและซิสตริงซ้ำๆ กัน 3 ซิสตริง

5. ใช้วิธีการท่องไปในต้นไม้ เพื่อหาแพ็ตอะเรย์ ซึ่งคือ โหนดภายนอกที่มีดรรชนีชี้ขึ้นจากทางซ้ายไปทางขวา แล้วจัดเก็บลงแฟ้มแพ็ตอะเรย์ และถ้าโหนดใดมีซิสตริงที่ซ้ำๆ กันให้จัดเก็บลงแฟ้มแพ็ตอะเรย์ซ้ำ

4.3.2.2 ข้อกำหนดเบื้องต้นของการสร้างแพ็ตอะเรย์

1. ต้องระบุชื่อแฟ้มตำแหน่งซิสตริง

2. ต้องมีหน่วยความจำไม่น้อยกว่า ขนาดของโหนดหนึ่งโหนด คูณกับ จำนวนซิสตริงทั้งหมด รวมกับ ขนาดของแฟ้มต้นฉบับย่อย ที่กำหนดไว้ในขั้นตอนการเตรียมเอกสาร เนื่องจากต้นไม้แพ็ตทั้งต้นกับแฟ้มต้นฉบับย่อยต้องอยู่ในหน่วยความจำ โดยมีโครงสร้างข้อมูล ดังนี้

```
typedef struct {
        short skipb;
        long left, right, displce, duplicate;
} KEYBUF;
```

skipb	left	right	displace	duplicate

รูปที่ 4.3 โครงสร้างโหนดต้นไม้แพ็ต

โดยที่ skipb คือ บิตข้าม (Skip Bit)
left คือ ดรรชนีชี้ทางซ้าย (Left Pointer)
right คือ ดรรชนีชี้ทางขวา (Right Pointer)
displace คือ ตำแหน่งซิสตริง (Position of Sistring)
duplicate คือ ตำแหน่งซิสตริงซ้ำ (Duplicate Position of Sistring)

<u>หมายเหตุ</u> อาจลดค่า skipb ลงได้อีก 1 ไบต์ (ขนาดซิสตริงยาวได้ไม่เกิน 256 ไบต์) เพื่อประหยัดหน่วยความจำในการสร้างต้นไม้แพ็ต

3. มีทางเลือก ให้เลือก ดังนี้

3.1 สามารถสร้างแพ็ตอะเรย์ เก็บลงแฟ้มข้อมูล <input>.Arr โดย A คือ ตัวระบุว่าเป็นแฟ้มข้อมูลประเภทเก็บแพ็ตอะเรย์

3.2 สามารถสร้างข้อมูลสถิติ เก็บลงแฟ้มข้อมูล <input>.Crr โดย C คือ ตัวระบุว่าเป็นแฟ้มข้อมูลประเภทเก็บข้อมูลสถิติ

3.3 สามารถแสดงโหนดต้นไม้แพ็ตทั้งหมด เก็บลงแฟ้มข้อมูล <input>.Lrr โดย L คือ ตัวระบุว่าเป็นแฟ้มข้อมูลประเภทแสดงโหนดต้นไม้แพ็ตทั้งหมด

3.4 สามารถแสดงรูปภาพต้นไม้แพ็ต เก็บลงแฟ้มข้อมูล <input>.Mrr โดย M คือ ตัวระบุว่าเป็นแฟ้มข้อมูลประเภทแสดงรูปภาพต้นไม้แพ็ต

3.5 สามารถเก็บต้นไม้แพ็ต เก็บลงแฟ้มข้อมูล <input>.Trr โดย T คือ ตัวระบุว่าเป็นแฟ้มข้อมูลประเภทเก็บต้นไม้แพ็ต

3.6 สามารถแสดงซิสตริงที่เรียงลำดับแล้ว เก็บลงแฟ้มข้อมูล <input>.Zrr โดย Z คือ ตัวระบุว่าเป็นแฟ้มข้อมูลประเภทแสดงซิสตริงที่เรียงลำดับแล้ว

4. แฟ้มแพ็ตอะเรย์ย่อย คือ <input>.Prr โดย P คือ ตัวระบุว่าเป็นแฟ้มข้อมูลประเภทเก็บแพ็ตอะเรย์

5. แฟ้มแพ็ตอะเรย์ซ้ำ คือ <input>.Srr โดย S คือ ตัวระบุว่าเป็นแฟ้มข้อมูลประเภทเก็บแพ็ตอะเรย์ซ้ำ

4.3.2.3 <u>โครงสร้างแฟ้มแพ็ตอะเรย์ย่อย</u>

ประกอบด้วย 2 ส่วน ดังนี้

1. ส่วนหัว (Header) มีองค์ประกอบหลักของข้อมูล คือ

ชื่อข้อมูล	รายละเอียดของข้อมูล
1. Large Text	ชื่อแฟ้มต้นฉบับ
2. Position of Sistring	ชื่อแฟ้มตำแหน่งซิสตริง
3. Duplicate PAT Array	ชื่อแฟ้มแพ็ตอะเรย์ซ้ำ
4. Size Small Text	ขนาดแฟ้มต้นฉบับย่อย
5. Base Large Text	ตำแหน่งเริ่มต้นแฟ้มต้นฉบับ
6. Length Small Text	ความยาวแฟ้มต้นฉบับย่อย
7. Number of Sistrings	จำนวนซิสตริงทั้งหมด
8. Number of PAT Array	จำนวนแพ็ตอะเรย์ทั้งหมด
9. Length Sistring	ความยาวซิสตริง
10. Flag Last Small Text	ตัวระบุแฟ้มต้นฉบับย่อยเป็นฉบับสุดท้าย

ตารางที่ 4.2 ส่วนหัวแฟ้มแพ็ตอะเรย์

2. ส่วนข้อมูล (Detail) เก็บแพ็ตอะเรย์ คือ ตำแหน่งซิสตริงที่ใช้ในการอ้างอิงถึงซิสตริงในแฟ้มต้นฉบับที่เรียงลำดับแล้ว กับ ตำแหน่งแพ็ตอะเรย์ซ้ำของแฟ้มแพ็ตอะเรย์ซ้ำ เพื่อใช้เป็นดรรชนีในการอ้างถึงซิสตริงที่ซ้ำกันในแฟ้มต้นฉบับอีกทอดหนึ่ง ดังรูปที่ 4.4

```
typedef struct {
        long dpat;
        long ddup;
} PAT_ARRAY;
```

รูปที่ 4.4 โครงสร้างแพ็ตอะเรย์

โดยที่ dpat คือ แพ็ตอะเรย์
ddup คือ ตำแหน่งในแฟ้มแพ็ตอะเรย์ซ้ำ

4.3.2.4 โครงสร้างแฟ้มแพ็ตอะเรย์ซ้ำ

ประกอบด้วย 2 ส่วน ดังนี้

1. ส่วนหัว (Header) เหมือนแฟ้มแพ็ตอะเรย์
2. ส่วนข้อมูล (Detail) เก็บแพ็ตอะเรย์ที่ซ้ำๆ กันเรียงต่อๆ กัน โดย มีค่าศูนย์เป็นตัวบ่งบอกจุดสิ้นสุดของแพ็ตอะเรย์ที่ซ้ำๆ กัน

สำหรับขั้นตอนการสร้างแพ็ตอะเรย์ต้องอาศัยหน่วยความจำเพื่อเก็บต้นไม้แพ็ต เพื่อสร้างแพ็ตอะเรย์ โดยประสิทธิภาพของโปรแกรมขึ้นกับปัจจัย ดังต่อไปนี้

1. เวลาทั้งหมดในการอ่านแฟ้มตำแหน่งซิสตริงที่นำมาสร้างแพ็ตอะเรย์เก็บในหน่วยความจำ

2. เวลาทั้งหมดในการสร้างต้นไม้แพ็ต แต่เนื่องจากประมวลผลในหน่วยความจำจึงใกล้เคียงศูนย์

3. เวลาทั้งหมดในการท่องไปในต้นไม้เพื่อหาแพ็ตอะเรย์ แต่เนื่องจากประมวลผลในหน่วยความจำจึงใกล้เคียงศูนย์

4. เวลาทั้งหมดในการบันทึกแฟ้มแพ็ตอะเรย์ และ แฟ้มแพ็ตอะเรย์ซ้ำ

4.3.3 การผสานแพ็ตอะเรย์

หลังจากทำการสร้างแฟ้มแพ็ตอะเรย์ย่อยของแต่ละแฟ้มตำแหน่งซิสตริง เสร็จแล้ว นำแพ็ตอะเรย์เหล่านั้นมา ผสานเข้าด้วยกัน เกิดเป็นแพ็ตอะเรย์ฉบับใหม่ โดยทำการผสานทีละแพ็ตอะเรย์ย่อย เข้ากับแพ็ตอะเรย์ฉบับเก่า โดยอาศัยหน่วยความจำที่มีอยู่อย่างจำกัดให้เกิดประสิทธิภาพมากที่สุด ซึ่งมีหลักการ คือ เก็บแพ็ตอะเรย์ย่อยทั้งหมดในหน่วยความจำ แล้วนำซิสตริงของแพ็ตอะเรย์ฉบับเก่า ไปค้นหาตำแหน่งที่จะแทรกเข้าไป ต่อกับซิสตริงของแพ็ตอะเรย์ย่อย เพื่อได้ซิสตริงของแพ็ตอะเรย์ฉบับเก่าผสานกับแพ็ตอะเรย์ย่อย เก็บต่อเนื่องกันไปได้อย่างถูกต้อง แล้วจึงจัดเก็บลงแฟ้มแพ็ตอะเรย์ฉบับใหม่ ส่วนที่เป็นซิสตริงซ้ำๆ กันจัดเก็บลงแฟ้มแพ็ตอะเรย์ซ้ำฉบับใหม่ และเก็บตารางการผสานแพ็ตอะเรย์ลงแฟ้มตารางการผสานแพ็ตอะเรย์ด้วย เพื่อใช้ในการอ้างอิงถึงซิสตริงในแฟ้มต้นฉบับได้อย่างถูกต้อง

4.3.3.1 ขั้นตอนการผสานแพ็ตอะเรย์

1. กรณียังไม่มีแพ็ตอะเรย์ฉบับเก่า ให้สร้างจากแพ็ตอะเรย์ย่อยที่ต้องการผสาน พร้อมทั้งสร้างตารางการผสานแพ็ตอะเรย์ด้วย

2. โหลดแพ็ตอะเรย์ย่อย กับ แฟ้มต้นฉบับย่อย ที่ต้องการผสานเก็บในหน่วยความจำ

3. จองหน่วยความจำตัวนับค่า เพื่อเก็บจำนวนซิสตริงของแพ็ตอะเรย์ฉบับเก่า ขนาดเท่ากับแพ็ตอะเรย์ย่อยบวกหนึ่ง

4. โหลดตารางการผสานแพ็ตอะเรย์เก็บในหน่วยความจำ

5. โหลดแฟ้มต้นฉบับย่อยของแพ็ตอะเรย์ฉบับเก่าชิ้นแรกตามตารางการผสานแพ็ตอะเรย์เก็บในหน่วยความจำ

6. อ่านซิสตริงของแฟ้มตำแหน่งซิสตริงของแฟ้มต้นฉบับย่อยของแพ็ตอะเรย์ฉบับเก่าตามตารางการผสานแพ็ตอะเรย์ แล้วนำไปค้นหาแบบทวิภาคกับแพ็ตอะเรย์ย่อยที่ต้องการผสานใน

หน่วยความจำ ถ้า ***ผลเท่ากัน*** ให้เพิ่มค่าตัวนับค่าที่เก็บ ***ค่าเท่ากัน*** ของ ***ดรรชนีตัวที่ค้นพบขึ้นอีกหนึ่ง*** แต่ถ้า ***ผลน้อยกว่า*** ให้เพิ่มค่าตัวนับค่าที่เก็บ ***ค่าน้อยกว่า*** ของ ***ดรรชนีตัวที่ค้นไม่พบตัวสุดท้ายขึ้นอีกหนึ่ง*** และถ้า ***ผลมากกว่า*** ให้เพิ่มค่าตัวนับค่าที่เก็บ ***ค่าน้อยกว่า*** ของ ***ดรรชนีตัวที่ค้นไม่พบตัวสุดท้ายถัดไปขึ้นอีกหนึ่งแทน***

7. ปฏิบัติจนครบจำนวนซิสตริงทั้งหมดของแฟ้มตำแหน่งซิสตริง แล้วปฏิบัติกับแฟ้มตำแหน่งซิสตริงถัดไปตามตารางการผสานแพ็ตอะเรย์จนครบหมด ซึ่งมีผลเท่ากับปฏิบัติจนครบแพ็ตอะเรย์ฉบับเก่า

8. ผสานแพ็ตอะเรย์ย่อยเข้ากับแพ็ตอะเรย์ฉบับเก่า โดยอาศัยค่าที่เก็บในตัวนับค่าเป็นดรรชนีชี้นำ เช่น ค่าในตัวนับค่าชุดแรกเป็น 3, 1 แสดงว่าให้อ่านค่าในแฟ้มข้อมูลที่เก็บค่าแพ็ตอะเรย์ฉบับเก่ามา 3 ค่า แล้วจึงต่อด้วยค่าของแพ็ตอะเรย์ย่อย 1 ค่า โดยมีค่าของแพ็ตอะเรย์ฉบับเก่าที่ซ้ำกับค่าของแพ็ตอะเรย์นี้อีก 1 ค่า จากนั้นปฏิบัติกับตัวนับค่าชุดถัดไปจนหมด ซึ่งผลที่ได้ คือ แฟ้มข้อมูลแพ็ตอะเรย์ฉบับใหม่

9. เพิ่มตารางการผสานแพ็ตอะเรย์ขึ้นอีกหนึ่งชุด โดยนำข้อมูลจากแพ็ตอะเรย์ย่อยที่ต้องการผสานเข้าไปเพิ่มต่อท้าย

<u>หมายเหตุ</u> การผสานแพ็ตอะเรย์ อาจเกิดกรณีการซ้ำกันของซิสตริงได้

4.3.3.2 <u>ข้อกำหนดเบื้องต้นของการผสานแพ็ตอะเรย์</u>

1. ต้องระบุชื่อแฟ้มแพ็ตอะเรย์ย่อยที่ต้องการผสาน

2. ต้องมีหน่วยความจำเพียงพอที่จะเก็บ แพ็ตอะเรย์ย่อยทั้งหมด แฟ้มต้นฉบับย่อย ตารางการผสานแพ็ตอะเรย์ และ อะเรย์ตัวนับค่า ซึ่งมีโครงสร้าง ดังต่อไปนี้

```
typedef struct {
        long less;
        long equal;
} COUNT_ARRAY;
```

less	equal

รูปที่ 4.5 โครงสร้างตัวนับค่าของแพ็ตอะเรย์ย่อย

โดยที่ less คือ ส่วนที่เก็บ ***ค่าน้อยกว่า*** ในอะเรย์ตัวนับค่า

equal คือ ส่วนที่เก็บ ***ค่าเท่ากัน*** ในอะเรย์ตัวนับค่า

3. มีทางเลือก ให้เลือก ดังนี้

3.1 สามารถสร้างข้อมูลสถิติเก็บลงแฟ้มข้อมูล Merge.sta

3.2 สามารถแสดงซิสตริงที่เรียงลำดับแล้วเก็บลงแฟ้มข้อมูล Merge.sis

4. แฟ้มแพ็ตอะเรย์ฉบับใหม่ คือ Merge.pat

5. แฟ้มแพ็ตอะเรย์ซ้ำฉบับใหม่ คือ Merge.dup

6. แฟ้มตารางการผสานแพ็ตอะเรย์ คือ Merge.dir โดยมีโครงสร้าง ดังนี้

```
typedef struct {
        long mdisp;
        char fname[16];
} MERGE_DIR;
```

mdisp	fname

รูปที่ 4.6 โครงสร้างตารางการผสานแพ็ตอะเรย์

โดยที่ mdisp คือ ตำแหน่งสะสมของแฟ้มต้นฉบับย่อย

fname คือ ชื่อแฟ้มตำแหน่งซิสตริง

4.3.3.3 <u>โครงสร้างแฟ้มแพ็ตอะเรย์ฉบับใหม่</u>

ประกอบด้วย 2 ส่วน ดังนี้

1. ส่วนหัว (Header) มีองค์ประกอบหลักของข้อมูล คือ

ชื่อข้อมูล	รายละเอียดของข้อมูล
1. Duplicate Large PAT Array	ชื่อแฟ้มแพ็ตอะเรย์ซ้ำฉบับใหม่
2. Table of Merge PAT Array	ชื่อแฟ้มตารางการผสานแพ็ตอะเรย์
3. Number of PAT Arrays	จำนวนแพ็ตอะเรย์สะสมทั้งหมด
4. Number of Duplicate PAT Arrays	จำนวนแพ็ตอะเรย์ซ้ำสะสมทั้งหมด
5. Number of Duplicate Merge PAT Arrays	จำนวนแพ็ตอะเรย์ซ้ำในตอนผสาน
6. Length Sistring	ความยาวซิสตริง
7. Number of Small PAT Array	จำนวนชิ้นที่ทำการผสานแพ็ตอะเรย์

ตารางที่ 4.3 ส่วนหัวแฟ้มแพ็ตอะเรย์ฉบับใหม่

2. ส่วนข้อมูล (Detail) เก็บค่าแพ็ตอะเรย์ใหม่ คือ ตำแหน่งซิสตริงที่เรียงลำดับแล้ว บวกกับความยาวแฟ้มต้นฉบับย่อยแบบสะสมในตารางการผสานแพ็ตอะเรย์ กับค่าตำแหน่งแพ็ตอะเรย์ซ้ำใหม่ ของแฟ้มแพ็ตอะเรย์ซ้ำฉบับใหม่

4.3.3.4 <u>โครงสร้างแฟ้มแพ็ตอะเรย์ซ้ำฉบับใหม่</u>

ประกอบด้วย 2 ส่วน ดังนี้

1. ส่วนหัว (Header) เหมือนแฟ้มแพ็ตอะเรย์ฉบับใหม่

2. ส่วนข้อมูล (Detail) เก็บค่าแพ็ตอะเรย์ใหม่ที่ซ้ำๆ กันเรียงต่อๆ กัน โดย มีค่าศูนย์เป็นตัวบ่งบอกจุดสิ้นสุดของแพ็ตอะเรย์ที่ซ้ำๆ กัน

4.3.3.5 <u>โครงสร้างแฟ้มตารางการผสานแพ็ตอะเรย์</u>

ชื่อข้อมูล	รายละเอียดของข้อมูล
1. Accumulate Length Small Text	ความยาวแฟ้มต้นฉบับย่อยแบบสะสม
2. Position of Sistring	ชื่อแฟ้มตำแหน่งซิสตริงย่อย

ตารางที่ 4.4 โครงสร้างแฟ้มตารางการผสานแพ็ตอะเรย์

สำหรับขั้นตอนการผสานแพ็ตอะเรย์ พบว่าประสิทธิภาพการผสานแพ็ตอะเรย์ ขึ้นกับปัจจัย ดังต่อไปนี้

1. เวลาทั้งหมดในการอ่านซิสตริงตามตารางการผสานแพ็ตอะเรย์
2. เวลาทั้งหมดในการค้นหาแบบทวิภาคบนแพ็ตอะเรย์ย่อย แต่เนื่องจากประมวลผลในหน่วยความจำจึงใกล้เคียงศูนย์
3. เวลาทั้งหมดในการอ่านแพ็ตอะเรย์ฉบับเก่าตามตารางการผสานแพ็ตอะเรย์
4. เวลาทั้งหมดในการบันทึกแพ็ตอะเรย์ฉบับใหม่ที่ได้จากแพ็ตอะเรย์ย่อยรวมกับแพ็ตอะเรย์ฉบับเก่าตามตารางการผสานแพ็ตอะเรย์

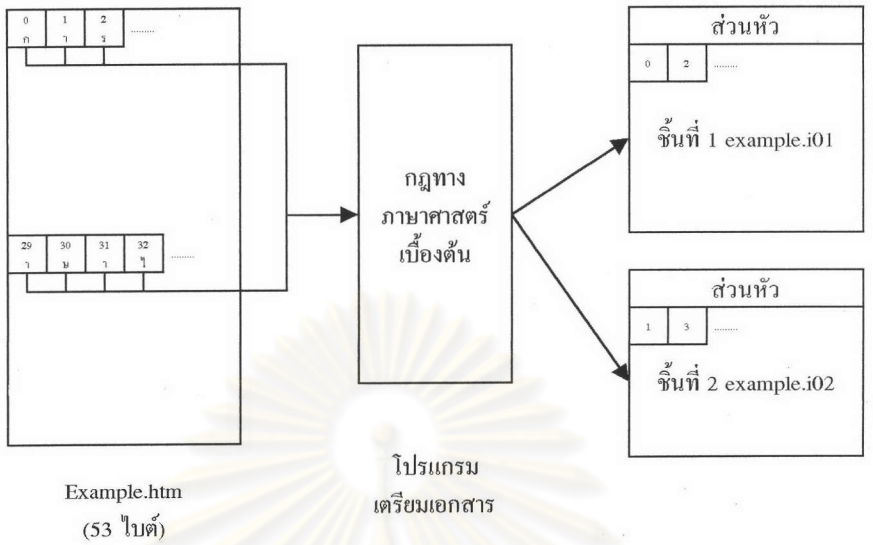
4.3.4 ตัวอย่างการสร้างแพ็ตอะเรย์

สมมติว่า ต้องการสร้างแพ็ตอะเรย์ ของแฟ้มต้นฉบับหนึ่ง โดยอาศัยวิธีการสร้างแพ็ตอะเรย์ดังกล่าวข้างต้น ซึ่งมีข้อกำหนดให้ดังต่อไปนี้

1. พารามิเตอร์กำหนดชื่อแฟ้มต้นฉบับ คือ Example.htm ขนาด 53 ไบต์ ซึ่งมีข้อความในแฟ้ม คือ “การพัฒนาระบบการค้นคืนข้อความภาษาไทยโดยใช้ต้นไม้แพ็ต”
2. พารามิเตอร์กำหนดขนาดแฟ้มต้นฉบับย่อย คือ 30 ไบต์
3. พารามิเตอร์กำหนดขนาดซิสตริง คือ 2 ไบต์

4.3.4.1 ตัวอย่างการเตรียมเอกสาร

การเตรียมเอกสาร คือ การจัดแบ่งแฟ้มต้นฉบับออกเป็นฉบับย่อย พร้อมทั้งค้นหาตำแหน่งซิสตริงตามกฎทางหลักภาษาศาสตร์เบื้องต้น แต่บันทึกเฉพาะตำแหน่งซิสตริงเท่านั้น เพื่อใช้เป็นดรรชนีอ้างอิงถึงซิสตริงในแฟ้มต้นฉบับ ดังรูปที่ 4.7



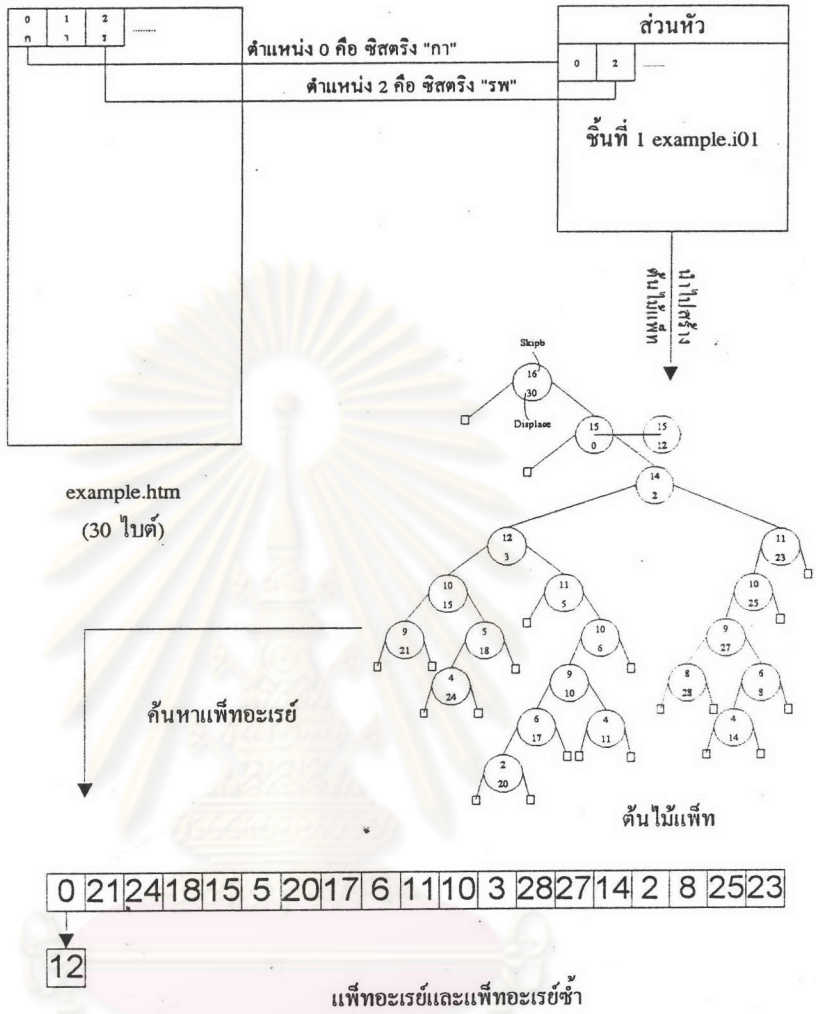
รูปที่ 4.7 ตัวอย่างการเตรียมเอกสาร

รูปที่ 4.7 เมื่อนำแฟ้มต้นฉบับ Example.htm ไปผ่านโปรแกรมเตรียมเอกสาร แล้วจะได้แฟ้มตำแหน่งซิสตริง คือ $\lceil 53/30 \rceil = 2$ ชิ้น ชื่อ Example.i01 กับ Example.i02 โดยขนาดแต่ละแฟ้มขึ้นกับจำนวนซิสตริงของแต่ละชิ้น (จำนวนซิสตริงทั้งหมด * ขนาดตำแหน่งซิสตริงในส่วนข้อมูล + ขนาดส่วนหัว) และแต่ละชิ้น (แฟ้มตำแหน่งซิสตริง) เก็บตำแหน่งที่อ้างอิงถึงตำแหน่งในแฟ้มต้นฉบับ เช่น แฟ้มข้อมูล Example.i02 ข้อมูลชุดแรก คือ 1 หมายความว่า เป็นตำแหน่งที่ 1 ในแฟ้มข้อมูล Example.i02 ที่อ้างอิงถึงตำแหน่งที่ 1 บวกกับข้อมูลตำแหน่งเริ่มต้นแฟ้มต้นฉบับในส่วนหัวของแฟ้มข้อมูล Example.i02 (29) คือ 1+29 = 30 คือ ตำแหน่งซิสตริง “ยา” ของแฟ้มต้นฉบับ เป็นต้น สำหรับตำแหน่งที่ไม่ถูกอ้างอิงผ่านแฟ้มตำแหน่งซิสตริงถึงเลยของแฟ้มต้นฉบับ แสดงว่าตำแหน่งนั้นไม่ใช่ซิสตริงตามกฎทางหลักภาษาศาสตร์เบื้องต้นดังกล่าวมาแล้ว

สำหรับแฟ้มตำแหน่งซิสตริง Example.i01 และ Example.i02 สามารถดูได้ โดยอาศัยทางเลือก ในข้อกำหนดเบื้องต้นของการเตรียมเอกสาร (ดูในภาคผนวก ก)

4.3.4.2 ตัวอย่างการสร้างแพ็ตอะเรย์

การสร้างแพ็ตอะเรย์ คือ การจัดเรียงลำดับซิสตริง ตามรหัสข้อมูล โดยอาศัยต้นไม้แพ็ตเป็นตัวช่วย โดยแพ็ตอะเรย์ คือ โหนดภายนอกของต้นไม้แพ็ตที่เรียงจากทางซ้ายไปทางขวา ซึ่งโหนดภายนอกเก็บตำแหน่งซิสตริงที่ใช้อ้างอิงถึงซิสตริงในแฟ้มต้นฉบับ ดังรูปที่ 4.8



รูปที่ 4.8 ตัวอย่างการสร้างแพ็ตอะเรย์

รูปที่ 4.8 นำแฟ้มตำแหน่งซิสตริงชิ้นแรก คือ Example.i01 ไปผ่านโปรแกรมการสร้างแพ็ตอะเรย์ เพื่อสร้างต้นไม้แพ็ต เพื่อหาแพ็ตอะเรย์ ตามขั้นตอนการสร้างแพ็ตอะเรย์ ดังนี้

1. โหลดแฟ้มข้อมูล Example.htm ขนาด 30 ไบต์เก็บในหน่วยความจำ

2. อ่านตำแหน่งซิสตริงจากแฟ้มตำแหน่งซิสตริง Example.i01 เก็บใน displace ของแต่ละโหนดในหน่วยความจำ

3. สร้างต้นไม้แพ็ตตามค่า displace ในแต่ละโหนด

4. ท่องไปในต้นไม้แพ็ต เพื่อหาแพ็ตอะเรย์ คือ 0, 12, 21, 24, 18, 15, 5, 20, 17, 6, 11, 10, 3, 28, 27, 14, 2 ,8, 25, 23

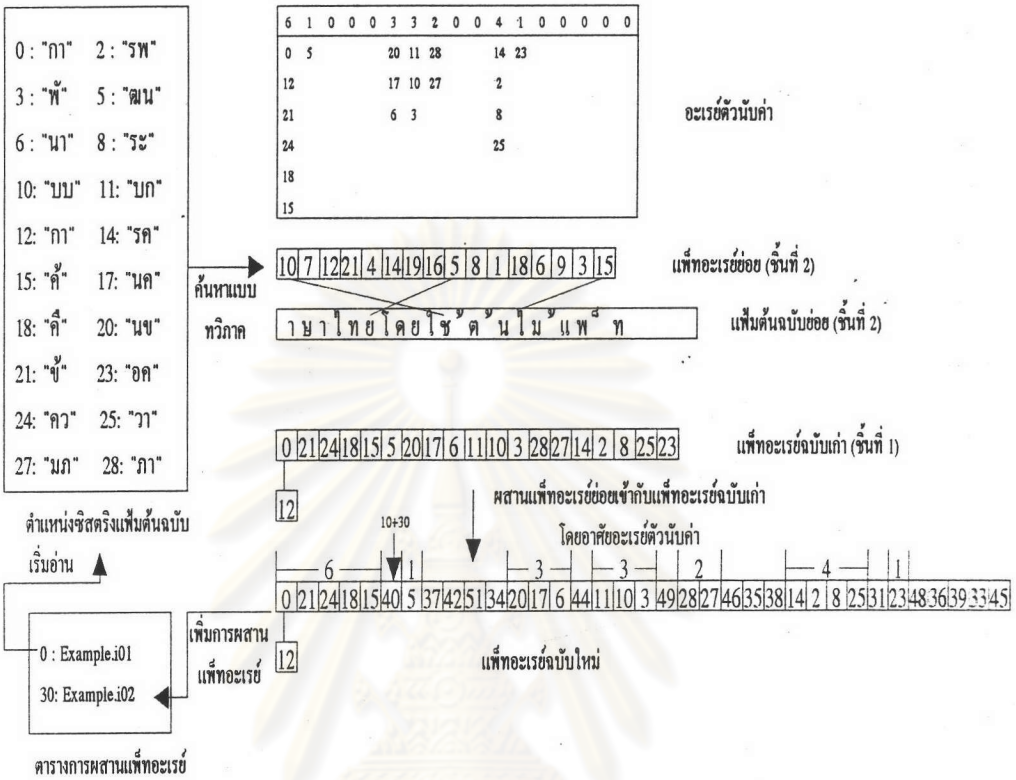
เมื่อพิจารณาค่าของแพ็ตอะเรย์ พบว่า ค่า 0 คือ ตำแหน่งของซิสตริง "กา" กับ ค่า 12 คือ ตำแหน่งของซิสตริง "กา" เช่นกัน ซึ่งเป็นซิสตริงที่มาก่อน ค่า 21 คือ ตำแหน่งของซิสตริง "ข้" ดังนั้น การสร้างต้นไม้แพ็ต เพื่อหาแพ็ตอะเรย์ คือ การจัดเรียงลำดับข้อมูลแบบดิจิตอลแบบหนึ่งที่มีประสิทธิภาพ

สำหรับแฟ้มตำแหน่งซิสตริงชิ้นที่สองคือ Example.i02 ทำในทำนองเดียวกับชิ้นแรก ซึ่งได้แพ็ตอะเรย์ คือ 10, 7, 12, 21, 4, 14, 19, 16, 5, 8, 1, 18, 6, 9, 3, 15

สำหรับแฟ้มแพ็ตอะเรย์ แฟ้มแพ็ตอะเรย์ซ้ำ และ แฟ้มอื่นๆ ในขั้นตอนการสร้างแพ็ตอะเรย์ สามารถดูได้ โดยอาศัยทางเลือก ในข้อกำหนดเบื้องต้นของการสร้างแพ็ตอะเรย์ (ดูในภาคผนวก ก)

4.3.4.3 <u>ตัวอย่างการผสานแพ็ตอะเรย์</u>

การผสานแพ็ตอะเรย์ คือ การนำแพ็ตอะเรย์ย่อยหลายๆ ฉบับที่ได้จากขั้นตอนการสร้างแพ็ตอะเรย์ มาผสานเข้าด้วยกันเป็นแพ็ตอะเรย์ฉบับใหม่ โดยอาศัยหน่วยความจำที่มีอยู่อย่างจำกัดให้เกิดประสิทธิภาพมากที่สุด ดังรูปที่ 4.9



รูปที่ 4.9 ตัวอย่างการผสานแพ็ตอะเรย์

รูปที่ 4.9 เป็นการผสานแพ็ตอะเรย์ของชิ้นแรก (Example.p01) กับชิ้นที่สอง (Example.p02) เก็บลงแฟ้มแพ็ตอะเรย์ฉบับใหม่ (Merge.pat) ตามขั้นตอนการผสานแพ็ตอะเรย์ ดังนี้

1. โหลดแพ็ตอะเรย์ย่อย (ชิ้นที่ 2) กับ แฟ้มต้นฉบับย่อย (ชิ้นที่ 2) เก็บในหน่วยความจำ

2. อ่านตำแหน่งซิสตริงตามตารางการผสานแพ็ตอะเรย์ แล้วนำไปค้นหาบนแพ็ตอะเรย์ย่อยแบบทวิภาคจนครบทุกตำแหน่ง เก็บค่านับจำนวนซิสตริงของแต่ละตัวลงในอะเรย์ตัวนับค่า ซึ่งอะเรย์ตัวนับค่า คือ อะเรย์เก็บจำนวนซิสตริงทั้งหมดที่มาก่อนและที่เท่ากันของซิสตริงขององแพ็ตอะเรย์ย่อยในแต่ละตัว (Element) เช่น ค่า 6 ตัวแรกในอะเรย์ตัวนับค่า หมายความว่า มีจำนวนซิสตริงของแพ็ตอะเรย์ฉบับเก่าจำนวน 6 ซิสตริงที่มีค่าน้อยกว่า หรือมาก่อนค่า 10 ตัวแรก

ในแพ็ตอะเรย์ย่อย (ชิ้นที่ 2) คือ ซิสตริง 0:"กา", 12:"กา", 21:"ช้", 24:"คว", 18:"คี", 15:"ค้" มาก่อน ซิสตริง (40-30): "ช้" หรือ 10:"ช้" โดย ค่า 40 ต้องนำมาหักออกจากตำแหน่งเริ่มต้นแฟ้มต้นฉบับ คือ 30 เนื่องจากตอนผสานได้บวกค่า 30 ไว้ เป็นต้น

3. ผสานแพ็ตอะเรย์ย่อย เข้ากับแพ็ตอะเรย์ฉบับเก่า โดยเริ่มอ่านแพ็ตอะเรย์ฉบับเก่าตามอะเรย์ตัวนับค่าทีละตัว แล้วตามด้วยแพ็ตอะเรย์ย่อย (ชิ้นที่ 2) ทีละตัว จนครบทุกตัว เก็บลงในแพ็ตอะเรย์ฉบับใหม่

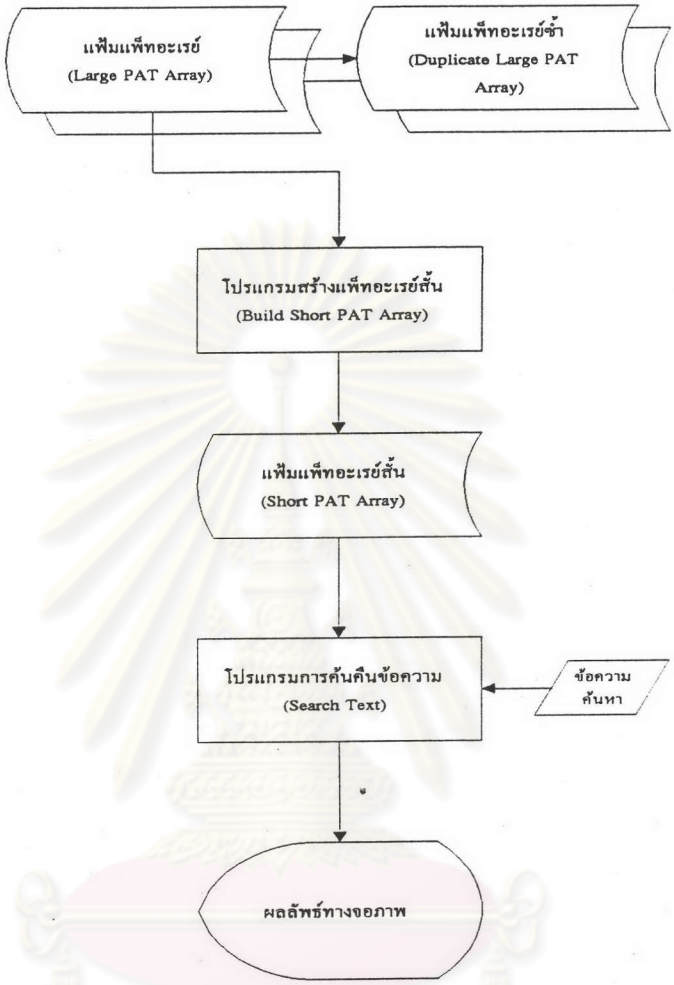
4. เพิ่มชื่อแฟ้มตำแหน่งซิสตริง (fname) ที่นำมาผสาน คือ Example.i02 กับค่าความยาวแฟ้มต้นฉบับย่อยแบบสะสม (mdisp) คือ ขนาดของแฟ้มต้นฉบับย่อยรวมกับค่าความยาวแฟ้มต้นฉบับย่อยแบบสะสมของชิ้นสุดท้ายในตารางการผสานแพ็ตอะเรย์ กล่าวคือ 0+30 เก็บลงในตารางการผสานแพ็ตอะเรย์

สำหรับแฟ้มแพ็ตอะเรย์ แฟ้มแพ็ตอะเรย์ซ้ำ และ แฟ้มอื่นๆ ในขั้นตอนการผสานแพ็ตอะเรย์ สามารถดูได้ โดยอาศัยทางเลือก ในข้อกำหนดเบื้องต้นของการผสานแพ็ตอะเรย์ (ดูในภาคผนวก ก)

4.4 การค้นคืนข้อความ

การค้นคืนข้อความเป็นการค้นหาข้อมูลที่ได้จัดเตรียมไว้ล่วงหน้าในขั้นตอนการสร้างแพ็ตอะเรย์ โดยแบ่งออกเป็น 2 ขั้นตอน คือ

1. การสร้างแพ็ตอะเรย์สั้น
2. การค้นคืนแบบเติมหน้ากับแบบบูล



รูปที่ 4.10 การค้นคืนข้อความ

รูปที่ 4.10 แสดงขั้นตอนการค้นคืนข้อความ โดยอาศัยแพ็ตอะเรย์ที่เตรียมไว้ล่วงหน้าจากขั้นตอนการสร้างแพ็ตอะเรย์ แล้วนำมาสร้างแพ็ตอะเรย์สั้น เพื่อเป็นดรรชนีคลุมแพ็ตอะเรย์อีกชั้นหนึ่ง ซึ่งสามารถค้นหาข้อมูลได้ทั้งแบบเติมหน้า หรือแบบบูล โดยมีหลักการ 2 ขั้นตอน คือ

1. สร้างแพ็ตอะเรย์สั้น เนื่องจากแพ็ตอะเรย์ทั้งหมด ไม่สามารถเก็บรวมไว้ในหน่วยความจำได้หมดในคราวเดียว ดังนั้นจึงสร้างแพ็ตอะเรย์สั้นคลุมอีกชั้นหนึ่ง เปรียบเสมือนเป็นดรรชนีของ

แพ็ตอะเรย์ เก็บไว้ในหน่วยความจำทั้งหมดในคราวเดียวได้ เพื่อใช้ในการค้นหาอีกทอดหนึ่ง โดยอาศัยโปรแกรมสร้างแพ็ตอะเรย์สั้นเป็นเครื่องมือในการสร้าง

2. ค้นคืนแบบเติมหน้า หรือแบบบูล หลังจากเตรียมแพ็ตอะเรย์สั้นเสร็จแล้ว สามารถนำไปค้นคืนแบบเติมหน้า (Prefix) ได้อย่างรวดเร็ว เช่น ค้นหาเอกสารที่มีข้อความขึ้นต้นด้วย "กา" ได้ เช่น "การพัฒนา" หรือ "การค้นคืน" เป็นต้น และสามารถค้นคืนแบบบูล (Boolean) ได้ เช่น ค้นหาเอกสารที่มีข้อความ "กา" หรือ "ษา" ก็ได้ เป็นต้น

4.4.1 การสร้างแพ็ตอะเรย์สั้น

ดังที่ได้กล่าวไว้แล้วเกี่ยวกับสาเหตุที่ต้องสร้างดรรชนีคลุมแพ็ตอะเรย์อีกทอดหนึ่งเพื่อให้ดรรชนี หรือ แพ็ตอะเรย์สั้นอยู่ในหน่วยความจำทั้งหมดได้ โดยสามารถกำหนดขนาดทั้งหมดของแพ็ตอะเรย์สั้น กับขนาดข้อมูลของแต่ละตัวของแพ็ตอะเรย์สั้น เป็นพารามิเตอร์ตัวหนึ่งของระบบ แล้วระบบสร้างแพ็ตอะเรย์สั้นตามพารามิเตอร์กำหนดขนาดทั้งหมดของแพ็ตอะเรย์สั้น ซึ่งค่าของแต่ละตัวในแพ็ตอะเรย์สั้น คือ ส่วนของข้อมูล (ซิสตริง) ของแพ็ตอะเรย์ตัวสุดท้ายของกลุ่มแพ็ตอะเรย์ ที่มีความยาวตามพารามิเตอร์ขนาดข้อมูลของแต่ละตัวของแพ็ตอะเรย์สั้น เก็บลงแฟ้มแพ็ตอะเรย์สั้น

4.4.1.1 ขั้นตอนการสร้างแพ็ตอะเรย์สั้น

1. กำหนดพารามิเตอร์ขนาดทั้งหมดของแพ็ตอะเรย์สั้น (M) กับพารามิเตอร์ขนาดข้อมูลของแต่ละตัวของแพ็ตอะเรย์สั้น (ls)

2. คำนวณหาจำนวนของแพ็ตอะเรย์สั้น (r) จากสูตร $r = M / ls$

3. อ่านค่าของจำนวนแพ็ตอะเรย์ทั้งหมด (n) จากส่วนหัวของแฟ้มแพ็ตอะเรย์

4. คำนวณหาค่าจำนวนกลุ่มแพ็ตอะเรย์ต่อหนึ่งแพ็ตอะเรย์สั้น จากสูตร

$$b = (n / r) = (n * ls / M)$$

เพื่อต้องการแบ่งแพ็ตอะเรย์ ออกเป็นกลุ่มๆ โดยมีจำนวนกลุ่มเท่ากับจำนวนแพ็ตอะเรย์สั้น

5. จองหน่วยความจำขนาดเท่ากับ พารามิเตอร์ขนาดทั้งหมดของแพ็ตอะเรย์สั้น (M)

6. อ่านส่วนของข้อมูล (ซิสตริง) ของแพ็ตอะเรย์ตัวสุดท้ายของกลุ่มแพ็ตอะเรย์ ที่มีความยาวตามพารามิเตอร์ขนาดข้อมูลของแต่ละตัวของแพ็ตอะเรย์สั้น เก็บลงแต่ละตัวของแพ็ตอะเรย์สั้นในหน่วยความจำที่จองไว้

7. เก็บแพ็ตอะเรย์สั้น ลงในแฟ้มแพ็ตอะเรย์สั้น

หมายเหตุ ควรปรับค่าจำนวนกลุ่มแพ็ตอะเรย์ต่อหนึ่งแพ็ตอะเรย์สั้น (b) ให้เป็นเลขลงตัว เพื่อประสิทธิภาพในการค้นหาข้อมูลดีขึ้น

4.4.1.2 ข้อกำหนดเบื้องต้นของการสร้างแพ็ตอะเรย์สั้น

1. ต้องระบุพารามิเตอร์ขนาดทั้งหมดของแพ็ตอะเรย์สั้น (M) กับพารามิเตอร์ขนาดข้อมูลของแต่ละตัวของแพ็ตอะเรย์สั้น (ls)

2. ต้องมีหน่วยความจำไม่น้อยกว่าพารามิเตอร์ขนาดทั้งหมดของแพ็ตอะเรย์สั้น (M)

3. มีทางเลือกให้เลือก ดังนี้

3.1 สามารถสร้างข้อมูลสถิติกับ ข้อมูลในแพ็ตอะเรย์สั้น เก็บลงแฟ้มข้อมูล Lschar.ls โดย ls คือ พารามิเตอร์ขนาดข้อมูลของแต่ละตัวของแพ็ตอะเรย์สั้น

4. แฟ้มแพ็ตอะเรย์สั้น คือ Spat.ls

4.4.1.3 โครงสร้างแฟ้มแพ็ตอะเรย์สั้น

ประกอบด้วย 2 ส่วน ดังนี้

1. ส่วนหัว (Header) มีองค์ประกอบหลักของข้อมูล คือ

ชื่อข้อมูล	รายละเอียดของข้อมูล
1. Large PAT Array	ชื่อแฟ้มแพ็ตอะเรย์
2. Duplicate Large PAT Array	ชื่อแฟ้มแพ็ตอะเรย์ซ้ำ
3. Table of Merge PAT Array	ชื่อแฟ้มตารางการผสานแพ็ตอะเรย์
4. Number of PAT Array	จำนวนแพ็ตอะเรย์สะสมทั้งหมด
5. Number of Duplicate PAT Array	จำนวนแพ็ตอะเรย์ซ้ำสะสมทั้งหมด
6. Memory	จำนวนหน่วยความจำทั้งหมด
7. Number of Last PAT Block	ตัวระบุจำนวนแพ็ตอะเรย์สั้นในกลุ่มสุดท้าย

8. Number of PAT Block	จำนวนกลุ่มแพ็ตอะเรย์
9. Number of Short PAT	จำนวนแพ็ตอะเรย์สั้น
10. Ls character	ขนาดข้อมูลของแต่ละตัวของแพ็ตอะเรย์สั้น
11. Length Sistring	ความยาวซิสตริง
12. Number of Small PAT Array	จำนวนชิ้นที่ทำการผสานแพ็ตอะเรย์
13. Flag Last PAT Block	ตัวระบุแพ็ตอะเรย์สั้นสุดท้ายไม่ครบ

ตารางที่ 4.5 ส่วนหัวแฟ้มแพ็ตอะเรย์สั้น

2. ส่วนข้อมูล (Detail) เก็บส่วนของข้อมูลของแพ็ตอะเรย์ตัวสุดท้ายของกลุ่มแพ็ตอะเรย์ ที่มีความยาวตามพารามิเตอร์ขนาดข้อมูลของแต่ละตัวของแพ็ตอะเรย์สั้น

4.4.1.4 ตัวอย่างการสร้างแพ็ตอะเรย์สั้น

รูปที่ 4.11 ตัวอย่างการสร้างแพ็ตอะเรย์สั้น

รูปที่ 4.11 เป็นการสร้างแพ็ตอะเรย์สั้น จากแพ็ตอะเรย์ที่สร้างขึ้นในขั้นตอนการสร้างแพ็ตอะเรย์ โดยมีขั้นตอนการสร้างแพ็ตอะเรย์สั้น ดังนี้

1. กำหนดค่า M = 10 ไบต์ กับ ls = 2 ไบต์
2. คำนวณค่า r = M / ls = 10 / 2 = 5 ตัว
3. อ่านค่า n = 35 ตัว
4. คำนวณค่า b = n / r = (n * ls / M) = (35 * 2 / 10) = 70 / 10 = 7 ตัว
5. จองหน่วยความจำขนาด M = 10 ไบต์
6. อ่านแพ็ตอะเรย์ตัวแทนกลุ่มแรก คือ แพ็ตอะเรย์ตัวสุดท้ายของกลุ่มแรก คือ ตัวที่ 7 มีค่าเป็น 5 นำไปตรวจสอบกับตารางการผสานแพ็ตอะเรย์ พบว่าอยู่ที่ค่าอ้างอิง 0 นำมาบวกกับ ค่า 5 ได้ 5+0 = 5 คือ ซิสตริง “ฒน” แล้วทำจนกระทั่งครบทั้ง 5 ตัว คือ
   - 6.1 ตัวที่ 7 มีค่า 5 คือ ซิสตริง “ฒน”
   - 6.2 ตัวที่ 14 มีค่า 6 คือ ซิสตริง “นา”
   - 6.3 ตัวที่ 21 มีค่า 27 คือ ซิสตริง “มภ”
   - 6.4 ตัวที่ 28 มีค่า 25 คือ ซิสตริง “วา”
   - 6.5 ตัวที่ 35 มีค่า 45 คือ ซิสตริง “ไม”
7. เก็บแพ็ตอะเรย์สั้น ลงในแฟ้มแพ็ตอะเรย์สั้น (Spat.002)

สำหรับแฟ้มข้อมูลต่างๆ ในขั้นตอนการสร้างแพ็ตอะเรย์สั้น ดูได้ในภาคผนวก ก

จากรูปที่ 4.11 ขั้นตอน การสร้างแพ็ตอะเรย์สั้น อาศัย หน่วยความจำ 10 ไบต์ ตาม พารามิเตอร์ที่กำหนด โดยมีประสิทธิภาพ ขึ้นกับเวลาทั้งหมดในการอ่านซิสตริงตามแพ็ตอะเรย์สั้นเก็บในหน่วยความจำ และเวลาทั้งหมดในการบันทึกแพ็ตอะเรย์สั้น ลงในแฟ้มแพ็ตอะเรย์สั้น

จากขั้นตอนการสร้างแพ็ตอะเรย์กับแพ็ตอะเรย์สั้น พบว่า วัตถุประสงค์หลักของการสร้างแพ็ตอะเรย์ คือ การเรียงลำดับซิสตริงให้ถูกต้องและรวดเร็ว เพื่อเพิ่มประสิทธิภาพให้กับขบวนการค้นข้อความ ซึ่งการสร้างแพ็ตอะเรย์ คือ การเรียงลำดับซิสตริงของเอกสารที่รวดเร็ว และแพ็ตอะเรย์จะไม่เก็บข้อความของเอกสารเพื่อใช้เป็นดรรชนี แต่แพ็ตอะเรย์จะเก็บตำแหน่งที่อ้างถึงซิสตริงในเอกสาร เพื่อใช้เป็นดรรชนีแทน โดยปกติแล้วปริมาณซิสตริงในเอกสาร มักมีปริมาณมากเกินกว่าหน่วยความจำจะจัดเก็บไว้ได้หมดต่อการประมวลผลครั้งหนึ่งๆ ดังนั้นจึงจำเป็นต้อง

สร้างแพ็ตอะเรย์สั้น เพื่อเป็นดรรชนีคลุมแพ็ตอะเรย์อีกทอดหนึ่ง โดยแพ็ตอะเรย์สั้นจะเก็บส่วนของซิสตริงของกลุ่มแพ็ตอะเรย์ที่ถูกจัดแบ่งตามความเหมาะสม กับหน่วยความจำที่มีอยู่ และแพ็ตอะเรย์สั้นสามารถเก็บไว้ในหน่วยความจำได้ทั้งหมด ต่อการประมวลผลครั้งหนึ่งๆ ซึ่งแพ็ตอะเรย์สั้นจะมีดรรชนีอ้างอิงถึงแพ็ตอะเรย์ที่อยู่ในหน่วยความจำสำรองอีกทอดหนึ่ง เพื่อใช้ในการค้นหาซิสตริงที่ไม่อยู่ในแพ็ตอะเรย์สั้น แต่อาจอยู่ในแพ็ตอะเรย์ได้ ดังนั้น การกำหนดจำนวนแพ็ตอะเรย์สั้น กับการกำหนดขนาดของแพ็ตอะเรย์สั้น ให้เหมาะสมต่อการค้นหาข้อความ เป็นปัจจัยสำคัญที่ทำให้การค้นหาข้อความรวดเร็วขึ้น ทั้งนี้ขึ้นอยู่กับหน่วยความจำเป็นหลัก

4.4.2 การค้นคืนแบบเติมหน้า

การค้นคืนแบบเติมหน้า คือ การค้นหาข้อมูลที่มีส่วนของข้อมูล ส่วนหน้าเหมือนกัน (ข้อมูลร่วม) เช่น การค้นหาข้อความที่ขึ้นต้นด้วย "คอม" เช่น "คอมพิวเตอร์" หรือ "คอมมูนิเคชั่น" เป็นต้น ซึ่งการค้นคืนแบบเติมหน้า โดยอาศัยแพ็ตอะเรย์มีความรวดเร็วมาก เนื่องจากแพ็ตอะเรย์ คือ ตำแหน่งซิสตริงของเอกสารที่เรียงลำดับแล้ว หรือคือ ซิสตริงของเอกสารที่เรียงลำดับแล้วต่อๆกัน ซิสตริงเหล่านี้มีข้อมูลร่วมที่เหมือนกันหรือใกล้เคียงกันเก็บเรียงต่อๆ กัน ดังนั้นในการค้นคืนแบบเติมหน้า จึงค้นหาซิสตริงที่มีข้อมูลร่วมตัวแรก กับ ซิสตริงที่มีข้อมูลร่วมตัวสุดท้าย ก็พอ (ไม่ต้องค้นหาหมดทุกตัว) และผลลัพธ์ของการค้นคืนแบบเติมหน้า คือ ซิสตริงทุกตัวนับตั้งแต่ซิสตริงตัวแรกที่มีข้อมูลร่วมจนถึงซิสตริงตัวสุดท้ายที่มีข้อมูลร่วม

4.4.2.1 ขั้นตอนการค้นคืนแบบเติมหน้า

1. กำหนดพารามิเตอร์ชื่อแฟ้มแพ็ตอะเรย์สั้น กับพารามิเตอร์ข้อความที่ต้องการค้นหา
2. อ่านแพ็ตอะเรย์สั้นเก็บไว้ในหน่วยความจำ (ทำหนเดียว)
3. ค้นหาแบบทวิภาคบนแพ็ตอะเรย์สั้น เพื่อหากลุ่มแพ็ตอะเรย์สั้นตัวแรก กับตัวสุดท้าย แล้วจะได้ช่วงของกลุ่มแพ็ตอะเรย์สั้น (br) ตามความยาวของแพ็ตอะเรย์สั้นหนึ่งตัว (b)
4. ถ้าความยาวของข้อความค้นหา(lq) มากกว่า ความยาวของแพ็ตอะเรย์สั้นหนึ่งตัว (ls) ให้ค้นหาแบบทวิภาคบนกลุ่มแพ็ตอะเรย์สั้นตั้งแต่กลุ่มแพ็ตอะเรย์สั้นตัวแรกถึงกลุ่มแพ็ตอะเรย์สั้นตัวสุดท้าย (br) เพื่อหากลุ่มแพ็ตอะเรย์สั้นตัวแรก กับตัวสุดท้าย แล้วจะได้ช่วงของกลุ่มแพ็ตอะเรย์สั้นช่วงใหม่ ตามความยาวของข้อความค้นหา (lq)
5. อ่านกลุ่มแพ็ตอะเรย์สั้นตัวแรกเก็บในหน่วยความจำ

6. ค้นหาแบบทวิภาคบนกลุ่มแพ็ตอะเรย์สั้นตัวแรกในหน่วยความจำ เพื่อหาซิสตริงตัวแรกตามข้อความค้นหา

7. อ่านกลุ่มแพ็ตอะเรย์สั้นตัวสุดท้ายเก็บในหน่วยความจำ

8. ค้นหาแบบทวิภาคบนกลุ่มแพ็ตอะเรย์สั้นตัวสุดท้ายในหน่วยความจำ เพื่อหาซิสตริงตัวสุดท้ายตามข้อความค้นหา

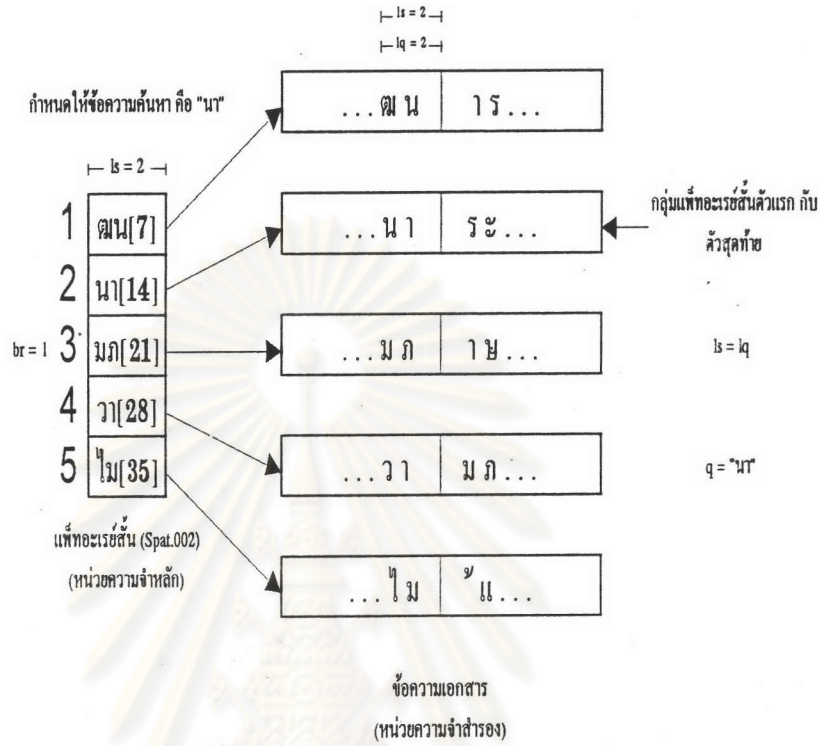
9. อ่านซิสตริงตั้งแต่ตัวแรกจนถึงตัวสุดท้ายตามที่ต้องการค้นหาแล้วเก็บบันทึกว่าซิสตริงผลลัพธ์อยู่ในเอกสารชิ้นใด โดยอาศัยอะเรย์เก็บจำนวนชิ้นเอกสาร ซึ่งใช้ค่า 1 บิต ต่อ เอกสารหนึ่งชิ้น

4.4.2.2 ข้อกำหนดเบื้องต้นของการค้นคืนแบบเติมหน้า

1. ต้องระบุพารามิเตอร์ชื่อแฟ้มแพ็ตอะเรย์สั้น กับพารามิเตอร์ข้อความที่ต้องการค้นหา
2. ต้องมีหน่วยความจำไม่น้อยกว่าแพ็ตอะเรย์สั้น
3. มีทางเลือกให้เลือก ดังนี้

    3.1 สามารถเก็บผลการค้นคืนแบบเติมหน้า เก็บลงแฟ้มข้อมูล Search.log

4.4.2.3 ตัวอย่างการค้นคืนแบบเติมหน้า



รูปที่ 4.12 ตัวอย่างการค้นคืนแบบเติมหน้า

รูปที่ 4.12 เป็นการค้นหาข้อความที่ขึ้นต้นด้วย “นา” ตามขั้นตอนการค้นคืนแบบเติมหน้า ดังนี้

1. กำหนดค่าพารามิเตอร์กำหนดชื่อแฟ้มแพ็ตอะเรย์สั้น คือ Spat.002 กับ ข้อความที่ต้องการค้นหาขึ้นต้นด้วย “นา”

2. อ่าน Spat.002 เก็บในหน่วยความจำ (ทำหนเดียว)

3. ค้นหาแบบทวิภาคบน Spat.002 ได้ br =1 และ ls = 2

4. ความยาวของ ls เท่ากับ lq ให้ข้ามขั้นตอนที่ 4 ไป

5. อ่านกลุ่มแพ็ตอะเรย์สั้นตัวแรกของกลุ่มแพ็ตอะเรย์ที่ค้นหาแบบทวิภาคได้ คือ กลุ่มที่ 2 เก็บในหน่วยความจำ

6. ค้นหาแบบทวิภาคบนกลุ่มแพ็ตอะเรย์สั้นตัวแรก ได้ซิสตริง คือ “นา”

7. อ่านกลุ่มแพ็ตอะเรย์สั้นตัวสุดท้ายของกลุ่มแพ็ตอะเรย์ที่ค้นหาแบบทวิภาคได้ คือ กลุ่มที่ 2 เก็บในหน่วยความจำ

8. ค้นหาแบบทวิภาคบนกลุ่มแพ็ตอะเรย์สั้นตัวสุดท้าย ได้ซิสตริง คือ "นา" เหมือนกัน

9. คำตอบ คือ ซิสตริง "... นา ! ระ ..." ซึ่งอยู่ในเอกสารชิ้นแรก คือ Example.i01 ของ Example.htm แล้วเก็บผลลัพธ์ลงในอะเรย์เก็บจำนวนชิ้นเอกสาร ได้ บิตแรกเป็น 1 คือ ซิสตริง "นา" อยู่ในเอกสารชิ้นแรก แล้วบันทึกผลการค้นหาลงในแฟ้มข้อมูล Search.log (ดูข้อมูลได้ในภาคผนวก ก)

จากรูปที่ 4.12 ขั้นตอนการค้นคืนแบบเติมหน้า อาศัยหน่วยความจำเท่ากับ ขนาดแพ็ตอะเรย์สั้นรวมกับขนาดของกลุ่มแพ็ตอะเรย์หนึ่งกลุ่ม คือ 10+(8*3) = 34 ไบต์ โดยมีประสิทธิภาพรวม ขึ้นกับปัจจัย ดังต่อไปนี้

1. เวลาทั้งหมดในการอ่านแพ็ตอะเรย์สั้นเก็บในหน่วยความจำ

2. เวลาทั้งหมดในการค้นหาแบบทวิภาคบนแพ็ตอะเรย์สั้น แต่เนื่องจากประมวลผลในหน่วยความจำจึงใกล้เคียงศูนย์

3. เวลาทั้งหมดในการอ่านกลุ่มแพ็ตอะเรย์สั้นกลุ่มแรกกับกลุ่มสุดท้ายเก็บในหน่วยความจำ

4. เวลาทั้งหมดในการค้นหาแบบทวิภาคของกลุ่มแพ็ตอะเรย์สั้นกลุ่มแรกกับกลุ่มสุดท้าย แต่เนื่องจากประมวลผลในหน่วยความจำจึงใกล้เคียงศูนย์

4.4.3 <u>การค้นคืนแบบบูล</u>

การค้นคืนแบบบูล คือ การค้นหาข้อมูลตามตรรกะ แอนด์ (And) กับ ออร์ (Or) เช่น การค้นหาข้อความที่ขึ้นต้นด้วย "คอม" แอนด์ "ระบบ" หมายความว่า ต้องการค้นหาเอกสารที่มีซิสตริง "คอม" และ "ระบบ" อยู่ในเอกสารชิ้นเดียวกัน เป็นต้น ซึ่งการค้นคืนแบบบูลโดยอาศัยแพ็ตอะเรย์จะช้ากว่าแบบเติมหน้า เนื่องจาก จำเป็นต้องค้นหาซิสตริงของแต่ละข้อความทุกตัวก่อน (อย่างน้อยสองตัว) แล้วนำผลลัพธ์ของแต่ละซิสตริงมาประมวลผลตามตรรกะนั้น (And, Or) ที่หลัง แล้วจึงแสดงผลลัพธ์สุดท้ายอีกทอดหนึ่ง

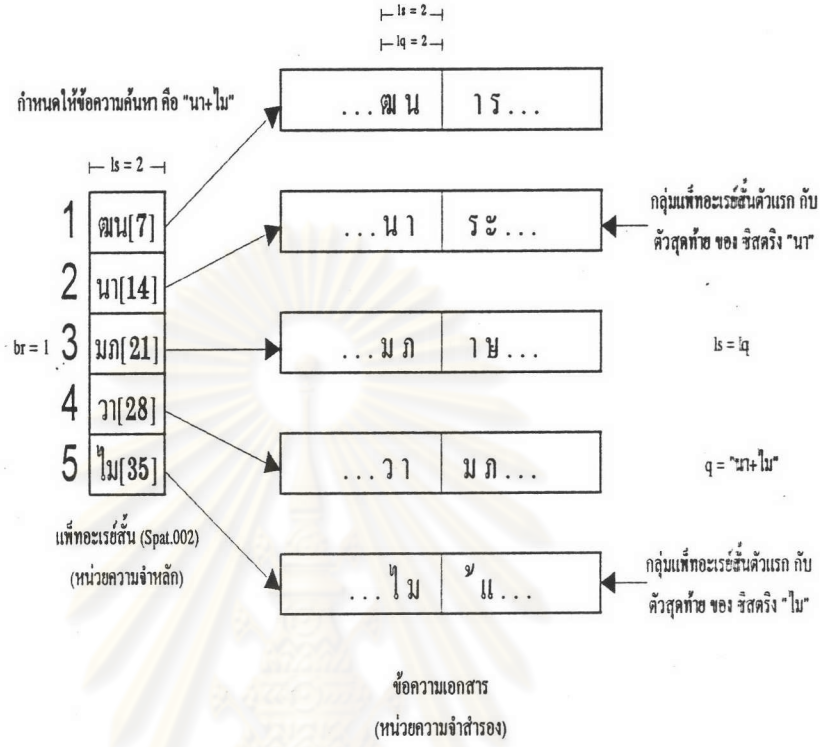
4.4.3.1 ขั้นตอนการค้นคืนแบบบูล

1. กำหนดพารามิเตอร์ชื่อแฟ้มแพ็ตอะเรย์สั้น กับพารามิเตอร์ข้อความที่ต้องการค้นหา
2. อ่านแพ็ตอะเรย์สั้นเก็บไว้ในหน่วยความจำ (ทำหนเดียว)
3. ตัดแบ่งข้อความที่ต้องการค้นหา ออกเป็น ข้อความย่อยแต่ละตัว
4. นำข้อความย่อยตัวแรกไปค้นคืนแบบเติมหน้า แล้วเก็บผลลัพธ์ลงในอะเรย์เก็บจำนวนชิ้นเอกสาร
5. นำข้อความย่อยตัวถัดไปค้นคืนแบบเติมหน้า แล้วเก็บผลลัพธ์ลงในอะเรย์เก็บจำนวนชิ้นเอกสารอีกตัวหนึ่ง แล้วนำอะเรย์ทั้งสองมาประมวลผลตามตรรกะของระหว่างซิสตริงทั้งสอง
6. นำข้อความย่อยตัวถัดไปประมวลผลตามขั้นตอนเดิม จนหมดทุกซิสตริง

4.4.3.2 ข้อกำหนดเบื้องต้นของการค้นคืนแบบบูล

1. ต้องระบุพารามิเตอร์ชื่อแฟ้มแพ็ตอะเรย์สั้น กับพารามิเตอร์ข้อความที่ต้องการค้นหา
2. ต้องมีหน่วยความจำไม่น้อยกว่าแพ็ตอะเรย์สั้น
3. มีทางเลือกให้เลือก ดังนี้
    3.1 สามารถเก็บผลการค้นคืนแบบเติมหน้า เก็บลงแฟ้มข้อมูล Search.log

4.4.3.3 ตัวอย่างการค้นคืนแบบบูล



รูปที่ 4.13 ตัวอย่างการค้นคืนแบบบูล

รูปที่ 4.13 เป็นการค้นหาข้อความที่ขึ้นต้นด้วย “นา” หรือ ขึ้นต้นด้วย “ไม” ตามขั้นตอนการค้นคืนแบบบูล ดังนี้

1. กำหนดค่าพารามิเตอร์กำหนดชื่อแฟ้มแพ็ตอะเรย์สั้น คือ Spat.002 กับ ข้อความที่ต้องการค้นหาขึ้นต้นด้วย “นา+ไม”

2. อ่าน Spat.002 เก็บในหน่วยความจำ (ทำหนเดียว)

3. ตัดแบ่งข้อความที่ต้องการค้นหา ออกเป็น ข้อความย่อยแต่ละตัว ได้ “นา” กับ “ไม”

4. นำข้อความย่อยตัวแรก “นา” ไปค้นคืนแบบเติมหน้า แล้วเก็บผลลัพธ์ลงในอะเรย์เก็บจำนวนชิ้นเอกสาร ได้ บิตแรกเป็น 1 คือ ซิสตริงอยู่ในเอกสารชิ้นแรก

5. นำข้อความย่อยตัวถัดไป “ไม” ค้นคืนแบบเติมหน้า แล้วเก็บผลลัพธ์ลงในอะเรย์เก็บจำนวนชิ้นเอกสารอีกตัวหนึ่ง ได้ บิตสองเป็น 1 คือ ซิสตริงอยู่ในเอกสารชิ้นที่สอง แล้วนำอะเรย์ทั้งสองมาประมวลผลตามตรรกะของระหว่างซิสตริงทั้งสอง คือ ออร์ จะได้ผลลัพธ์สุดท้าย คือ บิต

แรก กับบิตที่สองเป็น 1 แสดงว่าข้อความ “นา” หรือ “ไม” มีอยู่ในเอกสารทั้งสองชิ้น คือ Example.i01 กับ Example.i02 แล้วบันทึกผลการค้นหาลงในแฟ้มข้อมูล Search.log (ดูข้อมูลได้ในภาคผนวก ก)

จากรูปที่ 4.13 ขั้นตอนการค้นคืนแบบบูล อาศัยหน่วยความจำเท่ากับ ขนาดแพ็ตอะเรย์สั้นรวมกับขนาดของกลุ่มแพ็ตอะเรย์หนึ่งกลุ่ม คือ 10+(8*3) = 34 ไบต์ โดยมีประสิทธิภาพรวมมากกว่าแบบเติมหน้าเป็นสองเท่า