บทที่ 4

กระบวนการค้นข้อความไทยในเอกสารพีดีเอฟ

บทที่แล้วได้นำเสนอวิธีการในการถอดรหัสข้อความไทยในเอกสารพีดีเอฟไปแล้ว ขั้นตอนถัดไปของส่วนจำเพาะค้นข้อความไทยในเอกสารพีดีเอฟ ก็คือวิธีการค้นข้อความ ในบทนี้จะนำเสนอถึงวิธีการค้นข้อความไทย แนวคิดขั้นตอนและวิธีการทำงาน

4.1 การค้นข้อความไทยในเอกสารพีดีเอฟโดยวิธีการเปรียบเทียบสายอักขระ

การค้นข้อความไทยในเอกสารพีดีเอฟ จะใช้วิธีการค้นโดยวิธีการเปรียบเทียบสายอักขระตามแนวคิดของ Brute-force คือ ทำการเปรียบเทียบโดยเลื่อนอักขระที่ทำการเปรียบเทียบไปทางขวามือครั้งละ 1 อักขระ การค้นข้อความไทยในเอกสารพีดีเอฟเลือกใช้วิธีการตามแนวคิดของ Brute-force เนื่องจาก กลุ่มอักขระที่ใช้ในการเปรียบเทียบทั้ง 2 กลุ่มมีขนาดไม่ยาวมาก ไม่ว่าจะเลือกวิธีการเปรียบเทียบแบบใดก็ใช้เวลาในการค้นข้อความไม่แตกต่างกัน แต่การค้นข้อความไทยในเอกสารพีดีเอฟ ใช้การเปรียบเทียบตามวิธีการของ Brute-force อย่างเดียวไม่เพียงพอ ต้องมีการดัดแปลงการทำงานบางอย่างเพิ่มเติมเข้าไปด้วย ทั้งนี้เนื่องจากผลลัพธ์ที่ได้จากการถอดรหัสข้อความไทยเป็นปัญหาและอุปสรรคในการค้นข้อความ

กระบวนการถอดรหัสข้อความไทยในเอกสารพีดีเอฟ จะทำการแยกวัตถุข้อความออกมาจากเอกสารพีดีเอฟแล้วแปลงรหัสตัวอักษรให้ตรงตามข้อกำหนด มอก.620 ผลลัพธ์ของกระบวนการนี้จะได้กลุ่มของอักขระครั้งละหนึ่งกลุ่มอักขระ สำหรับภาษาอังกฤษ เอกสารพีดีเอฟถูกออกแบบมาให้สนับสนุนกับคุณสมบัติและข้อกำหนดต่างๆของภาษานี้ การนำข้อความออกมาจากเอกสารพีดีเอฟสำหรับภาษาอังกฤษแล้ว จะได้ข้อความออกมาเป็นคำหนึ่งๆ เช่น “Hello”, “word” แต่สำหรับภาษาไทยแล้ว คำๆหนึ่งในภาษาไทย อาจจะถูกแยกออกเป็นหลายกลุ่มอักขระ

ตัวอย่างเช่น ข้อความ “สวัสดีชาวโลก”

อาจจะได้เป็น “สวัสดีชาวโลก”

หรือ “สวัสดี” + “ชาวโลก”

หรือ “สวั” + “สดีช” + “าวโลก”

หรือ อื่นๆ

สระอำในเอกสารพีดีเอฟจะใช้รหัสตัวอักษร 2 ตัว เมื่อทำการถอดรหัสข้อความไทยในเอกสารพีดีเอฟ จะได้รหัสตัวอักษร คือ 237 และ 210

จากเหตุผลข้างต้น ทำให้การค้นข้อความไทยในเอกสารพีดีเอฟโดยวิธีการเปรียบเทียบอักขระแบบทั่วไปจะไม่พบอักขระที่ต้องการ ถึงแม้จะมีข้อความที่ต้องการอยู่ในเอกสารก็ตาม เนื่องจาก ข้อความไทยในเอกสารพีดีเอฟที่ได้จากถอดรหัสถูกแยกออกเป็นหลายส่วน การค้นข้อความไทยในเอกสารพีดีเอฟ จึงต้องมีกระบวนการและเงื่อนไขเพิ่มเติมในการที่จะตรวจสอบว่า พบกลุ่มอักขระที่ตรงกับข้อความที่ต้องการค้น แต่ยังไม่ครบทุกอักขระ เมื่อทำการเปรียบเทียบกลุ่มอักขระกลุ่มถัดไปในเอกสารพีดีเอฟ ให้นำผลการเปรียบเทียบในครั้งก่อนมาพิจารณาด้วย
โดยใช้วิธีการดังนี้

นำข้อความที่ต้องการค้นกับข้อความในเอกสารพีดีเอฟไปหาจำนวนอักขระที่ตรงกัน โดยวิธีการเปรียบเทียบสายอักขระ

ถ้า จำนวนอักขระที่พบเท่ากับขนาดของข้อความที่ต้องการค้น แสดงว่าพบข้อความที่ต้องการค้น

แต่ถ้า พบจำนวนอักขระที่ตรงกันตั้งแต่ 1 อักขระขึ้นไปแต่น้อยกว่าขนาดของข้อความที่ต้องการค้น ให้ลดจำนวนอักขระของข้อความที่ต้องการค้นจากทางซ้ายไปเป็นจำนวนอักขระที่พบข้อความ ให้ข้อความที่เหลือเป็นข้อความที่ต้องการค้น แล้วกลับไปทำการค้นใหม่

ตัวอย่างเช่น ข้อความที่ต้องการค้น = "ชาว" ข้อความในเอกสารพีดีเอฟ = "สวัสดีชาวโลก"

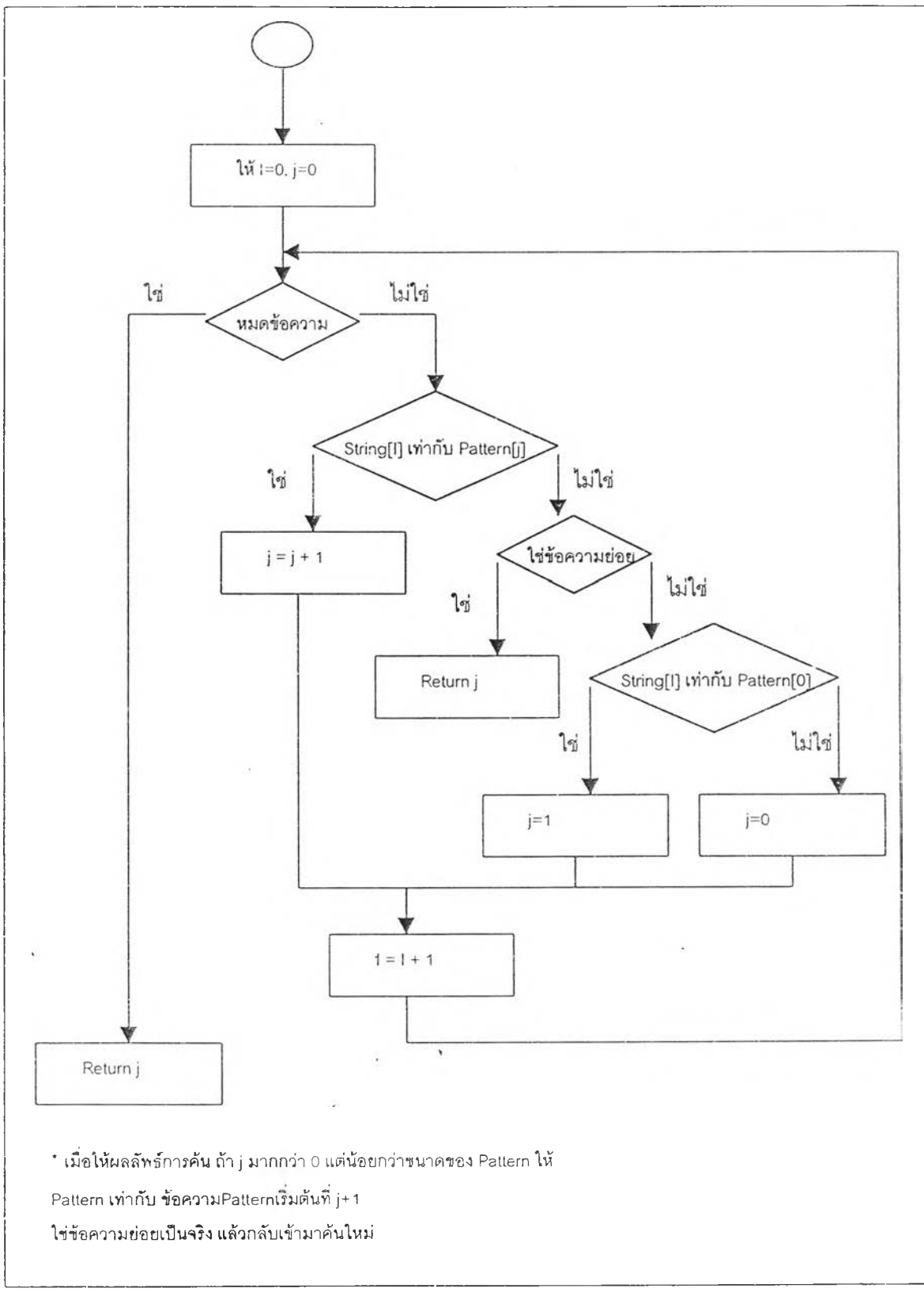
เมื่อแยกข้อความออกมาจากเอกสารพีดีเอฟ = "สวั" + "สดีช" + "าวโลก" จะได้ว่า

กลุ่มอักขระในเอกสารพีดีเอฟ = "สวั"
ข้อความที่ต้องการค้น = "ชาว"
ไม่พบอักขระ "ช" ในกลุ่มอักขระ "สวั"
กลุ่มอักขระในเอกสารพีดีเอฟ = "สดีช"
ข้อความที่ต้องการค้น = "ชาว"
พบข้อความที่ต้องการค้น จำนวน 1 อักขระ คือ "ช"
SubPattern = true
ข้อความที่ต้องการค้น = "าว"
กลุ่มอักขระในเอกสารพีดีเอฟ = "าวโลก"
ข้อความที่ต้องการค้น = "าว"
พบข้อความที่ต้องการค้นเป็นจำนวนอักขระเท่ากับขนาดของข้อความที่ต้องการค้น แสดงว่า พบข้อความที่ต้องการค้น

การค้นข้อความไทยในเอกสารพีดีเอฟ จะใช้วิธีการเปรียบเทียบสายอักขระโดยมีวิธีการดังนี้

```
WHILE(ไม่สิ้นสุดข้อความ)
    {
    เริ่มต้นอักษรที่จะเปรียบเทียบที่ตัวอักษรตัวแรกของข้อความ
    IF ใช่ข้อความย่อยของสระอำ และ อักขระข้อความจากเอกสารคือสระอา THEN
        ให้ อักขระข้อความจากเอกสาร เท่ากับ สระอำ
    END IF
    IF อักขระข้อความจากเอกสาร คือ ' ◌ำ'(237) และ อักขระข้อความที่ต้องการค้นคือ สระอำ
    THEN
        IF อักขระ ' ◌ำ'(237) ไม่ใช่อักขระสุดท้าย และ อักขระข้อความจากเอกสารถัดไปคือสระอา
        THEN ให้ อักขระข้อความจากเอกสาร เท่ากับ สระอำ
        IF อักขระ ' ◌ำ'(237) ใช่อักขระสุดท้าย THEN ให้ข้อความย่อยของสระอำ เป็นจริง
    END IF
    IF อักขระของข้อความที่ต้องการค้น ตรงกับ อักขระของข้อความจากเอกสาร THEN
        เลื่อนอักขระของข้อความที่ต้องการค้นไปด้านขวา 1 อักขระ
    ELSE
        IF อักขระตัวแรกของข้อความที่ต้องการค้นตรงกับอักขระของข้อความจากเอกสาร THEN
            เลื่อนอักขระของข้อความที่ต้องการค้นไปอยู่อักขระตัวที่สอง
        ELSE
            เลื่อนอักขระของข้อความที่ต้องการค้นไปอยู่ที่อักขระตัวแรก
        END IF
            IF ใช่ข้อความย่อย THEN Return จำนวนอักขระที่ตรงกัน
            END IF
    END IF
    IF จำนวนอักขระที่ตรงกัน เท่ากับ ขนาดของข้อความที่ต้องการค้น THEN
        Return จำนวนอักขระที่ตรงกัน
    END IF
    เลื่อนอักขระเป็นอักขระถัดไปของข้อความจากเอกสาร
    }
END LOOP
```

จากเหตุผลข้างต้น จะได้แนวคิดของกระบวนการเปรียบเทียบข้อความไทยในเอกสารพีดีเอฟ ดังนี้



รูปที่ 21 ผังงานกระบวนการเปรียบเทียบข้อความในส่วนจำเพาะค้นข้อความไทย

4.2 การแสดงตำแหน่งข้อความที่ค้นพบในเอกสารพีดีเอฟ

ส่วนจำเพาะเมื่อทำการค้นข้อความในเอกสารพีดีเอฟแล้ว จะทำการตอบสนองกับผู้ใช้ให้ทราบว่าพบข้อความหรือไม่ ในกรณีที่ไม่พบข้อความ เอกสารพีดีเอฟจะถูกค้นไปจนสิ้นสุดเอกสารแล้วแสดงให้ผู้ใช้ทราบว่าสิ้นสุดเอกสารแล้ว ถ้าผู้ใช้ต้องการค้นเอกสารต่อส่วนจำเพาะจะค้นข้อความต่อไปโดยจะกลับไปเริ่มต้นค้นข้อความที่หน้าแรกของเอกสารพีดีเอฟ ในกรณีที่พบข้อความส่วนจำเพาะจะแสดงข้อความที่ค้นพบ โดยการแสดงสีทึบทับตำแหน่งข้อความที่ค้นพบในเอกสารพีดีเอฟ ข้อความไทยในเอกสารพีดีเอฟอาจถูกแยกออกเป็นหลายส่วน ส่วนจำเพาะจะแสดงสีทึบทับตำแหน่งข้อความที่ค้นพบตามจำนวนของข้อความที่ถูกแยกออกเป็นหลายส่วนเหล่านั้น

วิธีการแสดงตำแหน่งข้อความที่ค้นพบในเอกสารพีดีเอฟ

1. เปิดไปยังหน้าเอกสารที่พบข้อความ
2. กำหนดให้ตำแหน่งเริ่มต้นในการแสดงสีทึบเท่ากับตำแหน่งข้อความที่พบ
3. กำหนดให้ขนาดข้อความที่จะแสดงสีทึบเท่ากับจำนวนข้อความที่ถูกแยกออกเป็นหลายส่วน
4. แสดงสีทึบทับข้อความที่พบ