

การวิเคราะห์ภาพลักษณะเอกสารภาษาไทยด้วยทฤษฎีกราฟ THAI DOCUMENT IMAGE ANALYSIS WITH GRAPH THEORY

สาธิต อินทจักร และ กิตติ ไพฑูรย์วัฒนกิจ

ภาควิชาเทคโนโลยีการวัดคุมทางอุตสาหกรรม ภาควิชาอิเล็กทรอนิกส์

คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบัง

Sathit Intajag and Kittti Paithoonwatanakij

Department of Industrial Instrumentation Technology

King Mongkut's Institute of Technology Ladkrabang

บทคัดย่อ

การวิเคราะห์ภาพลักษณะเอกสาร (Document Image Analysis) เป็นขั้นตอนต้นๆ ก่อนที่จะนำไปสู่กระบวนการรู้จักตัวอักษร (Optical Character Recognition : OCR) ในงานวิจัยชิ้นนี้ จะเน้นหนักไปที่ การวิเคราะห์เอกสารที่เป็นภาษาไทย ที่มีทั้งตัวอักษรและรูปภาพ สำหรับวิธีที่ใช้ในการวิเคราะห์เอกสารคือวิธี "Bottom-up" โดยพิจารณาจากลักษณะการต่อกันของจุดภาพ (Pixels) ในแต่ละตำแหน่งจนกลายเป็นวัตถุชิ้นหนึ่ง (Connected Components) และได้นำทฤษฎีกราฟในเรื่องของการหาค่าถ่วงน้ำหนักโดยใช้วิธี Minimum-Cost Spanning Trees มาใช้ในการวิเคราะห์การเชื่อมต่อกันของวัตถุเพื่อช่วยให้สามารถระบุได้ว่าเป็นตัวอักษรหรือรูปภาพ และโครงสร้างของเอกสารแต่ละหน้า สำหรับผลการทดลองจากการนำโมเดลนี้ ไปวิเคราะห์เอกสารไม่ว่าจะเป็นหนังสือหรือวารสารต่างๆ ปรากฏว่าให้ผลลัพธ์อยู่ในเกณฑ์ที่ดี แต่ถ้าพิจารณาลึกลงไปถึงการแยกตัวอักษร (Character Segmentation) ที่ติดกันก็ยังเป็นปัญหาอยู่และกำลังทำการวิจัยเรื่องนี้ต่อไป

ABSTRACT

The document image analysis is necessary for optical character recognition (OCR). In this paper, Thai document image analysis with graph theory is proposed. These documents have both text and picture. The method to segment each object is bottom-up approach. The connected components are employed to collect a group of pixels that represent an object (character and picture). And weighted graph algorithm is applied to analyze the objects connection to construct the page structure. For experimental results, the presented model can analyze documents such as book or magazine. For character segmentation, this model still has a problem with touching characters that needs further research.

บทนำ

การวิเคราะห์เอกสารภาพเป็นขั้นตอนที่สำคัญ ที่
จะนำไปสู่ขั้นตอน OCR ที่กล่าวเสนอนี้ก็เนื่องจาก การ
วิเคราะห์เอกสารแล้วนำมาเก็บไว้เป็นระบบ จะทำให้
สะดวกและมีความถูกต้องในขั้นตอนของ OCR ยังมี
การเก็บรายละเอียดที่สำคัญของเอกสารไว้มากเท่าไร
ความถูกต้องของ OCR ก็จะเป็นไปตามตัว
สำหรับรายละเอียดต่างๆ ที่สำคัญเช่น ตำแหน่ง ขนาด
และช่องว่างระหว่างตัวอักษร นอกจากนี้ก็จะเป็นข้อ
ปลื้มย่อยของเอกสารแต่ละชนิด เช่น การจัดย่อหน้า
การระบุส่วนไหนคือหัวข้อใหญ่ หรือหัวข้อย่อย และ
ตำแหน่งและขนาดของรูปภาพ เป็นต้น

การวิเคราะห์เอกสารภาพ สามารถดำเนินการได้โดยวิธีทั่วๆ ไป 2 วิธี¹ คือการวิเคราะห์แบบลำดับชั้น (Hierarchical Methods) และการวิเคราะห์โดยไม่เป็นลำดับชั้น (Non-hierarchical Methods) สำหรับวิธีการวิเคราะห์เอกสารแบบลำดับชั้น สามารถแบ่งย่อยลงไปได้อีกสองวิธีคือ Top-down และ Bottom-up และในบทความนี้จะใช้การวิเคราะห์แบบ Bottom-up เนื่องจากวิธีนี้สามารถวิเคราะห์รายละเอียดที่สำคัญของเอกสารได้มาก และยังวิเคราะห์รูปแบบของเอกสารได้หลากหลายกว่าวิธี Top-down แต่วิธีนี้จะมีขั้นตอนการดำเนินการที่ซับซ้อนกว่า

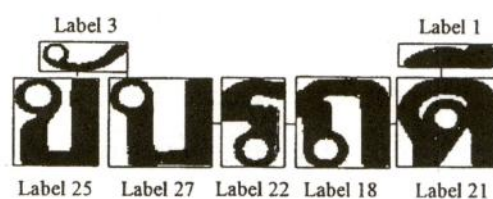
การวิเคราะห์เอนกสารภาพโดยวิธี Bottom-up จะเริ่มจากการพิจารณาการต่อกันของจุดภาพจนกลายเป็นวัตถุขึ้นหนึ่ง ซึ่งอาจจะเป็นตัวอักษร ส่วนใดส่วนหนึ่งของตัวอักษรหรือรูปภาพก็ได้ สำหรับวิธีการที่นำมาใช้วิเคราะห์การต่อกันของจุดภาพจนกลายเป็นวัตถุแต่ละชิ้นคือ วิธี Connected Components² เมื่อวัตถุถูกแยกและทำการติดป้าย (Label) ให้กับวัตถุแต่ละชิ้นแล้ว จะทำการใส่วัตถุเหล่านี้เข้าไปในโครงสร้างข้อมูลกราฟ³ หลังจากนั้นจะใช้ทฤษฎีกราฟในหัวข้อการเชื่อมต่อกันของโหนด (Connectivity)⁵ มาพิจารณาว่าวัตถุขึ้นใดเป็นตัวอักษร ส่วนใดส่วนหนึ่ง

ของตัวอักษรหรือรูปภาพและรูปภาพ โดยจะพิจารณาจากตำแหน่งและการเชื่อมต่อกันของวัตถุ ซึ่งจุดนี้จะเป็นขั้นตอนการรวมเอาวัตถุที่เป็นเซตย่อย (Sub-Set) ของวัตถุอีกชิ้นหนึ่งเข้าด้วยกัน ดังตัวอย่างที่แสดงในรูปที่ 1

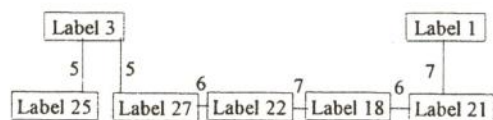
[illegible]

รูปที่ 1 แสดงวัตถุที่ซ้อนทับกันอยู่

ในรูปที่ 1 จะมีวัตถุสองชิ้นที่ซ้อนกันอยู่ โดยชิ้นแรกจะถูก Label ด้วยเลข 1 และชิ้นที่ 2 ถูก Label ด้วยเลข 2 วัตถุชิ้นที่สองเป็นเซตย่อยของวัตถุชิ้นที่หนึ่ง หลักเกณฑ์ที่ใช้ในการพิจารณาว่าวัตถุชิ้นใดเป็นเซตย่อยของวัตถุชิ้นอื่นหรือไม่ ก็จะใช้ขนาดของวัตถุเป็นหลัก โดยจะถือว่าวัตถุที่ขนาดเล็กกว่าจะเป็นเซตย่อยของวัตถุที่ใหญ่กว่า และเมื่อทำการรวมวัตถุที่เป็นเซตย่อยเข้ากับวัตถุแล้ว ก็จะลบวัตถุที่เป็นเซตย่อยออกจากโครงสร้างกราฟ ดังนั้นเมื่อผ่านการประมวลผลในขั้นตอนนี้แล้ว



11.



၂.

รูปที่ 2 แสดงการเชื่อมต่อระหว่างตัวอักษรด้วย
ค่าระยะทางที่สั้นที่สุด

วัตถุแต่ละชิ้นจะเป็นเซตที่จุดภาพเป็นอิสระต่อกัน คือ อินเตอร์เซกชันระหว่างวัตถุแต่ละชิ้นจะเป็นเซตว่าง

เมื่อวัตถุแต่ละชิ้นแยกกันโดยอิสระแล้ว ขั้นตอนต่อไปเป็นการพิจารณาถึงการเชื่อมโยงกันระหว่างวัตถุแต่ละชิ้นเพราะเราต้องไม่ลืมว่าวัตถุที่เรากำลังสนใจกันอยู่คือตัวอักษร ดังนั้นวัตถุที่อยู่ใกล้กันก็คือตัวอักษรที่ประกอบกันเป็นคำ เป็นวลี และ เป็นประโยคที่สามารถสื่อสารกันได้ ดังนั้นเทคนิคที่นำมาใช้ในขั้นตอนนี้ก็คือการหาระยะทางที่สั้นที่สุดระหว่างวัตถุแต่ละชิ้นในภาพ หรือที่เรียกว่า “Minimum-Cost Spanning Tress”^{3,4,5} เมื่อประมวลผลในขั้นตอนนี้แล้ว วัตถุแต่ละชิ้นจะเชื่อมต่อกันด้วยค่าระยะทางที่สั้นที่สุดดังแสดงในรูปที่ 2

ในรูปที่ 2 ถ้าเราพิจารณากันอย่างถี่ถ้วนแล้ว ขั้นตอนนี้ไม่ใช่ขั้นตอนสุดท้ายของการวิเคราะห์เอกสารภาพ เพราะถ้าเราเก็บโครงสร้างนี้แล้วนำตัวอักษรแต่ละตัวที่แยกออกมาได้ไปผ่านขั้นตอนการประมวลผลของ OCR โดยเรียงลำดับตาม Label ที่กำหนดก็จะได้ข้อความที่ไม่สามารถสื่อสารกันตามความหมายเดิมได้ อย่างในรูปที่ 2 เมื่อเรียงตัวอักษรแต่ละตัวตามลำดับ Label ก็จะได้ลำดับตัวอักษรดังนี้ “- ักตรขบ” ดังนั้นจะต้องมาวิเคราะห์กันต่อว่าจะเรียงตัวอักษรเหล่านี้อย่างไรเพื่อให้ได้ความหมายเหมือนเดิม

การเรียงลำดับตัวอักษรใหม่นั้น จะต้องใช้เทคนิคการวิเคราะห์ทางสถิติเข้ามาช่วย โดยใช้ข้อมูลของวัตถุแต่ละชิ้นที่เก็บอยู่ในโครงสร้างข้อมูลกราฟมาวิเคราะห์ คือเราใช้ตำแหน่งของวัตถุมาคำนวณค่าความสูงของตัวอักษร และใช้ค่าความแตกต่างของตำแหน่งของวัตถุในแนวแกนตั้งมากำหนดบรรทัด แล้วคำนวณค่าความถี่สะสมสูงสุด (ค่าฐานนิยม : Mode) ของความสูงตัวอักษรกำหนดเส้นกำกับบรรทัด ดังแสดงในรูปที่ 3

บรรทัด

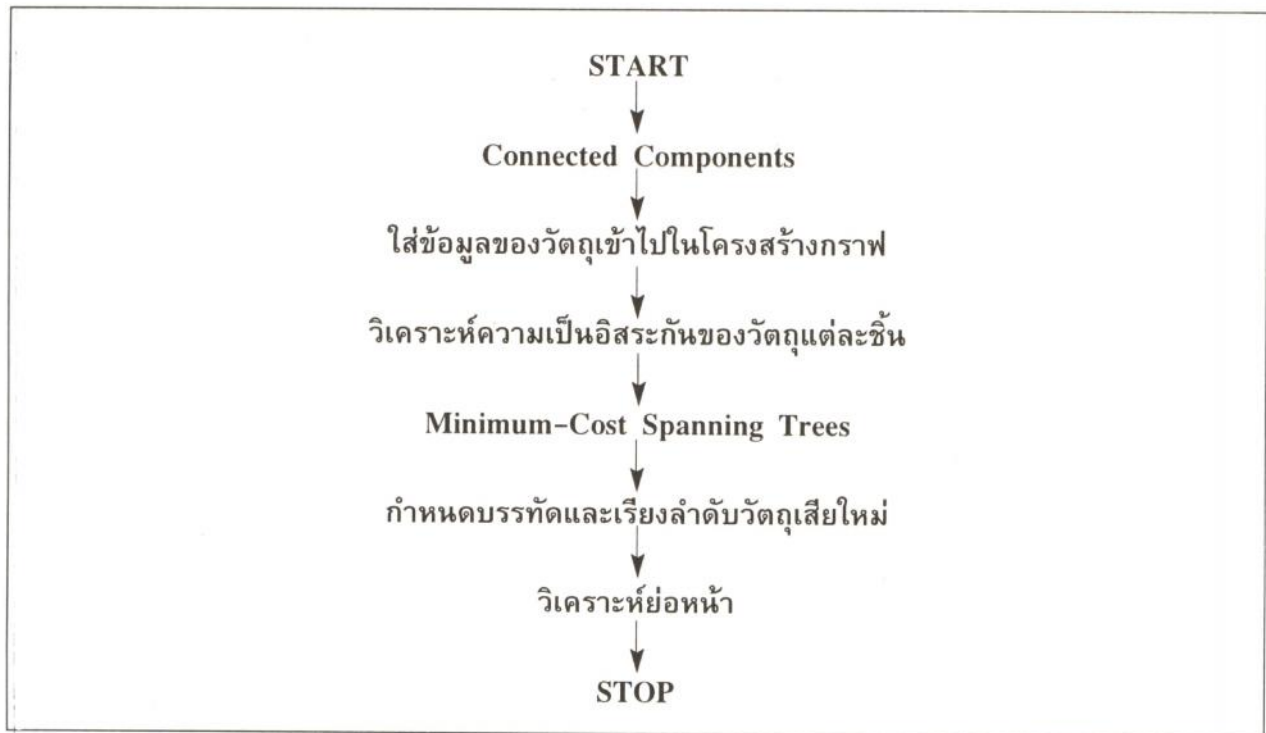
Base-Line

รูปที่ 3 แสดงเส้นกำกับบรรทัด

ในรูปที่ 3 เส้นกำกับบรรทัด ซึ่งมีทั้งข้างบนและข้างล่าง แต่เส้นที่สำคัญคือเส้นกำกับบรรทัดที่อยู่ด้านล่างหรือที่เรียกว่า “Base-Line” เนื่องจากวัตถุแต่ละชิ้นจะมีตำแหน่งซึ่งเป็นจุดที่ใกล้เคียงกันมากที่สุด ส่วนเส้นกำกับบนของวัตถุแต่ละชิ้นจะมีความแตกต่างกันเล็กน้อย แต่สามารถคำนวณค่าเฉลี่ยได้ เมื่อกำหนดเส้นกำกับบรรทัดได้แล้วก็ทำการเรียงลำดับตัวอักษรเสียใหม่ ตามแนวแกนนอน ก็จะได้ลำดับตัวอักษรที่ถูกต้องตามความหมายเดิม

ขั้นตอนสุดท้ายของการวิเคราะห์เอกสารภาพแบบ Bottom-up คือการจัดย่อหน้าในเป็นไปตามรูปแบบเดิมของเอกสาร ซึ่งในขั้นตอนนี้จะใช้เทคนิคการเขียนโปรแกรมเข้ามาจัด โดยใช้ข้อมูลตำแหน่งและขนาดของตัวอักษรมาเป็นข้อมูลในการตัดสินใจว่า บรรทัดใดเป็นหัวข้อหลัก บรรทัดใดเป็นหัวข้อย่อยหรือเป็นข้อความในย่อหน้าเดียวกัน

การวิเคราะห์เอกสารภาพโดยวิธี Bottom-up ที่เรื่อนำมาแล้วนั้น สามารถนำมาเขียนเป็นขั้นตอนการทำงานดังแสดงในรูปที่ 4



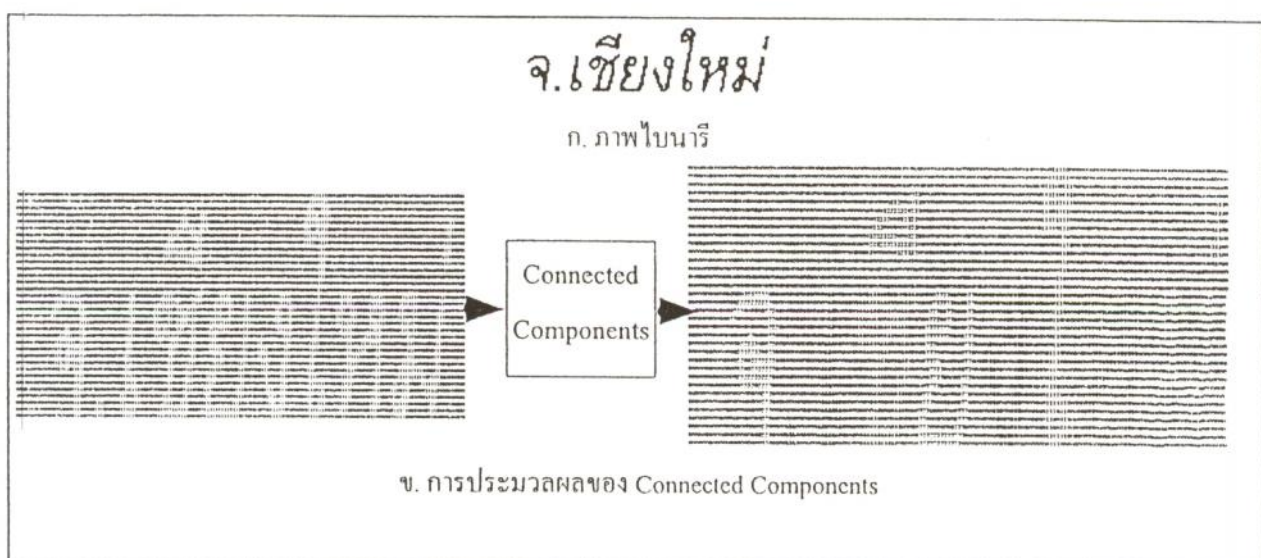
รูปที่ 4 ขั้นตอนการประมวลผลเอกสารภาพแบบ Bottom-up

ในหัวข้อถัดไปจะเป็นรายละเอียดของเทคนิคต่าง ๆ ที่นำมาใช้ในการวิเคราะห์เอกสารภาพแบบ Bottom-up ดังปรากฏในรูปที่ 4

CONNECTED COMPONENTS

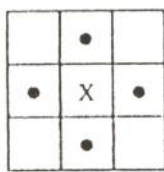
วิธีการดำเนินการของ Connected Components

มีอินพุตเป็นข้อมูลไบนารี และผลลัพธ์ที่ได้จากการประมวลผลโดยวิธีนี้คือ สัญลักษณ์หรือวัตถุที่ถูก Label ดังแสดงในรูปที่ 5 วัตถุแต่ละชิ้นจะถูก Label ด้วยเลขจำนวนเต็มค่าหนึ่งและวัตถุแต่ละชิ้นจะประกอบด้วยกลุ่มของจุดภาพไบนารีที่มีค่าเป็น 1 เชื่อมต่อกัน ซึ่งจะมีขนาดและรูปร่างขึ้นอยู่กับวัตถุ

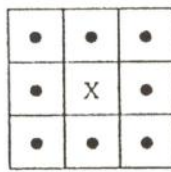


รูปที่ 5 ไดอะแกรมของ Connected Components

สำหรับอัลกอริทึมของ Connected Components มีผู้เสนอไว้หลายวิธีด้วยกัน² ซึ่งแต่ละอัลกอริทึมจะมีความแตกต่างกันในส่วนของการประสิทธิภาพในการทำงาน โดยเฉพาะความเร็วในการประมวลผล แต่สำหรับหลักการทำงานแล้วจะคล้ายกันคือ จะพิจารณาจากการเชื่อมต่อของจุดภาพในบริเวณใกล้เคียงภายในวินโดวขนาด 3x3 โดยมีการพิจารณาการเชื่อมต่อกันของจุดภาพอยู่สองแบบคือ แบบ 4 จุดกับแบบ 8 จุด ดังแสดงในรูปที่ 6



ก.



ข.

รูปที่ 6 การเชื่อมต่อของจุดภาพ ก. แบบ 4 จุด
ข. แบบ 8 จุด

สำหรับการวิเคราะห์เอกสารภาพ เราใช้การเชื่อมต่อของจุดภาพแบบ 8 จุด เพื่อลดเซตย่อยของวัตถุให้น้อยลง และขั้นตอนการทำงานของ Connected Components ได้แสดงไว้ในอัลกอริทึมที่ 1

อัลกอริทึมที่ 1 แสดงขั้นตอนการประมวลผลของ Connected Components

1. กำหนดค่าเริ่มต้นที่ไม่ซ้ำกันให้กับจุดภาพที่มีค่าเป็น 1
2. พิจารณาการเชื่อมต่อของจุดภาพในบริเวณใกล้เคียงทั้งภาพ จากข้างบนลงล่างดังนี้

```

for(int I=0; I<RowSize; I++)
    for(int J=0; J<ColSize; J++)
        if (LABEL[I][J])
            {

```

```

L = Min(Neighbors(I,J));
LABEL[I][J] = L;
    }

```

3. ตรวจสอบการเชื่อมต่อของจุดภาพในบริเวณใกล้เคียงทั้งภาพ จากข้างล่างมาข้างบนดังนี้

```

for(int I = RowSize; I>=0;I--)
    for(int J=ColSize; J>=0; J--)
        if(LABEL[I][J])
            {
                L = Min(Neighbors(I,J));
                LABEL [I][J] = L;
            }

```

โครงสร้างข้อมูลกราฟ

ข้อมูลของวัตถุแต่ละชิ้นที่แยกออกมาได้จากขั้นตอนของ Connected Components จะสามารถคำนวณตำแหน่งและค่าถ่วงน้ำหนักที่เป็นค่าระยะทางระหว่างวัตถุแต่ละชิ้น จากภาพไบนารีในรูปที่ 2 ที่ผ่านการประมวลผลของ Connected Components และสามารถระบุตำแหน่งของอักษรแต่ละตัว พร้อมกับคำนวณระยะทางระหว่างตัวอักษร ดังได้แสดงในตารางข้างล่างในรูปที่ 7 ค. สำหรับสมการที่ใช้ในการคำนวณระยะทางได้แสดงไว้ในสมการที่ (1)⁶



ก.

Label	x_1	y_1	x_2	y_2
1	259	20	316	35
3	30	22	82	39
18	190	42	249	98
21	255	42	319	98
22	143	43	183	99
25	12	44	64	99
27	72	44	173	99

ข.

LABEL	1	3	18	21	22	25	27
1	0	177	10	7	76	195	122
3	177	0	108	173	61	5	5
18	10	108	0	6	7	126	53
21	7	173	6	0	72	191	118
22	76	61	7	72	0	79	6
25	195	5	126	191	79	0	8
27	122	5	53	118	6	8	0

ค.

รูปที่ 7 แสดงข้อมูลที่จะใส่เข้าไปในโครงสร้างกราฟชนิดไม่ระบุทิศทาง ก. แสดงจุดที่ใช้บอกตำแหน่งของวัตถุ ข. ตารางแสดงตำแหน่งของวัตถุแต่ละ Label ค. แสดงระยะห่างระหว่างวัตถุแต่ละชิ้น

$$D(i,j) = \text{Max} (Dx(i,j), Dy(i,j)) \quad (1)$$

ที่

$D(i,j)$ คือระยะห่างระหว่างวัตถุที่ i , กับวัตถุที่ j

$Dx(i,j)$ คือระยะห่างระหว่างวัตถุที่ i , กับวัตถุที่ j ตามแนวแกน x

$Dy(i,j)$ คือระยะห่างระหว่างวัตถุที่ i , กับวัตถุที่ j ตามแนวแกน y

ข้อมูลที่น่าใส่โครงสร้างกราฟคือ ข้อมูลในตารางของรูปที่ 7ข. และ 7ค. และกราฟที่นำมาใช้ในที่นี้เป็นชนิดที่ไม่มีการระบุทิศทาง (Undirected Graph)

การวิเคราะห์ความเป็นอิสระกันของวัตถุแต่ละชิ้น

ในบางกรณีวิธี Connected Components ไม่สามารถเชื่อมโยงจุดภาพที่เชื่อมต่อเป็นวัตถุขึ้นเดียวกันด้วย Label เดียวกันทั้งวัตถุ ทำให้ตัวอักษรหนึ่งตัวถูกแยกออกเป็นสอง Label หรือมากกว่า ดังที่แสดงให้เห็นไปแล้วในรูปที่ 1 กรณีเช่นนี้เท่าที่ได้ทดลองมาจะเกิดขึ้นกับตัวอักษรบางตัวที่เอียง และขอบของเส้นตัว

อักษรไม่เรียบ สำหรับวิธีการแก้ปัญหานี้สามารถทำได้โดยการดึงวัตถุแต่ละชิ้น มาเปรียบเทียบตำแหน่งและขนาดของวัตถุ ซึ่งถ้าค่าระยะทางระหว่างวัตถุที่ i กับวัตถุที่ j ($D(i, j)$ ในสมการที่ 1) มีค่าเป็นศูนย์ แสดงว่าวัตถุที่ i กับวัตถุที่ j ไม่เป็นอิสระต่อกัน และการที่จะลบวัตถุที่ i หรือ j ทำได้โดยการเปรียบเทียบขนาดของวัตถุ ถ้าวัตถุชิ้นใดมีขนาดเล็กกว่าก็จะลบวัตถุชิ้นนั้นออกไป

การเชื่อมต่อของตัวอักษร

เมื่อวัตถุที่ซ้ำซ้อนหรือเป็นเซตย่อยของวัตถุชิ้นอื่นถูกลบออกไปหมดแล้ว ก็จะเหลือเฉพาะวัตถุที่ไม่ซ้อนทับกันแต่เรายังไม่ทราบการเชื่อมโยงกันของวัตถุแต่ละชิ้นว่าวัตถุหรือตัวอักษรใดอยู่ติดกันบ้าง ถ้าเราสังเกตในรูปที่ 7ค. ระหว่างตัวอักษรทุกตัวจะมีค่าถ่วงน้ำหนัก ซึ่งเป็นค่าระยะทางระหว่างตัวอักษรเชื่อมต่อกันทั้งหมด เทคนิคที่นำมาใช้หาการเชื่อมโยงว่าตัวอักษรใดบ้างที่อยู่ติดกันคือ Minimum-Cost Spanning Trees ซึ่งเป็นวิธีหาระยะทางที่สั้นที่สุดระหว่างวัตถุแต่ละชิ้น และ

อัลกอริทึมที่นำมาใช้กำหนดระยะทางที่สั้นที่สุดที่นำมาใช้ในที่นี่คือ อัลกอริทึมของ Kruskal^{3,4,5} อินพุตของอัลกอริทึมนี้ก็คือ เมตริกในรูปที่ 7ค. เมื่อการประมวลผลของอัลกอริทึมนี้สิ้นสุดลง กราฟที่ได้จากอัลกอริทึมนี้ได้แสดงไว้แล้วในรูปที่ 2ข. และตัวเลขที่ปรากฏอยู่ระหว่างโหนดแต่ละโหนด เป็นระยะทางที่ห่างกันน้อยที่สุด และผลลัพธ์ที่อยู่ในรูปของเมตริกได้แสดงไว้ในรูปที่ 8 ที่แสดงถึงการเชื่อมต่อกันของวัตถุในแต่ละ Label ซึ่งจะช่วยในการเรียงลำดับวัตถุและการกำหนดบรรทัด

LABEL	1	3	18	21	22	25	27
1	0	0	0	7	0	0	0
3	0	0	0	0	0	5	5
18	0	0	0	6	7	0	0
21	7	0	6	0	0	0	0
22	0	0	7	0	0	0	6
25	0	5	0	0	0	0	0
27	0	5	0	0	0	6	0

รูปที่ 8 เมตริกที่ใช้แทนการเชื่อมต่อระหว่างโหนด

การกำหนดบรรทัดและเรียงลำดับวัตถุเสียใหม่

ในการกำหนดบรรทัด สิ่งแรกที่ต้องกำหนดคือ Base-Line ซึ่งได้แสดงไว้แล้วในรูปที่ 3 เส้น Base-Line กำหนดได้โดยใช้จุด y_2 (ในรูปที่ 7ก.) ของวัตถุแต่ละชิ้น แต่จุดนี้สำหรับภาษาไทยแล้ว ในหนึ่งบรรทัดอาจจะมีตัวอักษรได้สูงสุดถึงสี่ระดับ (รูปที่ 9) ดังนั้นการใช้จุด y_2 ของวัตถุแต่ละชิ้นมากำหนด Base-Line โดยไม่มีการวิเคราะห์อย่างอื่นเพิ่มเติม ก็จะได้เส้นกำหนดบรรทัดดังที่แสดงในรูปที่ 9

สำหรับการกำหนด Base-Line ที่แท้จริงของแต่ละบรรทัด สามารถทำได้โดยการคำนวณความสูงของตัวอักษรของ Base-Line ในแต่ละระดับ และนำความสูงที่ได้มาคำนวณค่าความถี่สะสมเพื่อคำนวณค่าฐานนิยม และ Base-Line ที่แท้จริงคือ Base-Line ที่มีค่าฐานนิยมสูงที่สุด

เมื่อกำหนดบรรทัดได้แล้ว เราสามารถจัดลำดับของตัวอักษรเสียใหม่ เพื่อให้ได้ใจความเหมือนเอกสารเดิมได้โดยใช้จุด x_2 บน Base-Line เป็นเกณฑ์ในการพิจารณา คือ x_2 ที่มีค่าน้อยที่สุดก็จะเป็นอักษรตัวแรกของบรรทัด และอักษรตัวต่อไปก็จะมีค่า x_2 มากขึ้นไปเรื่อยๆ ส่วนตัวอักษรในระดับอื่นๆ คือ ระดับที่ 2 ระดับที่ 3 และระดับที่ 4 จะกำหนดลำดับได้โดยใช้ค่าระยะทางที่สั้นที่สุด ซึ่งเป็นตัวเชื่อมต่อที่ได้กำหนดไว้แล้วในหัวข้อก่อนหน้านี้



รูปที่ 9 แสดงตัวอักษรที่มีอยู่ 4 ระดับ

การวิเคราะห์ย่อหน้า

การวิเคราะห์ย่อหน้า เป็นการจับเก็บบล็อกของวัตถุต่างๆ ที่ได้แยกไว้แล้ว ให้มีความหมายเหมือนกับเอกสารเดิม ขั้นตอนการดำเนินการต่างๆ ในหัวข้อนี้ส่วนมากจะเป็นการคำนวณขนาดของตัวอักษร ระยะห่างระหว่างตัวอักษร ตำแหน่งตำแหน่งที่มีการย่อหน้า แล้วนำค่าเหล่านี้ไปเป็นเงื่อนไขในการจัดย่อหน้า โดยเปรียบเทียบขนาดของตัวอักษรคือ ขนาดของตัวอักษรปกติที่อยู่ในย่อหน้าจะมีจำนวนมากที่สุด ส่วนขนาดของตัวอักษรที่ใหญ่กว่าก็จะเป็นหัวข้อ และวัตถุที่มีขนาดใหญ่กว่าตัวอักษรปกติก็就会被แยกไปเป็นรูปภาพ

สำหรับการคำนวณขนาดของตัวอักษรจะใช้ความสูงของตัวอักษรเป็นหลัก ซึ่งสามารถคำนวณได้ดังนี้

$$h_i = \text{Label}[i].y_2 - \text{Label}[i].y_1 \quad (2)$$

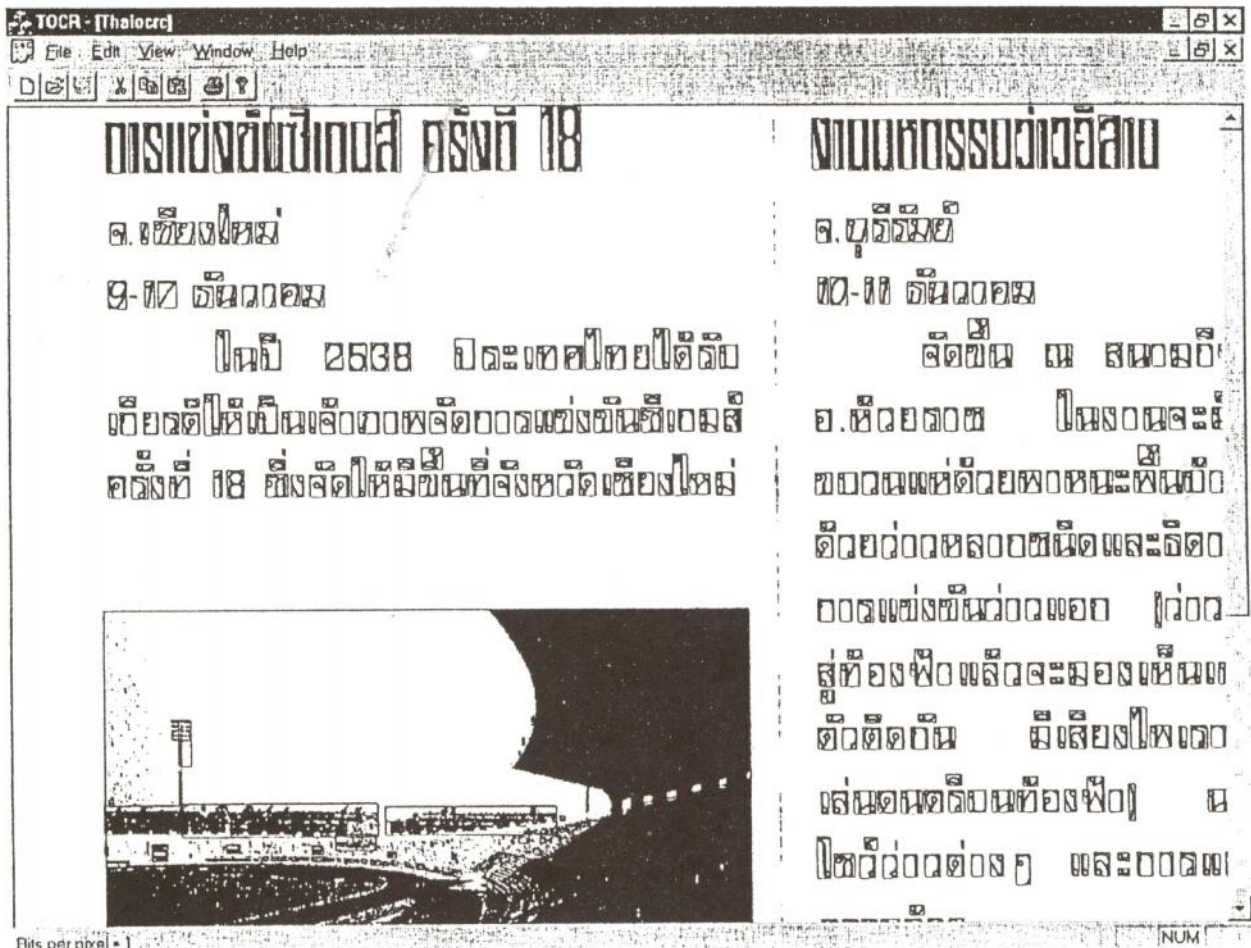
ที่ h_i คือความสูงของตัวอักษร Label ที่ i ส่วนระยะห่างระหว่างอักษรตัวที่ i กับตัวที่ j สามารถกำหนดได้ดังนี้

$$s_{ij} = \text{Label}[j].x_2 - \text{Label}[i].x_1 \quad (3)$$

ผลการทดลอง

ในการทดลองใช้โมเดลวิเคราะห์เอกสารแบบ Bottom-up ที่ได้เสนอไปแล้วนั้น เราได้ใช้ภาพไบนารีที่มีความละเอียดของภาพประมาณ 100 จุดต่อตารางนิ้ว โดยใช้สแกนเนอร์สแกนภาพไบนารีเข้ามา และเก็บไฟล์ที่สแกนเข้ามาในรูปแบบกราฟิกไฟล์ของไมโครซอฟท์ วินโดวส์ ในการประมวลผลวิเคราะห์เอกสารภาพ จะอ่านไฟล์นี้เข้ามา และโปรแกรมจะเริ่มทำการประมวลผลตามขั้นตอนต่างๆ ที่ได้เสนอไปแล้วตามในรูปที่ 4 จนได้ผลลัพธ์ออกมาดังแสดงไว้ในรูปที่ 11

ข้อมูลภาพที่นำมาทดสอบในรูปที่ 11 มีสองภาพด้วยกัน ภาพแรกคือรูปที่ 11ก. เป็นภาพที่สแกนมาจากนิตยสารที่มีความละเอียด 75x75 dpi ขนาด 1213x1348 จุดภาพ ชนิดของไฟล์เป็น Ms Windows 3 Bitmap ในภาพจะมีตัวอักษรสองขนาด และแบบของตัวอักษรมีทั้งที่เป็นตัวธรรมดาและตัวเอียง ส่วนรูปที่ 11ข. เป็นภาพที่สแกนมาจากโลโก้ที่มีความละเอียด 75x75 dpi ขนาด 1224x434 จุดภาพ ชนิดของไฟล์เป็นชนิดเดียวกับรูปที่ 11ก. ในภาพจะมีตัวอักษรสองขนาดแต่แบบของตัวอักษรจะมีเฉพาะตัวธรรมดา

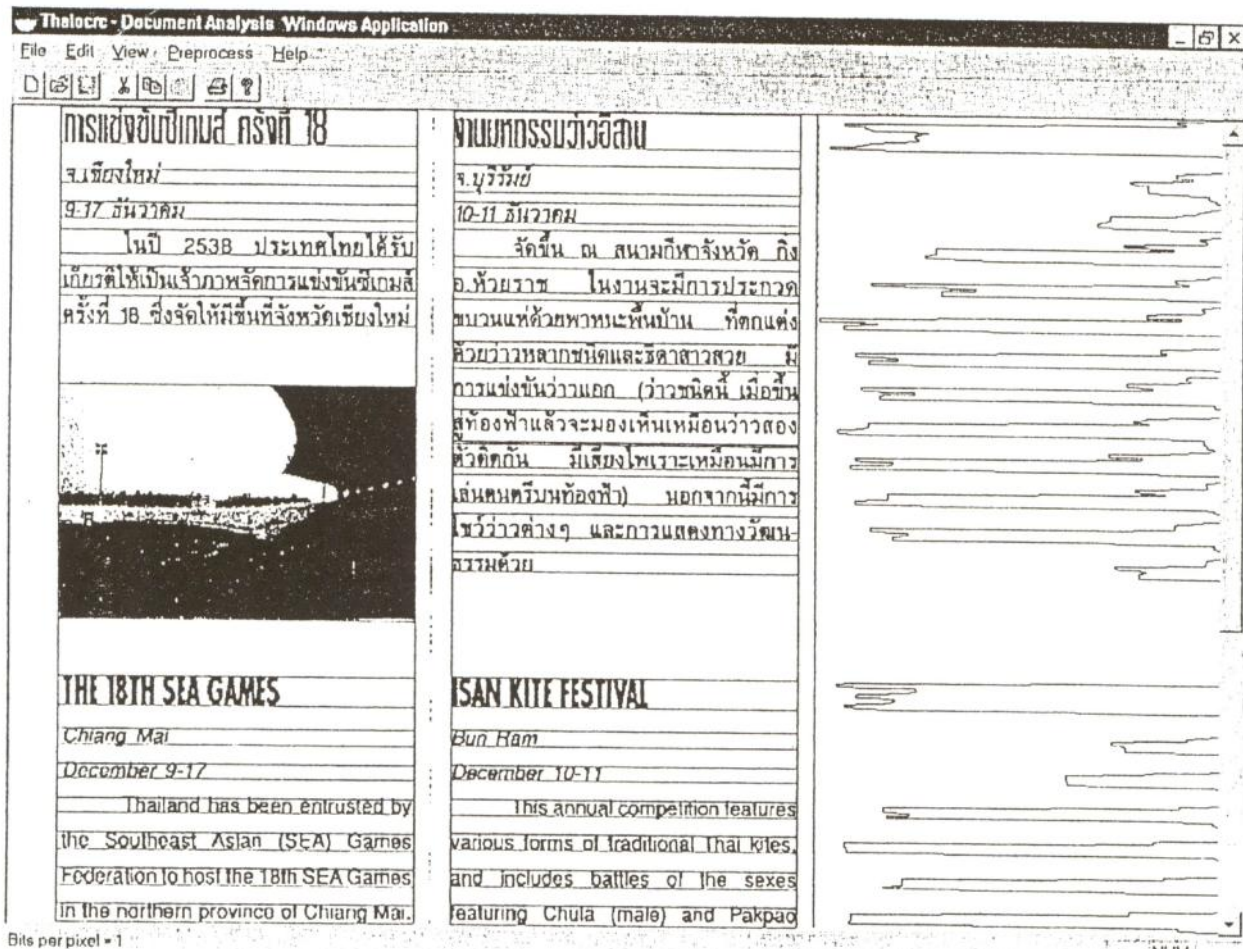


ก.

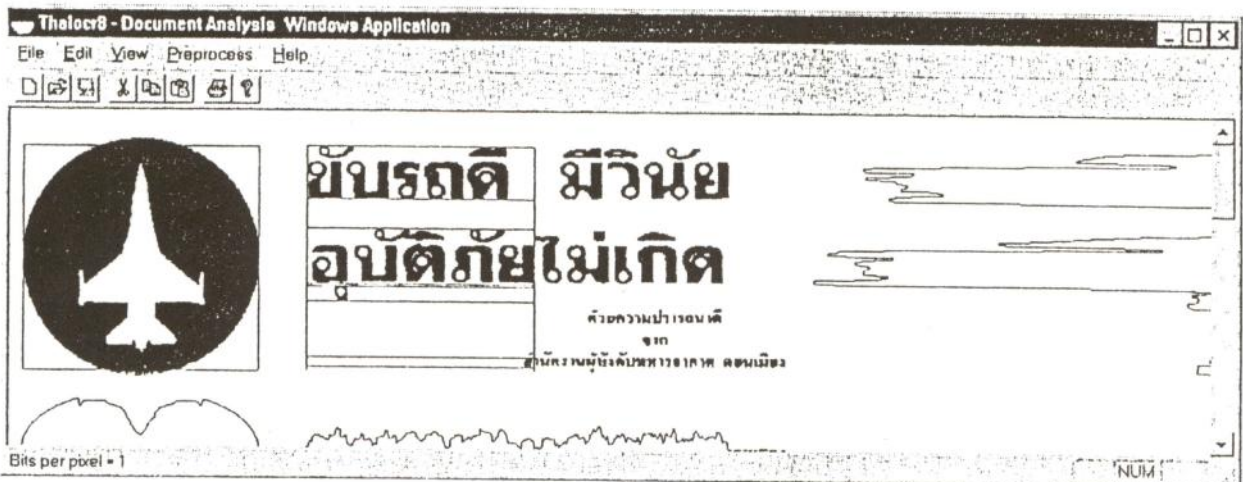


ข.

รูปที่ 11 แสดงภาพผลลัพธ์ที่ได้จากการวิเคราะห์เอกสารโดยวิธี Bottom-up



ก.



ข.

รูปที่ 12 แสดงภาพผลลัพธ์ที่ได้จากการวิเคราะห์เอกสารโดยวิธี Top-Down

ในรูปที่ 12 เป็นข้อมูลภาพอินพุตชุดเดียวกันกับรูปที่ 11 แต่ใช้วิธีการประมวลผลแบบ Top-Down เมื่อเปรียบเทียบผลลัพธ์ที่ได้ระหว่างสองวิธีนี้ ปรากฏว่า

วิธี Bottom-up ที่ใช้ Connected Components จะให้ผลลัพธ์ที่ดีกว่าคือสามารถใช้กับเอกสารที่มีรูปแบบหลากหลาย และให้ข้อมูลที่ได้จากการวิเคราะห์ที่

ละเอียดถูกต้องกว่าวิธี Top-Down ที่ใช้เทคนิค Projection Profile⁷ ดังผลลัพธ์ของทั้งสองวิธีที่ได้แสดงเปรียบเทียบกัน ส่วนข้อดีของวิธี Top-Down ก็คือมีขั้นตอนการประมวลผลที่มีความซับซ้อนน้อยกว่า จึงทำให้เวลาที่ใช้ในการประมวลผลน้อยกว่าวิธี Bottom-up

ในการเปรียบเทียบผลลัพธ์ระหว่างวิธี Bottom-up และวิธี Top-down ถ้าเราสังเกตรูปที่ 11ก. กับรูปที่ 12ก. จะเห็นว่าทั้งสองวิธีสามารถจัดรูปแบบย่อหน้าได้พอๆ กัน แต่ถ้าเราวิเคราะห์รูปที่ 11ข. กับรูปที่ 12ข. กันดู จะพบว่าภาพนี้มีรูปแบบการจัดย่อหน้าที่เปลี่ยนไปทำให้ผลลัพธ์ที่ได้จากวิธี Top-down ไม่สมบูรณ์คือ สองบรรทัดที่ตัวอักษรขนาดเล็กจะขาดหายไป ส่วนวิธี Bottom-up ยังคงจัดย่อหน้าได้ถูกต้อง ที่เป็นเช่นนี้ก็เพราะว่า วิธี Top-down จะต้องออกแบบอัลกอริทึมตามชนิดหรือรูปแบบของเอกสาร

ในการวัดประสิทธิภาพของวิธีที่เสนอ กรณีของความเร็วในการประมวลผล จากการทดสอบบนเครื่องคอมพิวเตอร์ Intel Pentium ความเร็ว 133 MHz แสดงไว้ในตารางข้างล่าง สำหรับความถูกต้องของผลลัพธ์ที่ได้ของวิธีที่เสนอโดยเฉลี่ยประมาณ 90% โดยไม่นับตัวอักษรที่ติดกัน และเอกสารที่สแกนจะต้องวางให้มีมุมเอียงไม่เกิน 10 องศา

ตารางแสดงความเร็วในการประมวลผล
(หน่วยเป็น วินาที)

รูป	Bottom-up	Top-down
ก.	25.32	4.07
ข.	2.14	1.76

สรุป

การวิเคราะห์เอกสารภาษาไทยโดยใช้วิธี Bottom-up ที่ได้เสนอไปแล้วนั้น เป็นการวิเคราะห์จากระดับล่างสุดขึ้นมาคือ จากจุดภาพ มาเป็นตัวอักษร คำ วลี ประโยค ย่อหน้า และเอกสารในแต่ละหน้า

สำหรับเทคนิคที่นำมาใช้ในการวิเคราะห์ในโมเดลที่เสนอคือ Connected Components ที่นำมาใช้แยกวัตถุแต่ละชิ้น เทคนิคการวิเคราะห์การเชื่อมโยงตัวอักษรด้วยการหาระยะทางที่สั้นที่สุดหรือ Minimum-Cost Spanning Trees โดยใช้อัลกอริทึมของ Kruskal และวิธีที่ได้เสนอไปแล้วนั้น เป็นวิธีที่ใช้ได้กับเอกสารหลากหลาย และสามารถแยกตัวอักษรที่แต่ละตัวที่ไม่ติดกันได้ด้วย แต่ข้อเสียของวิธีนี้เท่าที่พบมาคือมีขั้นตอนการประมวลผลที่ซับซ้อนจึงต้องใช้เวลาในการประมวลผลมาก และวิธีที่เสนอยังไม่ได้ทำในส่วนของการวิเคราะห์ตัวอักษรที่ติดกัน (Touching Character)

เอกสารอ้างอิง

1. Yuan Y. Tang, Seong-Whan Lee and Ching Y. Suen, "Automatic Document Processing: A Survey," *Pattern Recognition*, Vol. 29, No.12, pp. 1931-1952, 1996.
2. Robert M. Haralick and Linda G. Shapiro, *Computer and Robot Vision: Volume I*, Addison-Wesley, USA, 1992.
3. Clifford A. Shaffer, *A Practical Introduction to Data Structures and Algorithm Analysis*, Prentice-Hall, USA, 1997.
4. Gould Ronald, *Graph Theory*, Benjamin/Cummings, USA, 1988.
5. Robert Sedgewick, *Algorithms in C++*, Addison-Wesley, USA, 1992.
6. Steven Harrington, *Computer Graphics A Programming Approach*, second edition, McGraw-Hill, Singapore, 1987.
7. Akiyama T. and Hagita N., "Automated entry system for printed documents", *Pattern Recognition*, Vol. 23, No. 11, pp. 1141-1154, 1990.