บทที่ 1

บทนำ

ในปัจจุบันเราจะพบว่า เอกสารที่มีส่วนใหญ่อยู่ในรูปสิ่งพิมพ์กระดาษ เช่น หนังสือ บทความ คู่มือ จดหมาย ทะเบียนราษฎร์ แฟ้มประวัติ ฯลฯ ซึ่งถ้ามีจำนวนมากจะมีปัญหาทางด้านการจัดเก็บ อีกทั้งอายุการใช้งานของเอกสารเหล่านี้ก็สั้นด้วย (เสื่อมสภาพเร็ว) การนำเอกสารเก็บลงในคอมพิวเตอร์ (computerization) จึงเป็นทางออกที่ดีวิธีหนึ่ง สามารถทำได้โดยแบ่งเป็นวิธีใหญ่ๆ ดังนี้

1. พิมพ์ข้อความในเอกสารนั้นใหม่ทั้งหมดแล้วเก็บเอกสารในรูปตัวหนังสือ (text) วิธีนี้จะใช้เวลาในการทำมาก และมีปัญหาเรื่องรูปภาพ การจัดหน้า ตำแหน่งของรูปภาพกับตัวหนังสือ ฯลฯ

2. สแกนเอกสารที่มีอยู่แล้วและใช้การรู้จำตัวอักษร (optical character recognition, OCR) เพื่อเปลี่ยนข้อมูลภาพเป็นตัวหนังสือ ซึ่งจะใช้เนื้อที่ในการเก็บน้อยกว่าการเก็บแบบภาพ และสามารถค้นหาคำได้ แต่ปัญหาของ OCR คือ เรื่องความถูกต้อง และไม่สามารถจัดการได้กับเอกสารที่ประกอบด้วยอักษรพิเศษ ภาษาต่างประเทศ สมการคณิตศาสตร์ และมีปัญหาเกี่ยวกับการจัดตำแหน่งและความสัมพันธ์ของโครงสร้างเอกสาร เช่น ในกรณีที่เอกสารมีรูปภาพ หรือตารางประกอบ หรือมีหลายคอลัมน์

3. สแกนเอกสารที่มีอยู่แล้วและเก็บเป็นรูปภาพ (image) และใช้ OCR เพื่อเปลี่ยนข้อมูลภาพเป็นตัวหนังสือให้สามารถค้นหาคำทั้งหมดได้ภายในเอกสาร (แสดงเอกสารด้วยภาพ)

4. สแกนเอกสารที่มีอยู่แล้วและเก็บเป็นรูปภาพ (image) และสร้างดัชนีภาพของคำสำหรับค้นหา

5. สแกนเอกสารที่มีอยู่แล้วและเก็บเป็นรูปภาพ (image) อย่างเดียว วิธีนี้สามารถทำได้อย่างรวดเร็ว และสามารถเก็บส่วนที่เป็นรูปภาพกับข้อความรวมกันได้เหมือนต้นฉบับจริงมากที่สุด แต่มีจุดอ่อนคือใช้เนื้อที่มาก การค้นหาคำทำไม่ได้เลย

กรณีที่มีเอกสารจำนวนมาก และความต้องการในการจัดเก็บที่รวดเร็ว ประกอบกับปัญหาของ OCR ที่ได้กล่าวไว้ การเก็บเอกสารโดยเก็บเป็นรูปภาพจึงเป็นทางออกที่เหมาะสมกว่า แต่ปัญหาเรื่องความต้องการเนื้อที่ในการเก็บมาก ตัวอย่างเช่นในการเก็บเอกสาร 1 หน้าขนาด A4 (8.25x11.5นิ้ว) ถ้าสแกนด้วยความละเอียด 300 จุดต่อนิ้ว โดยไม่มีการบีบอัดต้องใช้ความจุถึง 1

MB สำหรับภาพสองระดับ (ขาวดำ) จึงจำเป็นต้องนำวิธีบีบอัดข้อมูลภาพ (image compression) ที่มีประสิทธิภาพมาใช้

เนื่องจากระบบการจัดเก็บภาพเอกสารที่มีใช้อยู่ในเชิงพาณิชย์เป็นระบบปิด การใช้งานขาดความยืดหยุ่น และวิธีการในการบีบอัดข้อมูลของภาพยังคงใช้อัลกอริธึม MH,MR ในมาตรฐาน CCITT Recommendation T.4(fax group 3) หรืออัลกอริธึม MMR ในมาตรฐาน CCITT Recommendation T.6(fax group 4) เป็นส่วนใหญ่ เช่น Adobe Acrobat Capture [1] เอกสารที่สร้างได้จะอยู่ในรูป PDF (Portable Document Format) โดยส่วนที่เก็บเป็นภาพจะอยู่ในรูป TIF ทั้งๆ ที่มีมาตรฐาน *JBIG*, Joint Bi-level Image experts Group (ITU-T Recommendation T.82) ออกมาซึ่งให้อัตราในการบีบอัดข้อมูลดีกว่า[2] ทั้งนี้เพราะมาตรฐาน JBIG ใช้อัลกอริธึม *Q-Coder* [3] ที่ได้รับการพัฒนาโดยบริษัท IBM ซึ่งได้รับการจดสิทธิบัตรสำหรับอัลกอริธึมในประเทศสหรัฐอเมริกา ทำให้ไม่มีผู้พัฒนาโปรแกรมที่ใช้มาตรฐานดังกล่าวออกมา แต่กฎหมายสิทธิบัตรของประเทศอื่นๆ ทั้งในยุโรปและญี่ปุ่นรวมถึงประเทศไทยไม่รับรองการจดสิทธิบัตรแก่อัลกอริธึมของตัวซอฟต์แวร์ (รับรองเฉพาะลิขสิทธิ์) โครงการนี้จึงต้องการพัฒนาโปรแกรมที่ใช้มาตรฐานดังกล่าวขึ้นมาเอง

ปัญหาสำคัญประการหนึ่งของการเก็บเอกสารแบบรูปภาพคือการค้นหาคำทำไม่ได้ จึงจำเป็นต้องมีการสร้างฐานข้อมูลแยกต่างหาก โครงการนี้ได้เสนอวิธีเก็บข้อมูลในส่วนคำสำคัญเป็นรูปภาพเพื่อใช้หาตำแหน่งของคำสำคัญทั้งหมดที่ปรากฏอยู่ในเอกสาร และลดภาระของผู้จัดทำเอกสารในการป้อนข้อมูล

ข้อดีของการแปลงเอกสารที่มีอยู่ ให้อยู่ในรูปสื่ออิเล็กทรอนิกส์ประการหนึ่งคือ สามารถเผยแพร่เอกสารผ่านทางเครือข่ายคอมพิวเตอร์ได้ง่าย เช่นระบบอินเตอร์เน็ตซึ่งกำลังเป็นที่นิยมอยู่ในขณะนี้

ในปัจจุบันเราจะพบว่า ข้อมูลส่วนใหญ่ในระบบอินเตอร์เน็ตจะเป็นภาษาอังกฤษ ความรู้เกี่ยวกับเรื่องไทยๆ ยังคงมีน้อยทั้งๆที่เรามีแหล่งความรู้ที่อยู่ในรูปสิ่งพิมพ์อยู่มาก โครงงานนี้จะช่วยให้สามารถเผยแพร่ข้อมูลของไทยได้ดีขึ้น

วัตถุประสงค์

1. เพื่อศึกษาและนำวิธีบีบอัดข้อมูลภาพตามมาตรฐาน JBIG มาใช้งาน

2. เพื่อพัฒนาระบบจัดเก็บและเรียกคืนภาพเอกสาร

3. พัฒนาโปรแกรมคอมพิวเตอร์เพื่อเผยแพร่ข้อมูลเอกสารที่มีอยู่ผ่านทางระบบอินเตอร์เน็ต

ขอบเขตของวิทยานิพนธ์

สร้างระบบจัดเก็บภาพเอกสารที่ประกอบด้วยส่วนจัดเก็บ และส่วนเรียกคืนภาพเอกสาร

ส่วนจัดเก็บเอกสาร

สร้างโปรแกรมสำหรับจัดทำตัวเอกสาร โดยให้คนป้อนข้อมูลที่จำเป็น และตัวโปรแกรมจะทำการสร้างเอกสารในรูป HTML

เอกสารที่ได้จะต้องเอื้ออำนวยต่อการใช้งาน ผู้ใช้สามารถค้นข้อมูลได้ระดับหนึ่ง

ส่วนเรียกคืนภาพเอกสาร

สร้างโปรแกรมเพื่อใช้แสดงภาพเอกสารที่อยู่ในรูปของ JBIG โดยสามารถแสดงภาพที่ความละเอียดต่างๆกันได้ และมีความเร็วในการทำงานเพียงพอ

ทำ plug-in เพื่อให้โปรแกรม Netscape Navigator สามารถแสดงภาพเอกสารที่อยู่ในรูปของ JBIG ได้

ประโยชน์ที่คาดว่าจะได้รับ

1. ได้ระบบจัดเก็บและเรียกคืนภาพเอกสารที่สามารถนำมาใช้งานได้

2. สามารถผลิตเอกสารในรูปสื่ออิเล็กทรอนิกส์ จากสิ่งพิมพ์กระดาษที่สามารถค้นหาคำได้บางส่วน

3. เป็นการเผยแพร่ความรู้ที่มีอยู่ในรูปสิ่งพิมพ์เอกสารผ่านทางระบบอินเตอร์เน็ตได้อย่างสะดวก และใช้เวลาจัดทำสั้น